電腦中文碼闡述

◎馬瑪莉（行政院主計處電子處理資料中心分析師）

電腦中文內碼是中文在資訊處理系統內部最基本一種表達型式，做爲儲存、處理等用途，本文係介紹電腦中文碼發展過程。

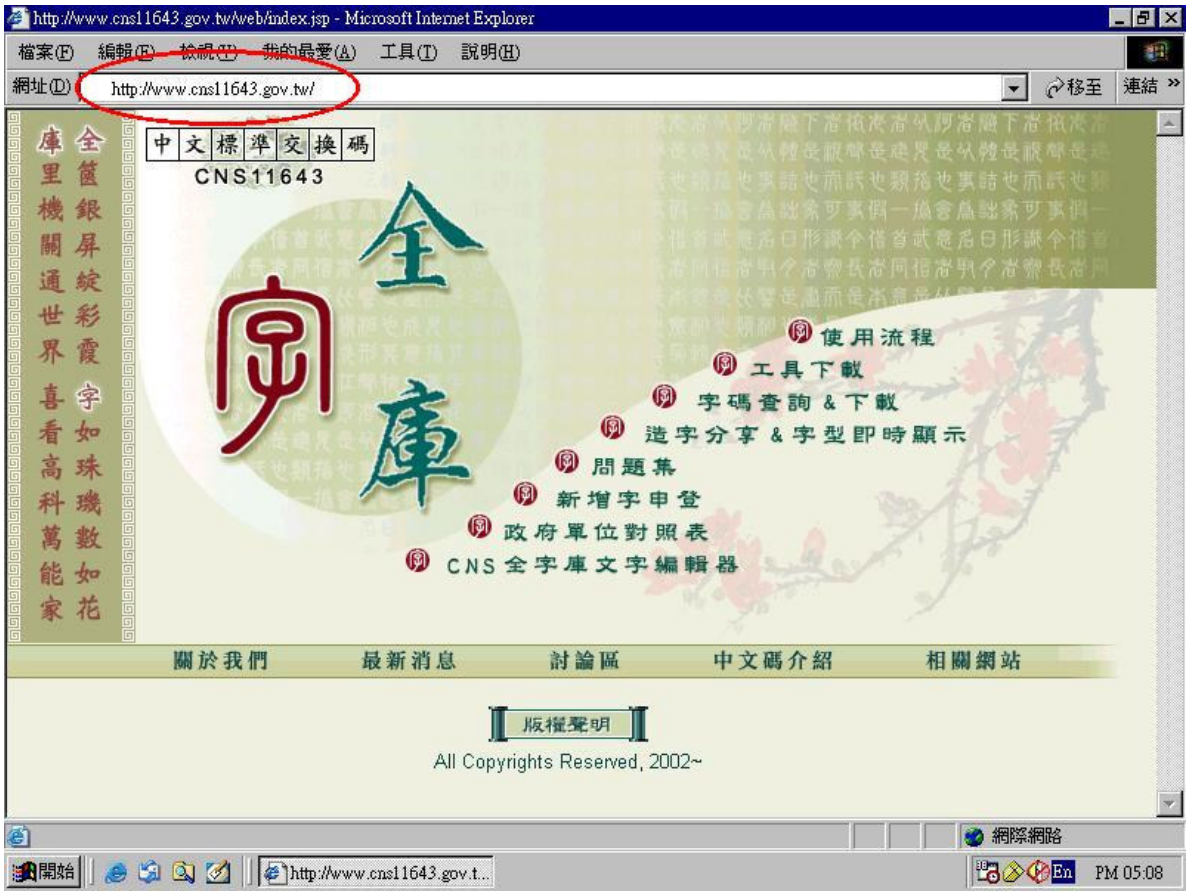
電腦文字處理系統實作必須具備：鍵盤碼、字碼和字型檔三樣東西。鍵盤碼是將鍵盤的按鈕轉換成字碼的橋樑，西文都是採用這種方式，東亞漢字語言則因漢字字數太多，無法採用按鈕直接對應字碼的方法，必須另外設計軟體將較複雜的按鈕轉換成字碼，這種軟體叫輸入法編輯器(input method editor，簡稱IME)，它的設計隨輸入法而定的，例如倉頡、注音、嘸蝦米等。字型檔則將字碼和字型連接起來，讓螢幕和列表機將字型呈現出來。以下介紹電腦中文碼之發展過程：

□電報碼：台灣最初引進電腦時，電腦硬軟體沒有中文系統，都係借用電報碼處理中文，電報碼原先是爲電報處理中文而編的碼，用四位數代表一字，因此一萬個編碼位置，實際上收入的只有八千多字。財政部財稅中心是最早使用電報碼的單位，但因爲四個位元組裡面每個位元組的256個可能編碼位置只用了十個，浪費電腦儲存空間，減低處理速度，不適用於大量文字的資料處理。如何取代電報碼，中文字碼究竟應採用二位元組或三位元組，產生相當大的爭議。二位元組的擁護者認爲字集裡有一萬多個字就很夠用，而且在技術層面上困難度和成本都低很多，獲得政府、工商業的支持，於是台灣前後有好幾套二位元碼出現。

□大五碼（Big5）：1984年資策會和五大電腦公司合作推出之Big5字集，有1萬3,053個字，推出後絕大多數的個人電腦都相繼採用，用字不夠時就在加字區內造字，因爲加字未統一，所以加字區的碼有些混亂；1986年中央標準局公佈CNS11643碼，字集和Big5碼同樣是1萬3,053個字，但字序則不一樣。上述這些碼和其他二位元組碼包含字數都在一萬三千字左右，使用者造字區五千字左右，因此可和ASCII英文碼混合使用。

□中文標準交換碼（CNS）：1980年9月，行政院國家科學委員會召集國內編碼專家學者開會達成初步協議，並報請行政院核定「國家中文資訊標準交換碼編碼原則」，翌年9月行政院函令國科會依據選定之原則，邀集教育部、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，積極推動編碼工作。1982年7月曾編定常用字碼一種，但所收字數不夠；1983年5月另組成編碼技術作業小組，進行編碼細則研討，10月底完成「通用漢字標準交換碼」，並決議試用二年；1986年3月獲行政院核定，正式公布實施，8月並獲得中央標準局審定頒布爲國家標準（編號CNS11643）；1992年5月中央標準局再應各界需要，由原２個字面 (1萬3,051字) 擴編爲７個字面（4萬8,027字），並更名爲「中文標準交換碼 (Chinese Standard Interchange Code)」，詳細資料請參閱國家標準碼全字庫網站(http://www.cns11643.gov.tw)中文

碼介紹單元。



□**EUC碼：**國內大型資訊系統所應用之字碼，以戶役政系統爲代表，戶役政系統建構在UNIX系統上，屬於主從模式架構，內碼採用UNIX系統之EUC碼。EUC碼雖與CNS11643長度不同，但卻採用CNS11643之編碼架構及字集。

□**CCCII碼：**1980年謝清俊教授和一群學者爲圖書館界發明CCCII（Chinese Character Code for Information Interchange）中文碼。最初編出4,808個字，其餘字碼分批陸續發佈，到1987年共編定5萬3,940個字。中央圖書館（國家圖書館前身）最早使用此碼，然後各大館陸續加入，目前國家圖書館、各國立大學圖書館以及台北、高雄、台中等三大公共圖書館都使用CCCII碼。CCCII用三位元組代表一個字，有八十幾萬編碼點，編碼空間很大。CCCII在當初設計時就被美國研究圖書館組群（Research Library Group；RLG）決定採用其架構和借用一部分字碼（只有1萬5,000字左右 ），並將日韓文編入該系統，這個版本即是美國國家標準Z39.64，稱East

Asian Character Code（簡稱EACC）。但CCCII經過多年使用曝露出二大缺點，一是重複字碼(即一字多碼)多達萬餘；另一缺點是它的三位元組架構不適用於現行Windows 95/98/2000。

□**BIG-5E字集：**1997年爲協助解決眾多使用BIG-5碼之政府單位，於進行公文電子傳遞時之自造字（Big5的13,053字以外自行造的字）無法轉換問題，決議辦理「BIG-5碼字集擴編計畫」，但擴編完成之「BIG5+碼」未獲多數廠商採用。不過擴編計畫所完成之「標準字集」，即政府單位一般文書最常用之自造字，如將其應用於BIG-5碼的造字區，可整合使用者常用的自造字、降低轉碼的頻率。於是由BIG-5E碼之「標準字集」中選取3,954個字，在BIG-5碼的造字區中建置「BIG-5碼補充字集(BIG-5 Extension Character Set，簡稱BIG-5E字集)」，並配合行政院「電子化／網路化政府計畫」之推動，於公文電子交換作業規範中訂爲可處理中文碼類別之一。

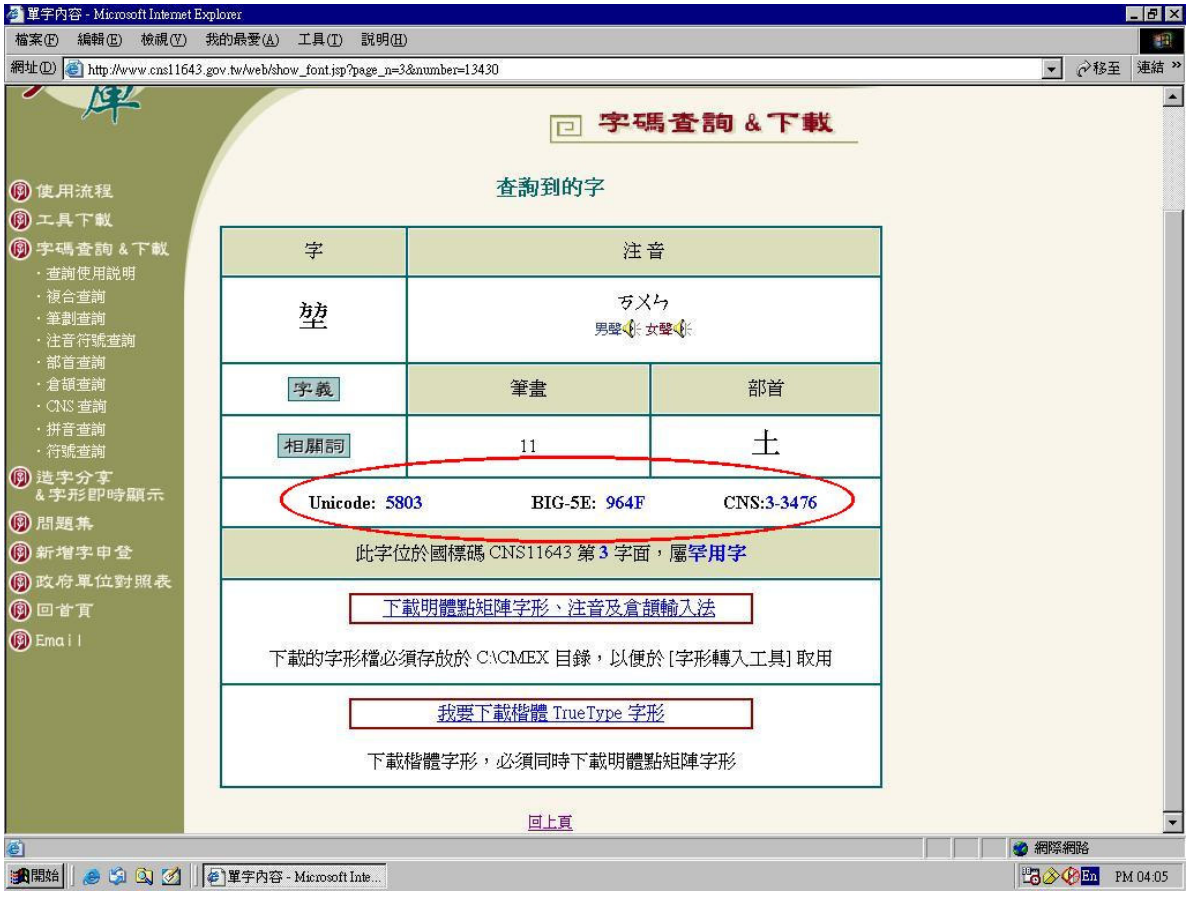
□**ASCII碼：**1960年代初期，美國國會圖書館(Library of Congress, LC)開始研擬機讀編目格式，同時也制訂了英文的字元集和交換碼，以做爲美國圖書館界書目交換的共同標準，LC交換碼隨後發展成爲美國的國家標準ASCII碼 (American Standard Code for Information Interchange)，並且還進一步演變成世界性的電腦字元編碼標準ISO646（其全名爲7-bit coded character set for information interchange）。時至今日，雖然一個位元組(byte)的長度已經從7個位元(bit)增加爲8個位元，ASCII和ISO646仍然是電腦與網路世界裡最重要的字元碼標準。

□**Unicode碼**：爲容納全世界各種語言的字元和符號，ISO的一些會員國於1984年發起制定新的國際字元集編碼標準，新標準由工作小組ISO/IEC JTC1/SC2/WG2負責擬訂（簡稱WG2），該小組所提出之ISO10646草案初稿一經公佈，其編碼結構立即遭到美國部份電腦業者的反對。1988年初，美國Xerox公司倡議以新的編碼結構，另外編訂世界性字元編碼標準：將電腦字元集編碼的基本單位由現行的7或8個位元一舉擴充爲16個位元，並且充分利用6萬5,536個編碼位置以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命名爲「Unicode」。

□**ISO10646碼：**WG2集各國專家之力共同整理全世界古今各種語言文字和符號，制定ISO10646編碼，並依語言特性區分爲表意文字和非表意文字兩類，表意文字其實發源於中國的漢字，主要使用於台灣、中國、日本、南北韓、越南、新加坡和港澳地區，除漢字之外的所有其他文字，一律歸類爲非表意文字，絕大部分爲拼音文字。非表意文字或符號，因爲字集小或是只有某個國家使用，通常直接在WG2會議上討論即可，但漢字字集規模龐大且爲多個國家或地區共同使用，則由WG2爲此設置表意文字書記小組(Ideograph Rapporteur Group, IRG)專責收集並予以編碼。漢字難免有些字型相同或極爲近似，爲了避免ISO10646編碼表出現重複字之困擾，IRG制訂了表意文字認同規則，凡是依規則應予認同的漢字，一律合併成一字賦予一個編碼，例如，我國中文碼國家標準CNS11643的字集裡就收編了兩個極爲相似的「圖」字，分別爲1-6837h和6-5B5Bh不同編碼，這兩個「圖」字，依認同規則必須合併爲一個，於是後者被前者認同掉了；詳言之，CNS的1-6837h和6-5B5Bh都對應到Unicode的5716h

（或ISO10646的0000-5716h），但是Unicode的5716h卻只對應到CNS的1-6837h。當我們把資料從CNS碼轉換成Unicode，再由Unicode轉回CNS碼時，將發生6-5B5Bh→5716h→1-6837h的結果，這種現象稱爲去回轉碼(round-trip conversion)錯誤。解決之道無他，必須在Unicode或ISO10646字集中多加上被認同掉的字元並另外賦予編碼（稱爲相容字元），做到CNS字集與ISO10646字集一對一。ISO10646第2字面所收容的CNS相容字集，就是我國爲了達到正確去回轉碼的目的，歷經多年力爭的成果，換言之，CNS11643現有中文字集已全數編入ISO10646編碼表中。

目前通用於我國的中文標準交換碼 CNS11643、業界標準內碼 Big-5，以及國際標準的 ISO10646 (Unicode)，其字集雖不相同，但實際上都有密切關連；Big-5 內碼的字集與字序，及國際標準的 Unicode 字集，都出自於標準交換碼 CNS。故無論是國內大型系統字集的需求，或國際上亞洲語系表意文字集（漢字）的擴充，都應以 CNS11643 的維護與擴充爲主，才能顧及整體性和涵蓋性。



【電腦中文碼小辭典】

字面：由縱、橫兩軸線(代表高、低位元組)所組成的一個平面中，每一個交叉點都是一個編碼位置。一般七位元環境下的一個字面，共有 94x94 個編碼位置；而雙八位元組環境下的一個字面則有 256x256 個編碼位置。

字元：即俗稱的「字」，是一組已獲認可之符號，做爲資料的構成或表示。字元可以是字母、數字、標點及其他的符號，常以分開或連接在一起的型態來表示。

字集：由不同字元所組合成之有限集，它們係一完整且已被認同的字元集。一般均採用已具共識的相同字集，再依不同需求訂定不同的字碼。例如：Big-5、EUC、TCA 碼的字集均相同，都採用 CNS11643 中文標準交換碼的字集。

字數：一完整字集中的全部字元的總數。例如：CNS11643(1~7 字面)中文標準交換碼共有 441 個符號和 4 萬 8,027 個中文字；Big-5 碼共有 441 個符號和 1 萬 3,053 個中文字。

字碼：表示每個字元的編碼。依照一套固定的規則，針對指定的中文字集內的每一個字或符號，編訂相對應的代碼，以方便電腦資訊之處理與應用。較常用的字碼表示方法有兩種；一種爲十進位數字的表示法，另一種爲十六進制表示法。

字序：一個完整字集中每個字元排列順序時所依據的方法。例如：CNS 的每個字面均依照先筆畫，後部首爲序；電信碼則是依先部首，後筆畫的字序排列。

字型：字元表現的型態，也就是以電腦能懂的方式，將字的型體表現在週邊設備上。例如：DOS 多使用點矩陣字型，Windows 環境則多使用向量外框字型。