統一的中文電腦環境

李國鼎 題

書同文

兩岸中文環境的統一

電腦書同文的內涵與關鍵

大陸通用中文代碼國際聯合會（ACCC）秘書長
張軸材

〝書同文〞是二千年前在中華大地上實現的偉業。漢字文化歷經上下幾千年、縱橫幾萬里，得到了宏揚光大。今天，人們又在談論〝中文電腦書同文〞的理想，兩岸中文代碼統一的呼聲日益高漲，中日韓漢字統一編碼的方案激起了不小的波瀾。在此，筆者願以個人之淺見，拋磚引玉，以期引起國內外同行的深入討論，達到更廣泛的共識。

首先，我們應當嚴格界定〝電腦書同文〞的範疇。事實上，當我們談到電腦書同文時，人們往往有著不同的理解，但不外乎下面三個課題（以人們談論的熱烈程度爲序）：

電腦輸入方法的簡化與統一
電腦輸出字形的規範與統一
電腦採用字符集的統一

在英語世界，上述三點幾乎已經完全不成問題；但在漢字文化圈中，上述三點幾乎全都懸而未決。理想的〝電腦書同文〞理應涵蓋這三項內容。但這三項任務並非具有同樣的重要性。

其次，我們應當明確電腦書同文的重點，較之輸入方法的快捷與否，輸出字形的優美程度，字符集的統一具有無可置辯的重要性。字符集是輸入／輸出的基礎，牽一髮而動全身。這一點曾長期被忽視、被淹沒，是不難理解的。因爲編碼字符的問題，不是像輸入、輸出問題那樣發生於人機介面上，而是發生在機器（電腦）的內部和機－機介面上。因此，不那樣直觀，不那樣爲人們所廣泛認識。

漢字輸入的確曾經是電腦中文化的瓶頸，但時至今日，鍵盤輸入技術已有了突破性的發展，達到了實用的水平。大陸均用拼音，臺灣慣用注音；成年人易接受形碼，年青人易接受音碼；專業輸入員不惜死記硬背一套規則來換取高速，而一般用戶則喜歡簡單易學的普及型方法；而隨著時間的推移和技術的進步，按詞語輸入又形成了一股肯定的潮流。我們不應當、也不可能在短時間內强求人們使用單一的方法。只能隨其自然，在時機成熟時，逐步引導，水到渠成，達成統一。即使到那種時候，也應當有幾種方法並存共用，以滿足不同用戶的不同需求。

中文之電腦輸出，當然也有求〝同〞的問題。但我們應當認淸字符(Character)、字形、圖符(glyph)和具體的字型（font image）乃是處於三個不同層次的概念。對具體字型的劃一，是理想化的苛求；統一圖符（glyph），亦非易舉（但ISO-IEC JTC1/SC18和AFII正在推進此事）；〝電腦書同文〞的當務之急，應是具有抽象字形的字符的認同。字符和圖符之間，存在著重要的差別，即字符主要反映概念（群），圖符則偏重形態。字符與字符之差，表徵爲構件（部

首、字根）之別；而圖符與圖符之差則局限於筆劃、筆形之別。據此，簡、繁、異體字均屬於不同的字符，不存在認同問題；而只有筆形之異的圖符，則應視爲同一字符，如廿／卄，亡／亡等等。我們理想中的字符集，應當是簡繁並存、兼收並蓄的，系統應提供各種可能性，用簡或用繁，或簡繁併用，則是用戶的選擇。

字符集統一之所以重要，是因爲：

字、詞、語，字是基礎。

輸入、輸出、處理、交換、存貯、傳輸均是以字符的編碼爲對象。

以不同字符集爲基礎研發的軟體一般是難以相通的，而從這些軟體誕生的軟體、數據又將以電腦的速度呈爆炸式的增長，從而分裂出迥然不同的〝電腦文化〞，隨著時間的推移和數據的增長，日後的整合將愈來愈困難。需知，兩個字集，比20種輸入方法給用戶帶來的困難還要大。

因此，我們應明確地指出，電腦書同文之關鍵，乃是建立一個各種電腦都共同採用的一套編碼字符集。

第三，即使有了一套通用的信息交換用字符集，還應整合交換碼與內碼（處理碼）的概念。因爲，迄今爲止的各個漢字字表、字集，都只是以7bit爲基礎，以94×94的方陣，定義了漢字的有序集，稱之爲交換碼；但是卻未規定在8bit多Byte的環境下怎樣表徵漢字。這便是大陸有CCDOS式漢字內碼、IBM DBCS內碼；臺灣有Big5，TCA，IBM 5550等內碼誕生的緣由。我們期望中的字符集，應當具有交換與處理的一致性，才能保證書同文達到交互運作（Interworking 無換碼交換）的水平。

第四，書同文，文者，文字(SCRIPT)也。按照ISO的最新定義，〝文字是用於一種或多種書面語的、有鑒別性特徵的圖形字符的完備集〞。漢字便是這樣一個集合，它用於中、日、韓（CJK）三種書面語，它具有表意性、方塊性等一系列有別於其他文種的特徵。由於漢字的國際性，我們便不應囿於兩岸的代碼統一（儘管它是首要的、主要的），而應進而推進CJK 漢字代碼統一。只有這樣，才能以最低的社會成本，將電腦資訊的本土化與國際化結合起來，並使兩岸的電腦軟硬體的發展共同踏上一個新的平台。

第五，電腦書同文，單是中文同還不夠，其他文種也要同。比如：大陸的GB2312和臺灣的CNS11643中的非漢字，大陸獨有的爲395個，臺灣獨有的爲180個，兩岸相同的有290個。其中，有的是標點符號，有的是希臘文、俄文、日文假名、數學符號等等，這些字符在中文爲主的文體中，也是所在多有的。如不爲此尋求一個共

同的并集，電腦信息的交換仍然會遇到不少困難。

第六，電腦書同文絕非空想，而是在可見的將來經過努力可以實現的理念。我們正面臨著重要的機會與挑戰，這就是在ISO10640多幾位(Multi-Octet)編碼字符集中，實現兩岸中文代碼的統一，乃至CJK代碼的統一。ISO DOS 10646的基本中文平面，爲漢字提供了兩萬多個碼位(Cell)，只要妥善安排，將兩岸、日韓常用、次常用的漢字排列進去，是完全可能的。主動權在很大程度上掌握在我們手裡，而不是外國人手裡。同時，基本多文種平面 (BMP)也爲統一的 CJK 漢字係留了空間，統一的字符集是在基本中文平面實現,還是在未來的BMP中實現？這是我們自己的選擇。一年前，當我們提出CJK漢字統一編碼的HCC(Han Character Collection)方案時，日本朋友曾〝斷然拒絕〞，韓國朋友亦不甚理解。但是，今天，日本和韓國都宣佈了參加CJK-JRG（中日韓聯合研究組），並爭當首次會議的東道主，這不能不說是一大進步。與此同時，在兩岸資訊業專家推動下建立的多字符集漢字庫正在日夜運轉，實施〝基於屬性的機助認同／甄別〞分析比較兩岸、日韓的字表、字集。而Unicode Level I 漢字字表的發表，又刺激和加速了這一進程。

饒有興味的是，ISO 工作組爲10646起了一個新的名稱:UCS-Universal Character Set（姑且譯作泛用字符集）。儘管UCS至今仍不盡人意，但它終究指出了一個方向，就是通用與統一，這與電腦書同文的理念是一致的。多幾位(Multi-Octet)，只是編碼方式，多文種、多用途（不只是交換）才是目標。我相信兩岸的同行，一定會克服重重困難，求大同，存小異，繼續合作，在官方、業者和廣大用戶的支持下，實現這一目標，爲電腦書同文奠定牢實的基礎。