等待新漢碼

——漢字的數位化與中華文化的衝擊

本文參考易符智慧科技所發表「中文資訊的表達與易符無限字庫」一篇，針對有關當今中文數位化之困局加以剴切剖析闡述，文中許多觀念乃源於中央研究院謝清俊教授之啟發及葉健欣先生之導入，特此銘謝。

2006.9月陳昌江（感謝張正一 等人協助校稿）

2007/5/9進一步修定：「用處理英文字的方式來處理中文字」更精確化，應是「用處理英文字母的方式來處理中文字」

2007/3/19更正：『康熙字典』的基本部首應是214個，感謝黃大一先生賜教（by正一）

摘要說明

一百多年來，中華民族在優勢的外來文明衝擊下，人民普遍喪失民族自信心，不僅使得中國傳統文化成了代罪羔羊，也使其更新的腳步停滯不前，無法受到應有的重視與發展。最無奈的是，許多中華文化的寶貴資產，就在這樣的時代洪流中無聲無息流失掉矣！

今天，兩岸三地的大漢民族普遍都富足了，然而這種文化上的自卑，仍然普遍存在著。在這樣的歷史洪流中，中華文化的更新與發展，當是這時代的歷史使命。其中

，在整個中華文化的數位化轉移中，漢字是中華文化的根本材料，其影響是無所不在的，因而漢字的數位化，便是中華文化進化到數位時代的基礎工程。
在漢字數位化工程中最基本的就是漢字表達的基礎結構。漢字數位架構的良窳，深深地影響到中文資料儲存成本、交換成本以及檢索效能等，這個基礎架構，也關係著中華文化的傳承與創新的能力，簡而言之，攸關著整體數位中文資訊處理的基礎成本，也牽動著中文資訊的表達能力，也就是影響著漢字數位內容的終極價值！
因此，這是一個重要而嚴肅的議題。

漢字資訊的五大要素

自古漢字就由「形、音、義」三個要素所構成，在資訊時代則必需加上「碼」和「序」二個要素。
「碼」是電腦認定一個漢字的一個相對數字，通稱為「字碼」，所有電腦的資料處理、資料交換都是針對「字碼」進行認定和處理。

「序」係人類認知的排列方式。由於需要查找排序和比對等資料處理，一個自然、共同認定的「字序」是一個文字系統重要而有價值的本質。 以查字典為例，查英文字典時的簡單方便而且準確，但查漢字字典就很不確定，這種問題相信你一定能感受到，這是因為漢字還沒有確定的字序的原故。

當前的漢字資訊表達的情況

形

字形就是人的眼睛所看到的。漢字字形的產生主要有點陣字和向量字兩種。

點陣字形

點陣字對電腦來說其實是一種「字圖」，就是在有筆畫的地方描上細細的點。點陣字的好處就是處理簡單，缺點就是每一種尺寸都需要一套點陣資料，因為一個點陣字就是一張點陣圖片，且資料量與字形的大小成等比級數上升，字形變大，資料量快速變大。這使得記憶體受限的小型數位裝置所能提供的字形就非常有限。

另一方面，要從這點陣資料圖中取得有關這個字形的特徵資訊不多，因此，除了進行高級的影像處理外，點陣資料的進階處理或應用都不容易。

向量字形

向量字則是只記錄各筆畫內容的位置、長度寬度等字形資料，而在最後展現時，才由電腦轉換成點陣圖來呈現。

向量字的發展主要為了解決點陣字資料量龐大的問題。但向量字形在呈現成點陣時所需要的轉換非常複雜，目前在機能不夠強大的小型數位設備上仍不易實現。

音

由於漢字是一種形意文字，與音韻並無緊密的連結，加上古今漢語音韻之變遷，形和音的對映是多對多的(多字同音，一字多音)，其中字音可以簡單地用建表的方式解決。但如果要處理破音和語境問題，就涉及自然語言處理的範疇，這方面學術單位已有相當多的有關研究。

義

形是義的視覺介面，音是義的聽覺介面，有形無音，則稱為「符號」，有音無形，叫作「語言」，只有同時具備形、音兩要素，才構成文字。

碼

中文在資訊時代的第一個挑戰是「編碼」，也就是為每一個漢字編上一個數字碼。一個漢字未被編上一個對應的字碼，就無法進行數位化處理，這也等於在數位世界中「不存在」，也就是在數位世界「沒有這個意思、沒有這個人、沒有這件事、沒

有這個地方」。

碼可分為「內碼」和「輸入碼」兩種，內碼係中文字的**數位代碼**，是方便電腦處理的代碼，沒有考慮人的記憶或邏輯，因此才衍生了各種方便人的記憶或辨識的輸入法來產生相應的內碼，輸入碼主要是針對輸入漢字的人機介面，也是作為人和機器溝通時的中介表達轉換之用。

內碼

內碼的主要考量是軟體的相容性、儲存的效率和程式處理的簡易性，因為在這數位世界中，漢字字碼是無所不在的，因此漢字的處理成本，這也就成了無所不在的成本負擔。

在早期電腦的文字模式(text mode)時代，為了遷就 ASCII碼表，故有 Big5、GB、 JIS 等雙字元的設計（一個**字元**就是一個BYTE，一個BYTE＝8位元，雙字元＝16位元；位元(bit)，就是一個表達0與1的單位）。然而，電腦進入圖形模式的現在，字形在螢幕上的顯示，已不再限定為固定寬度，加上當今電腦的容量與速度，因此對於實際儲存的字元數以及運算的複雜度已經不再是取捨的前提了，於是中文內碼的設計上，就有了很大的自由度。

目前電腦平台上涵蓋面最廣、最廣用的內碼 Unicode（統一碼），已經成為當今 Windows、Mac及 Unix-like等主流平台的內碼，因此Unicode 事實上已取代 ASCII 、Big5、GBK碼，成為各作業系統的預設編碼，並漸漸地成為國際間交換資料時主要的交換碼。

輸入碼

輸入碼可分為「拆形」和「拼音」兩大類。「電腦中文化」的歷程就是利用英文電腦的鍵盤，編上部首和注音的映對鍵位。然而中文部首的數目遠遠超過了鍵盤的鍵數（『康熙字典』的基本部首有214個），因此就必須在有限的鍵盤上，用一個鍵對應多個部首的方式來輸入。

由於這些分解動作，都加入了人為指定與巧思，並非來自文字的本質，因此需要很多的學習和記憶，於是成了漢字使用者的一個額外的負擔，這樣對漢字使用者無疑是建立了一個很大的門檻，**現在社會上還有很多人「不會電腦」，其實大部分都是「不會輸入」的意思。**這種現象不僅在大人的世界發生，在兒童

方面，也因為這個緣故，也在無形中電腦的啟蒙時間也被延後了，這使得使用華文的小孩在電腦應用與普及上與英語世界相較，也是有輸在起跑點的無奈。

一字一碼的時代困境

我們必須深刻地覺悟到，承載中文資訊的中文碼，其基礎架構對「數位中華」的影響是既深且遠的，若不深入觀察分析，大家也習以為常，不容易看出它無所不在的影響以及問題的嚴肅性。就以康熙字典為例，一萬多字的BIG5碼是做不出有四萬多字的康熙字典的，BIG5不行，兩萬多字的GBK也不行，如果字碼的根基沒解決，又要如何把中華文化數位化？

為了讓你發現這些在我們數位生活存在的諸多無奈事實，且讓我們來分析觀察英文字（word）的結構。

首先讓我們來看字序的問題。我們都知道，英文碼的基本定義是0～127的ASCII碼，其中有『A～Z』、『a ～z』 的52個『**英文字母**』（character），其餘為字符碼及控制碼。由ASCII碼的英文字母所構成有意義語素是 **word**，我們就以『**英文字**』稱之。各位請注意到，**英文字(word)**循著ABC的排序，就已經有了一個自然的、本質的排序。

在此一基石之上，舉凡字典的安排、資料庫的製作、物料的列舉、名單的列舉、二元搜尋（binary search）的方法、鍵盤的設計、作業系統表單的設計、快捷鍵（HOT KEY）的安排等，都離不開這ASCII 編碼的基本安排，其影響是無所不在的，可是中文字卻沒有這個序，只要稍有中文處理經驗的人，便可以知道，資料欄位沒有確定的排序，電話簿中的人名沒有確定的排序！

為了這樣的緣故，中文資料總是要另外自行設代碼或欄位編號等，以方便處理，這相對於英文，中文的資料處理，便增加了一層無所不在的額外成本。

發現潛藏在當今「一字一碼」架構中的意義

現在，再讓我們來看看當今中文字一字一碼的問題。

為了讓讀者發現這些潛藏在文字架構中影響力，讓我們來考慮下面的文字假設情況：

如果，我們把 ASCII碼拿掉，改用一個**英文字(word)也像中文一樣一字一碼**，那麼將會是個怎樣的景象？

我們先假設下列英文字**都**有了**內碼**：

PERSONAL　內碼是 $FF3A
CENTRAL　內碼是 $BB01
PROCESSING 內碼是 $FF3B
UNIT　內碼是 $FF3C
MACHINE　內碼是 $CC01
COMPUTING 內碼是 $DD02

那麼COMPUTING MACHINE（內碼為$DD02 $CC01）就**沒有機會**因為它的重要性日增而改稱 COMPUTER了。請注意：因為沒有「COMPUTER」這個內碼，如果要，就要某個標準單位提報，然後經過審核程序，在下一版中公佈新字碼！

好！假設真有那麼一天，「標準機構」「收錄」了COMPUTER這個新字：

COMPUTER 擴充新內碼是$AA01

一樣的問題又來了，在有了「COMPUTER」這個新字碼之後，PERSONAL COMPUTER（內碼$FF3A $AA01）仍不能馬上改稱PC，因為「標準單位」還沒有定下「PC」這個字碼！

同樣地，中央處理單元 CENTRAL PROCESSOING UNIT（內碼$BB01 $FF3B $FF3C）更不會簡稱 CPU了，**因為如果英文字也是像中文一字一碼的話，也就沒有機會創造出「CPU」這個新字了！**

當然這是個假設性的探索，英文文字事實上可以自然地隨著時代的需要「進化」，**這種進化機制可是關乎一個文化的根本活力！**

然而，這卻也正是這些年來，一字一碼的中文所經歷的過程。

諸位一定可以體會到，**所謂的一字一碼，就是拿處理「英文字母」(alphabet)的方式來處理中文字**，這是一個耗時費力而不切實際的過程！

然而，我們更需要嚴肅看待的是，這樣的困局所引發的嚴重後果：

漢字停止演化！

只因為在一字一碼的架構中，要增加一個新字，是一個令人無法承受的夢魘！

讀者是否可以看出來，當一字定成一碼的時候，由於是人為指定，於是，一個新字必須經過標準機構的公佈才有可能流通和使用，然而即便一個新字已經公佈了，無數已經在運行的系統又如何去更新呢？所以，這是成本非常高、過程複雜且時間漫長的過程！當然，由於太不切實際，其真正的結果一個就是「停止造新字！」，另一個就是已經做好的大型資料庫，還是用舊的碼，因為更新一個大系統，通常是一個重大的工程。這就是這幾十年漢字的僵化的景況。
以維基百科中的週期表 為例，其中的105~108元素是「有碼沒字」（未來可能會有字）和而第112~118元素則沒有碼，先前的元素在電腦出現之前，還可以造字，所以有字了，只好用英文表達或改用「元素-112」代表之。
為什麼？ **因為112~118已經進入一字一碼的時期，停止造字了！這也恰好說明了中文字因為一字一碼使得,漢字停止演化，也使得漢字漸漸無法表達新生的事物。**

於是，當今的一字一碼架構也就成了漢文字生機的死胡同！
然而，很無奈地，這卻是當今漢字數位化所存在的真實困局！

沈重的一字一碼

雖然現在這種人為的一字一碼並不是完全地不可行，問題就在必須每隔一段時間以人工審議的方式以追加新字碼，而在字碼尚未公佈前，中文數位資料的轉換、交換、搜尋比對都是不可能的，更別說是無法輸入和無法印出這樣的基本動作了。以佛教經典來舉例，佛教典籍有龐大數量古字未被編碼，早期佛教界

做了許多典籍的輸入，雖然耗費龐大的人力物力來造字，但由於各佛教系統各編自己的碼，因此至今卻仍是難以流通，即使今天要全面的更新這些既有的龐大系統又談何容易！

既使是在新標準公佈之後，由於許多已存在多年的系統無法隨著更新，所以要能全面地交換、搜尋和比對，仍然是一條漫漫長路，更別提UNICODE到2006年已經到了七萬多個漢字，表面上好像是解決了缺字的問題，但卻也是一個龐大的系統負擔（2006年Windows XP大部分的字型也只放了兩萬字）。因此，這些漢字只是「存在」但並非常用，這不僅是小型資訊設備無法承受記憶體的消耗（相較於英文文字系統是非常的龐大），就連我們在輸入時，也無法忍受輸入時每次從上百個字中挑選你要的字。

請想一想，如果一個龐大的資料庫，卻無法準確的搜尋、不能排序、難以粹取轉換、難以流通，請問這樣的「數位內容」的價值，是不是大打折扣，甚至在嚴肅的應用上如學術、戶政、刑事或醫療等應用上，就必須另建系統來保證其準確性或者乾脆捨棄，只能引用外文資料庫。

由於當今的BIG5、GBK、UNICODE等幾個主要中文碼都一樣是這種一字一碼的架構，所以都面臨相同的困境。

因此，我要說「**人為指定的一字一碼是漢字數位化進程中的歷史錯誤！**」。

中文在一字一碼的架構中固化了

我們中文漢字在每字指定一碼的架構下，「以筆書寫，自由創造」的漢字本有的生命力不見了，就因為這種架構讓漢字在數位世界中「固化」了！

這樣的固化現象是無所不在的，其效應也是無聲無息地不易被察覺的。為了更具體的剖析說明這種失去活力的「固化」過程，這裡再舉幾個例子來加以說明。

百年來，對人類非常重要的日常用具──電燈，按倉頡以來中文形聲造字的法則，最終應是進化為「**電登**」這個「字」（注意「電」「登」兩個部首寫併成一個漢字，因為「現在電腦還沒這個字碼」，所以這裡無法顯示！）（這個

新字應是唸做「登」）。（2007.8.1 阿江註Firefox 「新同文堂」模組很快就要提供 動態組字的PLUGIN了）

想一想，當電燈剛出現時，中國仍處於油燈的時代，所以借用火旁的油「燈」再加上一個電字來修飾當時的燈字。另外，像網際網路(互聯網)更已經是這數位時代生活密不可分的一部分，按倉頡造字進化的原理，它的新字應該是「***互罔***」，這便是一個文字活力成長的機制。

晚近幾十年，我小時候的油燈現在已幾乎看不到了，「電燈」普及了，我們已經不再需要說「開電燈」「關電燈」來與油燈分別，而直接說「開燈」「關燈」，這是語言本身隨著生活時代不斷演進的例子。你只要仔細觀察，這種例子俯拾皆是。

其實，這個新時代新增的字很多，像MODEM英文的這個字便是從「MODulation and DEModulation」複合而成，然而，由於目前中文碼是「一字一碼」，因此這個「調變解調機」（或用「數據機」簡化）就被「困住」了而不能隨需要進化，只因為中文沒有字碼也「不容易」另定字碼！

這都是因為現在使用的是一字一碼的定碼機制，我們所能做的，就只是**用現有的字碼來組新詞**，就是無法造新字！儘管時代不斷地演化，重要用品和概念不斷地出現，我們卻無法進一步跟著簡化。

於是，英文字隨著時代在進化中，中文字卻僵在原處！

中文在一字一碼的架構中僵住了

這在這個案例中，中文僵住了！OK，也許會有人說，「中文僵住了又怎樣？日子還不是一樣在過？」

當然，在BIG5時，用電腦、打手機簡訊也都可以啊！沒錯，但是，其結果就是下面的光景在不知不覺中大量地普遍地在進行著：

以「中央處理器」和「CPU」為例，許多人在生活中、文章中會不知不覺地會直接用「CPU」而放棄寫冗長的「中央處理器」，真的，實在太累了，可是一用CPU，就有許多小孩、老人和那些非資訊背景的人就更搞不懂了！（像ADSL、MODEM這種字也都是一樣的情形）。

而這樣的情況不只是發生在資訊界，也同樣發生在學術、工程、科學、醫療、

農業、生物、經濟、管理。。。等等所有進化中的領域。這樣的情況時間越久，中文所不能表達（或因不實用而被棄置不用而直接用英文）的字詞就會累積得更多，長久下去，中文就這樣無聲無息地漸漸地與時代脫節，也就漸漸失去一個語言的實用性與優越性！

各位要覺悟到，這種漢字的困局，是漢字的使用者必須自己關心解決的問題，外國人不會替你解決，UNICODE不斷地編碼，只是在解決跨國市場的全球化的需求而已，至於這架構的好壞，對漢字文化的未來的衝擊，外人怎麼可能替我們認真的面對！

字碼在無聲無息無所不再地影響著我們！

字碼的影響力是無聲無息的，無所不在的，我們都不知不覺中受到這種基本機制所制約而不自知。

文字是如何地在無聲無息中影響著我們？為了讓各位更清楚地看見中文碼對中文活力的影響，讓我們再舉下面的這些例子來觀察思考：

CPU 是電腦的心臟，在這個數位時代是如此的重要，所以常常被使用到。前面提到，大家寧可寫「CPU」而不用「中央處理器」，因為寫起來太冗長了。然而，換個角度，也是因為我們無法用「**电心**」（或「電心」注意，這是一個漢字，因為BIG5、GB、UNICODE裡還沒有「指定」這個字，因為沒有這個字碼，所以也無法用電腦顯示）這是個極簡潔而恰當的新字。

註：當有一天，「**电心**」這個字能夠自然地在一般系統中自然呈現時，也正是組字碼普遍使用之時。讓我們共同等待這一天的到來吧！（阿江，於2007.4.20註）

同樣的，就像英文可以把Personal- Computer 簡化成「PC」，但中國人卻得永遠寫成「個人電腦」，難怪會有很多人直接寫PC了，另外像「光碟」這個數位時代的關鍵儲存裝置，因為沒有「**光棐**」這個字，所以只能用「光碟」，但「**光棐**」（這是一個字）就明顯比「光」「碟」兩個字來的有效率。（至於「電腦」應該造成怎樣的新「字」你一定馬上可以想得到如何寫了！停下來想一想，其實，字的演化是這麼自然而且簡單。）

再如英文 BIT在電腦方面我們叫做『位元』或『比特』 BYTE 則譯做『位元組』或『字節』，但其實 BIT 的零一單位不就是中國既有的二進位系統易經八卦中的「爻」(音『姚』或唸成英譯的『必』也很好)，而8個BIT叫『**八爻**』(一樣，要併寫成一個漢字，唸『拜』，再也自然不過了)，依此原則可以進一步造出 16BIT WORD，32 BIT WORD的字，這樣的自然演化其實只是還給漢字本有的活力而已。

新中文碼的時代需求

我們需要一個能承載數位中文漢字的字碼架構。
前面的分析，應該能讓你感受到，數位漢字碼若要能承載中華文化中的活力，就必須具有新字詞的演化架構，因為這個質素代表著數位漢字在中華文化中能繼續具有重組與創新能力，而這些本來就是傳統漢字既有的本質機能，並且也是一個文化要能繼續生存發展所需具備的內涵要素。

這種獨特的構字能力，進一步來說，主要就是形聲造字法，這是漢字特質，也是漢字的活力和魅力所在。如果無法造新字，其結果就是迫使文辭變得冗長而生硬，因而將漸漸失去它的簡潔與優雅，除了減損了文字效率與實用價值，也會在不知不覺被逐漸的捨棄替換，最後，終將面臨被更簡潔有效的文字系統所取代的命運。

漢字是把概念分類和發音濃縮到小小的方塊內，這種二維的表達，比一維的英文字串，承載了更豐富而精緻的資訊，實在是有效而理想的文字表達方式。我們只有找出漢字在數位世界中進化的活路，才能夠讓漢字繼續保持她的實用與優雅。相較於英文，漢字的優點，其實俯拾皆是，這方面的探討很多，無庸贅述。在這裡僅舉一個簡單例來說：「鱐、鰫、鱞、魨、鰉、鱣、鱴、魦、鰷、鰗」，雖然你可能都沒見過，不過大概知道不是魚的名稱、就是跟魚有關係的事物，甚至已經可以想像，大概是屬哪一型的魚。有了魚的部首，「有邊讀邊，沒邊讀中間」，就算讀音不甚確定，也是八九不離十。反觀英文就沒這個好處，Tuna，crucian，salmon，bass, abalone，trout，scombroid，雖然都唸得出來，

但沒有事先學過，根本看不出任何關連，恐怕只有魚類學者才能弄明白真正的義涵。

結語

自從英文電腦發展以來的這幾十年來，我們進行了一場「電腦中文化」的努力。然而，在電腦普遍使用的今天，事實上我們已經漸漸地從硬體與技術的限制中解放出來，整個資訊產業正從「硬體」主導的產業轉移到由「內容」所主導的產業。因此「電腦中文化」也進入了「中文電腦化」的新階段，我們要從中文的真正本質與需求來運用電腦，而不再遷就於電腦硬體與技術。

當今的字碼，不管是BIG5、GBK、或Unicode都是人為指定的一字一碼架構，而使得數位化的漢字失去既有的生命力，不僅使得漢字變成一種僵化的文字，也使得漢字漸漸地降低了他的實用性，而由這些漢字所建構的數位內容的價值，也受到很大得限制。這樣的「歷史錯誤」是我們要嚴肅地重新審視的。

本文闡明了一字一碼的數碼架構在生活中的用字事實與其未來發展的困境，其目的在於讓我們發現，中文碼對中華文化的傳承與更新中所發生的關鍵作用。

中文碼對一個數位中華文化的發展，其影響可說是既深且遠，並且是無所不在的。在這中華文化邁入數位新世紀當中，中文字碼的架構正從根從本地影響了我中華文化的未來，希望我們能及早發現這個議題的嚴肅意義，期能引發各界深思熟慮，尋求解決之道。

作者註：本篇文章希望讓大眾發現潛藏在我們生活中的字碼是如何地影響著我們的中華文化的現在與未來。如果能獲得你的認同，歡迎轉載與拷貝，讓我們一起來等待新漢碼的未來。—2006.9

———— 全文完 ————