珍藏文獻整理與資訊科技應用研討會

談古籍檢索的字形問題

中央研究院資訊科學研究所

謝清俊

國家圖書館主辦

一九九七年四月廿一日

談古籍檢索的字形問題

中央研究院資訊科學研究所
謝清俊

一、前言

珍貴的古籍很不方便流傳利用，保存起來也要費許多心血。因此，將之數位化以保存、應用是很自然的趨勢。古籍的數位化，當然是越能保留原有的信息越好；其中對原書字形的忠實保留，是大家所盼望的。有些重刊重印的古籍，為了方便一般民眾閱讀，把原書的字形改成現在所用的字形，雖然一般人比較容易看得懂，卻失去了許多版本及考據上的信息，對作研究的人來說，並不恰當。可是，若以原字形呈現，在推廣上又確有不便。

為解此兩難之境，將古籍分兩種版本刊行，是目前常用的方法。這麼做，對於人也許問題算是化解了；可是，對電腦如何呢？也許有些人認為用影像存古籍字形與版面之真，佐以打字的現代字形版相互對映，應可解決問題。 這樣的構想是不錯的；尤其今日之電腦儲存空間夠大，又能做多媒體，這麼做似乎技術上並無困難。然而，稍深思之，就會發現問題不是如此單純；在這種情形下怎麼做檢索呢？如果只能用現代字形檢索，還是有所不便。最好是古今字形均可檢索、而且還能做比對、計數、以及作種種計算那是最好、最理想的。為了要將數位古籍做到這種程度，就不能不對字形多加了解。這就是撰寫本文的動機。

中文電腦發展之初，是每個字只有一個字形。這也許是因為當年電腦的能力有限，處理中文字形又很複雜，而且在應用上對字形也要求不高的緣故。現在，為處理珍貴古籍而提高了對字形處理的需求，探討字形問題正是時機。

二、字形問題概說

自古以來，漢字就不是一字一形的。雖然在歷史上每次整理文字時，都希望做到一字一形，但是實際上並沒有做到。歷代的字書都蒐集各體字形就是一項最好的證明。再者，從生活環境上說，任何時間的人民是不可能不接觸到古

字形的，這是因為社會現象本就是歷史的沈積，生活本離不了歷史。所以，從使用者的需求來看，中文電腦應有能力處理古今字形才是。縱然今天做不到，以後終將做得到。

然而，時下的電腦中卻是只存一套標準碼；碼中定有標準的字形。這個字形是經過整理和規範後的字形，並未涵蓋所有古今字形。以此字形處理珍藏古籍，有無問題是可想而知的。我們並不反對政府對當前的文字約以規範，替未來著想，定標準字形是有意義、有長遠的好處的。可是，標準字形只能顧到目前和未來，顧不了以往。這對處理珍藏古籍不只沒有幫助，反倒成為跼腳石。所以，對使用者（一般民眾）而言，電腦中不只應有規範的字形，還要有古時字形，這樣才能普惠眾生，解決日常生活中遇到的種種古今字形的問題。

從文字學史來看，漢字一字多形的情形比比皆是。成書在公元 770 年左右的《干祿字書》，顏元孫在自序中便將字形分為「通」、「俗」、「正」三者：

『所謂俗者，例皆淺近，惟籍帳、文案、券契、藥方，非涉雅言，用亦無爽……

所謂通者，相承久遠，可以施表、奏、箋、尺牘、判狀，固免詆訶。

所謂正者，　有憑據，可以施著述、文章、對策、碑碣，將為允當。』

這段話說明了字形的『約定俗成』生態。這是文字生命的源頭活水。如果禁絕了這約定俗成的生命，那麼文字或字形就固定死了，以後再無改進、適應的餘地，久而久之終將走上絕路。所以，字形有變化是常態，是正常的；僵化的強制推行標準，反是病態；電腦中不能處理一字多形，也是病態。

其實，漢字不只有一字多形的性質，還有幾個字共用一個字形的情形，文字學中稱為同形字。六書中的『假借』字，就是這種情形最佳的例子。依文字學，字是以字的音義來判別的，換句話說，音義相同者，雖字形不同，應判為同一個字。據此界說，不只可以判別同形字，也可判別異體字（即一字多形者）。可惜的是，現在的交換碼中字和字形不分，混淆為用，使得交換碼沒有判別異體字的能力。這也是推動數位珍藏古籍時，先要解決的問題。至於同形字，交換碼雖目前亦無能力處理，但電腦在還沒有能力處理字義之前，倒顯得並不是那麼急迫的要解決，暫緩無妨。

三、時下的字形與文獻檢索

一般而言，目前中文文獻檢索的方式都是利用字形，電腦只會比對字形，不會比對字！這不能不說是非常粗糙的、蠻不講理的檢索方式。電腦不會判斷異體字，這是眾所周知的，所以，若要檢索出可能的異體字，就全靠使用者了，然而，人的腦子卻又是極不容易詳舉所有的異體字。其實，人們要檢索的，大都是和文獻的內容有關的事物；所以，最好是能依句子、片語或詞的語義來檢索，然而目前只能用字串來檢索——說得更正確些應該是字形串。固然，句子、片語和詞皆由字構成，有其語義關連的方便處，但反過來說，若是電腦的文字知識不夠，像是不會認識異體字，那麼，句子、片語和詞也要受到連帶拖累。所以，解決字形問題，即在電腦中要明白的界定字和字形，使之能夠分辨處理，並將異體字的關係清楚地表達在電腦中，是做好檢索和文獻處理最基本，也是最重要的工作。

目前，電腦裏有的字型，是漢隸以後的印刷字體；從時間上來看，只適合處理漢以後文獻，從其形看來，不甚適合處理人工書寫的文獻。若從版本上來看，各時代的字形，各地的刻工，均有些許字形上的變化，這些都會造成字形處理的難題。譬如，敦煌文獻以及出土的歷代碑文，這些字形都不是目前電腦能夠應付了的。甚至四庫全書，雖是官方抄本，也難免有字形變化。如果要徹底解決字形變化的問題，勢必要建立一個時空座標，把不同時代、不同地區的字形變化安置在這個座標上，才能提供足夠的字形信息，讓電腦能順利地處理不同時空的珍藏文獻。這時空座標，也可不限於中國，凡漢字家族成員，均可概括在內。

要建立這樣一個座標的漢字體系，自非一蹴可幾。然而在珍藏文獻的數位化過程中，如果能有良好的軟體工具，把一本一本書的字形蒐集整理，長此以往的累積就是建立這個體系的一個可行之道。這麼做，不只可以促進字形的共享和利用，也將有益於數位文獻之流通。這是將珍藏文獻數位化過程中，大家可以考慮來施行的事。

現在用電腦處理珍藏文獻是很痛苦的事，市面上的字型和古代字形有很大的差異；如現在叫仿宋體的字形事實上已不是傳統仿宋體了，尤其是標楷體公佈後，有些廠商竟據以修正仿宋字形，簡直是把設計字體的原則和體例都攪亂了。更有甚者，有些字改、有些又沒改，真一個亂字了得！現在要找一個字集和傳統相近的，選來選去似乎只有『明體』比較接近。然而，明體卻是日

本的設計（不是明朝的字體），國內字形之亂一致於此，豈不哀哉？

四、字形變化的制式表達

如果只收錄漢以後的楷和仿宋兩種印刷字體，國字整理小組就曾集到七萬四千餘字形。若算上漢以前的各種字體、再加上書法的變化，那麼，漢字字形的數目實是不勝計數。處此情境，計算機要如何處理呢？字碼有那麼大的空間安放嗎？每個使用者都要背負這幾千年來所有漢字字形的沈重包袱嗎？顯然不能這樣，所以，必須設法以簡馭繁，有系統地來表達字形信息。

要解決這個問題，就必須稍深入地了解漢字的構形。將構形的規則以制式生成語法的方式表達出來，是解決的方法之一，譬如字形的字根式[1] 就是典型的例子。然而一個字的字根式可能會太長了，使用不便，能不能把一個字的構形只用兩個部件（至多三個，佔極少數），以及一個結構符號表示呢？答案是肯定的，目前的研究顯示：只要增加約八百個部件，配合一個約一千三百字的小字集，就可以用上述的方法描述約一萬二千個常用的字形。當然，此系統可以描述的字形實不止一萬二千，但究竟多少，尚待查驗。據估計，五千常用字加上約一千五百個部件，應至少可以描述五萬以上的字形。[2] [3] [4] 根據這種方法描述字形的「字形資料庫」雛型業已設計完成[5] 有些製作電子佛典的單位已經用它在解決佛典缺字問題。

如果做文獻檢索時，能接受這樣的字形描述，那麼就可解決一大部份的字形變異以及異體字相互參照檢索的問題。做珍藏文獻處理時，不妨一試。

五、其他相關問題

有一些異體字是須在特定的語意情境下才會發生的，例如：「淳樸」之於「純樸」，「十元」之於「拾元」，「梅雨」之於「霉雨」等。在語言學中，稱這些為異體詞。據《現代漢語詞典》及《新華字典》的收集，這類異體詞約有一千一百對左右。這異體詞產生的檢索問題，目前的系統雖無解，但解起來並不困難，只要將異體詞列表存入電腦中，便可寫簡單的程式處理。然而，若是慮及歷代的珍藏文獻，異體詞的數目就可能不止於此，這是處理古文獻時應留心收集的重要資料。其實，若電腦中有古文獻的索引典，異體詞是可以歸併入同義詞來處理的。

此外，古籍的標點方式和白話文不同，珍藏的古籍中可能有複雜的句讀符號（甚至包括幾種彩色和形狀），也可能有特殊的標示符號（如大正藏中的句法標示）。若是不能了解這些符號的用法和規則，則亦將造成檢索上的障礙或錯誤。所以，如何將這些格式用大家統一的制式格式表達出來，就成為解決這類檢索問題的前題條件。一個方法是用 SGML 或 HTML 將其構成（如 DTD）及各種符號的標示（ tagging ）描述出來，讓大家遵循採用。順便一提的是古籍的版面格式問題，如眉批，校勘，注疏等，這些體例所產出的問題，亦可如上用 SGML 或 HTML 來設法解決[6] [7] 。

六、結語

校讀古書時，文字學是必備的知識，用電腦處理古籍亦不例外，目前電腦中與文字學相關的基本資料和知識貧乏，幾等於無，處理古籍時之窘境不說也明白。處此情境，將文字學的知識有系統的納入電腦中，是唯一的解決之道。

依此前題推論，要解決古籍檢索時的字形問題，亦必須求助於文字學，尤其是構形的部份。本文討論的文字學部份其實都甚膚淺，若是真正要做好利用電腦來協助，漢學研究的工作，徹底治本之道是建立一個完整的文字學資料庫或知識庫在電腦中。這是一個浩大的文字工程，值得有心人一試。

參考文獻

【１】 謝清俊、黃永文、林樹,《中文字根之分析》,交大學刊,第六卷第一期, 1973 年 2 月

【２】 謝清俊,〈電子古籍中的缺字問題〉,第一屆中國文字學會學術研討會, 天津,1996 年 8 月

【３】 謝清俊,〈漢字的字形與編碼〉,漢字字碼與資料庫國際研討會,京都及東京,1996 年 10 月

【４】 謝清俊,〈 A Descriptive Method for Re-engineering Hanzi Information Interchange Codes 〉漢字字碼與資料庫國際研討會, 京都及東京, 1996 年 10 月

【５】 謝清俊、莊德明、張翠玲、許婉蓉,〈中文字形資料庫的設計與應用〉, 第六屆中國文字學會全國學術研討會,台中,1995 年 4 月

【６】 謝清俊、陳昭珍,〈談古籍之電子版本〉海峽兩岸中國古籍整理現代化技術研討會, 北京,1993 年 10 月

【７】 謝清俊、莊德明,〈古籍校讀工具『中文文獻處理系統』的設計〉中國古籍整理研究出版現代化國際會議,北京,1995 年 7 月