與文字共舞－中文數位化發展簡介

電腦科技創造出新的文明世代，國內中文碼的發展，亦造就出揚名於世的資訊產業與生氣蓬勃的數位化環境。從文字發展的介紹到中文碼發展歷程的描繪，以及數位化技術的闡釋，本文將與您一同分享這份由歲月與智慧所凝聚出的甜美成果。

壹、前言

中文數位化技術對大多數民眾甚至是科技專家而言，或許是個既陌生又專業的領域，早期中文電腦技術尚在萌芽階段，行政院主計處電子處理資料中心便率先研究發展中文相關的軟硬體應用系統，奠定了日後中文數位化發展的基礎，在政府與民間的合作下，發展出諸如政府大型資訊系統所引用的戶役政 EUC 碼、稅務碼、電信碼、圖書館專用的 CCCII 碼等中文化編碼，民間資訊業者研訂的 Big5 等，以及在國際上相繼出現的 ISO10646、Unicode 等標準碼，這些都是以內碼的型式存在的，而都已經普遍應用於現有的資訊環境中。至於「內碼」一詞，它就好像是電腦世界中的方言或母語一樣，在同一個內碼環境中，彼此之間可直接與立即的溝通。

爲因應國內各種中文資訊交換與整合不同自碼系統的殷切期盼下，中文標準交換碼 CNS11643 孕育而生，從摸索到成熟，在產、官、學、研各界的大力投入下，造就了國內的中文資訊環境整合的契機，也成爲國際漢字標準參酌的重要典範；而所謂的「交換碼」類似電腦界的官方語言，用以銜接不同內碼環境，使得溝通不致於產生障礙與隔閡。

因此身處於幸運的這一代，已承傳並享受著前輩辛勤得來的果實，這些手胼足胝的歷程是值得喝采的，就讓我們懷著一顆感恩的心來細數著這一幕幕美麗的故事。

貳、文字的演進軌跡

文字是文明的推手，不但活化了民族的生命，也創造出認同的歸屬，匯集著老祖先的智慧與生活經驗。

《易經・繫辭》記載著「上古結繩而治，後世聖人，易之以書契，百官以治，萬民以察，蓋取諸夬。」，不論是國家重大事務或是生民百姓的瑣碎家常，有了結繩這種小而美的記載方式，老祖宗們嘗試將溝通做系統化的整理；如果將結繩視爲一種訊號的話，廣義來說實與現今數位化方式相仿。

「倉頡之初作書也，蓋依類象形，故謂之『文』。其後形聲相益，即謂之『字』。文者，物象之本；字者，言孳乳而寖多也。著於竹帛謂之『書』。因此『文字』被創造出並保留在『書』中。甲骨文、金文、玉石文字、簡帛文字、孔壁古文、籀文、篆文、隸書等文字在歷史上相繼出現，成爲現今世上保存歷史最悠久、一脈相傳的文字體系。而秦始皇一統天下後，篆文及隸書全國通用，開創出文字標準化的新紀元。

另一個文字發展的關鍵是印刷術，加上蔡倫的蔡侯紙公諸於世，使得文化傳播的方式從私塾獨授的簡帛時代大步跨越到的智慧分享的紙張時代；到了宋仁宗慶歷年間畢昇發明了膠泥活字版，更促使民智大開。因此，文字擺脫傳統人工撰寫手抄既緩滯又辛苦的方式，知識得以大量的複製與流傳，文明的腳步也加快了。

參、電腦數位化之發展

電腦數位化的發展造就了新的人類文明，如排山倒海般的建立了人類新的願景。

但要如何讓只能分辨 0 與 1 訊號(1bit)的電腦能處理文字呢？那就必須有文、數字的編碼系統了，將人類語言中的字符(Characters)轉換電腦內部能處理的數位型態。1960 年代初期，美國國會圖書館制訂了英文的字元集和交換碼，作爲美國圖書館界書目交換的共同標準，這便是 ASCII(American Standard Code for Information Interchange)編碼系統的前身，ASCII 用 7 個位元(範圍 0~127)共 128 個字的編碼空間，將英文 A~Z、數字 0~9 及其它符號數位化，賦予其唯一的編號，像「A」在電腦內的代碼爲 41H，00H 到 1FH 位置編爲控制字元，20H 到 7FH 位置編爲圖形字元，作爲電腦之間交換文、數字資訊之用。基於互通原則，國際 ISO 組織(International Organization for Standardization)亦將此訂爲 ISO 646 標準，現今電腦所處理的純文字文件便是以 ASCII 爲基礎字集。

歐語或亞洲語言之編碼系統則需要更多字元來闡述文字，因此歐洲方面也發展出一套 8 位元的 Latin-1 字碼規則，以 ASCII 爲基礎，增加常用字元到 256 個字。IBM 公司就將 80H 到 0FFH 共 128 個字元編入框線、音標和其他歐洲非英語系的字母，一般稱爲 EBCDIC(Extended Binary Coded Decimal Interchange Code)。於是，逐漸形成現今電腦以 8 位元(1 位元組)來存取字元集(character set)的運作方式。

肆、中文碼的演進

由字母所發展出的「拼音文字」與我們一圖一字所呈現的「表意文字」，西方世界與東方文明對於字的詮釋有著不一樣的方式與表現，所以不論是 ASCII、EBCDIC、Latin-1 等 8 位元即可處理的字元集，無法處理爲數大量的表意文字，而到目前爲止，缺乏一種具體有效的方式，能將表意圖像文字拆解成類似拼音文字中字母的基本法則，做爲數位化的處理編碼依據，因此採用兩個以上位元組來描繪出編碼空間的大字集因應而生，擁有 13053 個字的大五碼(BIG5)如此，持續擴編新字目前九萬多字的國際通用編碼(UNICODE)亦如此，諸多這類電腦系統中所謂的「內碼」，因其不同的編碼，反而使得數位文件在網路發達的時代造成流通之不便，因而衍生了標準化的「交換碼」。國家標準交換碼「CNS11643」孕育而生，行政院主計處電子處理資料中心也應用資訊技術，研發出「全字庫」營運平台提供國內無障礙的中文文字交換環境。

早在民國六十一年，由於當時國內的資訊環境尚屬萌芽階段，行政院主計處電子處理資料中心便開始研發中文軟硬體的應用系統，像中文表報輸出系統(CROS)、字根輸入法中型中文鍵盤、第一代中文終端機、中文線上作業系統等，促使各界對發展出國內專屬的中文化電腦環境有了進一步的認知，並有初步的共識，六十七年行政院就中文電子計算機系

統之發展成立專案研究小組，六十八年在國家建設研究會議中，與會的專家學者在討論後達成編碼的共識，行政院主計處電子處理資料中心乃據以編定『中文資訊交換標準碼』之施政計畫，於是展開編定中文字碼的工作，亦開啓了中文電腦發展與應用的嶄新時代，民國六十九年由行政院國家科學委員會所舉辦的溪頭會議，各專家、學者達成國家中文資訊標準交換碼編碼原則並報院核定，翌年行政院函令國科會、教育部、中央標準局及行政院主計處電子處理資料中心組成專案作業小組，持續推動編碼工作。(另六十九年爲因應當時國外電腦處理東亞語文資料的需求，便由國字整理小組規劃並頒布 CCCII(Chinese Character Code for Information Interchange)以供使用，至今仍有許多圖書館系統採用。)民國七十一年編定常用字碼；民國七十二年行政院資訊推動小組成立編碼技術作業小組，針對已定之編碼原則完成『通用漢字標準交換碼』，並決議試用二年。

七十四年，國科會與行政院主計處電子處理資料中心邀集各相關單位與業者組成技術小組，檢討試用結果、進一步修訂編碼原則後重編，七十五年獲行政院核定，正式公布實施。同年中央標準局審定頒布爲國家標準，編號「CNS11643」；八十一年該局再因應各界之需要，由原兩個字面共 13,051 個字，大幅擴編爲七個字面共 48,027 個字， 公布並更名爲『中文標準交換碼 (Chinese Standard Interchange Code)』。現爲整合全國各資訊系統與用字上的需求，其字面數已擴充到十五字面，除增納國內諸如戶役政、工商、公路監理等重要行政系統的用字外，亦將國際標準編碼 ISO10646 中各國的拼音文字、CJK 字集收納其中。標準檢驗局預定於今(九十五)年編審公告 CNS11643 最新版本，總字數將高達九萬多字；同時，新版中亦規劃擴增字面至八十字面，用以搜納更多如古代漢字、各民族用字等的中文字。

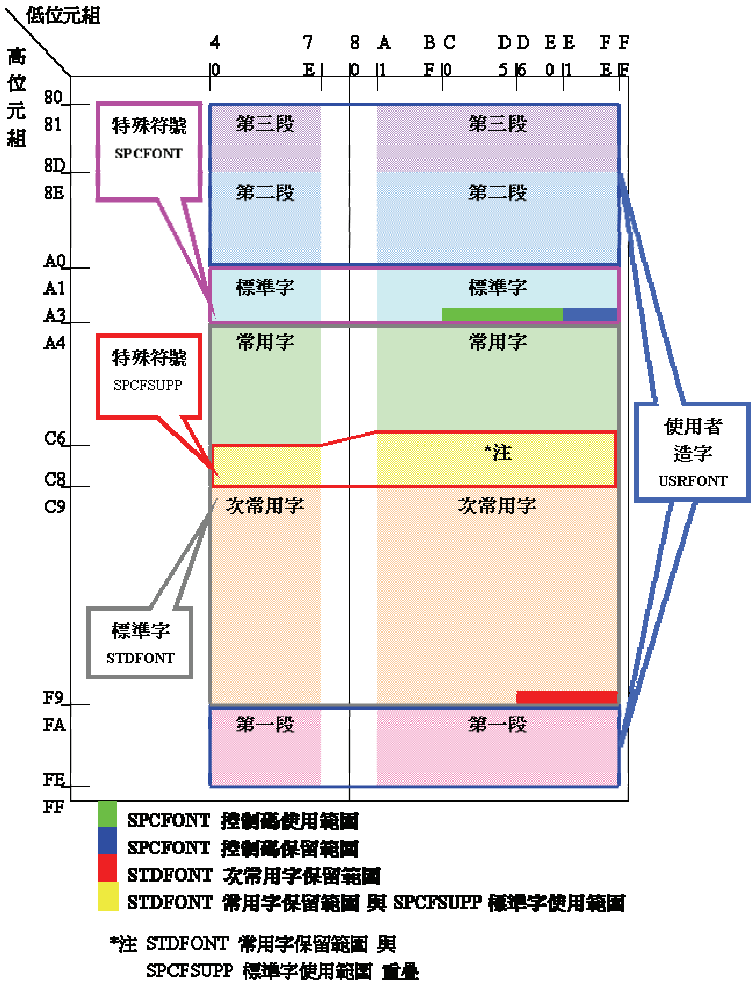
伍、中文碼的介紹

一、大五碼 (Big5)

民國七十二年出現了第一部擁有處理漢字功能的個人電腦—IBM5550，加速了國內發展中文電腦的熱潮。財團法人資訊工業策進會與國內 13 家業者合作進行的「五大軟體專案」，在政府的大力支持下，資策會以「通用漢字標準交換碼」之前身「常用字碼集」爲藍本，發表了一套專爲五大中文套裝軟體所設計的中文內碼，Big5 遂成爲我國中文電腦的業界標準。

Big5 內碼擁有 13,053 個中文字、408 個符號及 33 個控制字元的字集，雖引領風騷二十年，但由於無專責機構負責維護，微軟、倚天、宏碁等中文系統廠商基於擴充上的需要，推出了不同版本的 Big5 碼。政府有鑑於不同版本的 Big5 對於國人已產生使用上的困擾，於是經濟部標準檢驗局在民國 92 年委託財團法人中文數位化技術推廣基金會修訂 Big5 編碼字元表，正式定名爲「Big5-2003」。由於原 Big5 碼 13,053 個中文字中，發現『兀』字與『毉』字重複編碼，所以此版本調整後的總字數爲 13,051 個中文字，同時也新增了 30 個數字符號、24 個部首、14 個罕用符號、268 個日本假名，以及 34 個表格符號共 370 個符號。

又行政院主計處電子處理資料中心鑑於政府機關公文及資訊系統進行電子傳遞時，因各單位擴充 Big5 碼之自造字而無法交換處理，協商行政院研究發展考核委員會成立「Big5 碼字集擴編計畫」專案處理，搜納公務部門在一般文書上最常用之自造字加以編訂，民國八十六年七月擴編完成「Big5＋碼」。又爲能使該字集能在資訊業務上正常運作，於是從 Big5＋中挑選 3,954 個字編訂 Big5 碼造字區碼位，完成 Big5 碼補充字集(Big5 Extension Character Set)的建立工作，亦即是公務上所熟知的 Big5E 字集。

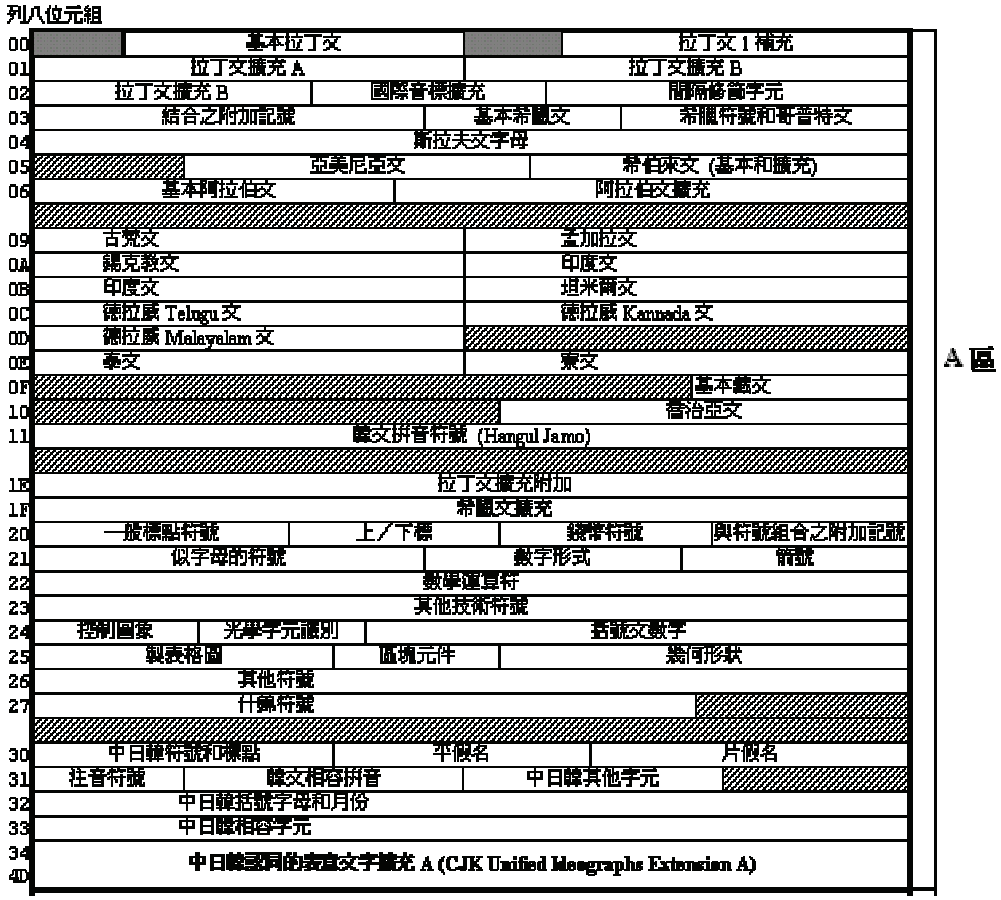


二、國際通用碼 (Unicode)

各國爲處理資訊化的工作大多制訂了本土使用的字元碼。其依各國語言的字元集的大小，決定採用單位元組或多位元組的編碼方式。單位元組字元碼多採 ISO/IEC 8859 系列的八位元單位元組字元碼，多位元組交換碼則多遵循 ISO/IEC 2022 的編碼結構。對於國際數位化資料的交流，無異衍生出諸多的困擾，於是有了發展多語言整合性字元集的共識。在七十三年，國際 ISO 組織正式開始制定國際字元集編碼標準。此項工作交由 ISO/IEC JTC1/ SC2/WG2 工作分組負責，編訂出 UCS(Universal Multiple-Octet Coded Character Set)，編號訂爲 ISO/IEC 10646。

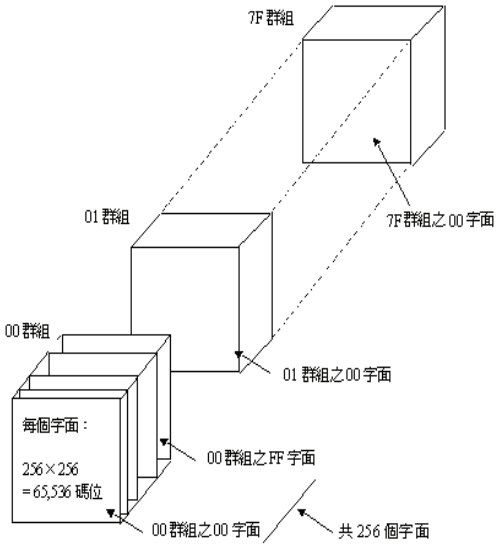
七十七年 Xerox 公司建議以將電腦字元集編碼的基本單位由現行的七或八個位元擴充到 16 個位元，利用 2^{16} 多達 65,536 個碼位容納全世界各種語言的字元和常用符號。新的字元集編碼標準被稱爲 Unicode。八十年由 IBM、DEC、Sun Micro、Xerox、Apple、MicroSoft、Novell 等公司共同成立 Unicode 協會(The Unicode Consortium)，並由 Unicode 技術委員會(UTC，Unicode Technical Committee)從事各國字元蒐集、整理、編碼等工作。於是在同年發表第一版(Unicode 1.0.0)的 Unicode 標準。之後 WG2 與 Unicode 協會達成協議，將 Unicode 併入 UCS 的 BMP(Basic Multi-lingual Plane 如圖)字面，並將字元的搜集、整理和編碼等工作交由 WG2 負責，所以後來 UCS(ISO/IEC 10646)與 Unicode 共同闡述同一個字元集。

UCS 亦即 Unicode 共搜錄拼音文字、表意文字、各種符號和控制字元等四種字元。其中 WG2 將表意文字蒐集、整理與比對工作，交由其下所設之表意文字書記組(Ideograph Rapporteur Group, IRG)專責處理。在 ISO 中所認定的表意文字，係指源自於中國，東亞各國正在使用或曾經使用過的漢字，包括台灣、中國大陸、日本、南北韓、越南、新加坡和港澳等地區，因此 IRG 成爲我國與國際標準編碼組織最爲重要的聯繫窗口，目前正透該組織申請國內所新增的兩萬多個戶政用字，同時我國代表也爭取到古漢字小組秘書組一職，對於參與國際活動及強化各國情誼用力良多。



UCS(或稱 UCS-4）採連續編碼，不再避開 C0 和 C1 控制碼區，以 32 個位元爲基礎的編碼方式，並劃分成四組八位元，以群（G-octet）、字(P-octet)、列(R-octet)和格(C-octet)，分別代表編碼結構中的群組（group）、字面（plane）、列（row）與格（cell）

的關係(如右圖)。每一群組由 256 個字面所組成，每一個字面由 256 列所組成，每一列則包含 256 格，每一格爲一個碼位。兩個碼位 FFFEh 和 FFFFh 保留不用。所以整個編碼空間總共 256×128 共 32,768 個字面，每個字面爲 256×254 共 65,534 個碼位，整個 UCS 可訂出 65534×32768 共 2,147,418,112 個碼位。目前僅有第 0 群組的第 0 字面的基本多語文字面(BMP，Basic Multi-lingual Plane)、第 1 字面、第 2 字面和第 14 字面真正收容編碼字元。



三、國家標準交換碼 (CNS11643)

在中文內碼或中文作業系統中，就現實環境而言一直存在著系統間中文資訊無法直接交換的問題，甚至在同樣的資訊環境內也會有因自造字而產生的同碼異字的現象，更加深數位資料交換處理上成本的虛耗。因此，解決中文資訊交換所衍生出的中文交換碼，便成爲大家一致的期待。國家標準交換碼 CNS11643 就在國人的殷切期盼下誕生了，也在大家的努力下逐漸成爲國內最大、最完整的字集，進而成爲國際表意文字編製內容的主要供應來源。

CNS 11643 第一版在七十五年正式公布實施，共收錄了 13,051 個中文字，分屬 1,2 兩個字面。由於施行起來無法滿足各電腦化業務所需，於是行政院主計處電子處理資料中心開始彙整各機關自造字，在八十一年由標準檢驗局公佈第二版，並更名爲「中文標準交換碼」(Chinese Standard Interchange Code，以下簡稱爲 CSIC)，總共搜納 48,027 字。此外，爲使 CSIC 與 ISO/IEC 10646 字集能同步化的編制，九十二年經濟部標準檢驗局再度著手擴編 CNS 11643，第三版的編碼字面數由第二版的 7 個字面擴編爲 15 個字面，中文字集也從原來的四萬八千餘字增加到約九萬多字。另爲一勞永逸的解決可編碼字面容量飽和的問題，CNS 11643 第三版也將編碼空間由第一版原先所規定的 16 字面增加到 80 個字面，預計九十五年公告施行。

CNS11643 遵循 ISO/IEC 2022 之規定，採兩位元組編碼，字元碼每一位元組使用 21h～7Eh 的 94 個編碼位置。因此，編碼空間爲 94 個字列，每字列 94 個字格（亦即碼

位），每一字面總計 94×94 共 8,836 個碼位，第三版增訂爲 80 個字面共 706,880 個碼位。

值得一提的是在各界對於中文碼整合工作寄予厚望下，行政院主計處電子處理資料中心承諾起這個重責大任。結合了政府與民間的力量，在多年研究與努力，建構了「全字庫」資訊服務，其爲包容與實踐了國家標準字碼集的資訊平台，用數位的技術來銜接日常生活與資訊系統間對中文用字同步的需求，發展至今，除字碼、字型及字詞屬性整合外，也引進新的數位觀念，創新研發文字互轉的機制與字碼應用方案，近年來更在電子化政府與 e 化台灣等重大政府資訊計畫中，提供了中文資訊互轉介面，將各種不同內碼、不同造字、不同系統中所產製的資料，透過該系統數位化處理，能正確與完整的完成各資訊系統間資料之互通，期待在全字庫所提供的各種數位化服務催化下，真正落實中文資訊環境的普及與無障礙。

陸、結語

資訊數位化的發展，是現階段人類文明發揚的推手，也是智慧進化的動力，文字演變的過程也述說著一代代對於傳統文化的那份堅持與理想，當科技與文化古今交會時，雖有著交融時的衝突，但更期待著是兩者的互利共生；在漫漫歷史的演進中，從結繩到數位編碼，對於文化生命的延續與民族圖騰的傳承，觸動著使命與挑戰的原始天性；不論是 Big5、Unicode 或是 CNS11643....，除保有在機器瞬間的脈絡邏輯中，更期待能烙印出深邃的文化內涵與傳承精神，讓文字見證著歷史的榮衰與成敗，希望、未來都是浪漫又美麗的榮耀。

【參考資料】：行政院主計處電子處理資料中心全字庫網站：http://www.cns11643.gov.tw

（本文由行政院主計處電子處理資料中心管理師余保倫 提供）

（本文轉載自主計月刊 95 年 6 月號）