



國立臺中圖書館館藏數位化作業

賴忠勤 國立臺中圖書館視聽教育課課長兼資訊小組執行秘書

一、前言

國立臺中圖書館（以下簡稱國中圖）成立於1923年日治時期，前身為臺中州立圖書館，因此具有為數不少的珍貴典藏。其中數量較為龐大的有日治時期所留下的日文圖書資料、民國20年代至今眾多種類之報紙資料。國中圖自民國90年開始進行相關數位化作業，其中就其館藏資料進行數位化作業，計有「舊版報紙」及「日文舊籍」兩項。本文簡介說明此兩項數位化作業相關內容及功能，並期望能與其他相同類型之數位化作業相互支援，避

免產生不同的作業規範及重複數位化的浪費。

二、舊版報紙數位化作業

國中圖整理館藏報紙資料，於91年將最為珍貴的《民報》進行數位化，並建置「舊版報紙數位化系統」、制定 metadata 格式、檔案命名格式、標題分類等規範；93年持續進行第2期的舊版報紙數位化，並特別挑選18種國家圖書館未曾數位化的報紙，避免有重複數位化的情形，總計1、2期數位化相關資料表列如下：

報紙名稱	收錄年代	版數	新聞標題
1. 民報	民國34年10月11日-36年2月28日	1,500	30,640
2. 正氣中華	民國43年10-12月，44年1、2、4-12月，45年1-12月，46年1-3月	3,366	31,358
3. 工人報	民國42年7-12月，43年1-6月，44年1-12月，45年1-6月	732	6,863
4. 外交部通報、外交部週報	民國35年11月-36年12月，37年1-4、8-12月，40年4-12月	589	1,763
5. 太倉明報	民國36年6月	120	820
6. 臺灣力行報	民國37年5-7月	104	882
7. 成都晚報	民國37年9月	168	1,200
8. 國際現勢周刊	民國45年1-6月	112	439
9. 青年新報	民國40年4、5月	56	330
10. 內外	昭和30年(1955)1-10月，昭和31年(1956)1-4、6-12月，昭和32年(1957)1月	2,860	23,264
11. 自強晚報	民國43年10-11月，44年1-10月	1,424	12,562
12. 天聲日報	民國44年3、5、6、10、12月，45年1、2、6-10月，46年1月	1,304	23,125
13. 大華新聞	民國45年9月-46年4月	924	8,098
14. 台東新報	民國45年1-5、7-12月，46年1-5月	2,450	29,186
15. 更生報	民國44年1月-46年1月	3,238	43,359
16. 攝影新聞	民國45年2-12月，46年1-7、9-12月	2,644	8,093
17. 中國日報	民國45年5-12月，46年1-8、11月	1,920	22,521
18. 商工日報	民國43年10月-44年9月，45年1-5月	3,058	48,373
19. 青光日報	民國45年5月-46年1月	870	13,259



(一) 數位化規範

《民報》進行數位化時，係採用全版掃描，版面內之所有新聞標題、廣告、公告等建置 metadata，對應至該則新聞所屬的版面影像，使歷史性報紙之版面編排樣式得以完成呈現。新聞標題

metadata 格式參考國內外相關資料庫的內容及文獻，例如「即時報紙標題索引系統」、「Metadata and Data Structures for the Historical Newspaper Digital Library」(註1)等，制定適合本案需求的 metadata 格式。其 metadata 欄位如下：

欄位名稱	欄位說明
System ID 系統識別號	報紙內容系統編號，為唯一之號碼。
Title 題名	新聞內容標題之主要題名、廣告產品名稱、公告、啟示題名等。
Sub-title 副題名	新聞內容標題之次要題名。
Author 作者	指撰寫新聞之作者、廣告刊登者、公告、啟示刊登者。
Report Location 報導地	指發布新聞之地點。
Source 報刊別	報紙名稱。
Publish Region 發行地	報紙發行地，藉以區分地方版，若非地方版則無須輸入。
Date 日期	報紙出刊日期。
Week 星期	報紙出刊日期之星期。
Page No. 版次	內容所在之版次。
Page Name 版名	報紙整版名稱。
Special Column 專欄名稱	專欄名稱。
Abstract 摘要	新聞內容摘要，每則約50字。
Category 類型	報紙內容之類型，如：新聞、廣告、公告、啟示、訃文、社論、漫畫、物價表等。
Subjects 主題	新聞分類主題、人物、地理名稱。
Graphical Objects 圖示物件	內容所附之圖示，如圖表、照片、插圖等。
Caption 圖示說明	圖示相關說明。
Area 內容所在區域	將報紙影像從右而左、上而下，區分為9個區域：右上、右中、右下、中上、中中、中下、左上、左中、左下。標示本內容於全版報紙影像中所在之區域。
Language 語文	內容語文。若為繁體中文則無須註記。
Notes 備註	此紀錄之其他註記。
Image 影像連結位置	報紙全版掃描影像檔或切割區域影像欄之路徑及檔名。
Full Text 電子全文	新聞內容電子全文。
Electronic Resource Location 電子資源位址	指網路上相同內容之電子資源所在位址。
Input Time 建檔 / 修改時間	建檔或修改紀錄之時間。

之後為配合行政院文建會「國家文化資料庫」的需要，擴增較為複雜的「報紙內容詮釋資料 (metadata) 格式及設定」，作為資料上傳交換的規範。

第1期《民報》掃描格式為300dpi黑白之TIFF檔，並轉為較低解析度之TIFF及JPEG格式供線上瀏覽觀看(檔案大小控制在1至2MB之間)。受到

報紙紙質泛黃及污漬的影響，黑白色調非常不易調整。因此第2期之數位化作業，永久保存之TIFF格式即改為彩色掃描方式處理。

報紙影像檔案命名考量到識別及唯一性，避免利用檔案目錄作為區分，所定的檔案命名結構如下：

單位代碼——物件類別——物件代碼——多部



份之序號——使用目的.附屬檔名

1. 單位代碼：本館為 cca100003，由文建會訂定。
2. 物件類別：新聞類之物件類別代碼為 News，即 nw 代表。
3. 物件代碼：報紙代碼+出刊日期（8位數字），例如：《正氣中華》民國 43 年 10 月 10 日出版之報紙，其物件代碼為 ccch 19541010。
4. 多部分之序號：
 - (1) 以 2 位數字代表版次，如第 1 版為 01，第 2 版為 02。
 - (2) 若為增張之附刊或特刊，以 a+2 位數頁次之方式呈現，如：a01、a02。
 - (3) 若切割其中之區塊為獨立之影像檔（例如：廣告），於版次後加英文 "a"，再加 2 位數數字之流水號，如：02a01、02a02、02a03。
5. 使用目的以 i 代表供下載及螢幕顯示用；t 表示預覽用；u 表示永久保存版。
6. 附屬檔名為 TIFF 格式者為 tif，JPEG 格式者為 jpg。
7. 檔名之英文字母一律以小寫呈現。

(二) 系統功能及加值功能

為吸引一般民眾使用，以資料庫檢索功能為主所設置的「舊版報紙資訊網」(<http://paper.ntl.gov.tw/>)，就報紙特色加以分析，設計幾項加值性功能及內容：

1. 歷史上的今天：由系統自動帶出同月同日的報紙標題，可直接點選至當天的報紙影像。
2. 舊廣告影像瀏覽及檢索：每則報紙廣告均以 metadata 建檔，並特別裁切廣告的圖檔，作為展示瀏覽之用。
3. 懷舊電影區：報紙上最常出現的電影廣告，亦另闢專區展示。
4. 民報燈謎區：《民報》在接近元宵節所出刊的內容中，有大篇幅的「元宵燈謎大會」專欄，整理燈謎的內容，設計線上猜燈謎功能。

(三) 未來發展

因報紙可以反映當時政治、社會、經濟狀況，包羅萬象的內容若能加以分析研究及設計，即

可以再創造出更多的加值性功能。而全文電子化也是未來努力的目標，從「北平世界日報內容數位化」的案例中，由於早期報紙用字與今日習慣的差異、排版與印刷技術的限制，再加上保存本身的損害，以及報紙內容缺字、漏字、難字、不常用字等問題，使用 OCR 光學字元辨識低於 50%，因此並不適合 OCR 自動辨識文字，而必須採用人工打字輸入。（註 2）未來若經費許可將報紙內容全文電子化，則可以再增加全文檢索機制及更進一步的應用。

三、日文舊籍數位化作業

國中圖典藏日治時期日文舊籍，為日本統治臺灣時期官方及坊間刊行之各種出版品，包括圖書、雜誌、官公報，總數共計 23,276 種，其中多數均已絕版，部分甚至為孤本，具有相當的價值性。當時配合文建會計畫成立「臺灣歷史數位圖書館」，國中圖所藏日文舊籍為數位化重要的標的，因此於 92 年由學者專家先行挑選出對臺灣歷史極具重要性、相關性資料，或在其他圖書館罕見之珍貴資料計 2,349 種，分年分期予以數位化。第 1 期（民國 92 年）將其中 809 種 1,189 冊數位化，並建置檢索系統；第 2 期（民國 93 年）數位化 320 種 426 冊，另外協助國立臺北師範學院圖書館典藏之日文舊籍 336 冊予以數位化，並建置專屬網站「臺灣歷史珍藏 e 點通」(<http://jobook.ntl.gov.tw>) 及設計加值功能。總計 1、2 期全部數位化掃描約 55 萬頁。

(一) 數位化規範

日文舊籍因其性質大部分為圖書資料，因此建檔之 metadata 直接採用「中國機讀編目格式（第 4 版）」，並將目次建置於欄號 327。為使日文字能完整呈現，建檔字碼採 Unicode。除了網站採用 Flysheet 作為書目檢索系統，並將書目紀錄轉入圖書館自動化系統之書目資料庫，在 OPAC 檢索顯示書目紀錄時，可透過 856 連結至該筆書目之全文影像展示畫面。

圖書逐頁掃描永久保存格式為 300 dpi 彩色 TIFF 格式，並轉換為 72 dpi 之彩色 JPEG 格式作為線上瀏覽用，為便於列印及多頁觀看，另外製作

PDF 格式提供使用。

(二) 系統功能及加值功能

1. 電子書格式

對於日文舊籍的展示功能，除了前述 JPEG、PDF 格式外，為增加使用者瀏覽閱讀日文舊籍內文之興趣，特別製作 100 本日文舊籍之電子書版本，採每月分類輪播不同之主題呈現。電子書採 FlipBook 電子書格式，需先下載 FlipViewer 閱讀軟體始能觀看。(圖 1)

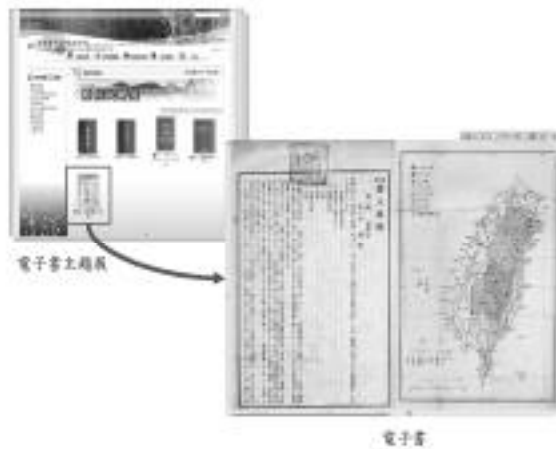


圖1 日文舊籍電子書主題展及電子書

2. 圖片瀏覽

擷取日文舊籍內文中精選的圖片影像檔，可按分類瀏覽。並且在網站首頁與電子書定期更換展示內容，以增加吸引力。

3. 青少年兒童學習館

將日文舊籍數位物件內容，以遊戲方式設計，吸引青少年兒童使用。

(三) 未來發展

將掃描之全文影像文字內容予以全文電子

化，亦是日文舊籍數位化作業的重要目標。與報紙影像不同之處，在於日文舊籍逐頁掃描的內容，較能夠進行 OCR 自動辨識方式處理，唯因為涉及日文內容及日文漢字辨識問題，需要再進一步評估成本效益。

目前國中圖仍有數量眾多的日文舊籍待後續經費的支持進行數位化，而國家圖書館、臺大圖書館亦有類似的計畫在進行，如何整合，而且避免重複數位化，將是未來進行數位化時必須特別注意的重點。

四、持續不斷的數位化作業

由於數位化作業所需經費龐大，而國中圖之經費較為拮据，因此數位化作業經費主要來源均為文建會建置「國家文化資料庫」補助之款項。有時受到補助經費不足或中斷的影響，使數位化作業無法依照既定計畫進行，此為國中圖進行數位化作業最大的困擾。

未來持續不斷進行數位化作業將是圖書館回溯館藏資料時必要的投資項目，而國立臺中圖書館具輔導全國公共圖書館的任務，未來也勢必會協助其他公共圖書館進行數位化作業，不過必須避免與國家圖書館或其他機構重複，讓數位化作業能夠互補有無，達到全面數位化且不重複的目標，需要各機關的協調與合作，以發揮最大的功效。

註釋

1. Robert B. Allen and John Schalow, "Metadata and Data Structures for the Historical Newspaper Digital Library", Conference on Information and Knowledge Management (NY: ACM, 1999), pp.147-153.
2. 余顯強，北平『世界日報』：民初歷史性新聞報紙數位化之研究，《圖書與資訊學刊》54 (2005.8)，頁87。