資訊公路上的「台灣資料庫」

史籍自動化室*

隨著資訊時代的全面來臨，電腦已成爲現代人的重要生活伙伴；尤其，經由網路擷取各種資訊，不僅可免奔波之苦，更可收即時即效。有形世界的距離，在網路中消失無形；反而透過網路，可以探索到另一個超越時空概念的世界，無遠弗屆。因應時代趨勢，台灣史研究所籌備處特別設立「史籍自動化室」，長期推展「台灣史籍自動化」工作，本文即第一階段——「台灣方志資料庫」的成果報告，並對第二階段之「台灣檔案資料庫」做一簡介。

一、何謂「全文資料庫」？

所謂「全文」，是指文獻的全部內容，包括文字、公式、表格、圖形、影像……等。將資料的「全文」，以儘量忠於原文的格式，一一轉換爲電子形式；並使之具備可瀏覽、可查閱、可檢索的特性。這便是所謂的「全文資料庫」。

如果要把中文史料製成「全文資料庫」，由於大部份的古籍係以文字叙述爲主，限於人力、技術的考量，所以目前只處理了一般文字及含文字的表格。將這些文字以原來的格式逐字繕打爲電子檔案，形成「原始文獻檔」；然後在「原始文獻檔」中添加標誌符號，用以描述資料的組織結構，便成爲「標誌文獻檔」；最後執行資料庫的建立程式，就可建成「全文資料庫」。

檢索全文資料庫的程式，提供了閱讀、檢索兩個主要功能。使用者透過個人電腦，配合倚天中文系統，連線至主機，便可如讀書般自然地去查閱任何篇章，

---

* 本文由自動化室同仁李素蓉、孫智明、林詩倫、洪雪卿共同執筆，並經詹素娟綜合改寫而成。

或檢索資料庫裡每一字、每一詞句。資料搜尋十分快速，短短的數秒至數分鐘，便能顯示出被檢索的字詞、出現在資料中的次數，及檢索資料量等統計數據；同時，亦可呈現出檢索詞的原文、出處及相關段落，或印出檢索所得的正文。這對研究者從事的資料搜集、整理、統計、分析等工作而言，實在是一大福音。

二、「台灣方志資料庫」製作緣起

「台灣方志資料庫」是「台灣資料庫」的先聲，其製作，緣起於民國七十三年，中央研究院歷史語言研究所與計算中心合作開發的「二十五史全文資料庫」。當時，計算中心藉處理二十五史檢索的機會，開發了一套通行的工具，適用於一般以文字爲主的中文典籍，稱爲「中文全文檢索系統」。本資料庫，即本處在台灣史田野研究室時期，參加史語所「史籍全文資料庫」的一個獨立項目。

鑒於「中文全文檢索系統」快速便捷的檢索功能，有助於各項台灣研究資料的建立；台灣史田野研究室遂於民國七十九年，在史語所臧振華先生擔任執行小組召集人時，開始規劃「台灣史籍自動化」的長期工作；同年二月七日，經「七十八學年度第三次執行小組會議」討論台灣史資料自動化案，作成決議，比照史語所「二十五史全文資料庫」，以完成「台灣方志資料庫」作爲第一階段的計劃目標。

此計劃隨即徵得史語所同意，及本院計算中心的大力協助；並立即於二月二十三日，邀請台灣史學界學者專家，群至計算中心，聽取丁之侃先生的示範講解。七十九年三月，計算中心輸入小組開始鍵入資料，迄九月，「台灣方志」全文四十六種、一百一十六册、八百萬字，輸入完畢。

八十年八月二十七日，由許雪姬小姐主持的「台灣史田野研究室第一次工作會報」中，決議與計算中心商討建立台灣方志全文資料庫之可行性。八十一年十一月，計算中心研發出新版的建檔程式，而當時已進行的部分仍爲舊版標誌，因此有新舊版標誌轉換的問題；其後中文系統改用倚天系統，與原先天龍系統不同，而有造字碼轉換的問題。然而，最後都獲得解決。

八十二年七月，台灣史研究所籌備處正式成立，台灣史籍自動化的工作，由黃富三主任繼續大力推動。同年年底，「台灣方志」第五次人工校對及修稿終於結束，又面臨參與連線建檔之電腦設備不足、建檔技術之訓練等諸多問題及難處，幸賴計算中心全力支持，而戴小琦小姐及林晰先生於建檔過程中之協助，更是「台灣方志資料庫」得以順利完成的關鍵。

三、「台灣方志資料庫」之簡介

「台灣方志資料庫」的方志版本，大部份選自台灣銀行經濟研究室於民國四十七年至五十七間所出版的「台灣文獻叢刊」標點本。但其中高拱乾、范咸、蔣毓英的《台灣府志》(簡稱：高志、范志與蔣志)，則採用1985年北京中華書局出版的版本；蔣志的標點，再以1985年廈門大學出版社出版的《台灣府志校注》爲參考依據。全文合計，共四十六種、一百一十六册，約七百六十萬字。

我們將此資料庫分爲三個部份：一爲通志、府志、縣志、廳志；二爲采訪册、一般志書和輿圖；三爲補闕。茲將各書有關的資料，分列如下：

表一　通志、府志、縣志、廳志

叢刊號	書　名	作(編)著	册數	出版年
68	清一統志台灣府		1	49
84	福建通志台灣府		6	49
130	台灣通志		4	49
65	台灣府志	高拱乾	3	49
66	重修台灣府志	周元文	3	49
74	重修福建通志台灣府	劉良璧	4	50
105	重修台灣府志	范　咸	5	50
121	續修台灣府志	余文儀	6	51
75	恆春縣志	屠繼善	2	49
103	台灣縣志	陳文達	2	50
113	重修台灣縣志	王必昌	4	50
140	續修台灣縣志	謝金鑾	4	51
124	鳳山縣志	陳文達	2	50
146	重修鳳山縣志	王瑛曾	3	51
141	諸羅縣志	周鍾瑄	2	51
156	彰化縣志	周　璽	3	51
159	苗栗縣志	沈茂蔭	2	51
160	噶瑪蘭廳志	陳淑均	4	52
164	澎湖廳志	林　豪	3	52
172	淡水廳志	陳培桂	3	52

表二 采訪册、一般志書與輿圖

叢刊號	書 名	作(編)者	册數	出版年
37	雲林縣采訪册	倪贊元	2	48
55	台灣采訪册	諸 家	2	48
58	嘉義管內采訪册		1	48
73	鳳山縣采訪册	盧德嘉	3	49
81	台東州采訪册	胡 傳	1	49
145	新竹縣采訪册	陳朝龍	2	51
48	苑裡志	蔡振豐	1	48
63	樹杞林志	諸 家	1	49
80	金門志	林焜熿	3	49
95	廈門志	周 凱	5	50
61	新竹縣志初稿	諸 家	2	52
101	新竹縣制度考		1	50
92	噶瑪蘭志略	柯培元	1	50
181	台灣府輿圖纂要		1	52
185	台灣地輿全圖		1	52
195	福建通志列傳選	陳 衍	3	53
233	泉州府志選錄		1	56
232	漳州府志選錄		1	56

表三 補闕

叢刊號	書 名	作(編)者	册數	出版年
104	澎湖台灣紀略	諸 家	1	50
109	澎湖紀略	胡建偉	2	50
115	澎湖續編	蔣 鏞	1	50
52	安平縣雜記		1	48
120	台灣通紀	陳 衍	2	50
243	清史稿台灣資料集輯		6	57
18	台灣志略	李元春	1	47
	台灣府志	蔣毓英	1	74

如前所述，古籍中的圖形仍是目前無法顯現而有待克服的困難。另外，在造字上，由於大五碼(Big5)的造字空間有限，而本院其他已完成的資料庫，又佔用大量的公共空間，導致本處發現的五百餘新字，只能選擇出現次數較多的字，納入公共造字檔；其餘的字，則用黑點替代、呈現。如下例：「紅花、胭脂米、●蘭米百斤，例六錢……」(《廈門志》，頁326)。本處將在所發行的使用手冊，一律附上缺字對照表，以方便使用者。

四、「台灣檔案資料庫」之賡續

本著「台灣方志資料庫」的經驗，史籍自動化室於八十二年底開始計劃第二階段的工作目標，計算中心亦同意繼續支援輸入及建檔工作；因此，本處目前正在進行第二階段的工作——「台灣檔案資料庫」。

此資料庫，仍取材自台灣銀行經濟研究室的「台灣文獻叢刊」。我們選取其中的實錄、檔案、彙錄等，做爲輸入的材料，合計九十七册，約八百萬字左右。參見表四。

有鑒於第一階段校對工作耗時經年，因而在第二階段進行項目中，已將人工的第一校，改爲程式比對的方式；亦即由兩個輸入單位分別繕打資料，然後藉程式比對兩批資料，以指出相異之處，再行修改；如此，將可節省不少時間。相信在技術的不斷改進與各單位的協助下，第二階段工作會進行的更加順利。

五、「台灣方志資料庫」上線——中文全文檢索系統之妙用

「台灣方志資料庫」在經過逐字輸入，及五年時光的比對、標誌、校對、建檔等流程後，利用中文全文檢索系統(ftms)製成。本文在此，對「中文全文檢索系統」之妙用，做一簡單的說明。

中文全文檢索系統可提供閱讀、檢索二大功能，其特性及使用方式如下：

(一)中文全文檢索系統之特性

1. 所謂中文全文檢索系統，係藉著層級式的目錄來反映書本上的章、節、段落，以利使用者調閱正文或訂定檢索的範圍。
2. 保留原書頁碼，以便調閱正文，提供檢索詞的出處，方便使用者參照書本。

表四　台灣檔案資料庫內容

叢刊號	書　　名	册數	出版年
27	劉壯肅公奏議	3	47
29	福建台灣奏摺	1	48
31	台案彙錄甲集	3	48
38	同治甲戌日兵侵台始末	2	48
49	東溟奏稿	1	48
62	楊勇愍公奏議	1	48
88	左文襄公奏牘	1	49
110	台灣海防檔	2	50
158	清世祖實錄選輯	1	52
165	清聖祖實錄選輯	1	52
167	清世宗實錄選輯	1	52
173	台案彙錄乙集	4	52
176	台案彙錄丙集	2	52
178	台案彙錄丁集	2	52
179	台案彙錄戊集	3	52
186	清高宗實錄選輯	4	53
187	清仁宗實錄選輯	1	52
188	清宣宗實錄選輯	3	53
189	清文宗實錄選輯	1	53
190	清穆宗實錄選輯	1	52
191	台案彙錄己集	3	53
192	法軍侵台檔	4	53
193	清德宗實錄選輯	2	53

200	台案彙錄庚集	5	53
203	籌辦夷務始末選輯	3	53
204	法軍侵台檔補編	1	53
205	台案彙錄辛集	2	53
210	清光緒朝中日交涉史料選輯	3	54
226	清會典台灣事例	2	55
227	台案彙錄壬集	1	55
228	台案彙錄癸集	1	55
231	吳光祿使閔奏稿選錄	1	56
236	籌辦夷務始末選輯補編	1	56
243	清史稿台灣資料集輯	6	57
247	清季申報台灣紀事輯錄	8	57
253	述報法兵侵台紀事殘輯	2	57
256	清奏疏選彙	1	57
262	東華錄選輯	2	58
273	東華續錄選輯	2	57
276	劉銘傳撫台前後檔案	2	58
277	光緒朝東華續錄選輯	2	58
278	清季台灣洋務史料	1	58
285	李文襄公奏疏與文移	3	59
288	道咸同光四朝奏議選輯	3	60
290	台灣對外關係史料	1	60
300	雍正硃批奏摺選輯	2	60

3. 正文具備橫、直二種格式，各段落均儘可能照原書加以編排。

4. 檢索條件由一個或多個檢索詞組成。各詞之間，可用「｜」(或)、「＆」(且)、「！」(且非)三種符號，規範彼此間的關係。檢索詞前後，亦得附加

排除字集 —— 以符號{ }表示，利於更精確的檢索。舉例如下：

例1：減免¦除

表示凡段落內出現「減免」或「除」任一詞，皆爲所求。

例2：減免¦除&田租¦口賦

表示凡段落內出現「減免」、「除」及「田租」、「口賦」任一詞，皆爲所求。

例3：減免！除

表示！後的條件不得成立，只有「減免」一詞爲所求。

例4：{國扶遺}風{伯后師}

表示爲蒐集較純粹的氣象資料 —— 風，凡風的前後出現{ }內的任一字，檢索時即遭到排除。

5. 檢索的範圍可小至段落、大至整個資料庫，檢索結果可列出含有檢索詞的句子或段落，或者將檢索詞置於行列中間，而兼取前後文來呈現。

(二)中文全文檢索系統之使用方式

1. 使用者的電腦，需有倚天中文系統或其他相容的系統。若屬初次使用，需先裝設造字檔。(操作步驟，詳見使用手册。)
2. 使用中文全文檢索，需先選定適當「資料庫」—— 如「台灣方志資料庫」，才能調閱正文。(指令使用，詳見使用手册。)
3. 檢索方式有二種：(1)一般檢索(2)引得檢索。(操作步驟，詳見使用手册)

以下，試以檢索「檳榔」一詞爲例，說明中文全文檢索系統之特性與功能。

以一般檢索輸入「檳榔」一詞，進行檢索。檢索結果，即出現如下畫面：

————————————《 計274項 》————————————

費時　12.00秒

檢索　76050082字

找到　274段，274項，502詞

意即，在12秒內，電腦已爲使用者檢索完畢整個「台灣方志資料庫」，並在四十六種方志、七百六十萬字中，找到274段含有「檳榔」一詞的資料。

繼續查閱後，則可以清楚列出如下清單：

－－－－－－－－－－－－－－－－－－－－《 計274項 》－－－－－－－－－－－－－－－－－－－－

【 詞 彙 】	【 出 處 】	【 路 徑 】	【 數 量 】
1.檳 榔	清一統志臺灣府・正文 44頁	/ 2.1.2.1.25	2
2.檳 榔	福建通志臺灣府・正文203頁	/ 2.2.2.19.1.1.1	5
3.檳 榔	福建通志臺灣府・正文213頁	/ 2.2.2.19.1.1.10.1	1
4.檳 榔	福建通志臺灣府・正文216頁	/ 2.2.2.19.1.1.10.4	1
5.檳 榔	福建通志臺灣府・正文221頁	/ 2.2.2.20.1.1	6

此時，若再應用引得檢索，可以呈現出另一種檢索結果，如圖例：

－－－－－－－－－－－－－－－－－－－－《 引 得 》－－－－－－－－－－－－－－－－－－－－

1.正 文	61 蕉：包刀花生：菰辣生	檳榔	干汁
2.正 文	31 柿 柚 鳳梨(一名黃梨)	檳榔	(和荖葉夾食之，
3.正 文	612 夜喃喃・泥人夢裡含雞舌，一椀	檳榔	出枕函(指甲花五
4.正 文	493 牽 手爲謀家室結潘、楊，花菓	檳榔	滿彩筐，祝罷無違
5.正 文	239 義縣菜園一所，徵銀一兩・蕭壠	檳榔	二十四宅，徵銀六

台灣方志，主要是清領時期台灣各級行政區劃 —— 府、縣、廳所編纂的志書。內涵山川形勝、建制沿革、城市街道、聚落分佈、政經設施、文教武功、風土人情、民俗語彙及人物典章等；既有包羅萬象、「百科全書式」的豐富，更因記錄翔盡，落實於各地鄉土，具體而微展現台灣社會的歷史風貌。因此，台灣方志可謂從事台灣研究必備的基本工具書。不僅歷史學，包含考古學、語言學、人類學、社會學、經濟學、政治學諸學科，凡有興趣於台灣者，皆不得不仰仗它；其使用率之高，引用之繁，早爲諸書之首。「台灣方志資料庫」的全文檢索功能，使台灣研究的基礎資料匯爲一爐、一舉而得，特此介紹。