2003 年資訊科技與圖書館學術研討會，台北淡水，頁 73-93，民 92 年 5 月 29 日

中文報業數位化技術與現況探討

－聯合知識庫數位化經驗

孫正宜
淡江大學資訊與圖書館學研究所研究生
聯合知識庫資訊部製作組副組長
E-mail: cy.sun@udngroup.com

林信成
淡江大學資訊與圖書館學研究所副教授
E-mail: sclin@mail.tku.edu.tw

摘要

泛黃的舊報紙記載著許多人類珍貴的文化資產，以及歷年社會生活的共同記憶，透過數位化處理後，可以在知識經濟時代裡創造無窮價值。本研究藉由文獻探討、實地參與和訪談調查等方式，對中文報業處理報紙文獻的數位化技術和現況進行瞭解，提供各界有關回溯報紙文獻的管理與技術，期望對新聞數位化資料庫的建置過程有所裨益；此外，本文亦以聯合報回溯新聞資料數位化流程作為個案研究對象，進一步了解以數位化技術挑戰大量文件處理的實務經驗與成果，最終期盼對於建構未來全球中文新聞資訊共享之網路機制有助益。

關鍵字：報紙文獻、掃描影像、OCR、光學文字辨識、聯合知識庫、數位典藏、詮釋資料、Metadata

壹、前言

「今日的新聞，明日的歷史」，報紙文獻內容的重要性、即時性，非一般書籍所能取代。然而，報紙卻也因體積龐大，佔用儲存空間，紙質差，容易變黃變脆等因素，造成圖書館收藏時的一大問題。報紙的利用，更是圖書館的夢魘，沒有合適確切及隨時更新的索引目錄，造成讀者於查閱報紙資料時，比大海撈針還難，而讀者在複印時，又常不慎破壞裝訂好的報紙。種種的問題，以致早期有些圖書館甚至不願收藏報紙。雖然報紙管理有其不便，但其使用價值於今日資訊蓬勃發展的社會裡，有其不可？滅的地位。報紙兼具新穎性、知識性、娛樂性、與休閒性，是生活中不可或缺的一部分，亦是查找事實真象的好工具，已成為「逐日之百科」；報紙反應時事及社會變遷現象，不但可作為與民眾雙向交流的媒介，更可做為歷史的佐證與研究的參考。欲洞察先機、掌握新知，瞭解過去、前瞻未

來，報紙文獻是不可或缺的最佳利器。

然而報紙文獻僅有一天？命，第二天又有新的新聞出現，在保存上是十分不容易的一件事，再加上報紙文獻量龐大，一天去除廣告約有 40 塊版面，文字約有 28 萬字，與一本書的份量不相上下。[1]若不即時數位化，日後再想回溯，其困難度必定與日俱增。故不論是回溯的或每日新增的報紙文獻，使其數位化而能永久典藏，方便使用者利用，實為圖書館界當務之急。

誠如大陸學者張琪玉所言，報紙文獻是指報紙上登載的消息、文章、廣告等一切文字和圖像資料，是非常重要的資訊源，具有特殊的參考價值和史料價值。其特點是：[2]

(1) 報紙文獻是全社會的檔案；
(2) 報紙文獻是第一手文獻；
(3) 報紙文獻內容異常豐富；
(4) 報紙文獻具備獨特性；
(5) 報紙文獻具備有序性；
(6) 報紙文獻具備可近性。

數位圖書館的發展，最重要的部份之一就是如何將傳統媒體轉換成為適當的數位格式，並透過整理與索引的工具，提供使用者簡單有效的使用途徑。[3]報紙文獻也不例外。透過資訊科技的輔助，報紙文獻數位化可達到以下目標：[4]

(1) 滿足傳播科系與中國近代史學者研究的需要；
(2) 增進歷史新聞事件資料查詢的便利性；
(3) 增進數位圖書館發展；
(4) 豐富華文網路資料庫。

有鑑於報紙文獻數位化的重要性，本研究乃著眼於探討中文報紙文獻之數位化現況與技術，以期日後對回溯報紙文獻的管理、資料庫的建置以及建構未來中文新聞世界資訊共享之網路機制有助益。然而，由於許多實務界的經驗並未公開發表，以致文獻資料取得即為不易，故本研究採用實地參與、訪談調查和個案研究為主要研究方法，而以文獻分析作為輔助。至於研究對象與範圍，則因民國 77 年以前台灣有報禁，大陸也未開放，兩岸報業成長各有特色，同時中文字碼在數位化的過程中有著許多待克服問題，故將研究對象與範圍設定在海峽兩岸有進行數位化工作與若干較知名的報紙。

---

1 孫正宜，《解嚴後聯合報合訂本（OCR）人力評估報告》，台北：聯合知識庫，民 90 年 1 月，頁 1。

2 張琪玉，〈報紙文獻是一種極？豐富而未被充分開發的資訊源—關於發展報紙文獻索引和資料庫的思考〉，《圖書館雜誌》，88 年 2 月，第 2 期，頁 7。可得自 <http://www.libnet.sh.cn/magzine/99-2/p7.htm> (民 92 年 3 月 22 日)。

3 楊曉雯，〈美國圖書館數位化技術之應用（下）〉，《國立中央圖書館台灣分館館刊》，民 89 年 9 月，6 卷 5 期，頁 40。

4 世新大學資訊傳播學系，〈北平世界日報內容數位化開發計畫工作流程簡介〉，可得自 <http://content.ndap.org.tw/result/process/12theme-shu01/12theme-shu01.htm> （民 92 年 2 月 10 日）。

貳、報紙文獻數位化技術

報紙文獻特別零散，即使關於同一事件、同一領域的資料，也往往刊載在多日甚至隔幾天再出版的報紙上，各報紙的報導既有交叉又不相同，成？有效利用報紙文獻的難題。過去，？了解決查找報紙文獻的困難，一般採用編製剪報資料庫和編列書目或卡片索引兩種方法。剪報資料庫無法收錄完整條目，也不可能從多種角度對報紙文獻進行檢索，但有可直接檢出文獻原件的優點。索引則可從多種角度對報紙文獻進行檢索，雖不能一步檢得文獻原件，但可用書本式出版，其成本大大低於剪報。即使是做卡片式索引，成本也比剪報低。剪報需要用兩份報紙，才能剪得正反兩面的有用文獻，還要貼在紙上，分類裝訂或放在紙夾中，用櫃子或架子存放，成本比做索引高。[5]

由於資訊科技的突飛猛進，當今報紙文獻的發展是全文數位資料庫，等於是「索引 + 剪報 + 電子化全文」，優點更多。因此，本節重點著眼於探討如何利用資訊科技將紙張上的文字轉成數位化，供使用者查詢。

一般而言，報紙文獻數位化處理方式有掃描影像、重新打字、光學文字辨識、每日新聞直接下載轉入資料庫等四種不同方式，分述如下：

一、掃描影像

直接將報紙版面掃描成為影像檔儲存，這種做法比較簡單且省時省力，且可提供仿真的資料原件複本給使用者，不過其內容電腦無法直接辨識，而無法提供檢索，對使用上的效益遠不如電子全文資料。

舉例而言，中國時報、中央日報過去的回溯報紙就是用影像掃描，但為了彌補全頁影像不能檢索的功能，該二報同時再用人工打字方式建置索引（包括標題、作者、日期、版次、第一段導言）資訊供使用者查詢。[6]而「數位典藏國家型科技計畫」之一的「國家圖書館期刊報紙典藏數位化計畫」[7]所成立之報紙影像資料庫，更是此種方式的代表，將報紙掃描後（含微片轉製 34 種，共有 445,584 頁影像檔），提供了報紙文獻的全頁影像與新聞標題查詢。

二、重新打字

此種方式是直接拿紙本原件或將過去已經掃描成影像或製成微縮影片的報紙重新輸出，再用人工方式重新打字建置資料；打好的字再經人工校對，把校對好的文字檔轉換成為資料庫格式，上網供使用者查詢。

舉例而言，「數位典藏國家型科技計畫」之一的「世新大學世界日報內容數位化開發計畫」便是一例，世新大學把民國 15 年的北平世界日報利用上述方式重新打字上線，供讀者查詢使用。[8]

---

5 同註 2，頁 8。

6 江紀祖，漢珍數位圖書公司業務經理，受訪於孫正宜，電話訪問，台北，民 91 年 9 月 10 日。

7 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〉，可得自<http://www.ncl.edu.tw/pub/c_news/92/05.html>(民 92 年 2 月 25 日)。

8 莊道明，世新大學「世界日報內容數位化開發計畫」主持人，受訪於孫正宜，電話訪問，台北，民 92 年 3 月 7 日。

三、光學文字辨識

所謂光學文字辨識（Optical Character Recognition，簡稱 OCR）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。

OCR 適合印刷清楚、資料量龐大的文獻，從聯合知識庫為聯合報系的報紙文獻的做 OCR 的經驗可知，在近十年來的報紙文獻經 OCR 後的文字資料其正確率可達 99.98%，但是由於報紙編排方式多樣、複雜，其錯誤分佈並不平均，故仍需要校對。但早期的報紙文獻由於紙張泛黃，掃描後的品質不佳，內容清晰度差，利用 OCR 技術還不如人工重輸入來的有效率。如果報紙文獻的品質不差，利用 OCR 的技術來還原文字，其成本還是大大少於人工輸入的成本。

四、電子報直接轉入資料庫

此種方式是把當日文字檔直接轉入資料庫，並建置 Metadata 供使用者查詢，這是在 Internet 普及之後才開始普遍。實際上，以國內而言，最知名的兩大報系聯合報和中國時報，早在二十年前便將報紙的編排方式數位化了，只是並未把當日的文字檔儲存至資料庫中，當時所謂的「數位化」僅止於將手寫文字改為電腦輸入之電子形式，及以關鍵字進行較無效率的地毯式全文查找，至於資料庫、Metadata 的建置、XML 的應用等則是後來才逐漸受到重視。其他各家報紙、雜誌情況亦然。

參、回溯報紙數位化概況

本節以海峽兩岸較重要之報紙為例（中國時報、中央日報、聯合報系、世界日報、大陸人民日報和大陸解放軍報），採用訪談法以瞭解中文報紙回溯資料數位化之建置現況。

一、中國時報

中國時報是最早將回溯報紙合訂本實行數位化的台灣報紙，由民國 1940 年 10 月 2 日到今天報紙作全頁影像掃描，有簡易版索引(僅有日期、版次，無標題索引)，而 1995 年以後的數位化作業包括全頁影像、導言、標題、索引、首段，但是無 full-text 的內文。目前所有的資料都存入資料庫中，僅供內部同仁使用，在 1997 年成立的中時電子報但也只收錄每天兩百則新聞而已。[9]

二、中央日報

中央日報的數位化工作大致分為二部份：第一部份為撤台前的資料(1928-1949 年)，為毅士達公司延續人民日報的作業，向北京大學圖書館商借合訂本及場地，在北京大學歷時約一年完成，製作範圍與人民日報雷同，仍是全版影像及標題索引；第二部份是來台後的資料，由民國 1996 年 1 月 1 日到今天，建置的格式是全頁影像、標題索引、與全文的 full-text 檔。[10]然而，1950 年至 1995

[9] 林榮松，中國時報資料中心主，受訪於孫正宜，電話訪問，台北，民 92 年 3 月 4 日。
[10] 同註 6。

年的報紙文獻則尚未進行數位化的作業。

三、聯合報系

聯合報系於 1999 年成立聯合知識庫，2000 年開始對旗下五報[11]全面展開報紙文獻數位化作業。由於報紙掃描品質隨年代愈遠逐年下降，故按年代分階段完成。最早開始的是聯合報，其次是經濟日報與民生報，聯合晚報與星報則尚未動工。目前已完成之文獻及相關數據資料如下表所示：

項目 / 報別	聯合報	經濟日報	民生報	總計
日期	1975~1999	1988~1999	1988~1999	27 年
版數	212,928	160,000	144,861	517,789 版
字數	898,999,279	616,548,308	425,144,762	1,940,692,349 字
則數	1,612,357	1,297,090	910,108	3,819,555 則
區塊	4,134,508	3,012,460	2,348,446	9,495,414 塊

表 1 聯合報系數位化相關數據[12]

聯合報系推行報紙文獻數位化工作不遺餘力，計畫自 2000 年至 2005 年止，完成旗下五報創刊至 1999 年的報紙文獻數位化工作，是最有決心將報紙文獻數位化的報社。

四、世新世界日報

世新世界日報包括了北平世日界日報(民國 15 年 5 月 1 日至民 26 年 6 月 民國 34 年 11 月至 38 年 2 月)、上海立報（民國 24 年 9 月 20 日創刊至民國 26 年 11 月 24 日止)、香港立報民國 27 年 4 月 2 日至民國 30 年 12 月 3 日止)。[13] 目前參加了「數位典藏國家型科技計畫」。世新大學世界日報內容數位化開發計畫預計在三年半內完三種報紙的數位化作業。世新大學曾經考慮用 OCR 方式製作，但報紙影像的品質極為不佳，故用人工重新打字輸入。現在先做北平世界日報，已完成 15 年到 18 年共四萬多則新聞。所遇到的困難是當時的報紙標點符號僅有句號，讀起來十分困難，為了上線滿足使用者，在製作時必需重新標註標點符號。再加上鉛排字的缺字，漏字，不常用字，簡體字等問題，都有待克服。[14]

五、大陸人民日報

人民日報是最早將回溯報紙合訂本數位化的大陸報紙。人民日報數位化自 1995 年開始，然僅限於人民日報一份報紙，全報系大規模數位化工作則始於 1997 年 1 月 1 日人民網開始，目前已完成 1948 年創刊以來的人民日報全部內容數位化工作。人民日報早期報紙數位化? 人工文字輸入，照片重新掃描，並無全頁版

[11] 聯合報系旗下五報指的是：聯合報、經濟日報、民生報、聯合晚報、星報。

[12] 孫正宜，《聯合報、經濟日報、民生報合訂本 OCR 內文辨識作業完成報告》，台北：聯合知識庫，民 91 年 1 月，頁 2。

[13] 世新大學資訊傳播學系，〈北平世界日報電腦全文資料試編〉，可得自 <http://icd.shu.edu.tw/lipo/> (民 92 年 2 月 10 日)。

[14] 同註 8。

面影像、表格；直至 1998 年人民網才開始製作人民日報的 PDF 版，主要困難是雷射排版前報紙數位化工程浩大，版面原型復原困難，電腦無法 100%復原（字體、花邊等不太一樣）。另外，內容分類難定，整個報系協調不易，需要對出報系統進行統一調整。[15]人民日報已出版的 1948-1995 資料光碟，包括文章 200 萬條、圖片 2 萬張、約 10 億字，統稱為 50 年圖文數據庫光盤，目前由漢珍公司代理銷售。

六、大陸解放軍報

大陸解放軍報是中共軍方的報紙，也是數位化最完整的報紙資料庫，自 1946 年到今天，內容包括報紙的全頁影像、新聞標題、Full-Text 的內文、內容摘要、分類，一共五億多個字，除了少部份用 OCR 文字辨識外，其餘全是重新打字。該資料庫的分類方式是依中國大陸軍方的分類法則，目前由聯合百科電子出版公司代理銷售。[16]

肆、個案研究－以聯合報系為例

為了對報紙數位化技術、現況有個更深入的瞭解，本節採用個案研究法，以國內進行數位化工作不遺餘力的聯合報系為對象，分別從其數位化動機、數位化現況和數位化技術三個層面來進行實務面之探討。由於本文作者之一是聯合知識庫成員，因此，藉由實際參與數位化工作的過程，透過實地走訪、觀察、紀錄等方式，相信必能呈現出報紙數位化工作最真實的一面。

一、數位化動機

聯合報系將舊報紙數位化的主要動機有二：

其一，五十年來，台灣社會跨越了戰後的困乏，經歷了無數經濟及政治的環境的變革，聯合報始終堅持著「正派辦報」的精神，不但忠實地提供讀者「知的權利」，也留下了寶貴的歷史紀錄。這些珍貴新聞資產正是台灣民眾過去五十年生活的共同軌跡與驗證。過去這些珍貴資料除了做成合訂本之外，也曾製作微縮片及縮印本，近幾年也有廠商與聯合報系合作將報紙掃描成影像檔，並輸入標題文字製成光碟版。除了在聯合報系之外，只有少數大型圖書館或研究單位才有，資料或有不全，使用也十分不便，充其量只能稱為資料，不能為社會大眾廣泛應用[17]。因此聯合報系決定以回饋社會的態度發展聯合知識庫。聯合報系執行這項知識工程，係從「社會價值」的角度出發。聯合報系希望藉由科技的協助，讓舊報紙不再只是躺在大樓空調房間裡的一大堆泛黃紙張，歷史也藉由科技的呈現而鮮活起來。

再者，網際網路（Internet）不但是趨勢，更是不歸路；如果無法退縮，那

[15] 楊武軍，人民日報社網絡中心信息部主任，受訪於孫正宜，電子郵件訪問，台北，民 92 年 3 月 4 日。
[16] 范揚松，大人物公司負責行銷解放軍報業務，受訪於孫正宜，電話訪問，台北，民 91 年 7 月。
[17] 何銘傑，〈一網看盡五十年－聯合知識庫的建置與運用〉，《全國新書資訊月刊》，民 90 年 5 月，29 期，頁 7。

就得積極的加入。聯合報就是在這不同的時代下有了這種動機。由於電訊與電腦的應用與發展正趨向迅速整合，網路用戶可以透過電腦既快速又低廉的成本取得資料。但綜觀目前供使用者查詢的資料庫或入口網站，均屬綜合性的資料庫，尚不能夠滿足以知識為需求的網路公民，故聯合報開始致力將歷來傳統資料以數位化的面貌呈現出一個全方位的知識庫－聯合知識庫。而聯合知識庫建置是不同於以前一般的數位資料庫，它除了基本的資料庫建置外、尚且涵蓋資料倉儲（Data Warehouse）的建構，與資料探勘（Data Mining）的管理。簡單的說是一個要以人為本的知識管理的知識庫，故而一個知識庫的籌建不但要將龐雜的資料（Data）轉換成有用的資訊（Information），進而成為有效的知識（Knowledge），提供給使用者。目前想要使用過去報紙資料，除了到圖書館查合訂本外，就是上網查詢。近兩年來的資料或許可以看到電子檔，但過去的報紙資料目前多以影像檔呈現，不僅閱讀不便，亦無法做搜尋，也要花費相當長的時間下載，不符合網路公民的期盼。為了迎接 e 世紀數位時代的挑戰，聯合報系投入巨資，成立了聯合知識庫 udndata.com，計畫在五年內（自 2000 年至 2005 年），將聯合報、經濟日報、民生報、聯合晚報、星報等聯合報系五十年來的完整新聞，建構成為華文世界最完整的數位化線上資料庫。

二、數位化現況

聯合報系每日見報新聞有 180 影像檔版面，1500 則新聞，一百餘萬字，每日固定有這麼龐大的資料流量進到資料庫。另外再加過去聯合報系內的五報，共有 77 萬 8 千版， 578 萬 6 仟則新聞 ，近 80 億的字。現在決定要將這紙本的資料在四年內全部數位化倒進資料庫內。對一個資料庫的儲存、管理或是日後大批使用者同時的取用、查詢是一大考驗。目前在聯合知識庫內有 480 萬筆資料[18]可供讀者查詢。這其中包括聯合報系每日見報新聞，與聯合報與經濟日報民國 77 年 1 月 1 日到民國 88 年 8 月 31 日經過光學文字辨識（OCR）後的文字資料。

聯合報系數位化工程分為兩部份：一是全頁影像掃描外加建置標題；二是將報紙文字經過電腦辨識後成數位字碼保存於資料庫之中。如今聯合報 50 年的報紙全部掃描完畢，接下是經濟日報、民生報、聯合晚報、星報依此類推。

全頁影像掃描數位化是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，再不然就像世新大學的世界日報是全部重新輸入，而目前市面上用 OCR 技術，將傳統紙張數位化工程，做得較透澈的就聯合報系的聯合知識庫了。

聯合知識庫的 OCR 作業，在過去 50 年的報紙，只要電腦能夠辨識出來的字，一律用 OCR 技術將文字數位化，若是電腦無法辨識再考慮用重新輸入，或其他方式解決。聯合知識庫將聯合報系五十年來報紙以時代的意義，分成幾個階段來做 OCR 的工作，第一階段先做解嚴以後的報紙，因為解嚴對台灣民主憲政是一個相當重要的分水嶺。

- **第一階段：**

[18] 以聯合知識庫首頁 <http://udndata.com/> 數據資料為依據，民國 92 年 02 月 14 日庫藏資料共計 480 萬筆資料。

- □ 第一步先做聯合報 77 年 1 月 1 日到民國 88 年 8 月 31 日的報紙。
- □ 第二步再做經濟日報 77 年 1 月 1 日到民國 88 年 8 月 31 日的報紙，因應市場需要。
- □ 第三步先做聯合報 76 年 1 月 1 日到民國 64 年 12 月 31 日的報紙。
- □ 第四步是做民生報 77 年 1 月 1 日到民國 88 年 8 月 31 日的報紙。
- □ 第五步接下來再做聯合晚報 77 年到民國 88 年 8 月 31 日的報紙。

■ 第二階段：

- □ 回溯聯合報 40 年創刊到民國 63 年的合訂本。
- □ 回溯經濟日報 56 年創刊到民國 76 年的合訂本。
- □ 回溯民生報 67 年創刊民國 76 年的合訂本。

聯合報系擬按此順序，計畫在五年內把聯合報系五十年來的五報報紙數位化工程全部完成。

三、數位化技術

在聯合報系的數位化工程中，工作份量最繁重、投入人力物力最多的當屬將大量紙本資料轉為數位化資料的 OCR 作業過程。因此，此處便就聯合報系的 OCR 辨識技術及作業流程加以說明，以提供有興趣進行報紙數位化工作的同道參考。

1. 數位化辨識技術

近年來電腦的技術日新月異，各種生動活潑的多媒體影音技術，以及資訊豐富的電腦網路陸續發展，的確帶給人們相當的方便性及舒適性。電子書、電子報等電子文件帶給出版業極大的衝擊，也讓人們對未來無紙張世界充滿了期待。然而，對於現存的大量紙張式文件，如何將其數位化，以方便保存與快速流通，則是愈顯重要的課題。將整張文件掃瞄經過壓縮存成影像檔是個解決法，但仍嫌太占空間，且不易修改。若將文件中影像部分挖出壓縮，文字部分以字碼方式儲存，則不但節省大量空間，且新增、刪除或修改文字內容均極為容易。文件分析與文字辨識的研究，也應運而生。[19]

聯合報系的數位化系統採用「多核心辨識」技術。[20]所謂的「多核心辨識」就是將一個內文影像檔同時交給多個 OCR 技術同時辨識，並以「投票」方式決定該字的正確性，有爭議的字與一致通過的字會分開標示。如果只需處理有爭議的文字，校對的效率自然可以大幅提升。

經過多「多核心辨識」後，接著進行「批次集字校對」，就是將辨識成同一個字的影像檔全部集中，不需要比對原稿的校對方式。例如：被辨識成「聯」字的報紙影像檔全部集中，可能會有數百甚至數千個放在一起，校對的工作只要找出不是「聯」的文字影像檔即可。再配合多「多核心辨識」技術，將有爭議的字

[19] 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，民 90 年 1 月，頁 1。

[20] 全景公司，《聯合報報紙內文自動辨識及校對系統導入階段系統評估規格書》，台北：全景公司，民 89 年 8 月，頁 1。

放在前面並另作標示，在這個技術下有數十倍於傳統效率，且任何人都可操作，不需要特別的文學基礎。系統也會記憶、學習曾經所做過的修改，逐步提昇系統的辨識率。

2. 數位化作業流程 [21]

實際的數位化作業過程如下圖所示：

1. 影像掃描 → 2. 切字辨識 → 3. 集字校對 → 4. 逐字校對 → …

… → 5. 版面校對 → 6. 轉文字檔 → 7. 資料進庫

圖 1 報紙文獻數位化作業流程

茲分述如下：

(1) 影像掃描：包括全版影像、單篇影像、小圖等，如下圖所示。

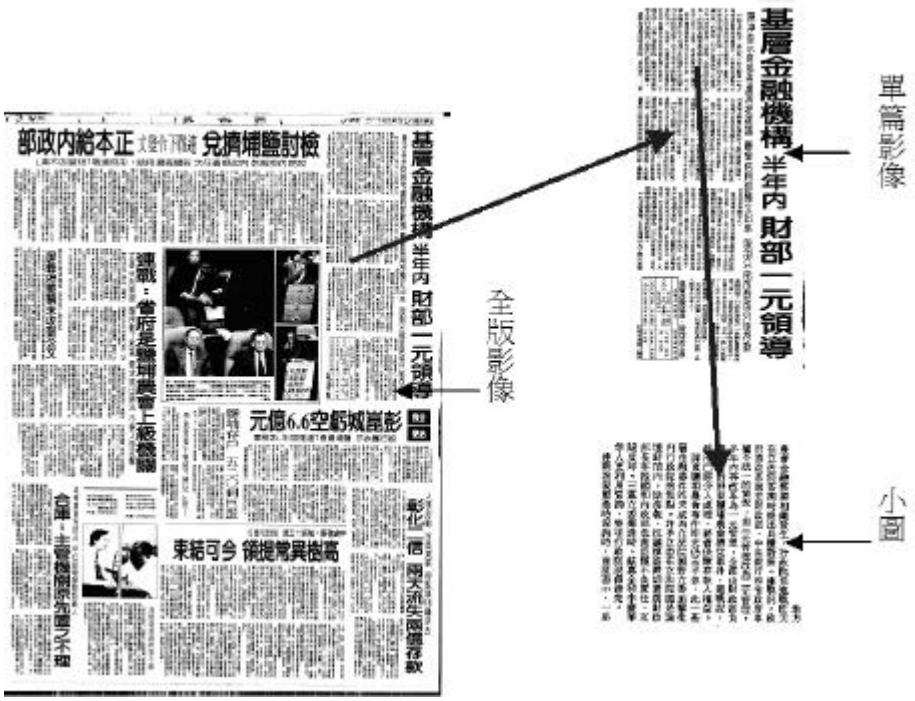

圖 2 影像掃描過程

(2) 切字辨識：將剪下的黑白內文影像辨識成文字，並存成集字校對的資料格式。

[21] 全景公司，《聯合報報紙數位化生產線規格及系統建置計畫書》，台北：全景公司，民 89 年 11 月，頁 15~19。

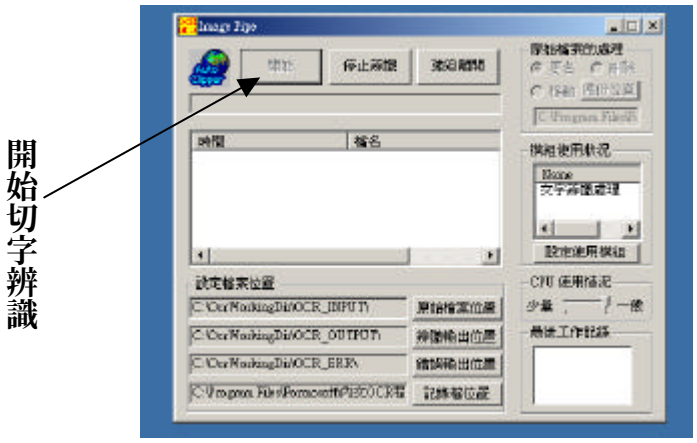

圖 3 切字辨識系統

(3) 集字校對：利用一群相同字的影像，容易看出不一樣的字的原理，來做辨識完成後的集字校對。集字校對的作法是將辨識出的字之中，相同的字集中一起，由人工檢核該字是否正確。

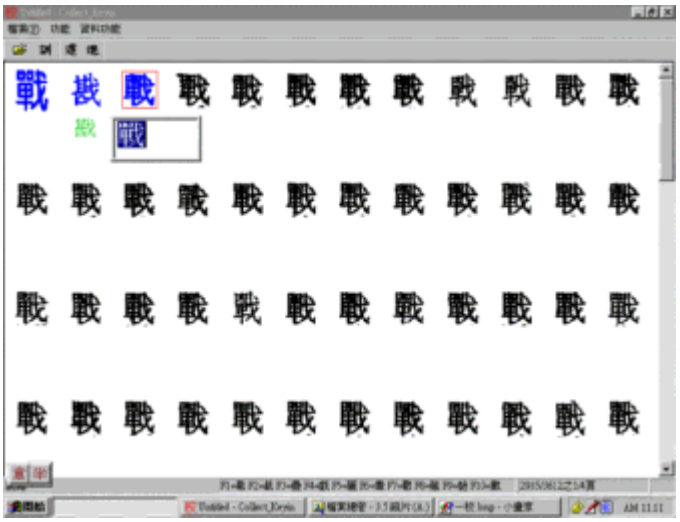

圖 4 集字校對介面

(4) 逐字校對：將每張影像與一校出來的結果，做最後比對。逐字校對的作法是，經過多核心比對後，系統認識的字通過，不認識或有疑問的字，以紅字顯現，逐字校對系統將影像檔原稿叫出，逐字校對。

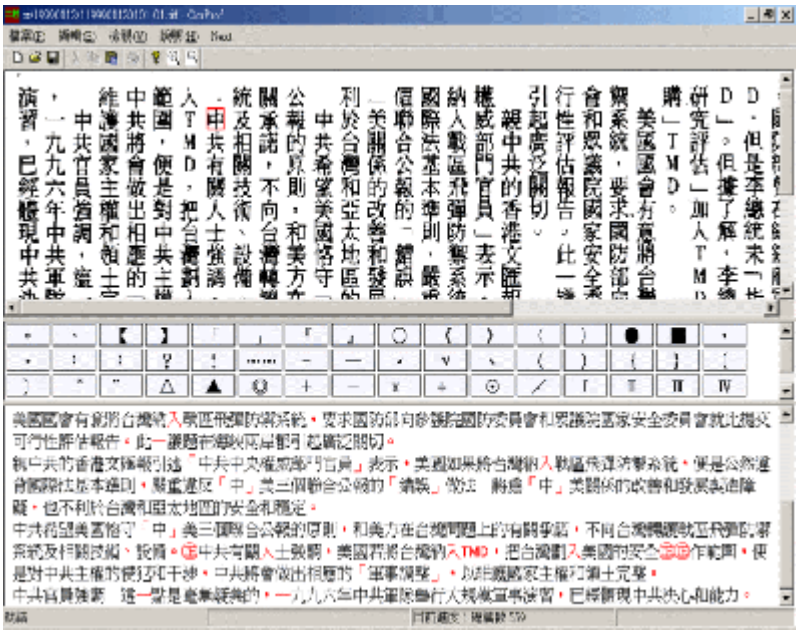

圖 5 逐字校對介面

(5) **版面校對：為預防小圖有所缺漏，故再增加一個版面校對，更確保數位化後資料的正確性。**

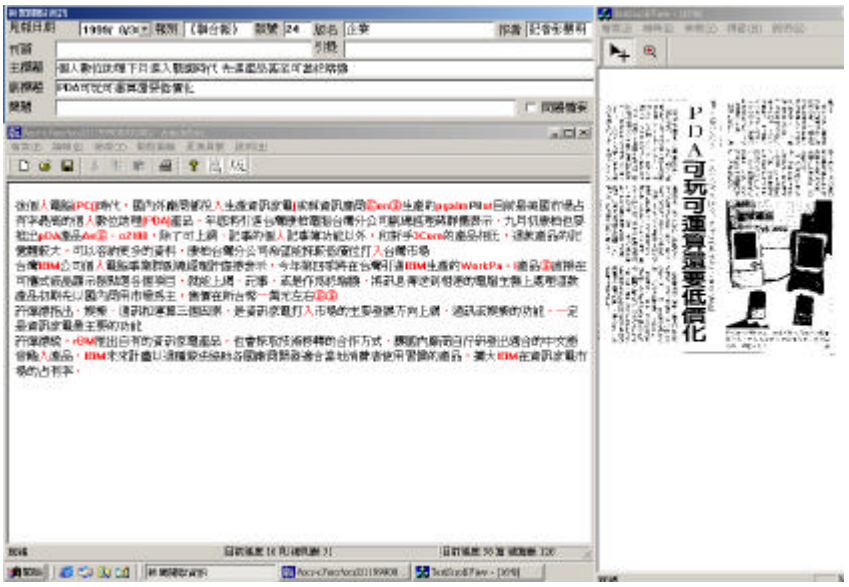

圖 6 版面校對介面

(6) **轉文字檔：將二校出來的資料轉成文字檔，並將以文章為單位重組這些文字檔，最後輸出的文字檔之文字碼，有** Big5 **與** Unicode **版本。**

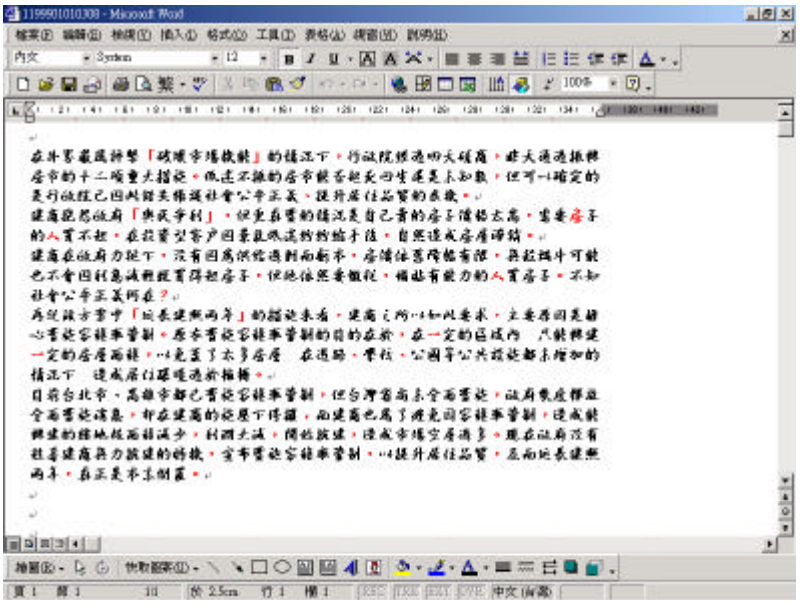

圖 7 文字轉檔介面

(7) 資料進庫：此時影像檔數位化作業完成，只要將資料轉入資料庫上網即可。至於資料庫系統的規劃與設計則是報紙數位化工程中的另一重點，限於文章篇幅有限，本文並未詳加介紹，此部份將留待日後探討。

圖 8 完成後聯合知識庫的介面

伍、報紙文獻數位化未來展望

一、有計畫有系統的發展

數位典藏不僅可以永久保存資料，更將影響知識累近過程，有鑑於此國科會決定將分開由各部會執行的數位典藏計畫，提升為國家型計畫，由國科會設置數

位典藏國家型計畫辦公室負責統籌推動，並將相關計畫經費等級列為第一優先，將我國固有文化以數位化方式向全球公開。

「數位典藏國家型科技計畫」在民國 91 年 1 月 1 日正式成立，是承襲行政院國家科學委員會「數位博物館計畫」、「國家典藏數位化計畫」、「國際數位圖書館合作計畫」三個計畫的經驗，依據國家整體發展，重新規劃而成。[22] 為了豐富華文報紙資料庫與中文新聞內容數位化內容，數位典藏國家型科技計畫於 91 年 7 月在內容發展分項計畫正式成立新聞主題小組，致力於報紙文獻之數位化工作。

二、資料規格的統一 － 詮釋資料（Metadata）的建置

Metadata 在資訊組織界最普遍的解釋是「data about data」，意指有關資料的資料，即資料之描述性資料，此一定義源自 1995 年 3 月由 OCLC 與 NCSA（National Center for Supercomputing Applications）兩單位共同主辦名為「Metadata Workshop」研討會，廣邀圖書館學、電腦科學、文獻編碼及相關領域學者專家參加，於會議中提出「資料的資料」作為 Metadata 的定義。例如圖書館的 MARC 記錄，即為一種 Metadata。另 Renato Innella 認為定義為「Structure data about data」，此結構二字使得採用 Metadata 做組織資訊的方式和全文索引用（full-text indexing）有所區隔。中文譯名有「詮釋資料」、「元資料」、「後設資料」等。[23] 詮釋資料是對藏品資料屬性的一組描述，目的在促進資訊系統中對於資料的檢索、管理與分析。其具有傳統目錄著錄功能，目的在使資料的管理維護者與使用者，可透過詮釋資料來了解並辨識資料，進而去利用與管理資料。目前全世界各國的數位圖書博物館計畫中，皆將 Metadata 列入必備的研究項目之一，而且形式不一。實有必探討與瞭解我國數位圖書館或博物館在「詮釋資料」或「後設資料」（Metadata）的使用情況。

現今國內報業界的報紙資料庫格式是百家爭鳴，各據山頭，毫無可能交互利用的可能性，對使用者與典藏而言非常不便。同時管理者對 Metadata 格式既無深入的認識與也無統一標準的行動，若有需求大都參考國家圖書館所制定的 Metadata 格式。故「數位典藏國家型科技計畫」新聞主題小組於 2003 年 3 月 7、8 兩日舉行中文新聞數位化研討會，討論新聞內容標誌 XML 之推廣應用、新聞詮釋資料(metadata)之建構以及 TEI(Text Encoding Initiative)及 GIS(Geographic Information System)應用於新聞資料庫之研發。[24]該會同時研討如何把國內所制定的 Metadata 與國際的新聞 Metadata 標準接軌。

---

[22] 謝清浚，〈數位典藏國家型科技計畫簡介〉，《數位典藏的意義與影響座談會》，台北：資訊工業策進會資訊科學展示中心，民 92 年 2 月 24 日，頁 2。

[23] 陳昭珍，〈一個 XML/Metadata 管理系統設計經驗淺談－Metalogy 之架構與功能簡介〉，《海峽兩岸第五屆圖書資訊學學術研討會論文集》，台北：中華圖書資訊學教育學會，民 89 年 8 月，頁 59。

[24] 「數位典藏國家型科技計畫」新聞主題小組，〈數位典藏國家型科技計畫內容發展分項計畫〉，《中文新聞內容數位化研討會論文集》，台北：中央研究院，民 92 年 3 月 7 日，頁 4。

三、建全系統效能與網路頻?

報紙數位化的最終目的在於長久保存與有效利用。有別於以往的圖書館微縮調閱、光碟查詢等方式，將數位化的報紙文獻儲存於資料庫系統中並上網供讀者查詢、調閱，讓更多使用者輕易取得新聞資料，再分析、組織進而創造出更好的內容，提供社會一個反省、改革、決策的有價資訊，不但是網路時代必要的，也是最好的運用方式。然而，要善用網路建立一個高附加價值的新聞資料庫，首先必須面臨網路頻寬與資料庫系統效能的挑戰。因此，加大網路頻寬，適當提高系統效能，是新聞資料庫因內容纍增而不斷加大的同時，所必須同步考量的重要議題。健全系統效能常用的方法有：硬體更換或升級、採用叢集系統、網路負載平衡等系統管理與開發技術。

總之，報紙資料庫在考量資料完整性、建置 Metadata 之外，若能有效提升系統效能，可以增加資料庫應用的效益，在實質上，發揮資料庫的真正價值。

四、報紙文獻權威檔建立

權威控制目標主要為二：一是決定目錄中索引點之形式建立一致性之人名、劃一題名、標題等標目，以維護目錄品質。權威控制的另一個目標在提供參照關係，由相關款目參照確立款目。[25]如此才能找全其所要文獻。

報紙文獻的題名、譯名、人名、地名各家報社的做法與說法不一，以檢索的目的來說，是非常需要權威控制，否則無法找全資料。國內圖書館界的權威控制之法都仿歐美，不盡符合我國之需求，更遑論報紙文獻的權威製作。台灣早期的權威紀錄都是由各個圖書館自行建立，較大規模有國立台灣大學、國立中央圖書館（國家圖書館的前身），但都僅針對圖書館的書目資料而設計，實在無法滿足報紙文獻資料庫之需求。故而要發展一個完善的報紙文獻資料庫，權威檔的製做，是不可少的。

陸、結論

舊報紙或紙張雖然泛黃，卻藏有不少人類珍貴的文化資產，而這些珍貴的文化資產透過數位化的處理後，可以在知識經濟時代裡能創造無窮的價值。然而這些珍貴的老骨董若再不處理時，就可能隨著時間灰飛煙滅，而這種損失是無以彌補的。馬亞文化、龐貝城、敦煌石刻尚有遺跡可尋，而一旦紙張的消失，不是人類用任何科技可以挽回。幸而產、官、學界有心人士利用科技的技術保存這些珍貴的文化資產。故不論全頁影像掃描、人工重新輸入、OCR 光學文字辨識技術，雖然它們不是盡善盡美，但是可以將乏人問津的舊資料數位化，成立資料庫上網，無遠弗屆的供世人使用。隨著資訊壓縮技術和復製技術的發展以及其成本的大幅下降，數位資料庫形式由文摘走向全文形態發展。[26]而儘管已經有現代的科

[25] 盧荷生、陳昭珍，〈台灣地區權威檔之建立與問題探討〉，《中國圖書館學會會報》，民 83 年 6 月，52 期，頁 64。

[26] 李學軍、鄧小川，〈報紙文獻的數字化建設〉，《四川圖書館學報》，民 90 年 1 月，1 期，頁 67。

技輔佐，但是要將過去的報紙文獻放到網路上去，還是要動用相當大的人力、時間與經費。一般網路業者認為回溯舊資料的商業價值不高，而且又很辛苦，國外很多新聞資料庫也不做新聞的回溯工作，但以保存人類文化資產的而言，然誠如聯合線上內容長易行所言：只要「知識是有價值的」，[27]這件事情在網路上被承認，那麼舊報紙數位化的工作所做的事就是有價值的。

[27] 楊瑪利，〈聯合知識庫一網看盡五十年報紙〉，《天下雜誌》，民 90 年 2 月，237 期，頁 218。