Meetings and Conferences

John W. Haeger

Keith Hazelton
MEETINGS AND CONFERENCES

I. Joint Advisory Committee on the East Asian Library Program

The Joint Advisory Committee on the East Asian Library Program of the ACLS, SSRC and ARL met December 6-7 in New Orleans. This was the fourth meeting of the committee, established in August 1978, now comprised of Robert E. Ward (Stanford), Patricia Battin (Columbia), Charles Churchwell (Washington University), F. W. Mote (Princeton), Eugene Wu (Harvard), Hideo Kaneko (Yale), and John Haeger (ACLS).

The committee heard a final report on the ACLS-sponsored conference on East Asian script processing held at Stanford in November (see next item) and strongly endorsed a proposal already submitted to the Ford, Mellon and Hewlett Foundations and to NEH by Research Libraries Group, Inc., to develop the capability to include East Asian vernacular characters in the Research Libraries Information Network, citing it as "the cornerstone of national planning for East Asian libraries." It also considered proposals from the University of Washington for a union list of East Asian serials and from Research Libraries Group for a National East Asian Serials Program. Further discussion of these issues is expected when the committee meets next in March.

(John W. Haeger)

II. ACLS Conference on East Asian Character Processing in Automated Bibliographic Systems

The American Council of Learned Societies' Conference on East Asian Character Processing in Automated Bibliographic Systems met at Stanford University, November Eighth through November Tenth, 1979. Through presentations and discussion, the thirty Conference participants addressed both specific technical questions and more general issues of policy and planning. This report, together with such papers as were distributed at the conference, is intended to provide a summary record of the proceedings. In view of this goal, the oral presentations and the discussions are given more attention here, while only the briefest of abstracts are provided for papers for which a printed text is available. Full sets of conference papers will be placed on deposit at the Hoover Institution and at the Harvard-Yenching Library.

Thursday, November 8, afternoon session: "The State of the Art"

Mr. John Schroeder, "Hardware and Software". Mr. Schroeder identified three aspects of the computer processing of East Asian scripts. The first is the adoption of shared conventions for the representation of non-Roman scripts within the automated system. Since data for computerized processing must ultimately be given a numerical representation, it becomes necessary to assign unique numeric codes to given Chinese characters. This assignment is a prerequisite for automated processing. "The primary problem is not a
If information sharing is to be possible, the choice of internal representation has to be made to conform with standards which exist or are emerging here and elsewhere."

The second and third areas, terminal hardware and hard-copy devices have to do with facilitating the entry of East Asian vernacular material into the automated system—input—and the production of a wide range of final products, from answers to individual queries to complete catalogs—output. Mr. Schroeder devised a typology for such devices, which are available in a bewildering variety, and concluded that "the fundamental technology exists in both the hardware and software domains [which with adaptation can] support the computerized processing of East Asian scripts." But he stressed that preliminary tasks include the detailed specification of what such a system will be asked to do and the careful examination of national and international standards to assure the possibility of future data sharing.

Discussion: The merits of various technical solutions to the problem of the display of East Asian scripts were argued, but the point was made that the question of input and output devices could be, and ought to be, treated as independent from the data base management system itself. This approach seemed especially advisable because the burgeoning development of "intelligent terminals" promises great advances in input and output capabilities. Representatives from the Library of Congress were asked about progress in the development of standards for the inclusion of non-Roman scripts in MARC records, a standardized format for bibliographic information interchange on magnetic tape. They responded that only minor adaptations were required to make it possible to include such data, but that actual implementation seems a long way off.

Mr. James Agenbroad, "Character Sets and Standards". In this paper, Mr. Agenbroad gave a lucid explanation of computers' internal representations of character data, noting the special requirements of East Asian scripts. With the development of a standard table of numeric codes for kanji, JIS or Japanese Industrial Standard, and its capacity to incorporate many additional characters as needed in the future, the Japanese have solved the problem in a way quite compatible with existing international standards for information interchange.

Chinese language materials will certainly require a large repertory of characters and the development of a standard for Chinese will have to be based on answers to two fundamental design questions: 1) What bibliographic data will the data base be attempting to handle, and what uses of it are contemplated? 2) Should the aim be an exhaustive enumeration of all possible characters initially or provision for the easy addition of new characters to the set as they are needed?

It was noted during the discussion that an organization was recently established in Peking, the Han-tzu tien-ma yen-chiu-hui, to work on character set conventions for the People's Republic of China.

Mr. Hideo Kaneko, "Current Bibliographic Applications of Automated Vernacular Character Processing in Japan". Mr. Kaneko summarized work over the last
The development of a Japan MARC is now under way. The National Diet Library expects to begin distributing tapes by 1981, with the goal of an on-line system by 1984-1985. Significantly, the Advisory Committee on Japan MARC gives very high priority to conformance with existing international standards.

Discussion: It was noted that the prevailing mode of access of bibliographic records containing both kana and kanji is to use only kana for the input of requests. This accords with the suggestion in the RLG proposal to use romanization in the formulation of search requests. Professor Nakayama estimated that about twenty university libraries in Japan have some form of information retrieval or data management services. On the technical question of how much computer memory was required to support the hard-copy output devices, Professor Nakayama gave the figure 720 kilobytes for 10,200 characters, but he noted that printers usually have their own local processor and memory so the central computer need send only the two-byte character codes.

Friday, November 9, morning session: "Reports on Work in Progress"

Dr. Ching-chun Hsieh, "Chinese Data Processing in Taiwan". The work in Taiwan in the seventies reported on here focused intensively on the formal analysis of Chinese characters into a set of components (wen) and a set of structural templates for their assembly into characters. This approach has led to relative simplicity and speed in input and relatively low memory requirements for output. With a set of 588 component wen, it is possible to identify a given character with the entry of no more than four components, so the keyboard can be much simpler than one containing all possible characters. On output, using a character-generating processor with vector-drawn components, it is possible to store the information necessary to print out 10,000 different characters with 68 kilobytes as compared to 1,280 kilobytes for a 10,000 character font represented in a more conventional 32 x 32 dot matrix.

Dr. Hsieh acknowledged the need for a standard Chinese information interchange code and observed that a committee of the Institute of Information Industries is now at work on this problem. He also observed that in bibliographic applications, the desired access points are often multi-character "words" and that a pre-occupation with characters per se may obscure the nature of this particular problem. In a word-oriented data base, romanization would perhaps be adequate for most purposes. In response to a question, Dr. Hsieh pointed out that the component analysis method could potentially be modified and expanded to handle simplified characters. He put the
cost of such a complete input device with display at $10,000, a very modest figure.

Mr. Alan Tucker, The University of Cambridge Project to Assess the Processing of Chinese Characters in Automated Library Catalogs. Cambridge University Library is moving toward a microfiche (COM) cataloging system which is scheduled to replace the old general catalog by 1981. The above project is to explore the possibilities for the inclusion of Chinese, Japanese and Korean materials within this automated system. This would be accomplished by expanding MARC records to include fields with non-Roman data and by providing subsidiary programs to allow editing of and access to this data. The technological questions of input and output are secondary. The most important aspect of this project is that if this adaptation is possible, for the first time in the library's history, Far Eastern materials would be processed by the same system that handles all the other library collections.

During the discussion, Mr. Tucker observed that the British Library and Cambridge University Library were the primary centers in England at present for Chinese character data processing. The discussion involved an amplification of Mr. Tucker's stress on the conceptual distinction between the input and output of East Asian scripts and the set of procedures and computer tools for integrating this material with planned and existing general purpose bibliographic data bases.

Professor Kazuhiko Nakayama, "Kanji Input/Output in Use at Tsukuba University". Professor Nakayama comes from Tsukuba University, a young institution with an ambitious set of goals which include "the development of a single, primarily on-line, integrated information-processing and sharing system [to be used] easily, anytime from anywhere." The University has collected seventy bibliographic data bases and has made them available at terminals at various locations around the campus. Work is currently progressing on an on-line library system (TULIPS) which, in addition to providing ordering, cataloging and circulation functions, will allow a user to "browse" the bibliographic data bases by author, title, subject and a variety of other character and numeric keys.

Sorting data into some analogue of "alphabetic order" is a basic function that must be provided by such a system. However, for kanji, there are a number of possible bases for ordering characters. What is needed is a range of options for sorting by kana reading, stroke count, radicals and other features from which the user is permitted to select the method most appropriate to a given task.

During the discussion, it was suggested that in addition to the other access points to bibliographic records, some form of abstract be included to facilitate keyword-in-context retrieval. Dr. Tsuneishi asked if there were a standardized system for the assignment of keywords. National Diet Library representatives said there was no explicitly formulated set of guidelines and Library of Congress representatives said they used the keys provided by the National Diet Library.

This issue will certainly grow in importance if the bibliographic data bases
are called upon to provide topic-oriented listings for researchers' individual projects. Without an explicit list of keywords and subject categories, it will be difficult to use the data bases for these ends.

Friday, November 9, afternoon session: "Work in Progress" (continued)

Mr. John Haeger, "Plan for the Inclusion of East Asian Records in RLIN". This draft plan, as detailed in the paper of the above title, was the result of a convergence of interests of the American Council of Learned Societies' East Asian Library Program and the Research Libraries Group in the national bibliographic control of East Asian materials. Within the last couple of years, at the same time as available technology has become capable of supporting such applications, over one-half of the major U.S. collections of East Asian materials came to be represented in the Research Libraries Group and its Research Libraries Information Network (RLIN). The goal of the draft plan, it should be stressed, is not a separate system for East Asian materials, but the design and implementation of only those modifications necessary to bring these materials into the mainstream of automated bibliographic system development.

One point of discussion was whether or not prior agreement was required on which characters would be handled. Mr. Haeger noted that a project was currently under way at Stanford University to determine character frequencies in a selected, representative sample of bibliographic records. This project's completion will be a first step toward the determination of the required character set. The draft plan calls for retrieval of character data by romanized keys, much as one does in using a manual card catalog for East Asian collections. It was asked how one would handle the problems of multiple romanizations for given characters as occur with some frequency in both Chinese and Japanese. Mr. Howard of the Library of Congress described the system currently employed there which is to follow the readings given in the Kuo-yu tz'u-tien. It was suggested that name differences would have to be handled by an automated authority file, but that the problems raised in this regard are present in manual as well as automated systems.

Ms. Yoo, "Korean Character Processing Systems". Korean script, like Japanese, is a mixture of native phonetic symbols and Chinese characters. By entering consonant-vowel and consonant-vowel-consonant sequences, the computer is able to determine the corresponding hangul character. As in Japanese, one hangul reading may map to a number of Chinese characters. One approach adopted in Korea is to enter the hangul into storage, then run a utility program which plays back the text, presenting the alternatives to each reading and allowing a human operator to pick the appropriate one in each case. As in all such systems, input is a labor-intensive process: in one case it took six man-months to input 25,000 records of bibliographic material including Chinese characters.

The Ministry of Education has directed the National Central Library to develop a Korea MARC. Ms. Yoo is a member of the working committee which is currently exploring the issue and which plans to consult with Japan on their MARC project.
Mr. Karl Lo, "Human-Machine Communication in an East Asian Library System".

Mr. Lo analyzed man-machine communication into two primary subdivisions: the actual operations performed at the terminal and the interactions with a data base management system in terms of requests and responses which are mediated through the terminal devices. His discussion of the first area included a critical survey of current technology for the inputting, storage and output of character data based on an extensive collection of literature gathered over the last few years. Current means of entering character data exhibit a wide range of learning curves and maximum rates: using keyboards with one key for each character in the set, one attains a maximum speed of 30 characters per minute after three months of experience. At the other extreme, working with a modified touch system, in twelve months, a speed of 100 characters per minute is realized and a full thirty months is required before attaining the maximum speed: 150 characters per minute.

Input by voice recognition or optical character recognition is not yet available. On the output side, laser-driven xerographic printers now attain speeds of 18,000 lines per minute.

With respect to the second subdivision, there is no off-the-shelf bibliographic data base system which will meet our needs. Such a system would have to have all the information and software required to perform the varied tasks of cataloging new entries and servicing a wide variety of user queries, from librarians checking on holdings to researchers looking for works relevant to a current project. While the terminal technology is in hand, an adequate overall information management system is still some way from realization.

Much of the discussion centered on the implications of such a comprehensive system for the training and staffing requirements of East Asian Libraries. Mr. Wu suggested that it would be necessary to train a corps of people to operate the system, but that this would not prove to be qualitatively different from the steps taken to implement automated bibliographic systems for western languages. Subsequent remarks centered on possible divisions of labor between technically trained data entry personnel on the one hand, and catalogers and other library users on the other. It was proposed that the initial entry of data into the system, which will be an ongoing, massive task that requires the full character input capability, is best carried out by specialized data-entry personnel working at a central site. Daily use of the system and, perhaps, error correction and other editorial functions would probably make do with a simpler terminal and a more "conversational" mode of communication, thus allowing individual researchers and the general library staff to use the system with less extensive training. Alternative approaches to this problem were also suggested. Dr. Tsuneishi noted that data entry personnel in the United States would have to have specialized language skills, which would make the labor costs higher than in countries where the language is the native tongue.
Saturday, November 10, morning session: "Costs and Priorities"

Mr. Warren Haas' presentation was an attempt to put East Asian automated bibliographic efforts in the context of general trends in library operations. First of all, the bibliographic data structure is evolving on an international scale: it now transcends the holdings of a given library and is becoming an independent entity. This trend is exemplified by the scientific abstract services and medical data bases which have emerged in the last decade. Secondly, more attention is being given to collection control. Libraries have long aspired to individual self-sufficiency, but with recent financial constraints affecting research libraries especially severely, a revolutionary change in the meaning given to self-sufficiency is taking place: it is now coming to mean the attainment of the capacity to tap internal and external resources to meet the needs of the local user. Such a change is leading to a major transformation in the areas of acquisition, retention and preservation. Through the planned sharing of resources, it will become possible to rationalize acquisition development by seeing individual libraries as parts of a whole.

The Council on Library Resources has undertaken a Bibliographic Services Development Program with the threefold objective of helping libraries to 1) better serve the needs of scholars, 2) serve society by promoting the cause of unrestrained access to bibliographic resources (countering the trend in some areas to capitalize on the value of information by making access to it contingent on "club membership") 3) serve the goal of cost containment by seeking to maximize the effect of the limited funding now available for research libraries.

The RLG proposal for the inclusion of East Asian records in RLIN will have to be tested against these three objectives. It must be shown how the proposed project will make a difference in these particular regards. Funding prospects will be enhanced if the proposal includes specific, detailed goals and targets which will facilitate assessment of the success of the project with respect to its stated aims.

Subsequent discussion centered on clarifying two areas in which an automated bibliographic system for East Asian materials could "make a difference": costs and services. Many voices were raised cautioning against making reputed cost savings a major rationale for automation. It was argued that in the short run, at least, capital requirements were likely to raise costs over those associated with the manual mode, and further, that East Asian bibliographic services, by their very nature, and in part due to relatively poor bibliographic control in the past, are costlier than their Western language counterparts, that these services are used by a smaller community and thus are likely targets for criticism in any narrowly defined cost-benefit analysis.

The only direct counter to this line of argument was that while associated costs may be higher initially, an automated system might be expected to change the rate of cost increases so that at some future point a cross-over will occur, after which the automated system will realize its potential for long-run costs savings.
Another group of participants held that cost-benefit analysis was too restricted a viewpoint from which to judge the advisability of automation efforts. The better question, they argued, would be to imagine what the research library situation would look like a decade hence if we maintain the status quo and undertake no automation efforts. With rising costs (Japanese materials have gone up 239% in ten years) libraries will have to curtail acquisition and researchers will often be unable to obtain access to needed materials. Also, as Mr. Haeger noted, the card catalog will become less and less suitable as a research tool for three reasons: 1) the volume of material it is attempting to control is growing rapidly and broadly, 2) the nature of scholarship on East Asia is changing: rather than an intense focus on a limited number of fundamental texts, social scientific research involves a broad search for information scattered in a wide range of documentary materials, and 3) in coming years, an increasing quantity of primary and reference materials will be provided naturally in machine-readable form; without automated systems, this material will be difficult to use. Thus, without automation, research libraries face a gradual, but increasingly severe deterioration in quality of services offered.

Automation has the potential actually to expand the range of services available, not least by making possible inter-library cooperation in acquisition and cataloging so that duplication of effort and expenditure can be minimized and total coverage maximized. Mr. Frodin cautioned the group not to over-estimate the ease of achieving wide-spread sharing of resources since it may often be perceived as counter to the interests of a particular library in maintaining its own independent status.

If a well-designed keyword facility were incorporated in the bibliographic records, the individual scholar would possess a powerful new mode of research: the ability to "browse" through the shared holdings of all participating libraries from a terminal located in the home library.

Saturday, November 10, afternoon session: "Japan MARC, Recommendations and Conclusions"

Ms. Kiyoko Tamura, "Recent Developments in Japan MARC". Ms. Tamura shared with the conference the rich experience of the National Diet Library in Tokyo in the automated processing of kanji and kana and of bibliographic records containing Japanese language data. Most importantly, she reported on the substantial progress in the definition and implementation of Japan MARC made within the last year. She called particular attention to the efforts of the Japan MARC committee to develop compatibility with international standards. The additions to the UNIMARC bibliographic record format are so designed as to dovetail with existing fields, thus minimizing the number and scope of changes that must be made to existing software to handle Japan MARC records. These MARC data files will eventually include Japanese publications back to 1969, a file equivalent to some 220,000 cards.

Ms. Tamura noted the high costs of initial setup and data input: 55% of the automation budget was used for equipment and the expense of key operators entering information. She repeatedly expressed the hope that since such time and effort were being put into the creation of this Japan MARC data
base, that other nations would find the product useful and would develop means of utilizing it. Copies of Japan MARC tapes, which should be distributed beginning in 1981, will be sent to the Library of Congress.

In discussion, the potential uses of Japan MARC tapes was explored. A minimal use would be to extract and use those fields which are provided in romanized form such as title, subject code and author. New hardware would be required to handle the kanji and kana portions of the record, but this technology is commercially available at present in Japan. Once such provisions were made, the MARC tapes would reduce the burden of the tasks of classification and cataloging in the United States. This saving could also be achieved with access to microfiche output from Japan MARC tapes.

Mr. Agenbroad noted the value of the three automated authority control files which are being developed in conjunction with Japan MARC: a 110,000 entry author file, a 20,000 entry subject heading file and a 6,000 entry publisher file. The meeting of the National Diet Library representatives with the Library of Congress scheduled for the week after this conference is expected to discuss more fully the means by which we may access the growing fund of data which the Japan MARC files will represent.

The conference participants discussed and approved a series of statements summarizing the conclusions reached in the three days of meetings:

1) That viable technology for machine processing of large Chinese, Japanese and Korean character sets, using a variety of strategies, has now been developed and is operational in many environments and continues to improve.

2) That application of this technology to the automated management of library bibliographic records is possible and has begun in Asia.

3) That development by individual libraries and networks in the United States of the capability to manage East Asian vernacular data and to adapt and utilize machine-readable catalog data produced in East Asia is possible and desirable and should be encouraged and supported.

4) That because East Asian character sets cannot be exhaustively defined, a continuing mechanism should be established to study, monitor and facilitate internationally acceptable standards.

5) That international collaboration on machine-readable formats for bibliographic data and on strategies for implementation of such formats by individual libraries and networks will facilitate international sharing of bibliographic data, which offers significant economic, social and scholarly rewards, and should therefore be planned and supported.

Thanks were tendered to Mr. Haeger and the ACLS for organizing this conference, and to the Japan–United States Friendship Commission and the Rockefeller Brothers Fund for providing the necessary funding.

(Keith Hazelton)