期刊報紙數位化工作流程指南

數位典藏與數位學習國家型科技計畫 拓展台灣數位典藏計畫

致 謝

感謝國家圖書館、磁軒資訊媒體行銷有限公司、相關單位之計畫主持人以及參與之工作同仁，撥冗協助本計畫調查與訪談，使得本文能有更詳盡的資料，並彙整各單位經驗爲其他欲進行數位化作業同仁之參考，讓典藏品更能彰顯其品質，並在資訊流通上更爲無遠弗屆。特別感謝玄奘大學資訊傳播學院院長郭良文教授，擔任此數位化工作流程指南評論人。本計畫主持人林富士先生及共同主持人邱澎生先生於本文撰寫期間，提供諸多鼓勵與指導，以及計畫辦公室同仁熱情協助，在此一併致謝。

出版序

「數位典藏國家型科技計畫」於西元2002年開始執行，衆多機構計畫與公開徵選計畫的工作夥伴紛紛加入我們的團隊，進行種類繁多而又數量鉅大的數位化工作，第一期五年計畫於西元2006年圓滿結束。次年，即與「數位學習國家型科技計畫」整合成爲「數位典藏與數位學習國家型科技計畫」(TELDAP, http://teldap.tw/)，以「呈現台灣的文化與自然多樣性」爲總體目標，繼續拓展數位典藏與數位學習內容，並更有系統地往教育、研究與產業等面向推廣數位典藏與數位學習計畫的成果；同時，也希望能更積極地結合民間力量，推動相關產業的應用與成長，既保存我國重要的文化資產，也促成數位時代新文化的創造。

做爲「數位典藏與數位學習國家型科技計畫」的分項計畫，我們也由第一期的名稱「內容發展分項計畫」改名做「拓展台灣數位典藏計畫」(http://content.teldap.tw)，更積極地拓展數位內容的來源，向民間公私立單位甚至是個人的收藏品，廣泛徵集有關檔案、考古、語言、地理、族群、藝術、民間生活與動物、植物等數位化的計畫，並努力促成這些有關自然與人文不同性質的數位內容能做更好的整合，製作成兼具趣味性與啓發性的數位典藏素材，既供民衆免費下載進行教育與研究之用，也便利廠商與公私典藏者發現彼此在商業加值方面的合作機會。「拓展台灣數位典藏計畫」與「數位典藏與數位學習國家型科技計畫」其他分項計畫的相互協力，將加速我國數位內容由典藏保存跨入教育、研究與商業加值的過程，以期呈現台灣的文化與自然多樣性，並讓更多國內民衆與國際人士體會並珍視我國歷史文化之富盛與自然生態之茂美。

在典藏與加值數位內容的同時，無論是於「內容發展分項計畫」或是於「拓展台灣數位典藏計畫」時期，本計畫同仁都針對公私立機關與公開徵選計畫等工作夥伴從事各類物件數位化的工作流程及相關技術進行調查與記錄，並且結合各項數位化技術與工作流程相關的國際標準，編撰成爲

一系列的「數位化工作流程指南叢書」。自西元2005年以來，我們即先精選諸如瓷器、書畫、古籍等單一種類的數位化物件，綜合不同典藏計畫從事此項單一物件數位化的工作經驗，並輔以國內外的相關理論與實務成果，陸續撰寫了21冊不同主題的數位化工作流程指南（可自「拓展台灣數位典藏」網站「虛擬圖書館：數位化書籍」欄位下載全部21冊的全文電子檔）。

自去年以來，我們即準備修訂並擴充這套「數位化工作流程指南叢書」，希望增加流通管道，以供更多博物館、圖書館、機構與個人參考。我們的準備工作，主要分爲修訂既有「精選物件」指南以及新撰「共通原則」指南兩方面：前者指的是修訂既有的21冊工作流程指南，特別是針對數位化新技術與規範的引進、更實用的軟硬體設備，以及數位內容保護機制等層面做修訂，預訂每年修訂出版七本專書，並於三年內出版完成。至於新編的「共通原則」指南，則重點在於導入數位資訊「生命週期」與品質管理等關鍵概念，以「跨物件」而非單一精選物件爲探究對象，採用共通原則做爲架構該指南的數位化工作流程內容；這裏所謂的共通原則，指的是諸如專案管理、工作流程管理、圖像管理、影音管理、文字管理、色彩管理、委外製作和國外資源分析等，這八個共通原則都成爲我們調查、研究與撰寫指南的主題內容，預計在三年間陸續出版這八本指南。

在我們的規劃理念上，精選物件指南與共通原則指南其實彼此間具有一種相輔相成的關係：共通原則指南著重在對數位化工作的各項重要主題做分析，引導讀者對數位化的利弊得失做通盤而深入的思考；精選物件指南則描述特定物件的數位化實務與技術，便利讀者針對單一物件選擇最合適、最有效益的數位化工作流程。透過這套「數位化工作流程指南叢書」的出版，相信可爲更多有志投入數位化工作的單位與個人，提供一套富有整體性思惟並且又能循序漸進的實用指南。要特別強調的是：這套叢書的主要立論基礎，仍在於多年來陸續加入我們的機構與公開徵選計畫工作團隊多年來所累積的各種寶貴經驗，這些經驗讓更多的數位內容可以用更精

緻的品質，以及更效率的成本來製成、展示與維護，從而也豐富了我國的數位典藏與數位學習事業。在陸續出版這套「數位化工作流程指南叢書」的同時，我們要謝謝接受訪問的工作夥伴以及參與寫作的同仁，也要衷心感謝協助我們審查與諮詢這些數位化工作流程指南的學者專家。最後，也盼望讀者隨時給我們指正與建議，讓我們的工作可以做得更好。

數位典藏與數位學習國家型科技計畫
拓展台灣數位典藏計畫・數位內容建置與整合子計畫

計畫主持人  敬誌

中華民國 98年2月10日

目 錄 | CONTENTS

壹、引言

Introduction

行政院國家科學委員會自民國91年起，依據「數位博物館計畫」、「國家典藏數位化計畫」，以及「國際數位圖書館合作計畫」等三個計畫的合作經驗，整合規劃了「數位典藏與數位學習國家型科技計畫」計畫，其首要目標是將國家重要的文物典藏數位化，建立國家數位典藏知識網絡。計畫項下設有8分項計畫，分別爲：1.拓展台灣數位典藏計畫、2.數位技術研發與整合計畫、3.數位核心平台計畫、4.數位典藏與數位學習之學術社會應用推廣計畫、5.數位典藏與數位學習之產業發展與推動計畫、6.數位教育與網絡學習計畫、7.語文數位教學計畫、8.數位典藏與學習之海外推廣暨國際合作計畫。而其中「拓展臺灣數位典藏計畫」負責數位典藏內容之管理、規劃及各機構間的橫向聯繫、協調等事宜，並將各計畫的典藏品依照其性質分成各種主題小組，詳見圖1-1。

新聞主題工作小組於民國91年正式成立，以報紙、期刊、新聞影音爲主要數位化典藏內容，典藏品形態包含平面報刊媒體與電視媒體之文字、圖像、照片、影音等各項種類。

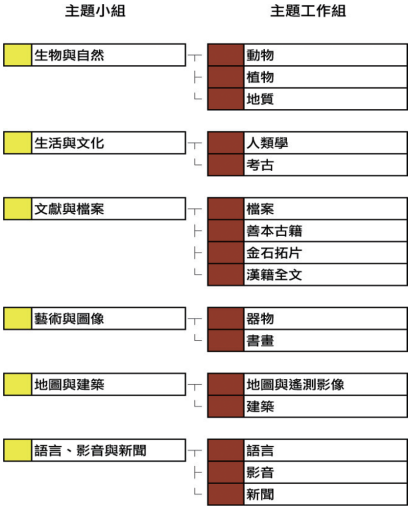


圖1-1、主題小組與工作組示意圖

歷年來參與其工作組進行數位化計畫的機構單位有：數位典藏國家型科技計畫「維運管理分項計畫—出版子計畫」（91年至94年度）、國家圖書館「國家圖書館期刊報紙典藏數位化計畫」（91年至95年度）、國立交通大學資訊工程系「電視新聞數位博物館」（91年度）、「世新大學北平世界日報內容數位開發計畫」（91年至94年度）、淡江大學「台灣棒球運動珍貴新聞檔案數位資料館之建置」（93年至95年度）、國立交通大學傳播研究所「蘭嶼原住民媒體資料庫建置與數位典藏計畫」（94年至96年度）、國立交通大學圖書館「雲門舞集舞作資產數位典藏計畫」（96年度迄今）、中央研究院社會學研究所「台灣『外省人』生命記憶與敘事資料庫」（96年度迄今）、 國立交通大學傳播研究所「達悟歌謠與庶民文化數位典藏計畫」（97年度）、淡江大學「台灣婦女新知運動史料數位典藏計畫－建置婦女新知雜誌與騷動季刊25年資料庫」（97年度）、國立政治大學廣播電視學系「《中國時報》新聞攝影底片之數位化—台灣政治民主化過程裡的政府與政黨新聞1988-2000」（97年度）。

以下簡略說明其內各計畫之數位化工作內容：

1. 維運管理分項計畫—出版子計畫主要負責《國家數位典藏通訊》發行，並以XML標誌語言加以分析進而建立檢索資料庫；
2. 國家圖書館則從事館藏之臺灣地區發行期刊約1,000種，與臺灣地區發行報紙約30種之數位化工作，其主要數位化工作項目爲期刊典藏影像數位化、報紙典藏數位化、期刊篇目後設資料分析建檔等；
3. 國立交通大學資訊工程系則有「電視新聞數位博物館」網路資料庫，典藏中華電視公司新聞影音資料；
4. 國立交通大學傳播研究所的典藏有蘭嶼在地刊物《蘭嶼雙週刊》、數位化幻燈片影像資料及蘭嶼地方廣播節目的聲音內容，並建置多媒體資料庫；
5. 世新大學資訊傳播學系則取得北平世界日報之微縮膠卷資料（報紙原件存放於北京圖書館），並陸續全文輸入典藏北平世界日報之新聞內容；

6.淡江大學與聯合報合作進行台灣棒球新聞之數位化，並建置「台灣棒球運動珍貴新聞檔案數位資料館」；

7.國立交通大學圖書館「雲門舞集舞作資產數位典藏計畫」則將國內外與雲門舞集相關的報導、評論...等以掃描的方式加以典藏；

8.中央研究院社會學研究所「臺灣外省人生命記憶與敘事資料庫」主要與社團法人外省台灣人協會合作典藏家書、返鄉照片、相關新聞資料以及社團法人外省台灣人協會與全台各地社區大學合作開辦蒲公英女性寫作班文章爲主，記錄保存和這個「外省人」類屬有關的記憶與生命敘事，並協助其生產出在地的、時代的文化意義與社會連結；

9.淡江大學「台灣婦女新知運動史料數位典藏計畫－建置婦女新知雜誌與騷動季刊25年資料庫」資料庫主幹將以台灣戰後婦運歷史最悠久的正式組織—婦女新知基金會之資料保存爲基礎。第一階段以婦女新知出版之雜誌《新知通訊》以及《騷動》季刊爲典藏對象，上述資料歷時25年，見證婦女新知在衆多性別議題的開創性位置，也提供了台灣婦運發展的記錄；

10.國立政治大學廣播電視學系「《中國時報》新聞攝影底片之數位化—台灣政治民主化過程裡的政府與政黨新聞1988-2000」是將《中國時報》（以下簡稱中時）編輯部新聞攝影中心所擁有的一組新聞攝影底片，進行數位化的工作，並將數位化後的影像檔案，建置爲可供政治大學校內與近用國科會數位典藏平台之研究者瀏覽檢索的公共資源。

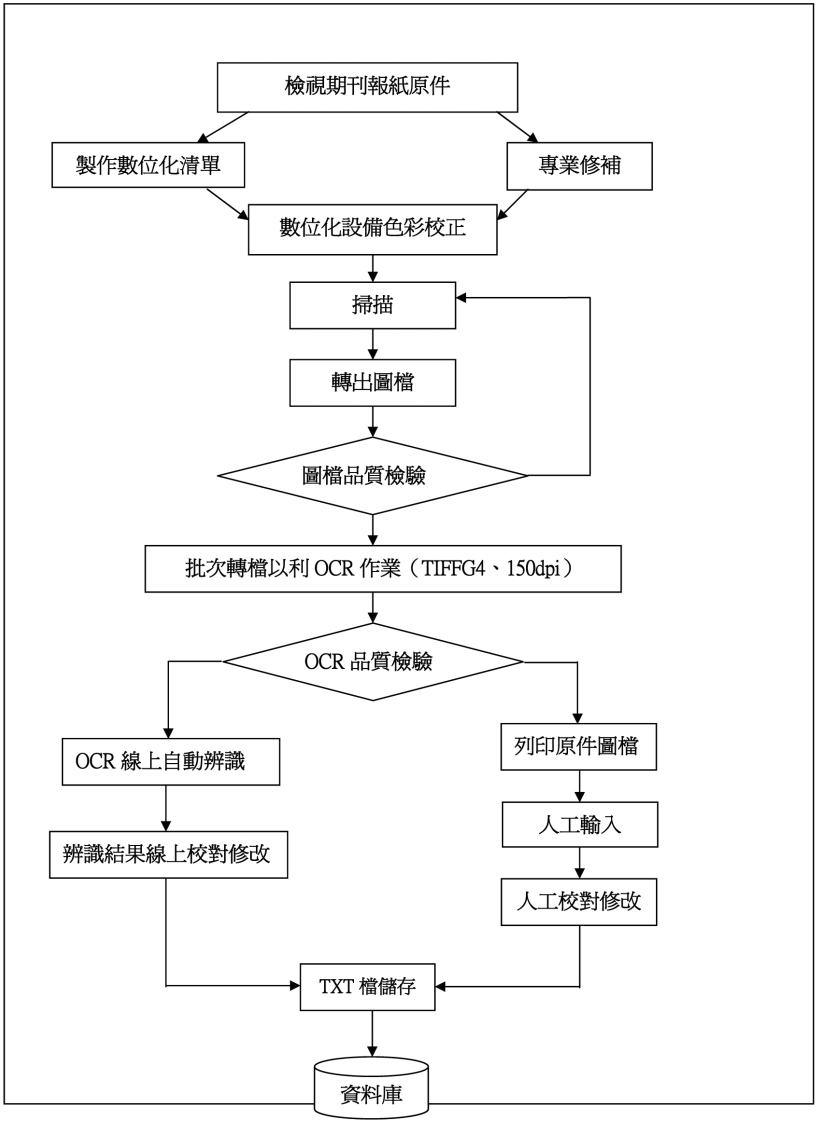
爲了解各機構單位典藏品內容以及數位化工作程序，拓展臺灣數位典藏計畫亦針對各主題小組進行數位化工作流程之調查，在93年曾經出版新聞主題小組數位化工作流程叢書，透過圖像及文字並陳方式來紀錄各計畫單位的數位化工作流程，以提供給其他數位化機構單位相關之參考經驗；而94年則預計將不同主題但爲相同數位化物件者，進行跨主題式之全面性整合，「物件」包括平面物件—相片（正片、負片、照片）、文書、檔案、期刊報紙、書畫、

拓片等；立體物件—動植物標本、考古遺物、地質標本、器物等，其中並以相同數位化方式（如：掃描、攝影翻拍）進行數位化工作流程指南之彙整，以提供一套完善的標準作業流程(Standard Operational Procedure，SOP)作爲數位化參考依據。

「期刊報紙數位化工作流程指南」的目標對象則針對以期刊、報紙爲數位化物件的機構單位或有興趣之個人爲主，並以全文輸入視爲本文數位化方式之重點來撰寫，調查方式則藉由採訪數位化執行廠商並實際測試操作，針對目前全文輸入的現況與技術進行分析及歸納，讓不同階層的使用者能依據實際情形、人力或時間成本等，選擇適合進行數位化的方案，也提供其對數位化工作流程更進一步的認識與瞭解。內容主要包括有：（一）引言、（二）數位化工作流程圖、（三）前置作業、（四）物件數位化程序、（五）後設資料與資料庫建置、（六）委外製作、（七）數位內容保護、（八）設備與成本分析、（九）結語及、（十）參考文獻。

貳、數位化工作流程圖

Digitization Flowchart



資料彙整：拓展台灣數位典藏計畫

圖2-1、期刊報紙之全文輸入數位化工作流程圖

參、前置作業

Preliminary Procedures

一、年度工作規劃

數位化工作進行之際，因考量到藏品數量、預定數位化進度與範圍及計畫進行期間數位化品質之一致性，故必須針對數位化工作各階段環節進行標準規格的制訂與嚴謹明確的作業規範，以避免無統一而具體的脈絡規則可遵循。概括而言，數位化工作大致包含以下步驟：檢視原件、製作數位化物件清冊、制訂標準與規範、資料影像數位化、全文輸入檢索建置、後設資料(Metadata)分析與著錄、數位化資料儲存與管理、數位化成果運用與加值等。

（一）原件檢視與類型

「期刊報紙數位化工作流程指南」擬定數位化物件爲期刊、報紙，而早期報紙除了以原件類型收藏之外，尚有彙集製作成微縮膠卷(Microfilm)及拍攝成單張黑白底片之形式，故本文在此將紙質的期刊報紙稱爲「直接原件」，而膠捲及底片型則稱爲「間接原件」。檢視「直接原件」必須注意其保存現狀、紙質與印刷品質、破損狀況、缺頁及裝訂方式等，若有需要進行修復者，則須依照物件性質的不同而使用專業修補方式。除此之外，尚需注意原件的完整性，建議以字跡清楚且富典藏價值的藏品作爲數位化物件之首選。

「間接原件」包含微縮膠卷，原理爲將「直接原件」經攝影方法縮攝於鹵化銀底片或其他適於長久保存底片中，進行微縮作業，其常見的型號有16mm和35mm，於溫度21°C、濕度50%下可保存長達100~500年，僅需簡單工具(如：放大鏡)即能閱讀，亦能減少保存空間，然而較不便之處爲製作及複製均需一定的標準程序和機器。此種典藏方法大量應用於圖書館、報社之保存或醫院儲存病人之數碼病歷。下列簡略介紹微縮膠卷的效益與優點：

1.技術成熟穩定：微縮技術具百年歷史，且擁有國際統一規格標準。

2.增加管理效率：體積小，易於管理或調閱。

3.節省儲存空間：比原件紙質資料節省約95%以上的儲存空間。

4.利於永久保存：屬銀鹽正片，可保存100年以上，適合圖書館作永久性的典藏。

5.利於取得複本：讀者可利用閱讀複印機將原尺寸的報紙影印出來，提供研究和傳閱。

表3-1、微縮膠卷蒐藏之報社

報紙名稱	微縮膠卷資料起訖時間	數量
聯合報	民國40年—92年12月	357卷
經濟日報	民國76年—92年12月	196卷
民生報	民國67年2月—92年12月	234卷
中華日報	民國35年2月—85年12月	269卷

這些古老且具有歷史價值的微縮膠卷，經過時間證明其保存時間較爲長久，然而隨著資訊科技的發展，微縮膠片技術也迫面臨淘汰的窘境，若沒有延續保留原始寶貴資料的轉換技術，將對資料的可用性造成威脅。

（二）製作清冊

根據各計畫單位所擬定的數位化物件，進行資料來源分類，因爲物件種類的性質不盡相同，則後續的數位化方式選擇也將依照典藏與使用目的作彈性變更。前述檢視原件過後，將數位化物件編列流水號，並製作數位化清單，再交由專業人員重新核對清冊。另外，物件進行修復者，則待修復完成後再編入清冊中。

（三）訂定標準規範

在進行數位化作業過程中，必須訂定嚴謹而明確的標準與規範。國家圖書館在執行期刊報紙數位化之相關計畫時，特邀請圖書資訊界專家與館內同仁，成立「文獻分析機讀格式計畫小組」，修訂期刊文獻資源建檔之後設資料格式，並共同訂定數位化作業的相關標準與規範。各項規範包含關於後設資料(Metadata)的《文獻分析機讀格式》及《資料數位化標準—檔案數位化與命名原則》、《國家圖書館期刊影像編碼原則》、《國家圖書館報紙影像編碼原

則》，其中編碼原則的制訂是國家圖書館爲避免日後期刊報紙連結後設資料時產生問題，所以依照期刊報紙卷期特性及編碼方式，訂定編碼原則各一份，以作爲數位影像檔案編碼的依據。（詳見附錄一、二）

1.確立施作方式與工作程序

一般在實際施行數位化工作時，考量到使用者的設備、使用的便利性、資訊檢索的需求、網路上資料的傳輸速度、資料的永久保存等問題，需依據工作內容等項目，區分爲「自行製作」以及「委外作業」兩種方式，並建立後設資料分析與著錄作業方式等，目的爲制訂前置作業至資料備份、建置Metadata與製作網站資料庫的整個工作流程順序，同時也可規劃並掌握數位化工作之進度。

2.製作文字輸入及校對規範

無論是選擇以「人工輸入」或軟體辨識之數位化方式進行「全文輸入」，都得事先製作文字輸入建檔及校對規範，其中包括標點符號及字級行距之訂定、折行處之標示、難辨識文字與缺字情況之處理方法、檔案格式、檔案命名等，這些標準的制訂是爲了確保檔案的一致性，同時也方便各執行單位進行內部控管，甚至可加入Metadata欄位，在做全文輸入時順便建置，以達事半功倍之效。如果資料內容較簡單易懂，僅需電腦打字輸入技能的話，則可考慮委外製作方式；而內容若以古字、變體字爲主的文件，則建議交由專業人員執行建檔及校稿。此外，在全文輸入、文字建檔、校對、修改電子檔之工作進行過程中，會經過反覆校稿、列印、改正電子檔等作業，爲確實掌控各部分資料之進展情形，可製作一份進度表供日常登錄之用，而比較詳細的工作記錄，仍以利用電腦軟體處理登錄，如此一來，將有利於追蹤掌握各工作環節實際進度或適時修正。

（四）確立數位化檔案規格及用途

1.訂定數位化檔案規格

依據典藏品資料性質，以及數位化方式的不同，需要考慮制訂不同的檔案格式。如果原始資料是以電腦打字的電子檔，則除了儲存一份文字的原始檔之外，另建議轉成HTML、PDF或RTF三種檔案格式。儲存文字檔的原因是爲了方便做全文檢索，若只有建立後設資料之需求，須先將原件掃描，並以不壓縮格式，儲存一份永久檔，再視需求轉存成其他目的之格式，如網路下載格式及預覽格式等。若原始資料爲照片、圖片、地圖等，則需以掃描器掃成影像檔，並以不壓縮格式儲存一份永久檔，同樣可視需求轉存成其他目的之格式。數位化後的檔案格式一般採用：TIFF不壓縮；TIFF G4；JPG 85%壓縮；PDF等格式。格式說明分別詳述如下：

(1)TIFF(Tag Image File Format)

TIFF的第一個版本是由ALDUS公司於1986年所創立，它利用標籤 (Tag)爲其組成的基本架構，具有極大的擴充性。每一個TIFF檔可以是單頁或是多頁，在編輯的過程中能達到影像資訊無失眞，已被大多數軟體所使用。TIFF格式具有豐富的色彩支援，包括全彩、灰階及黑白等影像格式亦或線條稿（純文字圖檔），並且提供多種壓縮模式，包括LZW（Lempel-Ziv-Welch Encoding，簡稱提LZW）、Huffman's Encoding、及變動長度編碼法等，能使檔案體積變小，但仍然不失眞。使用者可依照需求使用合適的壓縮策略。針對純文字圖檔，建議利用TIFF G4格式（256階、黑白TIFF），使檔案體積最小的情況下，獲得最佳影像品質。以TIFF G4、300dpi、A4尺寸的檔案爲例，每頁檔案體積爲50KB。

(2)JPEG(Joint Photographic Experts Group)

JPEG是由國際標準組織（International Organization for

Standardization，簡稱ISO）和國際電話電報諮詢委員會（International Telegraph and Telephone Consultative Committee，簡稱CCITT）所建立的一個數位影像壓縮標準，主要是用於靜態影像壓縮方面，其採用可失眞(Lossy)編碼法的概念，利用數位餘弦轉換法(Discrete Cosine Transform，簡稱DCT)將影像資料中較不重要的部份去除，僅保留重要的資訊，以達到高壓縮率的目的。雖然被JPEG處理後的影像會有失眞的現象，但JPEG的失眞比例可利用參數來加以控制，一般而言，當壓縮率在5%~15%之間時，JPEG依然能保證其適當的影像品質。其適合應用於壓縮全彩或是8位元的灰階影像，凡是照片或色彩連續的影像都非常適宜利用JPEG來壓縮，且同解析度的檔案體積也比TIFF格式小，更利於在網路上傳送閱讀，也由於JPEG壓縮率高，且影像品質在接受範圍內，所以目前支援JPEG的應用軟體相當多，是目前網路上使用最普遍的影像壓縮格式之一。

(3)JPEG2000

JPEG2000正式名稱爲「ISO 15444」，由JPEG(the Joint Photographic Experts Group)組織於2000年3月制訂完成。JPEG2000的壓縮率比傳統JPEG高約30%左右，並同時支援有損和無損壓縮，而JPEG只支援有損壓縮，且具有支援「感興趣區域」特性，可任意指定部份影像壓縮量或先解壓縮。然而目前支援JPEG2000的應用軟體並不普及，較完整軟體則屬LuraTech技術廠商，其與ACD Systems公司簽訂協定，在使用率最高的圖形管理軟體ACDSee上，提供JPEG2000 LWF格式的外掛元件演算法，如此只要安置此插件就可觀看並製作LWF格式檔。

(4)PDF(Portable Document Format)

PDF是Adobe公司所推出的一種跨平台軟體，爲Adobe系統中

Acrobat的原生性檔案格式，PDF格式獨立於原有製作這些文件的應用軟體、硬體、及作業系統之外，是不需用原有軟體就能閱讀的共用檔案格式。PDF能保存原始文件的字體、影像、圖形和版面，不受設備與解析度影響。目前常見的PDF爲單層PDF，而雙層PDF則融合了OCR辨識結果，即文件內容上層爲圖像，但底層包含OCR辨識的文字資料，可供搜尋之用，並具全文檢索功能，且能找出文字、書籤和資料欄的位置。因此，其不僅保存原始文件的外觀和完整性，又兼顧了文字資料檢索的需求，讓文件的相容性與閱讀性大增。此外，PDF檔案可經由設定密碼來保護文件，以避免被不當複製或未經授權的檢視和修改，同時又可以讓授權的審閱者用註解和編輯工具，因此除了微軟所出的Microsoft Reader之外，PDF也是目前世界上最通行的電子書(eBook)格式之一。

(5)其它格式

CEB格式(Chinese Electronic Book，簡稱CEB)是由北大方正公司所創Apabi Reader中文電子書格式，具有版權紀錄與鎖定的功能，同樣也是不需用原有軟體而能閱讀的共用檔案格式。

表3-2、常用格式的容量比較表（A4 300DPI）

	會否失真	彩色	黑白	容量
TIFF 不壓縮	不會	可	可	極大
TIFF LZW 壓縮	不會	可	可	大
TIFF G4	會（部分文字不會）	不可	可	極小
JPEG 不壓縮	會	可	可	大
JPEG 85% 壓縮	會	可	可	中
JPEG2000	不會	可	可	極小
PDF	不確定	可	可	中

資料彙整：拓展台灣數位典藏計畫

2.數位化檔案之用途

(1)印刷

A. 期刊報紙之印刷用途

a.原物重現、再版發行

b.宣傳展示

B. 解析度需求

簡單而言，解析度即圖檔的清晰程度，而解析度越高則所需儲存空間也就越大。上述印刷用途皆可依照原始尺寸、放大或縮小以進行印刷作業。要達到原始尺寸的印刷，其解析度至少要300dpi。若要放大印刷，則解析度必須相對提高，然而因爲報紙本身尺寸的關係，在掃描技術上就必須要克服提升解析度的困難；另外若放大的需求是大圖輸出，例如大型海報或外牆使用等，則解析度以72dpi爲基準數，依照實際需求將長寬等比例放大即可，其目的在於遠距離觀看，故近距離檢視下出現馬賽克是可被接受的，此做法較適合量少的宣傳品使用。至於縮小作稿方式，原則上建議在電腦設備可支援情形下，使用72dpi、原尺寸1：1或300dpi、縮小4倍進行輸出作業較不易產生馬賽克，成品質感也較佳。

(2)實體與數位化保存

對期刊報紙實體存放空間而言，不論是在何種場所、空間大小、溫濕度控制、照明亮度或是降低紙質成分的損毀度等，都是對於進行數位化工作相當重要的關鍵。簡單來說，期刊報紙必須在恆溫恆濕以及與空氣日光接觸少的環境空間下儲存，然而調閱瀏覽及操作掃描等人爲因素次數愈頻繁，造成原件壞損的機會便愈大，於是進行數位化工作便等於增加另一種保存原件的方式。而期刊報紙原件也因爲尺寸及數量的關係，累積蒐藏量體積相當龐大，需要絕對寬敞的儲存空間來存放，相對而言，儲存成本總

金額也隨之增加，故採取何種數位化格式也就刻不容緩且須謹慎評估之。例如國家圖書館在進行館藏期刊報紙資料數位化時，爲要求數位化內容清晰以及永久典藏，則依據「資料數位化與命名原則」之建議規格，決定採用文字檔及影像檔資料永久保存格式進行數位化。其中文字檔之永久保存格式建議規格爲TIFF不壓縮、300~600dpi；下載格式建議規格爲JBIG[1]、150~300dpi；預覽影像建議規格爲GIF、72dpi。（詳細數位化檔案建議格式請參閱附錄三。）

(3)網路瀏覽

網路瀏覽的目的在於使數位化圖檔能夠在網路上供大衆瀏覽，然而因爲網路頻寬的限制，所以必須選擇適合的檔案格式來進行數位化，而圖檔體積愈小，網路瀏覽便愈順利，相對地圖檔清晰度也會減少，尤其是圖檔內容以文字爲主時特別明顯，而目前可透過新掃描技術提供品質較佳的低容量圖檔體積並且降低文字清晰度的流失。

(4)電子書

期刊報紙進行數位化後的圖檔，可以依照所需主題組合而成電子書，以電子書形式提供予使用者下載、閱讀或列印。目前國際普遍檔案格式爲PDF，而中文電子書則以北京方正阿帕比技術有限公司發行之Apabi Reader軟體市佔率最高(http://www.apabi.cn)。

二、數位化執行方式之選擇

以往期刊與報紙的數位化處理方式，有影像掃描、人工輸入、光學文字辨

1 JBIG是一種無損圖像的壓縮標準，從Joint Bi-level Image Experts Group而來，並由ISO/IEC standard 11544 ITU-T recommendation T.82所標準化。資料來源：http://en.wikipedia.org/wiki/JBIG。

識(Optical Character Recognition，簡稱OCR)、電子報直接轉入資料庫等四種[2]，以下將以新聞主題工作組內計畫作爲範例，各數位化執行單位可依原始資料性質並評估成本預算後，再決定採行的數位化方式，或是數種方式搭配使用。

（一）影像掃描

影像掃描是將報紙版面掃描成爲影像檔儲存，可存爲JPG或PDF等圖檔格式，原則上解析度要到300dpi才夠清晰，爲目前市面圖書館與大型研究機構較常用的一種數位化作業，而目前爲止新技術已能滿足清晰度且高壓縮至150dpi，這種做法比較簡單而省時省力，且可提供仿眞的資料原件複本給使用者，例如「國家圖書館期刊報紙典藏數位化計畫」所成立之報紙影像資料庫，即是此種方式的代表：將報紙掃描後（含微片轉製34種，共有445,584頁影像檔），另外建置標題與相關後設資料與欄位，以提供報紙文獻的全頁影像與新聞標題查詢。然而倘若掃描的影像內容無法直接辨識進而提供檢索，在使用上的效益將遠不如電子全文資料。故現今已有雙層PDF融合影像內容及OCR辨識結果，以彌補純粹影像掃描而無法進行全文檢索之憾。

（二）人工輸入

人工輸入則是將紙本原件或將已經掃描成影像或製成微縮膠卷的報紙重新輸出，再用人工方式重新打字建置資料，完成的內容必須再經人工校對，例如「世新大學北平世界日報內容數位化開發計畫」，最後是把校對好的文字檔轉換成爲資料庫格式，上網供使用者查詢。這種全文輸入的方式，需要的是電腦打字輸入的技能，可以採外包的方式，再由單位內的人員進行檢校；若資料原件多異體字或有闕漏，則不建議交付外包。

2 孫正宜、林信成，〈中文報業數位化技術與現況探討－聯合知識庫數位化經驗〉，《2003年資訊科技與圖書館學術研討會論文集》，2003年5月。

（三）光學文字辨識

光學文字辨識是使用掃描設備將印刷文件讀入，並將文件上的文字辨認後轉換成電腦使用的文字編碼，例如ASCII 碼或BIG-5碼，再轉入資料庫供使用者檢索查詢，適合印刷清楚、資料量龐大的文獻，其正確率可達99.98%，若是期刊報紙的原件年代久遠、紙張泛黃，而產生漏字缺角、辨識模糊等缺陷，仍需要經由人工校對來提高正確率；有時掃描品質不佳，內容清晰度差，OCR效率反而比不上人工輸入，例如「世新大學世界日報內容數位化開發計畫」即在評估之下選擇使用人工輸入法。不過一般而言，在典藏品掃描後品質仍佳的情況下，利用OCR的技術來還原文字，其成本遠比人工輸入來得低廉。如果已有其他型式媒體備份，例如影本或微縮版，則第一階段之輸入建檔應利用影本或微縮資料列印文件。影本或微縮資料列印文件如有不清楚之處，再批次調取原件核對。要進行核對時，如果廠商數位檔已製作完成，則可利用數位檔進行核對；原則上是盡量減少提取原件的機會，以保護原件。

（四）電子報直接轉入資料庫

「辦公室維運分項：出版子計畫」則是直接將電子檔轉入資料庫，以《國家數位典藏通訊》電子報的形式發送，必須另外建置Metadata方能供使用者查詢。又如國內最知名的兩大報系—聯合報以及中國時報，早已將報紙編排方式數位化，並把當日新聞文字稿儲存至資料庫中，而所謂資料庫、Metadata的建置、XML的應用等則自從網路普及後才逐漸受到重視。

表3-3、期刊報紙數位化方式特性分析表

數位化方式	特點	弱點
影像掃描	提供原件複本	無法全文檢索
人工打字	可直接判斷出缺字或難字	耗費大量人力、時間成本
光學文字辨識	速度快、效率高	鉛字排版、印刷字與手寫字混排、注音體、影像檔品質不良等辨識率低
電子報	本身形式即已經過數位化	

綜合上表四種期刊報紙數位化方式之優缺點比較，因影像掃描方式若無法提供使用者內容的全文檢索，因此使用效益不大；人工打字方式雖僅需打字技能，相較於光學文字辨識則耗費了太多的人力與時間成本；而OCR數位化效率雖高，但若無適合的文件類型，則辨識率仍有待突破；電子報本身形式已經過數位化，暫不在此進行比較。

一般而言，執行單位在進行文字數位化時，較常遇見情形爲OCR辨識率過低，不得已改而採取較耗費成本之人工輸入法，然而，若是能對物件影像檔做些適當的處理以提高其辨識率，不僅能使大量文字圖像內容能夠重新引用並方便檢索，同時也能減少許多不必要的人力或時間成本（OCR辨識處理步驟將於下一章節詳細作說明）。因此，本文除了針對OCR光學文字辨識作一深入探討研究之外，也提供一些選擇人工輸入或OCR辨識的參考依歸，其中以OCR品質檢驗要則爲主要考量，利於使用者在進行全文輸入時，依據本身現有的實際情形自行斟酌並作調整。

就文件類型而言，適合進行OCR辨識的文件類型有常見的印刷體爲主、已清除雜點、傾斜校正且文字與底色反差明顯者。而不適合進行OCR辨識的文件類型則包括排版格式複雜、字體非一般常用字、帶有注音符號或數學運算公式等，甚至因爲紙張較薄（磅數較低）、油墨較深者容易造成背面文字顯現於正面文件上，這些因素都將對OCR辨識率造成影響。另外，民國50年左右的報紙是使用鉛字排版方式印刷，因排版字縫間有空隙，且因年代久遠或溫、濕度失恆而使紙張泛黃或毀損，導致掃描後品質不佳、內容清晰度差者，則建議使用人工輸入方式較有效率。

表3-4、數位化方式品質檢驗要點

數位化方式 / 品質檢驗要點	OCR光學文字辨識	人工輸入
字體	常見印刷體	純手寫稿、夾雜注音體、數學運算公式、印刷體或手寫字混排、古文或變體字多
排版格式	電腦排版、格式簡單、讀文順序清楚	早期鉛字排版、格式複雜、讀文順序不順暢
雜點	版面較為乾淨、無雜點	字體周圍較多標記或雜點
反差度	純黑白稿、字體清晰、反差度高	本身影像品質不佳、字體較為模糊、反差不明顯

資料彙整：拓展台灣數位典藏計畫

就圖檔格式而言，OCR軟體在個人電腦問世後不久即產生，然而當時僅能支援150dpi、黑白TIFF或BMP檔案格式。目前則因個人電腦處理能力大幅提升及改善，OCR也已能處理JPG格式。而爲確保辨識的精確性並提升辨識效率，建議將彩色或灰階文件圖檔進行影像處理，取得較佳的影像格式（150~200dpi、黑白TIFF），以利OCR作業之進行。目前測試結果顯示有利OCR之圖檔格式依序爲：黑白TIFF G4、150dpi；黑白TIFF G4、300dpi；全彩JPG／TIFF、300dpi。黑白圖檔因文字與底色的反差明顯度大於彩色圖檔，故OCR辨識度較高；而在同樣能進行OCR作業情況下，黑白TIFF G4、150dpi則因檔案體積及佔用資源空間較小，故較優於黑白TIFF G4、300dpi進行OCR文字辨識。

表3-5、微縮膠卷蒐藏之報社

圖檔格式	利於OCR辨識程度（依次排序）
黑白TIFF G4、150dpi	反差度高、體積較小
黑白TIFF G4、300dpi	反差度高
全彩JPG／TIFF、300dpi	底圖與文字反差不明顯，對OCR辨識造成干擾

資料彙整：拓展台灣數位典藏計畫

三、後設資料之建立

（一）確立檔案格式

目前新聞相關之後設資料格式不論平面或是電子媒體尚無統一標準（針對在新聞主題工作組中不同媒體類型之典藏品），可能需要不同的後設資料加以詮釋；近來許多新聞傳播相關計畫加入「數位典藏與數位學習國家型科技計畫」，各個子計畫或典藏單位的資料庫，都具有描述各自典藏品的後設資料與整理工作，期望未來能夠逐步跨資料庫與檢索系統間加以結合。

（二）後設資料需求訪談

不同類型數位化物件的後設資料不盡相同，若能訪查相關計畫或有經驗的單位，請專家們給予參考，建置符合使用者及管理者需求的後設資料，並參考國際相關標準，將可徵集多方意見，使得後設資料更加完備。

（三）訂定後設資料規範

將各類型資料加以分析比較之後，即可依照各典藏品特性來訂定後設資料規範與欄位建置；由於聯合目錄所採用的是都柏林核心集（Dublin Core，簡稱DC）做爲核心欄位，其普遍性雖然可以處理異質資料庫間的共通，但不同的媒介與計畫間應有適用於該主題更需被凸顯的核心欄位，由此整合的核心欄位再行對應DC欄位，並搭配個別資料庫欄位的分析，將可提高呈現內容的目錄價值。

肆、物件數位化程序

Object Digitization Procedure

一、色彩校正

（一）儀器之色彩校正

色彩校正之目的在於充分保留報紙期刊的原狀，尤其是色彩以及文字資訊部分，讓使用者能從閱覽數位化檔案便能獲取與原物件相同之資訊內容，並了解期刊報紙在掃描當時的保存狀況爲何。而色彩校正也一直是電腦繪圖及印刷最困難亦最不易解決的問題，因電腦螢幕上的顏色有許多根本就無法印出來，或者有嚴重的色偏等，其每一環節皆環環相扣，從螢幕、掃描器至輸出到印刷，每一層轉換步驟都有色偏的問題。造成色偏之因素如下：

1. 螢幕：螢幕校正需要使用貼在螢幕上之光學儀器，藉由讀取螢幕上特定色塊之顏色值來修正。
2. 掃描器：掃描器則必須使用該掃描器專用的校正用色卡，經由比對理論顏色與實際掃描得到的顏色來作修正。
3. 印表機、印刷機：依然必須執行色彩校正才能在可能範圍內得到最佳的輸出品質。

（二）色彩校正方式

就桌上型掃描器而言，是依照國際照明協會（法文Commission internationale de l'éclairage，簡稱CIE；或英文International Commission on Illumination，簡稱ICI）於1976年將CIE xyz以數理方式轉換成新的CIE Lab模型爲基準，並以色彩工業標準—IT8標準色彩導表來作爲桌上型掃描器校色之基礎。

而近年來則因爲數位相機的誕生，便出現取代傳統相機底片的電子光學元件，即感光耦合元件（Charge Coupled Device，簡稱CCD），而隨著CCD或互補性氧化金屬半導體（Complementary Metal-Oxide Semiconductor，簡稱CMOS）技術的進步，各設備皆有其相對專用之色彩導表以進行色彩校正，並產生裝置色彩描述檔ICC Profile，根據此影像標準格式檔與前、後端設備做連結，盡可能保持輸出的一致性。 倘若儀器設備狀況有任何變動，則必須重新進行色彩校正與調整。在此本文以專業多用途掃描器爲例（廠牌：I2S、型

號：DiGiBook 10000RGB）進行色彩校正。詳細色彩校正流程與專用色彩導表請參閱附錄四。

（三）特例說明

數位化過程中，若需要較大的亮度才能顯現掃描物件本身的細節與特性，則必須考慮需求與目的爲何，是否以物件本身色彩爲第一優先，或以清晰呈現細節爲優先考量。例如植物標本的掃描，若考慮使葉脈更爲銳利化，則物件本身顏色即會些微偏差。

（四）輸出應用模式

1.列印（印表機）

一般個人使用並不會特別注重印表機的色彩校正，然而以專業色彩校正而言，印表機本身及所使用紙張、碳粉或於墨水更換時都必須確實執行色彩校正，才能確保輸出之色彩品質均具有一致性。

2.印刷（印刷機）

爲確保印刷文件品質與原件相同，印刷機也必須執行色彩校正，因目前台灣市場上大部分的印刷機器並不支援色彩校正，所以實務上執行有其困難度存在。

3.網路瀏覽

經過螢幕及掃描設備色彩校正後之檔案可直接應用於網路瀏覽。

二、數位化掃描技術

回顧以往多數以數位化產出爲首要考量基礎的設備或技術，因在數位化過程中較少將重心放置於文物的保護上，導致原件因設備（如掃描機器離心力過大或燈光過熱等）、存放空間（如過於陰暗潮濕）或人爲因素（如無使用適宜手套翻閱掃描）而造成毀損或破壞。目前則因有專門適合期刊報紙進行數位化之機器設備（如書籍掃描器、專業多用途掃描器等），使得文物能兼顧數位化

產出及保持現狀之需求，以降低數位化過程中原件受傷害程度。目前市面上掃描器已能支援在不破壞原件的情形下，進行書背較厚的裝訂式期刊報紙之數位化，其過程不需接觸文物或拆卸裝訂，原理是運用180度書籍支架（圖4-1）或120度翻開面支架（圖4-2）來支撐物件左右兩邊重量之平衡。另外若物件本身裝訂處過於緊靠文字，則建議以盡量不傷害原件爲原則，使操作人員依然能清晰可見裝訂處之文字並進行掃描，例如使用手套將物件四邊拉平，而手套則需準備棉質與膠質二種，端視期刊報紙物件狀況而決定穿戴何種手套。[3]

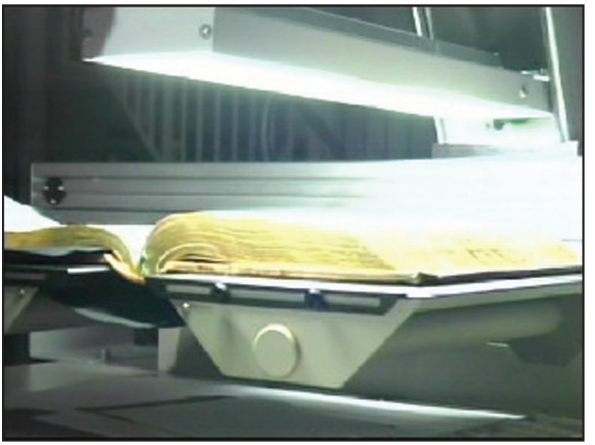


圖4-1、180度書籍支架

3 洪淑芬，《文獻典藏數位化的實務與技術》，台北：數位典藏國家型科技計畫訓練，2004年2月，推廣分項計畫，頁96。「棉質手套」：如果所處理之事項多為搬移作業，接觸部分多為資料之外包裝，或是翻動之資料狀況良好，極易翻掀，則棉質手套可防汗垢沾上資料，但是，棉質手套必須隨時清洗乾淨，避免使用已髒污之手套。「膠質手套」：最好是手套內無粉者。膠質手套不透氣，穿戴時間稍長會感到不舒服，但對於有蟲蛀之資料，必須使用表面光滑之膠質手套，以防止資料上的蟲損之處，黏附於手套上，反而對資料造成傷害。

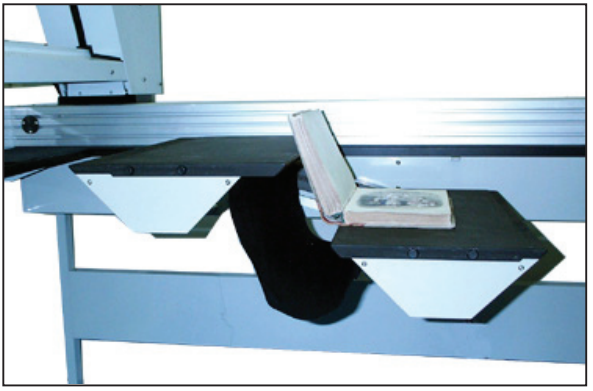


圖4-2、120度書籍支架

圖片提供：磁軒資訊媒體行銷有限公司

三、 光學文字辨識技術

（一）光學文字辨識系統說明

所謂光學文字辨識是利用掃描器或數位相機等光學輸入設備獲取印刷文件或手寫於紙上的文字圖片資訊，再以各種模式識別演算法逐一辨識分析文字形態特徵，並轉換成電腦可操作的文字編碼，例如美國資訊交換標準碼（American National Standard Code for Information Interchange，簡稱ASCII code）或BIG-5碼，然後轉入資料庫供使用者檢索查詢。

對OCR光學文字辨識而言，進行中文字辨識的困難度遠高過於歐美國家的拼音文字，因中文字字數特多，且需考慮字型架構、字型變化的複雜度等，故國內的中文OCR研究至近期才邁入實用的階段。傳統將整張文件掃描經過壓縮存成影像檔的儲存方法，不僅占用空間龐大，且內文不易修改、編排或複製，一旦涉及建檔、索引、歸類等資料庫處理時更是一項繁瑣且廢時的工作，若能將文件中影像部分壓縮，再利用OCR將文字部分加以數位化轉成字碼方式儲存，則不但節省大量檔案儲存空間，且新增、刪除或修改文字內容均極爲容易。

（二）OCR技術與產品現況

目前OCR的研究與技術開發，在台灣有力新國際科技、蒙恬科技、全景軟體，在大陸則以清華文通和北京漢王最著名。以下介紹上述OCR主要廠商之技術與產品現況。

1.力新國際科技

原本爲力捷電腦(UMAX)的軟體部門，負責開發掃描器驅動程式售軟體，後來於1987年獨立成爲「力新國際」公司。產品以影像處理（非常好色）、光學文字辨識（丹青）軟體與名片辨識系統爲主。其中丹青文件辨識系統[4]技術移轉自工業技術研究院電腦與通訊研究所，是國內最早技術達至成熟的產品，功能包括處理黑白、彩色文件、文件版面分析、表格抽取、印刷多種字體中英數字夾雜辨識。力新國際也積極以專案方式與各機構單位合作，例如國電訊發展室「傳眞文件的辨識與分類」、中華電子佛典協會（Chinese Buddhist Electronic Text Association，簡稱CBETA）與日本「大藏出版株式會社」簽約進行的《大正新脩大藏經》數位化，均與該公司合作。其中，力新國際科技研發部更專爲CBETA輸入作業需求而設計，進而發展出「丹青for CBETA版」的OCR辨識軟體。

2.蒙恬科技

蒙恬科技爲獨資企業，成立於1991年，由蔡義泰博士創辦，以手寫輸入系統切入市場，爲當時手寫辨識(Handwritten Recognition)技術最先進的中文手寫輸入系統。1994年自工研院電通所前瞻資訊技術中心（Advanced Technology Center，簡稱ATC）移轉OCR辨識核心，並中央大學資訊工程學系合作，開發OCR相關技術，於1996年推出與「認識王」可辨認手寫稿之OCR軟體。並自1997年開始研發語音辨識技術，經由IBM的ViaVoice語音辨識核心的授權，於1998年首推「聽寫王」彙集語音與手寫辨識系統。其它OCR的應用技術則有整合掃描、

4 目前（97年）丹青文件辨識系統已發展至第5版，本文則以4.5版為範例。

辨識、翻譯三種介面的「掃譯筆」以及名片辨識與編輯的「名片王」。並將辨識技術推爲爲全球產品「WorldocScan」，以輕巧型機身包含辨識文件、表格、名片與照片並轉製爲PDF檔案格式，並可建檔用關鍵字搜尋，解析度爲600dpi (optical)，加強許多功能且與其他裝置連結性更高。

3.全景軟體

全景軟體公司於1998年正式成立，創始人爲前國立交通大學校長、交通部長郭南宏博士，公司在創立初期藉由產學合作計畫自交通大學引進了OCR、文件影像分析、彩色影像處理、影像壓縮、音訊處理、檔案加解密等資訊關鍵技術，進行技術商業化及個人用套裝軟體開發，目的在於將實驗室內可商品化的實驗結果帶出，持續研發成爲商品。目前的產品領域包括與OCR相關文件影像、網路安全、與虛擬實境三類。而藉由企業化經營的過程，公司目前已成功發展出國內產學合作的良好典範。但其OCR部分爲企業解決方案形進行整合，並未包裝成商品套件上市。

4.清華文通

北京文通信息技術有限公司（Wintone，原北京清華紫光文通資訊技術有限公司）成立於1992年，是在中國科技部（原中國國家科委）與清華大學電子工程系的支援下，爲推廣應用國家「863高科技計畫」資訊領域多字體印刷漢字自動識別技術研究成果而形成之企業。TH-OCR是清華大學自1985年即開始研發，TH則是TsingHua（清華）之縮寫，文通資訊以工程院院士吳佑壽 爲首，在丁曉青教授領導下，長期致力於清華TH-OCR的研究與開發，目前能自動識別多體漢字、漢英混排文字、印刷及手寫體，其產品在大陸市場佔有率達65%以上，其中日、韓文與英文混排文字檔的識別水準甚至超過日本及韓國對其本國文字的識別水準，而亞洲文字（中文簡體、中文繁體、日文、韓文）識別技術也因此獲得微 軟高度認可，並在

Microsoft Office 2003中全面配裝。

5.北京漢王

漢王科技股份有限公司成立於1998年，以「中國國家文字識別工程中心」科技研究爲基礎，在中國「七五計畫」、「八五計畫」、「九五計畫」、「863高科技計畫」、國家自然科學基金等重點專案支持下，專注於手寫、語音、OCR、生物特徵等識別技術的研究和推廣，相繼推出了語音命令合成技術、OCR掃描輸入、名片識別管理系統、指紋識別、身份證識別、車牌號碼識別、銀行票防僞識別認證甚至人臉辨識等系列，與OCR相關的產品系列有漢王文本王、漢王E摘客、漢王名片通以及漢王文本儀等。

（三）OCR技術與實際操作

1.辨識操作程序：

評估掃描過後的影像圖檔是否需要進行去雜點或頁面傾斜校正，之後再經過OCR軟體做版面切割動作，並比對字形檔與圖像內之字樣，經檢索出對應字後，再就文句本身的詞義做詞庫之自動校正，待人工方式做對照校正後，即可儲存成一般的文字檔，最後依照各使用者之需求，運用其他應用軟體加以處理。

2.OCR技術分析：

OCR在技術研發方面以文件分析與光學文字辨識研究爲主，其中文件分析包括彩色背景的去除、文件區塊（文字、影像、表格）的分離、直橫排的偵測、閱讀順序的決定等；而光學文字辨識則包括文字切割、手寫或印刷字之判斷、印刷字體的偵測、手寫及印刷中文和英數字的辨認核心等。OCR的處理過程除了本身的辨識引擎之外，還可針對辨識前的影像圖檔或辨識後的結果做進一步的處理與分析。以下略爲描述前處理、辨識引擎及後處理等步驟。

(1)前處理

期刊報紙等物件經由掃描成爲影像檔至進入辨識引擎之前，

這期間的處理過程均屬於前處理範圍。此步驟又可分爲「影像處理」、「版面分析」與「字元切割」等三部分。

A.影像處理

本文曾說明物件本身的文字與底色反差明顯者較宜進行OCR，亦即直接以黑白文件且清楚而無雜點者進行掃描較佳，然而，爲避免因掃描品質不佳而使得黑白文件影像檔中的字元產生破碎或模糊不清，如今OCR辨識系統已能允許彩色或灰階的文件影像輸入，並利用影像處理技術[5]求得較佳的黑白影像檔，以提高辨識率的準確性。

B.版面分析

由於OCR辨識引擎通常只辨認單一字元，因此文件影像需先經過版面分析，而版面分析原理及使用技術敘述如下：

a.版面分析原理

將文件區分爲影像、表格與文字三種區塊，其中影像區塊是不可辨認者，可經過壓縮予以儲存；表格區塊則經過格線抽取、交點偵測、欄位抽取等，將表格的格線與欄位分離，而表格的欄位和文字區塊，則需經過文字行的抽取與字元的切割，將每個字元抽取出來後，可分爲「區塊分割」、「區塊型態判斷」及「傾斜校正」三種方法見表4-1再進入辨識引擎做辨認處理。

b.版面分析使用技術；可分爲「區塊分割」、「區塊型態判斷」及及「傾斜校正」三種方法，見下表4-1。

---

5 曾逸鴻，《光學文字辨識(OCR)技術整理報告》，台北：國防部電訊發展室，2004年1月，頁2。

6 同上註，頁3。區塊切割有兩種方法：「遞迴投影法」(Recursive projection analysis)或「相連元件偵測法」(Connected component detection)。若文件屬於版面較傾斜者，則前者「遞迴投影法」較無法獲得準確的切割位置。

表4-1、版面分析使用技術

版面分析使用技術	技術說明	
區塊分割[6]	在一般文件影像中，每個區塊均會以空白行（大小不定）做區隔，因此在理想情況下，可將每一文字行切 出，甚至切出每個字元。	
區塊型態判斷	黑白點比例	首先，先計算區塊內的黑白點比例，若黑點遠多於點，則可能為影像區塊白點，則可能為影像區塊。
	線段的存在	若區塊內可找到數段直線，則可能是表格區塊。
	相連元件的平均大小與間隔	區塊內相連元件的大小與間隔分佈平均，且找不點，則可能為影像區塊到直 線，則應為文字區塊。
傾斜校正	一般而言，OCR通常可進行些微傾斜字元的辨識（傾點，則可能為影像斜角度在正負0.5度以內），但若傾斜角度過大，將會影點，則可能為影像響版面分析與文字辨識率，因此在版面分析階段，會先做點，則可能為影像傾斜角度的偵測與校正。目前新技術「地理性校正」已能點，則可能為影像針對頁面或內容文字傾斜進行曲度修正，並盡量將影像點，則可能為影像頁面調整至水平以利後續OCR辨識作業。以下就期刊為數點，則可能為影像位化物件作範例，以影像掃描後製軟體Book Restorer進行地點，則可能為影像理性校正前後之比對。（圖4-3、圖4-4）	

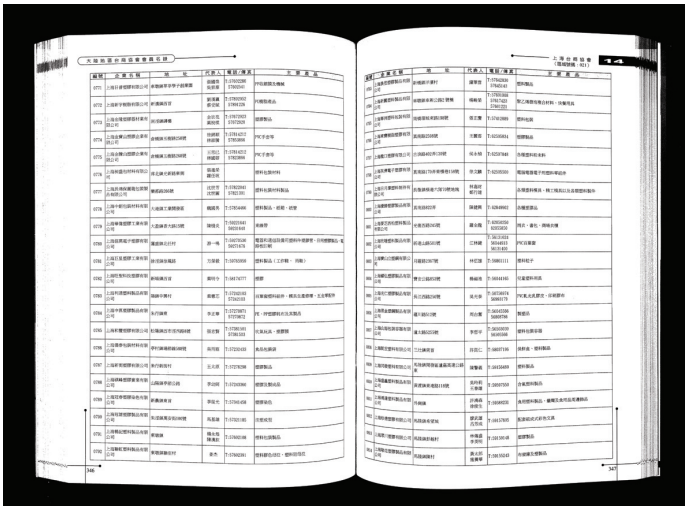


圖4-3、原始物件掃描之影像檔

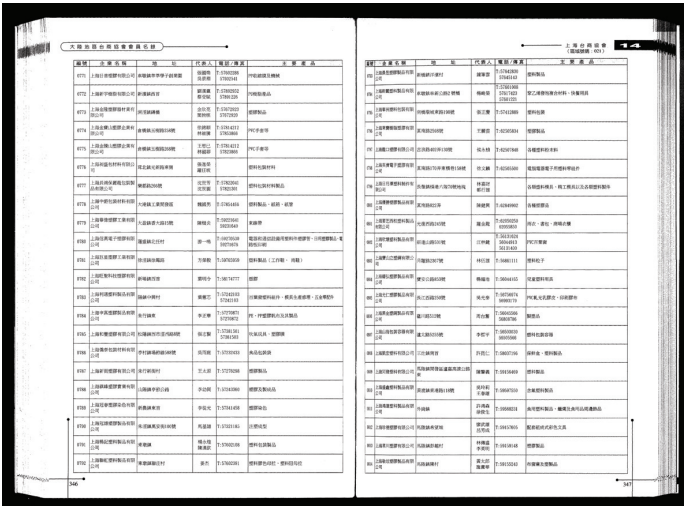


圖4-4、進行地理性校正之影像檔

圖片提供：磁軒資訊媒體行銷有限公司

C. 字元切割

當版面分析將每行或段落文字切出後，在進行辨識之前，尚須將每一文字元切割清楚。在此可利用一些文字特性，來決定哪些是正確的切割位置。例如，中文字乃方正字，若採用某切割位置，則可能導致切出太狹長的字元而無法採用。但若辨識文件爲中英文夾雜者，可將切出的非方正字先進行英文辨識，如果辨識結果符合原字元，則此切割位置方法將可採用。當辨識文件中的每行字元間距夠明顯，即可提高字元切割的效率與速度。

(2)辨識引擎

當字元切割完成後，即可將每個字元影像以辨識引擎進行辨認。最基本的辨認方式，即將字元影像與資料庫中每個中文字的影像比對，並計算相對位置的顏色是否相同，找出差異最小者即爲辨識結果。辨識引擎的內部技術有特徵抽取、特徵比對與加速技術。詳述說明請參閱附錄五。

(3)後處理

一般而言，在文件本身的影像品質不佳的情況下，辨識率其實不易達到令人滿意的效果，然而在後處理的技術方面，加強OCR系統學習功能是有可能微幅提高辨識率的。此部份可採取字典查詢或者前後文相關方法來進行：

A. 字典查詢法

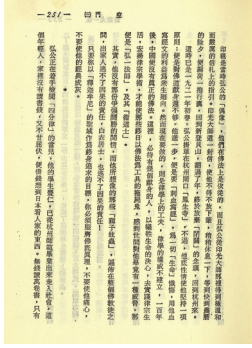
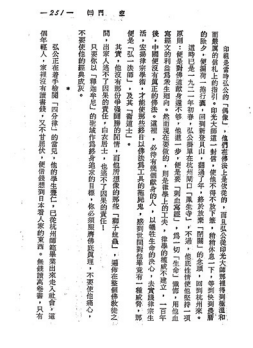
針對辨識內容特定的需求與用途（例如名片辨識、新聞字幕等），可事先內建辭典以提供候選字做更正的步驟。以名片辨識而言，通常會有一欄位爲「電話：」，而其後緊接的字元就可限制爲阿拉伯數字及特定字（如＃、轉、分機等），如此便能降低辨識系統誤認的情況。

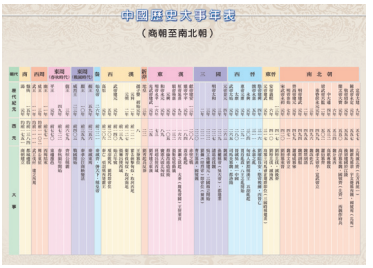
B. 前後文相關法

蒐集大量辨識字元，並統計每個字元前後相關聯字出現最頻繁者，讓OCR系統具備自動學習關聯字之功能，待完成辨識結果後，即可以本身字元的候選字加上前後文來判斷最有可能的辨識結果。

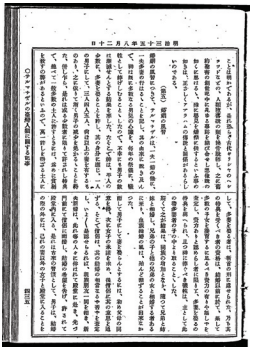
3.辨識範例說明：

進行OCR辨識測試物件有橫式中英文夾雜文件JPEG、TIFF；直式中文文件JPEG、TIFF；直式表格JPEG；直式中日文夾雜文件TIFF等。詳細測試圖檔列於下圖4-5：

	
橫式中英文夾雜 （彩色JPG）	橫式中英文夾雜 （黑白TIFF）
	
直式中文（彩色JPG）	直式中文（黑白TIFF）



直式表格（彩色JPG）



橫式中文夾雜 （黑白TIFF）

圖4-5 、OCR辨識測試圖檔

圖片提供：磁軒資訊媒體行銷有限公司

本文以實地採訪方式進行OCR辨識軟體的操作過程與結果分析，其中因全景軟體版本無商業發行版可茲比較，而北京漢王則無發行台灣版，故本文在此針對台灣的力新國際、蒙恬科技以及大陸清華文通三家廠商軟體進行操作介面、辨識速度及效果之測試及研究。下列爲OCR軟體測試系統版本：丹青中英日文文件辨識系統4.5、蒙恬認識王專業版V3.1、清華TH-OCR 2003錄入工廠。

在進行物件測試OCR辨識的過程中，可發現文字與底圖色差愈明顯，則辨識效果愈佳，並且以印刷體文字較適宜進行OCR。故物件圖檔格式建議轉爲黑白TIFF、解析度爲150dpi，如此一來便能提升OCR辨識率的速度及效率。

根據測試物件的版面分析及辨識結果差異較大者，本文以辨識進行畫面作說明：在橫式中英文夾雜文件測試結果中，以清華軟體辨識率較丹青及蒙恬軟體高；直式中文文件的測試結果則較無太大差異，唯獨清華軟體較能分辨出上下引號之符號（即「」）。至於直式中日文夾雜文件的辨識結果，因爲蒙恬軟體版本無法支援辨識日文，強制執行下的辨識率並不高；丹青軟體

在進行辨識時，版面會有亂碼出現，但仍可進行辨識，而清華軟體的中日文夾雜辨識結果則出現一堆問號，必須另存至TXT檔才能出現辨識結果，其辨識率高過於丹青軟體；以直式表格文件作測試，則發現丹青及蒙恬軟體皆辨識出表格內容之文字行，而清華軟體的辨識結果則包含表格框線和內容文字（圖4-6）。另外，值得說明的是在本文測試軟體系統中，清華軟體可移動影像內容與辨識結果中的橫隔線，這對進行後製處理步驟而言，相對較爲方便且人性化（圖4-7）。

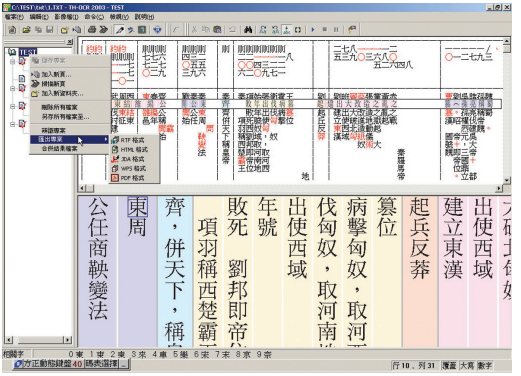


圖4-6 、OCR辨識測試圖檔

圖片提供：磁軒資訊媒體行銷有限公司

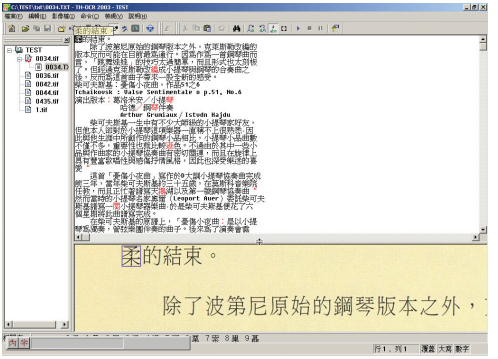


圖4-7、清華軟體—可移動式橫隔線

圖片提供：磁軒資訊媒體行銷有限公司

（五）OCR效能之分析與比較

OCR辨識最重要的指標是「辨識的正確率」，除了受內部辨識核心引擎系統強度之影響外，而圖檔清晰度、文稿排版樣式、不同字體與語系（如繁體中文、簡體中文、英文、阿拉伯數字及含表格的文件）混合編排的識別成功率，亦很重要。

表4-2、OCR辨識系統分析一覽表

		丹青中英日文文件辨識系統4.5	蒙恬認識王專業版V3.1	清華TH-OCR 2003錄入工廠
操作介面		較簡單	較簡單	較繁複
辨識種類	繁體中文	可，辨識率97%	可，辨識率97%	可，辨識率97%
	簡體中文	可	可	較佳
	英文	可	可	較佳
	中英混合	較差	較差	較佳
	日文	可，辨識率<50%	不支援	較佳，辨識率90%
	表格	較差	較差	較佳
辨識速度		快	快	稍快
輸入格式		*.pcx/ *.tif/ *.jpg/ *.bmp	*.pcx/ *.tif/ *.jpg/ *.bmp/ *.eps/ *.msp/ *.png/ *.psd/ *.tga/ *.wmf	*.tif/ *.bmp/ *.pcx/ *.fax/ *.jpg
輸入格式		*.txt/ *.rtf/ *doc./ *xls./ *slk./ *csv./ *html	*.txt/ *.doc/ *.xls/ *.html	*.rtf/ *.html/ *.txt/ *.jda/ *.wps/ *.pdf

資料彙整：拓展台灣數位典藏計畫

伍、後設資料與資料庫建置

Metadata and Database Establishment

一、後設資料與XML

（一）Metadata釋義與目的

所謂Metadata，在資訊界最普遍的解釋是「資料中的資料」(data about data)，意指與資料相關的描述性資訊，國內翻譯爲「元資料」、「詮釋資料」或「後設資料」等不同辭彙。國際圖書館聯盟協會（The International Federation of Library Associations and Institutions，簡稱IFLA）對Metadata之定義爲可用來協助對網路電子資源的辨識、描述、與定位其位置的資料。另外，較重視Metadata結構性概念者，則解釋作「結構性資料」(Structure Data About Data)，其以「結構」二字區隔Metadata資訊組織方式與全文索引(full-text indexing)，目的在於以結構化項目，經由人工或自動的方式來描述另一物件，而Metadata系統則會包含相關語法，並與所描繪的物件有密切相關之功能性，針對實體或數位化資料做描述，以方便資料的查詢、管理與再利用。

後設資料主要用途在於對無文字敘述的物件，例如實體的書畫、雕塑品或者數位影像、聲音、視訊資料以及平面書籍等提供檢索功能，其眞實涵義在於針對資訊的內容與外觀等特性作適當性的描述，就它的意義和功能來說，其實就是一種電子目錄(electronic catalogue)，而編制目的即爲描述資料的內容和特色，進而達成資料的檢索。在兼顧後設資料標準、實際著錄需求與資訊系統投資的情況下，後設資料標準並不適合當作各單位共通的著錄規範或資料庫規格，而比較適合做爲某特定領域典藏資料交換與查詢介面的標準。因此各單位可保留各自所需的著錄項目，再透過對應關係轉爲領域內共通的後設資料標準交換格式來交換典藏資料，才可達到後設資料標準國際化的目標。

後設資料約可分爲兩類，一種類型爲「描述資源或知識的資料」，此類後設資料並無明顯的標誌或符號，而是一種組織或表達知識的架構方式，例如日常生活中文書編撰所使用的文章組織架構與編排格式皆屬之。另一種類型爲「結構化與半結構化的描述資料」，意指資料是以電腦能了解的結構方式所表達，例如資料庫內所定義的欄位資料就屬於結構化描述資料，而可擴展標記語言（Extensible Markup Language，簡稱XML）與超文字標記語言(Hypertext

Markup Language，簡稱HTML）等則爲半結構化描述資料，可提供使用者有彈性的資料表達結構。

就後設資料分析的模式而言，數位技術研發與整合計畫之後設資料分析小組建議，從人、事、時、地、物五個角度來思考後設資料應包含哪些著錄項目，因此應結合與典藏物品本質相關的資料，及外在資料兩者間的資訊關係，以分析後設資料應包含哪些著錄項目。同時透過管理(administration)、取用(access)、保存(preservation)、應用(use of collections)等四個層面去思考建立後設資料的用途與後設資料使用者之需求，以使後設資料的分析盡可能包含各層面的需要。後設資料應滿足以下需求：

1. 促使系統互通，而不僅僅是提供摘要性資訊。
2. 當越來越多的資訊被電子化時，後設資料模組應能讓電腦連接資訊源並自動擷取詮釋資料。
3. 後設資料管理系統應能定期核對原始資訊源，以確保後設資料資訊的正確性。

後設資料可根據其在使用時功能性(Functionality)的不同，分爲管理(Administrative)、描述性的(Descriptive)、保存(Preservation)、用途(Use)和技術性的(Technical)等五大類Metadata（表5-1）。[7]

7 曾欣怡、潘育潔，〈新聞傳播多媒體資料庫Metadata分析研究〉，《中文媒體數位典藏與新聞標誌語言研討會論文集》，台北：數位典藏國家型科技計畫，2005年5月，頁3-4。

表5-1、Metadata功能類型定義及功能

類型	定義	例子
管理的 (Administrative)	資源的管理(Metadata used in managing and administering information resources)	物件權限、位置資訊、版本控制
描述性的 (Descriptive)	資源的描述及識別(Metadata used to describe or identify information resources)	編目資料、超連結、使用者註解
保存 (Preservation)	資源的保存管理(Metadata related to the preservation management of information resources)	資源的實際狀態文件、原件、數位物件的保存文件
用途 (Use)	資源的使用層次及類型(Metadata related to the level and type of use of information resources)	展示紀錄、使用紀錄、內容重複使用及多版本資訊
技術性的 (Technical)	描述系統及Metadata如何運作 (Metadata related to how a system function or Metadata behave)	軟硬體文件、數位化資訊

資料彙整：拓展台灣數位典藏計畫

就新聞主題工作組各計畫進行不同數位化物件而言，後設資料可能包含文字、畫面、聲音以及影像等多媒體資訊，而本文以針對期刊報紙之內容的後設資料作說明，而非內容本身的文字後設資料。物件本身內容的文字後設資料爲文字訊息，其包含則有文字的種類、頁數、文字的形成，以及其他有關章節數目與段落數目等資訊。文字也可以被加以註釋，雖然注釋大多用於聲音和影片資料，然而大量文字資料也需要包含重要資訊的注釋，尤其是以網頁爲基礎的系統，可以利用連結來取得特定被檢視的文字資料注釋。注釋也可以被視爲補充的資料，並且可被視爲一種後設資料。文字資料的重大發展爲國際標準組織（International Organization for Standardization，簡稱ISO）於1986年制訂了標準通用標記語言（Standard Generalized Markup

Language，簡稱SGML）。

因爲SGML，文字資料可以輕易地被標示並且截取出後設資料，可標示出文字資料中所包含的人與發生地點，因此可以用關鍵字來擷取後設資料，SGML後來即演變爲XML。

（二）XML的應用

1.何謂XML

網路上的新聞資料庫若要建立更有效的檢索、或進行跨平台使用，必須讓電腦辨識若干訊息內容的意義。第一個以結構和新興標準來支配後設資料的，就是所謂的可擴展標記語言（Extensible Markup Language, 簡稱XML）。標記(markup)是指在稿件或文章上添加一些特殊記號，以記錄各種不同的資訊，就像在中國古代書籍中打圈批改的眉批，或是平常我們閱讀文章時，會把重點特別註記起來，目的是用來突顯或是註解這些地方，這就是標記的原始概念。

日常生活中，我們在書寫時所用的語言，可以稱爲書面語言，如果在書面語言中爲了突顯某些訊息，而加入一些標記，那麼這種加了標記的書面語言就可以被稱做爲「標記語言」(markup language)。在這裡所說的標記語言，是一種爲了讓電腦能夠處理而設計的標記語言，而所使用的標記，通常選擇具有一定涵義的文字或數字來標記，一般的做法是依據需求，先定義一套助憶的標記，然後將這套標記添加到書面語言中，使書面語言變成標記語言。

全球資訊網協會（World Wide Web Consortium，簡稱W3C）於1998年2月正式公佈了XML的Recommendation 1.0版語法標準。XML掌握了SGML其延展性、文件自我描述特性、以及其強大的文件結構化功能，但XML卻摒除了SGML過於龐大複雜以及不易普及化的缺點。雖然字面上看來XML是一種標示語言，但嚴格來說它是一種「元語言」(meta-language)。換句話說，XML是一種用來定義其它語言的語法系統，這正是XML功能強大的主因。

XML主要有以下優點：

(1)延伸性：可自訂標籤以滿足不同應用的需求，它沒有固定的一組標記，允許使用者自行定義適用。

(2)跨平台、跨程式語言。

(3)利於網路環境下的傳送與使用。

(4)具有提供有意義的標記的能力。

(5)具有共通性與國際化的特性。

(6)結構化：用XML可以定義出文件的結構，複雜度不設限。

(7)具有自我描述資訊的能力：XML除了可使用標記與屬性來描述資料的意思外，也用來確認XML文件結構的正確性。

XML同時也具有以下缺點：

(1)標準尚未成熟。

(2)複雜度較高。

(3)工具軟體的支援度不高。

(4)可定義結構但無法限制語義(semantics)，亦即XML可用來描述文件的結構，但卻無法完整表達這些結構的語義。

2.用於新聞領域的XML[8]

科技與網路的蓬勃發展，使得越來越多新聞媒體利用電腦及網路相互傳播新聞，數位化新聞遠比傳統新聞需要更強而有利的資訊組織方法，以便能夠迅速有效的進行交換、傳遞與分享，因此對於新聞資料的保存及使用也就產生了新的技術與規格，以求能將新聞資源做最佳化的管理典藏，並且透過系統平台讓使用者快速且簡捷的獲得新聞資訊，加速資訊的散播。爲解決數位化新聞資訊組織的問題，許多專用於新聞事件的後設資料格式也就隨之產生，且各

8 林信成、康珮熏，〈報紙新聞數位典藏Metadata 轉換系統之設計與應用〉，頁B2-1

有不同用途。而使用後設資料格式描述新聞事件，可加強新聞的結構性且增加自我描述性，有助於迅速的交換、傳遞與分享數位化新聞。用 於新聞領域的XML簡述如下：

(1)NITF(News Industry Text Format)

由國際新聞通訊協會（International Press Telecommunication Council，簡稱IPTC）所制訂，著重在新聞內文的描述。

(2)NewsML(News Markup Language)

著重封裝多種不同的媒體，用於描述電子出版、傳送、典藏的新聞檔。

(3)SportsML(Sports Markup Language)

用於運動項目紀錄。

(4)ProgramGuideML(Program Guide Markup Language)

專用於廣播與電視新聞節目。

(5)PRISM (Publishing Requirements for Industry Standard Metadata)

由IDEAlliance(International Digital Enterprise Alliance)所發佈，主要是爲滿足雜誌、新聞、書籍和期刊等平面媒體出版者商業需求而設計。

(6)XMLNews

由XMLNews.Org所研擬，主要在描述新聞報導之實質內容，是借用NITF(News Industry Text Format[9])而來。

(7)RSS(Really Simple Syndication)

RSS衍生自Netscape推播技術(Push)，是一種用於互通新聞和其

9 NITF was developed by the International Press Telecommunications Council, an independent international association of the world's leading news agencies and publishers. It is a standard that is open, public, proven, well-used, well-documented, and well-supported. 資料來源: http://www.iptc.org/cms/site/index.html?channel=CH0107

他Web內容的資料交換規格，目前已普遍應用於入口引擎、新聞網站、Blog和WiKi等系統中。

(8) NRMF(News Records Metadata Format)

行政院文化建設委員所制訂的新聞紀錄Metadata格式。

(9) UdnML(UDN Markup Language)

台灣新聞業界聯合報系所訂定的「聯合新聞標示語言」。

(10) XinhuaML(Xinhua Markup Language)

中國新華社所發展的「新華標示語言」。

(11) CNTF(Chinese News Text Format)

由中國報業協會制訂的「中國報業電子新聞文稿格式」。

二、資料庫建置

資料庫的建置，初期在處理Metadata的統合工作、建置具有學科原理的分類架構等基礎建設，必定會耗費較大的心力，需要結合涉及內容領域之知識專家與資訊科技人才。

（一）數位化資料儲存與管理

由於數位化的格式種類多，且早期資訊儲存技術不發達時，報紙儲存方式除了原件之外，大多製作成爲微縮膠卷，但卻也因使用頻繁而受磨損。而目前在儲存技術的進步與發達之下，則可依據不同的目的，儲存與備份設備如DVD、CD-R、磁碟陣列及光碟櫃等多種形式；而數位化的品質需有專業人員定期檢驗，確認無誤後再轉入資料庫中，以提供使用者利用。惟在將網站資料庫開放之前，需先將版權問題妥善處理，以免觸法。

（二）撰寫規格需求書

在設計資料庫前，一般也會先撰寫需求規格書，尤其是當資料庫外包給廠商做時，需求規格書是取得共識的好方法，能讓資訊技術人員能正確的分析、規劃、設計出內容知識專家所需的典藏系統，從事Metadata分析與資料

庫管理之人員需要有良好之溝通，方可避免Metadata分析的結果與資訊系統分析產生矛盾的現象。

（三）資料庫設計

由於多媒體資料庫未來收錄內容繁多，一般的檢索條件有時仍會導致搜尋結果資料量過於龐大，對於進階搜尋的部分，可設計「搜尋結果範圍內查詢」的功能，以節省搜尋時間，提高精確度，也就是讓使用者下好關鍵字，並得到第一次檢索資料條列後，讓系統使用適當的程式來進一步發問，使用者再經由系統提供的答案，繼續搜尋自己想要的資料；分類架構的管理系統本身，不管是在分類的哪一個層次上，都要預留「修改」、「增加」、「刪除」等功能，使得編輯人員可以依照資料所呈現出的樣貌，隨時修改分類架構，甚至可發展爲離散式資料庫：每一筆資料的分類作業與管理系統是連動的，可讓編輯人員藉由開啓另一個視窗，直接在「分類管理」系統中，修改類目名稱，因此只要分類架構改變了，那麼資料庫中所有資料與欄位都會即刻改變分類位置，可能會有新增類目或者類目合併的狀況。

（四）資料庫維護

若是定期持續更新典藏品的資料庫，其資料庫維護必須由專人隨時待命，讓資訊內容持續更新與即時回訊，使系統安全維持穩定運作，以利資料庫的維護工作。這方面必須特別注意資料庫管理人員的工作交接。

陸、委外製作

Outsourcing Management

一、委外作業

委外服務[10]是「將組織運作需要部分（非關鍵功能）以合約方式交由外面服務者負責」。因應時代環境之變革，委外服務的定義擴大為：「假若有一份工作，外面的組織能做得比組織本身更有效率而且便宜，則此份工作應由外面的組織來做，假如組織本身能將此工作做得較好，則此工作應該保持自製。」以本數位典藏與數位學習國家型科技計畫下之機構單位—國家圖書館而言，委外是指將館內連續性生產流程中所涵蓋的各階段作業步驟，透過合約的簽訂，由館方轉包全部或一部分予外部機構或廠商代為處理。

政府為有效提升委外專案的執行績效及品質，由行政院研究發展考核委員會委託中華民國資訊軟體協會編撰「行政院所屬各機關資訊業務委外服務作業參考原則」[11]，提供委外相關之招標文件、契約、服務水準作業規範及經費計價標準等作業規範，例如：委外作業流程圖（圖6-1），除此之外，亦蒐集各機關現行招標資訊及國內外案例經驗分享，藉由該網站之推動，以建立公平、公開且透明的委外作業規範。

本指南所針對之數位化物件為期刊報紙，其間接原件為微縮膠捲/片，因此有關微縮資料之委外製作部分，可參考「微縮資料數位化工作流程指南」。與期刊報紙相關計畫中，具有委外經驗者為『北平「世界日報」內容數位化開發計畫』、以及國家圖書館期刊報紙典藏數位化計畫，前者計畫之報紙原件存於北京圖書館，繼取得微捲複製片後，因執行單位（世新大學）只有微縮膠捲閱讀機（具閱讀及列印功能），並無將微捲轉製成數位化圖檔的機器設備，因此轉製影像部份改以委外方式辦理；後者計畫因當初報紙原件拍攝成微捲時，有少數部分狀況不佳，影響後續委外廠商數位化品質，必須重新調閱原件掃描，此舉對館方整批作業模式造成困擾，因此便改而採取直接拿報紙原件委外進行掃描。

---

10 朱碧靜，＜圖書館館務委外之決策與管理探討＞，《大學圖書館》第2書第2期，1998。

11 中華民國資訊軟體協會所編撰之政府機關資訊委外作業參考手冊，檢索:2009年4月，http://web.cisanet.org.tw/cgi-bin/big5/cisa/aa02。

一般而言，委外方式大致上可分爲以下兩種作業模式[12]，各機構單位可依照本身設備或資源情形斟酌考量之：

（一）授權外製：將作業委由承包商處理完成，此種作業方式可以節省各單位人力、物力及空間設施等資源，但亦有監督不易之憂。

（二）派員駐館：代理商或承包商派員至圖書館協助館務，以解決人力不足之問題，此種作業形式較容易控制品管與交期。

12 同上註。

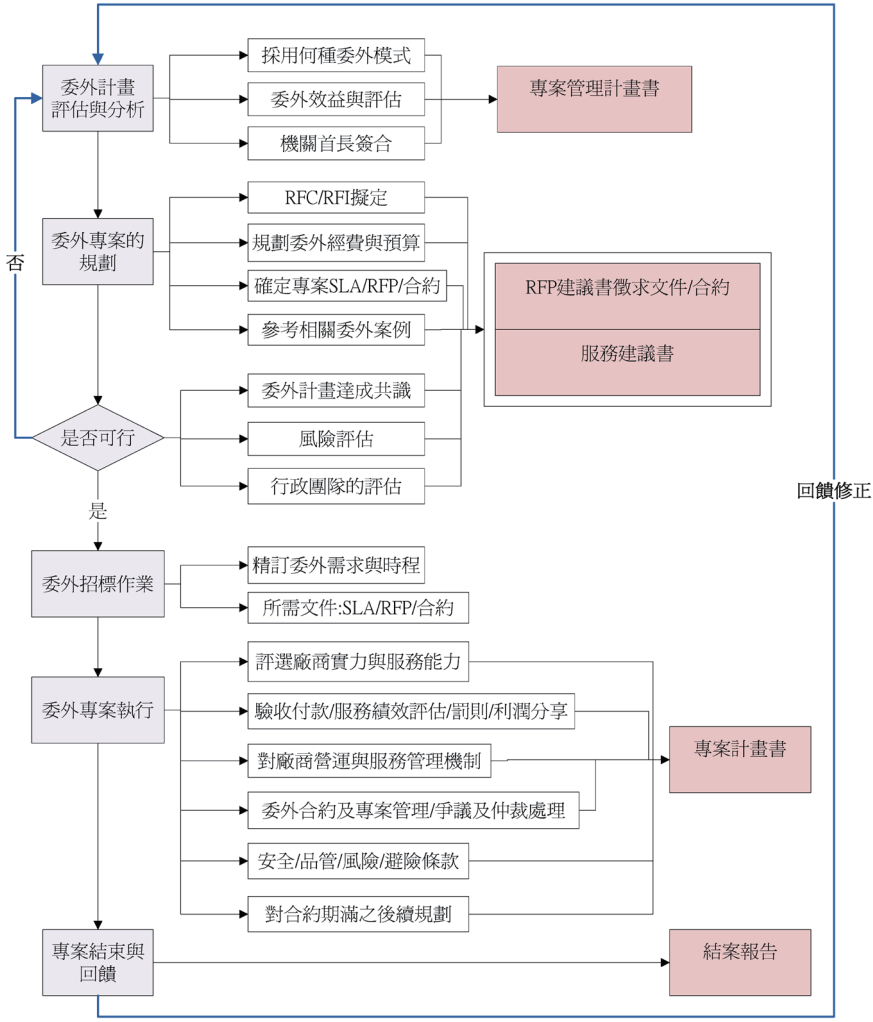


圖1-1、委外專案之作業之流程圖

資料來源：政府機關資訊委外知識網

對所有進行數位化工作的機構單位來說，如何以最低成本取得高品質的委外服務實屬一大考驗，在預算有限、人力及設備資源不足的情況下，仍必須提升作業績效並避免委外失敗而導致原件毀損或必須重新招標等，而透過與委外廠商的合作，機構單位依然得負起整體規劃、監督進度、管控執行方法、評估及修正等責任。因此應審愼評估數位化工作該以自製或委外的方式進行，而決策之結果正確與否亦將影響各機構單位整體的發展。

二、委外執行

（一）制定契約書

經過招標程序之後，得標之委外廠商必須依照委託機關所擬定之契約書進行作業，通常該時期廠商會針對委託機關的需求、期望、細部工作範圍等，進行詳盡的溝通，並依合約規定陸續交付各項文件及成品；而委託機關則必須驗收廠商所交付的文件、成品，進一步作審視與確認，並回應廠商所提出的需求，斟酌擬定相關之配合決策，或者定期召開工作會議，以掌握工作進度以及品質管理，排除任何可能延誤工作進度的因素。

制訂契約書之重點在於界定委託機關與委外廠商雙方之間的權利與義務，該份契約在簽約當時所規範之事項也許僅爲一些原則性，或不因時空，情境改變而改變的事項，因此，許多委外契約的內容也可能在雙方同意下，進一步協議作調整或增刪。

（二）驗收標準

1.檢驗流程

在進行驗收之前，應詳細閱讀契約書中明訂之驗收項目，例如：原件上既有的污點是否要保留或進行後製編修，而在數位化或驗收過程中更應清楚紀錄原件與影像檔有瑕疵或異處，以利後續重新製作或影像編修作業能順利進行。另外，各機構單位進行驗收時應爲影像品

質之二校，在此之前，委外廠商必需先進行初步校驗，且爲了確保影像品質，一校與二校應逐筆與原件對照校驗，以避免發生疏漏，在確認無誤之後，方進行燒錄存檔作業，以減少人力負擔及資源浪費。下頁圖6-2即爲一般數位化影像品質檢核流程圖。

2.驗收基準

根據《數位典藏技術彙編》所彙集資料，其中國家圖書館「古籍原書暨微縮資料轉製影像作業契約書」中明訂，驗收時除了核對交付清單所列數量及項目是否相符外，檢驗影像品質之基準也必須依照中國國家標準(CNS)2779 Z4006（數值檢驗抽樣程序及抽樣表）之規定，採用III 級一般檢驗水準進行驗收。而關於驗收基準，通常各委託機關可要求委外廠商製作出符合標準的數位影像檔案，使驗收人員在進行品質檢核時有所參考，也可作爲日後驗收的依據。

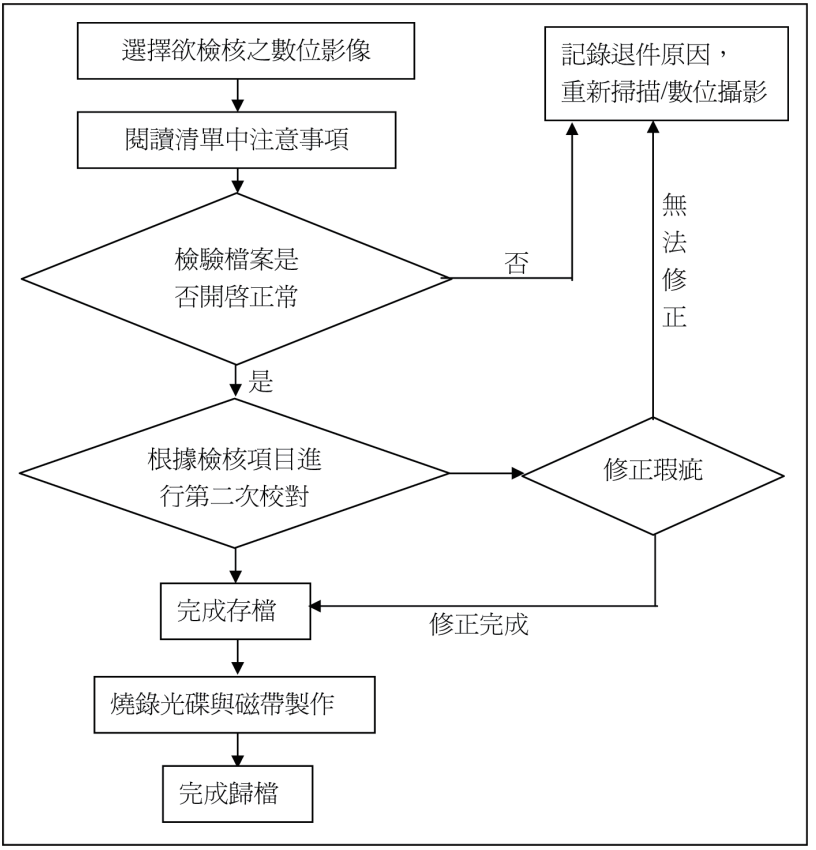


圖6-2、一般數位化影像品質檢核流程

資料來源：陳雪華、項潔、吳海如《國家檔案數位化影像品質之研究》

（三）品質管理

完成數位化工作流程規劃後，通常期盼後續作業能順利進行，並且製作出符合品質需求的數位影像檔。而數位影像品質的控管可以從工作流程、教育訓練以及委外作業之溝通與協調等三方面作探討，分別描述如下：

1.工作流程控管

根據過內數位典藏相關機構單位之經驗，工作流程控管模式主要分爲兩種：

(1)填寫紙本表單

各單位將其工作流程依步驟製成表單，執行各步驟之工作人員簽名以示負責，且採取相互審查的方式相互確認，以使每個環節都能夠正確無誤。

(2)電腦輔助控管

雖然各機構單位仰賴電腦系統的程度不一，但此方法其實具有較高效率，其能使工作流程一目了然、責任的歸屬清楚，更可進行統計分析，以利資源作有效配置。以下即爲電腦系統進行流程控管的優點：

A.追蹤工作人員狀況，針對重作率高的工作人員進行溝通，瞭解其工作上的困難並進行排除。正確率高的工作人員，也可請其分享工作經驗，提升工作團隊效率。

B.記錄常發生的品質問題，確實瞭解品質有瑕疵的原因，據以修正工作規範，並於教育訓練時加強此段工作能力上的訓練。

C.以電腦系統的方式控管工作流程，可以減少紙本作業繁複標記可能造成的錯誤。不同的工作於同一平台完成，也可降低錯誤發生率。

D.若另外與實體典藏系統進行結合，亦可強化檔案調閱的管理。

建議各單位如欲發展各自的流程控管系統，至少應結合「掃描／數位攝影」、「品質檢核與光碟」「磁帶製作」等工作，並且包含下列項目：

步驟項目	說明
掃描/數位攝影	1. 在工作清單中選取欲數位化的檔案時，系統可自動顯示出該原件數位化時特殊注意事項。 2. 可於系統中進行數位影像的修改，如歪斜、明顯污點、對比等。 3. 經過修改的數位影像可記錄其修改的項目。
檢核	1. 選取欲檢核的檔案後即可顯示對應的檢核項目。 2. 可於系統中進行數位影像的修改，如歪斜、明顯污點、對比等。 3. 經過修改的數位影像可記錄其修改的項目。 4. 若有退回重新掃描/數位攝影的檔案，可註記其退回的原因。
燒錄	1. 可顯示燒錄進度。 2. 可自動確認檔案是否可正常開啓使用。 3 完成燒錄後可印製對應之標籤。

表6-1、步驟項目說明

2.教育訓練

一般說來，委外廠商的人員流動率較難掌控，其多半雇用工讀生處理掃描等作業，因此建議各委託機關應給予適當的教育訓練，包括數位化工作目的、原件搬運及掃描、如何進行螢幕校正、品質檢驗基準等注意事項。盡量利用短期而密集的教育訓練，使新進工作人 員能瞭解其工作的重要性，並且迅速進入狀況。此外，當工作流程改變或數位化設備更新時，也應再進行一次教育訓練，藉由上課講 解和實際操作等練習，確認每一個步驟都有一致規範性。

3.委外作業之溝通與協調

在數位化委外作業中，廠商和委託機關之間雖然已有明訂契約和規範，但也常因認知上不同而產生許多問題。因此，除了充分溝通與協調之外，還可以實際測驗將文字具體化，使得雙方皆能取得

共識，並作爲驗收依據。而爲確保數位化產出之影像品質，各機構單位可依實際需求，現場檢驗委外廠商工作人員是否依照工作規範進並得抽驗影像品質是否合乎製作規格。無論數位化工作流程與規範之訂定多麼周密與嚴謹，皆有可能因種種因素而與期望不符，所以建議應與委外廠商定期作討論與檢討，以協調製作過程之例外處理或雙方配合事宜，並適度修正工作流程，並且依各單位情況而定，定期召開品質與進度檢討會議，以瞭解品檢狀況並掌握進度。

柒、數位內容保護

Digital Rights Management

一、數位內容保護概述

隨著資訊科技的發達，電腦能夠快速且大量地處理數位化資訊，而處於知識爆炸的二十一世紀，網際網路的無遠弗屆更是加速了資訊的傳遞及交流，如同一場新興革命般影響著每個人的生活觀念甚或工作模式，因此，在所有數位資料都得以快速、便利地複製與傳輸時，伴隨著而來的便是著作權保護與智慧財產權等問題，尤其是以現今提倡數位版權的時代，更須謹慎注意非創作性資料的來源及出處。就以本「數位典藏與數位學習國家型科技計畫」而言，各典藏計畫單位皆產出數量龐大且珍貴的數位內容，因其形式有別於傳統的有形著作，是以文字、圖像、影音等儲存媒介存在著，因此也勢必面臨如使用者隨意重製檔案而侵害智慧財產權等問題，所以各內容典藏單位無不希冀透過各種保護機制以防止非法複製及使用，可想而知，如何有效保護數位內容將成爲各數位典藏單位相當重視的一個環節。

所謂「數位內容」，根據經濟部工業局數位內容產業推動辦公室之定義爲「影、音、文字、圖像的內容經過數位化，整合運用成產品或是服務，而在數位化的平台上展現」，換言之，所有能以數位方式來儲存、傳播的內容皆爲數位內容，而其所涵蓋的範圍亦非常廣，包括數位遊戲、電腦動畫、數位學習、行動內容、影音內容、網路服務、內容軟體、電子出版、數位典藏[13]等領域，這些資訊的使用關係涵蓋著三種不同的角色與定位：內容提供者(Content Provider)、內容使用者(Content User)以及數位產權(Digital Rights)。數位內容創作所產出的成果屬於無形的智慧財產，通常除了以「後設資料」(Metadata)描述其相關資訊以方向檢索搜尋之外，如何使用與散佈方法也必須加以註記，以避免因複製容易而產生侵權行爲，同時對於使用者也應該要有相對應的身份確認與權限規範，以防止原創者作品受非法散佈或未經授權的侵害。

13 周宣光，〈數位內容產業的發展趨勢〉，http://www.inficyut.edu.tw/950321.ppt。

本章節主要針對數位內容保護與相關權利控管機制作探討，因數位化資訊的取得與重製過程過於簡單、快速而且幾近零成本，對數位內容創作者而言，除了智慧財產權受威脅外，也可能打擊到其創作意願，而非法重製行爲也大大地阻礙了內容提供商生產數位內容的意願，因此，未來關於數位內容保護技術與數位版權管理機制勢必爲相當重要的一門研究課題。

二、數位內容保護機制

近年來，產、學、業界在研究數位內容保護機制的發展趨勢已逐漸強調完整流程的保護，讓數位內容在其生命週期內，從製造開始，包含傳遞紀錄、使用狀態追蹤，以及與資訊安全相關技術（如：加解密技術、數位浮水印、數位指紋、數位簽章及使用者驗證）的整合等，皆同時受到保護，進而建構完整的數位內容保護環境。一般而言，一個完整的數位版權管理技術架構應當具備數位浮水印、密碼學、權利描述語言三大技術，其中數位浮水印技術是將版權資訊植入數位內容中，密碼學技術則是用於限制數位內容的存取，而權利描述語言是提供使用者有關數位內容的使用權利範圍。在此本章節先介紹整合型技術—數位版權管理（Digital Right Management，簡稱DRM），並依序針對數位浮水印、數位指紋、公開金鑰基礎建設、數位簽章及標準權利描述語言等詳加說明。

（一）數位版權管理(Digital Rights Management)

根據國際數據資訊中心（Internet Data Center，簡稱IDC）爲數位版權管理（Digital Rights Management，簡稱DRM）所下之定義如下：

The chain of hardware and software services and technologies confining the use of digital content to authorized use and users and managing any consequences of that use throughout the entire life cycle of the content. DRM is one kind of content protection technology.

其意指：結合硬體與軟體之存取機制，將數位內容設定存取權限，並與儲存媒體聯結，使得數位內容在其生命週期內（自檔案產生至刪除或無法開啓使

用的狀態下），均能受到保護。不管在其使用過程中是否有複製行爲的發生，仍然可以持續追蹤與管理數位內容之使用狀況。總而言之，在數位內容生命週期內，能提供完善保護數位內容、權利之管理技術，則稱之爲DRM。[14]

數位版權管理技術近年來引起廣泛的討論與注意，其所涵蓋的範圍相當大，從數位內容的產生、內容權利之授權、使用者管理與權限控管等，只要有某一環節發生問題，就會產生數位內容被侵用的危機。此概念前身即爲反盜版技術，是種控制數位檔案使用權的技術，其可保護數位內容在散佈、傳遞或進行商業交易時的安全，而基本原理則是利用加密保護，當使用者取得解密金鑰時才能使用數位權限等。

初期數位版權管理的重點在於資料加密與安全性，以解決非法授權盜用的問題，演變至今則包含數位內容記錄、識別、交易、保護、版權所有者的管理以及各種版權利用情況的監測與跟蹤。總括來說，數位版權管理應當涵蓋了控制並追蹤數位內容的存取、管制存取對象、確保重要內容不受更改且於有效期限內不外洩、防止未經授權的使用等，讓數位內容在其生命週期內，透過數位版權管理機制提供較完善的文件存取及使用策略，使軟、硬體在最佳狀態下相互結合，進而保障機密資訊無法輕易被盜用、修改或外流，以確保數位內容受到完整保護，並維護原創者的權益。數位版權管理之所以興起的原因大致列舉如下：

1.保護智慧財產權

原創者創作內容或公司資產機密檔案，必須具備良好之控管機制，使數位內容無法任意被重製或盜用，而透過完善的智慧財產權保護機機制能保障其內容的完整性與價值。

14 楊大廣口述、林雅玲整理，〈數位權利管理的市場趨勢及技術展望〉，《智慧財產權管理》，頁6-11。

2.保護隱私權與機密內容

尤其是重要資訊在傳遞時，往往擔心中途被攔截或遭竊取，因此希冀能透過相關安全管理機制，以掌握資訊傳送的安全性與使用記錄。例如：總統府或國安局之國事機密檔案，可善加運用數位版權保護技術限制使用期限、地點或使用者權限等，以保護機密檔案的隱私性。

3.創造新商機

透過數位版權管理機制的建立，將帶來不同的商業營運模式，而此機制也加入了許多消費者使用的觀念，除了數位內容本身產品之外，也能從其伴隨而來的資訊與服務獲得利益。

4.重視版權與品質

建置一套有效的數位版權管理機制，能同時兼顧數位內容提供者與使用群之間的權利，保護兩者皆不受侵害，如此也樹立消費者尊重正版的授權觀念，也能獲得更多高品質的數位內容資源。

5.統一標準

當具有相當遠景的數位版權管理機制產生之時，各家業者也紛紛積極搶佔市場，希望能建立吸引生產數位內容與使用者加入的機制，朝向整合性的數位內容服務邁進。

數位版權管理主要功能包括數位內容保護加密、使用者認證與授權、數位版權管理發行以及版權安全交易等。而整體數位內容保護的架構之下，此機制對於著作權人提供了相當可靠的智慧財產權保護方案，主要有以下三項保護方向[15]：

1.避免智慧財產權未經授權而複製且使用

2.有效控管智慧財產權

15 陳心渝、李政宏、邱一航、林韋伶，〈數位版權管理機制實作—以數位典藏管理系統為例〉，《第4屆數位典藏技術研討會論文集》2005年9月，頁93-100。

3.偵測及追蹤侵權行為

在數位版權管理流程裡，數位內容可透過加入浮水印、設定數位權限、加密保護等技術而成為受保護的數位化資料：[16]

1.浮水印：請參照下一節「數位浮水印」之介紹。

2.設定數位權限：包含存取、播放（使用）、內容複製、編輯、備份存取、列印、刪除、出借有效期限、使用狀況追蹤…等。

3.加密保護：以加密保護程序識別使用者的身分，以憑證下載使用權限的方式，獲得解密金鑰解開對應的加密資料，而該特定權限才得以使用數位內容資料，以保護數位內容不被非法盜取，避免不必要的數位資產損失。

（二）數位浮水印(Digital Watermarking)

數位浮水印是將能代表原創者符號或圖騰（如註冊商標、識別標誌）植入受保護的數位內容之中，以期日後發生版權爭議而欲進行侵權認定時，可作為版權歸屬的依據，只要能提出有效證明標記者便是合法擁有者，因此可對想要逾權使用的人造成一定程度的嚇阻作用，而若以此技術為保護核心的架構下，數位浮水印的有效性及強健性，將是整體數位內容保護是否成功的關鍵因素。

依照浮水印的可見程度，可分為顯性與隱性兩種：顯性浮水印(Visible Watermarking)在視覺上是可察覺的，具有宣示及嚇阻作用；而隱性浮水印(Invisible Watermarking)則是視覺無法察覺的，具有版權保護及安全作用，而一般所稱的浮水印技術，大部分則是指隱性浮水印。根據數位典藏與數位學習國家型科技計畫—技術研發分項與中央研究院資訊所聯合主辦「2004浮水

16 黃世昆、林宗伯、洪偉能〈數位內容保護與追蹤機制〉，http://datf.iis.sinica.edu.tw/Papers/2002datfpapers/sessionD/D-1.pdf。

印技術評比」[17]活動中，測試結果探討可發現關於數位浮水印設計考量因素涵蓋層面如下：

1.透明程度：植入浮水印後，不能影響閱聽人的視覺品質。
2.防禦性：所植入的浮水印必須具有不可偵測的特性。即便浮水印架構已被破解，還必須擁有相對應的解秘金鑰才可盜取。
3.明確性：浮水印應清楚表示版權爲何人所有。
4.強健性：浮水印儘管經過蓄意攻擊，仍能完好存於受保護的數位內容。
5.容納程度：能加入浮水印的多寡程度。此條件通常和透明程度的要求背道而馳。

通常浮水印是針對不同的需求而具有不同類型的應用方式，例如：版權保護、驗證及追蹤[18]等，如下所述：

1.版權保護：在版權上發生爭議時，事先植入的浮水印可辨識所有權者。
2.驗證：用以偵測或察覺出數位資訊是否已遭截取或竄改，以此驗證資料之眞確性。
3.追蹤：另外植入獨一無二的識別碼—數位指紋(Fingerprinting)，此機制與數位浮水印皆同屬資料隱藏(Information Hiding)技術，其能確切找出相對應的紀錄，以便日後追蹤並證明版權被非法盜取之證據。

數位浮水印在過去曾被視爲完整的智慧財產權保護解決方案，如今在各種不同需求要求之下，因現有技術強健性的不足而顯得力有未逮，所以也僅能視爲數位內容安全機制的一部份，作爲財產權宣示作用，但浮水印卻也不是唯一能證明版權所有的證據，而且無法保障數位內容絕對不被竊取，其只能作到事後追蹤而無法防範於未然，屬於較消極的防範措施，也可算是最後

17 蕭人豪、林欣慧、林金龍、林麗虹，〈數位浮水印技術發展現況:以數位典藏計畫為例〉，《第三屆數位典藏技術研討會》2004年8月，頁163-169。

18 同註15。

一道防線，雖然如此，若能善用其嚇阻作用，還是能發揮保護的效果。

（三）公開金鑰基礎建設(Public Key Infrastructure)

本章節曾提及在數位版權管理機制中，通常是以密碼學技術來限制數位內容的存取，而在密碼學系統中則有可運用許多加密及解密的方法以達到秘密通訊之目的。目前研發密碼系統中，依照加、解密金鑰設計的不同可分爲以下兩種方法：[19]

1.對稱式加密法（傳統加密）

加、解密端之使用者雙方皆擁有同一把金鑰且不可外洩，若其中一方欲與多方通訊時，則必須先與對方各自產生私鑰才能傳遞，所以此法金鑰在團體中管理並不容易，然而因該系統執行速度較快， 因此常被用於保護數位 內容物件本身的安全。

2.非對稱式加密法（公開金鑰加密）

加、解密端之使用者雙方皆擁有一對兩把不同的金鑰—公開金鑰、私密金鑰，前者公諸於世，而後者則由使用者持有保管不對外公開，這對金鑰是具相對應關係的數位密碼，只有成對的金鑰對才能相互加、解密，其中一支金鑰對資料加密後，在進行傳輸的過程原始訊息內容。這樣的意義在於某把公（私）鑰編碼過的資料唯有利用其相對應的私（公）鑰才能解碼，而這產生方法也具有不可逆性，以防止有心人士由公鑰推算其相對應的私鑰而竊取資料。

目前研究發展中，以公開金鑰加密技術製作而成的電子簽章稱之爲數位簽章(Digital Signature)，其法定效力相等於一般傳統簽名，甚至具有更大的法

19 廖鴻圖、郭明煌、林金龍、陳貴青，〈Wrapper-based 數位版權管理機制〉，《第五屆數位典藏技術研討會論文集》，2006年9月，頁31-38。

律效力，因一份文件若只手寫簽名於最末頁，則並不能保證其他頁數沒遭受竄改，因此若此份文件爲數位簽章模式，則可經由公立之第三者驗證是否遭修改。所謂的的公立第三者是爲了因應必須有一套制度以管理公開金鑰與使用者身份確認等問題而建立，該制度是以公開金鑰密碼學技術爲基礎而衍生的架構，其稱之爲「公開金鑰基礎建設」(Public Key Infrastructure，簡稱PKI)，其中公鑰存放於認證機構裡，主要用來加密與驗證；私鑰則儲存於使用者晶片中，用來簽章與解密，是目前被公認爲網路通訊及商業交易安全需求中最成熟的方案，而在訊息傳遞與交換的過程中，主要能提供以下四大功能：

1.訊息資料的完整與隱密性

2.鑑識使用者的身分

3.確認交易簽章資料的不可否認性

4.具有法律效力

（四）權利描述語言（Rights Expression Language，簡稱REL）

數位內容其包含實體與權利兩面，實體指的是數位內容本身如何取得、應用加值等；權利則是指有關著作權管理的部分。而前文除了介紹數位內容保護機制之外，在此也介紹表達使用者與數位內容之間權利、義務範圍的權利描述語言。目前較常見的權利描述語言與發展組織有下列五項：

1.XrML(eXtensible rights Markup Language)[20]

XrML爲國際標準組織作爲數位版權描述語言標準，源自於1994年Xerox PARC，是ContentGuard發展爲音訊產業的通用標準。其可供數位化內容的DRM、Metadata、內容管理、內容傳遞及安控等服務，並成爲各式媒體的內容版權管理標準語言，例如電子書、數位出版、數位廣播、音樂、影像、數位電服務、數位電視服務等。目前已有如：Microsoft、Adobe、SONY、HP和Xerox及著名出版商等採用。數位版權的管理可用以對任何具備內容性質的數位資料加上版

20 http://www.xrml.org/about.asp。

權簽章資訊，藉以控制該數位資料的流通和拷貝。

2.ODRL(Open Digital Rights Language)

澳洲地區所延伸出的標準。

3.EBX(Electronic Book exchange)

EBX 技術框架的核心爲「使用許可證」，其描述用戶對於eBook所擁有的權利，主要包含資訊如下：

(1)eBook的唯一識別碼

例如書號ISBN或者數位物件識別號(The Digital Object Identifier，DOI®)。

(2)eBook的加密金鑰

只有可得到閱讀授權的機器才能取出金鑰，通常加密金鑰會以使用許可證擁有者的公鑰加密，只有符合該使用權限者，才可以EBX專有的閱讀軟體解密。

(3)eBook的複本使用權限

此爲作者對eBook所擁有的操作權利，包括能否複製、轉印、可閱讀期限等。

4.MPEG(Moving Picture Expetrs Group)

MPEG爲業界的組織，自1998年發展至今，已陸續有MPEG2、MPEG4、MPEG7、MPEG21等。根據1998年歐盟委員會指出MPEG21是爲了解決下述問題[20]而發展：

(1)數位內容易被複製或修改，收費和偵測非法使用卻很困難；

(2)消費者無法透過符合國際標準的機制，購買合法的數位內容；

(3)版權擁有者無法透過符合國際標準的機制，獲得相對的報酬；

(4)無法百分之百確認數位內容的合法性；

(5)數位內容版權的擁有者不易確認，降低數位內容商品的流通性；

(6)數位化環境中，缺乏安全機制與法律層面的認同；

21 同註17。

(7)買賣雙方合約與條款的確認動作較難，須得面對面才能促成交易；(8)現有的數位內容發展系統仍欠缺整合和安全性。

5. SDMI(Secure Digital Music Initiative)：SDMI 爲The Secure Digital Music Initiative MPEG4(SDMI)帶給全球錄音、消費電子產品和資訊技術工業，以用來保護數位音樂的開放式規格，任何希望使用被保護的格式來創作和傳播音樂的人，甚或以匿名方式發表者，隨著這個規格發表的利基，將不會被限制。

三、現況與未來趨勢

「數位典藏與數位學習國家型科技計畫」首要目標是將國家重要的文物典藏數位化，並且秉持提升人文教育與知識普及的意義，進而鼓勵產業加值，推動社會經濟的發展。然而當數位內容於傳遞過程中卻衍生出許多困擾，如非法侵權，對原創者而言，未經授權而擅自複製或散佈，實屬一大打擊。因此，因應數位內容保護的課題，本文也陸續增修篇幅以介紹數位版權管理機制，若數位內容可受到完整且安全的保護，則激勵原創者繼續創作，且內容提供者也才能無顧慮地開放內容，使得數位內容市場更加蓬勃發展，讓具有珍貴文物蒐藏者將其典藏物品數位化，國家文化的傳承也變得更加有深刻意義。

然而近年來因數位內容日趨熱門，許多新興的數位產業如雨後春筍般紛紛崛起，面對這樣複雜的數位化轉換過程，業者所需投入的人力資源與經費其實也相當龐大，因此國內關於數位內容的商業模式、版權認證及交易機制等仍尚未成熟，目前大部分廠商皆爲獨立發展，其擁有不同的管理機制與不同格式的數位內容，且往往希望各自的數位格式或保護機制能成爲標準或規範，以此提高市場佔有率，造成各家系統之間均不能相容，且數位檔案格式的轉換也相當不便，對使用者而言，已形成許多困擾。以Apple的iTunes音樂商店[22]爲例，從iTunes下載的數位音樂檔案都受Apple的FairPlay技術保護，因此只能在Apple的

22 台灣網路危機處理暨協調中心—技術專欄，〈數位內容保護技術〉。

iPod上面播放；而唱片公司Sony BMG就在去年（2005年11月）因被軟體專家Mark Russinovich揭露其音樂光碟爲了防止使用者盜拷，於是在數位版權管理軟體中採用Rootkit技術，在使用者不知情狀況下潛入電腦，此舉因過度保護機制而引起的挨告風波在當時也鬧得沸沸揚揚。然而，內容提供業者以數位版權管理機制保障自身權益的同時，該如何也兼顧消費者的期望與需求呢？總括來說，數位版權管理機制的最高境界至少必須勝任以下兩種挑戰：

（一）達到眞正數位內容的控管，以保護機密資訊

（二）在正常使用範圍內，使用者感覺不到此系統存在，唯有侵犯授權時，才會出現警告。

因此，如何擬定一個嚴謹且具彈性的數位內容保護機制，以確保原創者的權益不受損，而在不改變使用者原本使用工具的情況下，還能達成數位內容保護之目的，不至於對消費者的權益造成影響，在這樣自由與限制的兩端中如何尋求平衡點，實屬所有未來有意於數位內容產業發展者值得省思之議題。

四、智慧財產權使用與權利歸屬

目前加入新聞主題工作組之計畫日益增多，而典藏品的範圍與種類也相對擴大，從以往單純掃描報紙元件，到保存書信甚至新聞照片，再加上全文輸入，版權對於這些藏品與欲將其數位化以及公開使用的單位而言，無疑是亟需注意的事項，以免造成侵權而不自知。一般常見於新聞主題工作組的版權相關問題如下：

（一）報紙新聞掃描後除保存影像檔之外，另以全文輸入；

許多相關計畫在建立資料庫之時，會因舊報紙新聞保存年限已久，當時印刷的字體或許已模糊而採用全文輸入的方式，或者加以分詞技術方便學術研究上檢索使用，但基本上全文輸入的方式已屬重製，在著作權法第三條第一項第五款中針對重製之定義爲「五、重製：指以印刷、複印、錄音、錄影、攝影、筆錄或其他方法直接、間接、永久或暫時之重複製作。於劇本、音樂

著作或其他類似著作演出或播送時予以錄音或錄影；或依建築設計圖或建築模型建造建築物者，亦屬之。」因此在進行輸入之前，最好先取得針對藏品之授權與重製權，以免資料庫公開使用之後衍生問題。

另因典藏單位多屬學術機構，符合著作權法第四十八條[23]內容者，可主張合理使用，但原則上還是以先釐清典藏品的版權，再進一步取得授權爲保險之做法。

（二）報紙新聞/評論應取得記者或是報社授權？

著作權法第九條第一項第四款規定，單純傳達事實新聞報導之語文著作，不得爲著作權之標的，因此需要先釐清其新聞內容，所謂的「單純傳達事實」意指如：『日經平均指數以上漲112.39點的8859.56點作收』、『今年是1972年首度實施閏秒以來，第24次增加閏秒。』、『東部幹線和跨線列車，明年1月7日開放訂票，西部幹線8日開始訂票』...等，諸如此類僅針對事實做出陳述的報導方式，不得爲著作權之標的，無著作權；但多數的新聞，比例上依然有加入記者對於新聞事件的描述性字眼，此類型則不屬「單純傳達事實」，如：『高鐵春節假期也一口氣加開了308班列車』、『市場憂心中東情勢影響原油供給，國際油價29日一度暴漲超過12%』、『多數民眾也不認爲未來半年是購買耐久性財貨的好時機，例如房屋、車子等』，因不屬單純傳達事實新聞報導之語文著作，所以在進行典藏作業之前需要取得授權。

---

23 著作權法第四十八條

供公衆使用之圖書館、博物館、歷史館、科學館、藝術館或其他文教機構，於下列情形之一，得就其收藏之著作重製之：

一、閱覽人供個人研究之要求，重製已公開發表著作之一部分，或期刊或已公開發表之研討會論文集之單篇著作，每人以一份為限。

二、基於保存資料之必要者。

三、就絕版或難以購得之著作，應同性質機構之要求者。

下一步則是要釐清向誰取得授權，究竟是發行的報社或是撰稿的記者？此部份則牽涉到勞雇關係，根據著作權法第十一條第一項[24]規定「受雇人於職務上完成之著作，以該受雇人爲著作人。但契約約定以雇用人爲著作人者，從其約定。」報社記者除自由撰稿人之外多數爲報社雇用，因此所撰寫新聞報導之著作財產權應屬報社所有，所以如需使用或引用報社相關報導應向該報社取得授權，並標明該則新聞之時間、作者與出處；如社論或是民意信箱投稿之類，則情況與前述新聞報導不盡相同，社論屬評論性質，如爲該報社記者所撰寫，則著作財產權同爲報社所有，但如該社論/報導爲自由撰稿人撰寫，而由報社支付其費用並在無其他特殊條件下，則該文章之所有權爲原作者所持有，而報社可以自由使用該文章，如著作權法第十二條規定；[25]而民意信箱此類的投稿，屬一次性質，如需引用此類文章，最好先取得該文作者之授權同意，因文章之所有權並非報社所擁有，如著作權法第四十一條規定「著作財產權人投稿於新聞紙、雜誌或授權公開播送著作者，除另有約定外，推定僅授與刊載或公開播送一次之權利，對著作財產權人之其他權利不生影響。」

---

24 著作權法第十一條

受雇人於職務上完成之著作，以該受雇人為著作人。但契約約定以雇用人為著作人者，從其約定。依前項規定，以受雇人為著作人者，其著作財產權歸雇用人享有。但契約約定其著作財產權歸受雇人享有者，從其約定。

前二項所稱受雇人，包括公務員。

25 著作權法第十二條

出資聘請他人完成之著作，除前條情形外，以該受聘人為著作人。但契約約定以出資人為著作人者，從其約定。依前項規定，以受聘人為著作人者，其著作財產權依契約約定歸受聘人或出資人享有。未約定著作財產權之歸屬者，其著作財產權歸受聘人享有。

依前項規定著作財產權歸受聘人享有者，出資人得利用該著作。

目前台灣報業發行量在數位文化衝擊下逐漸萎縮，而國內外報社多以電子報形式取代平面報紙，而引用報導則應如同平面報紙，需取得授權並作出明確的標示，若只引用該篇報導之連結，則不在此論。

（三）期刊雜誌內之評論文章，應取得文章作者或是發行出版社授權？

期刊雜誌之引用與授權，與新聞報導接近，如撰文者爲該出版社所聘用，則文章所有權多半爲出版社所有，需向出版社取得授權方得使用；但如文章爲出版社聘用之自由撰稿人，則需詢問其契約關係，方能正確取得該文授權並使用。

（四）新聞攝影之照片應與原作者或是報社取得版權？

報章雜誌相關報導都會配合利用新聞照片，但因新聞照片往往爲報社攝影記者或部份爲新聞稿投送單位附件，或是特約攝影…等，此三種情形版權使用情形均有不同，如攝影記者爲報社所雇用，則依著作權法第十一條第一項規定依雇用人爲著作權所有人；若爲新聞稿件所附之圖片，則最好尋求發稿之單位授權，釐清是否爲該發稿單位有所有權，或是純爲該圖攝影者所有；特約攝影之類圖片取得，應循雇用人（如：報社）跟受聘人（如：攝影記者）所訂定之契約內容判定，如約定其作品歸於雇用人，則須取得雇用人同意授權方得使用，若作品僅供配合文章使用一次，應向原作者取得授權。已獲授權使用的圖片，必須清楚標示圖片來源與作者，或是圖片提供的單位名稱。

（五）如欲典藏私人書信，其版權歸屬應如何釐清？

書信類型文件的問題較上述報導類型多，因書信內容多數涉及個人隱私，且爲寄件者與收件者之間往來，原則上雙方均有著作財產權，因此須取得雙方同意爲佳，以中央研究院社會學研究所「臺灣外省人生命記憶與敘事資料庫」爲例，包含家書投稿及家書附件、家書原件，因多數爲捐贈，但如發生信件之中寄件者與收件者其中一者因地理關係無法跨海取得聯繫，此種情況最好請捐

贈人協助處理，如其中一方因年事已高無法取得同意，則至少要取得繼承信件之所有權人授權，例如：直系或其他親屬，如無法跟繼承人取得聯繫，因隱私權無年限的限制，爲避免爭議，則文件以採取不公開爲佳；而已獲授權且欲公開的信件中，如有涉及個人隱私的部份，也最好不公開，可採部份公開的方式，或是以後設資料說明。

智慧財產權問題看似相當複雜，釐清相關權利歸屬之後則較易獲得解答，雖然多數典藏資料爲學術研究或是非營利用途，但並非如此即可主張合理使用，即使是合理使用也有比例上的限制。而在進行典藏工作前，工作人員最好對各藏品之智慧財產權做出分類，對於授權的年限與使用範圍，也最好有清楚的瞭解，因數位化資料流通迅速，除了需要作好數位內容保護之外，對於資料庫內的內容之著作權背景也需釐清，以防有心人士不當使用而產生訴訟。而授權書內容擬定時也要嚴加注意，無論是去取得授權或是授權給他人使用，均要避免侵權或是權利受損。

捌、設備與成本分析

Equipment and Cost Analysis

一、數位化設備分析

（一）期刊報紙適用之數位化設備

1.直接掃描期刊報紙原件

(1)桌上型平台式掃描器

(2)桌上型自動進紙式掃描器

(3)桌上型無邊縫書籍掃描器

(4)滾筒掃描器

(5)仰面式書籍掃描器

(6)專業多用途掃描器

2.原件製作成微縮膠卷

(1)微縮膠卷掃描器（單頁式/捲片式）

3.原件製作成單張黑白底片

(1)翻拍類

A.數位相機

B.數位機背

(2)掃描器類

A.具備光罩之桌上型掃描器

B.專業多用途掃描器

表8-1、數位化物件與設備對照表

數位化物件	可使用設備
期刊報紙原件	1.桌上型平台式掃描器　4.滾筒掃描器 2.桌上型自動進紙式掃描器　5.仰面式書籍掃描器 3.桌上型無邊縫書籍掃描器　6.專業多用途掃描器
微縮膠卷	微縮膠卷掃描器（單頁式/捲片式）
單張黑白底片	《翻拍類》 1.數位相機　2.數位機背
	《掃描器類》 1.具備光罩之桌上型掃描器　2.專業多用途掃描器

資料彙整：拓展台灣數位典藏計畫

（二）各數位化設備功能簡介

1.掃描器類

(1)桌上型平台式掃描器

此種掃描器爲目前市面上最爲普遍且單價較低之機型，主要用於一般文件及印刷品等影像掃描，少數含光罩之桌上型平台式掃描器則用來掃描照片或正片，其尺寸最大範圍至A3，若掃描物件大於A3尺寸，則必須進行圖檔影像銜接之後製工作，且書背較厚之物件經掃描後，影像圖檔中書縫間的陰影也必須花更多的時間與技術去克服。且每掃一頁均須重複掀開遮光蓋板，將整本書反轉後依序翻頁以進行掃描動作，而此步驟則需注意掃描物件是否裝訂堅固、紙質狀況良好等。

(2)桌上型自動進紙式掃描器

此種掃描器是將掃描資料放置於自動機械裝置，並由機器依序逐張進行掃描，速度較快，其適宜掃描資料類型包括紙張狀況良好、格式尺寸一致之資料，若爲較破舊之古書，則不建議重新拆卸裝訂，以避免花費太多人力、經費及時間，且無法保證書刊是否能恢復原貌。

(3)桌上型無邊縫書籍掃描器

此機型爲改良式桌上型掃描器，有一斜邊裝置助於書籍期刊之掃描，可掃描尺寸爲A4，但爲確保書縫間的影像更爲清晰，在掃描過程中難免施予重力以壓平物件，此動作對裝訂老舊之書籍而言，則容易造成書頁脫落的情形。

(4)滾筒掃描器

滾筒掃描器爲專業印刷用之掃描器，只針對單頁或單張物件進行掃描，解析度可達4800dpi，但掃描速度較慢，且滾筒捲軸的離心力易對原件造成傷害，因此，目前市面上生產率已不高。

(5)微縮膠卷掃描器

此型掃描器有單頁式或捲片式之機款，是專門爲數位化物件

爲微縮膠卷者所設計，其掃描速度快。

(6)仰面式書籍掃描器

此種掃描器以翻拍的理論設計，將掃描資料面朝上放置，並自機器上方投射光源以攝取掃描物件之影像，掃描尺寸可到A2或A1，進行書籍掃描時，可翻動書頁即可，不至於對原件造成太大傷害，機器並隨附玻璃蓋板，以便將書籍壓平，使書縫間的文字影像更爲清晰，掃描速度快。

(7)專業多用途掃描器

此型機器體積較大，兼具翻拍以及傳統掃描之特色，將掃描資料面朝上，並以移動式光源對物件進行掃描，掃描尺寸可到A1，可掃描物件範圍較廣，包含期刊、報紙、書籍、地圖、書畫、紡織品、植物標本、玻璃畫、皮影戲偶、立體物件等，當掃描書籍時，可不需玻璃蓋板而將書縫間的文字影像顯現至清楚可閱讀，掃描速度快。

2.翻拍類

(1)數位相機

數位相機較適合用來翻拍少量的圖像原件，若物件數量過於龐大時，則並不適宜以此方式進行數位化，因其原始設計並非以大量使用而取勝，若使用頻率過於頻繁，則容易造成相機快門的故障率高。當翻拍較大尺寸之物件時，因焦點聚焦於物件正中心，而四周影像則略爲模糊化，此部分的光線處理也較需要專業技術與經驗來控制。

(2)數位機背

數位機背是在傳統的專業單眼相機後方再加掛一個CCD或CMOS感應器，較高階之數位機背可翻拍的尺寸達A1以上，而此款機器也適用於少量翻拍，使用頻率不建議過於頻繁，在光線控制方面也需專業人員操作才能達到較佳數位化品質。

表8-2、數位化硬體設備樣式

設備種類	機器樣式
滾筒掃描器[26]	
桌上平台式掃描器[27]	

26 資料來源：國立歷史博物館委外工作照片。

27 資料來源：中央研究院歷史語言研究所考古組設備。

設備種類	機器樣式
具備光罩的桌上掃描器[28]	
桌上自動進紙式掃描器[29]	

28 同註25。

29 資料來源：全友相片文件掃描器。

設備種類	機器樣式
桌上型無邊縫書籍掃描器[30]	
仰面式書籍掃描器[31]	

30 資料來源：台灣虹光掃描器。

31 資料來源：磁軒多媒體行銷有限公司。

設備種類	機器樣式
專業多用途掃描器[32]	
數位相機	

32 資料來源：磁軒多媒體行銷有限公司。

設備種類	機器樣式
數位機背[33]	

表8-3、硬體設備比較表

適用性 / 機型	掃描尺寸	掃描速度（A2以上）	最高解析度	垂直線是否變形	適合物件	大量生產	傷害情形	機器單價
桌上型平台式掃描器	A3	—	600	不會	單張	可	須拆書、接圖	10萬~20萬
	A4	—	600	不會	單張	可	須拆書、接圖	3,000~6,000萬
具備光罩之桌上掃描器	A3	—	600	不會	單張	可	須拆書、接圖	15萬
桌上型自動進紙式掃描器	A3	—	600	不會	單張	可	須拆書、接圖	20萬
桌上型無邊縫書籍掃描器	A3	—	600	不會	單張、書籍	可	書頁容易脫落	8~10萬
滾筒掃描器	A1	慢	4800	不一定	單張	可	離心力	100萬

33 資料來源：中央研究院歷史語言研究所金石拓片小組。

適用性 / 機型	掃描尺寸	掃描速度（A2以上）	最高解析度	垂直線是否變形	適合物件	大量生產	傷害情形	機器單價
微縮膠卷掃描器	—	—	—	不會	微縮膠卷	可	—	300～350萬
仰面式書籍掃描器	A1	一分鐘內	300	不會	單張、書籍	可	光線過熱、紅/紫外線傷書、玻璃壓力	450～600萬
專業多用途掃描器	A1	一分鐘內	1600	不會	平面物件、可平放立之體物件	可	傷害程度較低	160～350萬
數位相機	視原件大小	快	—	邊角可能變形	不限	不可	光線過熱、紅／紫外線傷書	20～40萬
數位機背	視原件大小	快	—	邊角可能變形	不限	不可	光線過熱、紅／紫外線傷書	100~150萬

資料彙整：拓展台灣數位典藏計畫

本文針對全文輸入OCR之需求，特地於數位化設備中加註說明使用OCR軟體等成本考量，下表即爲此次研究OCR主要軟體之比較。

表8-4、軟體系統一覽表

軟體型號	公司廠牌	產出地點	軟體價位
丹青中英日文文件辨識系統4.5	力新國際	台灣	$6,600
蒙恬認識王專業版V3.1	蒙恬科技	台灣	$3,990
無發行商業版	全景軟體	台灣	無發行商業版
清華TH-OCR2003錄入工廠	清華文通	大陸	$120,000

資料彙整：拓展台灣數位典藏計畫

二、數位化成本分析

數位化成本包含設備、人工、維修等，也依照方案不同而有所變動。數位化方案有計畫單位自行數位化及委外廠商進行數位化。本文先以單位自行數位化方案爲例說明，因委外方案必須考慮公開招標金額，較前者複雜，故暫不列於此詳述。

（一）數位化成本項目估計

1.掃描設備成本（租用或採購）

2.設備操作所需空間及水電：依照租金乘以使用比例

3.掃描所需人力：所使用人次

=預計掃描總數量/所使用的掃描器每小時可掃描數量/預計完成天數

4.掃描所需人力時間：薪資*時間

5.檢查與重新掃描所需人力：所使用人次

=預計檢查總數量/每小時可檢查數量/預計完成天數

6.檢查與重新掃描所需時間：薪資*時間

7.影像相關資訊輸入建檔所需人力：所使用人次

=預計輸入總數量/每小時可輸入數量/預計完成天數

8.影像相關資訊輸入建檔所需時間：薪資*時間

9.儲存設備成本估計：總DVD張數或硬碟空間之金額

（二）舉例說明

下列以期刊與報紙爲物件進行數位化以計算成本，本文稍略以設備及人工掃描成本爲基礎僅供參考，而人員教育訓練時間、評估試掃品質、後製修圖人力及時間、機器故障維修費用等因素，則暫不列入考量。

1.掃描物件爲裝訂式期刊（A4尺寸）

(1)設備成本：桌上型平台式掃描器（A3尺寸）估計爲15萬元、電腦設備兩台各3萬元，丹青辨識軟體6,600元，預計攤提時間爲三年

(2)人工成本：正職掃描及辨識人員各一人

（一天實際工作六小時，月薪3萬元）

(3)掃描速度：規格爲全彩、300dpi；A4尺寸一頁掃描速度爲2分鐘（含人工翻頁之時間），則一人一小時可掃描30頁，每月（20個工作天）產出量約爲30*6*20=3,600（頁）

(4)平均成本：

設備攤提（150,000+30,000*2+6,600）/3年/12月=6,016元/月

每張成本=（6,016+30,000*2）/3,600=18元/頁

2.掃描物件爲現今發行之報紙（A1尺寸）

(1)設備成本：專業多用途掃描器（A1尺寸）估計爲350萬元、電腦設備兩台各3萬元，清華辨識軟體12萬元，預計攤提時間爲三年

(2)人工成本：正職掃描及辨識人員各一人

（一天實際工作六小時，月薪3萬元）

(3)掃描速度：規格爲全彩、300dpi；報紙A1尺寸（一張2頁）掃描速度爲40秒，則一人一小時可掃描3600/40=90張，每月（20個工作天）產出量約爲90*6*20=10,800（張）

(4)平均成本：

設備攤提（3500,000+30,000*2+120,000）/3年/12月=102,222

每張成本=（102,222+30,000*2）/10,800=15元/張

玖、結語

Conclusions

「期刊報紙數位化工作流程指南」希望能對欲進行數位化之機構單位或個人蒐藏者提供明確而清楚的數位化流程與整體概念，期待藉由淺顯易懂的標準作業程序來提升數位化工作效率，並降低初步摸索數位化工作流程的時間，使各機構單位在教育訓練上面花費較短時間與人力且有效率地進行數位化工作。由此工作流程指南與實際進行工作流程作評估與比較，並從中截長補短，以加速並確實掌握數位化之工作進度。對於以期刊、報紙或平面書籍等作爲數位化物件的計畫單位，希冀本文中的光學辨識系統OCR研究與分析能提供執行全文輸入時作參考，以有效運用人力與時間，達成事半功倍的效果。本文「期刊報紙數位化工作流程指南」盼望能有以下效益：

（一）提升數位化進行過程之工作效率。

（二）可作爲教育訓練工作流程手冊之用。

（三）降低數位化進入門檻。

（四）提供數位化硬體設備及OCR軟體系統比較分析，以節省人力與時間成本。

「期刊報紙數位化工作流程指南」因研究範圍有限，故無法針對缺字技術與委外情形做進一步的分析，且礙於OCR辨識軟體發行版本的限制，如全景軟體無發行商業版、北京漢王則無發行台灣版，以致無法使台灣、大陸的光學文字辨識系統作更深入的研究與全面性的評估，此點深感遺憾，然而，本文也希望在現有的設備及軟體技術之下，提供一份適當的數位化工作流程指南以供各界參考。展望未來，因OCR軟體的應用仍持續進步中，印刷體辨識系統已逐漸成熟且應用廣泛，因此，我們仍可樂觀預見多種全文輸入數位化的方式，甚至是手寫體辨識或同步語音辨識的發展，在不久的將來，其軟體及技術皆能趨於穩定且具普及性，以期高效率地輸入大量文字資料，並提供全文檢索及查詢等便利性。

數位化工作流程在整體規劃上必須是嚴謹而縝密的，在執行過程中也盡可能使每一個環節具有連貫性且可調整，並能充分掌握數位化的進度。在科

技日新月異的今天，機器硬體設備不斷地升級更新，或許每一份參考作業流程只能配合當時的設備與技術，但我們仍然寄予無限的希望，對於數位典藏的未來需要更多的努力與試驗，進而不斷修正而找出最適合物件本身進行的數位化方案。

參考文獻

References

專書

洪淑芬著，《文獻典藏數位化的實務與技術》，台北：數位典藏國家型科技計畫訓練推廣分項計畫，2004年2月，初版。

曾逸鴻，《光學文字辨識（OCR）技術整理報告》，台北：國防部電訊發展室，2001年1月。

Konstanze Bachmann，《藏品維護手冊》，劉藍玉譯，台北：五觀藝術管理，2001年。

《新聞主題小組數位化工作流程》，台北：數位典藏國家型科技計畫內容發展分項計畫，2005年1月，初版。

期刊論文

林信成、康珮熏，〈報紙新聞數位典藏Metadata 轉換系統之設計與應用〉，《中文媒體數位典藏與新聞標示語言研討會論文集》，台北：數位典藏國家型科技計畫，2005年5月，初版，頁B2-1~B2-23。

孫正宜、林信成，〈中文報業數位化技術與現況探討－聯合知識庫數位化經驗〉，《2003年資訊科技與圖書館學術研討會論文集》，2003年5月，頁73~93。

莊樹華、張凱達，〈檔案數位影像製作之流程與管理〉，收錄於《檔案季刊》第2卷第1期，2003年3月，頁57~67。

陳同孝、張眞誠，〈淺談影像壓縮〉，《資訊與教育》，第63期，1997年，頁20~27。

陳心渝，〈JPEG 2000及浮水印批次套印技術〉，數位典藏國家型科技計畫技術發展組，2005年。

黃國倫、蕭人豪、李家豪、陳心渝，〈數位典藏系統缺字處理及應用〉，《第三屆數位典藏技術研討會論文集》，2004年8月，頁79~85。

黃耀輝，〈淺談中文字及其輸入、辨識之比較〉，《中研院計算中心通訊》，第14卷第25期，1998年7月，頁233~234。

曾士熊，〈中文輸入法概述〉，《中文輸入法專題》，第13卷第8期，1997年4月。

曾欣怡、潘育潔，〈新聞傳播多媒體資料庫Metadata分析研究〉，《中文媒體數位典藏與新聞標示語言研討會論文集》，台北：數位典藏國家型科技計畫，2005年5月，初版，頁B3-1~B3-43。

廖運尙，〈國史館採用無失眞壓縮實作經驗談〉，《國史館館刊》，第35期，2003年　12月，頁184~200。

謝育平、吳政泓、項潔，〈可攜式字集資源架構—用以解決缺字問題〉，《第三屆數位典藏技術研討會論文集》，2004年8月，頁71~78。

林淑芬、宋美珍，〈期刊報紙數位化問題淺探〉，《國家圖書館館刊》，第1期，2002年6月，頁197~213。

朱碧靜，〈書館館務委外之決策與管理探討〉，《大學圖書館》，第2卷第2期，1998年4月。

楊大廣口述、林雅玲整理，〈數位權利管理的市場趨勢及技術展望〉，《智慧財產權管理》，第35期，2002年12月，頁6~11。

陳心渝、李政宏、邱一航、林韋伶，〈數位版權管理機制實作—以數位典藏管理系統爲例〉，《第四屆數位典藏技術研討會論文集》，2005年9月，頁93-100。

余顯強、楊曉農，〈數位權利管理應用於歷史性新聞報紙之架構研究〉，《第四屆數位典藏技術研討會論文集》，2005年9月。

廖鴻圖、郭明煌、林金龍、陳貴青，〈Wrapper-based 數位版權管理機制〉，《第五屆數位典藏技術研討會論文集》，2006年9月，頁31-38。

蕭人豪、林欣慧、林金龍、林麗虹，〈數位浮水印技術發展現況：以數位典藏計畫爲例〉，〈第三屆數位典藏技術研討會論文集〉，2004年8月，頁163-169。

項潔、陳雪華、鄭惇方、鄭雅惠，〈數位典藏加值應用之探討〉，《圖書資訊學刊》， 第2卷第1期，2004年6月，頁1~17。

林禎吉、賴溪松，〈數位浮水印的技術〉，《資訊安全通訊》，第4卷第3期，期，1998年6月。

余顯強，〈歷史性新聞報紙數位權利管理之研究〉，《圖書資訊學刊》，第2卷第1期，2004年6月，頁73~83。

謝東明，〈電子簽章及電子憑證應用介紹〉，《中華電信研究所—電子商務/交易人才培訓課程》。

吳韻宜，〈數位視訊內容服務版權使用介紹〉，《數位視訊多媒體月刊》，2005年5-6月，頁4~5。

謝志祥，〈數位著作權管理是個幻想〉，《數位時尚》，2006年10月。

陳昭珍，〈數位出版發展現況與趨勢研究〉，《圖書館學與資訊科學》，第30卷第2期，2004年10月，頁107~115。

何建明，〈網路上的缺字解決方案〉，中央研究院資訊科學研究所。

張文村，〈數位內容的機會與展望〉，《資訊尖兵雜誌》，第180期，2003年4月，頁127~130。

規範

國家圖書館，〈數位資料委外製作需求規範〉，《數位典藏技術彙編》，2004年，頁2.3.6.1~2.3.6.12。

中央研究院資訊科學研究所，〈數位化掃描與文字辨識〉，《數位典藏技術彙編》，2004年，頁2.2.13.1~2.2.13.17。

施文祥，〈我國數位內容產業環境發展重要議題〉，《2004台灣數位內容產業白皮書》，頁5-2~5-16。

網路資源

林淑芬，〈期刊文獻資訊網新服務－「全國報紙資訊網」及「國家圖書館期刊影像資料庫」上線服務〉，民國92年2月。http://www.ncl.edu.tw/pub/c_news/92/05.html。

范紀文、何建明，〈數位典藏系統與工具—輕鬆建立屬於您的典藏管理系統〉，PNC 2000年數位典藏及TEI研討會。http://pnclink.org/events/2000dlm/news.html吳政叡，〈Dublin Core繁體中文譯介〉，http://dimes.lins.fju.edu.tw/dublin/。

CBETA中華電子佛典協會，《CBETA電子佛典集成》，2005年2月，http://w3.cbeta.org/index.htm政府機關資訊委外知識網，行政院研究發展考核委員會，http://www.rdec.gov.tw/。

周宣光，〈數位內容產業的發展趨勢〉，http://www.inf.cyut.edu.tw/950321.ppt

邱郁芬，〈數位內容保護技術〉，台灣網路危機處理暨協調中心—技術專欄，2006年1月，http://www.cert.org.tw/document/column/show.php?key=97。

黃世昆、林宗伯、洪偉能，〈數位內容保護與追蹤機制〉，http://datf.iis.sinica.edu.tw/Papers/2002datfpapers/sessionD/D-1.pdf。

項潔、陳雪華、陳昭珍、郭筑盈，〈數位典藏產業商業模式之探討〉，www.lac.org.tw/admin/ArticleFolder/2/75期/3643-75-05.pdf。

邱迪先，〈電子書的發行平台與版權管理〉，永豐紙業—數位出版事業處，http://dataserv.teldap.tw/modules/PDdownloads/visit.php?f_id=5185。

遲嫻儒，〈尋求數位內容的終極保障〉，收錄於《數位內容產業週報》，2005年6月。http://www.digitalcontent.org.tw/e/temp/940608/DC.htm。

唐瀟霖，〈守護數位文檔 數位版權管理：一個商業難題〉，2006年7月，http://big5.xinhuanet.com/gate/big5/news.xinhuanet.com/newmedia/2006-07/07/content_4805798.htm。

林克寰，〈再論DRM〉，2006年9月，聯合新聞網數位文化誌，http://mag.udn.com/mag/dc/storypage.jsp?f_MAIN_ID=1&f_SUB_ID=425&f_ART_ID=44607。

李彥璋，〈DRM應用於數位出版的突破與趨勢〉，全景軟體股份有限公司，http://of.openfoundry.org/rt/Ticket/Attachment/48458/33650/DRM%E6%87%89%E7%94%A8%E6%96%BC%E6%95%B8%E4%BD%8D%E5%87%BA%E7%89%88%E7%9A%84%E7%AA%81%E7%A0%B4%E8%88%87%E8%B6%A8%E5%8B%A2.doc。

程蘊嘉，〈DRM數位版權管理與圖書館Lib News圖書資訊網誌〉，全國高中職圖書館電子報，http://140.111.115.88/epaper/epaper15/E5%9C%96%E6%9B%B8%E9%A4%A8%E6%96%B0%E7%9F%A5%E5%85%A8%E6%96%87.htm。

〈數位典藏的缺字解決方案及應用〉，http://daal.iis.sinica.edu.tw/document/word_intro.doc。

玄奘大學圖書資訊處，〈校園著作權百寶箱-新聞報導有受到著作權法保護嗎？〉，http://hinfo.hcu.edu.tw/front/bin/ptdetail.phtml?Part=IPO_004&Rcg=15。

蔚理法律事務所，〈最新著作權法實用〉，第三章 誰是著作人與著作權人，http://www.weli.com.tw/floors/Chinese/Floor/03Floor/Ch03/P9.htm。

章忠信，〈記者所寫的新聞稿，著作權歸誰？〉，http://www.copyrightnote.org/crnote/bbs.php?board=3&act=bbs_read&id=687&reply=687。

力新國際官方網站，http://www.newsoft.com.tw/。

全景軟體官方網站，http://www.formosoft.com/index.jsp。

蒙恬科技官方網站，http://www.penpower.net/。

北京文通信息技術有公司，http://www.wintone.com.cn/index.aspx。

漢王科技股份有限公司，http://www.hw99.com/。

Iannella, R., Mostly Metadata, A Bit Smarter Technology, http://www.dstc.edu.au/RDU/reports/VALA1998/。

W3C，Extensible Markup Language (XML)，http://www.w3.org/XML/。

附錄

Appendix

《附錄一》期刊影像掃描檔案編碼原則

參與研發單位：國家圖書館
提供單位：國家圖書館
使用單位：國家圖書館
國家圖書館閱覽組（期刊）93 年4月第13次修訂

1. 期刊批次掃描以掃描全本期刊爲原則。即時期刊影像掃描則以單篇爲掃描單位，但皆適用本編碼原則。本掃描之期刊影像需與本館相關資料庫系統自動產生關連，以利影像調閱及文獻傳遞，故編碼過程需配合本館「中華民國出版期刊指南系統」、「中華民國期刊論文索引影像系統」、及「國家圖書館新到期刊目次服務系統」等書目資料的著錄原則。資料庫網址：http://readopac.ncl.edu.tw/。

2. 每本期刊其檔案目錄分爲三層：期刊識別號、卷期總號、出版年月。再以頁碼區分檔名，檔名中英文字母皆爲小寫。

例: 研考月刊第1 卷1 期民國85 年1 月第1 頁

→00000001/1n1/8501/00000001.tif

說明：

2.1 第一層：期刊識別號

共 8 bytes，由中華民國期刊指南系統查出期刊識別號。

例: 研考雙月刊 → 00000001

2.2 第二層：卷期總號

由期刊之封面與書名頁查出該期之卷期總號時參考本館「中華民國期刊論文索引影像系統」及「國家圖書館期刊目次系統」之卷期著錄方式。

卷期總號長度不受限於8bytes，應完整編碼。

2.2.1 凡卷期總號中含有特殊符號或文字者，請以下列英文字母代替之。

卷 : → n 例: 3 卷1 期 3:1 →3n1

合刊 / → x 例: 4、5 期合刊 4/5 →4x5

合刊 - → - 例: 62 卷1-2 期 62:1-2 →62n1-2

總號 = → e 例: 3 卷1 期總號495

3:1=495 →3n1e495

試刊號, 試刊 →t

創刊號 → f

第十章 典藏品識別碼暨數位檔案命名規範

1.10.2.2

特刊 → s 例: 特刊16 → s16 5(特刊) → 5s

復刊 → r 《 r 之後請勿加_ 》 例: 復刊16 →r16

增刊 → a

專刊 → b

革新 → j

索引 → i 例: 1-12 期索引 → i1-12

上 → u 例: 70 期上 70(上) → 70u 去除括號()

中 → m 例: 70 期中 70(中) → 70m 去除括號()

下 → d 例: 70 期下 70(下) → 70d 去除括號()

外編、別冊 → c 例: 別冊1 →c1

外編第四種上冊 →c4u

副刊、附冊、附輯 → g

補編 → h

總目錄 → o(英文)

新，新刊 →y 例: 新3:2 →y3n2

凡無卷期者,請輸入0(數字)

春 → sp 例: 1994 春季號 1994:春 → 1994nsp

夏 → su 例: 87 夏季號 87:夏 → 87nsu

秋 → au 例: 84 秋季號 84:夏 → 84nau

冬 → wi 例: 84 冬季號 84:冬 → 87nwi

2.2.2 凡卷期外有標示學科分冊者代碼如下:

特 刊 → s

例 第5 期特刊 5(特刊) → 5s

人文分冊 →hu

例: 1 卷1 期人文分冊 1:1(人文分冊) →1n1hu

人文社會篇 →hs

科技人文篇 →sh

社會科學分冊 →so

例: 1 卷1 期社會科學分冊

1:1(社會科學分冊) →1n1so

管理科學分冊 → ma

例: 1 卷1 期管理科學分冊

10-2 期刊影像掃描檔案編碼原則（原10-2 更新）

1.10.2.3

1:1(管理科學分冊) →1n1ma

文學院 →li

例: 35 期文學院 35(文學院) →35li

理學院 →sc

例: 35 期理學院 35(理學院) →35sc

工學院 →te

例: 35 期工學院 35(工學院) →35te

管理學院 →ma

例: 35 期管理學院 35(管理學院) →35ma

社會科學學院 →so

例: 35 期社會科學學院 35(社會科學學院) →35so

農學院 →ag

例: 35 期農學院 35(農學院) →35ag

例: 14 期文學部門 14(文學部門) →14li

商學部門、商學・管理部門 →bi

例: 14 期商學部門 14(商學部門) →14bi

理工部門 →sc

例: 14 期理工部門 14(理工部門) →14sc

區域研究部門 →ar

例: 13 期區域研究部門 13(區域研究部門) →13ar

文商理工部門→lb

例: 16 期文商理工部門 18(文商理工部門) →16lb

文學與商學部門 →li

例: 12 期文學與商學部門

12(文學與商學部門) →12li

社會科學學院 →so

例: 35 期社會科學學院 35(社會科學學院) →35so

科技・醫學篇 →st

例: 32 期科技・醫學篇 32(科技・醫學篇) →32st

文史・社會篇→ lh

例: 32 期文史・社會篇 32(文史・社會篇) →32lh

第十章 典藏品識別碼暨數位檔案命名規範

1.10.2.4

軍事社會特刊→ mi

中國系列 → ch

行政革新專號→ ad

2.2.3 凡無卷期編號者，掃描時編碼爲0

2.3 第三層：出版日期

由期刊之封面與書名頁查出該期之出版日期，同時參考本館「中華民國期刊論文索引影像系統」、「國家圖書館期刊目次系統」之日期著錄方式，以

求一致性。出版日期長度不限於8bytes，以詳盡著錄爲原則，如年月日。但須配合以上系統之著錄方式。出版日期採民國紀元。

2.3.1 凡出版年月日中含有“民”字者，請省略不予註記。

例: 民87 年1 月 →8701

2.3.2 年月日間之“・”號逕行省略，不輸入亦不空格

例: 87.01 →8701

2.3.3 下列文字請以英文字母代替之：

春 → sp 秋 → au

夏 → su 冬 → wi

例: 民87.春 →87sp

2.3.4 合刊的年月處理如下

23-24 民76.11-12 → 23-24 76.11-12

民75.12-76.01 → 7512-7601

3. 頁碼(檔名)編碼

頁碼檔名長度一般以8bytes 爲原則，少數特例可長達9bytes。

例如：第100 頁 → 00000100.tif

第100 頁後之插頁→000100_1.tif

以內文頁碼加上“.tif”作爲檔名。如內文第 1 頁，其檔名爲“00000001.tif”。

注意事項：

3.1 內文第 1 頁前面之各頁(即非正文部份)，如封面、目次、封底等，請自封面起依序計數，頁碼第一位加“ α”以區別之，如： α0000001.tif ，α0000002.tif...

3.2 內文後面多出且原本未編頁碼之各頁，請依原文最後之頁碼繼續編號下去即可。

3.3 原文編有頁碼或實際有佔頁碼但未編頁碼之空白頁或廣告頁等請仍依原順序掃入。

3.4 原文未編頁碼且爲多餘之空白頁請予跳過不掃。3.5 內文中之插頁，如原文未編頁碼，則於接續之前頁後加"_"連續編碼。如：

在86 頁至87 頁間插頁 2 頁但未編碼，請以"000086_1.tif"、"000086_2.tif"編號。

3.6 期刊分左、右版次者，以右版爲主爲原則，但仍需先查核期刊索引及期刊目次系統之編碼，以配合之。左版頁碼需以L(小寫)區別，右版頁須以R(小寫)區別。

如：頁左33-左40"，檔名爲"L0000033.tif"~L0000040.tif

如：頁右12-右20"，檔名爲"R0000012.tif"~R0000020.tif

注意：一本期刊不須同時區分左、右版，應取其一爲主，另一版加註區別即可，原則上以加註左版者居多。但須配合國圖期刊索引與目次系統之著錄方式。

3.7 凡標明"頁中"或"中"者請轉換爲"m"。如"頁中13-14"，輸入檔名爲"m0000013.tif"~"m0000014.tif"

3.8 凡正文中每篇文章皆以"1"起頁者，依篇序頁碼前分別以 ()冠各篇序號，頁碼轉換時規則如下：

□□ □□ □□□□. tif

附錄 篇 頁 碼

例: 第一篇1-17 頁

(1)1-(1)17 →00010001.tif-00010017.tif

第二篇1-18 頁

(2)1-(2)18 →00020001.tif-00020018.tif

第21 篇1-18 頁

(21)1-(21)18 →00210001.tif-00210018.tif

頁(A)27-(A)33 → 00010027.tif-00010033.tif

頁(y)1-(y)5 → 00250001.tif~00250005.tif

附錄(α)7~附錄(α)10

→ap010007.tif-ap010010.tif*附錄 → ap

*a、b、c……依英文順序轉換例a=01 b=02 ……z=26

第十章 典藏品識別碼暨數位檔案命名規範

1.10.2.6

3.8.1 前述情形若又有左右起頁之橫直版之不同，則須多加一碼，冠以L或R 分別區分左起頁版或右起頁版，此種編碼會有9 位。頁碼轉換時規則如下：

R□□□□□□□□. tif

L□□□□ □□□□. tif

例: 左起頁 第一篇1-17 頁

L (1)1-(1)17 →L00010001.tif-L00010017.tif

右起頁 第二篇1-18 頁

R(2)1-(2)18 →R00020001.tif-R00020018.tif

3.9 凡正文有兩組頁碼標示者，一組各篇從1 編頁，一組爲總頁碼者，依總頁碼編。但若有兩組總頁碼，一組自1 編，一組是接前期續編者（頁數號碼較大），則依第一頁起始者編，但仍應先查核本館期刊索引及期刊目次系統之著錄方式，或請示館方負責人員。

3.10 凡頁碼編排有疑義應先參考期刊索引系統或期刊目次系統登錄方式，如仍有問題應先請示館方負責人員。

《附錄二》報紙影像編碼原則

參與研發單位：國家圖書館

提供單位：國家圖書館

使用單位：國家圖書館

國家圖書館閱覽組（期刊）民國90年1月第二次修訂

1. 本報紙編碼原則適用於紙本報紙掃描爲影像檔，及微縮捲片(35mm)報紙轉製影像檔之檔案編碼處理。
2. 紙本報紙影像掃描以每日爲單位。
3. 其影像檔案目錄分爲二層：報紙識別號、出版日期。再以版次區分檔名，檔名中英文字母皆爲小寫。

   例：臺灣新生報 民國50 年1 月1 日 第1 版

   → /68600106/19610101/00000001.tif

   3.1 報紙識別號

   檔名長度爲 8 bytes，由本館中華民國期刊指南系統查出報紙識別號。

   例：臺灣新生報

   識別號 → 68600106

   3.2 出版日期

   不限檔名長度，原則上以完整著錄爲原則，並將出版日期轉換爲西元紀元。

   例：民國50 年1 月1 日 → 19610101

   3.3 版次

   檔名長度共8bytes，以一版面單位爲一頁。

   例：第一版 → 00000001.tif

   非定期專刊、增刊、特刊 例: 專刊4 版 → s0000004.tif

   單獨編頁碼之廣告 → ad 例：廣告第8 版→ ad000008.tif

   3.4 編碼實例：

   民生報

   現代生活：a0000003.tif

體育戶外：b0000005.tif

影視娛樂：c0000006.tif

第十章 典藏品識別碼暨數位檔案命名規範

1.10.6.2

影視快訊：cs000007.tif

家庭消費：d0000008.tif

旅遊專刊：e0000009.tif

行程專輯：f0000010.tif

大成報

體育報：b0000002.tif

影劇報：c0000003.tif

經濟日報

金銀島：sb000003.tif

科技島：ss000005.tif

其他專刊：s0000003.tif

同一天第二種專刊 s1000004.tif

同一天第三種專刊 s2000003.tif

China Post

增刊：s0000004.tif

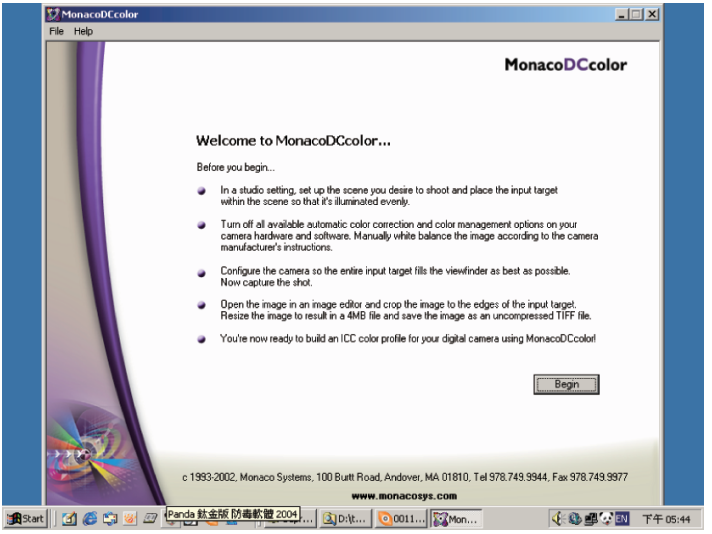
《附錄三》國家圖書館數位化檔案建議格式

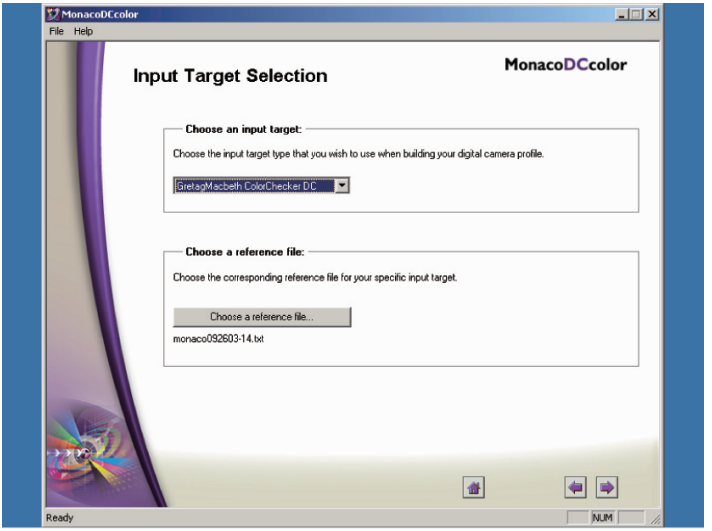
檔案格式	建議規格	說明
文字檔		
資料永久保存格式	檔案格式： TIFF 色調深度：黑白；灰階-每像素8-bits；彩色-每像素24-bits 壓縮：不壓縮 解析度:300~600（或更高）dpi（依原始資料品質及重要性選擇適當解析度，一般印刷品可採300dpi）	將資料數位化典藏，保持原有風貌。提供使用者作重製、壓縮處理或其他圖像處理交換之用。
網路下載格式	檔案格式：JBIG or JBIG2 色調深度：黑白；灰階-每像素8-bits；彩色-每像素24-bits 壓縮：JPEG（灰階壓縮比約10:1，彩色壓縮比約10:1） 解析度或影像大小：150dpi~300 dpi，或影像大小從500×400至1000×700pixels	提供使用者網路上觀看及列印用。
預覽影像	檔案格式：GIF 色調深度：每像素8-bits 壓縮：原生影像至GIF 解析度或影像大小: 72dpi，或影像大小從150×100到200×200 pixels	提供使用者預覽及選擇欄位用。
影像檔		
資料永久保存格式	檔案格式：TIFF 色調深度：灰階-每像素8-bits；彩色-每像素32-bits 壓縮：不壓縮色彩濃度值4.0D以上（color），3.2D（B&W） 解析度： 300~600（或更高）dpi（依原始資料品質及重要性選擇適當解析度，一般印刷品可採300dpi，美術品供複製畫使用建議採600dpi，供印刷出版使用採350dpi）	將資料數位化典藏，保持原有風貌。提供使用者作為重製、壓縮處理或其他圖像處理交換之用。
資料服務／參考格式	檔案格式： JFIF（JPEG交換格式） 色調深度：灰階-每像素8-bits；彩色-每像素24-bits 壓縮：JPEG（灰階壓縮比約10:1，彩色壓縮比約20:1） 解析度或影像大小：150~300 dpi，或影像大小從500×400至1000×700 pixels	提供使用者網路上觀看及列印用。

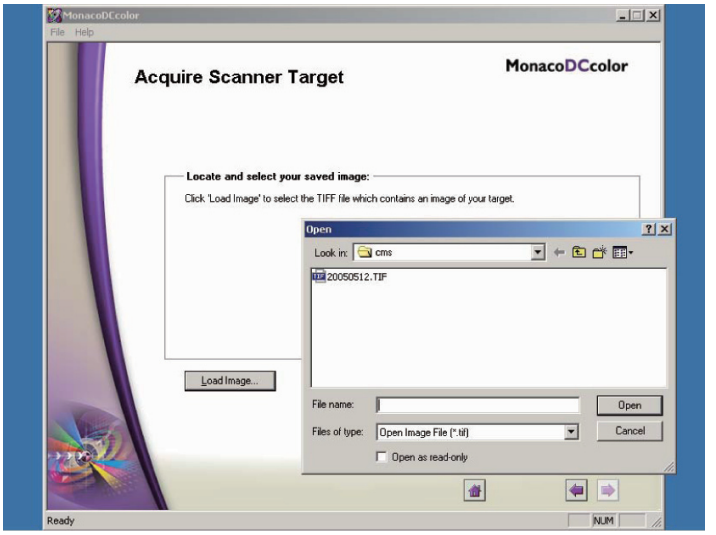
檔案格式	建議規格	說明
縮圖影像	檔案格式：GIF 色調深度：每像素8-bits 壓縮：原生影像至GIF 解析度或影像大小：72dpi，或影像大小從150×100到200×200 pixels	提供使用者預覽及選擇欄位用。

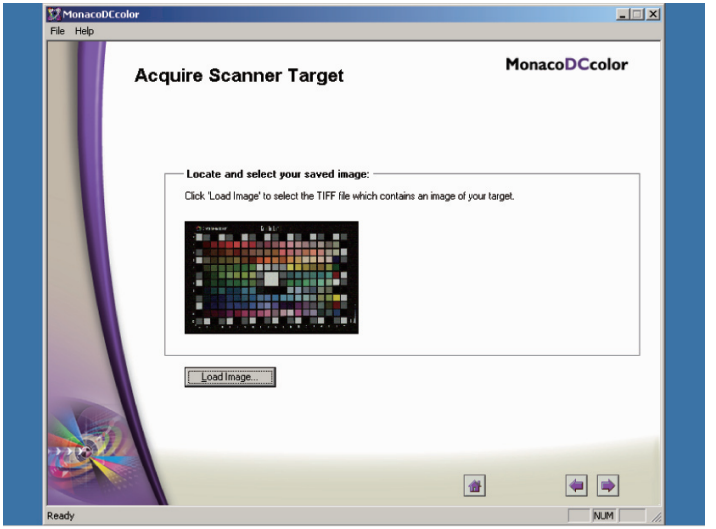
《附錄四》專業多用途掃描器色彩校正流程

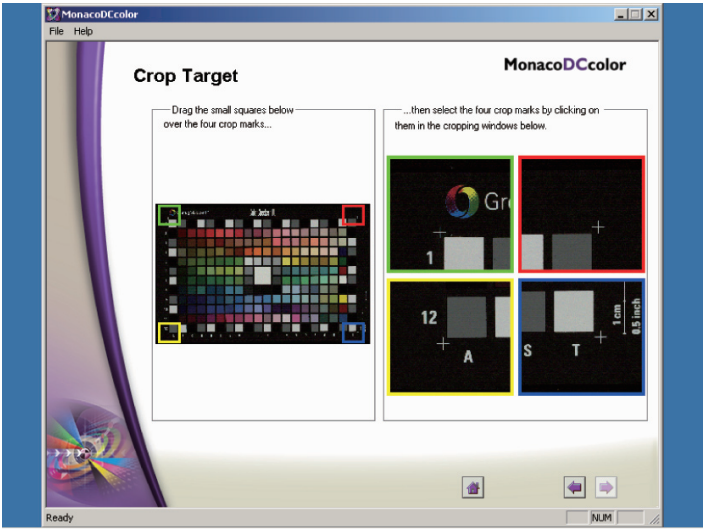
資料來源：專業多用途掃描器代理商 — 磁軒資訊媒體行銷有限公司

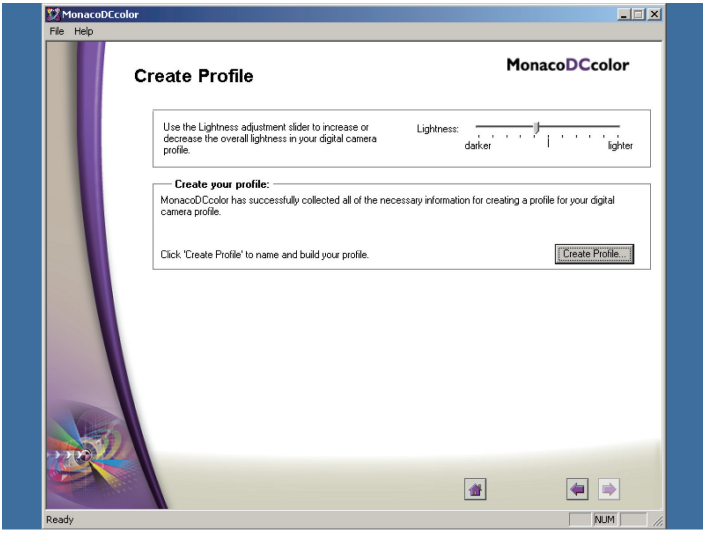





MonacoDCcolor
File Help
Acquire Scanner Target
MonacoDCcolor
Locate and select your saved image:
Click 'Load Image' to select the TIFF file which contains an image of your target.
Load Image...
Open
Look in: cms
20050512.TIF
File name:
Open
Files of type: Open Image File (*.tif)
Cancel
Open as read-only
Ready
NUM


MonacoDCcolor
File Help
Acquire Scanner Target
MonacoDCcolor
Locate and select your saved image:
Click 'Load Image' to select the TIFF file which contains an image of your target.
Load Image...
Ready
NUM


MonacoDCcolor
File Help
Crop Target
MonacoDCcolor
Drag the small squares below over the four crop marks...
...then select the four crop marks by clicking on them in the cropping windows below.
1
12
A
S
T
1 cm
0.5 inch
Ready
NUM


MonacoDCcolor
File Help
Create Profile
MonacoDCcolor
Use the Lightness adjustment slider to increase or decrease the overall lightness in your digital camera profile.
Lightness:
darker
lighter
Create your profile:
MonacoDCcolor has successfully collected all of the necessary information for creating a profile for your digital camera profile.
Click 'Create Profile' to name and build your profile.
Create Profile...
Ready
NUM

《附錄五》辨識技術

資料來源：大葉大學資訊管理系 曾逸鴻助理教授

《光學文字辨識（COR）技術整理報告》

當字元切割完成，即可將每個字元影像丟入辨識引擎做辨認。最基本的辨認方式，即是將字元影像做大小的正規化(Normalization)，然後與資料庫中每個中文字的影像（亦已經過正規化）做模版比對(Template matching)，計算相對位置的顏色是否相同，找出差異最小者即爲辨識結果。此種模版比對方式爲確實掌握文字特性，且所需的記憶體空間較大，比對速度也慢，所以並不被大多數OCR系統所採用。在辨識引擎的內部技術，我們可分特徵抽取、特徵比對與加速技術三部分來描述。

1.特徵抽取

特徵抽取是辨識引擎最重要的一節，要找到最少的特徵，來得到最佳的辨識效果，常採用的特徵可分爲結構特徵與統計特徵，結構特徵包括文字影像內的線段(line segment)、筆畫(stroke)、曲線(curve)、環路(loop)等，通常文字影像需先經過細線化(thinning)，將字元轉成只剩一個像素的寬度，再來抽取結構特徵。經過實驗，利用結構特徵所建構的OCR辨識引擎，較適合辨認印刷清楚且筆畫較少的字元，不太適合於建構商用OCR軟體。統計特徵則將文件影像的像素分佈作分析，利用大量的學習影像來計算特徵向量的平均值與變異度。只要學習影像收集的夠完整、數量夠多，利用統計方式建構出的OCR辨識引擎較能做較廣泛的應用。常採用的統計特徵如下：

(1)筆畫數目(Stroke count)特徵：對於某個參考點(reference point)，往上下左右延伸，計數可通過多少筆畫。此處筆畫的定義爲，延伸線上的點「由白變黑」再「由黑變白」，算是一個筆畫。因此對於每個參考點，我們可得到四個特徵值。

(2)邊緣像素數目(Contour pixel count)特徵：由於不同文件切出的字元影像擁

有不同的筆畫寬度，此特徵乃計數字元的邊緣點數目。

(3)邊緣方向數目(Contour directional count)特徵：考慮邊緣像素，計算四個方向（水平、垂直、左撇、右捺）的邊緣點數目，可得到四個特徵值。

(4)網眼特徵(Cellular feature)：對於某個參考點，往上下左右延伸，計算要延伸多長的距離始可碰到第一個黑點，可得到四個特徵值。

(5)周圍背景面積 (Peripheral background area, PBA)特徵：由字元邊界往內走，走到第一個黑點便停止，記錄其距離，將所有距離累計，即爲此特徵值。由於此種特徵不管字元中心部分，只描述其周圍的白色背景面積，認因墨水過多導致中心部分容易糊成一坨的字元。

(6)周圍背景差異 (Peripheral background difference, PBD)特徵：與PBA類似的計算法，只是此特徵記錄的事兩距離的差異，而不是累計距離。因此，可分辨雖然累積距離相同，但距離先長後短與先短後長的不同。一樣適於辨認中心部分易模糊的字元。

(7)橫越個數特徵 (Crossing counts feature)：由字元左邊界往右邊界走，計算通過的筆畫數，加以累計，垂直方向亦同。

(8)投影特徵 (Projection feature)：將字元影像分別往四個方向（水平、垂直、左撇、右捺）投影，設適當的門檻值，分別在此四個投影圖中，計算投影量高於門檻值的筆畫的個數，當作特徵值。

另外，由於要找到效果很好的特徵不易，一旦找到適當的特徵，爲求更精準描述字元，通常會將字元做切塊，例如邊緣方向數目特徵雖然只有四個特徵值，若先將字元切成8x8塊，在每一塊抽出四個特徵值，則此字元總共可得到8x8x4=256個特徵值。字元的切塊方式有兩種：

(1)等分(uniform)切割：直接以字元的寬或高等距切成數等分。

(2)不等分(non-uniform)切割：先將所有黑點往X軸投影，將投影圖切成數份，使得每一份內的的黑點數目相同，在對Y軸投影，以同樣方式切成數份。此方式切出的區塊大小不同，但較可容許手寫字的變異度，及印刷字的雜訊。

2.特徵比對

特徵抽取完，成爲一個多維的特徵向量(Feature vector) $\tilde{x}$ 後，就要與資料庫中經過學習各字（中文字常用字數爲5401字）的代表特徵向量 $\tilde{\mu}_1, \tilde{\mu}_2, \ldots, \tilde{\mu}_{5401}$ 作比對。由於學習與辨識所採用抽取特徵的過程都一樣，因此，比對方式爲兩特徵向量間，計算相對維度特徵值的差異和。

假設特徵向量共256維，未知字元影像抽出的特徵向量爲 $\tilde{x} = (x_1, x_2, \ldots, x_{256})$ 字元j的代表特徵向量爲 $\tilde{\mu}_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{j256})$，其標準差(standard deviation)爲 ，計算兩特徵向量差異值 的方法有下列幾種：

(1)Minimum distance: $d(\tilde{x}, \tilde{\mu}_j) = \sum_{i=1}^{256} |x_i - \mu_j|$

(2)Euclidean distance: $d(\tilde{x}, \tilde{\mu}_j) = \sum_{i=1}^{256} (x_i - \mu_j)^2$

(3)Cross correlation distance: $d(\tilde{x}, \tilde{\mu}_j) = \frac{\langle \tilde{x} - \tilde{\mu}_j \rangle}{\|\tilde{x}\| \cdot \|\tilde{\mu}\|}$

(4)Modified Mahalanobis distance: $d(\tilde{x}, \tilde{\mu}_j) = \sum_{i=1}^{256} \left[\frac{(x_i - \mu_j)^2}{\sigma_j^2} + \log(\sigma_j^2) \right]$

(5)Li and Yu distance: $d(\tilde{x}, \tilde{\mu}_j) = \sum_{i=1}^{256} \left[\frac{(x_i - \mu_j)^2}{\sigma_j^2} + \log(\sigma_j^2) \right]$

3.加速技術

由於中文字數量極多，辨識特徵取出的維度亦不少，使得如何加速比對過程，也成爲相當重要的研究課題，常採用的方法有下列幾種：

(1)分群法(Clustering)：先以簡單特徵將中文字分成數群，不同群內字元可重複或不重複。未知影像抽完簡單特徵後，先決定此未知影像會落於哪一群，再以較複雜的特徵，與該群內的字元做細部比對。此方式需先決定哪些字元屬於同一群，且不同未知影像只要落於同一群，其細部比對的候選字元均相同。

(2)候選字選擇法(Candidate selection)：此法不必事先決定哪些字元屬於同一群。未知影像抽完簡單特徵，就與所有字元做比對，取前幾名（如前百分之一）再以複雜特徵做細部比對。因此，不同未知影像其細部比對的候選字元必定不同。

(3)分支界定法(Branch and Bound)：前兩種加速法均致力於降低比對的字元數目，因此會降低整體辨識率，此法則設法加速特徵向量的比對速度，

主要用於複雜特徵的細部比對過程。首先，先按照重要性將特徵向量的各維特徵值做重排列，以最重要的幾個特徵值與代表特徵向量作距離的計算，按照此累計距離將候選字元的比對順序重排。在來求出未知字元與第一個候選字元的完整距離，以此爲一門檻值，在計算第二個候選字元以後的完整辨識距離的過程中，每累計一個維度特徵值的差異時，便與此門檻值做比較，若超過門檻值，則未計算的維度也不用再計算，便可跳到下個候選字。若累計完所有維度得到完整距離，仍未超過門檻值，則將門檻值改爲此完整距離。此加速法的最大優點爲完全不會降低整體辨識率。

國家圖書館出版品預行編目資料

期刊報紙數位化工作流程指南/李珮瑛，程婉如作. - - 初版. - - 臺北市：數位典藏拓展台灣數位典藏計畫， 民98.04
面； 公分
參考書目：面
ISBN 978-986-01-8160-9(平裝)

1.文獻數位化 2.文物典藏 3.報紙 4.期刊 5.工作說明書

028.026 98006342

期刊報紙 數位化工作流程指南

指導單位：行政院國家科學委員會
發 行 人：林富士
總 編 輯：邱澎生
執行編輯：林彥宏、林慧菁、高芷彤、林芳志
作 者：李佩瑛、程婉如
審 稿 者：私立玄奘大學資訊傳播學院 郭良文院長兼任所長
發行單位：數位典藏與數位學習國家型科技計畫 拓展台灣數位典藏計畫
地址：115 台北市南港區研究院路二段128號
中央研究院歷史語言研究所
電話：886-2-2782-9555轉288
傳真：886-2-2786-8834
網址：http://content.teldap.tw
Email：content@gate.sinica.edu.tw
封面設計：李維創意工作室
排版印刷：禾古精緻印刷有限公司

中華民國98年4月初版
ISBN 978-986-01-8160-9