Transformation of Palace Archives of Ming and Ching Dynasties onto CD-ROM and Internet : An Exemplification of Building up Chinese Databases in Taiwan

Tracy Shih

Transmission Information System Co., Ltd.

Taipei, Taiwan, R.O.C.

1. Preface

While the new era of Internet has been dawning on us, nearly every sector in Taiwan society, the government, academic institutes, industrial enterprises, etc., is devoted to reaching the goal of making Taiwan a true Sci-tech island. One of our latest passions and common efforts has been the development of various network resources, and consequently more and more academic and commercial databases are becoming available via WWW nowadays. Since Transmission Information System Co., Ltd. (TISC) has been for years providing foreign database services to academic circles in Taiwan, it is among those first recognizing the need of a Chinese information retrieval system. Accordingly, TISC made it one of its major business objectives years back to setup and publish Chinese/English databases. It has then successfully developed the "Transmission Full Text Retrieval System (TTS)," and introduced it to the public in 1993. Ever since then, TISC has assembled and launched on the market quite a number of different databases equipped with TTS in various electronic formats, such as CD-ROM, Internet, and Intranet.

With its outstanding Chinese/English full text retrieval system, TISC had the pleasure to be chosen by many local government branches and schools to digitize their current databases. One of our typical major undertakings of this sort was the work of digitizing the Palace Archives of Ming and Ching Dynasties, entrusted by the Institute of History and Philology (IHP) at Academia Sinica in 1995. We are here more than happy to share with you of what we have experienced under this meaningful mission in

the past four years.

2. TTS Full Text Retrieval System

A retrieval system capable of meeting user's needs is the core requirement for building up any database. Therefore, TISC has put quite a lot effort into modifying user interface and functions while digitizing the aforesaid archives. To further elaborate on this point, a general introduction on the current retrieval system known as TTS is provided here in the following.

2.1 What is TTS

TTS (Transmission Text Retrieval System) is a product of our long-term R&D efforts and is a highly efficient and powerful system specially designed for Chinese language. Its compatible feature allows it to work on various platforms and to retrieve databases of all disciplines up to millions of records from CD-ROM and Internet databases. It is one of the most prominent technologies in the Chinese processing environment.

2.2 Features of TTS

- It is a full text retrieval system compatible with Chinese in the simplified forms, Traditional Chinese, English, and Japanese.
- It has powerful retrieval functions, such as cross-field retrieval, specified-field retrieval, index browsing and cross-database retrieval.
- It has mature image integration capacity.

2.3 Major functions of TTS

- Operation Platform
  User End: Windows/NT, Win95/98, or more advanced.
  Server End: Sco-Unix, Sun-Solars； IBM-Aix, Linux, Free BSD, NT.

- User Interface
  WWW interface can be connected for retrieval via Netscape.
  Navigator or Microsoft Explorer.
- Retrieval Functions

- It has full text retrieval capability.
- It has cross-database retrieval capability.
- It has Bullin logic (AND、OR、NOT) calculation retrieval capability.
- It provides retrieval capability on specified fields.
- It provides field index.
- It keeps retrieval history.
- It allows free combination of retrieval terms.
- It allows free setting of record fields to be displayed, printed or downloaded.
- It allows free choice of printing and downloading all or partial retrieved result.

3. Digitizing the Ming and Ching Palace Archives

3.1 Background and Current Situation

The progress of this project can be divided into three phases.

Phase I: File revising, indexing and copy printing

The Archives of Ming and Ching Dynasties at the IHP contain over 310,000 items with more than 3 million pages in printed format. They have been under an everlasting process of filing, indexing and shelving routines ever since the Institute moved to Taiwan from the Mainland China. In addition to having managed to preserve these historical documents rather properly, IHP ventured itself into collaboration with a local printer, Linking Publishing Co., Ltd. (LPCL) to have those revised documents become available in printed materials for researchers at large all over the world.

Phase II: Indexing and Digitizing

It is a common sense that the best way to preserve valuable documents is to reduce the frequency of their being physically accessed. Thanks to the latest advancements in technology, the provision of scanning the documents into electronic images to users in place of direct physical contacts became feasible and, therefore, the main objective of this phase of the project. With this in mind, IHP also asked TISC to work on the scanning of all those documents on a page-by-page basis, besides the building up of a retrievable electronic database for the bibliographic index that has already completed, by the way. As soon as the proof reading of all images is done, TISC will further integrate the system by linking the indexing and images together, and make them accessible on the Intranet.

Furthermore, by taking into account of avoiding the estimated huge production cost, yet gaining on the ground of better user-friendliness, LPCL asked TISC to transform into electronic images a selection which was originally to be printed by LPCL. The same client also requested a multi-function CD-ROM database that would integrate indices and images. This outstanding and very useful database is expected to be available on the market real soon.

Phase III: Building up the on-line service system

At present, IHP has managed to post about one million pages of the full text images of its Ming and Ching Palace Archives on its Intranet. A CD-ROM version will be available in the marketplace shortly. In the meantime, the old document processing routine is still kept on going at IHP as usual in indexing, scanning and re-organizing. Eventually, IHP would like to have this service available to the public. However, due to the large capacity and unique historical values of these archives, IHP is implementing a system to facilitate both administrators and users in terms of more convenience and online security in order to fully benefit from the digitalization of these documents.

3.2 Problems encountered

1) Document Scanning

Since the main purpose of the scanning is for the sake of public access, its legibility is thus the bottom line as well as the first priority. At the same time, in consideration of its large volume, limited production budget and storage capacity, as well as the technical simplicity and readiness, a black and white format is adopted as the major approach above others with gradation and full color. However, due to the timeliness of these documents, the scanning is far more difficult when compared to jobs involving other documents. There are two major reasons for this:

Poor quality of the original copies

Many of the original copies are shattered in pieces and require re-binding before scanning can be carried out. Other common problems are 1) the background being too dark, 2) paper being stained or perforated by pests, and 3) paper being just worn out from over handling in the past. All these inevitably reduce the legibility of the output. Many a document has to be scanned four or five times before a decent enough copy is obtained. In order to achieve the best result, sometimes a full-color copy is preferred instead of the normal black and white format. Consequently, the process turned out to be a very time-consuming and labor-laden practice, and it becomes very hard on the worker to master such expertise, both of which have unfavorably resulted in a much higher production cost.

Complexity in Proof-reading

Each scanned output has to be carefully examined by our technicians to see if re-scanning is necessary to follow. If the verdict is affirmative, needed modifications to improve the quality of subsequent output, like places where could use some masking and a better layout angle for instance, are decided at the same time. Moreover, a copy would still not be final without a satisfied second proofreading.

2) System Developing

Initially, the index and image databases were built under the now dated DOS/Windows. They eventually have been upgraded to a Web interface of an Internet/Intranet environment. In addition, in the name of protecting copyrights, a special program has been developed for such security purpose. As both the hardware and operational system are rapidly changing, the retrieval systems have to be upgraded constantly to fill the gap. Foreseeing an increasing need of online service system, we will from now on be more committed to the development of integrated systems with features such as image description and security control.

3) Image managing

It is infeasible or impractical to have the document scanned in numerical order as long as a team rather than one individual does the work. However, in order to facilitate the proofreading, the image files have to be centralized later for re-numbering. The size of the image files will sometimes be as big as over a hundred GB that takes some hardware with tremendous capacity.

3.3 The publishing of CD-ROM

Since the CD-ROM version of the Ming and Ching Palace Archives, which we assembled on behalf of LPCL, went on the market, different operation systems on retrieval and image processing have been adopted. The CD-ROM version can thus be subdivided now into PC version and network version. The publisher's logo as well as the disclaimer statement has been added to the image files. Moreover, each image file has been attached with its primary document title by means of bibliographic index so that both will be displayed whenever the image is presented or printed.

3.4 Future Development -- Online service system

After these consecutive years of keen efforts, TISC has been able to upgrade the digitized archives into a retrievable electronic database available on the IHP Intranet. However, due to security reasons and the extent of network transmission, its availability is limited only to IHP itself. Security measures have to be further enhanced for lay people access to put an end to abuses of these invaluable documents. Nowadays, image files are mostly protected with watermarks and/or security codes. However, on top of these, the provisional service of image files of such historical documents calls for a complete system of copyright management and monitoring. It will cover areas like usernames, copyrights, authorization administration, and all kinds of records and statistics on usage.

Although the web interface has been very user-friendly, protection of the image files is relatively tedious. Some suggestions in this respect for the future online service system dealing with the Ming and Ching Palace Archives are listed below:

User End:

- To apply for usage authorization from the server.
- To retrieve the bibliographic database under proper authorization.
- To access the original image under proper authorization.
- To display, print or download the image of the original under proper authorization.
- A password is required to open each image file at the user's end to prevent any breach of copyright.

Service End:

- To build up a bibliographic index database with retrieval function.
- To set limitations on the retrieval usage, image output and download, validity of password, etc.

- To maintain usage information.
- To encode the image files (filing coding).
- To add a second encoding for the image files (usage coding).
- To limit the number of records downloaded within a consecutive period of time .
- To monitor the number of records being downloaded by authorized users.
- To provide records and statistics on usage.
- To prepare relevant management reports.

4. The application of TTS on Chinese databases in Taiwan

TTS has been in use for more than six years since its creation. It has been upgraded constantly to meet the international standards in respect of the system design, functions and efficiency. Aside from the aforesaid project for Ming and Ching Palace Archives, we have also processed many other historical documents and publications of databases in either Japanese or the simplified Chinese. Below are some of the database products for our major clients:

4.1 Major cases of database building up

1) National Cheng Kung University
   - Index to Oracle-bone or Shell Writing Rubbings and Full Image Database.
2) Academia Sinica Fu Ssu Nien Library
   - Bibliography of Rare Books and Full Image Database.
3) National Central Library - The Union Catalog of Rare Books in Taiwan.
4) National Central Library - Bibliography of Research on Han Philosophers.
5) National Central Library - Bibliography of Research on the Classics.
6) National Central Library - The Union Catalog of Ming Dynasty Writers' Works in Taiwan.

4.2 Database products

1) National Taiwan University Research Library
   - The Taiwan JIHO Full Image Database.

- It is the first digitized publication of historical studies of Taiwan.
- It composes of two interfaces, in Chinese and in Japanese as well.
- Its features include full-text retrieval and image linkage.

2) China Renda Social Sciences Information Center
   - Chinese Social Science Digest Data on CD-ROM.

* It is compatible for both the simplified and traditional Chinese.

3) Shanghai Information Institute of TCM
   - Chinese Materia Medica Database

* It is compatible for both the simplified and traditional Chinese.

5. The exemplification of international collaboration in setting up the Chinese database

Database for the classification catalogues of Chinese documents at Institute of Oriental Culture (IOC), University of Tokyo

5.1 Background and current situation

In 1971, the Institute of Oriental Culture (IOC) published a printed classification catalogue covering 20 thousand volumes of their collection in Chinese. The collection includes Jing (Classics), Shi (History), Zi (Philosophy) and Ji (Literature). All entries are in the traditional Chinese since they are all ancient Chinese documents. At the time when IOC decided to digitize the catalogues into an electronic database, the BIG5 internal code system was obviously the best choice for the traditional Chinese working environment. As a result after a series of evaluation, TISC was appointed in 1998 to work on this project which includes data filing development and design. By the end of 1999 the part of Jing (Classics) and Shi (History) were completed in their setups and

can now be retrieved online from the server of IOC.

5.2 Major Features

1) Data Entry

Since the utmost objective of the digitized database is for retrieval purposes, the data entry has to be set in the format of on-going field layouts. This, however, is accompanied by a major drawback of insufficiency to fill in all necessary data. In contrast to this, the ancient approach appears comparatively more flexible in terms of consecutive description of details since the concept of field layout was not conceived as yet. But this feature has added quite a burden to the network in sorting out each and every data entry. Moreover, additional fields or modifications are necessary to fit in for some specific data especially in the preliminary stages. Much time and manpower were thus spent for the process of setting up and the proofreading that follows.

2) Resolution for obsolete characters

Quite a number of ancient characters cannot be converted under the BIG5 code system due to their obsoleteness. The resolution for those long-missing characters is to encode each of them with a specific number primarily based on a unicode system. In case there is no unicode available, code from the Ta Han He Dictionary is to be adopted as the last resort. These coded characters will then be able to be converted as the unicode environment is to be set up.

3) Retrieval by Chinese romanization

This online catalogue database is expected to be accessible to all researchers worldwide in the near future. Therefore, the major titles and authors have undergone a Chinese romanization under a separate system. For users who are not familiar with BIG5 system, they can now retrieve the catalogue information by simply keying in the alphabets instead.

5.3 Future Development

IOC considers the setup of this database as one of its long-term projects. Meanwhile,

the Zi (Philosophy) and Ji (Literature) collections will be ready shortly, adding to the existing electronic database of Jing (Classics) and Shi (History). All these will be transferred to an online database under a unicode system once unicode becomes prominent at the users end. On the other hand, IOC is also seeking after the contribution of Chinese catalogues from other institutes and organizations in Japan, as well as through collaboration opportunities such as data exchange with other, especially Chinese spoken, countries.

6. Conclusion

The prevalence of education on information related subjects and the readiness of infrastructure of relevant facilities have contributed much to the growth of information industry in Taiwan which is right on its way of becoming an info-tech island in the 21st Century. As part of this bright scenario, we are committed to take a leading role in the forefront of the information industry with competitiveness in expertise, efficiency and creativity. We are much honored to have engaged in the mainstream of setup process of the Chinese database in the past. In the meantime, TISC has many similar requests from reputable academic institutes in Taiwan for future online database setups, production and publication of CD-ROM databases, and other products featuring processing of various image files.

After all, TISC has accumulated so far a vast pool of hard-to-come-by, precious, and very unique experiences in the field of building Chinese database, not to mention the pleasure we are having in the course of assisting our treasured clients to solve problems of various nature. TISC is aiming at a broader spectrum of collaboration on an international level from now on. In other words, globalization of the Chinese database eventually will be our number one major business concern.

The End