Abstract

Since the implementation of the "NCL System for Retrieval of Digitized Images of Ancient Books" much work has been spent on further development of its functionality, taking human interface design and ease of use as primary consideration. This article introduces new features that are available since the completion of this development stage in 2002, namely the handling of missing characters, conversion between traditional and simplified characters, browsing of scanned images, full text search of prefaces and postscripts, name authority control, and the conversion of MARC format data into Metadata format.

Keywords

Handling of missing characters, Synthetic Chinese Character Format, Conversion between traditional and simplified characters, Browsing of scanned images, Full text search, Name authority control, MARC format, Metadata

1 Preface

The NCL Ancient Books and Documents Digital Archives Project will digitize the National Central Library's precious collection of rare and ancient books as well as its bronze and stone rubbings, all together more than 12,000 pieces. During initial development of the system in 2001 the project completed a "Metadata Catalog of NCL Rare and Ancient Books" as well as a "Retrieval System for Digital Images of Ancient Books". Bibliographic information on ancient books held at the NCL could thus be made available in the form of Metadata.

To improve the usability, flexibility and scalability of the system, the NCL introduced a plan for further development of the system's functions in early 2002. This plan was accomplished with great success at the end of 2002. The following is an introduction to the features of the system as they were completed last year, namely the handling of missing characters, toggling between simplified and traditional characters, browsing of images, full-text search of prefaces and postscripts, the implementation of personal name authorities, and the capability to convert MARC format into Metadata format.

2 Introduction to the system

2.1 Handling of missing characters

One of the major problems in the process of digitizing Chinese sources is the problem of missing characters. On the one hand a routine for creating missing characters has to be in
place before starting to build the database. Only thus can the quality of the data be ensured. On the other hand the user has to install special software that is able to handle these characters before searching the database. Only then will he be able to read the contents without problems. This is a major obstruction in the usability of the system. Furthermore, new characters created by different systems are usually not compatible, which causes difficulties in data exchange and data management. In addition there is only limited space available for newly created characters, which results in character code conflicts if this space is exceeded. All this adds to the problems of handling ancient books and documents.

Since the data of this system are imported from the electronic catalog of the NCL, which uses CCCll encoding, they have to be converted to Big5, resulting in a large number of missing characters. To avoid the problems of creating new characters, and to make usage of the system possible without having to install a special character file or any other additional software, we decided not to create an own set of new characters. The system uses the following method:

1. Data saving: While importing data from the library's catalogue, missing characters in Big5 encoding are replaced by \{CCCII code number\}.

2. Recording occurrences of missing characters: A software is used to record every occurrence of \{CCCII code number\} within the data set. Thus it will be possible to know which missing characters appear how often and in which records. It will also be possible to identify the number of missing characters in each record of the data set. This information is important for optimizing the handling of missing characters during data management.

3. Management of missing characters: For every single character the system operator can choose between offering a similar Big5 variant for the missing character or using the Synthetic Chinese Character Format (SCCF) of the Academia Sinica [fig. 1].
4. Results handling: When delivering the results to the end user, the system is able to identify missing characters and to offer either a similar Big5 variant or an image created by the SCCF (a SCCF Java Applet is integrated into the Web page) [fig. 2].

5. Retaining flexibility: Since there is currently no accepted standard for handling missing Chinese characters, and since computer soft- and hardware is developing rapidly, the method used by this system is adjusted to the current working environment while still enabling transfer to future standards at the lowest costs possible:

   A) With Win2000 and WinXP replacing Win98 and becoming general standard, the system's coding will have to be switched to Unicode. At that time data encoded in Big5 can automatically be converted to Unicode. Next, the records about missing characters will be searched for characters already available in Unicode; these can also be converted in a batch. Since SCCF does not currently support Unicode, missing characters not available in Unicode will automatically be converted into an SCCF image file. Instead of using the SCCF Java Applet this image will be made available through a link on the Web page.
Fig. 2: On this search results screen, the character 恒 has already been transformed into an image. When moving the mouse on the image, the SCCF of the character is shown.

B) If by that time SCCF has become Unicode compatible, method A) will still be used, implementing the Java Applet instead of a link for displaying missing characters not available in Unicode.

C) In both cases described above the main work involving human labor is to check whether missing Big5 characters are available within the Unicode standard. Since there are only about 700 such characters it is estimated that this work can be done manually. Writing a software for automatically converting codes and missing characters is also not very difficult. For these reasons the current method for handling missing characters has been applied.

2.2 Automatic conversion between traditional and simplified characters

The coding used in the current system is Big5. However, to improve the usability for Chinese users all over the World a functionality to change from traditional characters to simplified characters has been added. The user can thus select the encoding of the characters displayed, which significantly improves the service character of the project. The following features are implemented:

1. Within the interface the user can select between a traditional characters edition and a simplified characters edition. The system will then automatically use Big5 or GB encoding to display the contents. Thus the user does not have to worry about reading either traditional or simplified characters that might not be familiar to him [fig. 3 and 4].
2. When the user selects the simplified characters edition he has to enter his search terms in GB encoding. The system will automatically show the results in GB encoding. Using the simplified characters edition will thus be in no way different to using the traditional characters edition, search results will also be identical, only that characters are displayed in different encoding.

3. Having implemented the capability of automatic conversion between traditional and simplified characters it will only be necessary to maintain one data set encoded in Big5. When a user is searching in simplified characters, the system will automatically and instantly perform conversion into GB. There is no need to maintain two different versions of data sets, one with traditional and one with simplified characters, which saves disk space. There is also only one piece of software needed instead of having to cope with two, one each for the traditional and for the simplified characters edition. This will reduce maintenance costs.
2.3 Browsing of scanned images

Besides providing a facility to search bibliographic records on ancient books, the system also offers scanned images of these books for the user to browse. Using this convenient Web service researchers do not have to rely on the cumbersome microfilmed versions anymore.

There are various possibilities to browse these images:

1. Searching:
   
   a. After entering a search term, the system will produce a list of hits in front of which the following note will appear: "此篇目有影像資料，如要看影像請登入" ("There are scanned images for this title; please enter the system for viewing") [fig. 5].
   
   b. If you switch to the full view mode of the record, the system will produce not only the biographical data of the book but also the titles of all chapters (卷) within. Upon selecting one chapter by mouse click all headings contained within this chapter will be shown. After selecting one heading the corresponding scanned image will be shown [fig. 6-8].
   
   c. Using the advanced search or the refine search interfaces search may be limited to scanned images only by selecting "yes" (是) at the "Search scanned images only" (只查有影像) option [fig. 9].
Fig. 5: Login ID and password have to be entered before viewing the scanned images.

Fig. 6: After entering login ID and password chapter titles are open for selection.
Fig. 7: Display of headings within chapter five; any heading or any other chapter title may be selected here

Fig. 8: Scanned image of the first page of "贈朱射陂考最序" (Preface presented to Mr. Zhu from Xiepi in honor of his excellent performance)
Fig. 9: To improve search efficiency it is possible to search for scanned titles only

2. Browsing:
   a. If a search term is entered within the browsing interface books available as scanned images are available immediately.
   b. Scanned books can be browsed by subject (hierarchical or by category), edition or headings. Since data in the system are already arranged according to category, the user can easily do research browsing through related topics [fig. 10].

3. When browsing through the scannings, the following features are available: zoom in, zoom out, rotate, print, select page [fig. 8].

4. Besides viewing the scannings of single chapters the system also offers the possibility to browse through the complete volume of the electronic ancient book. By clicking on the title of a subchapter the system will automatically jump to the according page of the scanning. This functionality is similar to browsing through the printed edition, thus greatly enhancing the usability of the system [fig. 11].

2.4 Full text search of prefaces and postscripts

The system also offers a full text version of prefaces and postscripts. These can be read following the steps below:

   1. Search: the interfaces for advanced search and refine search offer a field for full text of prefaces and postscripts. After entering a search term the system will
perform a full text search within all available prefaces and postscripts and show a table of results [fig. 12].

Fig. 10: Selecting the browsing feature (影像瀏覽) the user can browse through related resources predefined within the system

Fig. 11: Electronic ancient book where you can automatically browse the pages by selecting the title of a subchapter
Fig. 12: If there are full texts of prefaces and postscripts as well as scanned images, both will be listed simultaneously for the user to select.

2. Results screen: on the detailed results screen the system will show the full text of the prefaces and postscripts of that title together with a scanned image, if available, enabling the user to compare both at the same time. Especially in the case of prefaces or postscripts written by collectors in cursive script (草書) inexperienced readers will not be able to decipher these. Since such texts usually also lack punctuation marks they are very hard to read. After adding this feature to the system the user can easily read and examine these texts [fig. 13].

3. New feature: Beginning with 2003 the system will automatically show both full text and scanned images of prefaces and postscripts without the user having to select this option manually.

2.5 Name authority control

Since bibliographic data are entered as they appear on the original work, names of authors or writers of prefaces or postscripts etc. might appear in various forms. To allow search for all works of one author, the system will perform name authority control during the process of search. This is how it works:

1. Creating name authorities: name authorities are imported from the catalog of the NCL and then indexed.

2. When performing a search names in the personal name field are handled first and compared with the name authority file. In case of an occurrence all name variants of the person such as his zi, hao, sobriquet, etc. will be used as search value. Example: the Yuan dynasty writer and calligrapher Zhao Mengfu 趙孟頫 (1254-1322)
is known under several different names such as Ziang 子昂, Wenmin 文敏, Shuijinggong 水晶宮, Shuijinggong daoren 水晶宮道人, Songxuezhai 松雪齋, Songxue daoren 松雪道人, Oubo 歐波, Weiguogong 魏國公, Ouboting 歐波亭, Zhao Ziang 趙子昂, Zhao Wenmin gong 趙文敏公, and Zhao Songxue 趙松雪. If the user searches for Zhao Ziang in the personal names field the system will first find out that this name is a sobriquet of Zhao Mengfu and will then use his original name together with all his name variants as search value [fig. 14, 15].

Fig. 13: Full text and scannings of the prefaces and postscripts can be compared directly

Fig. 14: Entering "Zhao Ziang" in the personal names field
2.6 Metadata

The system’s data are imported from the catalog of the NCL, which is using the MARC format. Relying on the chapter on “Metadata and sample records for ancient and rare books” of the first edition of the *Compendium of the Chinese Metadata format* (ed. by the NCL Metadata research group) as reference, necessary Metadata and their corresponding tag fields have been defined. Furthermore, based on the original ancient books and the Catalog of Rare Books at the NCL, additional Metadata fields that do not appear in the MARC format as well as a DTD for ancient book records have been defined as reference for entering and verifying records as well as for import and export of data (cf. appendix 1-4). The DTD is mainly based on the fifteen core elements of the Dublin Core set. However, in order to better describe the various attributes of the material and to express the meaning of these descriptions more accurately, extra element qualifiers have been added to the respective fields. This increases its application range while at the same time following international standards.

After consulting other digital library projects, the capability of XML import and export has been added to the capability of converting MARC format into DC Metadata format [fig. 16, 17]. This facilitates data exchange with other institutions. Since the DTD does not define the format of the various fields, nor any necessary data conversion of data contained therein, nor whether or not they should be indexed, the system operator may freely define the character of any field as date, time range, file, URL, etc. According to the need of the user the system operator may also modify the setup as to whether or not any field should be repeatable, mandatory, automatically created, indexed, or shown on the short list or in
full view mode. Thus the usability of the system can be optimized without any software changes [fig. 18, 19].

Fig. 16: Import functionality: XML data compliant with the current DTD for ancient books can be imported; the system will automatically ingest these data

Fig. 17: Data export: after selecting the name of the DTD and entering the system's ID range or time range of last modification, the data will be exported
Fig. 18: List of fields for the ancient book database that offers detailed configuration options

Fig. 19: Refining configuration of a single field
On completion of the digitization of important holdings of the NCL the project intents to add its complete holdings to the database system described above. By then these resources will not only be available to the academic community for research, but the general public as well may enjoy the richness of China's ancient books and documents. The texts on bronze and stone rubbings will continue to be passed on to future generations, and our age old art and culture will be carried forward without borders. All kinds of valuable online content can be offered for the use of Chinese throughout the World, constituting a precious digital historical resource for the generations to come.

Bibliography

1. 中文詮釋資料（METADATA）格式彙編 (Compendium of the Chinese Metadata format) / NCL Metadata research group [ed.] (2000)
2. metadata、影像規格、編碼原則 (Metadata, standards for scanning, and rules for coding) / http://readopac.ncl.edu.tw/ndap/rar/ndap-rar-rul-01.doc
3. 古籍善本 MICI-DC 及 Metalogy 相關設定 (Metadata Interchange of Chinese Information - Dublin Core for ancient and rare books and Metalogy setup) / http://datas.ncl.edu.tw/catweb/metadata/mici00110819.PDF
4. 國家圖書館善本古籍與 CMARC 欄位對照表 (CMARC table for NCL ancient and rare books) / http://readopac.ncl.edu.tw/ndap/ndap-doc-03.htm
5. 數位圖書館分散式協定(SOAP, OAI, Open URL) (Standards for data distribution in digital libraries - SOAP, OAI, Open URL)
6. 資訊檢索服務與協定(CNS-13416, Z39.50) (Standards for information retrieval - CNS-13416, Z39.50)