期刊報紙全文輸入
工作流程指南

數位典藏國家型科技計畫　內容發展分項計畫

研究助理　程婉如

spinner@gate.sinica.edu.tw

中 華 民 國　95　年　12　月　04　日

目　　錄

壹、引言

民國九十一年一月一日，行政院國家科學委員會依據「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「數位典藏國家型科技計畫」；計畫的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。計畫辦公室下設有五分項計畫，分別爲：內容發展、技術研發、應用服務、訓練推廣及維運管理分項計畫，協助總計畫相關業務的推動。而其中「內容發展分項計畫」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，至民國九十四年止已成立 16 個主題小組，包括：動物、植物、地質、人類學、檔案、地圖與遙測影像、金石拓片、善本古籍、考古、器物、書畫、新聞、影音、語言、漢籍全文與建築等主題小組。

爲因應「內容發展分項計畫」所規劃之主題分類，新聞主題小組於民國九十一年正式成立，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。歷年來參與新聞主題小組進行數位化計畫的機構單位有：本國家型計畫「維運管理分項計畫─出版子計畫」（九十一至九十四年度）、國家圖書館「國家圖書館期刊報紙典藏數位化計畫」（九十一年度迄今）、國立交通大學資訊工程系「電視新聞數位博物館」（九十一年度）、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」（九十四年度迄今）、「世新大學北平世界日報內容數位開發計畫」（九十一至九十四年度）以及淡江大學「台灣棒球運動珍貴新聞檔案數位資料館之建置」（九十三年度迄今）。

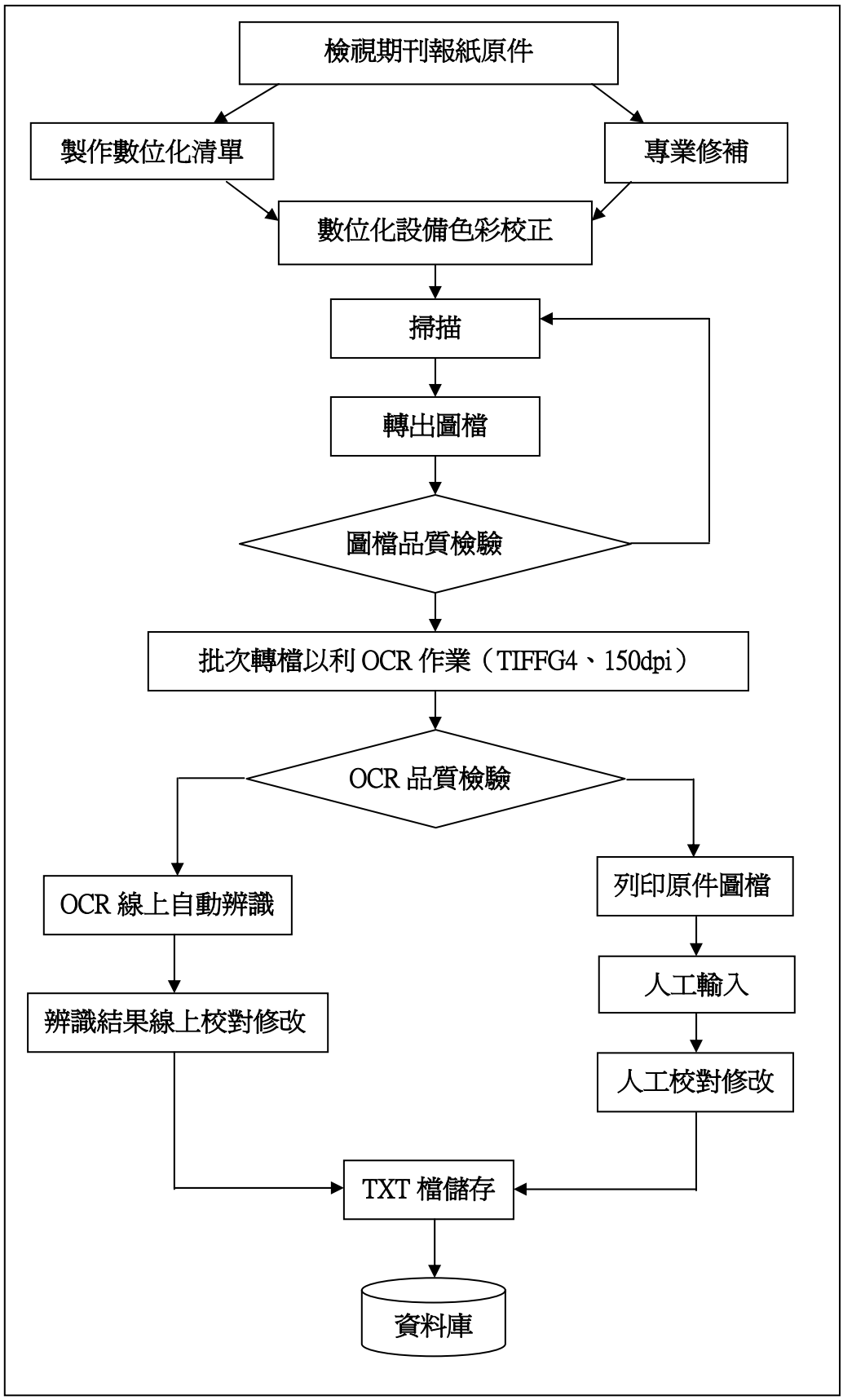
以下簡略說明新聞主題小組內各計畫之數位化工作內容：維運管理分項計畫─出版子計畫主要負責《國家數位典藏通訊》發行，並以 XML 標誌語言加以分析進而建立檢索資料庫；國家圖書館則從事館藏之臺灣地區發行期刊約 1,000 種，與臺灣地區發行報紙約 30 種之數位化工作，其主要數位化工作項目爲期刊典藏影像數位化、報紙典藏數位化、期刊篇目後設資料分析建檔等；國立交通大學資訊工程系則有「電視新聞數位博物館」網路資料庫，典藏中華電視公司新聞影音資料；國立交通大學傳播研究所的典藏有蘭嶼在地刊物《蘭嶼雙週刊》、數位化幻燈片影像資料及蘭嶼地方廣播節目的聲音內容，並建置多媒體資料庫；世新大學資訊傳播學系則取得北平世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；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」。

爲了解各機構單位典藏品內容以及數位化工作程序，內容發展分項計畫亦針對各主題小組進行數位化工作流程之調查，在九十三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；而九十四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，簡稱 SOP）作爲數位化參考依據。

「期刊報紙全文輸入工作流程指南」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。

本文以目前新聞主題小組下的各機構單位數位化計畫爲例作說明：「國家圖書館期刊報紙典藏數位化計畫」主要數位化工作爲影像掃描，並針對期刊部份建置單篇篇目後設資料；「世新大學北平世界日報內容數位開發計畫」則將世界日報微縮膠捲以人工輸入方式建立新聞資料；「蘭嶼原住民媒體資料庫建置與數位典藏計畫」數位化內容之一爲蘭嶼廣播帶，計畫預計以達悟語及漢語全文輸入方式來記錄廣播節目內容；「台灣棒球運動珍貴新聞檔案數位資料館」計畫則將聯合報棒球新聞進一步做更深入的後設資料分析。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）設備與成本分析（七）效益與侷限（八）結語（九）參考文獻等。

貳、數位化工作流程圖


檢視期刊報紙原件
製作數位化清單
專業修補
數位化設備色彩校正
掃描
轉出圖檔
圖檔品質檢驗
批次轉檔以利 OCR 作業（TIFFG4、150dpi）
OCR 品質檢驗
OCR 線上自動辨識
辨識結果線上校對修改
列印原件圖檔
人工輸入
人工校對修改
TXT 檔儲存
資料庫

參、前置作業

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

「期刊報紙全文輸入工作流程指南」擬定數位化物件爲期刊、報紙，而早期報紙除了以原件類型蒐藏之外，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 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

資料來源：漢珍圖書數位公司

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真，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 LZW）、Huffman's Encoding、及變動長度編碼法等，能使檔案體積變小，但

仍然不失真。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，其採用可失真（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真的現象，但 JPEG 的失真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 3.0 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，PDF 不僅保存了原始文件的外觀和完整性，另一方面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使用註解和編輯工具，因此除了微軟所出的 Microsoft Reader 之外，PDF 也是目前世界上最通用的電子書（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 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

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、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眾瀏覽，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，而中文電子書則以方正 Apabi Reader 軟體市佔率最高。

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、光學文字辨識（Optical Character Recognition，簡稱 OCR）、電子報直接轉入資料庫等四種[1]，以下將以新聞主題小組內計畫作爲範例，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，且可提供仿真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。

（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 Internet 普及後才逐漸受到重視。

表 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，民國五十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 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	版面較爲乾淨、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、字體較爲模糊、反差不明顯

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 5、利於 OCR 辨識之圖檔格式

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

三、後設資料之建立

（一）確立檔案格式

目前影視新聞相關之後設資料格式尚無統一標準，這裡指的是在新聞主題小組中不同媒體類型的典藏品，可能需要不同的後設資料加以詮釋；有許多新聞傳播相關主題加入了數位典藏計畫，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，期望能與不同的資料庫與檢索系統加以結合。

（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

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（Commission Intornation De'l E' clairage，簡稱 CIE 或 International Commission on Illumination，簡稱 ICI）於 1976 年將 CIE Yxy 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的話，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：DiGiBook10000RGB）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。值得一提的是，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（圖一）或 120 度翻開面支架（圖二）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 [2]。

圖一、180 度書籍支架

圖二、120 度書籍支架

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，且需考慮字形架構、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、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

原本爲力捷電腦（UMAX）的軟體部門，負責開發掃描器驅動程式與搭售軟體，後來於 1987 年獨立成爲「力新國際」公司。目前產品以影像處理（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的辨識。力新國際也積極以專案方式與各機構單位合作，例如國防部電訊發展室「傳真文件的辨識與分類」、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，並與中央大學資訊工程學系合作，開發 OCR 相關技術，於 1996 年推出「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、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」。

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 相關的名片辨識系統、影像剪輯、網路安全、與虛擬實境四類。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 部分爲專案方式進行整合，並未在市場上發行 OCR 軟體。

4. 清華文通

北京文通資訊技術有限公司（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

北京漢王科技有限公司成立於 1993 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、銀行票據防僞識別認證等系列產品。

（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[3]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再進入辨識引擎做辨認處理。

（B） 版面分析使用技術

a. 區塊分割[4]

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。

b. 區塊型態判斷

上述區塊分割之後，通常會以下列三種區塊特性進行區塊的型態判斷：

(a) 黑白點比例

首先，先計算區塊內的黑白點比例，若黑點遠多於白點，則可能爲影像區塊。

(b) 線段的存在

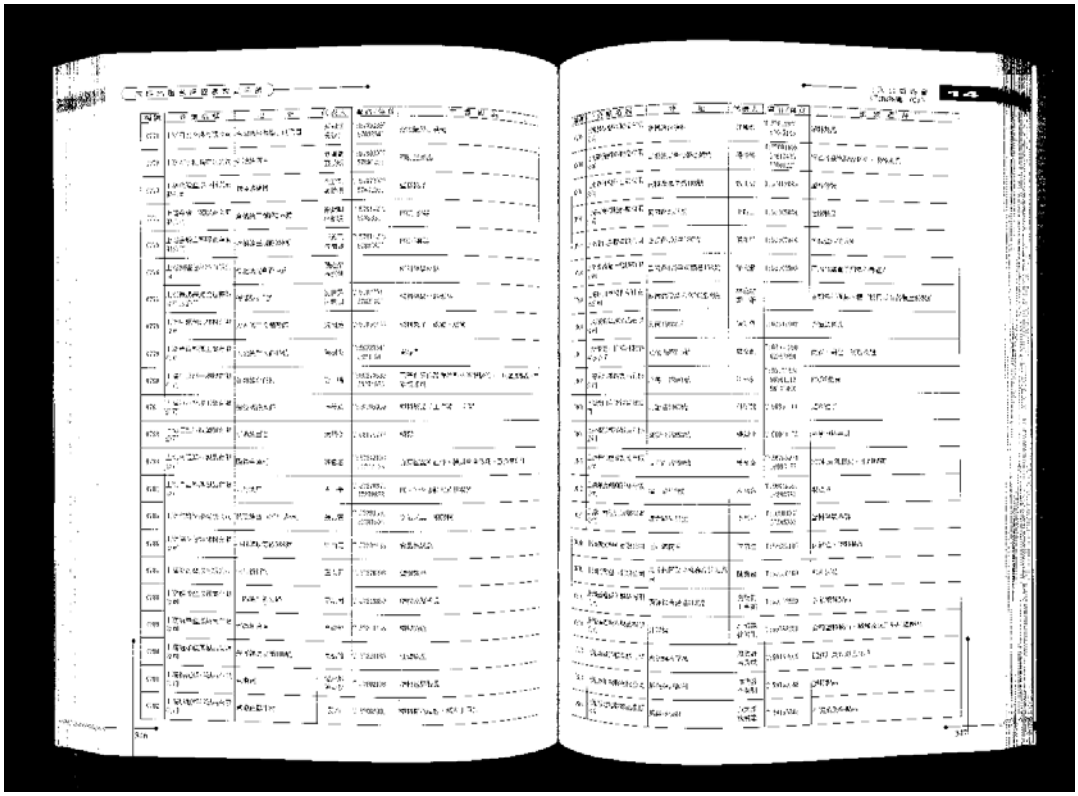
若區塊內可找到數段直線，則可能是表格區塊。

(c) 相連元件的平均大小與間隔

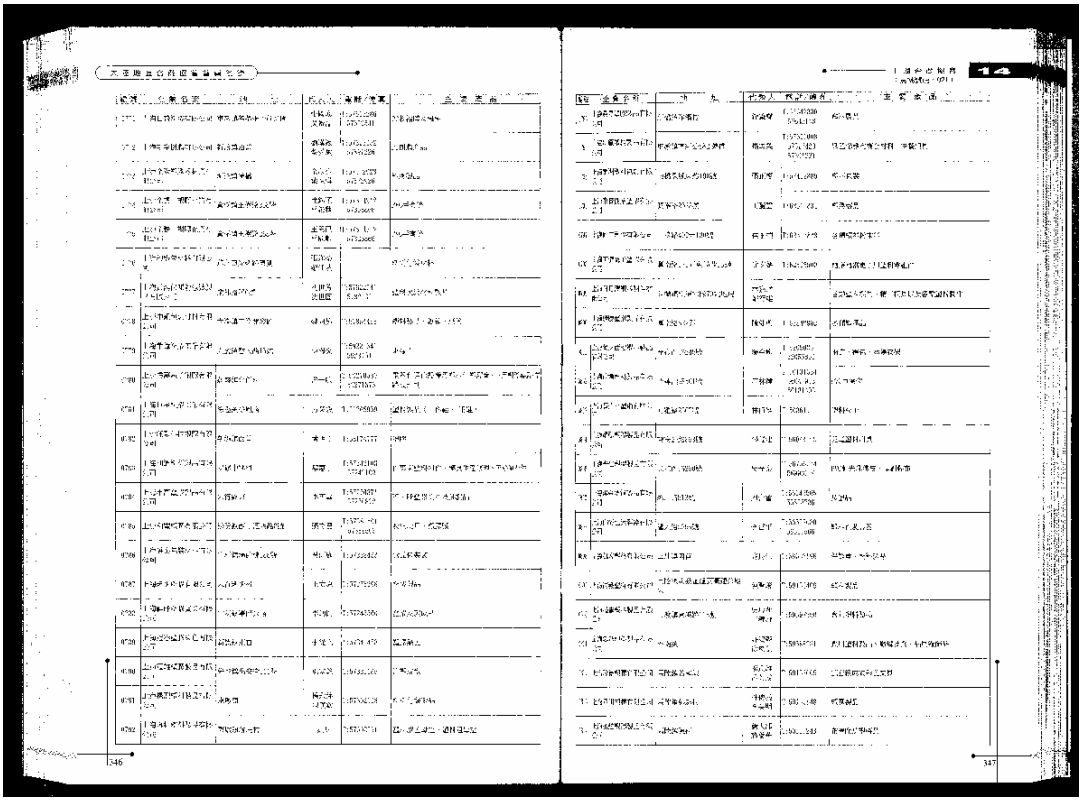
區塊內相連元件的大小與間隔分佈平均，且找不到直線，則應爲文字區塊。

c. 傾斜校正

一般而言，OCR 通常可進行些微傾斜字元的辨識（傾斜角度在正負 0.5 度以內），但若傾斜角度過大，將會影響版面分析與文字辨識率，因此在版面分析階段，會先做傾斜角度的偵測與校正。目前新技術「地理性校正」已能針對頁面或內容文字傾斜進行曲度修正，並盡量將影像頁面調整至水平以利後續 OCR 辨識作業。以下就期刊爲數位化物件作範例，以影像掃描後製軟體 Book Restorer 進行地理性校正前後之比對。（圖三、圖四）



圖三、原始物件掃描之影像檔



圖四、進行地理性校正之影像檔

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

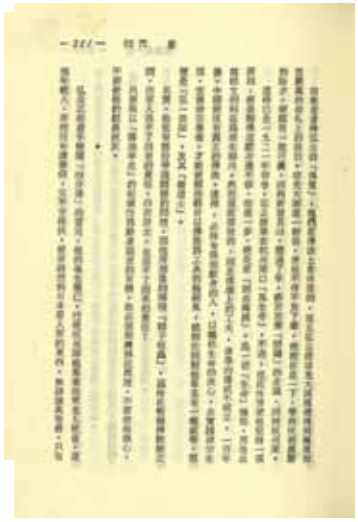
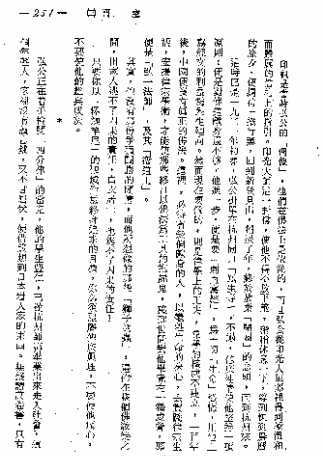
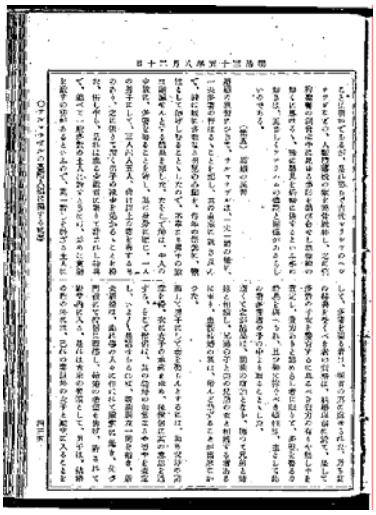
針對辨識內容特定的需求與用途（例如名片辨識、新聞字幕等），可事先內建辭典以提供候選字做更正的步驟。以名片辨識而言，通常會有一欄位爲「電話：」，而其後緊接的字元就可限制爲阿拉伯數字及特定字（如＃、轉、分機等），如此便能降低辨識系統誤認的情況。

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五：

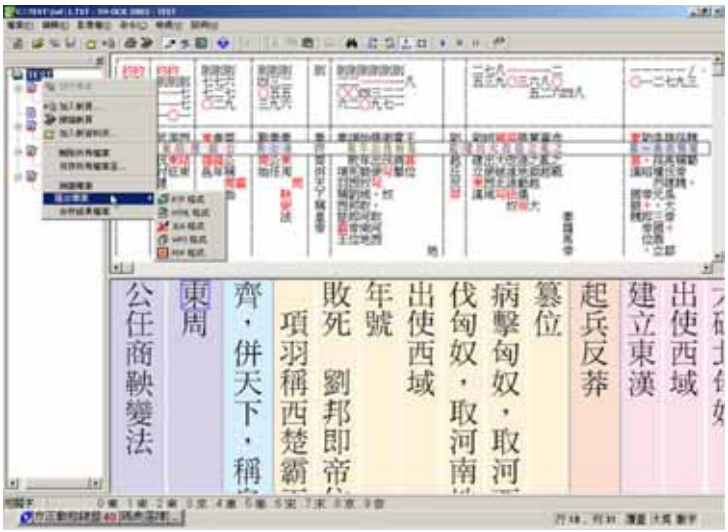
	
橫式中英文夾雜 （彩色 JPG）	橫式中英文夾雜 （黑白 TIFF）
	
直式中文（彩色 JPG）	直式中文（黑白 TIFF）
	
直式表格（彩色 JPG）	直式中日文（黑白 TIFF）

圖五、OCR 辨識測試圖檔

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。下列為 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為黑白 TIFF、解析度為 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，因為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（圖六）。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，無疑較為方便且人性化（圖七）。



圖六、清華軟體辨識表格內容及框線

圖七、清華軟體─可移動式橫隔線

（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 7、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 91.5%	可，辨識率 97%
	簡體中文	可	可	較佳
	英文	可	較差	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
儲存格式		*.txt/*.rtf/*doc./*xls./*slk./*csv./*html	*.txt/*.doc/*.xls/*.html	*.rtf/*.html/*.txt/*.jda/*.wps/*.pdf

伍、後設資料與資料庫建置

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，其真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，中央研究院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 6）[5]。

表 6、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

就新聞主題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文字型後設資料作說明。物件本身內容的文字後設資料包含文字訊息，而非內容本身的文字後設資料則有文字的種類、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。文字也可以被加以注釋，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[6]

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來相互傳播新聞，數位化新聞遠比傳統新聞需要更強而有力的資訊組識方法，以便能更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利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料，加速資料的散播。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，且各有不同用途。而利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，有利於更迅速的交換、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、目錄、書籍和期刊等平面媒體出版者的商業需求而設計。

（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 而來的。

（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）

大陸新華社所發展的「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

一、委外作業

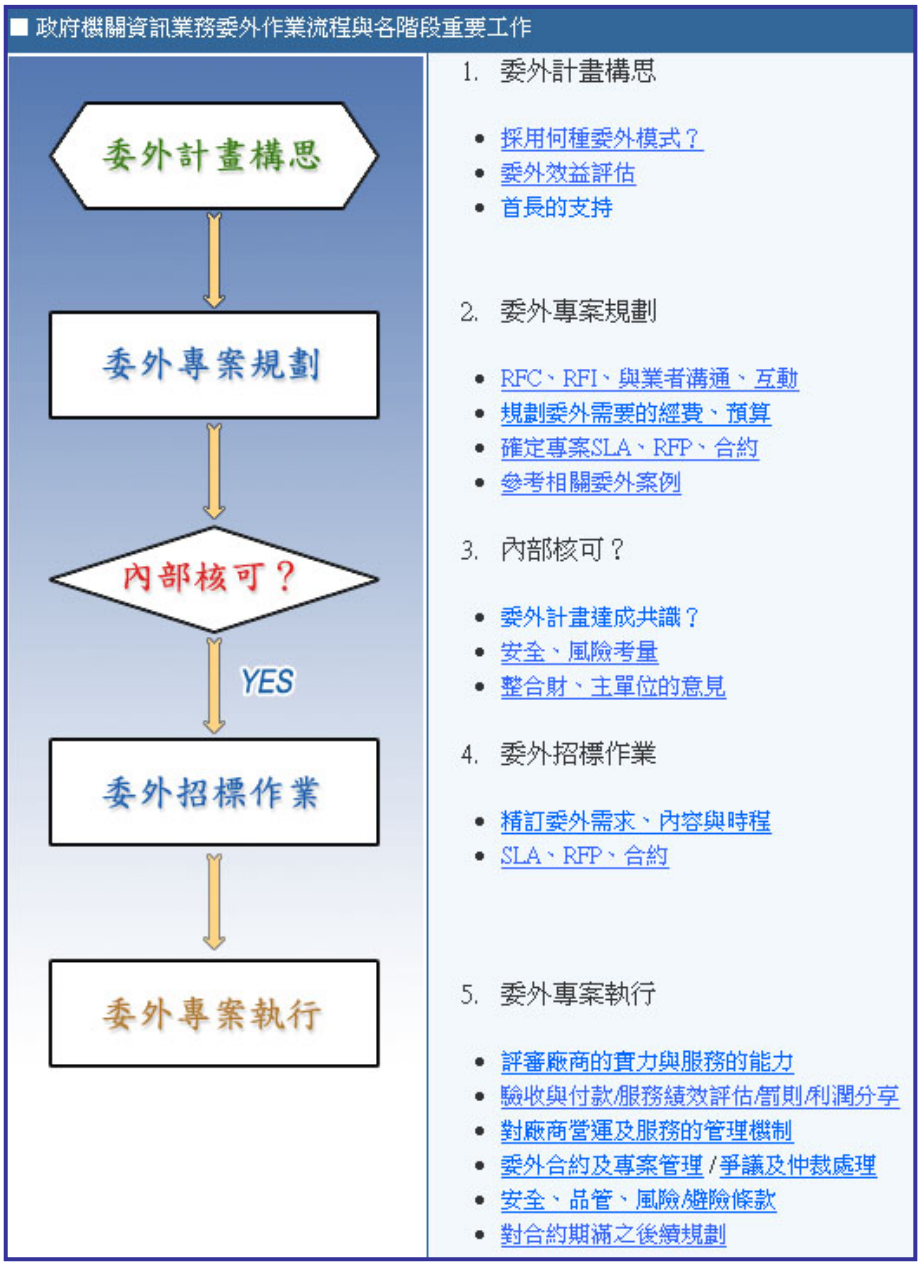
委外服務[7]是「將組織運作需要部分（非關鍵功能）以合約方式交由外面服務者負責」。因應時代環境之變革，委外服務的定義擴大爲：「假若有一份工作，外面的組織能做得比組織本身更有效率而且便宜，則此份工作應由外面的組織來做，假如組織本身能將此工作做得較好，則此工作應該保持自製。」以本國家型科技計畫下之機構單位─國家圖書館而言，委外是指將館內連續性生產流程中所涵蓋的各階段作業步驟，透過合約的簽訂，由館方轉包全部或一部分予外部機構或廠商代爲處理。

政府爲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研訂「行政院所屬各機關資訊業務委外服務作業參考原則」，並於民國九十一年十一月一日公布，以作爲各機關擴大辦理資訊委外作業之依據，並且特別成立「政府機關資訊委外知識網」（http://web.rdec.gov.tw/cisa/），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（圖八），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爲期刊報紙，其間接原件爲微縮膠捲/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」中第捌章節。與期刊報紙相關計畫中，具有委外經驗者爲『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。

一般而言，委外方式大致上可分爲以下兩種作業模式[8]，各機構單位可依照本身設備或資源情形斟酌考量之：

- 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。
- 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。



圖八、委外作業流程圖

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。因此應審慎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，隨著時移事遷，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。而數位典藏國家型科技計畫出版的《數位典藏技術彙編》手冊中，亦蒐集相關機構計畫委外之相關招標規範與契約書，提供各數位化典藏單位在進行委外作業時的一大參考依據。

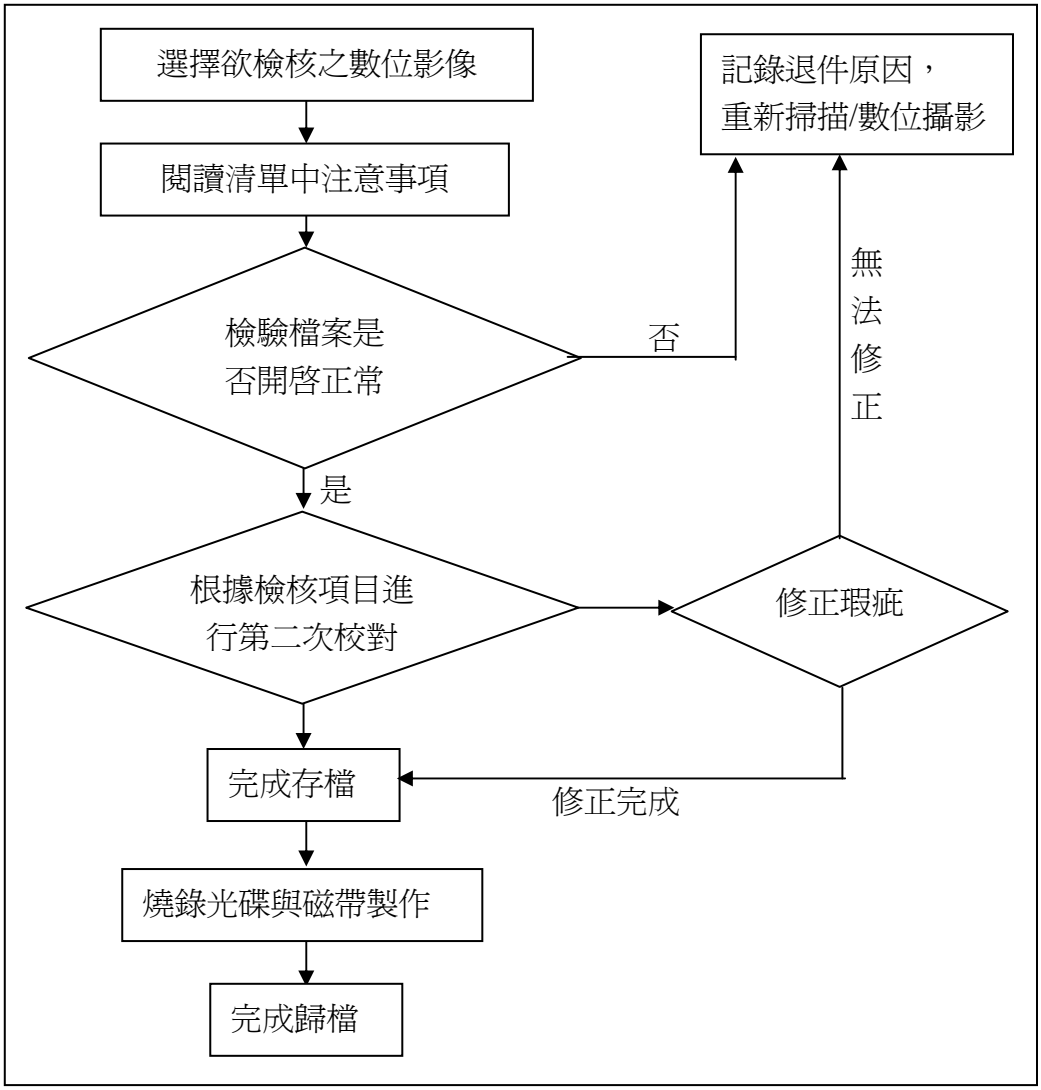
（二）驗收標準

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。下頁圖九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。



圖九、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按部就班順利進行，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：

（A）掃描/數位攝影

a.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。

b.可於系統中進行數位影像的修改，如歪斜、明顯污點、對比等。

c.經過修改的數位影像可記錄其修改的項目。

（B）檢核

a.選取欲檢核的檔案後即可顯示對應的檢核項目。

b.可於系統中進行數位影像的修改，如歪斜、明顯污點、對比等。

c.經過修改的數位影像可記錄其修改的項目。

d.若有退回重新掃描/數位攝影的檔案，可註記其退回的原因。

（C）燒錄

a.可顯示燒錄進度。

b.可自動確認檔案是否可正常開啓使用。

c.完成燒錄後可印製對應之標籤。

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行作業，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

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濫用，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[9]等領域，這些資料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便檢索搜尋之外，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。

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，密碼學技術則是用來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)為數位版權管理(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)，都能受到保護。不管在其使用過程中是否有複製行為的發生，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，則稱之為 DRM[10]。

數位版權管理技術近年來引起了廣泛的討論與注意，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。此概念前身即為反盜版技術，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制能保障其內容的完整性與價值。

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，而未來的趨勢也必定走向整合性的數位內容服務。

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[11]：

1. 避免智慧財產權未經授權而複製且濫用

2. 有效控管智慧財產權
3. 偵測及追蹤侵權行爲

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爲受保護的數位化資料[12]：

1. 浮水印：可參照下一節「數位浮水印」之介紹。
2. 設定數位權限：包含讀取、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，可作爲版權歸屬的依據，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，而若以此技術爲保護核心的架構下，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，可分爲顯性與隱性兩種：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 浮水印技術評比」[13]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[14]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，以此驗證資料之真確性。
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[15]：

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，因此常被用來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，在進行傳輸的過程中無法輕易由非收取者解讀；而另一支金鑰則作爲解密用途，以獲得原始訊息內容。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）

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 發展爲音樂產業的通用標準。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、數位電影服務、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）

澳洲地區所延伸出來的標準。

3. EBX （Electronic Book exchage）

EBX 目前已併入 NIST 下的 Open eBook Forum，成爲 OEB 標準的一部分。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 或者數位物件標識 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

此爲讀者對 eBook 所擁有的操作權利，包括能否複製、列印、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[16]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；

（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）現有的數位內容發行系統仍欠缺整合和安全性。

5. SDMI（Secure Digital Music Initiative）

SDMI 爲 The Secure Digital Music Initiative(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，任何希望使用不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益，將不會被限制。

三、現況與未來趨勢

「數位典藏國家型科技計畫」首要目標是將國家重要的文物典藏數位化，並且背負著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[17]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 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）達到真正數位內容的控管，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找尋一平衡點，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。

捌、設備與成本分析

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、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

（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 9、數位化硬體設備樣式

	
桌上型平台式掃描器	具備光罩之桌上型掃描器
	
桌上型自動進紙式掃描器	桌上型無邊縫書籍掃描器

	
滾筒掃描器	微縮膠卷掃描器
	
仰面式書籍掃描器	專業多用途掃描器
	
數位相機	數位機背

表 10、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
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
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

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 11、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
無發行台灣版	北京漢王	大陸	無發行台灣版

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

（一）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

「期刊報紙全文輸入工作流程指南」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，以避免浪費過多的人力與時間。因此本文「期刊報紙全文輸入工作流程指南」盼望能有以下效益：

1. 提升數位化進行過程之工作效率。
2. 可作爲教育訓練工作流程手冊之用。
3. 降低數位化進入門檻。
4. 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。

「期刊報紙全文輸入工作流程指南」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，因爲 OCR 軟體的應用仍然持續在進步中，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。

[1] 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〉，

頁 3～4。

[2] 洪淑芬，《文獻典藏數位化的實務與技術》，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。

[3]曾逸鴻，《光學文字辨識（OCR）技術整理報告》，頁 2。

[4]曾逸鴻，《光學文字辨識（OCR）技術整理報告》，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。

[5]曾欣怡、潘育潔，〈新聞傳播多媒體資料庫 Metadata 分析研究〉，頁 B3-4。

[6] 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。

[7]朱碧靜，＜書館館務委外之決策與管理探討＞，1998。

[8]朱碧靜，＜書館館務委外之決策與管理探討＞，1998。

[9]周宣光，〈數位內容產業的發展趨勢〉。

[10]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。

[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為例〉。

[12] 黃世昆、林宗伯、洪偉能〈數位內容保護與追蹤機制〉。

[13] 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〉。

[14]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〉。

[15]廖鴻圖、郭明煌、林金龍、陳貴青，〈Wrapper-based 數位版權管理機制〉。

[16]何佳欣、陳映后，〈數位版權管理綜觀〉。

[17] 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〉。