Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

Preserving the literary past, looking to the future: the first Hong Kong Literature Database

MA Leo F.H., WONG Rita, LAU Paul

(*University Library System, The Chinese University of Hong Kong, Hong Kong, China*)

E-mail: leo-ma@cuhk.edu.hk; rita-wong@cuhk.edu.hk; Paullau@cuhk.edu.hk

Received Aug. 5, 2005; revision accepted Sept. 10, 2005

Abstract: In the last two decades of the 20th century, there has been an increasing interest in and emphasis on the study of the Hong Kong literature in both the academic and general public in Hong Kong. Recognizing the emergent need of the resources on Hong Kong literature, the University Library System of the Chinese University of Hong Kong set up the Hong Kong Literature Database (the "Database"), which was the first Chinese literature database in the Internet in 2000. The paper will examine how the database is constructed using XML technology and metadata schema. The database also employs Unicode UTF-8 as the internal code. A mapping table for traditional and simplified Chinese characters was created based on Unihan and is used behind the scene so that a user can either input traditional or simplified Chinese characters and retrieval will give both traditional and simplified Chinese characters. Currently 65% of journals use OCR technology so that full-text searching is possible. The Chinese OCR technology will be examined in greater detail. Special features of the Database such as, page-by-page browse mode, position-highlight for full-page newspaper, linking Table-Of-Contents and book jackets from the Library catalogue, etc. are described. The paper will also bring out the problem of massive downloading and compare the state-of-the-art technology and their shortcomings. This paper shows how the Hong Kong Literature Database facilitates future collaboration and data exchange by using open standard, shareable structure and the latest technology.

Key words: Hong Kong Literature, Hong Kong Literature Database, XML, Metadata schema, Database structure, Unicode UTF-8, OCR technology

doi:10.1631/jzus.2005.A1341 **Document code:** A **CLC number:** TP391

INTRODUCTION

In the last two decades of the 20th century, there was an emerging interest in and emphasis on the study of Hong Kong literature among academic scholars. Since the first biennial conference "Conference on Taiwan and Hong Kong Literature (Tai Gang Wen Xue Xue Shu Tao Lun Hui)" held in 1982 in Mainland China, various conferences with diverse themes on Hong Kong literature have been held both locally and overseas in these twenty some years (Huang, 1988; Liu, 1997). The Chinese University of Hong Kong successfully held "the First International Conference on Hong Kong Literature", one of the largest of its kind, on April 15–17, 1999. A total of 69 papers were presented at this conference.

There are increasingly more undergraduate and postgraduate courses on Hong Kong literature offered by universities in Hong Kong. For example, the Chinese University of Hong Kong officially launched the first undergraduate courses on Hong Kong literature titled "Introduction to Hong Kong Literature" in the second semester of the academic year 1999/2000. Numerous Ph.D and Master's theses on Hong Kong Literature are written by the universities in Hong Kong and beyond.

HISTORY OF THE DEVELOPMENT OF HONG KONG LITERATURE

Despite the fact that Hong Kong was a colonial

city of the British Government before 1997, the development of Hong Kong literature in the 20th century tied in closely with the development of literature in Mainland China (Zhao, 2003). Uninterrupted contact and interaction between writers of Hong Kong and Mainland China occurred even during Civil War and the Second World War. The continuous flow of information and writers between Mainland China and Hong Kong greatly influenced the shaping of the "landscape" of Hong Kong literature (Chen, 1991). As Hong Kong is a city with immigrants who arrive and leave, many writers descended from them are now found overseas in the four corners of the world. The Hong Kong Literature Database conducted a user survey from August 2004 to February 2005 showing that over 30% of users are based outside Hong Kong, 30% are based in Mainland China and 40% are based outside Mainland China and Hong Kong. In constructing the database, we have to think of interoperability, and retrieval of literary material in traditional or simplified Chinese characters to suit different needs of the users.

With such a clear perspective, it is easier to target the primary user group.

DESIGN OF THE ONLINE HONG KONG LITERATURE DATABASE

In view of the emerging research interest of academic circles, there is a need to develop a more comprehensive collection of Hong Kong literature to support their academic activities. Targeted at local and global users, the University Library System of the Chinese University of Hong Kong decided in 2000 to develop the Online Hong Kong Literature Database, the first of its kind on the Internet, accessible not only to local researchers and readers but also to the global community. There are three main objectives in building up this database:

(1) To serve the teaching and research needs of the university community and beyond on Hong Kong literature;

(2) To provide easy access to materials on Hong Kong literature anytime anywhere in the world;

(3) To promote Hong Kong literature as a subject discipline to a wider audience globally.

Up-to-date, the Online Hong Kong Literature Database contains more than 330 000 records. Broadly speaking, there are five main categories of materials available in the database, namely:

(1) Hong Kong literary journal articles (243 000 entries);

(2) Book jackets and table of contents of monographs on Hong Kong literature (13 000 entries);

(3) Hong Kong newspaper literary supplement articles (36 000 entries);

(4) Newspaper articles and pamphlet materials (38 000 entries);

(5) Theses and dissertations on Hong Kong literature (400 entries).

Besides these materials, the Database also provides links to useful Web resources such as literary organizations, literary journals, authors' Web page, etc. The key idea is to build up a hub of resources on Hong Kong literature so that researchers can go to the Database to search for materials on Hong Kong literature (Joshi and Vyas, 2002).

Two retrieval interfaces are provided (Fig.1), one for rapid search by beginners while advanced search is for more sophisticated searchers.

(a)

(b)

Fig.1 Two retrieval interfaces. (a) Simple search; (b) Advanced search

ACQUIRING COPYRIGHTS

Like all full-text databases, the most challenging task in developing the Online Hong Kong Literature Database is to acquire the copyright permission from the copyright owner. Our experience showed that it is the most time consuming job to lobby the copyright owner to give us permission to digitize their publications. To-date, more than 30 journal and newspaper publishers have granted us the right to make their publications available full-text to global users. Sixty-five percent of the database contains full-text records.

CONSTRUCTION OF THE DATABASE

To allow interoperability of code and data across and within platforms and application boundaries, XML and Java are used. The database runs on Tamino (http://www.softwareag.com/corporate/prodcuts/tamino/default.asp) XML server. The Tomcat application server was set up to process user requests and translate it into queries over the database. The results of these queries are then transformed into HTML pages and returned to the user via the Web server (Schoning, 2001).

HARDWARE AND SOFTWARE USED

Server and Components are shown in Table 1.

All the above servers are under two levels of network protection: at Network layer firewall software is Netscreen 204 firewall and at Server layer firewall software BlackIce is installed on individual machine (Fig.2).

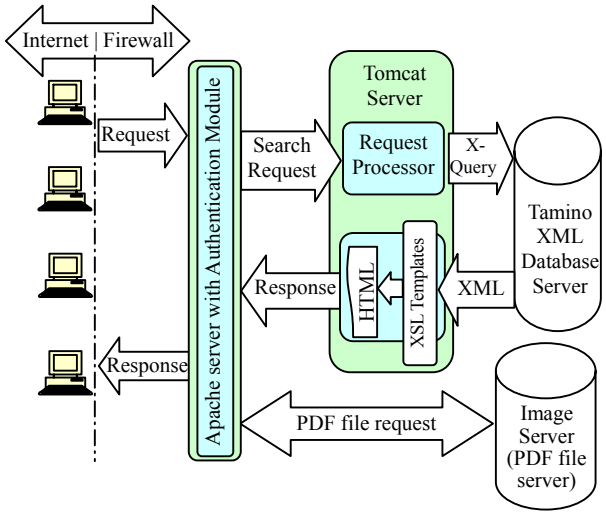


Fig.2 Data flow diagram of Hong Kong Literature Database system

Table 1 Server and components list (on Win2K system)

Server	Hardware configuration	Software	Version
Application server	Dell 2650 CPU: Xeon 2.8 GHz Memory: 1 GB Hard disk: 70 GB (Raid 1)	Tomcat	3.1.3.2
		Apache Web server	1.3.26
XML Database server	Dell 2650 CPU: Xeon 3.2 GHz Memory: 2 GB Hard disk: 16 GB (Raid 1) for OS; 100 GB (Raid) for data	Tamino XML database server	4.2.1 with patch ts42104
ASA database server for membership management	Dell 2650 CPU: Xeon 1.8 GHz Memory: 1 GB Hard disk: 16 GB (Raid 1) for OS; 100 GB (Raid) for data	Sybase Adaptive Server Anywhere	8.0
Image server for PDF file	Dell 2450 + 220S CPU: Xeon 1.8 GHz Memory: 1 GB Hard disk: 101 GB (Raid 5) for OS and system backup; 683 GB (Raid 5) for image file		

MECHANISM TO BLOCK MASSIVE AND SYSTEMATIC DOWNLOADING

In order to prevent systematic and massive downloading, a mechanism is in place to analyze the activities of any IP address that downloads files exceeding our prescribed limit, the IP address will be added to the denial list (Fig.3) (Selingma and Smith, 2004).

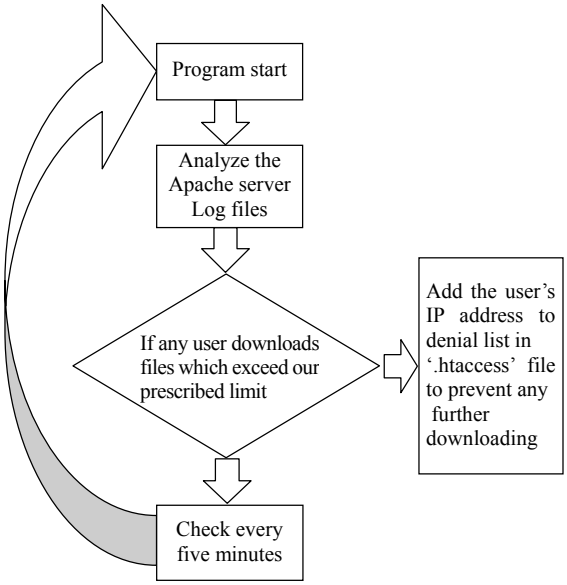


Fig.3 System diagram of the blocking system

METADATA STRUCTURE

Metadata based on Dublin Core (http://www.dublincore.org) is used. Since Dublin Core does not provide property for identifying the link for position highlight and the link for PDF file, two distinct identifiers are added to the Online Hong Kong Literature Database. In addition, Dublin Core description of a journal article has only one property, i.e. dcterms: bibliographicCitation, we find it necessary to identify journal title, journal volume, journal issue, journal identifier and pagination separately. Table 2 shows the relationship (Afonso de Sousa *et al*., 2004).

Sample Records:

```
<?xml version="1.0" encoding="utf-8"?>
<article>
<title>牛津公園感秋</title>
<creator> 錢鈡書</creator>
```

Table 2 HKLIt database to Dublin Core mapping

Level	Property	DC Element
Article	Subject	Dc.Subject
	Title	Dc:Title
	Creator	Dc:Creator
	Key	Dc:Desciption.key
	Abstract	Dc:Description.abstract
	Fulltext	Dc:Description.fulltext
	Type	Dc:Type
	Format	Dc:Format
	Loc_id	Dc:Identifier.id
	URI	Dc.Identifier.uri
	URL	Dc:Identifier.url
	Ppdflink	Dc:Identifier.ppdflink
	Pdflink	Dc:Identifier.pdflink
	Language	Dc:Language
	Search date	Dc:Date.search_date
	Display date	Dc:Date.display_date
	Journal_title	Dc:Relation.Journal_Title
	Journal_volume	Dc:Relation.Journal_Vol
	Journal_issue	Dc:Relation.Journal_Issue
	Journal_id	Dc:Relation.Journal_Identifier
	Pagination	Dc:Relation.Pagination
	Publisher	Dc.Publisher

```
<subject>HKLit</subject>
<description>
   <category>古典詩詞</category>
   <abstract/>
   <fulltext/>
   <key/>
</description>
<publisher>大公報</publisher>
<contributor/>
<date>
   <search_date>1992-07-08</search_date>
   <display_date>1992 年 7 月 8 日</display_date>
</date>
<type>
   <source_en>newspaper</source_en>
   <source_zh>報章文藝版</source_zh>
</type>
<format>text/pdf</format>
<identifier>
   <loc_id>16631</loc_id>
   <pdflink>T002a</pdflink>
   <ppdffile/>
   <uri/>
```

```
   <url/>
</identifier>
<source/>
<language>cn</language>
<relation>
   <journal_title>《大公報·文學》</journal_title>
   <journal_vol/>
   <journal_issue>第二期</journal_issue>
   <journal_identifier>68</journal_identifier>
   <pagination>第 18 頁</pagination>
</relation>
<coverage/>
<rights/>
</article>
```

SPECIAL TECHNICAL FEATURES OF THE DATABASE

1. Ability to treat traditional Chinese characters and simplified Chinese characters as one

Since the source materials from Mainland China will be in simplified Chinese, whereas from Taiwan and Hong Kong will be in traditional Chinese characters, there is a need to show all the results no matter whether the searching is in traditional Chinese characters or simplified Chinese.

Unicode UTF-8 is used to handle traditional Chinese characters and simplified Chinese characters. Since each traditional Chinese character and simplified Chinese character has different internal code, a mapping table is needed.

Behind the scene, a mapping table (Table 3) for traditional Chinese and simplified Chinese characters based on Unihan (Unihan Database, http://www.unicode.org) is built. For example the words in Table 3 are the same.

Table 3 Traditional Chinese to simplified Chinese mapping

Traditional Chinese characters		Simplified Chinese characters
俠	→	侠
亂	→	乱
萬	→	万
與	→	与
麼	→	么

Querying the title with '*乱*' would return all records that contain a word '乱'or '亂' (Fig.4).

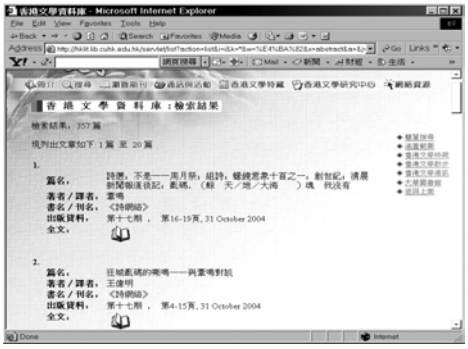


Fig.4 Result screen

2. Page by page browse mode

To allow a user who would like to flip through pages of newspaper, a page by page browse mode is added that can be a full page mode, 2-pages mode, 8-page mode (Fig.5).

3. Position highlights

A page of newspaper may have several articles on a page. It is important to know the position of the article and its actual size on the page. Position highlights will help (Fig.6).

4. Table-Of-Contents and book jacket

Table-Of-Contents from the book and the scanned book jackets, are attached to the bibliographic record in the Library OPAC. It would be useful to link them to the Hong Kong Literature Database, thus giving users additional information (Fig.7).

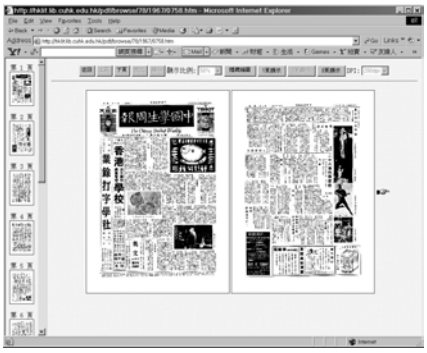
OPTICAL CHARACTER RECOGNITION

Sixty-five percent of journals and newspaper have digital full text. Chinese OCR software used is from Tsinghua Tong Feng. It can achieve 95% accuracy. A number of errors occur in recognizing punctuations and special characters. Considerable time and manpower is spent in correcting the errors (Fan, 2002).

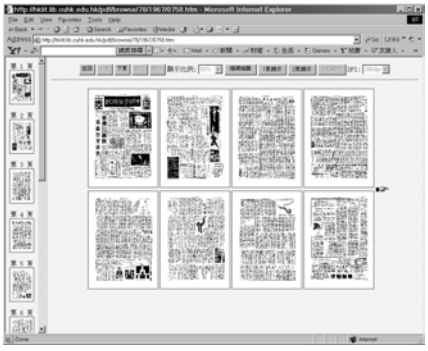
USAGE

The database proves to be very popular. About 5

(a)



(b)



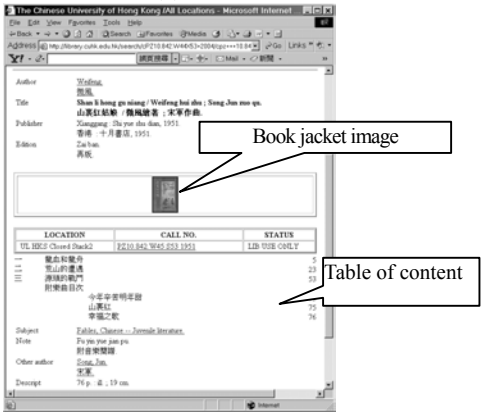
(c)

Fig.5 Page by page browse mode. (a) full page mode; (b) 2-pages mode; (c) 8-page mode

Fig.6 Position highlights

(a)

(b)



(c)

(d)

Fig.7 Table-Of-Contents and book jacket. (a) Search screen for "韋婭"; (b) Resulting page display; (c) Table-Of-Contents display at the Library OPAC; (d) Miniature of full-sized book jacket

million accesses since its production in 2000 at an average of 200000 hits a month (Fig.8).

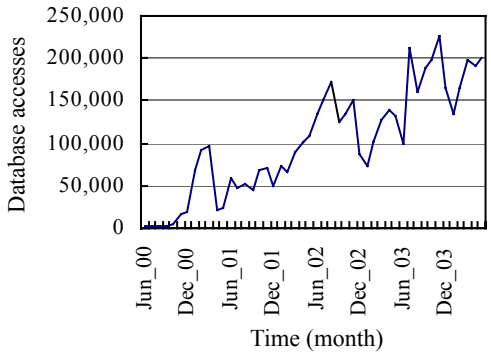


Fig.8 Usage statistics since June 2000

CONCLUSION

The Online Hong Kong Literature Database proves to be a popular database with over 2 million accesses a year. With the structure of the database in an open standard, language independent and sharable structure using DTD, it is hoped it will enable inter-operability and allow other databases to share the information and data.

References

Afonso de Sousa, A., Duarte, P., Pereira, J.L., Carvalho, J.Á., 2004. Topics on XML Data Storage and Management. Proceedinds of the IADISac2004, Lisbon, Portugal.

Chen, B.L., 1991. Hong Kong Literary Critisism. Joint Publishing Company, Hong Kong (in Chinese).

Fan, J., 2002. Off-line Optical Character Recognition for Printed Chinese Character–A Survey. http://www.ee.columbia.edu/~junfan/E6880_Final5.pdf.

Huang, W.L., 1988. Critique of Hong Kong Literature I. Wah Hon Publishing Co., Hong Kong (in Chinese).

Joshi, H., Vyas, M., 2002. Framework for a federated digital library. http://dspace.inflibnet.ac.in/bitstream/1944/353/1/04cali_45.pdf.

Liu, D.H., 1997. Hisotry of Hong Kong Literature. Hong Kong Writers Publishing, Hong Kong (in Chinese).

Schoning, H., 2001. Tamino–A DBMS Designed for XML. Tamino–17th International Conference on Data Engineering (ICDE'01), p.0149. http://www.comp.nus._edu.sg/~jaga/papers/Arch-Tamino-ICDE01.pdf.

Selingma, P., Smith, S., 2004. Dectecting Unauthorized Use in Online Journal Archives: A Case Study. http://www.ists.dartmouth.edu/library/securing-systems-software/duu1004.pdf.

Zhao, X.F., 2003. Fiction Hong Kong. Joint Publishing Company, Beijing (in Chinese).