Collaboration to a Brave New World – Reflections on the development of Digital Library Projects at The Chinese University of Hong Kong Libraries

By Rita Wong
Head, Systems & Communications, University Library System, The Chinese University of Hong Kong

I. Historical Background

The Past since 1995 onwards

Newspaper Image Project

The first general Internet service link in Hong Kong was set up at The Chinese University of Hong Kong in 1992. The following year, the Library web server was set up and went into production in 1994.

In 1993, a request was made to the University Grants Committee of the Hong Kong Government to seek funding to scan the local newspapers and to be provided with a retrieval system so that news topics could be searchable. The Committee, consisting of representatives from all universities in Hong Kong, endorsed the proposal as they were convinced it would save a lot of manpower from other institutions from cutting newspaper articles. A sum of $6 million dollars was granted. After recruitment of librarians and technical staff, the Project kicked off in late 1994 and went into production in June 1995.

A proprietary software BASIS which was capable to handle both Western languages and Chinese languages was purchased. A 6000-term bilingual thesaurus with built – in relationship was created in addition to personal, organizational and geographical names.

Each newspaper article was cut and scanned (stored in TIFF format) and indexed manually. If the client machine can handle Chinese characters (Big-5 encoding), Chinese terms can be input. An image viewer, such as Photoshop, was required in order to display the newspaper images.

Since more and more local newspapers offered internet version from 1996 onwards and since too much manpower was needed for assigning keywords, from 1998 onwards, only keying in of the title of the article was continued. It is then hyperlinked to the internet version of the newspaper together with the scanned image.

A local company, which sprang off from the Engineering Faculty of the University, employed the similar concept. The company obtained the rights from various newspaper publishers to purchase their electronic data The Library assisted in the design of the user interface and pilot testing. Later on, full text searching with scanned images was offered as a fee charging service.

CUHK Examination Paper Image Database

The CUHK Examination Paper Image Database also employed the same image technology by scanning in all examination papers. The database was searchable by course name, course number and name of teacher. Permission was obtained from the University Examination Unit for the Library to undertake to scan the papers.

Hong Kong Index of Chinese Periodicals (HKInChip)

The name of the author, the title, abstracts and the citation of articles appeared in some 200 periodicals published in Hong Kong were keyed in. Permission was obtained from some of the publishers to include the full text. As a result, scanned images in PDF format would be included. As of May 31, 2002, the total number of articles included was 205,179. The database is open to all internet users. There are over 100,000 visits per month with over half of the visits come from outside Hong Kong. Frequent visitors are from Taiwan, Mainland China and USA.

II. The Present as from 2000

United College General Education Program Senior Seminar Paper Database

The database includes not only full text in PDF format but also slides digitalized in JPEG format. Copyright permission has been sought from the students.

Hong Kong Literature Database

The database includes not only scanned images of journal articles but also book citation and newspaper articles. The image in PDF format can be viewed by Acrobat Reader. Watermark was added to prevent unauthorized duplicating. The system is also capable of detecting massive downloading.

When book citation is available from CUHK libraries, there is a link to the OPAC so that not only the current status of the book is shown but also Table-of-Content or book jacket will be available. The total number of records as of end of August 2002 is about 90,000 entries. The run of the database is from 1901.

Recently Chinese OCR has just been employed for journal articles so that full text searching will be available. Tsinghua University OCR software is used and the error rate of about 1% is acceptable. It is also possible to edit the PDF image file after it is cut and paste to Word file.

The database is a perfect example of cooperation between the Library and the academic. Faculty staff from our Chinese Department assist in locating and analyzing the materials while the Library helps with indexing and hosting the database.

Database on Chu Bamboo Manuscript of Guodian

The database allows searching on the web the electronic version of our Faculty Prof. Cheung Kwong-Yue's scholarly publication. The book published in 1998 is a study of Guodian Bamboo Manuscript which describes the bamboo strip fascicles newly excavated in Hubei in 1993. This is another example of cooperation between the academic and the Library.

Retrieval is by bamboo strip number, title and content.

There are 702 bamboo strips and 27 broken strips with 12072 Chinese characters. About 800 Chinese characters are not found in Big 5 code and have to be created as images.

Big 5 code is the most commonly used Chinese character internal code while the most commonly used in Mainland China is GB2312 which contained 6763 Chinese characters.

III. Problems Encountered and How Collaboration will help

1. Copyright

Much time was spent in negotiating for the copyright of the journals, books or newspapers. If there was a Copyright Clearing House in Hong Kong and China, the process of negotiation could be shortened. Alternatively, collaboration can be sought from publishers for joint projects.

2. 2. Data Exchange

If data were stored in a common format, such as Dublin Core or the metadata standard that can easily exchange with each another, the flow of information and data. Organizations must collaborate together to allow free flow and exchange of data. Steps are now taken at The Chinese University Libraries to convert all databases to Dublin Core standard and will eventually reside on an XML server.

3. Common Users Interface

From the user's point of view, they would sometime prefer a common interface so that they do not need to search for multiple databases. Z39.50 protocol can solve searching multiple databases if each database does have Z39.50 protocol. Z39.50 is a good example to show how collaboration helps the flow of information.

4. Intelligent context analysis and searching

Hopefully there will be more research on context analysis and machine learning using common terms to search full-text meaningful. Researches on intelligent context analysis show both researchers from computer science, information engineering and languages need to work together for a common course.

5. 5. Common Chinese dictionary for Chinese information processing

GB2312-80 has 6763 Chinese characters. In 1995 the government of PRC announced a new set of Chinese characters, ie. GBK which contained 20902 characters. CCCII maintained by Taiwan Character Research Group has about 70,000 Chinese characters. The superCJK has 70,000 traditional and simplified Chinese characters while Big5 has 13,000+ traditional characters. . Would the ISO/IEC 10646.1 Unicode be able to solved the problem and how to handle all Chinese characters, particularly characters from classic Chinese texts, remains to be seen.

If organisations from Mainland China, Taiwan or other places can widely adopt a common unified Chinese dictionary to cater for both traditional and simplified Chinese characters processing, such collaboration will further advance the Chinese information processing.

IV. Conclusion

The Digital Library Projects at The Chinese University of Hong Kong Libraries shows the importance of collaboration, between individuals and organization, within one's own organization, and externally between organizations are vital for further development of the digital world. By removing barriers among organisations, publishers, libraries, etc. can we reach a bright brave new world of digital library.

Notes and References

1. Smith, G *Info2clear in UK*. Information World Review 9175) Dec. 2001, p. 10. The paper suggested an international copyright clearance structure.

2. Town, B *How do you provide access to electronic content without creating* barriers for users? Information World Review (175) Dec 2001, p. 53. The paper discussed ways to approach the problem of copyright without creating barriers for users.

3. Davis, D.M. & Lafferty, T. *Digital rights management: implications for libraries.* The Bottom Line: Managing Library Finances, (15), 1, 2002, p. 18-23.

4. *Information Retrieval (Z39.50)- Application Service: Definition and Protocol Specification (Version 3).* The National Information Standards Organisation, 1995.

5. Arant, W. & Payne, L. *The common user interface in academic libraries: myth or reality.* Library Hi Tech, (19), 1, 2001, p. 63-76.

6.Neil Beagrie. *The Digital Preservation Coalition*. Ariadne, (27) Mar 2001.

7. Beagrie, N. *The JISC Digital Preservation Focus and the Digital Preservation Coalition.* New Review of Academic Librarianship, (6), 2000, p. 257-67.

8. Eden, P & Gadd, E. Cooperative preservation activities in the UK: findings of a *research project*. Library Management, (20) 3 &4, 1999, p. 220-7.

9. Berthon, H, Thomas, S & Webb, C. *A cooperative approach to building a digital preservation resource*. D-Lib magazine, (8), Jan. 2002.

10. Dublin Core. Available at <http://dublincore.org> .

11. World Wide Web Consortium. Available at <http://www.w3.org>.

12. Paul Miller. *Metadata (2): Towards consensus on educational metadata.* Ariadne, (27) Mar 2001.

13. Allen, Robert B & Schalow, John. *Metadata and data structures for the* historical newspaper digital library. Int. Conf. Inf. Knowledge Management, ACM, New York, 1999, pp. 147-153.

14. Kesong, H & Yongcheng, W. *Methods of keywords and subject concept indexing to Chinese full-text.* Journal of China Society for Scientific & Technical Information, (200, 2, 2001 p. 212-6.

15. Wang, M & Cao, S. *The system for automatic text categorization based on Chinese character vector.* Journal of China Society for Scientific & Technical Information, (19), 6, Dec. 2000, p. 644-9.

16. Chien, Lee-Feng. *PAT-tree-based adaptive keyphrase extraction for intelligent Chinese information retrieval.* Information Processing & Management, (35), 1999, p. 501-521.

17. Qili, W & Zhaohui, L. *A discussion about automatic classification and indexing.* Journal of the China Society for Scientific & Technical Information, (18), 1, Feb 1999, p. 33-36.