談字碼和圖書館作業

宋　玉

一、圖書館作業和電腦文字處理

人類文明和文字有不可分的關係。總觀歷史，文明幾次的大進步都是和文字有關。紙筆的誕生，使文字紀錄的製作、流傳和傳播比以前竹簡或甲骨方便簡易很多。印刷術可以容易地複製書籍和文件，不必再用人工抄寫。電腦雖然是爲便利數學計算而發明，但由於它的快速和準確，不久它就被用作處理文字的工具。再加上電腦有關技術的快速發展，它的功用已經擴展到無數領域，不論是科技、商業、教育、娛樂，都依靠電腦計算、通訊、圖形繪製，和文字處理功能開闢大片天地。所以我們可以說，電腦用作文字處理是人類文明又一次的革命。

電腦文字處理的好處是快速，可以處理大量資料，準確性極高。自從網路技術成功地和電腦結合以後，可以說是無遠弗屆。最早的電腦只能呈現文字，處理圖形的能力很小，但蘋果公司的麗莎型電腦，在 1983 年開始用視窗及滑鼠以來，使用電腦不再是技術困難的操作，而變得友善易學。電腦功能和新的應用也就如雨後春筍一般，蓬勃成長。世界上電腦能處理的文字也越來越多，如中文系統在七十年代就已出世，目前幾乎所有的語言都可以以電腦處理。

電腦也不是沒有缺點。凡需要複雜判斷的工作較難處理，如機器翻譯雖有若干成果，但仍不能和人工翻譯品質相比。語音辨識仍停留在實驗或少數特殊用途上。

圖書館作業主要是文字處理。書目的編製、維護、查閱、權威控制，閱覽流通的借還作業、預定、催還，遠距作業的查詢、館際互借，採訪作業的選書、訂購、催查、會計出納，參考作業的檢索資料庫等，每一項都需要電腦文字處理，也就是說都需要電腦作業。電腦系統並可藉圖形介面讓館員能迅速有效地執行上述的各種作業。網路則將整個館各部門連在一起，並且和他館也連上以求擴大資源，對讀者的服務也可連線到家或辦公室。

文字處理當然要處理「字」，字的問題在下一節內討論。

二、中文文字處理的特點

中文的文字處理和拼音文字的不同，它是方塊字，原是上下縱行排列，但也可作橫行排。橫行排時沒有一定的方式，傳統的排法是從右至左；但受西方文化的影響我們也開始自左向右排。現在電腦作業時，絕大多數是橫行從左向右排列。

中文字集是一個開放集，沒有人知道中文到底有多少字。對一般人而言，萬把字也就夠用了。每個應用領域又有其習用的字，如佛經譯本就充滿了許多平常見不到的字，在臺灣還有人爲取名而創造新字，科學也會需要造新字，此外，古籍中也有許多現在不常見的字。圖書館因爲要收藏各種學科古今中外的圖書，接觸的字也就比平常要多許多，因而它的文字處理系統必須要處理這些字的能力。一般性質的圖書館大概要處理兩

萬五千字，但仍會遇到所謂「缺字」（missing character），其實字集大到五萬、六萬還是會碰見缺字，不過頻率少一點罷了。

中文是靠字形（glyph）辨識的。但是一個中文「字」有時會有不一樣的字形出現，有的是筆畫上的小變化，有的是結構上較大的變化。這些不一樣的字形仍保有相同的字義、讀音和用法，所以通常稱之爲「異體字」。在幾個異體字之中，我們選擇原始字形或是一個最常用的字作爲「正體字」。異體字的產生和正異體字間的關係實際上是非常錯綜複雜的。認識正異體字的關係有助於圖書館作業如編目、檢索法等等。

下圖是正異體字表中的一頁。橫行最左邊的字是正體字，其他同列的字都是異體字。大陸用的簡體字也是一種異體字，在本表中的位置是正體字右邊第一個異體字。大陸的簡體字常常把部首省去，如「麵」字簡化成「面」字。還有省得更多的，如「餅乾」省成「餅干」，「幹部」字也簡化成「干部」，當然「干涉」的「干」字仍沒有變。所以有時好幾個字簡化成同一個字。這樣繁體字轉成簡體還可按對照表做，但簡體轉回到繁體則需了解全文的意義才能轉得過去。

異體字是怎樣來的呢？以下是幾種不同方式產生的異體字。「俗字」有的是因爲語文（或稱口語）的產生通常是在文字之先，民間俗字和後出的官方文字產生分歧；也有俗字是官方文字簡化而成。俗字官字的例子如「逃」和「淚」；「吊」和「弔」；「叨」和「饕」；「做」和「作」等。

「假借」是口語中沒有字形的字借用其他已有的字，如「亟」字是「極」字的本字，「極」字本義是屋棟的意思，但常被借用爲極致之意，本義反而罕見。亟字現作急迫義。

「簡體字」也是俗字之一種，存在已久，是爲簡化書寫筆畫而生。大陸將之正式化，改變許多字的字形和寫法。

「省體字」也是簡化之一種，將原字結構中省略其中一部分的文字。如「準」字省成「准」；「　」字省成「凜」。

「後起字」原有本字，因爲被假借義佔用，久假不歸，於是另造一字來還其本原，這個另外造的字就是後起字。如「暮」是「莫」的後起字。

字形小變化的一個來源是筆寫的楷書和印刷刻板的字體（如明體，仿宋體）之間字形變化。譬如「戶」字在楷書上面是一點，而印刷體上面則是一撇。其他的小變化也都類乎此，筆畫稍有不同，或是省去一點等等。此類小變化現在稱之爲「微變化」（microvariation），在一般情況下應視作同一字形。下圖列出「微變化」的實例。

綜上所述，中文字數繁多，字形又不是很統一，我們圖書館作業人員究竟應該如何處理呢？我的想法如下：

1. 圖書館系統所有的字集應包含館藏圖書內的字集。當年國家標準碼（CNS）的字集只有一萬三千的時候，圖書館界拒絕採用它，而用專爲圖書館作業發展出的中文資訊交換碼（CCCII），因爲後者的字集可達五六萬。

2. 依上所述，圖書館作業應以中文圖書的字集爲對象。但不必爲戶政中姓名怪字而困擾，除非該姓名有著作列入書目資料庫中。科學所產生的新字也應如此看待。換言之，當這些字成爲圖書館系統的缺字時我們再求補救，沒發生前大可不必爲此操心。

3. 缺字問題（missing character problem）—字集大的系統當然遇見「缺字」的機率比較少，但這不應該是主要考慮。因爲中文字集是個開放集（注：開放集是數學名詞），沒有邊，中文圖書也有各式各樣的內容，無數的變化，所以再大的系統也會碰見缺字。我們應該考慮的是有缺字時如何去處理。缺字通常解決的方法是造字，在電腦系統上加入缺字的字形、輸入法和字碼。但這條路如果大家不協調一起做，就會造成資料不能交換的問題。所以這需要集合大家的智慧和力量，商量出一套辦法來解決問題。

   另外我認爲中央研究院謝清俊教授的「缺字」研究也是一個很值得考慮的方案，不過這計畫目前停留在利用已造好的字型階段，我認爲應「當場」（on the fly）造字更合乎實用。

4. 另外一個問題是協調大家對字的認知。有微變化的字應仍屬一字，但在有些字碼表中被分別列爲不同的字。如果沒有協調一致的認知，檢索和查詢就會有遺漏錯誤。實際上「認知」的問題應該從書作者起，到印刷排版人、圖書館員、一直到讀者都應該一致，否則就會產生錯誤，而使資訊檢索操作得不理想。我提起這一點的原因是因爲在我整理書目問題字時，發現印刷人員有過排字錯誤，館員也有過書目編製錯誤。至於作者和讀者的錯誤，我相信也是有的。

三、字　碼

當文字處理系統付之實作（implementation）時，有三樣東西是必須具備的：字碼、鍵盤碼和字形檔。這三樣之中，最重要的是字碼（character code），因爲它是將文字和電腦內部運作連接在一起的主要關鍵。鍵盤碼是將鍵盤的按鈕轉換成字碼的橋樑。（西文都是用這種方式。）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動作轉換成字碼。這種軟體叫輸入法編輯器（input method editor，簡稱IME）。它的設計是隨輸入法而定的，如倉頡、注音、嘸蝦米等。字形檔則將字碼和字形連接起來，使螢幕和列表機將字形呈現出來。點陣式圖形檔較粗糙，且不易放大縮小。新式字形檔如 TrueType 描繪出字形較美觀，可放大縮小。通常輸入法及字形檔都可由獨立廠商開發，本文因篇幅關係不討論。

什麼是字碼？字符（character）對西方拼音文字而言就是字母，再加上輔助的標點符號。對漢字而言則是我們的方塊字和標點符號。碼（code）是電腦內用以代表事物的二進位數值。所以字碼就是電腦用以代表字的數值。每個電腦原則上只認識和處理機內字碼的字。所以電腦之間交換文字資料時，不論是用磁碟片或連線，有「亂碼」產生多半是由於雙方用的字碼不一致的緣故。

當我們必須用文字表示二進位數碼值時，我們常用十六進位數字以便利閱讀和節省空間，如「不」字 CCCII 碼是 21302A 和 Unicode 碼是 4E0D。因爲一位元組有八個位元，所以能代表字的碼點只有 256 個，西方拼音語言的字母和符號可以放在一個位元組之內。但東亞漢字語言字數少則數千，多則數萬，根本不是一個位元組所能容納的，於是就有多位元組碼出現。譬如在臺灣個人電腦普遍使用的 Big 5 碼是二位元組碼，它不算使用者造字區已只有 13,051 字。而圖書館界用的 CCCII 碼是三位元組碼，字數有 54,090 個字。

字碼經過四、五十年來的發展，我們可以看出一個成功、普遍使用的字碼系統必須滿足下列的幾個條件：

* 它的字集（character set）必須滿足當時使用者的需求。請注意這項需求終會隨使用而有變遷。譬如一方面 Big 5 使用者很高興有一萬多字可以使用時，但很快會對某些新聞人物名字不能處理感到不便。

* 它必須有明確不混淆的字形對字碼的關係，最好是一對一的關係。如果因特殊情況要有例外時，也應有明確不混淆的使用規則。這樣才會使運作明確順當，不易發生錯誤。

* 它應是當時的技術跟價格限制下的最佳產物。使用它所做出來的系統必須有足夠的商業後盾，這樣才能有良好的維護支援。

以下回顧字碼的發展，先討論歐美拼音文字的字碼。

（一）英文及其他拼音文字的字碼

在真正的電腦還沒有發明前，就已經有了用穿孔卡片和穿孔紙帶的半電腦型計賬機器（accounting machine）和印字電報（teletype），其中有些是用寶多（Baudot）五孔碼。

這種碼因爲用五孔，所以只有 32 種組合，它用了「轉入」（shift in） 和「轉出」（shift out）增加了一倍的編碼空間，才勉強包括英文 26 個字母，10 個數字和少數標點符號。

電腦發明後所用的字碼極不統一。原因是那時代的電腦的架構紛歧，如一個 word（電腦內部資訊的一個單元，與文字無關）有 12 位元、24 位元、32 位元、36 位元、48 位元、60 位元不等。那時代較多用所謂 BCDCC（binary coded decimal character code）是一種六位元爲一組的字碼，它的字碼集和前述的寶多碼大了許多，增加了一些標準符號和控制碼，字母仍沒有大小寫。

當時的情況是字碼混亂，沒有公認的標準，並且 BCDCC 也不甚實用，所以 1963 年美國國家標準學會推出了七位元的 ASCII（American Standard Code for Information Interchange 美國資訊標準交換碼）解決了大部分的問題。ASCII 是七位元碼，所以有 128 個編碼位置（或稱編碼點），其中 34 個碼是用於通訊或電腦的控制碼，94 個碼則係代表圖文字符，如英文大小寫字母、數字、符號等。除了 IBM 之外大部分電腦廠商都採用 ASCII 碼。在英文環境裡它已經夠用，所以在近四十年之今日，ASCII 仍然大家都採用。

但從八十年代起，跨國電腦資料交換及連線通訊逐漸興起，ASCII 限制在英文環境內使用就漸漸遭受挑戰。歐洲各國擴充 ASCII 爲八位元，但不甚一致。所以這段期間裡，電腦與電腦間的通訊常常有部分亂碼的情況出現。隨後，世界標準組織發展出ISO-8859 系列的字碼表以解決問題。ISO-8859 包括一系列的八位元碼表，每種碼表涵蓋一個語文地區，如西歐拉丁語系德法西意瑞典文等、東歐俄文語系（Cyrillic）、波羅的海國家的語文、阿拉伯文、希伯來文、希臘文等。這些碼表前一半都是 ASCII，後一半才是地區語文的字和符號。這系列的碼表總算大致解決了歐洲各地區的問題，但也沒有全部解決，譬如俄國就不用 ISO 8859-5 碼表，而獨用 KOI8-R，並且跨語文地區仍是問題。

綜觀拼音文字字碼因爲英文或拉丁文的基本字母有限，所以七位元的碼表就夠用了。但歐洲各地區有在基本字母上加上的各種「特殊讀音附加符號」（diacritic），因此要用到八位元的碼表。但即使是八位元的碼表也只夠容納一個地區的「附加符號」，也無法包括全部「特殊讀音附加符號」。 ISO-2022 是電腦間轉換碼表的一個標準，它使電腦文件可以包含多過一種字碼。但它的使用不是很方便。其他拼音文字如泰文、馬來西亞文、藏文都有自己的碼表，但不在 ISO-8859 系列內。拼音文字的字碼不論是七或八位元的，都算一個位元組。

（二）中日韓文和其他表意文字的字碼

中日韓文都用漢字 ，其他如越南、新加坡等地區也用漢字。漢文字碼的問題和西方拼音文字不一樣。主要的區別在漢文字集很大，需要用兩個或更多位元組的字碼，才能裝得下。但各國的漢字之間，字集的大小出入很大，少數字的字形也不盡相同。以下主要介紹臺灣中文字碼發展的經過。

在臺灣最初引進電腦的時候，電腦硬軟體沒有中文系統。最先借用電報碼處理中文。電報碼原先是電報處理中文而編的碼，用四位數代表一字，因此有一萬個編碼位置（或稱編碼點），實際上收入的只有八千多字。雖然字集不是很夠，但已經可以用了。財政部財稅中心是最早使用這個字碼的單位，後來他們又陸續加了一些字。

電報碼的空間效率很低，因爲在四個位元組裡面每個位元組的 256 個可能編碼位置它只用了十個，因此它浪費電腦儲存空間，增加成本，減低處理速度，所以不是非常理想的安排。在不是很多文字的應用系統裡還算可用，但不適用於含大量文字的資料處理。

怎樣取代電報碼卻有相當大的爭議—中文字碼究應採用二位元組和三位元組。二位元組碼的擁護者認爲字集裡有一萬多個字就很夠用，並且二位元組碼在技術層面上困難度和成本都低很多，於是獲得政府、工商業的贊同。在臺灣前後有好幾套二位元碼出現。

IBM 從事研究東亞寫意文字字碼多年，在日本有漢字產品，所以很容易地在臺就推出 IBM 兩套二位組碼，一套爲 IBM5550 用，一套爲 IBM 主機用。這兩套的字集、字序都相同，字碼則依其作業系統環境而有不同的放置。

除了 IBM 碼之外，臺灣在市面上還有好幾套碼。資策會爲了減少混淆，就和五個大電腦公司合作，在 1984 推出所謂「大五碼」（Big5） 字集、字序都和 IBM 碼相同，有 13,051 個字，字碼則有不同的放置法。推出以後甚爲成功，絕大多數的個人電腦都採用它。用字不敷時就在加字區域內造字，但因不是統一加字，所以加字區的碼有些混亂。

在 1986 年中央標準局也公布 CNS11643 碼，字集和前述的 IBM、Big5 碼相同，字序則不一樣，因此它也有 13,051 個字。

上述這些碼和其他二位元組碼包字數都有一萬三千左右，使用者造字區五千左右，它們都可和 ASCII 英文碼混合使用。但由於它們碼的安排都是分成幾段，所以不能作排序。如果要作排序，則需要用輔助排序表。

但 1992 年中央標準局將 CNS11643 擴充到 48,027 字，分置於七個字面，每個字面容納 5,401 到 8,603 字不等，使用時轉換字面需要用「脫離字符串」（escape sequence）。它不適用作電腦的內碼，所以算作交換碼。

三位元組字碼其實發展得很早。CCCII 是謝清俊教授和一群學者在 1980 年爲圖書館界所發明的中文碼。最初編出 4,808 個字，以後分批陸續發布。到 1987 年共編定 53,940 個字。國內圖書館由中央圖書館（即國圖前身）最早使用此碼，以後各大館陸續加入。直到今日，國家圖書館、各國立大學圖書館以及三大公共圖書館（北市、高雄、臺中）都使用 CCCII。

CCCII 用三位元組代表一個字，所以編碼空間很大，有八十幾萬編碼點的空間。CCCII 容納了五、六萬字之外，還用編碼位置來表示正異體字的關係，以方便異體字的檢索。並且因爲它是一個大表，所以它不需要「脫離字符串」在不同字面之間跳來跳去。

CCCII 在當初設計時就被美國研究圖書館組群（Research Library Group，簡稱 RLG）決定採用其架構和借用一部分字碼（只有 15,000 字左右），並將日韓文編入系統內。這種版本即是美國國家標準 Z39.64，叫 East Asian Character Code，簡稱 EACC。臺灣廠商在做 CCCII 系統時，其實是將 EACC 也一併做在內，所以中、日、韓文都有。

但 CCCII 經過十多年的使用也曝露出一些缺點，它最大兩項缺點，其一是正異體字定義不明確，以至於重複字碼（即一字多碼）多達萬餘；另一大缺點是它的三位元組架

構不適於現在流行的圖形使用者介面（graphical user interface，簡稱 GUI），如 Windows 95/98/2000。次要的缺點包括：字形缺乏一致性、日韓文字形不完全正確、與 ASCII 碼混用時需要脫離字符串、異體字碼表無索引、以及由於使用者不多，因而系統價格高，廠商服務品質欠理想等等。

大陸一般用的字碼是 GB 1312-80（簡稱 GB 碼）。這個碼表是 1981 年發布，包括 6,763 個大陸用的漢字，連同符號和外語字符一共 7,445 個字符碼。它是二位元組碼，每個位元組的 0x00 - 0x20 編碼點都留給控制碼，0x21 - 0x7E 等 94 個碼點用來編碼，即所謂區點制（kuten, or row-cell）。在實際應用時，兩個位元組的首位元都改成 Set，便利和 ASCII 混用。這碼叫作 Shift GB。GB2312 碼後來經過三次修正，字集增到 8,150 字符（7,399 個漢字）。GB 碼雖然字數較少，但據了解運作沒有問題。

1993年大陸爲了配合 Unicode 的到來，將 GB 碼擴充到 21,886 個字符（21,003 個漢字）。Windows 95/98/2000 的大陸版都已在用 GBK 碼。

大陸另有 GB 12345-90 標準，是 GB 2312-80 的繁體字對應碼表，共包括 7,709 個字（其中 6,866 個漢字）。大陸簡體字和傳統繁體字間的轉換有一個根本的問題，就是前面所說簡繁體字之間有一對多的關係。這個問題很難用電腦去完美解決，因爲字義要看上下文和語意而定。

日本從事電腦文字資訊運作很早。早在 1976 年就作出只有半型片假名的一位元組碼表。1978 年 JIS C 6228-1978 是第一個包括假名、漢字、符號的二位元組的日文碼，有 169 個假名、6,349 個漢字和 284 個符號和外國字母。其後經過兩次增修，現在的 JIS X 0208-1990 其實只加了 6 個漢字。另外 JIS X 0212-1990 增列了 5,801 個輔助漢字、266 個符號和外國字母。這些碼表間的轉換要用「脫離字符串」。日文所用漢字較中文爲少，再加上日文文字資訊系統發展較早，運作也成熟，所以沒有多大問題。但是碰上要轉換到其他語文碼表時，雖然可用 ISO 2022 的機制，卻也不是很方便。另一方面，像所有二位元組字碼一樣，因爲要和 ASCII 混合使用，切割分析處理速度比較慢。

韓國最早的字碼表是 KS C 5601-1987，現用的是 KSX 1001-1992。其中包括 2,350 個韓語拼音組合（Hangul），4,888 個漢字，986 個符號。另有 KS X 1002-1991 延伸集字碼表，包括 3,605 個韓文拼音組合，2,856 個漢字，1,188 個其他字符。這些碼表都是二位元組碼，其中有好多重複字。我們對韓國文字資訊系統不夠熟悉，就不多討論了。

四、Unicode

所有現用的字碼，不論是西方拼音文字或是漢字，都有一個共同的缺點，就是不能很方便地跨越語文，將不同的語文混合處理。ISO 2022 雖然有轉換語文的機制，因爲它是將就現實而設計的，所以不是很好用，並且沒有一個電腦系統是把它納入在操作系統內的。現在文字在網路上的應用越來越普遍，因之對跨語文的需求也越強烈。我們圖書館本來就是處理多國語文的地方，所以對這個需要感覺更深切。那麼圖書館對字碼問題應採取什麼抉擇呢？我們的建議是採用 Unicode，因爲它的架構是根據以往字碼發展的經驗，從電腦科技的根本重新開發出來的結果，解除原來的限制，大大地增廣了字碼的天地。

Unicode 的特點簡述如下：

1. Unicode 是用二或四位元組代表一個字符的碼，它將世界上主要語言編在二位元組的 BMP 字面（Basic Multilingual Plane 基本多國語言字面）。在 Unicode 3.1 標準內，在 BMP 65,536 個編碼位置中，編入 11,830 個拼音文字字母和符號、27,786 個中日韓越用的漢字和 11,172 個韓語拼音組合，共 50,788 個字符，另有 6,400 個碼位指定爲使用者造字區。至於四位元組字面，Unicode 用兩個特殊二位元組的替代碼組合（surrogates，前替代碼之範圍爲 D840-DBFF，後替代碼爲 DC00-DFFF）來表示。這種四位元組的碼去年編定的 Unicode 3.1 公布了第一批，計漢字 42,711 個，所以漢字的總數已達 70,207。另外還有四位元組的使用者漢字造字區 65,536 個碼位。我們認爲應該是夠用的了。

2. Unicode 在 BMP 內，一律用十六位元的定長字碼，不用八位元字碼，所以每個字碼的長度都是固定的，方便電腦切割分析（parsing）字串。在四位元組區時，它用兩個 BMP 內二位元組的替代碼的組合代表一個字符，所以切割分析仍是定長，並且不用繁複的「脫離字符串」。這種處理方式簡單明瞭，並且快捷。

3. Unicode 是以十六位元爲一個單元，所以控制碼（control codes）也是十六位元。它一共有 64 個控制碼，而其他 65,472 個碼位都可以用來編碼。這和傳統的二位元組碼很不一樣。因爲傳統碼爲和 ASCII 混用，並避開每個位元組內控制碼位置，因此只有 19,000 個碼位左右。這就是 Unicode 的空間效率大很多的原因。但這也是有代價的。第一，原本只用一位元組的拼音文字現在必須變成十六位元（即二位元組）。其次，Unicode 的作業系統須要重新編寫，或是需要作大幅更動。譬如 Microsoft Windows 2000 即是專門爲支應 Unicode 而設計的作業系統。

4. 在可能情形下，Unicode 企圖合併語言中的重複部分。因此所有西歐語言（德、法、意、西、丹麥、瑞典、挪威）都併入拉丁語系碼表，斯拉夫語言也全併入 Cyrillic 等。所有漢字的語言（中國大陸、臺灣、日本、韓國、香港、新加坡等）全都合併成統一漢字集（Unihan）。凡是一語系中字形相同的字符只給一個碼。字義字音相同，但字形有較大出入時給不同的碼。這樣節省不少的編碼空間，如在 BMP 內，我們用的中文字有 23,989 個字，大陸有 21,087 個字，日本有 12,815 個字，韓國有 9,310 個字，其他 375 個字，加起來有 67,576 個字，但合併以後總共只佔 27,786 個碼位，節省約 59% 的空間。

5. Unicode 將各種語言併在一個大碼表中，所以處理和顯示任何一個語言中的字，或是混合不同語言中的字都不是問題。不像以前的系統中，雖然有多個碼表可以更換，但在某一個碼表換入時，只有 ASCII 和該碼表的字符能夠顯示。換言之，從前的碼表作業不能將多種文字同時顯示在螢幕上，更遑論作混合處理。

6. Unicode 有一個理想，就是把世界上所有的語言即使用人口超過五百萬的全部容納在內，再加上歷史上的重要的語言。目前它已包括三十餘種主要語言，尚有五十餘種在審核程序中。現 BMP 中尚有 7,827 個碼位待編，而四位元組的一百萬餘碼位僅編定了 42,711 個漢字，所以假以時日，Unicode 會成一真正的「世界通用」（universal）碼。在 BMP 內，二位元組碼需要貯存空間較小，處理速度較快，但餘下的空間已所剩無多，所以在 BMP 內納入的機會已不大。四位元組區域的語言和字，雖然在處理上較吃虧，但能和世界上所有語言文字一起處理，已經是一項不錯

的選擇了。我們在 BMP 內有 27,786 個漢字 ，一般而言，應屬夠用。

7. Unicode 只是一套字碼系統，但如前所述在電腦上應用時，尚需要鍵盤碼或輸入法編輯器來作輸入系統，和字形檔作輸出系統。前者是隨語言特質而不同，較複雜的語言如中日韓文需要輸入法編輯器。這些語言需要一或多種 IME，如我們中文有注音、倉頡、大易等不同的輸入子系統，用時隨使用人喜好而定。輸出則需要字形檔，每種語言需要一或多種字形檔，如我們有細明體、楷體和粗黑體等等。有了上述的兩項子系統，文字處理才能運作。這些子系統通常都有專業公司在開發，Unicode 系統雖然龐大複雜，我們不認爲在輸入輸出方面會有重大困難。

8. 在目前網路上通訊時，各系統必須遵守一些現行的通訊協定（protocol），也就是說，在網路上流通的位元流（bit stream）必須受到一些規定限制。爲此，Unicode 有好幾種爲網路通訊的轉換格式：UTF-8、和 UTF-16，最近還有 UTF-32 正在討論中。現在在網際網路上有許多網頁是用 Unicode 編寫的，並且 Netscape 及 IE 都有接受 Unicode 的能力。

9. 支應 Unicode 的軟體作業系統正在增加和改進中，但目前真正用 Unicode 的應用軟體仍然不多。據悉，美國 VTLS（淡大已採用）、DRA、以色列 Alef 都有用 Unicode 的圖書館系統，此外 Innovative、Dynix 也都在開發中。國圖爲了因應 Unicode 的到來，已開發出 CCCII、Big 5、GBK 等碼對 Unicode 的雙向轉換表和轉換軟體，同時也在發展 Unicode 書目試驗系統。

10. Unicode 不是沒有缺點。它因爲一律用十六位元爲一單元，所以對西方拼音語言原來只要用一個位元組的就增加了一倍的貯存空間。對臺灣圖書館書目而言，原來是 ASCII 和 CCCII 混用，ASCII 和 CCCII 間的轉換則用三位元組的「脫離字符串」，改成 Unicode 以後，因爲有增有減，書目記錄大約是原來的長度多四成。

11. Unicode 的另一個缺點是增加新字和修正錯誤的機制，因爲需要國際上合作制定，過程非常緩慢。所以，我們認爲應有一個單位，統一蒐集缺字錯字，在讀者造字區來做加字和修改錯誤的工作。一面通知大家，供應必需的輸入輸出資訊檔，並且一方面也要向國際管道提出申請，等到國際方面正式承認加字和修改後，再將結果通知大家。

五、結　語

以上簡述 Unicode 的特點，包括優點和缺點。我們相信這個碼將會取代 CCCII 及 Big 5，因爲它能在電腦和網路上傳遞和處理世界上各國的語文資訊，所以特別適於圖書館用。 至於它甚麼時候能普遍應用，這將看許多因素而定， 很難預測。不過我們深信 Unicode 是我們要走的路。國圖正在爲它做準備工作，我們也願意和大家分享開發出的軟體還有 Unicode 經驗。