淺談編目的轉檔與字集

魏令芳

壹、前言

許多讀者在作線上公用目錄查詢時，經常會在螢幕上看見一些怪字，尤其是一些關於歐洲語文方面的題名，如一些變母音等，更是常見，這最主要的形成主因是因爲轉檔與字集的因素所導致的。編目除目前的原始分編外，抄錄編目也佔有相當大的比例，供轉載和抄錄的光碟資料庫來源增多，再加上網際網路的發達，彼此間的抄錄和轉錄也逐漸的熱絡，雖說技術和成本上尚有許多爭議及待解決的問題，但就『資源分享』這一個大的目標來講，即『一館建檔而他館分享』的目的上，確也逐步的在落實。是故資料的轉出轉入，即轉檔的功能，是書目進行交換的一個大前提及要項，以下謹就編目的轉檔和書目進行交換時所產生的字集字碼問題作一個淺要的說明。

貳、轉檔功能

資料的迅速建檔方式有多種，除外包建檔作業外，尚可自光碟資料庫(如西文的Bibliofile、CIP、OCLC等)、書目網路(NBINet的CATSS)或是

和他館進行書目交換來達成，但在這當中，都牽涉到不同系統或不同媒體間資料的交換與利用，能否達成交換分享的目的，就視資料的規格是否相同一致而定。目前國內外各圖書館所採用自動化系統不一，各系統所採用的機讀編目格式和使用的字碼也不盡相同，單就機讀格式而言,自1965年美國國會圖書館研發MARC以來,據統計已有二十餘種以上，是故轉檔功能是編目模組重要而不可缺的一環。

就本館系統URICA 2.6X版的編目模組而言，轉檔的來源目前僅可接受轉成ISO-2709格式的書目資料，目前的各種光碟資料庫,如OCLC、Bibliofile、 SinoCat，均可將資料以ISO-2709格式轉出,可大量批次載入館內書目主檔中。此格式基本上包有三大要項：

一、資料結構(Record Structure)
依據ISO-2709規定的資料結構,資料記錄被分成四個部份，即記錄標示、指引、書目資料登錄欄和記錄分隔，這個格式是各種機讀格式被設計的準則。

二、內容標示(Content Designations)
乃機讀交換格式中所含有的欄位、指標、分欄之定義；用各種的數字、符號及字母來說明不同的書目記錄項目。

三、資料內容(Content)
也就是資料欄內的書目資料所在。

轉檔功能的執行重點，也就是將要交換的雙方機讀格式中，將資料結構和內容標示定義不同的地方做比對，依對照表互轉至適當的定義位。不過在不同的機讀格式中，仍有許多需由人工來作比對判斷，程式的轉換尚不能作到百分之百，還需由編目館員針對轉換後否決的地方作修正後，作第二次的轉換作業。

參、字集(Character Set)

字碼是電腦在處理文字資訊時的必要媒介。在做轉載交換時，就單一系統所採用的字集而言，無論採用那一種字集，資料的建檔並不會有任何的問題存在，但一旦要進行轉檔轉換，就必須有約定的字集才能順利地進行資料的轉載，否則就無法解讀出轉出的文字資訊爲何。在目前的網際網路上也有這個問題的存在。每一個系統都有其規定的使用字集，但目前的字集種類繁多，尤其是在中文字集方面。

就英文字集方面來說，英文字集目前採用最多的是ASCII及EBCDIC，尤其是ASCII七位碼施行已有三、四十年之久，早已被大家接受採行，在英文資料的轉換上比較沒有問題。

而就中文字集來講，字集目前有二十種之多，如八千多字的電報碼、一萬三千多字的國家標準碼(CNS 11643-1986)、國內市場上佔有極大比例的BIG5碼和圖書館界所採行的中文資訊交換碼(CCCII)等，這麼多種的字碼造成了今日中文字在做資訊的處理及交換上的一個極大的困擾。舉例來說，BIG5碼收有約一萬三千多字，CCCII則有五萬多，若是以少轉多(即BIG5轉CCCII)，應不致產生對應上多大的問題，但若以多轉少(即CCCII轉BIG5)，則會有些字無法對應，更何況書目資料中會有一些怪字或符號，尚未收集在內，必需靠造字而成，更形成字碼的分歧，轉換上的混亂。

目前爲了解決此轉換上的字集問題，由教育部電算中心主導，在民國八十三年元月成立了『書目共享字集字碼問題工作小組』，國家圖書館宋玉顧問擔任總召集人，成員有教育部電算中心、國家圖書館書目中心及其資訊室人員、字碼專家、電腦及自動化系統廠商和圖書館界。基本上，此小組以解決圖書館自動化書目資料交換時所產生的字集字碼問題爲主，並整合各圖書館所遭遇到的問題字及整理一些書目中常遇到而難以處理的外

國文字、符號和標點，其工作重點有以下幾方面：

一、中文及標點圖形符號方面：

彙整及整理諸多不一致的問題字、標點、圖形符號，使其在CCCII上有統一規則可尋。

二、日韓文方面：

目前市場上的中文系統已把中文資訊交換碼(CCCII)和美國國家標準的ANSI/NISO Z39.64的東亞字集(EACC, East Asia Character Code)做在系統內，雖說CCCII尚未對日韓文給碼但已預留空間給日韓文，東亞字集內則有日韓文可用，但夠不夠用及輸入法的統一等，是目前要探討的。

三、歐洲文字方面：

目前國外的UNICO組織和ISO-10646小組發佈及推行一套可容世界各國文字的通用碼，已發佈的有拉丁語系、希臘及俄文等字集，工作小組正把其部份符號和字集加入中文資訊交換碼中。

目前館內的書目資料是以4 bytes的復盛碼儲存(但可以3 bytes CCCII轉出)，中文依注音、中文倉頡等輸入法做建檔及檢索，大陸簡體字則採化簡爲繁方式，以繁體字作輸入；日文則須以日文倉頡作建檔及檢索，即使是日文漢字，仍須以日文倉頡作檢索；韓文資料非常稀少，基本上以翻譯爲中文題名後，才做建檔。其次在歐洲語文方面，對於CCCII尚未規定之歐洲字形，自八十五年起，依國家圖書館採行方式，過濾其附加符號，改成一般字形或最接近的英文字母，以利讀者檢索查詢與館員建檔，如umlaut u，則以u建檔，目前尚有許多資料待修正，未來讀者在作線上公用目錄查詢時經常會遇見的怪字，相信會越來越少。

肆、結語

走向標準化、統一化及國際化是爲了讓彼此的溝通更暢行無阻，不管

是在編目的格式、字集或是其它各相關標準也是，大家都深切期望有一套作業的標準、協定可爲依據，當然礙於目前各館系統不一，大家無法以統一的格式建檔，惟有仰賴轉檔功能的轉出轉入，實際上也花費了不少時間和人力，大量的批次交換尤甚，但相信這些問題，在標準化、統一化及國際化的前提下，指日可待，總會解決的。

參考資料：

1.宋玉，「書目共享字集字碼問題工作小組報告」，全國圖書資訊網路通訊，3卷3期(民83年6月)，頁1-4。

2.宋玉，「書目共享字集字碼工作小組83年度工作報告」，全國圖書資訊網路通訊，4卷2期(民84年3月)，頁11-14。

3.鄭恆雄、林淑芬，「全國圖書資訊網路系統轉錄資料問題之探討」，圖書館自動化系統及機讀格式轉換研討會會議論文集，民83年，頁18-36。

4.江綉瑛、鄭玉玲、許令華，「國立中央圖書館編目新系統簡介」，國立中央圖書館館訊，17卷3期(民84年8月)，頁20-22。