

Digitization and Knowledge Base Construction of Archives: A Case Study of *Tan-Hsin Archive*

Shu-Fen Hung, National Taiwan University Library, 1 Sec.4, Roosevelt Rd., Taipei, Taiwan, R.O.C. 106,
shufen@ntu.edu.tw

Abstract: A well-planned digitization project could be of use to establishment of a sophisticated searchable database and to more of other useful applications. The author suggests that strategic planning should be done at the beginning stage and the specifications be seriously followed during the whole working process. The author states that for a digitization project, the standards and the strategic planning should cover image digitization, metadata schema, and specifications for metadata, etc. Taking the Judiciary *Tan-Hsin Archive Collection* as an example, considerations and correspondent approach as follows are discussed and demonstrated: useful information to be included in the inventory list, selection of digitization methods, file formats and resolution, reasons for the creation of typed full-text and utilization of it, mechanism to achieve automatic or semi-automatic link of related different types of sources, creation and maintenance of a precise metadata schema, tasks to be done in supporting of programming, and methodology to incorporate the educational e-learning objects into the main knowledge base.

Keywords: image digitization ; knowledge base construction ; archive collection ; Metadata ; digitization specification ; digitization workflow

General introduction about *Tan-hsin Archive*

Contents of the *Tan-Hsin Archive*

Tan-hsin Archive is a collection of administrative and judicial documents from central to very northern part of Taiwan, namely, Hsin-chu Hsien, Tamsui Ting, and Taipei Fu in Qing Dynasty dating from 1776 to 1895. This archive contains nearly 20,000 documents of about 1,600 cases. It is the most sizable and complete collection of government documents from Taiwan during the Qing Dynasty, and it is also the one that covers the longest time span. The archive is invaluable for the study of Taiwan history because of its plethora of materials related to government, economics, society, and agriculture during the Qing Dynasty. For the study of Chinese history it is also a world-renowned archive that provides important information about the Chinese judicial system.

The importance of sharing the digitization experience of *Tan-hsin Archive*

Due to the critical physical condition and complicated contents, standards of both image digitization and metadata specification are not easy to decide. The sound NTUL's specifications are acquired from high costs of both monetary and human effort investments. The NTUL's experiences could be useful for other similar archives of the same history era held in China. Having more institutions that hold same kind of documents adopts the knowledge base schema and same set of specifications would ease the share and integration of the resources. For most institutions, if the person in charge of the specification and workflow of digitization project could learned about the most complicated and difficult case, he/ she will be capable in planning for other easier cases.

History about the digitization and application of the archive

The digitization effort for the *Tan-hsin Archive* began at a very early stage of digitization activities in Taiwan. The first experimental digitization project was conducted as a sub-project of the Digital Museum and Library Project sponsored by National Science Council, Taiwan in 1987-88. During the two-year project, the awarded fund was only enough for a small portion of the archive, therefore, only selected document had been digitized in the project. From 2001, the government launched a National Pilot Project for NDAP. The NTUL was the only one academic institution that awarded funding among other eight national institutions. In conducting the project under the Pilot Project, the library has finished the image digitization of the whole archive, reviewed and revised the metadata schema and started to establish a complete metadata knowledge base for the archive. The Library also continued the full-text transcribing project and make use of the full-text in the *Tan-hsin Archive* database. The metadata construction and full-text transcribing/ publishing work had been lasted through the First Phase of the NDAP (2002-2006) and the current Second Phase of the NDAP (2007-). As most digitization projects, for *Tan-hsin Archive*, the time needed for metadata construction is many times of that for image digitization (see Figure 1).

Since 2007, with immense amount of digital image and great portion of metadata available, the NTUL started to undertake an application project *E-Learning Portal for Tan-hsin Archive* based on the archive.

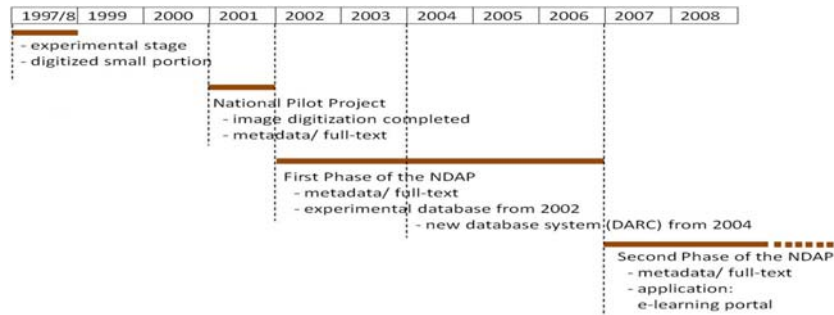


Figure 1. Time table for different tasks for *Tan-hsin Archive*.

Image digitization

Physical condition of *Tan-hsin Archive*

The documents contained in *Tan-hsin Archive* are varied in size and, in each ‘case’, all the varied-size documents had been glued one next to the other. The paper used was of different materials and colors, and some of them had definite form printed and, some paper was with special patterns. Many of the documents have turned brittle especially those had been paper-lined fixed. In addition to the materials deterioration, many documents had been damaged by tremendous worm-bites. Although the fumigation done in recent years has stopped the worm damage, the critical condition is irreversible. Therefore, during the digitization process, besides exercising necessary preservation care for some of the documents, the digitization-related tasks should be done at the consideration of minimizing the further damage to the documents.

Methods to minimize further damage to the *Tan-hsin Archive* while ensuring success

Planning for an overall workflow

The two major parts of work to be done for a digitization project of a documentary archive before going to the database programming stage are digitization of documents images and construction of metadata. Different archives may need different metadata schema. Besides, it is necessary to have an overall understanding about the physical condition of the archive in terms of the size and color of documents, scripts and font of characters/ letters (in terms of size and evenness of color and size) on the documents, organization of the documents, whether the coding system is enough or applicable for digitization, etc. Only when all such aspects are completely grasped could it be possible to carry out a sound planning for naming system, image digitization method and digital file formats, metadata schema and specification, and an executable workflow that enables all the tasks been undertaken in seamless coordination. In the digitization project of *Tan-hsin Archive*, in order to achieve this, the overall workflow for all related tasks had been planned at initial stage, with an inventory list as the essential foundation for all the related tasks (see Figure 2).

Careful inspection and establishment of an expandable preliminary inventory list

Under the new coding system that Professor Dai created, each document in an ‘case’ was labeled with a tag that includes the ‘case code’ (three-layered classification code + case number within the third layer of class) and the ‘item code’ (a case may contain series of documents, i.e., items. ‘Item code’ indicates the order the document appears in the whole ‘case’.) As the documents were very fragile, it is important that all the need-to-know information be acquired at ‘one shot’, thus documents could be free from frequent open up for various kinds of check, hence could minimize possible damage. Under such consideration, in the preliminary inventory list, the following aspects have to be included so the inventory list could be used for other purposes:

1. A complete list on which each ‘case’ occupies a single row.

The preliminary inventory list was established in ‘case’ level. Number of documents (items) to be indicated in the parentheses follows the ‘case code’.

2. Whether or not the ‘case’ includes a summary statement sheet (‘案由’).

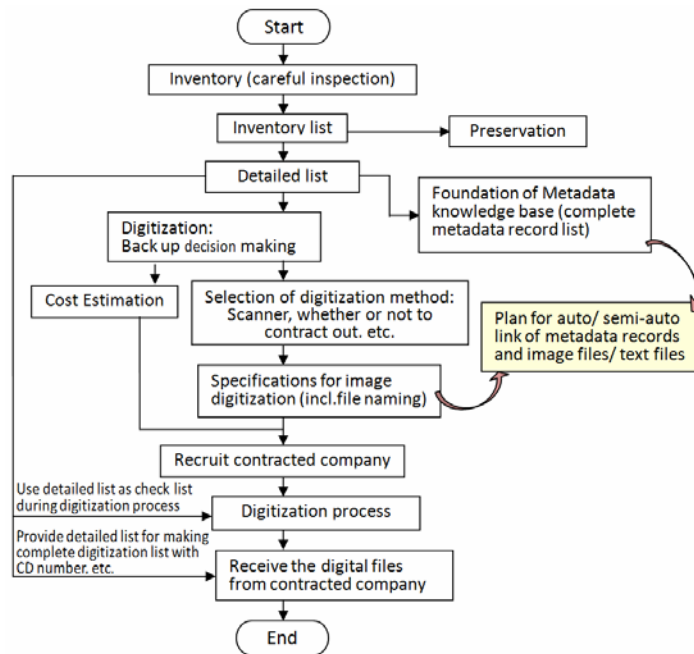


Figure 2. Workflow for all tasks for *Tan-hsin Archive* digitization project.

Most summary statement sheets are independent from series of documents that had been glued one document next to the other. Summary Statement is an important part for each 'case', therefore, deserves a single column for quick check up.

3. Indications about preservation needs.

In *Tan-hsin Archive's* case, three columns were made for (1) need to repair Summary Statement, (2) need to repair the whole case, (3) number of items need to be repaired, (4) on which item are there attached slips need to be repaired (need to indicate number of slips need to be repaired) .

Lined-fixing preservation care is very time consuming. And, those documents in critical condition were not possible to be digitized either because of too many wrinkles or because the worm-bites to strengthen the whole sheet of paper to an ideal condition for digitization. As the digitization work and preservation care had to be done simultaneously, it is necessary to make a priority list for preservation. With the different level of preservation need indicated in independent columns, it is possible to sort and get an easy-to-hard preservation care list. This was helpful in speeding up preservation care for most light-damaged documents so that in overall process image digitization could be smoothly done in order.

Designated the expanded detailed inventory list as foundation for all tasks

The preliminary inventory list was constructed with each 'case' in a row. However, the metadata would be constructed in item level, i.e., each document of a 'case' will be given a metadata record, and the full-text PDF file also need to be saved at item level. Therefore, it is important to create an item-level detailed inventory list before starting image digitization and individual full-text PDF files saving works. In *Tan-hsin Archive's* case, the detailed inventory list was expanded from preliminary inventory list. By the inventory information contained in the preliminary list and by using add-rows function provided by Excel, a precise detailed inventory list was made to include file names in each row for the following conditions: (see Figure 3)

1. a row for each document 'item',
2. a row for each attachment,
3. an additional row for documents with attached floating label,

The detailed list created could be utilized as a check list during digitization process, the foundation of metadata knowledge base (complete metadata record list), the foundation for the image digitization contracted

Item ID	Description
TH 33502_003_00_01	前立全理大中國書部 (送新檔案)
TH 33502_003_00_01R	前立全理大中國書部 (送新檔案)
TH 33502_004	前立全理大中國書部 (送新檔案)
TH 33502_005	前立全理大中國書部 (送新檔案)
TH 33502_006	前立全理大中國書部 (送新檔案)
TH 33502_006_00_01	前立全理大中國書部 (送新檔案)
TH 33502_006_00_01R	前立全理大中國書部 (送新檔案)
TH 33503 (7)	前立全理大中國書部 (送新檔案)
TH 33503_000	前立全理大中國書部 (送新檔案)
TH 33503_001	前立全理大中國書部 (送新檔案)

Figure 3. Example of detailed inventory list.

company to make a complete digitization list with CD number, etc., and it is also useful for quick random check of digital files when received from contracted company, need not to mention that it is always good for future reference. With such approach, it is possible to ensure correctness and consistency of file naming, and we rarely need to consult to the original documents for verification during the later works, thus could best protect the fragile documents.

Selection of digitization methods and digital files' formats

For image digitization of the *Tan-hsin Archive*, in considering that there is large amount of documents to be done during the definite period of time, and documents are in critical condition, it was decided that the Library need to recruit a contracted company that equipped with special scanner. The follows are the special decision for image digitization for this archive:

1. Using planetary scanner or well-designed stand accompanied with digital camera back instead of flatbed and feed-through scanners.
2. To digitize at higher resolution and convert to smaller files according to different purposes. Files for different purposes and their formats/ resolutions need to be carefully set at most reasonable size (see Table 1, Figure 4,5).

Table 1. Types of digital files saved for *Tan-hsin Archive*: purposes, formats and resolutions.

Collection	File format	resolution	"File size"	"Document size"
<i>Tan-hsin Archive</i> scan area: A2 (42 x 59.4 cm, i.e., 16.5 x 23.4 in)	Tiff -- for long-term preservation	300 dpi	55 MB	54 MB
	JPEG -- for e-commerce	300 dpi	12.4 MB	54 MB
	JPEG -- for Display and Net-print compression rate: 1:75	150 dpi	368 KB	was 13.5 MB became 6.77 MB after reduced width/height
	JPEG -- for Display only	75 dpi	151 KB	1.7 MB

→ see Figure 4,5

The 'Document size' of both 300 dpi Tiff file and JPEG file are the same regardless the conversion from Tiff file to JPEG file. 'Document size' is the cache space that would be taken each time a file is opened. It is necessary to keep 'Document size' as small as possible so that those workstations without large memory will not experience slow performance or even, shut-down trouble when the user needs to browse as many images. The effective way to reduce file size is to reduce the setting of 'Document size' (see Figure 4,5).

However, documents with small characters or uneven ink color may need to do several tests to find out best compression ratio and most suitable 'document size'.

The above example was based on *Tan-hsin Archive* example. It is expensive to use planetary scanner for digitizing images. If the materials to be digitized are not as delicate, then a less expensive method could be considered. When the NTUL conducted a digitization project for an old magazine collection with more than 400,000 pages to be digitized, higher-level digital camera was enough to carry out good-quality PDF files. The initial format

of file created is DNG Archive file (RAW). Such file could be flexibly converted into various kinds of formats and with various resolutions. In NTUL's experience, 24 bit/ color/ Jpeg files had been created and saved for there are color pages in the collection. And black and white/ Tiff files also had been created and saved. Both later created color/ Jpeg files and black and white/ Tiff files could be used to create PDF files to accommodate with the trends of magazine/ periodical. The digital camera solution could reduce cost to 1/4 of the cost using planetary scanner.

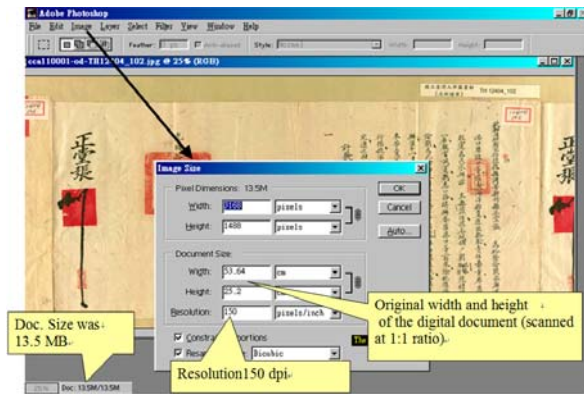


Figure 4. Before changing 'Document size'.

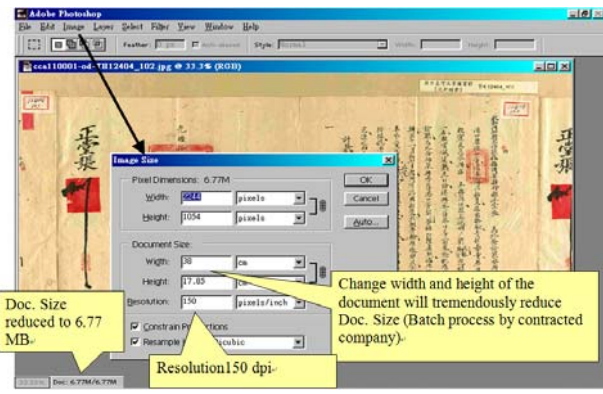


Figure 5. After changing 'Document size'.

Creation and use of text files

In 1995, a typed full-text compilation and publication project for *Tan-hsin Archive* was initiated for the considerations that original documents are hard to read/ comprehend on account of many old characters and errata contained in the archive, and some documents were written in running hand-writing, besides, the archive contains documents of varied formats and it is not easy to recognize some definite information in the documents. It is believed that printed version can promote use of the important archive and enable easy access to the contents.

The printed typed-version of this archive later was utilized in enhancing the *Tan-hsin Archive* database-- both image file and typed full-text are provided for users. The typed full-text are for easy and quick reading, and the image files are for researcher to use in verification, etc. Since 2007, such typed full-text had been further used in the e-learning portal for the archive. It is also possible that full-text files would be used to support full-text search. It was estimated that there would be 36 volumes in total when all of the documents in the archive have been typed and published. The NTUL has already published 24 volumes. The successive 4 volumes are to be published in recent months.

Metadata construction

Metadata can be exercised to fulfill the following functionalities: (1) For conservation inventory purpose, i.e., as the identification of important culture/ history relics (via physical description), (2) For grouping functionality (enabling easy link of all related materials of various kind of media via well-designed file name formula and some fields that specify the concept of 'collection' from different scopes, etc.), (3) For enabling search and search result display (via content-related metadata fields and ontological commitment during input process), (4) For enabling browsing search (via structured topic/ subject tree and indexes). For *Tan-hsin Archive*, the practices for these functionalities are as follows:

General considerations for selection of software and inputting approach

When tackle with a static archive such as *Tan-hsin Archive*, considerations for some decision could be different with tackling with a dynamic case. For establishing a metadata knowledge base for *Tan-hsin Archive*, the following factors had been taken into consideration for decision making: The archive is static hence it is not a must to have a sophisticated knowledge management system. Besides, by the time the Library had to have metadata inputting quickly move on, the data model was not yet concluded to be a final version. Therefore, the Microsoft Office Excel was selected for the very fundamental knowledge base construction work. There are some good points about Excel for quick construction of a knowledge base. It is easy to construct a set of template for individual collection, and the software is easy to learn. Besides, the Excel working sheet is easy for making quick check, it allows 'drag down' for successive numbers as well as 'fill down' same properties for those metadata records that share the same properties so that it is possible to minimize typographic error. Other important features include that it

allows doing changes to the metadata schema ⁽¹⁾ and the knowledge base build up with it is independent from system. Having knowledge base be independent from database software is important for flexible utilization, data migration, security reason, etc. It is also compatible with metadata format conversion programs thus enable easy conversion to XML format.

By using Excel for metadata construction, the Library not only had saved the cost for buying a software package but also saved time for long discussion for a definite specification. For a documentary archive with complicated content, an ideal metadata schema may not be easy to achieve. It is time consuming and sometimes not fruitful holding several discussions to acquire consensus from all content experts, information service provider, and programmer. Therefore, with the premise that there is conventional data model or good similar examples for reference, and the following up programming is possible to be done by information staff or the institution allows contracting out the work, then it could not be a risk to consider using Excel.

As for the metadata inputting work, because *Tan-hsin Archive* is such a special collection, its metadata construction has higher knowledge background prerequisite. The NTUL therefore decided to have history major students work for it. The faculty and students of the NTU History Department is genuinely an asset to such project that the NTUL had long been maintaining good collaboration relationships with them.

**To empower functionality of metadata (1):
Mechanism to assure automatic/semi-automatic link of all related images: file naming**

Metadata could be exercised to having functionality beyond enabling searching and browsing. A very important functionality but rarely mentioned is the utilization of the metadata identification number's naming system. A naming system that is shared among different digitization tasks will make possible automatic/semi-automatic link of all related images. It is well agreed that the success of *Tan-hsin Archive*'s digitization and the associated database construction owe a great deal to a sophisticated naming system.

File name to cover all possible situations

When metadata records are ready for establishing a database, all the related sources including image files, full-text files, etc. are to be incorporated into the database system. As stated, the crucial strategies to ensure automatic or semi-automatic link of all related resources is to have all parts of works adopt the 'detailed inventory list' as the essential foundation. The file name in the list therefore should be sophisticated enough to cover all possible conditions. In other words, in *Tan-hsin Archive*'s case, sequential documents are glued one next to the other, and some documents may contain attachment document(s) or attached floating slip(s) (see Figure 6). We have to be conscientiously digitize all the attachments and, for the documents with attached floating slips, digitize twice having one image with the slip cover the underneath words and having one image with the slip flip open for displaying the underneath information. The naming system should abandon sequential accession numbers because that numbering system does not have the mechanism to 'group' related images together for automatic link purpose. For the extra large or extra long documents contains in the *Tan-hsin Archive*, there is not any scanner could capture the complete image. Therefore, the ideal alternative is to strive for a digital file naming system that allows scanning in parts while maintaining good coherence coordination among different parts. To fulfill all these requirements, the solution is to create a 'file naming formula' that covers all possible conditions (see Table 2) with a '9-square matrix' (see Figure 7) as supplement to ensure a logical and consistent naming for part by part (segment by segment) digitization (i.e., designate a number for definite area of a document).

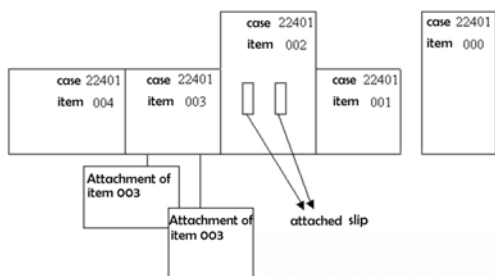


Figure 6. Various conditions of *Tan-hsin Archive*.

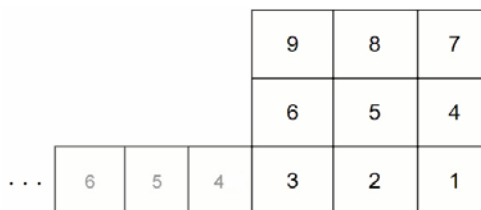


Figure 7. 9-square matrix for extra large documents.

File name formula and principles involved

Table 2. File name formula for *Tan-hsin Archive*

Segment of file name	1st segment (coll. code + case no.)	2nd segment (item no.)	3rd segment (attachment)	4th segment (attached slip)	5th segment (segment no.)	6th segment A/ B env.
explanation	Fundamental segment			_01 : with slip cover the text _01R : with slip flip open	Segment code is according to the 9-square matrix	
Specification for each segment (in terms of number of digits)	T H 2 2 4 0 1	_ 0 0 2	_ 0 1	_ 0 1	_ 1	_ A
	T H 2 2 4 0 1	_ 0 0 2	_ 0 1	_ 0 1 R	_ 9	_ B
Example 1	T H 2 2 4 0 1	_ 0 0 2	_ 0 0	_ 0 1		
Example 2	T H 2 2 4 0 1	_ 0 0 2	_ 0 0	_ 0 1 R		

The file name formula for *Tan-hsin Archive* is designed to have as many as 6 segments for covering all possible conditions. Underscore ('_') is used to separate different segments. (see Table 2) Segments of file names are varied. Some file name could contain only two segments. The important principle for assigning a file name is that when assign a file name for an image of certain condition, for example, for an image with the attached slip cover the underneath words, then the file name should be as shown in Example 1 (The image is for the main text but not for the attachment document, then, the 3rd segment space that is reserved for 'attachment' condition should be given '00' so that the 'attached slip' information could be consistently appeared at the 4th segment.)

A metadata record could be shared by all the related images (image of attachment document(s), image of the document with attached slip cover the text, image of the document with attached slip flip open, different area/ segments of a large or lengthy document, etc.). Since digital images of all these conditions share the same 'fundamental segment' ⁽²⁾, it is possible that through programming all the related images could be automatically group together with the metadata they belong to. Even if there are some special cases need human involvement, the human involvement for link of all the related images or different types of sources could be very much minimized.

The file name formula for *Tan-hsin Archive* was established following the principals as: (1) To retain the original organization logic/coding system of the archive. In *Tan-hsin Archive*'s case, many users have long been relied on the original coding system. Therefore, it is especially important to retain the original coding system and to allow it be searchable for users. (2) To reserve a definite 'position' for each specific condition, i.e., attachments, attached floating slips, parts of a document (reserve this space for the situation when need to scan document part by part), etc. It is also necessary to strictly follow the rules as how many digits should be input for definite segment. This will make it easier to make quick review and error check. (3) Segments of file name to be ordered from most frequently used segment to least used segment. (4) Number of digits of each segment should be enough to cover all the documents, e.g., if the total page/ item number is over a few thousands, then need to consider about if 4 or 5 digits is more adequate. If the collection is a growing collection, then 5 digits may be a better decision. (5) File names should include collection code. In *Tan-hsin Archive*'s case, the collection code 'th' is based on the first letter of the words of the collection name. This code and institution's code can be added to file names through batch process. (6) File naming system should be established in the very early stage so that different parts of work (metadata, image, text, audio-video materials, etc.) could base on the same naming system for the generated digital files. Because the file name is quite long that in order to avoid human error, it is suggested that the formula to be posted at easy-to-see place for anytime consultation.

To empower functionality of metadata (2): Maintaining precise metadata elements

International/ national standards, local inner-institution standards, shared standards

At initiation stage, the NTUL had adopted one of the agreed standards-- Dublin Core Elements ⁽³⁾, and broke down most core elements to many more precise elements so as to well represent the knowledge domain. There are about eighty fields in metadata schema for *Tan-hsin Archive*. In practice, some fields are often used and some are not. Besides, the DTD provided by the institution in charge of union catalog or gateway portal is much simpler. It may be quested to omit some rare-used fields from the schema. In fact, it is better to maintain the set of detailed precise elements for the metadata schema. This is good not only for the management of the collection but also for a more sophisticated database system and future mapping with other similar databases. Therefore, it is recommended to maintain inner-institution standards at more precise level and export data according to simpler shared DTD for inter-institution project(s).

Establishing item/ collection level metadata according to the property of objects

As stated above, the NTUL not only has established a searchable database for the *Tan-hsin Archive* but also has tried to make application use of it-- establishing an *e-Learning Portal for Tan-hsin Archive*. For this e-learning portal, in addition to a searchable *Tan-hsin*-related research materials database, four groups of learning objects are to be created: (1) E-learning materials regarding the categories of documents of *Tan-hsin Archive* (To explain about the key features of specific category of documents). (2) E-learning materials regarding the document format of *Tan-hsin Archive* (To explain about the equivalent format in the printed full-text file, i.e., the contemporary format). (3) Theme stories from the *Tan-hsin Archive*. (4) Self-evaluation tests. Item level metadata and collection level metadata are respectively established for material of different property. For theme stories, a complete story is based on all the document items belong to a same 'case', therefore, collection level approach is adopted. For other three groups of materials, in consideration of the needs to provide detailed information about each individual object, item level approach is adopted.

Making teaching resources for e-learning project requires highly civilized knowledge about the domain and good writing capability. It takes time to accumulate one by one from zero. With very few learning objects been created in the current early stage, it does not seem to be necessary to establish metadata for each e-learning object. However, a plan for future is import, besides, with the metadata schema for the e-learning objects available from early stage could help the design of an integrated system for the knowledge domain and all of the different groups of e-learning objects.

To establish metadata knowledge base for e-learning objects, making the database an independent one or having the e-learning objects share the same base with the domain database are both fine. The key-point to ensure clear differentiation of different groups of materials or to ensure successful relation establishment among related materials is that we have to be conscientiously input information in the associated fields as: (1) If the learning object is of item level or collection level, (2) The code of the 'collection' (i.e., code of the 'case'), (3) The object group an object belongs to. Such information in definite fields is useful for the purpose of bridging the related materials in an integrated system in which auto-link approach is preferable.

To empower functionality of metadata (3): Be conscientious about subject and keywords

For historical documents like *Tan-hsin Archive*, knowledge base of subject and keywords might be the hardest to build up. But these fields can enable browse search and index search and are more useful for users especially when they have no idea about what to start with. The problem for historical archive collections holders is that such archive collection may lack of a ready-to-use ideal subject classification system. For *Tan-hsin Archive*, there is a subject classification system created by Professor Yen-hui Dai, a professor of the Department of Law in NTUL. However, when Professor Dai conducted the first large scale preservation and cataloging project in 1950s, he abandoned the original organization of the archive⁽⁴⁾ and re-classified it according to the contemporary concept of administration and law-- reclassified all the documents into three major sections: administrative, civil, and criminal. Under each section, there are two layers of more precise classes. The classification system and the associated coding system, though is not as ideal as the one that maintains the original organization logic, has long been the most well consulted access 'gateway' for all the users, therefore is adopted for the *Tan-hsin Archive* database. Maintaining this well-used current convention is helpful for users to switch to the electronic environment. It is also of great help to new users. Therefore, systematic subject or keywords information is what a practitioner should think about and strive for.

To fully exercise the functionality of metadata

Learning about the functionality of each metadata element is helpful for the right and good application of metadata fields. All of the broke-down metadata elements from Dublin Core Elements could be re-categorized into different functional groups. During the re-categorization, it is possible that we break the boundary of some core elements and put the elements that belong to a same Dublin Core Element to different functional groups. For example, the core element 'description' has been broken down to include elements that describe the physical conditions/ features of the object and elements that describe about the content (e.g., in *Tan-his Archive*'s case, one of the 'description' field is to record all the related persons involved in the event described in the document, and the job title or roles of those persons in the event.) These two kinds of descriptions are to be categorized into functional group of 'inventory' and 'support of search' respectively. The knowledge base model constructor is clearer about precise content of each field. Therefore, it is suggested that the knowledge base model constructor classifies all metadata elements according to their different functionalities as to be used for searchable field, grouping purpose, sorting and filtering purpose, enabling browsing search, as index, etc. Allocating each metadata element into the

functional group it belongs to in a tidy way is very helpful for the discussion and communication with programmer. Therefore, even though each element of the Dublin Core Elements seems to have its own 'boundary', it is not necessary to rigidly adhere to their boundary when we make working sheets materials for discussion.

Conclusion

The *Tan-hsin Archive* is very complicated both in its physical conditions and documents' contents. The lesson we have learned from its digitization include: (1) Strategic planning should be well done at the beginning stage. (2) The specifications should be seriously followed during the working process so that the outcomes could enable knowledge reuse and share, various applications, easy integration of different databases, besides, to minimize the necessary revision. (3) Creation of a metadata schema is expensive and time consuming. It is suggested that if there is any good practice or community accepted standard, than it is better to try to adopt a well-developed specification instead of starting from scratch. (4) The quality of metadata is much influenced by the available amount of funding and manpower. The fields that require intellectual input such as subject and abstract fields are comparatively hard to input but these fields are more helpful for users. In summary, it is important that we try to carry out all parts of works at one right shot and make good application utilizations so that we can bring the investment to the utmost fruitful results.

Endnotes

- (1) When doing changes to metadata schema, it is suggested to add new tag in the back new column but not to insert new column(s) among the in-use columns. Field tag change is allowed as long as the content under the tag and related specification is unchanged. Such tag change will not require an overall reengineering of the database program but only need some minor revision of tag labels. When add new metadata element(s) to the knowledge, if we add new tag in the back new column, then the programmer can just make use of it to create new database functionality or to enhance the existed functions but need not to do tremendous changes to the existed programs.
- (2) In *Tan-hsin archive's* case, 'fundamental segment' is used as the metadata identification number. If an attachment is considered as a document that needs a metadata record of its own, then the metadata identification for it will include 3 segments. The third segment is to indicate what number it is as the subordinate attachment to the main document.
- (3) (1) Contributor, (2) Coverage, (3) Creator, (4) Date, (5) Description, (6) Format, (7) Identifier, (8) Language, (9) Publisher, (10) Relation, (11) Rights, (12) Source, (13) Subject, (14) Title, (15) type. (<http://dublincore.org/documents/dces/>, accessed October 19, 2008)
- (4) The original documents were classified into five categories: 'Li' (Administration), 'Hu'(census register), 'Lee'(ceremonial rules), 'Ping'(military), 'Hsieng'(criminal), 'Kung'(construction and fortifications).

References

- Hung, S. F. (2004). 文獻典藏數位化的實務與技術 = *Tasks and techniques for digitization of documents*. Taipei: 數位典藏訓練推廣分項計畫 (Human Resource Development & E-learning Division, National Digital Archives Program). (Available at http://dlm.ntu.edu.tw/01_1_2.htm)
- Hsiang, J., Chen, H. H., Wu, H. J., & Hung, S. F. (2004). 各國檔案數位化之探討 = A survey of digitization methodologies of archives. *檔案季刊(Archives Quarterly)*, 3(3), 1-20.
- Hung, S. F., & Chiu, W. J. (2004). 國立臺灣大學圖書館數位典藏 metadata 之設計與資料庫之建置彙整 = Metadata schema of National Taiwan University Library digitization projects: specifications and integration of different databases. In *技術研發分項計畫93年數位典藏技術規範會議——後設資料在數位典藏之研究發展：回顧與前瞻研討會論文集* (pp. 237-269). Taipei: 數位典藏國家型科技計畫技術分項計畫.
- Hung, S. F. (2007). 紙質文獻類的雜誌書籍之數位化 = Digitization of magazines and monographs. *佛教圖書館館刊(Buddhist Library Bulletin)*, 45, 19-25.

Related Websites

1. Database for the *Tan-hsin Archive*: <http://www.darc.ntu.edu.tw>
2. *Tan-hsin Archive E-Learning Portal*: <http://140.112.114.10/tanhsin/>

Acknowledgments

The author, on behalf of the NTUL, would like to express gratitude to Dr. Hsiang who has been advising his team in the Research Center for Digital Humanities to collaborate with the library staff to construct the database system for this and some other archive collections for the NTUL. Our gratitude also extends to the National Science Council, Taiwan, for the consistent monetary support for the projects for *Tan-hsin Archive*.