漢字數位化的困境及因應：談如何建立漢字構形資料庫

莊德明

中央研究院資訊科學研究所研究助技師

壹、漢字的形體及字數

這些年來電腦科技蓬勃發展，數位化的呼聲不絕於耳，很少有一門學科能置身事外而不受影響。然而各門學科數位化的進展快慢不同，漢學在這方面是比較落後的，尤其是文字學。

從殷商的甲骨文算起，漢字的使用已達三千四百年之久。漢字從歷史演變出發，可以分成兩大階段。這兩個階段以秦代的小篆作為分界：自甲骨文到秦代的小篆，通稱古漢字；自秦漢隸書以後，通稱今漢字。當代正在使用的漢字，稱作現代漢字，在形制上也屬於今漢字。[1]古漢字包括甲骨文、金文、戰國文字和小篆，今漢字包括隸書、草書、行書和楷書。今漢字與古漢字相比，由於形體變化太大，漢字的形義關係已很不明顯，甚至被完全隱沒。所以要想瞭解某個字的構形寓義，必須找到它的古文字形體[2]。

例如「員」字本為方圓的「圓」的本字。表一列出「員」的古文字形體，其中 1 為甲骨文，2-3 為金文，4 為說文籀文，5 為說文小篆，6 為楚系簡帛文字。「員」的甲骨文和金文都从鼎、从〇，〇亦聲。林義光《文源》:「〇，鼎口也，鼎口，圓象。」《說文》:「員，物數也。从貝，口聲。鼘，籀文从鼎。」說文籀文仍然保留从鼎、从〇的寫法，小篆「鼎」已訛為「貝」，字義也由「圓」而借用為「物的數量」，於是後人又於「員」外加「口」作「圓」以還其原。[3]

表一、「員」的古文字形體

					
1.合集 10978	2.員父尊	3.或方鼎	4.說文籀文	5.說文	6. 秦 1.2

現代漢字的字數已很難統計，而出現在各種典籍裡的，以及出現在甲骨、鐘鼎、簡帛、玉石等材料上面的古漢字個數就更加難以統計

[1]見《漢字漢語基礎》頁 7。

[2]見《漢字漢語基礎》頁 63。

[3]見《金文詁林》卷 6，頁 335。

了。表二從歷代一些具影響的字書中列出一個大略的估計。[4]從表二也可看出，古漢字最大的特點是異體字（重文）多，例如《金文編》收錄的「尊」字有 252 個形體，而「寶」字則有 273 個形體。

表二、古漢字個數的大略估計

書名	成書時間(西元)	作者	字數	備註
《說文解字》	100	許慎	10516	小篆字 9353 個，重文 1163 個。
《甲骨文編》	1965	中國科學院考古研究所	4672	正編收單字 1723 個，附錄收單字 2949 個。
《金文編》	1985	容庚	24261	金文字頭 2420 個，重文 19357 個；附錄收字 1352 個，重文 1132 個。
《先秦貨幣文編》	1983	商承祚等	8215	正編收錄 313 個字，同文異體字 5726 個；合文 63 個，同文異體字 232 個；附錄 534 個，同文異體字 1347 個。
《漢印文字徵》	1978	羅福頤	10239	收錄漢魏官、私印文字 2646 個，重文 7432 個；附錄收字 143 個，重文 18 個。
《楚系簡帛文字編》	1995	滕壬生	19250	

表三、歷代部分字書收字情況

書名	成書時間(年代)	作者	收字頭數	備註
《說文解字》	西元 100 年(東漢)	許慎	9353	加上重文 1163 字，共計收字頭 10516。
《玉編》	西元 543 年(南朝梁)	顧野王	22726	
《廣韻》	西元 1011 年(宋)	陳彭年等	26194	
《集韻》	西元 1067 年(宋)	丁度等	53525	
《字彙》	西元 1615 年(明)	梅膺祚等	33179	
《康熙字典》	西元 1716 年(清)	張玉書等	47035	加上古文字字頭共 49030，其中重收字頭 81 個。
《漢語大字典》	西元 1986 年—1990 年	徐中舒等	54678	

現行漢字是由各個歷史時期的漢字發展積澱而成。它的總體數量、單字筆畫、結構、讀音以及體勢都在不斷變化。從現存的主要字

[4]見周曉文〈建立「信息交換用古漢字編碼字符集」的必要性及可行性〉，古漢字數位編碼暨現代化應用研討會，台北，2005 年 10 月。

書來看，漢字的總數在不斷地增多。5表三列出歷代部分字書收字情況6。這些新增字有相當數量是異體字，它們是由兩個原因造成的：一是共時的個人書寫漢字因隨意性而產生的變異；二是前代不同形制的漢字積澱到後代而產生的差異。例如《漢語大字典》收錄了 3 個「員」的異體字：「  」、「  」、「  」，其中「」為說文籀文「」的楷化，「」出自袁博殘碑，「」為簡化字。

貳、漢字數位化的困境

1942 年，第一部電子計算機誕生後不久，拉丁文字的計算機處理便開始起步，並獲得迅速發展。1960、1970 年代，西方實現文字初級自動化處理的時候，漢字承受著機械化與自動化的雙重重壓。直至 1981 年，IBM推出第一部個人電腦後的十來年間，電腦的漢字處理技術才實現了全面突破。然而今天漢字與電腦的適應，應該說只是基本的，或是初步的，它還不完全，不完善，不完美。許壽椿認為突出的問題和難題至少有：[7]

1.缺字困擾普遍存在。現今任何一個系統都不能表達全部漢字，都有缺字問題。

2.排序混亂。現有的各漢字編碼方案均沒有認真處理好排序問題。

3.多種中文字碼並存，造成傳輸、交流的障礙。

4.輸入法的規範、優選仍無成效。高效率、易學用、與漢字基礎教育相協調的輸入法久喚不出。

5.字量龐大使字庫的設計、生成、存貯、調度多有困難。

以下分別說明這些問題的成因及現況：

一、缺字。缺字問題由來已久，是因為目前電腦中文字碼字形不足而引起的，在處理古文獻時尤其嚴重。例如 Big5 只收繁體字「員」，簡體字「」即為缺字；GB2312 只收簡體字「」，繁體字「員」即為缺字，詳見表四。由表四也可看出，GB2312 收錄的漢字只有 6763 個，實在不夠，大陸於是又制定 GBK，收字 21003 個，GBK 已包含 Big5 的繁體字。

中文字碼的後續發展，Unicode 已成為關注的焦點。自從微軟視窗 2000 開始支援 Unicode 2.0，台灣及大陸的中文系統已可同時使用

[5]見《漢字漢語基礎》頁 65。

[6]見《漢語大字典》冊 8，頁 5460。

[7]見許壽椿〈網絡時代的漢字全面解決方案和漢字本體研究〉，'99 漢字應用與傳播國際學術研討會，北京，1999 年 6 月。

繁體和簡體字，而且內碼完全相同，不同的只是使用者界面選用繁體或簡體的差別而已。

表四、中文字碼 GB2312、Big5、GBK 及 Unicode 的比較

字碼	版本	發表年	字數	編碼空間	收字示例	備註
GB2312		1981	6763	8836	员	大陸第一個中文字碼
Big5		1984	13051	19782	員	視窗 95/98(繁體)採用為內碼
GBK		1995	21003	23940	員 员 貟	視窗 95/98(簡體)採用為內碼
Unicode	1.1	1993	20902	2147418112 (21 億多)	員 员 貟	視窗 2000/XP(正繁體版)採用為內碼，目前已支援 2.1 版。
	3.0	2000	27484		員 员 貟	
	3.1	2001	70194		員 员 貟 鼎	

Unicode 3.1 的後續版本Unicode 3.2 及Unicode 4.0 分別於 2002 年及 2003 年發表，這兩個版本收錄的漢字並未增加，但這不表示漢字已收齊了。Unicode 3.1 的 70194 個漢字絕大多數來自於《漢語大字典》,《漢語大字典》所收字數為 54678 個，然而教育部國語會編輯的《異體字字典》[8]所收字數竟高達 106230 個，比《漢語大字典》多了 51552 字。

表五、「鵝」及其異體字

鵝 1.	鵞 2.簡化字	鵝 3.集韻	鵞 4.說文	鵞 5.字彙補	鵞 6.玉編
鳱 7.中華字海	鵞 8.宋元以來俗字譜	鵞 9.佛教難字字典	鵞 10.佛教難字字典	鵞 11.佛教難字字典	鵞 12.玉編

表五列出《異體字字典》收錄的「鵝」的異體字，其中 1-6 可見於《漢語大字典》，其中「鵞」、「鵞」、「鵞」和「鵝」的差別在於部件「我」、「鳥」的相對位置不同，而「鵞」和「鵝」的差別在於部件「鳥」的異寫；7-12 為《異體字字典》新增的異體字，除了「鳱」以外，其他五個也都是部件「鳥」異寫而形成的異體字。由「鵝」的例子可看出，漢字存在著大量可類推的異體字，簡化字只不過是其中一種，這些異體字也讓缺字問題變得更複雜。

Unicode 3.1 雖然收錄了 70194 個漢字，但仍未涉及古漢字。古漢字由於具有十分典型的表意文字特點，以及所含有的文化內涵，被不斷引進文化教育領域，迅速走向普及，成為世界各國瞭解中國文化的一個重要窗口，因此古文字進入電腦勢在必行。[9]Unicode由於編碼空

[8]見《異體字字典》網路版說明(2004 年第五版)，網址http://140.111.1.40/start.htm
[9]見王寧〈漢字研究與資訊科學技術的結合〉，漢字與全球化國際學術研討會，台北 ，2005 年 1 月。

間極大，目前相關單位也正積極展開古漢字的編碼工作。[10]古漢字的編碼工程繁複，難以在短時間內速成。目前先試著以材料、時間、空間將古漢字分成甲骨文、金文、玉石文字、簡帛文字、小篆及其他六類，二十二個區塊的區塊體系進行編碼，詳見表六。[11]

表六、古漢字編碼分類及區塊

材料	時間	空間
甲骨文	商甲骨文(1)	
	西周甲骨文(2)	
金文	商金文(3)	
	西周金文(4)	
	春秋金文(5)	
	戰國金文	戰國楚系金文(6)
		戰國晉系金文(7)
		戰國齊系金文(8)
		戰國燕系金文(9)
		戰國秦系金文(10)
玉石文字	商玉石文字(11)	
	西周玉石文字(12)	
	春秋玉石文字(13)	
	戰國玉石文字	戰國楚系玉石文字(14)
		戰國晉系玉石文字(15)
		戰國齊系玉石文字(16)
		戰國燕系玉石文字(17)
		戰國秦系玉石文字(18)
簡帛文字		戰國楚系簡帛文字(19)
		戰國秦系簡帛文字(20)
小篆(21)		
其他(22)	古陶文字、古璽文字、貨幣文字、漆器文字、兵器文字、雜器文字	

儘管 Unicode 的編碼空間極大，但是缺字問題的癥結並不只在編碼空間的大小，而在於漢字本身是個開放字集，存在著大量可類推的異體字，古漢字重文眾多，字數難以作固定的限量。擴大交換碼收錄的字形，對於缺字問題，雖有紓解的作用，但並不能完全解決問題。

二、排序混亂。中文字碼排序混亂的根本原因，在於現行的漢字字集是由幾個子集合組成，雖然子集合有排序，但是不同子集合的字卻沒有排序。表七列出 GB2312、Big5 及 Unicode 的字集子集合及子集合的排序方式。表八列出百家姓中的「芮羿儲靳汲邴糜松」八個姓

[10]見中推會於 2005 年 10 月召開的「古漢字數位編碼暨現代化應用研討會」

[11]見許學仁〈古文字資料庫與古文字數位編碼〉，古漢字數位編碼暨現代化應用研討會，台北，2005 年 10 月。

氏在 Big5 及 Unicode 的排序，若按筆畫排序，「芮」應在「糜」前，但是 Big5 的排序卻是「糜」在「芮」前，這是因為「糜」歸在常用字，而「芮」為次常用字；而「芮」、「糜」同時歸在 Unicode 的 CJK 認同表意文字區，所以排序仍維持先部首後筆畫。但 Unicode 仍然沒有解決排序的問題，因為擴充 A 區的 6582 個字(內碼 3400-4DFF)，這些字不論部首、筆畫，還是會排在原先的 20902 個字(內碼 4E00-9FFF)之前。

表七、中文字碼 GB2312、Big5 及 Unicode 的排序

字碼	字數	字集子集合	子集合字數	內碼	排序
GB2312	6763	第一級漢字	3755	B0A1-D7F9	按漢語拼音字母/筆形排序
		第二級漢字	3008	D8A1-F7FE	按部首/筆畫排序
Big5	13053	常用字	5401	A440-C67E	按筆畫/部首排序
		次常用字	7652	C940-F9D5	按筆畫/部首排序
Unicode 3.1	70194	CJK 認同表意文字區	20902	4E00-9FFF	按部首/筆畫排序
		CJK 認同表意文字擴充 A 區	6582	3400-4DFF	按部首/筆畫排序
		CJK 認同表意文字擴充 B 區	42710	20000-2A6D6	按部首/筆畫排序

表八、百家姓「芮羿儲靳汲邴糜松」八個姓氏的排序

姓	芮	羿	儲	靳	汲	邴	糜	松
部首	艸	羽	人	革	水	邑	米	木
筆畫	8	9	17	13	7	8	17	8
Big5	CDBA	ACFD	C078	E0DA	A856	CDD4	C153	AA51
Unicode	82AE	7FBF	5132	9773	6C72	90B4	7CDC	677E
先部首後筆畫	6	5	1	8	3	7	4	2
先筆畫後部首	3	5	7	6	1	4	8	2
Big5 排序	6	3	4	8	1	7	5	2
Unicode 排序	6	5	1	8	3	7	4	2

三、多種中文字碼並存，造成傳輸、交流的障礙。在網際網路盛行之今日，瀏覽不同內碼的中文網站、網頁，常常看到一堆不知所云的亂碼，收發中文電子郵件更是如此。單單在 1991 年前，世界各國發表的漢字交換碼就有 14 個。[12]

[12]見謝清俊〈談中國文字在電腦中的表達〉，「中國文字的未來」學術研討會，台北，1991 年 6 月。

表九、CCCII 及 CNS 11643 的最新版本及應用現況

字碼	發表年	字數	編碼空間	應用現況
CCCII	1989	75684	830584	國內外圖書館
全字庫 4.0	2002	76067	141376	戶役政系統

台灣除了Big5 及Unicode外，仍在使用的中文字碼，至少還有CCCII及CNS 11643（全字庫）[13]。表九列出CCCII及全字庫最新版本的相關資訊。

CCCII 由於主要使用者僅為圖書館界，使用層面不廣，圖書館界擬改用 Unicode，但是 CCCII 收錄的字數高達 75684，Unicode 仍無法全面取代。隨著 Unicode 不斷的擴充及普及，CCCII 或可改用 Unicode，然而 CNS 11643 已為國家標準，仍會持續擴編，並與國際標準接軌。GB2312 以及後續的 GB13000、GB18030 也是一樣。

四、高效率、易學用的輸入法久喚不出。中文輸入法可分成手寫、語音及鍵盤輸入三類，雖然手寫和語音辨識技術近年已有進步，但仍然有不少缺點和限制，以鍵盤為基礎的中文輸入法仍是目前的主流。鍵盤輸入法又可分成字音取碼和字形取碼兩大類。其中，字形取碼又可細分成筆順拆字及部件拆字兩類。[14]表十列出這三類鍵盤中文輸入法的比較。由表十可看出，字音取碼及筆順拆字的輸入法容易學習，但是速度較慢；部件拆字雖然輸入速度快，但是較難學習。

表十、三類鍵盤中文輸入法的比較

輸入法	字音取碼	筆順拆字	部件拆字
例子	注音輸入法、漢語拼音輸入法	字原、十二鍵	倉頡、快碼
輔助字形的數目			一般由大約 100 個至 400 個不等
熟練者的速度	每分鐘 10 至 35 字		每分鐘 20 至 60 字，個別用戶可達每分鐘百餘字。
優點	較易學習。		重碼字少，不需怎樣選字，輸入速度快。
缺點	重碼字較多，速度較慢。		較難學習，牢記輔助字形的階段最痛苦。若不經常使用，很容易生疏或忘記。
適合對象	除了職業輸入員外，無所謂不適合。		特別適合職業上需要大量輸入中文的用戶。

[13]見全字庫網站，網址http://www.cns11643.gov.tw

[14]見薛偉傑〈鍵盤輸入三大主流〉、〈輸入法如此多嬌引無數英雄競折腰〉，網址 http://input.foruto.com/introduce/index.html

自 1979 年朱邦復創出倉頡輸入法後，兩岸三地推出的中文輸入法已有上千種。如何發展一個高效率、易學用、與漢字基礎教育相協調的輸入法，已成為中國人一個歷久不衰的研究課題。

五、字庫龐大。中文字型可分成點陣字(Bitmapped Font)及伸縮字(Scalable Font)兩大類[15]。然而，由於中文字型的字數非常的多，因此為因應市場的需求，又產生若干字型描述技術[16]，總括而論，伸縮字又可分類為向量字、純外框式描邊字及組字式描邊字，詳見表十一。由於個人電腦的運算速度愈來愈快，儲存空間也愈來愈大，外形美觀的描邊字早已成為個人電腦使用者的優先選擇；但是在多媒體播放機、PDA、手機的螢幕顯示，則仍以點陣字居多。

表十一、中文字型的分類與比較

<table>
<tr><th colspan="4">字型分類</th><th>比較</th></tr>
<tr><td colspan="4">點陣字</td><td>將字形的資料以一點一點的方式儲存，放大時會有鋸齒狀的失真，並且太佔儲存空間，但是產生速度快。</td></tr>
<tr><td rowspan="5">伸縮字</td><td colspan="3">向量字</td><td>利用向量的特性，以直線來描繪文字的外緣，使得字形在放大之後不會產生鋸齒狀的邊，雖較點陣字美觀，只是小字會糊在一起，大字如果放的太大時，也會有鋸齒。</td></tr>
<tr><td rowspan="4">描邊字（外框字）</td><td colspan="2">純外框式描邊</td><td rowspan="4">1. 利用數學公式來記錄文字的外框，以求文字在放大之後仍能保持美觀的外表；並且利用微調技術以免小字糊在一起。字形所佔的空間比點陣字小，但是產生速度較慢。PostScript 和 TrueType 字形都屬於純外框式描邊。
2. 一套 Big5 碼純外框描邊字，約佔 3-10M。組字式字形佔用的空間比純外框式少很多，有些號稱 1M 可裝 15 套字，其他的則一套字從 200K 到 1.5M 都有。
3. 字根組字則由於同一個字根用在不同字形時，無法作細微調整，以致字形品質較差。</td></tr>
<tr><td rowspan="3">組字式描邊</td><td>筆畫組字</td></tr>
<tr><td>字根組字</td></tr>
<tr><td>圖元組字</td></tr>
</table>

參、漢字構形資料庫的範疇

漢字有形、音、義三個要素。廣義的文字學包括形、音、義三者。後來由於研究「音」的部份，獨立為「音韻學」；研究 「義」的部份，

[15]見游振昌〈字裡乾坤〉，第三波，1994 年 5 月，頁 50。
[16]見吳福生〈字形面面觀〉，0 與 1 科技． BYTE 中文版，155 期，1994 年 3 月號，頁 242。

獨立為「訓詁學」；於是文字學便成了狹義的文字形體研究。[17]從第二節漢字數位化所遭遇的困境可看出，無論是缺字、排序混亂、輸入法或字庫的問題，主要的癥結都在於目前電腦中的漢字知識過於貧乏，尤其是字形。本文對於漢字構形資料庫的探討，也是以字和字形為主。

為了字碼的排序及索引，再加上字碼能夠承載的漢字知識相當有限，在制定中文字碼的同時，常需同時整理漢字屬性或建立文字資料庫。例如表十二為CCCII的中國文字資料庫索引[18]，主要為部首索引、字音索引及字碼交互索引。

表十二、CCCII 中國文字資料庫索引

索引	項目
部首	部首、部首外筆畫數、筆順、總筆畫數
字音	國語注音、韋氏音標、劉氏音標、耶魯音標、漢語拼音、國語羅馬字
字碼	四角號碼、三角號碼、林樹頻率、數據通訊碼、通用漢字交換碼、五大碼、天龍碼、CDC 碼、凌群碼、首次尾輸入碼、神通碼、IBM 5550 碼、HP 碼、昭和碼

CCCII同時可將漢字的正異體字關係用字碼位置表示出來，如表十三的CCCII異體字表。CCCII的簡體字字碼比正體字字碼在第一位元組(B_3)碼值多 6，而其餘第二(B_2)、三位元組(B_1)的碼值完全相同。其他的異體字也和正體字有位置關係，即異體字第一位元組碼值比正體字的第一位元組碼值多 6 的倍數。這是因為正體字佔 6 字面，而異體字則放在以後的各字面，並要和它對應的正體字有上述的位置關係。因此，CCCII能提供極方便的異體字轉換與檢索功能。

表十三、CCCII 異體字表

	B_2	5	5	5	5	5	5	5	5	5	5
		1	1	1	1	1	1	1	2	2	2
	B_1	7	6	6	6	7	7	7	3	3	2
B_3		8	B	D	E	7	3	4	1	8	D
2	3	鶻	鶇	鷀	鷁	鶪	鷇	鷌	鷖	鷕	鷙
2	9	鹘	鸫	鹚	鹢	䴗	[illegible]	䳍	鹥	[illegible]	鸷
2	F			鶿	艗	[illegible]				[illegible]	
3	5				榏						
3	B				鶂						
4	1				[illegible]						
4	7				鷊						

[17]見中國國學小學館，網址http://www.superlogos.com.tw/main9/index_02.htm

[18]見《中國文字資料庫》，國字整理小組，1985 年 5 月

表十四為全字庫文字的屬性規範。[19]全字庫無法用字碼位置表示正異體字關係，因此另附簡繁對照。表十五為Unicode 4.1 的漢字屬性分類。[20] Unicode的漢字由於匯集自中、日、韓、越，所以也加入各國的字典索引及字碼等屬性。

表十四、全字庫文字的屬性規範

屬性	項目	備註
排序	Unicode、Big5(含 Big5-E)、EUC、稅務碼、財稅碼、電信碼、筆畫數、倉頡輸入碼、注音、部首、台語音	1-15 字面
字義	注音、字義	1-2 字面
拼音	注音、漢語拼音、注音第二式、耶魯、韋氏、劉式	
簡繁對照		

表十五、Unihan Properties by Category

Dictionary Indices	kCowles, kDaeJaweon, kFennIndex, kGSR, kHanYu, kHanyuPinlu, kIRGDaeJaweon, kIRGDaiKanwaZiten, kIRGHanyuDaZidian, kIRGKangXi, kKangXi, kKarlgren, kLau, kMatthews, kMeyerWempe, kMorohashi, kNelson, kSBGY.
Dictionary-like Data	kCangjie, kCantonese, kCihaiT, kDefinition, kFenn, kFrequency, kGradeLevel, kHDZRadBreak, kHKGlyph, kIICore, kJapaneseKun, kJapaneseOn, kKorean, kMandarin, kPhonetic, kTang, kTotalStrokes, kVietnamese.
IRG Mappings	kIRG_GSource, kIRG_HSource, kIRG_JSource, kIRG_KPSource, kIRG_KSource, kIRG_TSource, kIRG_USource, kIRG_VSource.
Numeric Values	kAccountingNumeric, kOtherNumeric, kPrimaryNumeric.
Other Mappings	kBigFive, kCCCII, kCNS1986, kCNS1992, kEACC, kGB0, kGB1, kGB3, kGB5, kGB7, kGB8, kHKSCS, kIBMJapan, kJIS0213, kJis0, kJis1, kKPS0, kKPS1, kKSC0, kKSC1, kMainlandTelegraph, kPseudoGB1, kTaiwanTelegraph, kXerox.
Radical-Stroke Counts:	kRSAdobe_Japan1_6, kRSJapanese, kRSKanWa, kRSKangXi, kRSKorean, kRSUnicode.
Variants	kCompatibilityVariant, kSemanticVariant, kSimplifiedVariant, kSpecializedSemanticVariant, kTraditionalVariant, kZVariant.

從表十三、表十四、表十五可看出，無論是 CCCII，或是全字庫、Unicode 的漢字屬性表，都沒有包含漢字的構形知識，而全字庫及

[19]見《全字庫文字、屬性規範 94.1 版》，行政院主計處電子處理資料中心，2005 年 5 月

[20]見Unicode Han Database, John Jenkins & Richard Cook, 網址：http://www.unicode.org/Public/UNIDATA/Unihan.html

Unicode 也無法在編碼架構上反映正異體字，提出的異體字表也僅以簡繁對照為主。因此，如何在電腦中建立漢字的構形知識，以及異體字處理機制，便成為目前建立漢字構形資料庫的重點。

肆、如何建立漢字構形資料庫

建立漢字構形資料庫的目的，在於改善電腦貧乏的漢字知識，進而提升電腦處理漢字的能力，以因應漢字數位化的困境。本節則以表十六「員」的相關字形為例，說明如何建立漢字構形資料庫。

表十六的 12 個字形中，1 為甲骨文，2-3 為金文，4 為為楚系簡帛文字，5 為說文籀文，6-8 為說文小篆，9-12 為楷書。表十七是這些字形的構形分析及使用關係，以下分別說明這些文字知識如何在電腦中表達，以及漢字知識庫的建立步驟。

表十六、「員」的相關字形

 1.合集 10978	 2.戓方鼎	 3.中山王嚳鼎	4.秦 1.2	 5.說文籀文	 6.說文
 7.說文	 8.說文	9.	10.	11.	12.

表十七、「員」相關字形的構形分析及使用關係

字體	字頭	主體字	字形	出處	直接部件	部件集	基礎部件集
甲骨文				合集 10978	○	○	○
金文				戓方鼎	○	○ 阜	○ 阜
				中山王嚳鼎	阜		
楚系文字				秦 1.2	口	口	口
小篆				說文	▢	○ ▢	○ ▢
				說文籀文	○		
				說文			
				說文	▢		
楷書		圓	員		口 貝	口 口 貝	口 口 貝
		員	鼎		口 鼎	貝 員 鼎	鼎
			隕		員		
			圓		口 員		

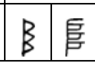
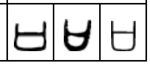
一、建立不同字體的構形資料庫。例如「員」在金文、小篆及楷書的直接部件，沒有兩個是一樣的，因此必須分別建立金文、小篆及楷書的構形資料庫，分別記錄這些字形結構。每個字體的構形資料庫

都會有各自的部件集及基礎部件集。

二、建立不同字體的異體字表。例如小篆的異體字表，「  」為字頭，籀文「  」為重文；楷書的異體字表，「圓」及「員」為主體字，而「員」及「⿱口鼎」為異體字。

三、銜接古今文字以反映字形源流演變。詳見表十八。

表十八、「員」相關字形的字形源流演變

正體字	員	隕	圓	鼎	貝	阜	囗	口
古漢字								

四、制定楷書字形結構表達式。詳見表十九。楷書字形結構表達式（簡稱構字式）是根據部件及部件的相對位置來表達字形的結構[21]，部件的相對位置可簡化成橫向（⿰）、直向（⿱）及包含（⿴）三種。構字式可用來表達缺字，例如Big5 缺字「⿱口鼎」可用「口⿱鼎」表示。

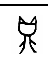
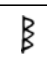
表十九、「員」相關楷書字形的構字式

楷書	員	⿱口鼎	隕	圓
構字式	口⿱貝	口⿱鼎	阝⿰員	囗⿴員
Big5	ADFB	（缺）	B96B	B6EA

五、設計缺字字形產生器。缺字除了透過造字外，還可利用構字式來產生字形，可參考第二節表十一的字根組字。

六、設計古漢字電腦字型。依據古今漢字的字形源流演變，設計古漢字電腦字型，詳見表二十。

表二十、「員」相關字形的古漢字字型

正體字 / 古漢字字型	員	隕	圓	鼎	貝	阜	囗	口
北師大說文小篆								
中研院楚系簡帛文字								
中研院金文								
中研院甲骨文								

七、制定古漢字風格碼。依據古漢字出處及古今漢字的字形源流

[21]見謝清俊〈電子古籍中的缺字問題〉，第一屆中國文字學會學術討論會，天津，1996 年 8 月

演變，制定古漢字風格碼，詳見表二十一。風格碼的型式為「形字或字形體出處」。風格碼和電腦字型的差異，在於電腦字型適合處理規範字，如小篆；風格碼則適合處理形體不固定，書寫帶有隨意性的古漢字，如甲骨文、金文、楚系簡帛文字等。

表二十一、古漢字「員」的風格碼

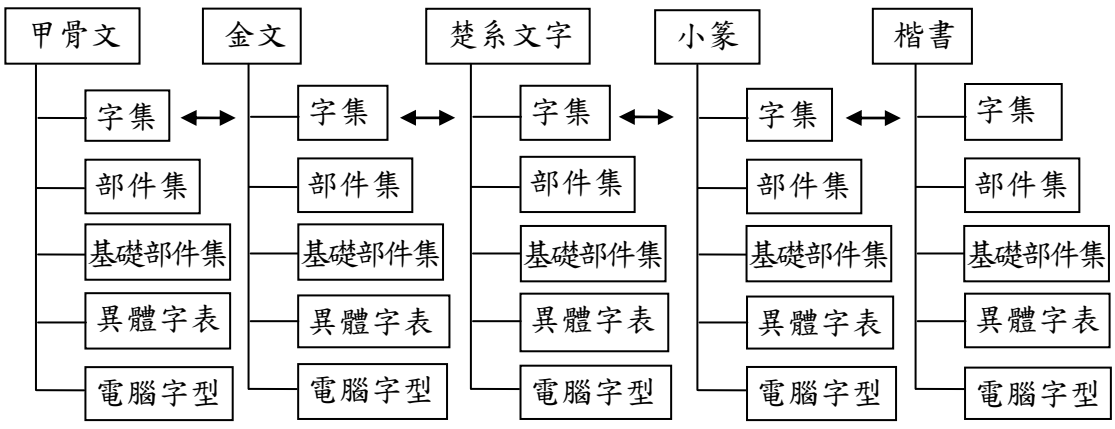
古漢字				
風格碼	形 員 體 合集 10978	形 員 體 集成 5861	形 員 體 集成 2789	形 員 體 集成 9534
古漢字				
風格碼	形 員 體 集成 2695	形 員 體 秦 1.2	形 員 體 說文籀文	形 員 體 說文

以上為建立漢字構形資料庫的概要說明，下節談到漢字構形資料庫的研究成果。

伍、漢字構形資料庫的研究成果

中研院資訊所文獻處理實驗室自 1993 年起，在謝清俊教授的指導下，著手建置漢字構形資料庫。漢字構形資料庫早期收錄的字形是以楷書的現代印刷字體為主，其後陸續增加小篆、金文、楚系簡帛文字及甲骨文。因此，現今的漢字構形資料庫是由甲骨文、金文、楚系簡帛文字、小篆及楷書構形資料庫組合而成，如圖一。小篆構形資料庫已完成，金文、楚系簡帛文字、甲骨文及楷書構形資料庫則是持續的在擴充。建立構形資料庫之前，必先挑選適當的字書，其次建立古今文字銜接及異體字表，隨後進行構形分析，整理出各自的部件集及基礎部件集，再依需要設計電腦字型。

表二十二說明漢字構形資料庫 2.5 版的內容，其中列出各個構形資料庫的主要參考字書、已收字數、已分析字形的部件數及基礎部件數、異體字表組數、製作的電腦字型、提供的字書索引、合作單位等。



圖一、漢字構形資料庫的組成

中研院史語所負責甲骨文、金文、楚系簡帛文字的構形分析，而北京師範大學則提供了小篆字型。

表二十二、漢字構形資料庫 2.5 版內容

	甲骨文	金文	楚系文字	小篆	楷書
主要參考字書	殷墟甲骨刻辭類纂	金文編	楚系簡帛文字編	說文解字詁林	漢語大字典
已收錄字數	2197	20069	16801	11100	62366
部件數	296	804	704	2004	5224
基礎部件數	228	469	464	367	982
異體字表組數	1762	2614	2206	1081	12208
電腦字型	中研院甲骨文及重文	中研院金文及重文	中研院楚系簡帛文字及重文	北師大說文小篆及重文	標楷體及細明體外字集
字書索引	殷墟甲骨刻辭類纂、甲骨文字詁林、甲骨文字集釋	金文編、金文詁林、殷周金文集成引得	楚系簡帛文字編	說文解字詁林	漢語大字典、中文大辭典
其他屬性		殷周金文集成器號、器名	出土墓號及簡號		Unicode、Big5
合作單位	中研院史語所	中研院史語所	中研院史語所	北京師範大學、台灣師範大學	中研院史語所

綜合來看，漢字構形資料庫有以下四個特色：

1.銜接古今文字以反映字形源流演變。

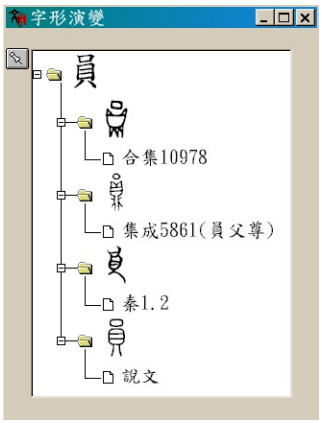
2.收錄不同歷史時期的異體字表，以表達不同漢字在各個歷史層面的使用關係。

3.記錄不同歷史時期的漢字結構，以呈現漢字因義構形的特點。

4.使用構字式及風格碼來解決古今漢字的編碼問題。

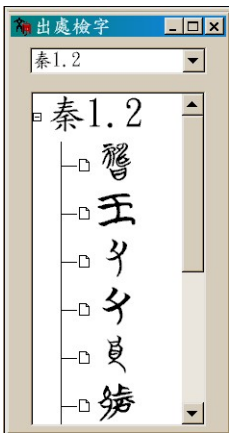
以下透過漢字構形資料庫的使用界面，舉例說明這些特色，並探討如何利用漢字構形資料庫來因應漢字數位化的困境。

一、銜接古今文字。圖二「員」的字形源流演變是參考《漢語大字典》的編排，以楷書「員」作字頭，於其下列出甲骨文「  」、金文「  」、楚系簡帛文字「」及小篆「」，並標示古文字出處。



圖二、「員」的字形源流演變

圖三、出處檢字(集成 5861)

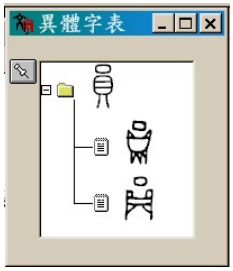


圖四、出處檢字(秦 1.2)

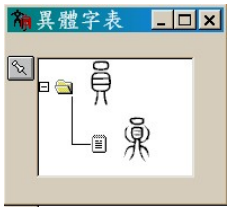
漢字構形資料庫的古今文字銜接，是以楷體字為核心，銜接的不只是古漢字和楷體字，古漢字也可透過楷體字而彼此銜接。古漢字進入電腦勢在必行，Unicode 古漢字編碼方案一旦付諸實施，無論是輸入或檢索，古今文字銜接都扮演了重要的角色。

圖三到圖四說明如何利用古漢字的出處，來檢索漢字構形資料庫中的字形。圖三列出《殷周金文集成》器號 5861 的金文，圖四列出江陵秦家嘴一號墓二號竹簡的楚系文字。

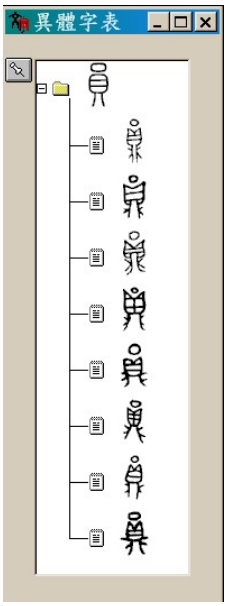
二、收錄異體字表。圖七的金文異體字表是參考《金文編》的編排，以小篆「」為字頭，並於其下列出金文「」、「」、「」、「」、「」、「」、「」、「」、「」。圖五是參考《類纂》，圖六是參考《說文》，皆以小篆「」為字頭，並分別於其下列出甲骨文「」、「」及籀文「」。圖八的楷書異體字表則是參考《漢語大字典》異體字表的編排，



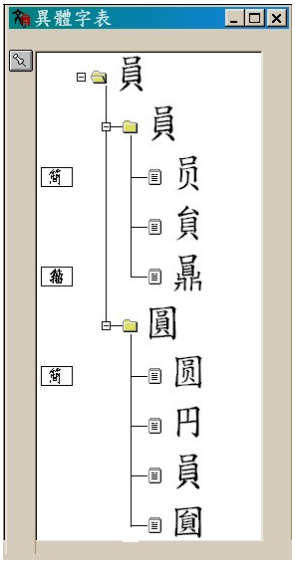
圖五、甲骨文異體字表(員)



圖六、小篆異體字表中(員)



圖七、金文異體字表(員)



圖八、楷書異體字表(員)

以「員」作主體字，再列出異體字「  」、「  」、「  」，其中「」為簡化字，「」為說文籀文；另外，「員」同時是「圓」的異體字。

由於現行漢字是由各個歷史時期的漢字發展積澱而成，楷書異體字表遠比甲骨文等古漢字的異體字表複雜。目前楷書異體字表是《漢語大字典》異體字表的擴充，仍然沿襲《漢語大字典》採用由主體字（其實就是正體字）統領異體字的編排方法，將同一主體字統領的簡化字、古今字、全同異體字（指音義全同而形體不同的字）和非全同異體字（指音義部分相同的異體字），集中在該主體字下編為一組。[22]

除了《漢語大字典》原有的簡化字標示外，我們在建立古漢字構形資料庫時，另外再加入說文小篆、說文或體、說文古文、說文籀文、說文奇字、金文、甲骨文、楚系文字等古今字標示。[23]表二十三是參考古今字標示，再利用古今文字銜接以小篆改寫白居易的〈問劉十九〉，其中「」為說文古文，楷化成「弋」。[24]

表二十三、以小篆改寫白居易的〈問劉十九〉

綠	蟻	新	醅	酒	紅	泥	小	火	爐	晚	來	天	欲	雪	能	飲	一	杯	無
	蟻								鑪					雪		飲	弋	桮	無
																			

隨著中文字碼的不斷擴編，大量異體字進入電腦，異體字表的重要性已不言而喻。例如 Unicode 3.1 的 70194 個漢字絕大多數來自於《漢語大字典》，而《漢語大字典》的異體字約為 11900 組，36000 多個字，這些字無論是檢索或輸入，異體字表都扮演著重要角色。

三、記錄字形結構。圖九到圖十二分別列出漢字構形資料庫中小篆「  」、楷書「圓」、金文「  」、楚系簡帛文字「」的字形結構。圖九小篆「」的字形結構完全反映《說文》的釋形：「圓，圜。全也。从口，員聲，讀若員。」「員，物數也。从貝，口聲。」圖十楷書「圓」的字形結構和小篆相似，除了二級部件由「」換成「口」。圖十一金文「」的字形結構指出小篆「」的部件「」在金文作「」（鼎），而標示部件「○」在小篆作「」，在圖十二的楚系簡帛文字作「口」。

圖十三到圖十七說明如何利用結構中的各級部件，來檢索漢字構形資料庫中的字形。圖十三說明「員」雖然歸在「口」部，但仍可用部件「貝」檢索。圖十四說明「員」雖然不是部首，仍可用來檢索「圓」。圖十五說明可同時使用部件「貝」、「口」來檢索「員」、「圓」。圖十

[22]見《漢語大字典》冊 8，頁 5333。

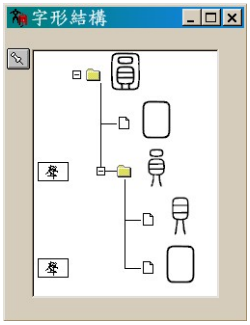
[23]見拙著〈漢字構形資料庫的建置與應用〉，漢字與全球化國際學術研討會，台北，2005 年 1 月。

[24]王心怡曾以小篆改寫這首五言詩。

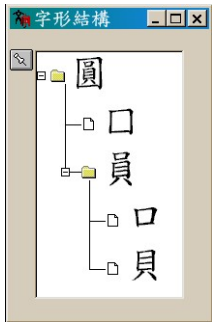
六說明使用小篆部件「  」可檢索出「」、「  」。圖十七說明使用金文部件「  」（鼎）可檢索出「  」（員）、「  」（隕）。

部件是漢字的構形單位，部件觀念對於漢字教學具有重要的意義。[25]部件不同於部首，現在字典的 214 個部首是依循自《康熙字典》，而《康熙字典》的 214 個部首則是根據《說文》的 540 個部首而刪併。例如「員」本是《說文》的部首，但《康熙字典》刪減部首「員」，並將它併在「口」部。部首檢字必須知道所查字形的部首，然而部件檢字可使用該字的任何一個部件，甚至可同時用好幾個部件檢索。從圖十三到圖十七的這些例子可看出，部件檢字實在比部首檢字方便許多。隨著電腦收字的增加，對於採用注音或漢語拼音輸入的使用者，部件檢字也可快速的幫他們找到不會讀的字。

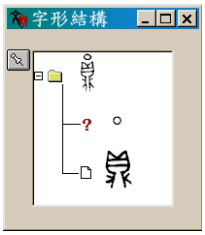
字形結構是漢字構形資料庫的核心，無論是輸入法的部件拆字，電腦字型的字根組字，甚至連交換碼的設計，都和它息息相關，能夠



圖九、小篆「  」的字形結構

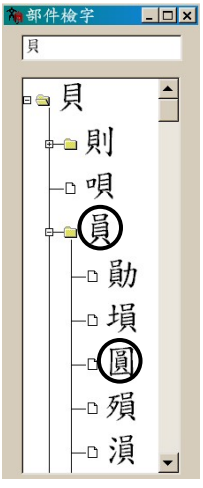


圖十、楷書「圓」的字形結構

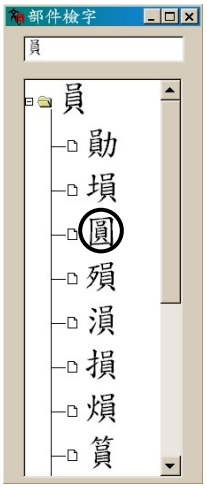


圖十一、金文「」的字形結構

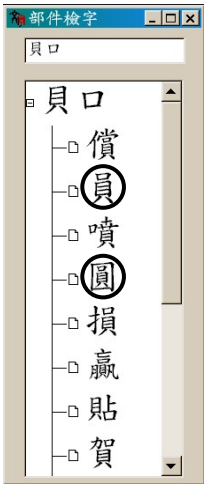
圖十二、楚系簡帛文字「  」的字形結構



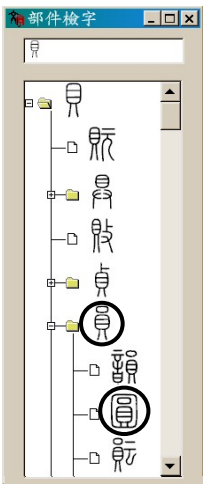
圖十三、楷書部件檢字(貝)



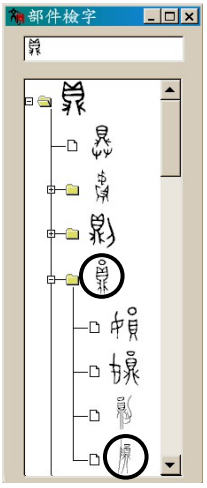
圖十四、楷書部件檢字(員)



圖十五、楷書部件檢字(貝口)



圖十六、小篆部件檢字()



圖十七、金文部件檢字()

[25]見《漢字教學的理論與實踐》，頁 90。

衍生的運用也最多。

四、古今漢字的編碼問題。漢字構形資料庫是採用構字式及風格碼來解決古今漢字的編碼問題，構字式適用於楷書，風格碼則適用於金文、甲骨文等古漢字。

表二十四、構字符號及構字式

類別	符號	說　明	構字式範例
連　接	⿰	當部件的連接順序由左至右	順＝川⿰頁
	⿱	當部件的連接順序由上至下	含＝今⿱口
	⿴	當部件的連接順序由外至內	圍＝口⿴韋
部件序	⿸	按部件書寫順序輸入，前後以起始符號（⿸）和終止符號（⿹）包夾。	解＝⿸角刀牛⿹
	⿹		
方便符號	⿲	二個相同部件直連	炎＝⿲火
	⿳	三個相同部件直連	
	∞	二個相同部件橫連	朋＝∞月
	⿾	三個相同部件橫連	
	⿻	三個相同部件呈三角狀排列	焱＝⿻火
	⿿	四個相同部件橫連	
	⿵	四個相同部件直連	

構字式即字形結構式。一個字的字形結構式，是該字極佳的識別符號；因為字形若不一樣，則字形結構必不相同；反之，字形結構若相同，其形也必相同。

字形結構的差異主要在於部件的選擇，其次才是部件的位置。例如「員」和「貢」的差異在於部件「口」和「工」的不同，而「員」、「唄」和「  」（甲骨文「  」）的差異在於部件「口」和「貝」的相對位置不同。絕大多數因部件位置不同而差異的字形，它的部件都只有兩個，而部件的相對位置為左右、上下或內外。於是我們定義表二十四的構字符號，並且用構字符號和部件來表達字形結構。

現行的中文字碼最基本的是「字」，而構字式最基本的則是「部件」。相較於現行的中文字碼，構字式更適合用來表達缺字，尤其是大量可類推的異寫字。例如「鵝」的異寫字「  」、「  」、「  」的構字式分別為「鳥⿰我」、「鳥⿱我」、「我⿱鳥」；金文「隕」作「  」，而「  」的楷化字作「」，構字式為「  」。

現行中文字碼的最大問題，是將漢字視同拉丁字母，採用相同的編碼架構，而完全忽略漢字是表意文字，是由有限的基礎部件所組成的開放字集。由於字碼是電腦的根本，中文字碼不完善，中文電腦的架構自然不健全，漢字在數位化過程中的缺字問題，也就不令人感到

意外。

構字式在漢字知識庫的建立過程中，一直扮演著重要角色。Unicode 的編碼空間雖然很大，跟構字式比起來，畢竟還是有限，構字式才真的是無限。十幾年前，我們開始在 Big5 的環境下建立漢字知識庫，Big5 只有 13051 個字，但是我們不但收了五萬多個字，還陸續增收小篆、金文、楚系簡帛文字、甲骨文等古漢字，這都是拜構字式之賜。

制定風格碼的目的在於解決古漢字重文的編碼問題。風格碼是構字式的延伸，構字式是利用字形結構來區分字形的字形碼，而風格碼則是利用出處來區分同一個字形而風格迥異的異寫字。表二十五列出漢字構形資料庫引用古漢字時，常見的出處。風格碼的型式為「形字或字形體出處⊙」，其中「形」、「體」、「⊙」為標示符號，「體」表同一個字（或字形）的不同形體，夾在「體」、「⊙」中的為出處。例如金文「」的風格碼為「形員體集成 5861⊙」。

金文的引用常以器名標註出處，然而器名並不是唯一的。例如《集成》器號 23 到 30 的青銅器，器名都是「中義鐘」。另外，由於金文楷定的不同，青銅器的命名，也常和《集成》有所出入。例如器號 38 的青銅器，《集成》的器名為「  」，然而「」的後起字形為「荊」，「」為「曆」，所以「」也可稱作「荊曆鐘」。因此，金文的風格碼改以器號為主。

表二十五、古漢字出處

古漢字	出處
甲骨文	合集(甲骨文合集)、屯(小屯南地甲骨)、英(英國所藏甲骨集)、懷(懷特氏等所藏甲骨集)。[26]
金文	集成器號或器名(約 12000 件)[27]
楚系簡帛文字	牌 406、仰 25、常 2、望 1、望 2、天卜、天策、雨 21、馬 1、磚 370、秦 1、秦 13、秦 99、范 27、滕 1、包 2、信 1、信 2、曾、帛甲、帛乙等出土墓號及簡號。[28]
小篆	說文、說文或體、說文古文、說文籀文、說文篆文、說文俗字、說文奇字

風格碼的概念也可再延伸來處理合文，如表二十六。這時只須將起始符號「形」改成「合」即可。

[26]見《殷墟甲骨刻辭類纂》所引甲骨著錄書目。

[27]見《殷周金文集成》目錄

[28]見《楚系簡帛文字編》所引文字的出土墓號簡稱表

表二十六、合文的表達方式

合　文	1.牝牡  合集 19987	2.無疆  邾伯祀鼎	3.寶尊  作旅寶彝卣
風格碼	合牝牡體合集 19987	合無疆體集成 2602	合寶尊體集成 5121
合　文	4.公孫  信 1.06	5.招財進寶 	6.唯吾知足 
風格碼	合公孫體信 1.06	合招財進寶	合唯吾知足

除了上述四個特色外，漢字構形資料庫還提供豐富的字書索引以省卻讀者翻閱部首檢字表的麻煩。另外，我們也在資訊所同仁的協助下，開發了微軟Office文件及網頁的缺字應用程式[29]。

陸、漢字構形資料庫的未來研究方向

漢字構形資料庫的建立，除了因應漢字數位化的困境外，它的長遠目標則是推動文字學的數位化。若不管音、義，只把文字學看作狹義的文字形體研究，漢字構形資料庫的未來研究方向如下：

一、持續建立甲骨文構形資料庫。《類纂》收錄字頭 3556 個，字形 4488 個，目前已完成字形分析，並加入《類纂》、《甲骨文字詁林》、《甲骨文字集釋》的索引，同時也掃描了《類纂》的字形總表，製作甲骨文描邊字型。甲骨文構形資料庫目前只先收可楷定的字頭及重文 2197 個，其他的甲骨文將持續收錄。

二、透過風格碼持續增收金文及楚系簡帛文字重文。《金文編》收錄字形 24261 個，《楚系簡帛文字編》為 19250 個字形，而目前漢字構形資料庫已收金文 20069 個、楚系簡帛文字 16801 個，其餘的重文仍需持續收錄。

三、提供缺字圖片下載。漢字構形資料庫自 2002 年 10 月開始，即可由網路免費下載。下載人次由剛開始的每日兩次到 2007 年 8 月的每日九次，累計下載人次 8257 次。然而漢字構形資料庫由於描邊字型的檔案太大，整個資料庫的下載壓縮檔約為 80M，再加上下載後的安裝問題，也會讓部分使用者卻步。其實對大多數人而言，缺字出

[29]見《漢字構形資料庫使用手冊》，莊德明、許婉蓉，中研院資訊所，2002 年 7 月。

現的機會並不多，於是我們打算同時推出缺字圖片下載，設計原則如下：

1.缺字有字集的針對性，對於未裝漢字字型的電腦而言，所有的漢字都是缺字。這個網站提供的漢字個數將超過十一萬，包含楷書字形六萬個，甲骨文、金文、楚系簡帛文字、小篆等古漢字五萬個。

2.採用部件及出處來檢索古今漢字，並透過異體字表及字形源流演變，銜接異體字及古今字。

3.提供任意大小，任意解析度的漢字圖片以滿足文書及美編的需求。這些圖片可由現有的描邊字型自動產生，其中楷書字形還可分成標楷體和細明體兩種。

四、開發 Unicode 版的漢字知識庫。漢字知識庫由 1993 年開始，一直都在 Big5 的環境下研發，因此目前還無法適用於微軟視窗簡體中文版。隨著 Unicode 的逐漸普及，開發 Unicode 版的漢字構形資料庫，方可適用於微軟視窗簡體中文版，甚至日文版。從 Big5 轉換成 Unicode，不只是轉碼而已，連系統開發工具都要更換，程式幾乎到了要全面改寫的地步。

五、整合字形產生器。字形產生器可以根據構字式來產生字形，圖十八的字形產生器是馬來西亞華僑葉健欣所設計[30]。葉健欣和本實驗室有長期合作的情誼，願意提供出來讓大眾免費使用。圖十八「鉤」的簡化字「钅厶句」並未收錄在《簡化字總表》，這類可類推的異體字數量極多，適合採用構字式來表達，並透過字形產生器動態產生字形。由於目前字形產生器為Unicode版本，所以這項整合工作會在Unicode版的漢字構形資料庫開發完成後再進行。

圖十八、字形產生器

六、建立戰國文字構形資料庫。戰國文字依現有的研究成果，可分為齊、燕、晉、楚、秦五系；若依傳統書寫材料分類，則可分為銅器文字、石器文字、貨幣文字、璽印文字、陶器文字、簡牘文字、漆器文字、縑帛文字八類。[31]戰國文字最突出的特點是形體歧異多。這個構形資料庫的開發時間，應在Unicode版的漢字構形資料庫完成後。

七、建立楷書筆畫資料庫。筆畫是楷書字形的書寫單位，它包括

[30]易符無限組字編輯器 2.0.2 版下載網址：
http://www.eforth.com.tw/localdown.htm
[31]見《戰國文字通論》，何琳儀著，江蘇教育出版社，2003 年 1 月

筆形、筆順及筆畫數等屬性。筆形即筆畫的樣式，筆順即一個漢字中在書寫時各筆畫的先後順序，筆畫數即組成一個漢字的筆畫數目。書寫單位不同於構形單位，漢字最小的構形單位是基礎部件，而基礎部件是由筆畫按筆順書寫累積而成的。[32]建立楷書筆畫資料庫可利用楷書構形資料庫打下的基礎，先將基礎部件拆成筆畫，但是一個字的筆順和基礎部件的筆順不一定相同。例如「圓」由基礎部件「口」、「ロ」及「貝」構成，但是「口」的最後一筆要等到「ロ」、「貝」寫完以後才寫。楷書筆畫資料庫的開發時間，也在Unicode版的漢字構形資料庫完成後。

柒、結語

世界上的文字分為表意文字與表音文字兩大類，而漢字是最典型又成熟的表意文字。現在中文電腦的文字編碼架構是依據表音文字來設計的，擁有的漢字知識自然不夠，因此漢字在數位化的過程中，總是不盡如人意。為了解決缺字問題，中文字碼不斷的擴編，排序混亂、罕用字的輸入及異體字的檢索等問題，又更突顯出建立漢字構形資料庫的重要性。

自從微軟視窗 2000 開始支援 Unicode 2.0，台灣及大陸的中文系統已可同時使用繁體和簡體字，而且內碼完全相同，不同的只是使用者界面選用繁體或簡體的差別而已。面對漢字數位化的困境，不管是漢字簡化產生的問題，或是繁體字本有的問題，目前大家都使用同一個系統，都需共同來面對。

在建立漢字構形資料庫的過程中，我們不但利用電腦來整理漢字知識，也同時利用這些知識來改善電腦處理漢字的能力。漢字的構形知識毋寧是目前漢字構形資料庫中最重要的，不但可用來表達缺字，也可用來改善或設計新的輸入法及字庫，或輔助漢字電腦教學。

這些年來使用電腦的經驗讓我深深體會，越早使用電腦，會讓後續的工作進行得更順利。利用構字式或風格碼來表達古今缺字，就是一個很好的例子。或許現在構字式用起來還不是很方便，但在處理缺字上，已能節省不少時間。即使 Unicode 要加入古漢字，也不可能因為電腦沒有這些字，就不用電腦處理，這時風格碼就可派上用場。

中文電腦發展至今，文字學者和語言學者不是參與太少，就是讓步太多，再加上電腦工程師又常自以為是（個人雖也引以為戒，但有時仍會犯此錯），不但讓電腦裡的漢字知識頗為不足，也讓新的漢字處理技術遲遲無法推出。風格碼就是在我終於體會古漢字重文的重要性後，才設計出。

從 2003 年底開始處理金文，2004 年底處理楚系簡帛文字，2006

[32]見《漢字漢語基礎》，頁 99。

年初處理甲骨文，有幸與中研院史語所的同仁共事，讓我這幾年來能浸淫於古文字的世界裡。每當看到字書中的古文字，一個個進入到電腦裡，我的心中就充滿了快樂。

參考文獻

1. 中國社科院考古所，《殷周金文集成》，中華書局，1984-1994 年
2. 王寧，《漢字漢語基礎》，北京科學出版社，1996 年 7 月
3. 王寧，〈漢字研究與資訊科學技術的結合〉，漢字與全球化國際學術研討會，台北，2005 年 1 月
4. 何琳儀，《戰國文字通論》，江蘇教育出版社，2003 年 1 月
5. 吳福生，〈字形面面觀〉，0 與 1 科技・BYTE 中文版，155 期，第三波文化事業，1994 年 3 月
6. 周法高等，《金文詁林》，香港中文大學，1974 年
7. 周曉文，〈建立「信息交換用古漢字編碼字符集」的必要性及可行性〉，古漢字數位編碼暨現代化應用研討會，台北，2005 年 10 月
8. 姚孝遂，《殷墟甲骨刻辭類纂》，中華書局・北京，1989 年
9. 容庚，《金文編》，中華書局，北京，1985 年 7 月
10. 徐中舒，《遠東・漢語大字典》，遠東圖書公司，1991 年 9 月
11. 徐中舒，《甲骨文字典》，四川辭書出版社，1989 年 5 月
12. 國字整理小組，《中國文字資料庫》，行政院文化建設委員會，1985 年 5 月
13. 張亞初，《殷周金文集成引得》，中華書局，北京，2001 年 7 月
14. 莊德明等，《漢字構形資料庫使用手冊》，中研院資訊所・台北，2002 年 7 月
15. 莊德明，《漢字構形資料庫的建置與應用》，漢字與全球化國際學術研討會・台北，2005 年 1 月
16. 許壽椿，〈網絡時代的漢字全面解決方案和漢字本體研究〉，’99 漢字應用與傳播國際學術研討會，北京，1999 年 6 月
17. 許學仁，〈古文字資料庫與古文字數位編碼〉，古漢字數位編碼暨現代化應用研討會，台北，2005 年 10 月
18. 游振昌，〈字裡乾坤〉，第三波，1994 年 5 月
19. 黃沛榮，《漢字教學的理論與實踐》，樂學書局，2001 年 12 月
20. 楊家駱，《說文解字詁林》，鼎文書局，1994 年 3 月
21. 滕壬生，《楚系簡帛文字編》，湖北教育出版社，1995 年 7 月
22. 謝清俊，〈電子古籍中的缺字問題〉，第一屆中國文字學會學術討論會，天津，1996 年 8 月
23. 謝清俊，〈談中國文字在電腦中的表達〉，「中國文字的未來」學術研討會，台北，1991 年 6 月
24. 《全字庫文字、屬性規範 94.1 版》，行政院主計處電子處理資料中心，2005 年 5 月

參考網站

1. Unicode Han Database, John Jenkins & Richard Cook, 網址 http://www.unicode.org/Public/UNIDATA/Unihan.html
2. 中國國學小學館，網址 http://www.superlogos.com.tw/main9/index_02.htm
3. 全字庫網站，網址http://www.cns11643.gov.tw
4. 教育部國語會《異體字字典》，網址http://140.111.1.40/main.htm
5. 易符無限組字編輯器 2.0.2 版下載網址 http://www.eforth.com.tw/localdown.htm
6. 漢字構形資料庫下載網址 http://www.sinica.edu.tw/~cdp/
7. 薛偉傑，〈鍵盤輸入三大主流〉、〈輸入法如此多嬌引無數英雄競折腰〉，網址http://input.foruto.com/introduce/index.html