View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by espace@Curtin

JAPANESE APPROACHES TO MULTI-SCRIPT DATABASE PROVISION

David Wells
Bibliographic Services Unit, Library and Information Service,
Curtin University of Technology

1. Introduction

This paper is based on research carried out in Japan during 1996 with the support of a Travelling Scholarship from the Victorian Association for Library Automation.[1] I was able to visit approximately thirty libraries and information services and to conduct detailed interviews with staff in about half of these. I chose institutions known to have large multi-lingual collections in various parts of Japan. These are drawn mostly from the university and public library sector, though they also include special libraries and centralised information agencies. The purpose of my research was to investigate the provision in Japan of multi-language and multi-script bibliographic databases, that is systems which combine records in different languages and scripts within the same database. This paper discusses practical and theoretical issues in multi-script database creation, describes sample implementations and examines prospects for future development.

2. Background

There are several reasons why the question of multi-script bibliographic databases should be particularly pressing in Japan.

- The Japanese language itself requires four different scripts in its written form: Chinese characters (*kanji*), the *hiragana* syllabary for Japanese words and grammatical endings, *katakana* for the transcription of foreign words, and roman letters for abbreviations, acronyms, etc.

- Japanese libraries, particularly research libraries, have, for historical reasons, relatively high numbers of foreign language books, chiefly works in Chinese, Korean and English, but also in other west European languages and Russian. Specialist institutes of course have collections in many other languages and scripts as well.

- Because of the nature of the Japanese writing system, Japanese computer engineers have already developed considerable expertise in multi-lingual

[1] An earlier version of this paper was presented at the 9th Biennial VALA Conference and Exhibition in Melbourne in January 1998. VALA retains copyright in the text.

text processing and are thus well placed to take rapid advantage of advances in hardware.

Traditionally, Japanese libraries have accepted the principle that the catalogue record should be in the language of the text, and in a card catalogue environment, separate catalogues were and are provided for different scripts. In many libraries a distinction was maintained simply between CJK languages (which were treated essentially as Japanese) and non-CJK languages with data presented in its original roman form or else transliterated into roman letters. Larger or more linguistically diverse libraries ran additional separate catalogue sequences for other scripts. This practice can be clearly seen, for example, at the library of the Institute of Social Science at Tokyo University, which although now automated, still has separate retrospective card catalogues for materials in Cyrillic and Thai scripts. The government Institute of Developing Economies has not yet automated its catalogue sequences for Arabic, Persian and Indian language materials. The fundamental split between Japanese and 'western' books, often reflected in the existence of separate cataloguing departments for Japanese and western materials, has sometimes been maintained even after automation. The National Diet Library, for example, maintains distinct computer systems for Japanese and roman scripts.

The implementation of fully multi-lingual catalogue systems is in practice inhibited by the limitations of the easily available technology. Although in principle, with the development of standards such as Unicode, it is now possible to store and display any form of script, many computer systems and bibliographic utilities still in use date essentially from the 1970s. The Japan Industrial Standard for character encoding (JIS X 0208), which is used by many library applications, for example, only includes a limited number of Chinese characters, and of non-Japanese scripts only basic roman, Greek and Russian.

3. Current Installations

Some of the main approaches of Japanese libraries to multi-lingual database provision are outlined below.

3.1. NACSIS Based Systems

The main bibliographic utility employed by university libraries is NACSIS-CAT, administered by the National Center for Science Information Systems in Tokyo. This operates something like ABN as a cooperative cataloguing pool, though its exact relationship to individual institutions varies. NACSIS uses JIS X 0208 together with a set of extension characters, which are, however, not necessarily supported by local library systems. NACSIS derived databases are able to contain data in Japanese, roman, Greek and Russian scripts. Chinese is strictly speaking not supported, although Chinese records are in fact frequently created using the available Japanese characters. Records in languages using other scripts are entered in roman transliteration using LC standards. Additional headings are provided in *katakana* for Japanese names and in LC romanisation for Greek and Russian. Examples include

Tsukuba University Library (running a Ricoh Limedio local system), Sapporo University Library (running on NEC LICS-U UX) and the University of Tokyo, which seems to dispense with a local system altogether and whose OPAC is an *extract* of NACSIS-CAT.

3.2. Waseda University Library

Waseda does not use NACSIS for cataloguing, but takes copy data from Japan MARC (produced by the National Diet Library) for Japanese books and from LC MARC for western materials. Chinese and Korean works are excluded from the computer catalogue because of the difficulty of reconciling different character forms across Japanese, Chinese and Korean. Other languages are given in LC romanisation. The local DOBIS system is able to display Japanese and Roman scripts, and thanks to a locally written program unlike many other systems also supports the full range of ALA diacritics. CJK names are given additional headings in *katakana*, using Japanese readings in all cases.

3.3. Osaka Prefectural Library

Osaka Prefectural Library receives MARC records for all Japanese materials from its supplier, TRC Library Service (Toshokan Ryûtsû Center). Card catalogues only are provided for Chinese, Korean and Vietnamese, using the appropriate scripts and diacritics. For other languages NACSIS is used, which means that of non-Roman scripts only Greek and Russian are available in their original alphabets. The number of items in other scripts in this library is, however, very small. Japanese and 'western' computer catalogues operate independently within a locally designed system operating on an NEC mainframe.

3.4. Tenri Library

The Tenri Library functions simultaneously as a university library, a public library and the library of the Tenrikyô sect of Shintô (though in practice it is its university role which predominates). It follows NACSIS practice for western languages (i.e. only roman, Greek and Russian scripts are allowed). It also uses a variety of other sources for MARC data in all languages, though for reasons of system incompatibility does not currently import them diectly. Tenri is distinctive in its provision of electronic records for Chinese and Korean. In order to do this within the constraints of the available technology, Tenri cataloguers have recourse to the following strategies. With Chinese, Chinese characters are input in their Japanese form and any characters not found in the available character set (JIS + Fujitsu supplementary character set, JEF) are input in a *katakana* version of pinyin romanisation. Access points are provided in pinyin, in character form, and by Japanese readings (in *katakana*) of the Chinese characters. With Korean, Chinese characters are similarly input in their Japanese form and the *hangul* syllabic script is transliterated into roman. Access is provided by roman transcription of the Korean and by Japanese readings of Chinese

characters. It is recognised that this ingenious if cumbersome system is extremely difficult for the untrained user to operate, and Tenri continues to maintain its card catalogues of Chinese and Korean materials as well.

3.5. Observations

The following observations can be made about the current state of multi-lingual bibliographic databases in Japanese libraries.

- I have not been able to identify any fully multi-lingual, multi-script databases in operation. The installations which do exist are in practice limited to certain scripts or groups of scripts.

- The most urgent practical concern in the Japanese library community is for the incorporation of Chinese and Korean bibliographic records into the existing system structures designed for Japanese.

- The practice in some libraries of using Japanese readings and Japanese forms of Chinese characters for transcribing data in Chinese and Korean has historical precedent and answers the practical need to provide database records for these languages within the restraints of the available technology. However, the practice does not allow for complete data integrity in bibliographic records and is confusing and culturally inappropriate for the relatively large Chinese and Korean populations living in Japan.

- At present there is little attempt to link name authorities across scripts. Consequently, Japanese and non-Japanese databases often function in practice as distinct entities even within the same database structure. There are perhaps certain advantages in this: readers are able to extract records in particular languages by searching for forms of names in particular scripts. However, the absence of multi-script authority structures makes comprehensive searching difficult.

4. Prospects for Future Development

Current Japanese initiatives in the development of multi-script databases include the following.

4.1. Imaging

One line of development which is possible in a web-based catalogue environment is to make a basic catalogue record in, for example, romanised form, but to give access to the original script by providing a link to an image of the title page of a work. This approach is being adopted by the Tokyo University of Foreign Studies, which holds books in some 120 different languages and relies heavily on input from non-librarians

in the preparation of bibliographic records. TUFS hopes to begin implementing this system later this year using a Hitachi LOOKS-U21 system.

4.2. Unicode-Based Systems

The development of the Unicode universal character set has the potential to revolutionise multi-lingual text processing, including bibliographic databases, in the medium term. In Japan there are numerous attempts to link this technology with existing systems. Notably:

- The Asian Languages Department (i.e. other than CJK) of the National Diet Library is experimenting with a PC-based system (Gemma Unitype) that will combine the scripts of South and South-East Asian languages within a single database, using romanised access points.

- Many groups, including teams at Waseda, Tokyo University and Tokyo University of Foreign Studies, are working on the creation of multi-lingual information systems, including the automatic translation of one character encoding system to another. This is particularly important for the sharing of MARC data for languages where multiple encoding standards exist.

- NACSIS is currently developing software to allow Chinese to be added fully to its existing coverage. This will involve adopting a Unicode based standard and, in conjunction with software to translate MARC formats, will allow for the importing of MARC data from China. This project is scheduled for completion in spring 1999 and will be followed by the incorporation of Korean, then South-East Asian languages.