

美國研究圖書館資訊網 東亞語文圖書資訊系統之發展

黃鴻珠

The Development of RLIN's Automated System in East Asian Languages Material

Hong-chu Huang

Associate Professor

Department of Educational Media & Library Sciences

Tamkang University

Taipei, Taiwan, R. O. C.

Abstract

The major problem of data processing for library material in East Asian languages are the design of character set, hardware for inputting, storing, processing, transmitting etc. and MARC Format. RLIN is the first institution to implement online network cataloging system in East Asian languages. In this paper RLIN's achievement and effort on this area are described and what can we learn from it to develop the Chinese library automation system in the R. O. C. are suggested.

一、前 言

一九七四年，哈佛、耶魯，及哥倫比亞大學與紐約公共圖書館，本著資源共享的宗旨成立「研究圖書館組織」(Research Libraries Group 簡稱 RLG)，其主要目標是在藉此組織共同發展館藏，並維護、保存館藏。一九七八年，RLG 會員們認為要達到真正合作的目的，必須建立自動化書目資訊系統，於是開始尋求適用的系統，最後他們選定史坦福大學 (Stanford University) 發展的

BALLOTS (Bibliographic Automation of Large Library Operations Using a Time Sharing System) 系統；同年將 BALLOTS 更名為「研究圖書館資訊網」(Research Libraries Information Network)，簡稱 RLIN。

RLIN 發展至今，有下列六種功能：

(一)採訪。

(二)編目。

(三)館際互借。

(四)會員館藏概要。

(五)建立特種文獻資料庫：建築文獻資訊系統 (Avery Index to Architectural Periodicals) 及「十八世紀出版品簡略書目」(The Eighteenth Century Short Title Catalog)。

(六)電傳郵件。

美國各大研究圖書館，除西文資料蒐藏豐富以外，東方語文資料亦不少。因之，在西文資料處理全面進入電腦化之後，接著開始尋求東亞語文資料電腦化處理之道，其中最積極的為 RLIN。RLIN 於一九七九年開始從事此方面的探討，一九八〇年七月獲得麥倫基金會 (Mellon Foundation)、福特基金會 (Ford Foundation)，和全國人文發展基金會 (National Endowment for the Humanities) —— 一〇萬美元資助之後 (註一)，便全力推展此項工作。

二、RLIN 對東亞語文圖書資料 電腦化處理所作的努力

為解決東亞語文圖書資料電腦化的處理，RLIN 所作的努力如下：

(一)解決硬體的問題：硬體設備方面須有輸入、處理及輸出東亞語文的能力，為此他們與美國麻州傳技公司 (Transtech) 合作，簽訂一項美金三十二萬元的合約，以製造CJK (Chinese, Japanese, Korean) 終端機 (註二)，即同時具有處理中文、日文，及韓文的能力，實際上此種終端機亦有處理英文的能力。

(二)排除資訊交換碼的困擾：交換資訊為圖書館業務電腦化最主要的目標，也是最大的特色。因之，一套適用於東亞各國語文的交換碼實為不可或缺的工具；為此，RLIN 採用中華民國、中國大陸、日本，及韓國等國所設定的交換碼融合於一體，名為「研究圖書館資訊網東亞語文字碼」(RLIN East Asian

Character Code，簡稱 REACC)。

(二)修訂美國機讀編目格式：美國國會圖書館所訂定之機讀編目格式 (U S MARC)，雖廣為各方所採用，並為世界各國設計其機讀編目格式的基礎，然以它來處理東亞資料時，才發現不符所需。依研究圖書館組織的規定，東亞資料至少在題名 / 著者敘述項、版本項、出版項和集叢項等處，須以羅馬拼音及原文同時顯現 (見舉例)。為適應此項需要，RLIN 曾向美國圖書館學會提出修訂 U S MARC 的建議書，建議有關之欄位採用並列數據 (parallel data) 的方式，即重複欄位，俾以拼音語文及東亞語文同時記載編目資料；另一建議為增設一字集 (character set) 識別欄位欄號066，以表達所載為東亞語文資料。

TERMINAL: C00

```

BKS/PROD          FUL/BIB      CSU083-B59          Cat Maintenance  CSU0-0AK
Record 1 of 1 - SAVE record
+
ID:CSU083-B59      RTVP:c      ST:s      FRN:      NLR:      MS:      EL:      AD:05-12-83
CC:9114      BLT:am      DCF:a      CSC:      MOD:      SHR:      ATC:      UD:05-18-83
CP:ch      L:chi      INT:      GPC:      BIO:      FIC:0      CON:
PC:c      PD:1981/1982      REP:      CPI:0      FSI:0      ILC:      MEI:0      II:0
040      CSt-H#cSt-H
043      -a-cc---
066      #c$1
110 20      Chungókung yenóchíu tsaóchihóshe.
110 20      中共 研究 雜誌社.
245 10      Taólu tióhsia k'anóhuóhuiópien / #cChungókung yenóchíu tsaóchihóshe.
245 10      大陸 地下 刊物彙編 / #c中共 研究 雜誌社.
260 0      [T'ai-peil i:#bChungókung yenóchíu tsaóchihóshe, #c1981?-
260 0      [台北] :#b中共 研究 雜誌社, 1981?-
300      v. ;#c21 cm.
440 0      "Chungókung yenóchíu" ts'anók'ao ts'ungóshu.
440 0      "中共 研究" 參考 叢書.
590      H. I. has : v. 7-10
650 0      China#xPolitics and government#y1976-      #xPeriodicals.
650 0      Chinese literature#y20th century#xPeriodicals.
650 0      Democracy#xPeriodicals.
650 0      Youth#xChina#xPeriodicals.

```

RLIN 東亞語文資料目錄舉例

(一)修訂美國權威機讀目錄格式，以適於處理東亞語文資料：其理由仍為同時記載羅馬拼音、英文，及東亞語文之需。

(二)研擬東亞語文要語索引 (keyword index) 的製作：要語檢索為 RLIN 西方語文資訊系統下的一大特色，為使東亞語文亦能保有此項特色，決定在輸入資料時即加以控制。西方語文因字與字之間有一空格隔開，實為利用電腦提取各

字最好的識別標記，然東亞語文無此方便，且東亞語文係以辭為主，而辭的字數並不一定；因之，要語須由人工加以斷定，為此研究圖書館曾數度聚會研商，並訂定提取字辭的原則名為RLG Chinese Aggregation Guidelines。

(六)修改 RLIN 的程式，以便處理東亞語文資料：RLIN 的程式原係為西文資料而設計，為處理東亞語文資料得加以修改，方能運作。

RLIN 上述各項努力中，最為各方屬目的成果則為 CJK 終端機的製作，及交換碼的設定；兩者皆為東亞語文資料電腦化處理的基本工具，而且關係密切。其處置方式與西方語文截然不同，特說明如下。

三、東亞語文字集的發展

RLIN 在探求字集設定之初，本著四個基本原則：

(一)東亞圖書館業務電腦化的作業，不能等待東亞各國發展出標準字集來以後再實施。

(二)未來電腦化書目資料記錄，應可與東亞各國的書目資料中心交換。

(三)所選定或設計的字集，在啓用之初，其字符不僅應有足夠的數量，同時應可擴充，以適應未來需要。

(四)同字異體、同字異語之間應可互相查詢（註三）。

字集的設計，RLIN 分二部份進行：一是配合終端機作業的需求；另一係為整個系統的運作。終端機所需之字集，由於當時的成品均不能符合上述原則，因之，RLIN 對美國、臺灣、香港，及其他地區的三十餘家電腦廠商索求建議書（request for proposal），回覆的有八家，最後選定傳技公司。該公司決定以其中華一號（Sinoterm）終端機作基礎，並將其已設定之 10,934 個字符與當時中華民國發展的「中文資訊交換碼」（Chinese Character Code for Information Interchange 簡稱 CCCII）第一冊、中國大陸的「中文資訊交換字集」（The Code of Chinese Graphic Character for Information Interchange）、日本的「工業標準字集」（Japan Industrial Standard Set），及韓國的「資訊處理字集」（Korean Information Processing System）相比；其結果如下：

1. 「中文資訊交換碼」第一冊（即教育部所公告 4,807 個常用字）的全部字符在中華一號內均可查獲。

2. 中國大陸的「中文資訊交換字集」有 2,337 個字符在中華一號內無法查

獲。

3. 日本的「工業標準字集」有 785 個字符在中華一號內從缺。

4. 韓國的「資訊處理字集」有 7 個字符為中華一號所無。

為此 RLIN 與傳技公司簽約，委託修訂中華一號，使其字集至少要擴大為 14,063 個字符，即以中華一號原有字數為主，加上上述 2. 至 4. 項從缺部份（註四）；並將該機器命名為 CJK 終端機。首部符合此項需求的終端機於一九八二年十月送至 RLIN 總部（註五）。其特點除將 14,063 個字符存於溫徹斯特（Winchester）磁碟機中，名為「隨機字典」（Local Dictionary）外，每部終端機配有一部列表機，一部羣組控制器（cluster controller）。該羣組控制器可控制此四部終端機。至一九八五年十一月，隨機字典已達 15,693 個字符（註六）。

為滿足中文圖書資料處理所需，究竟應有多少個字符才够？這是個無法回答的問題，因為中文圖書資料迄今已有數千年歷史，而且新資料可能產生新字符。為求字符編碼一致，以便各單位間交換資料，RLIN 除與傳技公司製造 CJK 終端機以供各館使用外，另在總部建有字彙（Thesaurus）檔。此檔至一九八五年已容 35,000 個字符（註七）。其目的在控制字符編碼的一致性；凡各館 CJK 終端機上所缺的字符，可查總部的字彙檔，如能查獲，則轉錄到各館終端機使用；無法查獲者，則由字彙檔的管理者（thesaurus administrator）造字編碼，存於總部，然後傳到所需的終端機。此種設計不僅可達字彙編碼的一致性，同時可免除各地 CJK 終端機存放大量字符的負擔，不失為一實惠、經濟的設計方式。另為達同字異體及未來與東亞各國交換資訊之計，此字彙檔各字記載的欄位如下：

1. 字型：該字符在 CJK 終端機上顯現的字型，採 16 點×18 點。
2. REACC: 研究圖書館資訊網東亞語文字符（RLIN East Asian Character Code）的編號。
3. SRCE: 為 SouRCE 的代碼，記載字符記錄（character record）的來源，例以阿拉伯數字的 1 代表此字符存在 CJK 終端機的隨機字典中。
4. ST: Status 的代碼，表示字符記錄使用的現況。
5. AD: Add Date 的代碼，說明字符加入字彙檔的日期。
6. CCCII: Chinese Character Code for Information Interchange 的代碼，存放該字符在 CCCII 中的編號。
7. PRC: People's Republic of China 的代碼，存置該字符在 PRC 的編號。
8. UD: Update Date 的代碼，即該字符最近一次被更新的日期。
9. JIS: Japanese Industrial Standard Code, 即存放該字符在 JIS 的編號。

10. KIPS: Korean Information for Processing System Code, 即存放該字符在 KIPS 的編號。

11. CTC: 爲 Chinese Telegraph Code, 即存該字符在中國電報明碼的編號。

12. DD: 爲 Distribution Date, 即該字符由總部發佈供各地終端機使用的日期。

13. RAD: RADical, 即記載該字符部首, 採用康熙字典的部首, 各部首以阿拉伯數字爲代碼。

14. XRAD: eXact RADical, 即記載該字符所屬部首, 採用康熙字典的部首。

15. AS: Additional Strokes, 即該字符除部首以外的筆畫數。

16. TS: Total Strokes, 即該字符的筆畫數, 含部首在內。

17. GL: Graphic Link, 即同字異體的連接, 在 CCCII 中同字異體另設編號, 以利資訊傳輸。

18. WAD: WADe-Giles, 即威翟氏羅馬拼音, 含聲調的序數(四聲中所屬的聲調)例如「考」載爲 k'ao3, 即第三聲。

19. PIN: PINyin, 即漢語拼音, 含聲調的序數。

20. ON: ONyomi, 即日文漢字音讀的羅馬拼音。

21. KUN: KUNyomi, 即日文漢字訓讀的羅馬拼音。

22. KANA: 標出日文漢字音讀的片假名, 及訓讀的平假名。

23. MCR: McCune-Reischauer, 即韓文的羅馬拼音。

24. KSS: Key Stroke Sequence, 即中文字符輸入時的字根組(拆)法, 例如「伝」拆爲人=亻。

25. CI: Conflict Indicator, 即衝突指標, 記載少部份字符字根組(拆)法相同時, 或某一異體字有二個以上正體字的指標。

26. CL: Conflict Link, 即連接二個字符具有相同的字型(16點×18點)。

27. DMC: Dot Matrix Chinese, 即中文字符資料檔 32×32 點的字型。

28. DMJ: Dot Matrix Japanese, 即日本工業標準24×24點的字型。

29. RC: Related Character 相關字符。

爲便於此字彙檔的利用, RLIN建有四種類型的索引以供檢索:

1. 部首、筆畫索引, 計有五種:

(1) 部首編號。

(2) 部首: 簡體字或日本常用漢字的部首。

(3) 部首: 正體字的部首。

(4) 除部首外的筆畫數。

(5) 全字的筆畫數。

2. 讀音索引, 有七種:

(1) 威翟氏羅馬拼音。

(2) 漢語羅馬拼音。

(3) 日文漢字音讀的羅馬拼音。

(4) 日文漢字訓讀的羅馬拼音。

(5) 日文漢字的假名，即日文漢字之讀音如無法分辨為音讀或訓讀，則任擇其一，以平假名或片假名檢索。

(6) 肯定的音讀的片假名、訓讀的平假名。

(7) 韓語的羅馬拼音。

3. 字符索引：

(1) 字根。

(2) 指定字符。

4. 各種字集字符編號：

(1) 研究圖書館資訊網東亞語文字集各字符編號 (REACC)。

(2) 中文資訊交換碼各字符編號 (CCCII)。

(3) 中國大陸字集各字符編號 (PRC)。

(4) 日本工業標準字集各字符編號 (JIS)。

(5) 韓國資訊處理字集各字符編號 (KIPS)。

(6) 中文電報明碼 (CTC) (註八)。

一九八三年九月十二日，美國國會圖書館以連線作業方式，將其資料輸入 RLIN 系統 (註九)，寫下歷史上東亞語文圖書館首度以連線編目的世界記錄；到一九八五年八月，計有廿一個單位輸入 120,000 餘筆編目資料。美國國會圖書館正以每年 30,000 筆的速度，將其資料輸入 RLIN 系統 (註十)，另欲與其連線作業的圖書館數量不斷成長。RLIN 在東亞語文圖書資料連線作業的成果已遙遙領先世界各國。

四、東亞語文圖書資訊系統快速成長的因素

RLIN 東亞語文圖書資訊系統發展如此快速，主要基於下列幾項因素：

(一) 世界人士共同努力的成果：東亞語文圖書資料電腦化，處理最棘手的問題即為字集的設定及終端機的製作，誠如前述，RLIN 在這方面係融合美國、中國、日本、韓國的研究成果而成。

(二) 建立在健全的西方資料處理系統基礎上：RLIN 在解決字集及終端機的問題後，所以能快速進入實用階段，主要原因是其西文系統已有二十餘年的發展歷史。RLIN 源自 BALLOTS，而 BALLOTS 係以史坦福大學於一九六〇年

所發展的公用資訊檢索系統（Stanford Public Information Retrieval System，簡稱 SPIRES）為根基。因之，修訂既有的西文系統並配合字集及 CJK 終端機即能迅速進入實用階段。

(三)有充裕的經費作為研究發展的後援：電腦化新產品的開發，在研究階段須仰賴充裕的經費作為後援；前述 RLIN 曾獲各方一百一十萬萬美元的資助，是最佳的例證。

(四)周密考慮東亞語文的特性：一九八一年二月在臺北舉行中文圖書資料自動化國際研討會，美國國會圖書館館員 Agenbroad 先生，提出中文終端機各自擁有造字、造碼的能力，如何使之統一，以便交換資訊的問題（註十一），始終困擾著各方人士。RLIN 在總部增設字彙檔的政策，徹底解決此項問題。除此之外，RLIN 的字彙檔又容同字異體、同字異語（日人採用的漢字，韓人採用的漢語，及中、日、韓對等的字）的檢索。其設想之周密，亦是廣為各方接受的原因。

(五)具有要語（keyword）檢索的功能：要語檢索首先由 IBM 公司的 Luhn 先生提出，由於資料輸入電腦後，即全交電腦處理，其快速、容易處理的特性深受各方喜愛，已成為西方最普遍採用的索引法。在東亞語文圖書資料處理，RLIN 的系統第一個採用此項索引技巧。此項應用對東亞圖書館而言，不但可彌補主題標題（subject heading）的不足，同時兼具符合便於讀者使用（user friendly）的原則。

五、結 語

RLIN 東亞語文圖書資料電腦化處理的成就已為世界之冠。由他們發展時所掌握的原則及努力，筆者建議，我國發展中文圖書資訊系統應從事下列數點改革：

(一)中國機讀編目格式記載羅馬拼音資料的部份改用並列數據的方式，以符合國際資訊交換的需求：中國機讀編目格式自民國七十年公告以來，國立中央圖書館立即採用，以建立書目資料檔，目前雖受該館主機的限制未能開放供其他單位連線檢索及線上編目，然國際間請求交換書目資訊的要求已陸續開始。為便於真正符合國際書目資訊的交換，凡需以羅馬拼音或其他語文同時記載的欄位，如欄 200 號題名及著者敘述項分欄 r 題名之羅馬拼音及 5-- 段相關題名段中各欄之分欄 r 羅馬拼音等，應採用並列數據方式，方能符合國外的需求。

(二)採用 RLIN 字彙檔的建法，以期達到一次檢索所需資料的境界：依編目規則，記述性編目資料 (descriptive cataloging) 著錄時依書上原字樣記載。由於書上採用字體不一，如正體字或簡體字間，像「臺灣」可書寫為「台灣」、「台湾」、「臺灣」四種型式，雖然字型不一，然實質內容相同。為便於系統使用者無論採用何種型式，均能一次找到所有的資料，RLIN 字彙檔的設計甚可效法。

(三)字集的設計應採各家兼容的原則：東亞語文圖書資料最大的難題之一即為字集的設計，迄今東亞各國未必有自己國家的標準字集公佈，我國即屬於此種情況，因之，有人喻為中文電腦的春秋戰國時代。在此混亂情況下，RLIN 採用中文資訊交換碼以 3 個數元組代表一個字符的基本原則，同時將中文資訊交換碼、電報明碼、中國大陸、日本、韓國等有關的字集編碼匯成一個字集；其各家兼容的原則可謂已在混亂中掌握正確的發展方向。我國中文字集標準的問題目前仍未公佈，RLIN 各家兼容的原則可供借鏡。

(四)積極研擬中文要語索引的可行性：要語索引的方法為圖書資訊系統中目前最為風行的索引法之一，其最大優點已如前述。中文要語的處理，先天上缺乏空格為界的特性，因之 RLIN 訂定出揭取要語的原則；此原則我國能否採用，建議集合圖書館界人士及語言學家予以研究。

美國 RLIN 投入東亞語文圖書資料電腦化處理的宗旨之一，為以此提昇研究的品質。日本一向把資訊視同天然資源一般重要，圖書資訊亦包括在內。基於這些觀念，先進國家積極投入發展圖書資訊系統的行列。我國雖早在十數年前即開始此項工作，然而發展速度緩慢，其原因之一是未有具體的系統模式可供參考。RLIN 東亞語文圖書資訊系統的建立，堪稱為東亞語文圖書資訊系統的範例。我們不但可以效法，同時應積極發展，迎頭趕上。

附 註

註一 Russel G. Fischer, "The CJK Terminals: RLG and Transtech's Achievement." *Library HI Tech*, Spring 1984, p. 28.

註二 同上。

註三 Karen Smith-Yoshimura and Alan Tucker, "RLIN East Asian Character Code and the RLIN CJK Thesaurus." paper presented in the Second Asian-Pacific Conferenc on Librarys Science. Seoul, Korea, May 1985, p.4.

註四 Alan Tucker, "The East Asian Project for the Research Libraries Group." In Proceedings of Symposium on Computer Processing of Chinese Library Materials and Computer-Assisted Chinese Language Instruction at ASIS-1982, Columbus, Ohio, U. S. A. October 19, 1982, p.2-5~p.2-6

註五 Russel G. Fischer, "The Computer Revolution in Chinese Language Collections." *Wilson Library Bulletin*, February 1984. p. 404.

註六 *CJK Enhancements Overview*. November 19, 1985. p. 2.

註七 同註六。

註八 同註三，頁 22-23.

註九 同註一，頁 34.

註十 同註六，頁 3.

註十一 James E. Agenbroad, "Personal Names and Chinese Character Code for Information Interchange." 民國七十年二月於臺北舉行之「中文圖書資料自動化研討會」之論文。