[演講報導]

中文字集字碼簡介

編目組：江敏妮記錄整理

演講人： 編目組魏令芳、江敏妮
日　期： 86 年 7 月 22 日

內容摘要如下：

茲將此次由編目組負責的主題，分爲四大部份來作扼要報告：即基本概念、中文資訊交換碼意義歷史、產業界發展的幾個重要中文字碼及一些相關標準作說明。

壹、基本概念：

一、字/字符(character)是書寫的基本單位。如日中大小 ABCD1234

二、字集/字符集(character set)則是指某書寫語文全部或其部分的集合。如中文字集、英文字集、日文字集等。

三、字或字集是人和電腦溝通的主要媒介，電腦須有內建之字碼才能處理(processing)文字資料。

四、內建字碼即電腦內碼，基本上可分爲兩類：一是圖形字符(文字標點或符號等)，另一爲控制字符，爲控制電腦執行某一特定動作或指令(如方向鍵之游移等)。

五、字碼有內碼(internal code)及交換碼(interchange code)之區別。內碼爲電腦系統或其應用程式所使用的碼，僅拘限在該系統或該程式使用，會隨系統之不同而異。交換碼則是不同系統間通訊或作資料交換所使用的碼。

六、系統可因需要而擇其內碼，但在作資料的交換時，不同系統則需爲因應不同系統之內碼而備有不同多套之轉換程式，增加成

本。故若本身爲一交換碼，或有共通標準交換碼之轉換程式，則可減低作業之成本。

這也是教育部之所以希望大學級以上的圖書館系統，採用中文資訊交換碼(CCCII)的原因，因其字集字量大(約 5 萬字，擬將所有中文字 8 萬多字收齊))且廣度廣(包數字符號、中日文標點、零星貨幣及符號、化學名詞用字、英、日、希臘文、拉丁語系、及俄文語系等字母)，又爲國內唯一經國際認可之標準。

七、目前中文字碼之難以統一的原因。

一爲字集未有統一之標準。無統一之字集何來統一之字碼。

目前中文字估計約有 8 萬字左右， CCCII 收了 5 萬多，國家標準 CNS11643 收有 4 萬多，個人電腦使用最多的 BIG-5 碼則約有 1 萬 3 千多字。

二爲廠商研發之應用程式因業者或市場需求，各取所需未能一致。

八、US 使用的是字碼標準是 ASCII ，國際版爲 ISO646 。

大陸目前使用標準爲 GB2312 。

日本爲 JIS6226 ；日本國會圖書館在標準未定時使用的內碼爲 NDL-70 Code 。故使用 NDL-70 Code 建檔資料需先轉爲 JIS6226 才能和其它系統作資料交換。

國內即國家的標準爲通用交換碼(CNS11643)；而圖書館界在中央圖書館(即國家圖書館的前身)和中國圖書館學會於民 69 年採用 CCCII(Chinese Character Code for Information Interchange)爲自動化標準之一。

貳、 中文資訊交換碼：

一、中文名稱爲「中文資訊交換碼」。

但民 72 年行政院公佈名稱爲「全漢字標準交換碼」。

英文全名採用張鼎鍾教授建議之：

「 Chinese Character Code for Information Interchange 」。

二、民 68 年美國學術團體審議會，曾討論是否採用 JIS6226 來處理其東亞資料，即「東亞資訊交換碼的採用」。但因之前對東亞的考查，也了解到國內中文字碼蓬勃發展的狀況，故決定先在作些時日的觀查。當時由國科會派謝清俊教授參加，得知此項

消息後，深知其對圖書館界的影響，回國後即拜會中央圖圖館館長王振鵠館長，後得行政當局的支持，結合文字、資訊及圖館界以國字整理小組整理出的 5 萬多字爲基礎，制定出「中文資訊交換碼」，並於民 72 年經行政院公佈，且於民 78 年申請爲國際認可之標準交換碼「 ANSI/NISO Z39.64-EACC 」。

三、其字碼空間是三位元組碼，字碼空間有 830,584 個。

四、其內容收有中文字 53,940 字，仍繼續收集當中。 包含有中文字、英、法文字、數字(中、英、羅馬)、符號(化學、貨幣等)、及中英標點，並預留空間給日、韓文。 除此之外，也收錄有 UNICODE 所編碼之外國語文字，如拉丁語系字母、希臘文、及俄文語系字母。也是朝 UNICODE 目標發展。

五、之前爲國字整理小組主導，文字經文字學者考證，目前爲國家書目中心書目共享字碼字集問題工作小組負責。

六、使用者大部份爲學術性圖書館，爲圖書館字自動化系統建檔使用。

參. 業者發展之中文碼

目前台灣資訊業者所發展之中文碼，包括 IBM5550 Code 、宏碁天龍碼、零與壹倉頡碼、神通簡易碼、 HP 簡易碼、五大碼、電信傳輸碼、通用漢字交換碼、王安碼、中文資訊交換碼.....等；其中以 IBM5550 Code 、五大碼、及通用交換碼三者，最具代表性。 以下就此詳作介紹。

一、 *IBM5550 Code*

(一)1983 年， IBM 公司爲因應 IBM5550 PC ，而制定了一套中文內碼，即 IBM5550 Code ；

(二)在早期中文碼尚未成熟階段，此碼在中文文書處理方面，品質堪稱優良。

二、 BIG-5(五大碼)

(一)1985(民 74)年，由資策會參考 IBM5550 Code 所制定的，但不合國際交換碼標準。

(二)所謂 BIG-5 ，指 PC 系統五大模組，即：系統公用程式、字的組成、狀態列設定、字形、及列印設定等模組。

(三)BIG-5 爲二位元組碼。，字碼空間僅 16,000 個左右，共包含了 13,053 個字，及 IBM5550 之內碼符號。

(四)其包含有中文字、英文字、數字(中、英、羅馬)、符號(化學、貨幣等)、及中英標點。

(五)其在個人電腦內之使用占有率，於目前國內實屬最高。

三、 CNS11643

(一)CNS11643(即 Chinese National Standard 11643)中文名稱爲「通用交換碼」。

(二)爲 1986(民 75)年，行政院主計處委託資策會，參考 BIG-5 基礎所制定的。

(三)共包含了四萬字左右。

(四)是國內唯一經過中央標準局認可的國家標準。

肆. 相關國際標準

目前有關字集編碼之國際標準，較具規模者有： ISO646 、 UNICODE 、 ISO10646 、 ISO2022 等，以下就此詳作介紹。

一、 ISO646

(一)1950 代中期(民 39-)，美國國會圖書館(Library of Congress)的 Henriette Avram ，制定了國會圖書館機讀編目格式(LC/MARC)；而同時 James Agnebroad 亦創制了英文交換碼，以配合機讀格式發展圖書館自動化。

(二)該英文交換碼全名爲 American Standard Code for Information Interchange ，簡稱 ASCII ，後來即成爲美國國家標準。

(三)其後經 ISO(International Standard Organization 國際標準組織)通過，轉爲 ISO646 世界標準。

二、 UNICODE

(一)1980 後期(民 69-)，全錄(Xerox)公司的 Joe Becker 首先提出了 UNICODE 的構想，希望由傳統 7/8 bit 編碼擴充至 16bit 編碼，以期能容納世界各種語言。

(二)此構想後經 Microsoft 、 Xerox 、 IBM 等公司支持，成立了基金會(The Unicode Consortium)來籌備此一計劃。

(三)經過十年多的努力，分別於 1991 、 1992 出版了 UNICODE 第一版之 1 、 2 冊。

三、 ISO10646

(一)1980 年(較 UNICODE 爲晚)，日本、中共等 ISO 會員國提出新編

碼的構想，欲從傳統之 7/8 bit 編碼擴充至 16/32 bit 編碼，以打破現行字集之編碼方式，此構想即 ISO10646 的由來。

(二)ISO10646 欲超越 ISO646 ，希望同一套程式不必做任何修改，可應用到全世界每一個角落。

四、 ISO2022

(一)ISO2022 與 ISO10646 目的相同，亦欲完成一套全世界語言字符(中日韓英法德西...等)、符號(＃、＊、※...等)、控制字符(Del 、Esc...等)通用之編碼標準。

(二)唯一不同於 ISO10646 、 UNICODE 的是： ISO2022 僅規範編碼規則，而不包含字集。