文獻的索引數位化製作與研究—以「引得市」為例

Research Of Literature Digitization - Take INDEX as an Example

陳信良

Chen Shin-Liang

國立臺北科技大學文化事業發展系兼任講師

摘 要

林慶彰先生說：「學術資料是從事學術研究的基礎，能善用資料的人作研究時比較會有新的見解或發現，學術資料上從天文下至地理，上下古今縱橫數千年，累積之多，不是用人腦或電腦所可能完全掌握的。因此，歷來要檢索學術資料，往往先利用工具書。」

筆者一直想要探究，當我們在幾秒內就能獲得如此豐富資訊時，對本身的專業研究能有什麼創新或應用，研究者又要如何因應與面對，都是值得再探討的課題。創建「引得市」這樣文獻索引數位化整合的資料庫，期盼營造新的研究環境，提供更便捷的文獻索引，讓真正投入研究的學者，更能有效率的發揮所長。

【關鍵詞】索引、引得、引得市、檢索、資料庫、文獻應用、檢字表數位化

一、前言

林慶彰先生說：「學術資料是從事學術研究的基礎，能善用資料的人作研究時比較會有新的見解或發現，學術資料上從天文下至地理，上下古今縱橫數千年，累積之多，不是用人腦或電腦所可能完全掌握的。因此，歷來要檢索學術資料，往往先利用工具書。」[1]

在文史哲研究的領域中，絕大多數人還是以紙本書籍作為文獻的來源，針對字辭典文獻的查詢，使用全文資料庫來查詢，其「快速與便利」的優點或許已經不需要再多作說明。但是，一般人往往只會停留在「就是能比翻書查的快而已」類似的想法。筆者一直想要探究，當我們在幾秒內就能獲得如此豐富資訊時，對本身的專業研究能有什麼創新或應用，研究者要如何因應與面對，都是值得再探討的課題。

資料取得的方式或研究方法與手段，無論學習的管道是哪裡，最後都是取決於研究者個人的習慣，並沒有什麼方法是絕對好或壞，或許，未來應該加強訓練的是研究者對於資訊獨立的思考與分析處理的能力，容易獲得資料（資訊），不代表後續完成的「研究」就能有一定的水平。背後還需要有一顆清晰明辨，能剔除多餘資訊的腦袋。所以，無論資料來源取得便利與否，重點還是在研究者本身，

[1] 編者序，林慶彰，《學術資料的檢索與利用》，臺北：萬卷樓圖書股份有限公司，2003 年 3 月。

因為網路資料庫的大量使用，而讓研究者資料蒐集能力的降低，或者研究沒有創新與獨立的見解等問題，都不是本文所探討的範圍，我們期盼能夠營造新的研究環境，讓真正投入研究的學者們，有效率的發揮所長。因此創建了「引得市」[2]這樣文獻索引數位化整合的資料庫。

二、關於「引得市」

楊菁說：「無論從事哪一個領域的學術研究，都必須在專業知識和理論基礎以外，熟悉相關文獻，掌握蒐集資料和查考問題的基本方法；因此，藉助於工具書，以最便捷的方式檢索到完整的資料，是每一個研究者都需具備的條件。」[3]

一直以來，筆者不斷的在尋求更為便利與效率的文獻處理與檢索方式，經由重複的資料處理，嘗試改善過去的缺陷或效率低落的原因。而「所想即所得」，是筆者在自己的研究領域中一直期望也能夠實踐的一種研究手段，「想得到」就要能夠「查得到、看得到」，也要能夠「用的到」。能達到「用」的階段，也表示該資料的來源清楚，而且也已經「數位化」完成，已經是掃瞄成的圖檔或者文字檔。

此網站主要內容係筆者近期所製作的古文字相關字典索引資料。從紙本書籍數位化製作，再轉成資料庫提供網路檢索使用。因資料庫以書籍檢字內容為主，過去以來 INDEX 英文諧音為「引得」，INDEX 日本稱為「索引」（さくいん）。索引是分析著錄文獻事項和單元知識，並按一定系統組織起來的檢索文獻工具，目的是幫助讀者快速查到文獻中的相關資料。[4]兩者都是 INDEX 之意。

自取「引得」一詞，筆者增添一「市」字，定名為「引得市」。或許也有「引德士」之諧音。「市」有眾多含意，引《老子》：「美言可以市尊。」「換取」之意。研究者查書時，若能使用便利的電子索引檔、甚至資料庫，便利之餘，就能多出

[2] 由筆者創建的文獻索引資料庫，成立於 2012 年 7 月。網址：http://www.mebag.com/index

[3] 楊菁，〈中國哲學資料的檢索與利用〉，刊《學術資料的檢索與利用》，臺北：萬卷樓圖書股份有限公司，2003 年 3 月，159 頁。

[4] 彭斐章・喬好勤・陳傳夫，《目錄學》，武漢：武漢大學出版社，2003 年 11 月。

更多研究的時間。資料庫是創辦者依舊程式自行修改而成，程式陽春，但使用起來簡單，穩定與否也需多人測試才能知曉，文獻索引製作耗時，雖可藉助軟體工具，節省許多時間，但畢竟一個人的力量有限，若更多有心人若能齊心協力，協同製作各類書籍索引，再集中一處。日後或許就能有真正「市」的規模了。

（一）「引得市」資料來源

資料庫主要分成「字典」與「辭典」兩大類，依性質再分為「古文字」與「書道」，如此分類純粹是筆者個人的認知，係為方便資料庫管理，並不影響檢索。

字典類目前收錄：《秦印文字彙編》、《古文字通假字典》、《古文字類編・增訂本》、《秦簡牘文字彙編》、《馬王堆簡帛文字編》、《新金文編》、《楚文字編》、《睡虎地秦簡文字編》、《戰國秦漢簡帛古書通假字彙纂》、《簡牘帛書字典》。另有《中國草書大字典》與《中國隸書大字典》提供書法創作者檢索使用。辭典類則收錄是《中國書道文化辭典》。

主要的資料庫（九本書籍檢字表）字例數量已超過四萬餘字。未避免單次檢索顯示的資料數量過多，十一月新完成的《新金文編》已是另外獨立檢索的資料庫，並沒有和其他文獻並列，資料庫將陸續新增的有《甲骨文字編》、《中國行書大字典》與《中國正書大字典》等。

（二） 「引得市」使用方法

因為「引得市」係以資料庫的形式存在於網路，所以必須能夠上網連線，才能夠使用這些資料。使用者可用 Google 或其他搜尋引擎輸入「引得市」，或於瀏覽器網址列輸入「www.mebag.com/index」。主要提供：文字搜尋、筆畫、頁碼、檢字表頁碼等檢索方式，使用者可選擇其中一種方式來查詢。

⊙「引得市」資料查詢方式示意表

書名	文字搜尋	筆畫	漢語拼音	日文	頁碼	檢字表頁碼	備註
秦印文字彙編	○	○			○	○	一次搜尋九本書的檢字表
簡牘帛書字典	○	○			○	○	
馬王堆簡帛文字編	○	○			○	○	
楚文字編	○	○			○	○	
戰國秦漢簡帛古書通假字彙纂	○	○			○	○	
秦簡牘文字彙編	○	○			○	○	
古文字類編・增訂本	○	○			○	○	
古文字通假字典	○	○			○	○	
睡虎地秦簡文字編	○	○			○	○	
新金文編（上中下）	○	○	○		○	○	
中國草書大字典	○	○			○	○	
中國隸書大字典	○	○			○	○	
中國書道文化辭典	○			○	○	○	

特別值得一提，「引得市」能夠一次同時搜尋九本書的檢字表。這對資訊專業的人來說這樣的功能並不稀奇。但是，對從事文史哲研究的朋友來說，如此搜尋方式應當可以省去不少時間。例如，輸入「得」字，即顯示出 14 筆位置資料。配合紙本書籍，查得頁碼之後，立即翻到該頁閱讀。當然，既然能夠查得頁碼，原來文獻資料內容的連結自然也可以同時獲得。「引得市」成立以來，筆者都是以圖文對照的方式來作檢字表的內容校正，長期以來也習慣於螢幕上瀏覽內容，思考分析研究文獻的內容，目光快速的移動於文獻資料之間。

無論在世界的任何角落，筆者使用智慧型手機或上網的手持裝置，隨時隨地想查那個字都可以做得到。還能夠設定直接瀏覽圖檔（文獻內容）。無論是作為研究考察或者單純的文字藝術賞析。但是，在沒有獲得著作人授權之前，這些文獻的內容都是無法公開的。所以，像這樣便利方法也只能暫時由筆者一人獨享。一般使用者就只能「搜尋」檢字表。

文字	頁碼	檢字表頁碼
得	0232	0903
得	0231	0903
得臣	1494	0026
得	0263	0013
得	0035	0317
得	0036	0349
得	0074	0009
得	0120	0924
得	0024	0239
得	0394	1034

「引得市」後端介面 （未公開），游標點選「眼睛」圖示，即可瀏覽文獻內容。

NO.	書名	筆劃	文字	頁碼	檢字表頁碼	缺字圖示
1	古文字通假字典	11畫	得	0232	0903	☒
2	古文字通假字典	11畫	得	0231	0903	☒
3	古文字類編_增訂本	11畫	得臣	1494	0026	☒
4	古文字類編_增訂本	11畫	得	0263	0013	☒
5	秦印文字彙編	11畫	得	0035	0317	☒
6	秦簡牘文字彙編	11畫	得	0036	0349	☒
7	馬王堆簡帛文字編	11畫	得	0074	0009	☒
8	楚文字編	11畫	得	0120	0924	☒
9	睡虎地秦簡文字編	11畫	得	0024	0239	☒
10	戰國秦漢簡帛古書通假字彙纂	11畫	得	0394	1034	☒
11	戰國秦漢簡帛古書通假字彙纂	11畫	得	0389	1034	☒
12	戰國秦漢簡帛古書通假字彙纂	11畫	得	0383	1034	☒
13	戰國秦漢簡帛古書通假字彙纂	11畫	得	0048	1034	☒
14	簡牘帛書字典	11畫	得	0312	0982	☒

1 第 1 頁 / 共 1 頁 (總計 14 筆)

「引得市」古文字類，「得」字的搜尋結果（14 筆）

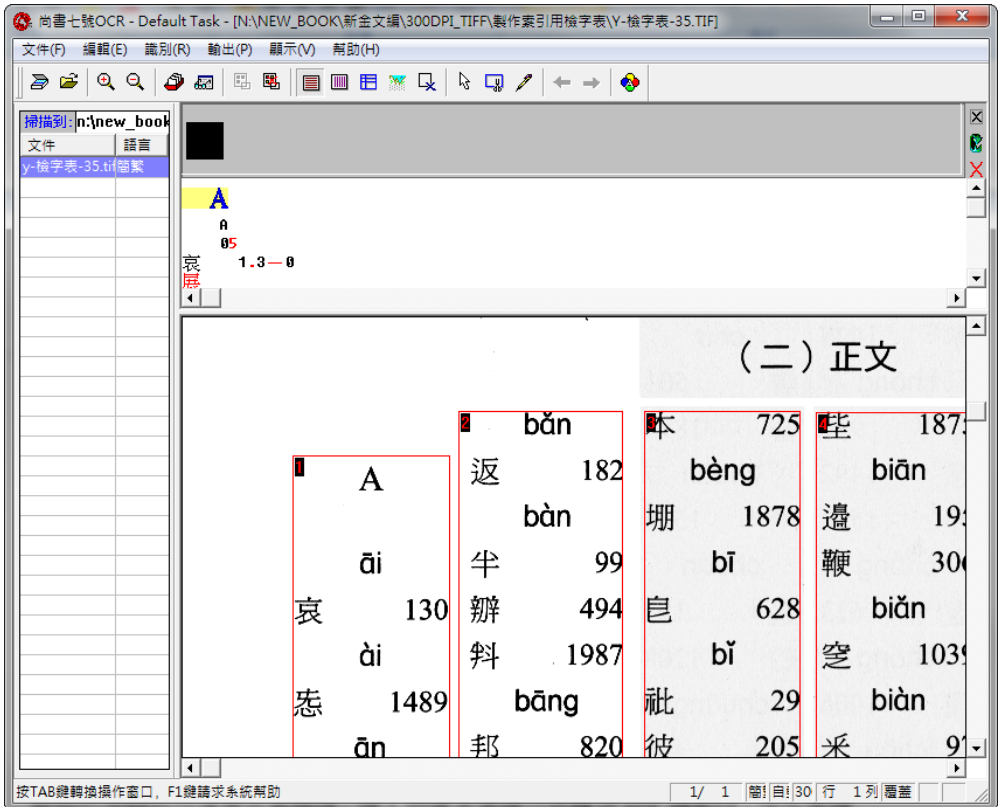
三、檢字表索引的數位化

「引得市」提供的檢索資料，數位化的過程大致上會經過四個步驟。即：1.紙本文獻掃瞄建檔、2.OCR 文字辨識、3.文字校正與缺字檢查、4.資料轉換與缺

字處理。略過第一步驟，我們直接從第二步驟開始說起：

（一）OCR 文字辨識

當紙本文獻已經掃瞄成圖檔儲存在電腦中，為了方便管理與查詢，也需要將檔案名稱處理成原本文獻相同的頁碼，處理的方法簡單，但若解說過多恐又離題，故請另參拙文[5]。OCR 辨識軟體的應用，我們以《中國書畫文獻索引》下冊 2924 頁為例[6]，OCR 使用軟體是「尚書七號」[7]。



尚書七號介面

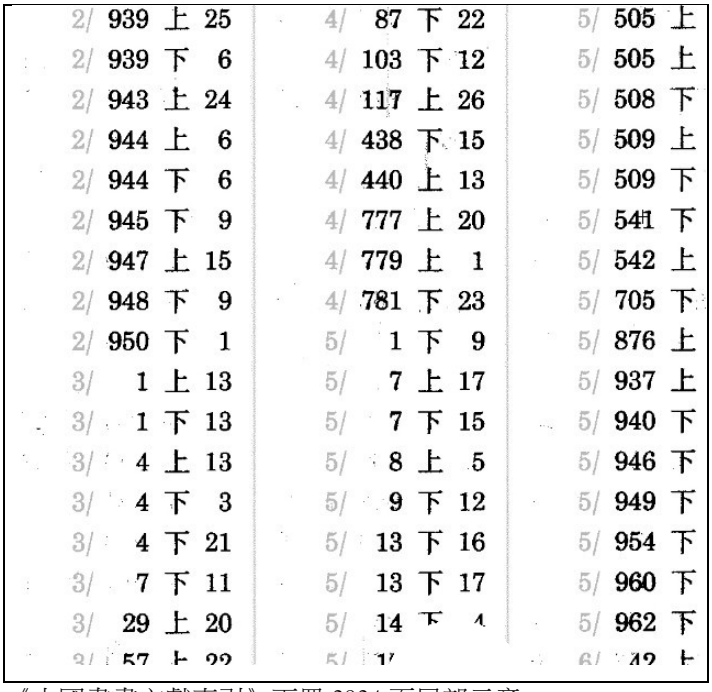
點選軟體左上角的功能列：文件→打開圖像（Ctrl+O）選擇需要辨識的檔案，或以「拖拉放」的方式把圖檔置入於軟體視窗左側。右側視窗中，以游標圈選需

[5] 筆者部落格文章：〈檔名修改應用示範-「甲骨文合集」圖檔為例〉（2011/12/13）http://blog.yam.com/ebag/article/45416338

[6] 筆者部落格文章：〈文獻處理 OCR 的應用——以《中國書畫文獻索引》下冊 2924 頁為例〉（2012/5/25）http://blog.yam.com/ebag/article/50275274

[7] OCR 文字辨識軟體，辨識率高，能同時處理數個甚至數百個以上的圖檔，辨識後的內容可匯出成文字檔。

要辨識的範圍（亦需視橫式或直式內容來調整）。辨識以前，圖檔是否先作編輯處理、刪除部分資訊，都需要長期使用的經驗才能作判斷。因為若選擇的內容太過複雜，反而會讓 OCR 識別不完整或者不精確，後續文字校正花更多時間重新調整。這裡，我們會把前段「2/」～「5/」先用影像編輯軟體刪除，這個部分因為是有規則連續的數字，於 EXCEL 處理時，也能很快的輸入。



2/	939	上	25	4/	87	下	22	5/ 505 上
2/	939	下	6	4/	103	下	12	5/ 505 上
2/	943	上	24	4/	117	上	26	5/ 508 下
2/	944	上	6	4/	438	下	15	5/ 509 上
2/	944	下	6	4/	440	上	13	5/ 509 下
2/	945	下	9	4/	777	上	20	5/ 541 下
2/	947	上	15	4/	779	上	1	5/ 542 上
2/	948	下	9	4/	781	下	23	5/ 705 下
2/	950	下	1	5/	1	下	9	5/ 876 上
3/	1	上	13	5/	7	上	17	5/ 937 上
3/	1	下	13	5/	7	下	15	5/ 940 下
3/	4	上	13	5/	8	上	5	5/ 946 下
3/	4	下	3	5/	9	下	12	5/ 949 下
3/	4	下	21	5/	13	下	16	5/ 954 下
3/	7	下	11	5/	13	下	17	5/ 960 下
3/	29	上	20	5/	14	下	4	5/ 962 下
3/	57	上	22	5/	17			6/ 42 上

《中國書畫文獻索引》下冊 2924 頁局部示意

辨識後的文字檔中，找出文字的規則，目的是能夠拆解分別「數字」與「文字」，詳細可參拙文：〈搜尋「\d」與「[^\d]」〉[8]。將文字與數字分別後，在 EXCEL 就能有順序的輸入其他資訊內容。

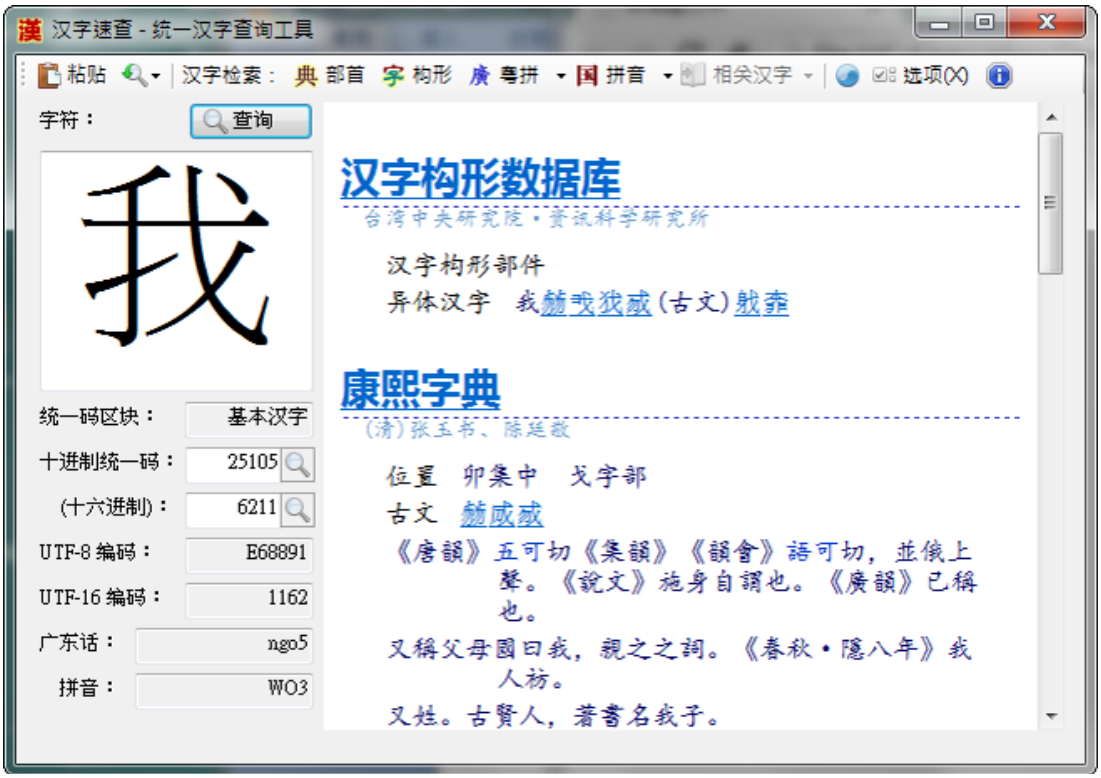
利用幾分鐘的時間，就可以快速的將內容數位化完成。此原稿中約有二百三十多筆，如果以打字的方式處理，一定很容易打錯字，打完字又得校正，不花上個一、二小時以上都很難。目前的 OCR 軟體辨識率高，尤其是英文或數字的部分，幾乎不會錯誤，圖檔使用 OCR 軟體辨識，再進行人工校正是相當有效率的方式之一。

[8] 筆者部落格文章：〈搜尋「\d」與「[^\d]」〉（2010/5/27）http://blog.yam.com/ebag/article/28828908

研究者花幾分鐘處理一份資料，應該是其有必要，如果需要一個小時以上才處理一頁，就值得商榷。研究文獻的數位化，不只是單純的掃圖或打字而已，除了軟體的應用之外，處理的先後順序也需要靠長期的經驗累積。

（二）文字校正與缺字檢查

二〇一二年六月中旬以前，筆者都是以「CHISE IDS 漢字検索」[9]網站來查詢缺字，但網站必須連結到網路，而且網路速度影響顯示的結果，檢查每個字的速度實在不是很快。此外，也會使用「漢字構形資料庫 2.7」[10]查詢，但是這軟體的缺點是，輸入欄位並不支援 Unicode，很多偏旁的內容都會變成「？」。為了不誤作缺字，我們會經由兩種以上的方法的檢查，才能確定是否為缺字，再新增造字。



「漢字速查」介面

[9] 查詢異體或特殊字漢字的資料庫。網址：http://chise.zinbun.kyoto-u.ac.jp/ids-find

[10] 中央研究院文獻處理實驗室所提供下載的漢字資料庫。網址：http://cdp.sinica.edu.tw/cdphanzi/

「漢字速查」介面 構形部件

自從在「國學數典」[11]知道了這個軟體後，就一直使用至今。六月期間，利用「漢字速查」軟體，先後完成了《秦印文字彙編》（1392 字例）與《楚文字編》（4894 字例）兩本書籍檢字表中「缺字」的檢查。因為軟體相當實用，所以特別推薦。此軟體作者也是「PDF 補丁丁」[12]的作者。亦是筆者推薦愛用的軟體之一。漢字速查（HanziSearcher）是一個支持統一碼（Unicode 6.1）七萬漢字的全漢字集檢索工具。其功能有：

- 使用部首筆畫、國語拼音、廣東話粵拼、漢字部件組合、內碼輸入等方式檢索漢字，可用於輸入難檢字、生僻字。
- 集成可擴展的字典功能（現有康熙字典、Unihan 等字典數據庫），可用於查詢漢字字音、字義。
- 音韻檢索功能可用於檢索同音字、同韻字。
- 字典查詢結果帶有超級連接，可在相關漢字之間靈活跳轉。[13]

[11] 對岸知名的網路論壇，是各種研究資料的交流平臺。網址：http://bbs.gxsd.com.cn/forum.php

[12] 係 PDF 書籤製作的工具軟體。網址：http://pdfpatcher.cnblogs.com/

[13] 軟體特色介紹引用自「漢字速查」網站。網址：http://hanzisearcher.cnblogs.com

軟體提供「部首」、「構形」、「粵拼」、「拼音」等方式。在「構形」檢索方法中，輸入「東鳥」就能查得「鶇」字、「馬 2」能查得「驫」字、「馬 3」能查得「驫」字。是當相直覺與便利的查字方式。使用者不需要有深厚的古文字背景，就能輕鬆面對諸多奇形怪狀的古字，與紙本對照，亦能校正處理古文字字典內的檢字表。突破校正者一定得具備古文字專長的門檻，解決輸入古文字必須知道「讀音」才能打字的問題，因此，古文字的應用層面將更加擴展。除了「查字」之外，也能夠利用這個軟體來當「輸入」古文字的工具，當我們想查詢「 龍」這個字時，卻又不知道發音或如何輸入，可以在該軟體「構形」檢索中，輸入「車龍」，查出後複製此字，再貼到「引得市」的搜尋欄中，就可以檢索。

絕大多數的漢字，只有古文字學者才用得到，一般人不認識也不需要學習，利用這個軟體，在一秒之內，就可以知道文字的字義，以及目前電腦上是否存在與使用。現在輸入法往往無法對應這些特殊罕用字體，能夠補強的，就是這個軟體了。即使，有人能夠完全記得這七萬多個漢字，在這個時代恐怕也沒有發揮的舞臺。筆者以為研究者無須急著擴充自己的「記憶體」或「硬碟」，而是要能夠獨立思考，強化自己浮點運算（分析及統一整合能力），善用周邊的軟體並發揮到極致。傳統的研究方法之外，軟體的應用也是相當重要的研究輔助手段之一。

（三）資料轉換與缺字處理

1.資料轉換

這個部分，筆者以《楚文字編》檢字表[14]的處理過程來說明。《楚文字編》筆畫檢字表（911-944 頁）共 34 頁，經一次校正後的結果，在 EXCEL 所呈現的，總計 4894 筆，缺字部分有 2159 筆。將近一半是缺字，應是目前古文字字典的極端特例。可輸入非缺字則為 2739 筆。這樣的整理，粗估約花費 60 個工作小時，（不含掃瞄圖檔）因為使用了「漢字速查」這樣方便的軟體，加上筆者長期以來處理古文缺字的經驗。換做他人或許要花上二、三倍以上的時間。

[14] 筆者部落格文章：〈《楚文字編》筆畫檢字表的索引數位化〉（2012/6/26）http://blog.yam.com/ebag/article/51399795

有了《秦印文字彙編》檢字表數位化等文獻的處理經驗，目前「《楚文字編》筆畫檢字表」已效率地完成前段工作，缺字量是前者的一百倍。光是前段的處理時間也花了兩倍以上。經估計，若要再完成缺字的處理，經粗估計算還需要 35.9 小時。即使，2159 個缺字字中，有一小部分是「印學相關缺字資料庫」已經製作完成的。

將來，若能完成「《楚文字編》所有缺字的建置，對於先秦時代，楚系的缺字大致上應該也已經能包含在內了。學者們就不必為了寫文章，又臨時組字合成，使用「醜醜難看」的字型圖檔了。期待這份索引檔全數完成的一天。完成辨識並校正二次左右以後的 xls 格式（EXCEL），即以複製貼上的方式轉換至「引得市」資料庫中。

	筆畫	起筆	文字	頁碼	原書頁碼
317	六畫	[一]	吉	70	913
318	六畫	[一]	正	91	913
319	六畫	[一]	共	160	913
320	六畫	[一]	舁	163	913
321	六畫	[一]	臣	190	913
322	六畫	[一]	寺	194	913
323	六畫	[一]	攷	200	913
324	六畫	[一]	●	203	913
325	六畫	[一]	百	226	913
326	六畫	[一]	再	248	913
327	六畫	[一]	死	253	913
328	六畫	[一]	过	283	913
329	六畫	[一]	荆	313	913
330	六畫	[一]	机	343	913

《楚文字編》檢字表 xls 格式

2.缺字處理

凡是電腦無法輸入或顯示的，我們稱為「缺字」。很多字體並非「缺字」，只是我們平時極少使用，也並不是電腦打不出來，這樣的字。通常我們會稱作「罕用字」。如果使用者不清楚目前自己的電腦對於罕用字體支援程度，可以連上網路至「中國哲學書電子化計劃」字體測試頁瀏覽對照。[15]例如，筆者的電腦使用的作業系統是微軟的 WINDOWS 7 家用進階版（32 位元），CJK 擴展 C 區與 D 區基本上是無法顯示的。雖然今日網路已提供 CJK 擴展 C、D 下載路徑，因筆者尚未使用，故於此不論。關於「引得市」收錄文獻中，總字例、缺字數量統計與處理情形，請詳見文後附錄。

「中國哲學書電子化計劃」字體測試頁

「引得市」延用「印學缺字資料庫」對缺字編號與處理的方式，「s085-060」s 代表缺字，s 之後的號碼是缺字的編號，如前半段的「085」代表「水」部，「060」代表第 60 個缺字。相關部首代號，請參考「印學缺字資料庫」[16]。因紙本不清晰，又以「?」符號表示對於該字不確定。「q」亦是缺字的代號，但目前尚未整理編

[15] 「中國哲學書電子化計劃」字體測試頁：http://ctext.org/font-test-page/zh

[16] 筆者 2004 年創建，已收錄超過 1000 餘字，包含古文字學、印學、書學、等缺字，是一個缺字集中管理的資料庫，網址：http://www.mebag.com/word

號。本資料庫缺字製作與技術支援由「印學缺字資料庫」與「glyphwiki」[17]。「glyphwiki」提供使用者自行製作缺字，因為是日文介面，或許一般人不容易使用，筆者曾有介紹。需要自作字體者可詳參部落格文章。[18]

NO.	書名	漢語拼音	筆劃	文字	頁碼	檢字表頁碼	缺字圖示
1	新金文編	xing2	6畫	s018-012(刑)	0628	0047	刱
2	新金文編	jia2	10畫	s050-008	1063	0040	帢
3	新金文編	ju1	15畫	s075-042	0731	0040	稟
4	新金文編	kang1	18畫	s115-003	0951	0040	穅
5	新金文編	li4	22畫	s108-013(盭)	1782	0041	盭
6	新金文編	qi4	18畫	s024-001(龕)	0457	0043	龕
7	新金文編	gu3	13畫	s066-024(鼓)	0390	0038	鼓
8	新金文編	wu2	16畫	s037-006(無)	0746	0046	橚
9	新金文編	zhi4	16畫	s072-011(智)	0429	0049	斲
10	新金文編	xuan2	16畫	s120-035	1285	0047	縣
11	新金文編	yan1	12畫	s094-006(猒)	0540	0047	猒
12	新金文編	yi4	11畫	s030-019(啻)	0253	0047	啻
13	新金文編	yi4	13畫	s058-001(緆)	0356	0048	緆
14	新金文編	ying1	19畫	s104-031	0482	0048	癭
15	新金文編	ying4	19畫	s149-032	0252	0048	鷹
16	新金文編	shu2	19畫	s123-007(鞫)	0317	0045	鞫

1 第 1 頁 / 共 1 頁 (總計 16 筆)

引得市《新金文編》缺字示意。《新金文編》一共有 16 字缺字，目前為止，這些「缺字」筆者都是以「圖形」作處理方式。

四、文獻處理與數位化文獻的應用

（一）文獻處理

研究的文獻蒐集與整理，過去多數人都是將書籍文獻逐頁影印、剪貼，或逐

[17] 網址：http://glyphwiki.org/wiki/。此字型製作網站是由「上地宏一」（かみち こういち KAMICHI, Koichi）先生所創建，提供登錄者自由創造與建立字型，亦是電腦缺字的解決與處理漢字異體字的工具之一。目前超過 280000 字例，上地先生目前任教於大東文化大學中文系，筆者曾多次當面請益關於資訊處理與缺字等問題。網站所製作的「花園字型」已超過八萬字，提供免費下載，網址：http://fonts.jp/hanazono/。

[18] 筆者部落格文章：〈「glyphwiki」線上製作字型（新帳號建立與登錄篇）初級〉（2011/7/7）http://blog.yam.com/ebag/article/39505788。〈「glyphwiki」線上製作字型（製作字型建立群組及使用字型篇）高級〉（2011/7/7）http://blog.yam.com/ebag/article/39506051

頁用標籤貼紙等方式，或許已經不符合追求效率的現況。面對資訊科技的時代，我們認為「熟練的軟體應用」才能達到「文獻蒐集儲存」與「整理」雙方面兼備的效果，也能夠加快研究的時程。

有人或許會說：「研究工作已經很忙碌了，還要另外花時間學軟體，不是很奇怪嗎？」筆者認為，唯有學會這些軟體，才能更快、更精確的完成研究。軟體種類眾多，只要挑選自己適用的，經常使用就能很容易上手。而且，絕大部分是自由下載的「免費軟體」，所以，沒有速成方法，還是需要多作嘗試並且經常使用。面對一本約五百頁的重要參考文獻，一般研究者的處理方法是？[19]

A.翻閱目錄，快速的閱讀並把重點部分貼上標記。（幾小時內完成）

B.沒有速讀的能力，慢慢讀，再把需要的地方影印起來。（24 小時內完成）

C.擺著，有空再翻翻看，或許有一天會用上。（時間未知）

筆者所知道的，A 方法通常是具備豐富研究經驗的學者才能做到，而且有「博學強記」的真功夫。但是，過目不忘的能力也不是每人都有，我們怎麼才能做到這樣的程度，甚至「超越」？筆者平時有空，就會持續地整理研究相關的文獻，最常使用的方式，就是「掃瞄」。

如果這份文獻相當重要，且值得收藏，那就值的花幾個小時來掃瞄，以現在的掃瞄機的速度，幾乎等於影印的時間，甚至更快。據筆者瞭解，仍有很多人寧願花幾個小時影印書籍，也不願「掃瞄」書籍，掃瞄完後的圖檔，可以製作成 PDF。PDF 還可以轉成文字檔，達到「全文搜尋」的使用狀態。同樣的時間花費，影印完的資料，就只是一疊的 A4 紙，既不容易收藏，若沒有認真讀，這些文獻跟你的「關係」還是「零」，如果又沒有裝訂，沒幾天可能就散落在各角落了。

原本五百頁的文獻從紙本變成了 PDF 格式的電子檔，所花費的時間，其實很短，最多只有幾個小時，明確的說，用 A3 的掃瞄機來處理，約只要二小時。

[19] 筆者部落格文章：〈面對文獻〉（2011/7/3）http://blog.yam.com/ebag/article/39396872

若要再進一步使用，利用「讀取革命」[20]轉成 doc。（使用 WORD 就能開啟，而且能修改內容）。

這時 500 頁的文獻，對使用者來說就是可以全文檢索的資料庫，我們可以輸入「關鍵字」查詢需要的內容，透過這樣不斷的「搜尋」與「瀏覽」，對這份文獻的印象與瞭解必定更加深刻，也許僅有五百頁的文獻並不夠展現出全文搜尋的特色，如果有十本、二十本這樣的文獻都是這樣的全文檢索，研究的核心瞬間就在指尖竄出，研究者的「個人分析與理解能力」就成為重點。一段時間經過，研究者自己就能開拓出獨立的研究領域。

即使把同樣的資料丟給十個人，作出來的研究內容也不會一樣，如果全部相同，那表示其中必有抄襲，沒有加入新的研究與看法。假使古人有電腦可以輔助，應該也會很自然的用滑鼠鍵盤，現代我們利用適當的軟體來輔助研究也是很合情合理的。能夠善用工具，轉變文獻使用的方法，五百頁的文獻，與它的關係在一瞬間就變緊密了，不會僅是眼睛所看到的排列整齊條列的方格字而已。

沒有實際參與製作整個流程，一般人很難瞭解文獻數位化實際上所需要花費的時間。關於數位化文獻的時間概算，為此筆者於部落格曾有一文說明：〈研究用書籍紙本數位化時間試算表 1.0 計算表〉，[21]以《楚文字編》為例，以灰階 600dpi 的解析度，本文 910 頁、檢索表 34 頁、缺字 2158 字，計算的結果是：掃瞄時間 3.03 小時、檢字表製作 51 小時、缺字製作 17.98 小時。一共需要約 72 小時。一天八小時，則需要約九個工作天才能完成。

由此可知，掃瞄所佔的時間並不是最多，主要是在檢字表數位化的過程很花時間，無法由機器的效能上加快速度，必須人工的逐字檢查，而我們使用的 OCR 文字辨識後再校正的方式，也一定比重新打字的方式快。檢字表的數位化，才能

[20] 原產品名稱為「読取革命」（日文版），能將 PDF 格式的檔案轉換成 DOC、XLS 等常用格式，試用版網址：http://panasonic.co.jp/snc/pstc/products/yomikaku/index.html

[21] 筆者部落格文章：〈研究用書籍紙本數位化時間試算表 1.0〉（2012/8/5）http://blog.yam.com/ebag/article/52617385

把把字辭典的發揮到極致。否則掃瞄後的圖檔，不容易搜尋，自然也不常使用，如此這些圖檔也只是佔據硬碟空間而已。當我們完成了字典檢字表的數位化，轉成資料庫來使用難度就不高了，在「引得市」就是最好的印證。

此外，在「檢字表數位化」校正的時間上，一般的字典處理的時間應該是低於 60 分（每頁）[22]，但若是古文字字典，缺字與罕用字很多，在考察所花費的時間與精神則是不容易估計，依照經驗平均下來都能在 90 分鐘內處理好。缺字製作方面，線上編輯的時間也能在每字 30 秒內處理完成，有些字甚至能在十秒以內。主要的時間會使用在對照編號方面，為了避免重複造字，這些步驟與時間也節省不了。

筆者認為，研究者應該找出自己研究中「重複的動作」是什麼？如果常需要使用到這本文獻，它就最好就是「電子檔」，而且附有完整的索引。期盼未來的研究者養成「掃瞄」的好習慣，如果這份文獻經常引用，就不要只是需要用哪裡，影印哪裡，既不環保也浪費紙張，並不是文獻處理的好方法。筆者一直認為：「東西好用方便，大家才會常用」。也絕不是任何書籍文獻都需要「數位化」，若它是具備研究參考價值的文獻，就一定值得花時間處理與製作。

（二）數位化文獻的應用

在文獻處理的效率與應用上，依照優劣排序，個人的看法是：影印文獻＜翻拍的文獻圖檔＜掃描的文獻圖檔＜有頁碼命名順序的文獻圖檔＜文獻圖檔轉成 PDF＜有 OCR 文字辨識的 PDF 文獻＜具備書籤的 PDF 文獻＜數位化的檢字表＜書籍文獻文字檔。

目前筆者比較推薦倒數第二種的文獻處理，這種方式也一直持續進行中，最後的一項的「書籍文獻文字檔」因礙於著作權，目前能實現的機率相當低。從中也可對比出「前」不如「後」的方便與實際。總結來說，「經過 OCR 文字辨識又有書籤的 PDF 格式檔案」，就能實現檔案內容的「全文搜尋」，書籤的建立也能

[22] 因每份文獻每頁檢字表字例數量不一，這裡的一頁約指數量 200 左右的字例。

快速的連結到想找的位置。

1.PDF 的應用

目前 PDF 已是相當普及使用的電子格式，透過網路傳遞分享交流都相當方便。紙本文獻的數位化，我們是以「掃瞄」方式，將圖檔儲存在電腦中，也會再將圖檔依序命名再轉檔成為 PDF 格式。筆者認為將圖形檔案轉換成 PDF 的理由是：

(1)方便管理，有利於開啟瀏覽。[23]

掃瞄後的文獻圖檔若沒有整理，就只是儲存於資料夾中，不僅無法搜尋索引，當數量龐大時，電腦的瀏覽執行速度也會變慢。若擔心佔據硬碟空間過多，在不影響圖檔品質的原則下，原來這些圖檔所存在的資料夾，也能合併壓縮成單一個 RAR 檔案。如果使用習慣用「紙本」來閱讀，PDF 檔也能隨時列印。

(2)PDF 裡面包含書籤[24]註釋等，能對檔案再做編輯與修改。

將文獻掃描成零散的圖檔儲存在電腦中，偶爾使用看圖軟體瀏覽內容，就是沒有把「數位化」的功能發揮到淋漓極致。當我們把圖檔整合成單一個 PDF 檔後，再經過 OCR 的文字辨識，整個文獻除了能在螢幕瀏覽之外，也能作初步的搜尋檢索（雖然無法達到百分之百）。單一本文獻的 PDF 檔案或許感受不到「搜尋」的重要，當有數個重要的文獻 PDF 檔案同時間一起搜尋時，就能讓使用者感受到全文檢索對研究的便利性。

我們曾以「中鋒」為關鍵字，搜尋範圍涵蓋了十六個以 PDF 格式儲存的書

[23] 當我們知道頁碼，使用 PDF 也能迅速「跳頁」到該頁面，詳參筆者部落格文：〈PDF 檔案跳頁連結 以「中國書畫文獻索引」為例〉（2011/5/18）http://blog.yam.com/ebag/article/38339180

[24] 關於 PDF 書籤的製作，可參考筆者部落格文章：〈PDF 書籤製作示範——以《宣和書譜》〉為例（2012/5/24）http://blog.yam.com/ebag/article/50236986

論相關的書籍，大約只花費十分鐘左右，結果共有五百多處[25]。點選條列隨即跳至該頁內容，隨選（點）隨看，顯示的卷期與頁碼，再一筆一筆手寫抄錄下來。如此，也一定比我們從頭到尾的翻閱檢查再記錄的來的精準與省時。這就是以 PDF 搜尋的應用之一。這樣對於文獻能夠深入、清楚地察知，已經是古人所說的「洞察」。藉助種種工具軟體，以及研究者個人獨立的思考分析與研究整理，就能達到古人用盡一生都難以做到的研究成果。

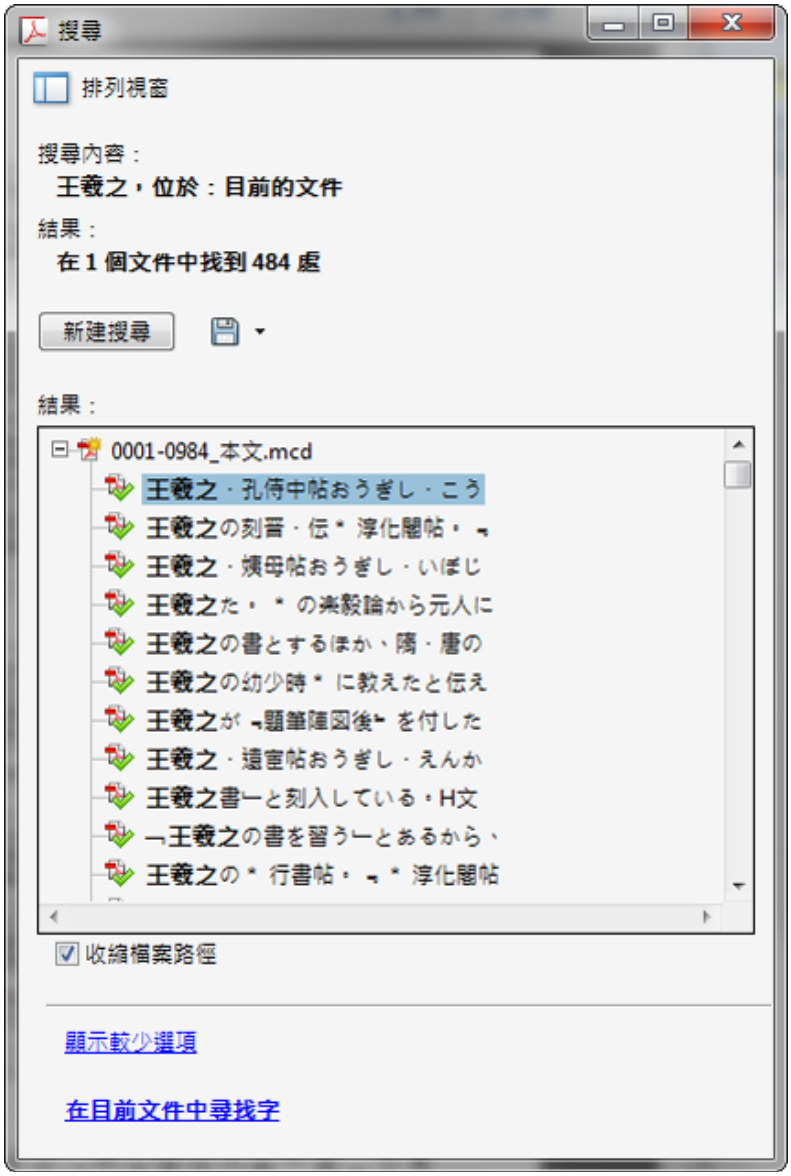
2.PDF 與「引得市」檢索性質的差異

「引得市」目前唯一收錄的辭典類索引是日本學者西林昭一所編輯的《中國書道文化辭典》[26]。該書在二〇〇九年六月出版，隨書附有一張光碟，內含的 PDF 格式文字檔案（非圖形檔，可全文檢索）。這樣書籍附錄光碟的作法，算是一種相當大的創舉，這樣紙本與光碟並存的情況，在文史哲或書畫研究領域中相當罕見，很值得推展。多年前，筆者曾提出這樣的一種概念，學術著作即使不想把全文做成光碟，至少也把「索引頁」的文字檔，放在網路或者透過其他管道來分享使用。

筆者發現，『中國書道文化辭典』的紙本（書），有一部份的文字檔沒有附在光碟中，就是按照五十音排列的「索引」共有 75 頁。令人不解的是，為何全文內容都已經放在光碟中了，為何最重要的「75 頁索引」卻沒有放在裡面？或許，書籍作者認為利用開啟 PDF 軟體的索引就能代替索引，但經過實際操作的經驗是，PDF 格式的檔案全文檢索有時後並不會比較快。因為如果輸入的關鍵字是一般常見的詞彙，可能出現的數量就會多的嚇人。全文檢索適合想要查詢該名詞在全書的各個位置，若只想要知道「印学」的解釋，翻到「い」開頭的頁面，然後找到「印学　いんがく　11」在 PDF 檔案中，輸入「11」就能直接跳頁到該處。

[25] 我們以《中國書論大系》、《精萃圖說書法論》、《國譯書論集成》三套書籍一共 16 個 PDF 電子檔作為搜尋的來源。並且已先確認「中鋒」二字在 PDF 的辨識文字裡正確無誤。筆者部落格文章：〈Acrobat 搜尋功能的應用之一〉（2011/5/14）
http://blog.yam.com/ebag/article/38227494

[26] 筆者部落格文章：〈《中国書道文化辞典》的索引製作和應用〉（2010/6/12）
http://blog.yam.com/ebag/article/29050664



《中國書道文化辭典》PDF「王羲之」搜尋結果

使用附錄的 PDF 檔案，我們若以「王羲之」作為關鍵詞查詢，即出現了 484 處，但事實上我們只是要查詢「王羲之」的介紹，並不是要查尋所有出現過「王羲之」的位置。但是如果搭配索引頁，能更快的查詢到想找的內容解釋。所以筆者在二〇一〇年已經把 75 頁的索引數位化完成，詞彙文字內容並且應用於日文輸入法中[27]。如果以「引得市」來查詢「王羲之」，查詢結果僅有 20 筆，即能有效率的獲得想要的資訊。雖然筆者很強調書籍的數位化和全文檢索的重要性，但

[27] 筆者部落格文章：〈「辭典索引檔」在 Google 日文輸入法的應用〉（2011/5/9）
http://blog.yam.com/ebag/article/38099901
筆者部落格文章：〈google 日文輸入法匯入中國書道文化辭典索引〉（2011/5/24）
http://blog.yam.com/ebag/article/38509195

也不會執著在任何書都一定要做成「電子檔」，其實，只要「索引」能有電子檔，快速的檢索，會比花很多時間製作全書的電子檔還實際。紙本索引頁和電子書的相互搭配，會比電腦笨笨的從頭搜尋到尾更有效率。

因此，目前紙本文獻還不能完全的被取代，仍有存在的必要。例如：筆者習慣將以黃色螢光筆在紙本檢字表上的「缺字」畫上記號，若日後發現此字並非缺字，也能再使用「橘色」的螢光筆補上，以作分別。這些淺色的記號，日後紙本書籍若再掃瞄或影印都不會影響字跡。隨時翻頁檢查缺字內容，在四周空白處寫上說明或註記，這些並不是單單用螢幕瀏覽、檢索資料庫所能取代的。

現在仍有許多文獻的索引工作持續進行中，例如《中國歷代人名大辭典》索引的數位化，約有五萬四千餘筆資料，還需從頭到尾與紙本對照校正。我們也可以說，當索引頁數位化製作完成時，就幾乎等於文獻可以達到近乎全文檢索的目標了。

NO.	日文發音	五十音	名詞(日本漢字)	頁碼	檢字表頁碼	缺字圖示
1	おうぎし-しゅうげつじょう	お	王羲之-秋月帖	0048	0005	☒
2	おうぎし-いまじょう	お	王羲之-姨母帖	0046	0005	☒
3	おうぎし-えんかんじょう	お	王羲之-遠宦帖	0047	0005	☒
4	おうぎし-かいせつじせいじょう	お	王羲之-快雪時晴帖	0047	0005	☒
5	おうぎし-かんおうじょう	お	王羲之-干嘔帖	0047	0005	☒
6	おうぎし-かんこうとうようじょう	お	王羲之-桓公当陽帖	0047	0005	☒
7	おうぎし-かんせつじょう	お	王羲之-寒切帖	0047	0005	☒
8	おうぎし-こうじちゅうじょう	お	王羲之-孔侍中帖	0048	0005	☒
9	おうぎし	お	王羲之	0046	0005	☒
10	おうぎし-ししじょう	お	王羲之-此事帖	0048	0005	☒

「引得市」《中国書道文化辞典》搜尋「王羲之」

五、小結

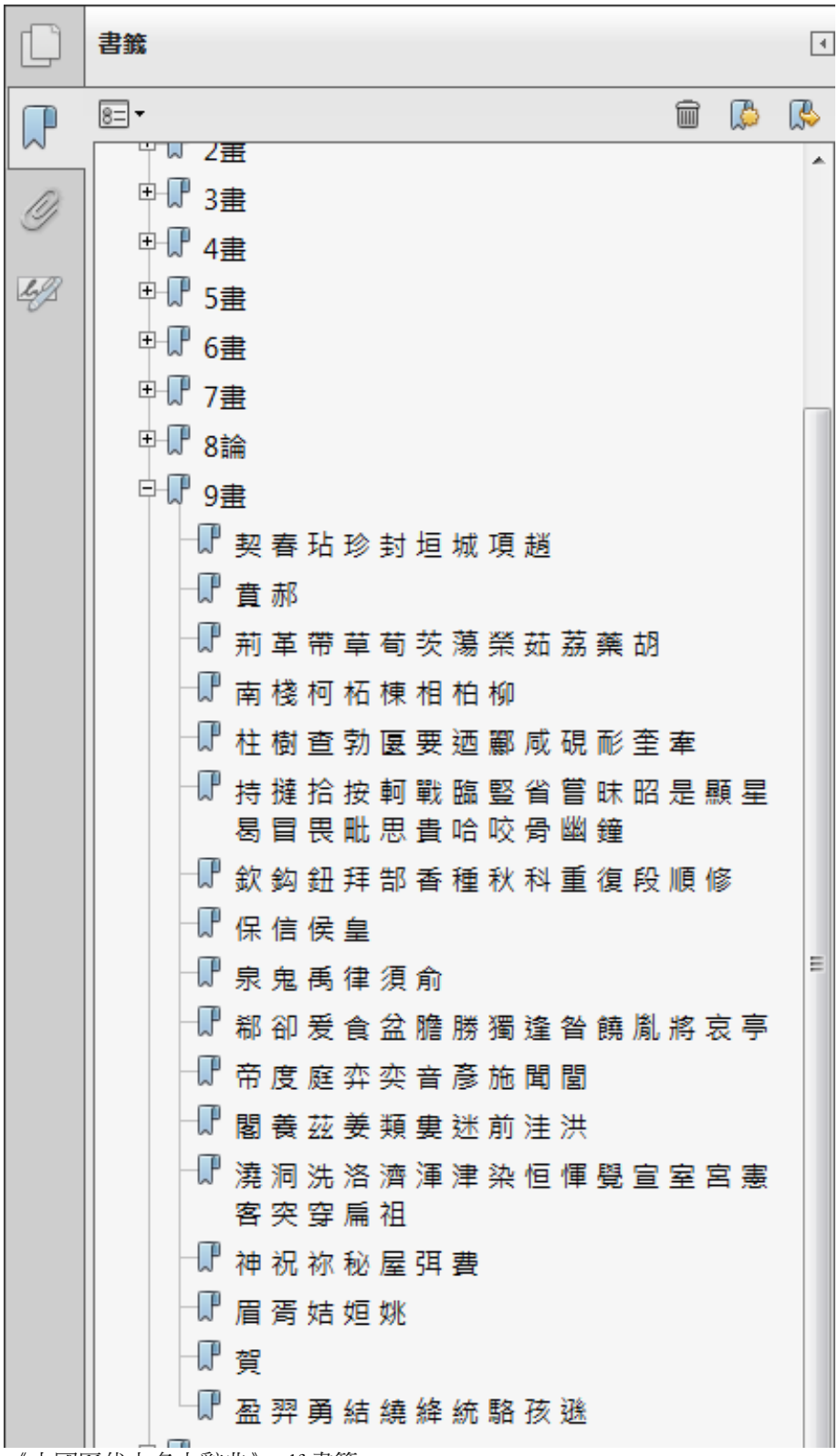
光陰似箭，又是一年之末，筆者只能以這樣的文章來回顧研究，檢視過去曾經作了什麼、還沒完成什麼，也常常有著原來曾經作了這些、那些資料的驚嘆。文獻資料掃瞄建檔，一次又一次文字的校正，這樣重複繁雜的作業，筆者稱為是「枯燥的有趣」；而網路遊戲中打王、打寶練功升等，則稱為是「有趣的枯燥」。兩者都是興趣，覺得性質上似乎也沒有什麼差別。

研究中面臨的困難或新發現的種種滋味也只有當事人感受最深。某一方面來說，研究可以不需要那麼的辛苦，只要找到方法或手段。但是，研究也不是那麼的輕鬆，隨意用文字剪剪貼貼就能交差的，沒有興趣的人還是不用急著來這個領域參一腳。現階段筆者只能透過文獻資源共享的方式，感謝師長與諸多朋友們長期以來精神與物質層面上的幫助與鼓勵。

文獻的數位化工作的範圍跨越多領域，期待更多人投入文獻處理的創新與研發，一同創造更好的學術研究環境。文獻與資訊應用的整合仍未成熟，拋磚引玉，祈請方家不吝指正。

1	筆畫數	人名	頁碼	目錄頁碼
54741	十九畫以上	灌何	2585	297
54742	十九畫以上	灌頂	2585	297
54743	十九畫以上	灌賢	2585	297
54744	十九畫以上	灌嬰	2585	297
54745	十九畫以上	娥阿	2585	297
54746	十九畫以上	驤兜	2585	297
54747	十九畫以上	齧缺	2585	297
54748	十九畫以上	霸突魯	2585	297
54749	十九畫以上	贛遂	2585	297
54750	十九畫以上	夔	2585	297
54751	十九畫以上	懿磷質班	2585	297
54752	十九畫以上	囊瓦	2585	297
54753	十九畫以上	囊加歹	2586	297
54754	十九畫以上	嚳拳	2586	297
54755	十九畫以上	嚳熊	2586	297
54756	十九畫以上	毯圖	2586	297
54757	十九畫以上	毯拜	2586	297
54758	十九畫以上	麟慶	2586	297
54759	十九畫以上	蠖蠖	2586	297
54760	十九畫以上	爨習	2586	297
54761	十九畫以上	爨云	2586	297
54762	十九畫以上	爨能	2586	297
54763	十九畫以上	爨琛	2586	297
54764	十九畫以上	爨瓚	2586	297
54765	十九畫以上	爨龍顏	2586	297
54766	十九畫以上	爨歸王	2586	297

《中國歷代人名大辭典》索引



《中國歷代人名大辭典》pdf 書籤

六、附錄

「引得市」收錄文獻的字例、缺字數量表

字典類（古文字）

書名	作者	本文起迄	檢字表起迄	字例數量	缺字數量	備註
秦印文字彙編	許雄志	1-309	311-322	1392	49	
簡牘帛書字典	陳建貢・許敏	1-958	978-987	2867	2	
馬王堆簡帛文字編	陳松長	1-611	1-20	3229	290	
楚文字編	李守奎	1-910	911-944	4894	2161	缺字待補
戰國秦漢簡帛古書通假字彙纂	白于藍	1-924	1011-1064	9493	1297	缺字待補
秦簡牘文字彙編	方勇	1-342	343-357	2088	109	缺字待補
古文字類編・增訂本	高明・涂白奎	1-1564	1-27	6512	332	缺字待補
古文字通假字典	王輝	1-806	886-923	7955	1051	缺字待補
睡虎地秦簡文字編	張守中	1-226	227-253	1807	59	

總字例：40237　缺字：5350

書名	作者	本文起迄	檢字表起迄	字例數量	缺字數量	備註
新金文編（上中下）	董蓮池	1-2232	1-50	2128	16	

總字例：2128　缺字：16

字典類（書道）

書名	作者	本文起迄	檢字表起迄	字例數量	缺字數量	備註
中國草書大字典	李志賢・蔡錦寶・張景春	1-1423	1-30	3744	2	
中國隸書大字典	范韌庵・李志賢・楊瑞昭・蔡錦寶	1-1292	1-28	3532	1	

總字例：7276　缺字：3

辭典類（書道）

書名	作者	本文起迄	檢字表起迄	辭例數量	缺字數量	備註
中國書道文化辭典	西林昭一	1-1030	1-75	7845	39	

總辭例：7845　缺字：39

「引得市」資料來源檢字表內容校正

書名	作者	檢字表錯誤位置與說明
秦印文字彙編	許雄志	檢字表・正文 316 頁，九畫「係」字 159 頁，應更正為 158 頁
秦簡牘文字彙編	方勇	檢字表 345 頁，七畫「谷」字，原 263 頁應更正為 264 頁。 檢字表 353 頁，十五畫「徵」字，原 20 頁應更正為 202 頁。
古文字類編・增訂本	高明・涂白奎	檢字表 5 頁右下右中，十四畫「□」（1268）與十五畫「□」（1268）重複，應刪除後者。 檢字表 5 頁右下，十五畫「壓」（845）與 6 頁左下，十六畫「壓」（845）重複，應刪除後者。 檢字表 14 頁右下，「锐」（1312）係簡體字，應改為「銳」。 檢字表 22 頁左上，「纫」（1026）係簡體字，應改為「紉」。
古文字通假字典	王輝	檢字表 904 頁，十一畫「廇」字（117）第二列與四列重複，後者應刪除。 檢字表 912 頁，十四畫第三列「□」字（338），檢字表 915 頁，十五畫第二列「□」字（338），後者應刪除。 檢字表 920 頁，十九畫第四列「𩰫」字（715），檢字表 922 頁，二十三畫第四列「𩰫」字（715）重複，後者應刪除。
睡虎地秦簡文字編	張守中	檢字表 233 頁，「定」字「七-七」誤。應為「七-六」。 檢字表 235 頁，「音」字「三-五」誤。應為「三-四」。 檢字表 237 頁，「桑」字「六-四」誤。應為「六-五」。 檢字表 244 頁，「寬」字「七-八」誤。應為「七-七」。 檢字表 246 頁，「綰」字誤。應為「掊」。 檢字表 250 頁，「旞」字「七-二」誤。應為「七-一」。 檢字表 251 頁，「 羈」字七-一一」錯誤。應為「七-一〇」。 檢字表 252 頁，「□」字「一一-二」錯誤。應為「一一-三」。 檢字表「為」字於 235 頁（9 畫）及 241 頁（12 畫）重複出現。
新金文編（上中下）	董蓮池	檢字表 50 頁，「尃」應改為「專」。
中國草書大字典	李志賢・蔡錦寶・張景春	筆畫檢字表 1 頁，第三行三畫「卅」（461）應更正為「川」（461）。 筆畫檢字表 1 頁，第五行四畫「川」（175）應更正為「卅」（175）。 筆畫檢字表 16 頁，第三行十二畫「童」（831）應更正為「童」（830）。
中國隸書大字典	范韌庵・李志賢・楊瑞昭・蔡錦寶	筆畫檢字表 11 頁，第二行十畫，補「□」（684）。 筆畫檢字表 19 頁，第一行十三畫「鄢」（1161）應更正為「鄢」（1162）。 筆畫檢字表 28 頁，第二行二十四畫，補「□」（866）。

參考資料

- 林慶彰，《學術資料的檢索與利用》，臺北：萬卷樓圖書股份有限公司，2003 年 3 月。
- 楊菁，〈中國哲學資料的檢索與利用〉，刊《學術資料的檢索與利用》，臺北：萬卷樓圖書股份有限公司，2003 年 3 月。
- 彭斐章・喬好勤・陳傳夫，《目錄學》，武漢：武漢大學出版社，2003 年 11 月。

筆者部落格文章（依時間順序）

- 〈搜尋「\d」與「[^\d]」〉（2010/5/27）http://blog.yam.com/ebag/article/28828908
- 〈《中国書道文化辞典》的索引製作和應用〉（2010/6/12）http://blog.yam.com /ebag/article/29050664
- 〈「辭典索引檔」在 Google 日文輸入法的應用〉（ 2011/5/9）http://blog.yam.com /ebag/article/38099901
- 〈Acrobat 搜尋功能的應用之一〉（2011/5/14）http://blog.yam.com/ebag/article/ 38227494
- 〈PDF 檔案跳頁連結 以「中國書畫文獻索引」為例〉（2011/5/18）http://blog.yam. com /ebag/ article/38339180
- 〈google 日文輸入法匯入中國書道文化辭典索引〉（ 2011/5/24）http://blog.yam. com /ebag/article/38509195
- 〈面對文獻〉（2011/7/3）http://blog.yam.com/ebag/article/39396872
- 〈「glyphwiki」線上製作字型（新帳號建立與登錄篇）初級〉（2011/7/7）http://blog. yam.com/ebag/article/39505788
- 〈「glyphwiki」線上製作字型（製作字型建立群組及使用字型篇）高級〉（2011/7/7）http://blog.yam.com/ebag/article/39506051
- 〈檔名修改應用示範-「甲骨文合集」圖檔為例〉（2011/12/13）http://blog.yam.com/ ebag/article/45416338
- 〈PDF 書籤製作示範——以《宣和書譜》〉為例（2012/5/24）http://blog.yam.com/ ebag/article/50236986
- 〈文獻處理 OCR 的應用——以《中國書畫文獻索引》下冊 2924 頁為例〉

（2012/5/25）http://blog.yam.com/ebag/article/50275274

- 〈《楚文字編》筆畫檢字表的索引數位化〉（2012/6/26）http://blog.yam.com/ebag/article/51399795
- 〈漢字速查軟體的使用與介紹〉（2012/6/29）http://blog.yam.com/ebag/article/51480611
- 〈研究用書籍紙本數位化時間試算表 1.0〉（2012/8/5）http://blog.yam.com/ebag/article/52617385