作業報導

●電腦中文應用環境平台—CNS11643 中文標準交換碼全字庫介紹

國家發展委員會助理設計師 黃柏盛

壹、編碼簡介

字元或符號必須編碼後才能被電腦處理，在不同時期發展不同之平臺及應用而有不同的編碼系統，早期的電腦系統使用 7 位元的 ASCII 編碼，為了處理漢字，於是有了用於簡體中文的 GBK 和用於繁體中文的 Big5。國內目前常見與中文有關的編碼如 Big5、EUC、 Unicode 及 CNS11643 分別說明如下：

一、Big5 碼

Big5 又稱為大五碼，是使用繁體中文最常用的電腦漢字字符集標準，1 個字(符)編碼長度固定 16 位元(2 個位元組)。最早由資策會於 1984 年為五大中文套裝軟體所設計的中文共通內碼，當時各資訊廠商皆以此為內碼發展軟體，也以此延伸多個版本。後於 2003 年由財團法人中文數位化技術推廣基金會接受經濟部標準檢驗局委託，召集國內業者代表、專家和學者，就 Big5 編碼字元表原始版本和各主要業界版本予以重整之最新版本，即 Big5-2003。

Big5 包含教育部常用字及次常用字共 13,053 中文字，因各機關有自造字，經「Big5 碼字集擴編計畫」擴編完成「Big5+碼」後再從其中選取 3,954 個字編於 Big5 造字區，即為 Big5 碼補充字集(Big5-E)。

二、EUC 碼

EUC 碼為 UNIX 作業系統使用之內碼(Extend Unix Code，EUC)，字碼長度為 4 Bytes，主要用於表示及儲存漢語文字、日語文字及朝鮮文字，在國內比較少使用，但為前一代戶役政資訊系統使用之內碼。

三、Unicode 碼

Unicode 發展由非營利機構統一碼聯盟負責，其對世界上大部分的文字系統進行了整理、編碼，使得電腦可以用更為簡單的方式來呈現和處理文字，又稱萬國碼或統一碼，現為電腦業界標準。目前分為 17 組編排，每組稱為字面或平面（Plane）。第 0 字面為基本多文種平面(簡稱 BMP)，編碼範圍為 0000~ffff。中文收納於 Unicode 第 0 字面及第 2 字面，第 2 字面為表意文字補充平面，編碼範圍為 20000~2ffff。另有保留作為私人使用區(Private Use Area，簡稱 PUA) 可供自造字使用置於第 0 字面及第 15、16 字面。

統一碼變換格式(Unicode Transformation Format，簡稱 UTF) 即把 Unicode 字符集的抽象碼位對映為 8、16 或 32 位元字(符)碼的序列(UTF-8，UTF-16)，用於資料儲存或傳遞。1 個字(符)的 UTF-8 為變動長度自 1 個位元組(byte)至 6 個位元組表示，UTF-16 為 2 個位元組或 4 個位元組，UTF-32 為固定 4 個位元組。

四、CNS11643 碼

CNS11643 為經濟部標準檢驗局所審訂的國家標準之中文交換碼，用於中文資訊處理系統資料交換及數位通信系統的資訊傳輸。其標準編號為 11643。編訂 1 個字符之字碼為 2 個位元組，每一位元組以 16 進位自 21 至 7E 表示，即編碼範圍自 2121 至 7E7E 為 1 個字面，每一字面為 94*94(16 進位 21 至 7E 長度為 94)共 8836 可用編碼。該標準編訂第 1 字面至第 80 字面。目前公布使用至第 15 字面。

貳、CNS11643 中文標準交換碼全字庫介紹

一、緣起

由於漢字多變化且難以計數，在電腦系統中並無法收納所有漢字，因此為建設我國的中文電腦應用環境，解決個人電腦中文字數不足、自造字交換、機關或企業組織團體內部同字不同碼及網頁上罕用字顯示等問題，最早由原主計處電子處理資料中心開發「CNS11643 中文標準交換碼全字庫」(簡稱全字庫)網站，提供個人電腦上中文字集自造字決解方案及管理工具，隨著組織改造作業，於 102 年起全字庫網站的維運改由國家發展委員會負責。

全字庫目前已有納編字形(符號)約 10 萬 7 千餘字。字形納編來源有教育部標準常用、次常用、罕用字、部分異體字及閩客母語，另包含 CNS14649 各種語文字母、符號及 CJK 認同表意文字，並包含戶政、地政、經濟部工商及財稅等用字。現行全字庫可對映至 Unicode5.0 版，分別對映至 Unicode 第 0 字面 39,143 字及第 2 字面 47,512 字。其餘無法對映之字形，如不符國際 ISO 組織收納漢字原則或待送審納編之字形即編入 Unicode 第 15 字面。

二、全字庫網站功能

全字庫網站為國家中文標準交換碼(CNS11643) 字碼與屬性資料服務平台，用以解決電腦中文字碼問題。平台內提供各項查詢工具及軟體，可快速查詢及免費自由下載使用。全字庫提供明體、楷書和宋體 3 種字型，屬性查詢有注音、倉頡、筆畫、部首、部件、筆順、CNS、Big5、Unicode、拼音、符號、拼音文字、複合查詢等，另建有字碼對照表，及顯示、造字、共通平台及輸入法功能的應用系統程式，提供各單位免費加值應用，以下就網站提供之功能說明如下：

(一) 字碼查詢與下載

除依總筆畫、注音、部首、倉頡、CNS 碼、Big5、Unicode、拼音、符號、音文字、部件或筆順序等單一查詢條件外，另提供複合查詢，可以透過筆畫數、部首、注音、倉頡、拼音、CNS 字面、部件及筆順序等 8 種查詢型態，依需要作不同組合的查詢。

(二) 轉碼互通

由於各政府機關導入資訊設備時程不一，可能使用不同之中文系統與中

文內碼，或由於機關內部需要而自行造字。為了確保各機關不會因所用中文內碼不同，在電腦資訊交換時發生無法對應之中文字型，因此資料在交換前可先轉為 CNS11643 國家標準交換碼，避免造成漏字、缺字或亂碼。

全字庫除可線上查詢 UTF-8、UTF-16 及 Big5 與 CNS 碼之對照外，並依經濟部標準檢驗局制訂 CNS7654「字元碼結構及延伸技術」轉換中文標準交換碼(CSIC)，提供使用者以文字檔上傳進行線上不同編碼之轉換後即時下載轉碼結果。另為利使用者開發之應用系統執行，提供可與應用系統互動之 Web Service 轉碼。

(三) 中文共通平台元件

由於中文的字數遠遠過個人電腦系統字數，所以某些中文字形無法在日常使用的作業環境下被正確的顯示出來，對於機關網頁系統有自行造字或顯示有問題、原有資料庫資料更新或修改不易者，全字庫提供中文共通平台 JAVA Swing 元件，讓遠端使用者在瀏覽器下得以正確輸入和顯示出字型，並且可以將輸入的資料傳送至伺服器端資料庫，對資料庫進行新增、修改、刪除及查詢等動作。全字庫提供中文共通平台 JAVA Swing 元件如下：

元件名稱	功能說明
CNSJButton　按鈕元件	可設定按鍵上的文字
CNSJComboBox　下拉式選單元件	可設定下拉選單上的文字、編輯文字、取得下拉選單文字內容
CNSJLabel　標籤元件	可設定標籤上的文字
CNSJList　選單元件	可設定選單上的文字、取得選單中的文字內容
CNSJPassword　密碼輸入元件	可設定文字、編輯文字、取得文字內容、顯示則以黑點表示
CNSJTable　表格元件	可設定文字、編輯文字、取得文字內容
CNSJText　單行輸入元件	可設定文字、編輯文字、取得文字內容

除上述中文共通平台 JAVA Swing 元件外，對於網頁文字有顯示罕見字型需求者，全字庫亦提供字型即時顯示，將網頁所需顯示之罕見字型轉為 PNG 圖檔，以便即時顯示於使用者端瀏覽器。

(四) 應用工具

全字庫除提供上述網站服務功能外，依不同作業平台提供個人電腦造字

處理工具，對 Unicode 平台如 Linux 作業系統或 MS Windows 2000 之後版本之作業系統(Windows XP、Windows 2003、Windows Vista、Windows 7 或 Windows 8 等) 提供全字庫體包及全字庫單機版轉碼工具；另對 Big5 平台如 MS Windows ME 以前版本之作業系統(Windows 98 及 Windows 95 等) 提供全字庫應用工具 4.0、個人自造字集整合工具 4.0 及網頁造字轉換工具 4.0。

(五) 新增造字解決方案

目前全字庫納編字型對映到 Unicode 第 15 字面(私人使用區，PUA) 約 2 萬餘字，就多數使用者所使用之 MS Windows 作業系統造字區僅於第 0 字面 E000-F8FF 區間共 6,400 字，遠少於全字庫對映到 Unicode 第 15 字面之字數而無法完全容納，因此常有使用者之新增造字需求，目前較佳的處理方式即將需使用到之自造字建於造字區，如需電子文件交換再以轉碼方式轉換成 CNS11643 中文標準交換碼作為中間傳遞資料。

全字庫提供相關的程式工具，讓用戶可直接將全字庫網站所提供的中文字下載並轉入至自己電腦的造字區，使用者即可於電腦上利用舊注音、舊倉頡或內碼輸入該字，用以解決使用者電腦的缺字問題，同時也提供相關工具讓使用者方便管理及使用所下載的造字。

以 Windows 2000 以後的版本為例，使用者可安裝全字庫軟體包後，至全字庫網站之「字碼查詢與下載」查詢所需的字碼，並下載安裝該字型後，重新啟動電腦後，即可用舊注音、舊倉頡、Big5 或 Unicode 內碼輸入該新增造字。

(六) 造字分享

為避免機關或團體因各自從全字庫網站下載字型，產生自造字「同字不同碼」現象，可透過全字庫軟體包安裝相同的自造字(共用)字集後，機關或組織團體可指定專人集中處理內部電腦所使用的自造字,當機關有造字需求時，由造字管理者從全字庫網站下載所需字碼，以建立內部共用自造字集，供內部使用者複製到自已的電腦使用；另外內部使用者亦可安裝全字庫軟體包，利用「造字分享工具」與造字管理者進行造字同步，以維護自造字與碼位的一致性(同字同碼)。

參、結語

由於國內民眾對中文字型有不同需求，而國際上對納入中文標準字碼有其規範及限制，為符合國內民眾使用需求，因此全字庫提供多項的字碼與資訊服務，並提供機關造字需求，除可透過網站查閱編碼資訊與下載正、楷、宋字型外，亦提供簡便字碼工具，節省使用者查詢及造字的時間與精神。此外為解決資料交換時中文碼之對應關係，爰在交換時由傳送端轉換為中文標準交換碼(CNS11643)，接收端再將其轉換為機關內部使用之編碼，此方式雖然可以解決機關間使用不同編碼字型不一致問題，但如果該字型係機關自行造字，因該自造字未在中文標準

交換碼(CNS11643)內，且未能流通於其他機關，因此仍可能發生無對應之中文字型造成缺漏情事。

目前全字庫網站所收納編製之字型資料，亦配合政府資料開放政策開放，可無償加值使用全字庫提供的資源，如有任何中文字碼使用上的困擾，可洽全字庫客服中心諮詢服務。