



Chinese Classics Full-Text Database Digitization Procedures Guideline


Ya-Ping Wang
Hsiao-Lin Hsieh

Taiwan e-Learning and Digital Archives Program
Taiwan Digital Archives Expansion Project



Acknowledgements

We are grateful to the members of the Archives and Databases Thematic Group of the “Taiwan e-Learning and Digital Archives Program Taiwan Digital Archives Expansion Project” that provided their results and experiences with digitization work over the years, including the “Buddhist Lexicon Digital Resource Establishment and Research – Buddhist Reference Books and Integrated Services in a Digital Era” of Dharma Drum Buddhist College, “Research and Building of the Chinese Buddhist Tripitaka Electronic Text Collection, Taipei Edition,” “TEI Localization” Project, Chinese Classics Full-Text Database Team of the Institute of History and Philology, Academia Sinica and the Chinese Buddhist Electronic Text Association. This guideline would not have been possible without their great efforts. We hope that this guideline will provide aid to the many project units and personnel also engaged in digitization work. We would like to specially thank Vice President Tu Cheng-Ming of the Dharma Drum Buddhist College for reviewing this digitization procedures guideline, and for offering precious advice on its contents and direction.



Publisher's Preface

After the “National Digital Archives Program” was implemented in 2002, members of numerous institutional projects and request-for-proposals projects joined our team to engage in digital work that covered countless categories and massive amounts of content. The first phase of the five year project was successfully completed in 2006. The following year, the “National Digital Archives Program” and “National Science and Technology Program for e-Learning” were integrated into the “Taiwan e-Learning and Digital Archives Program (TELDAP, <http://teldap.tw/>)”, striving to achieve the ultimate goal of “presenting Taiwan’s cultural and natural diversity” as it continued to expand digital resources in various fields, and systemically promoted digital achievements in education, research and industries. TELDAP is preparing to actively collaborate with the private sector to drive growth in related industries, not only preserving important cultural assets, but also accelerating the development of a new culture in the digital age of today.

Originally named the “Content Development Division” during the first phase, we were renamed “Taiwan Digital Archives Expansion Project” (<http://content.teldap.tw>) as a subproject of TELDAP, and took more active measures to expand the sources of digital content, extending our reach to the collections of private institutions and even individuals. We have widely requested proposals for digitization projects related to archives, archeology, philology, geography, ethnicity, art, daily life, animals and plants, and hope to better integrate digital content with different characteristics, to develop them into fun and inspiring digital materials, and to provide them free of charge to the public for education and research; this will also help firms and public or private holding institutions to find cooperation opportunities in value-added applications. Collaboration between the “Taiwan Digital Archives Expansion Project” and other projects under the “Taiwan e-Learning and Digital Archives Program” will help speed up development of educational, research and commercial value-added applications of digital content, which will benefit the presentation of Taiwan’s cultural and natural diversity, and allow people everywhere around to understand and appreciate the richness of our history and culture, as well as the beauty of our natural ecology.

While collecting and developing value-added applications of digital content, whether it may be during the “Content Development Division” or “Taiwan Digital Archives Expansion Project” period, members of this project have continuously followed up on digital workflow related technologies used by public and private institutions and open request-for-proposals projects, and compiled a series of “Digitization Procedures Guideline Books” that introduce various international standards and provide information on digitization technologies and workflows. Since 2005, we have written 20 digitization procedures guidelines on different themes (the full text of all of the 20 books can be downloaded from the “Taiwan Digital Archives Expansion Project” website under “Digitization Books”), selecting exquisite digital objects, such as ceramics, paintings, calligraphy, and string-bound books, combining the experiences of different institutional projects, and supporting them with domestic and foreign theories and practice results.

Since 2008, we have continuously revised and expanded our “Digitization Procedures Guideline” book series, hoping to expand distribution channels so that




they may be provided to even more museums, libraries, institutions and individuals for reference. Our preparations are mainly divided into revising existing guidelines for “selected objects” and compiling new guidelines on “common principles”. The former refers to revising the existing 20 guidelines with a focus on introducing new digitization technologies and specifications, more practical software and hardware, and digital content protection mechanisms; we expect to complete the publication of all 20 books by the end of this year. As for compiling guidelines on “common principles,” our emphasis will be on the introduction of key concepts, such as the “life cycle” of digital information and quality control, studying multiple types of objects instead of a single type of object, and adopting common principles as the guideline framework. The so called common principles refer to project planning, integrated workflow, audiovisual data, text data, color management, outsourcing management, and digital content protection and authorization. We will investigate, study, and write guidelines for these eight common principles, and expect to complete the publication of all eight guidelines by 2012.

While planning and writing guidelines for selected objects and guidelines on common principles, we set a complementing relationship for them. Guidelines on common principles emphasize on the analysis of important topics in digitization work, guiding readers to thoroughly consider the advantages and disadvantages of digitization. Guidelines on selected objects describe practices and techniques for digitizing specific objects, helping readers to select the most suitable, most effective digitization workflow. By publishing this “Digitization Procedures Guideline” book series, we believe that we are providing institutions and individuals with the intention to engage in digitization work with a series of practical guidelines that provide an overall view, while guiding them step by step through the digital workflow. Here we must stress that the theoretical foundation of this book series is the precious experiences of institutional and request-for-proposal project teams accumulated throughout the years. These experiences allow higher quality digital content to be produced, presented and maintained with less cost, further enriching our digital archives and e-learning content. As we continue to publish our “Digitization Procedures Guideline” book series, we must give special thanks to working partners who were interviewed and colleagues who were a part of writing the guidelines, and are grateful to the scholars and specialists that reviewed and provided their advice on the book series. Finally, we hope that readers will not be reluctant to correct any mistakes or make recommendations that will help us be even better.

Taiwan e-Learning and Digital Archives Program
Taiwan Digital Archives Expansion Project
Digital Archives Sub-project of Project Integration

Project Director
March 29th, 2011



Contents

2	Acknowledgements	
3	Publisher's Preface	
7	One. Introduction	
11	Two. Digitization Flowchart	
14	Three. Preliminary Procedures	
	I. Select Materials.....	15
	II. Establish Standards.....	16
28	Four. Object Digitization Procedures	
	I. Scanning.....	29
	II. Input.....	31
	III. Proofreading.....	36
44	Five. Metadata Establishment	
	I. Metadata Exclusive for Corpora – TEI.....	45
	II. TEI Core Elements.....	45
	III. TEI Best Practices.....	55
60	Six. Database and Other Applications	
	I. Database Establishment.....	61
	II. Produce Optical Disks.....	61
	III. Develop Related Encyclopedias and Dictionaries.....	62

63 Seven. Digital Rights Management

I. Creative Commons Licensing.....64
II. Digital Rights Managements (DRM).....67
III. Access Rights Restrictions.....67

69 Eight. Equipment and Cost Analysis

I. Equipment Selection Considerations.....70
II. Cost Analysis.....72

74 Nine. Outsourcing

I. Outsourcing Copying and Scanning.....75
II. Outsourcing Key-in and Initial Markup.....77
III. Outsourcing First and Second Proofreading.....79

81 Ten. Conclusions

83 References



One. Introduction

When flames reach 451 degrees Fahrenheit, all ancient books and records will be turned to ashes, but our suppressed souls will emerge shining.....this is a world without fire, and the work of firemen is to set fire. This is a world where all books are banned, and the duty of firemen is to “burn books.”

This paragraph appeared in the science fiction novel “Fahrenheit 451” published by Ray Bradbury in 1953. In the Western world of the future, physical books will share the tragic fate of books during the Qin Dynasty in ancient China. Book lovers volunteer to become carriers of knowledge to preserve the heritage of humanity and memorize books in their heads. If someone wanted to learn about the Bible, they visited the “Bible” person, and if they wanted to know about Shakespeare, then they would visit the “Shakespeare” person.

The concept of “everyone is a book” in the preservation and dissemination of knowledge was proposed by Ray Bradbury in 1953, when the internet did not exist. The idea was a solution to books not being able to exist in physical form, but had to be passed on in a different form. However, if this book were to be written in the twenty first century, people as carriers of knowledge might be replaced by electronic books, which allow knowledge to be more objectively, scientifically and safely preserved through computers and the internet, where it can rapidly and effectively circulate.

Digitization of books and records is the transformation from physical form (e.g. paper copies) into digital form. At present, there are three relatively common approaches to digitization. The first is to produce black and white or color digital images of books according to their original appearance, number of pages and size; this method is also known as “full-text image digitization” and has the advantage of showing the original text, texture and color of books. The second is “full-text digitization,” which inputs and proofreads text on books and outputs electronic text files; this method allows annotations or ambiguous parts of text to be identified, ancient characters no longer in use to be replaced with current characters, and provides full-text retrieval, thus increasing the research value of ancient books and records. Of the two digitization approaches above, the former emphasizes the reproduction of the form of books and allowing readers to see images of the books appearance; the later emphasizes the presentation and index of book “contents,” as well as allowing users to read

the text of books and understand its meaning. There is still a third approach that combines considerations of the first two approaches, and stores both appearance and contents in digital form in a database; the two are displayed side by side or a hyperlink is used to connect the two, allowing appearance and content to complement one another.

Digital data is characterized by AAA (Anyone Anywhere Anytime) and offers four benefits: 1. Capable of being stored in portable and convenient optical disks, hard disks, or disk arrays, saving space and benefiting preservation; 2. Capable of displaying new data types, such as hypertext and multi-media, giving users a refreshing experience; 3. Inspires new research directions. Using Buddhist scripture full-text digitization as an example, word frequency statistics allow researchers to understand word use in Buddhist scriptures, and enable the generalization of Buddhist decrees and regulations; 4. Easy to copy, disseminate and propagate, driving knowledge circulation.¹

A full-text database is a database that uses all text of books or records as materials and preserves the layout of each page.² The earliest electronic full-text database in Taiwan is “Scripta Sinica,” a database developed by Academia Sinica in 1981, and collects electronic full-text of various ancient books and records, including the twenty five historiographies, thirteen classics, Veritable Records of the Qing Dynasty, novels and theater, etc., making it the most complete Chinese full-text database in Taiwan. National art and culture resources collected in the “National Repository of Culture Heritage” of the Council for Cultural Affairs include full-text and images of ancient, modern and local literatures. With regards to religious books, there is the “Zhengtong Daoist Canon” full-text digitization work led by Research Fellow Li Feng-mao of the Institute of Chinese Literature and Philosophy, Academia Sinica; Buddhist Tripitaka catalogues include the Digital Database of Buddhist Tripitaka Catalogues of the Chinese Buddhist Electronic Text Association (CBETA), the Buddhist Tripitaka Catalogues Integrated Search System of the Luminary Buddhist Institute, and the “Taishō” and “Āgama” full-text databases

¹“Before and After – A Technique of Data Digitization” by Huang Hung-Chu, “Information Management for Buddhist Libraries,” Issue 15, September 1998.

²“A Survey of Full-Text Databases and Related Techniques for Chinese Ancient Documents in Academia Sinica” by Hsieh Ching-Chun and Lin Shih, Academia Sinica Institute for Information Science Chinese Documents Processing Lab, March 1997. Search: September 2010, website: <http://dbo.sinica.edu.tw/~tdbproj/handy/thesis.html>.

established by Foguang Shan Monastery and CBETA. The “Buddhist Electronic Text Database” of CBETA can be downloaded for free, and contains the highest quality and quantity of Chinese Buddhist electronic text.

Since the “Taiwan e-Learning and Digital Archives Program (TELDAP)” was implemented by the National Science Council in 2002, it has been devoted to the digitization of Taiwan’s cultural heritage and assets. The “Chinese Classics Full-Text Database Sub-group” was organized in July 2005 to integrate Chinese full-text digitization in Taiwan; Mr. Tu Cheng-Min serves as the convener of the Sub-group and has actively promoted research and technology sharing between full-text digitization projects in Taiwan. The sub-group has spared no effort in the promotion of full-text markup, because information suitable for the digital age must have high quality, high application value and can be shared with the world. For this reason, this guideline focuses on “full-text markup.”

This “Chinese Classics Full-Text Database Digitization Procedures Guideline” gathers together experiences of Chinese classics full-text digitization projects that have been implemented, and generalizes domestic and foreign technologies and standards into a set of full-text digitization procedures. This guideline hopes to serve as reference to supervisors and operators of project units, and allow future digital archive projects to more efficiently implement digitization work.



Two. Digitization Flowchart

Digitization procedures for full text can be divided into four parts. The first part is preliminary procedures and includes planning, evaluation and preparations before digitization work; the second part is text digitization procedures, encompassing document scanning, printing, keying in text and proofreading; the third part is markup, which records typesetting and content information of documents, creating academic research and application value; the fourth part is converting source books into electronic full-text for future application and development. See Fig.2-1 for detailed procedures. This guideline introduces digitization procedures in the order of the flowchart below.

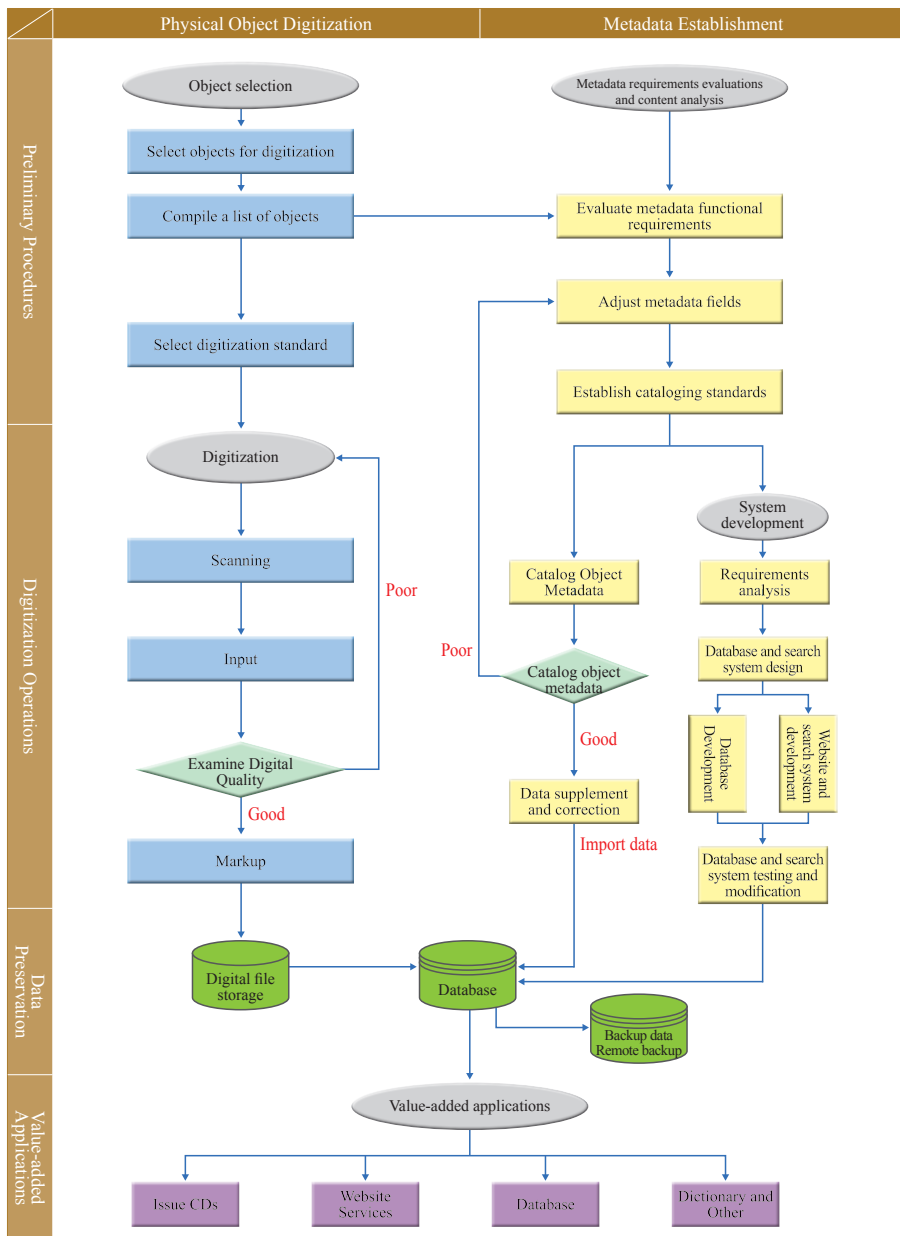


Fig.2-1 Chinese Full-Text Digitization Flowchart
 Compiled by the Taiwan Digital Archives Expansion Project



Three. Preliminary Procedures

Digitization is a costly and time-consuming task regardless of the quantity of digitization materials or project scale. Therefore, planning, evaluation and preparations before digitization operations are not to be overlooked.

There are two preparations to make before actual digitization can start—“select materials” and “establish standards.” The former requires understanding of existing materials and planning of overall objectives; the later benefits implementation results and quality management.

I. Select Materials

(I) Digitization Material Selection

Selection standards for digitization materials may vary based on different digitization purposes. Moreover, due to the varying and often limited funding of projects, materials should be arranged by priority to maximize results of labor and funds. Priorities may be based on rareness, importance and cost effectiveness,³ which can be divided into the following six items:

1. Rating of the physical object, e.g. standards of the Ministry of Education, such as national treasure or important relic, and classified document.
2. Preciousness of the physical object, e.g. uniqueness, rareness, epochal value and irreplaceable.
3. Preservation difficulty of the physical object, including fragility, cannot be copied, or might disappear.
4. Cost effectiveness after digitization.
5. Research, education and economic application value after digitization.
6. Other.

In addition, version is a matter of great concern and should be considered in full-text digitization. When knowledge entered the written era but before printing was invented, books were manually transcribed and often had mistakes or missing parts. Even after typesetting was invented, contents were misinterpreted or the wrong characters were used. Traditional China attaches great importance to handwritten literature, not only emperors throughout

³“Paintings and Calligraphy Digitization Procedures Guideline” by Kao Lang-Hsuan and Chen Hsiu-Hua, Taipei City: Taiwan Digital Archives Expansion Project, April 2009, pages 19-20.

the ages, but also scholars emphasized the investigation and preservation of literature. There is a great deal of government and private published history books, ranging from the three historiographies of the Three Kingdoms, to the thirteen historiographies and further on to the twenty-four historiographies designated by imperial order of Emperor Qianlong, showing that differences in version and records are important clues to analytical research by scholars. Furthermore, digital reproductions involve copyright issues, so digitization can only be carried out if copyright is successfully licensed; if not, switch to another version for which licensing can be acquired, otherwise project units can only reselect materials for digitization.

(II) Catalog Digitization Materials

After selecting books for digitization, a detailed booklist should be compiled for input. Since the digitization object is books, the smallest unit should be used for cataloging, meaning that one book should have one entry and contain the following information:

1. Title
2. Author
3. Publication Place
4. Publisher
5. Publication Date
6. Version

Besides the information above, the quantity of books should also be recorded. This booklist (as shown in Table 2-1) should be properly preserved for use in digitization operations.

II. Establish Standards

Standards and file formats should be established to ensure the successful link between each operation and maintain stable quality of digitization results.

Most projects or units refer to standards already established by related projects or units, while others rely on their own experiences. However, the purpose of standards establishment is to reach the ultimate objective of each individual project, not for copying the methods of others. Although operating standards may vary according to project objective, there are still some common principles. In the following section we will introduce standards and file formats associated with digitization work.

Table 2-1 Chinese Classics Full-Text Database Booklist of the Institute of History and Philology, Academia Sinica

	書 名	版 本
1	七才子詩選	
2	九卿議定物料價值 四卷	(清)工部編,清乾隆元年(1736)刊本
3	二十五史外人物總傳 要籍集成	董治安主編;濟南:齊魯書社,2000
4	入告初編 一卷, 二編 一卷,三編 一卷	(清)張惟赤撰,清順治(1644-1661)末刊本
5	入幕須知 五種	(清)張廷驥輯,清光緒十八年(1892)浙江書局刊本
6	八旗人口冊	不著編人,清光緒間(1875-1908)排印本
7	十科策略 十卷	(清)劉文安著
8	三流道里表	(清)唐紹祖等纂修,清乾隆年間武英殿刊本
9	三流道里表	不著編人,清同治十一年(1872)江蘇書局重刊本
10	三流道里表 不分卷	(清)查克順等纂,清乾隆四十九年(1784)刊本
11	三省曠防考 二卷	(明)劉應元撰,明隆慶元年(1567)刊本
12	三朝聖諭錄 三卷	(明)楊士奇輯錄,明鈔本; 漢籍資料庫有建置《國朝典故》
13	三賢政書 三種	(清)吳元炳輯,清光緒五年(1879)序刊本
14	上諭內閣 一百五十九卷	(清)允祿等輯,清刊本
15	上諭合律鄉約全書 一卷	(清)聖祖諭,(清)陳秉直解, 清康熙間(1622-1722)刊本
16	于山奏牘 八卷	(清)于成龍著,清康熙年間(1662-1722)刊本
17	于清端公政書 八卷, 外集一卷,續集一卷	(清)于成龍撰,(清)蔡方炳編次, 清乾隆二十六年(1761)刊本
18	于肅愍公奏議 十卷	(明)于謙撰,明嘉靖二十年(1541)杭州重刊本
19	于肅愍公集	
20	大明九卿事例案例 不分卷	不著編人,明鈔本
21	大明令 不分卷	(明)太祖撰,鈔本
22	大明律 三十卷	胡瓊集解,胡效才增附,日本蓬左文庫藏明嘉靖刊本 (本院圖書館查無此書)

(I) Digital Image File Format

Copies used for text key-in include digital images obtained from scanning books and copies of the original book.

Image files obtained from scanning books can not only be used for key-in and proofreading, but also be linked to electronic full-text to establish a full-text and image database. The Taiwan e-Learning and Digital Archives Program (TELDAP) divides digital file formats into three categories – browsing, commercial and preservation:

1. Preservation: Images in this category are for permanent preservation. They are undistorted and show the true appearance of objects in detail.
2. Commercial: Images in this category are provided for value-added applications, such as publishing, printing, copying, exchange or sales; image quality must meet printing requirements.
3. Browsing: Images in this category are for display online, so image quality must meet computer monitor browsing and internet transfer requirements.

Preservation images have a resolution of 300dpi or higher and are in uncompressed RAW or TIFF format, which are suitable for different software and platforms. Commercial images have the same resolution of 300dpi and are in uncompressed TIFF format. Browsing images have a resolution of 72dpi for the convenience of circulation but with low utilization value; such files are in JPEG format. The color mode for all three is RGB (24bit/pixel). (As shown in Table 2-2)

Table 2-2 Digital Image File Format of TELDAP

Category	Resolution and Size	Color Mode	File Format
Permanent Preservation	Original size, 300dpi or above	RGB (24 bit/pixel) or above	RAW or TIFF (uncompressed)
Commercial Purpose	Original size, 300dpi	RGB (24bit/pixel)	TIFF (uncompressed)
Online Browsing	Depends on website design, 72dpi	RGB (24bit/pixel)	JPEG (uncompressed)

Considering that images scanned for full-text digitization might serve different purposes in the future, the source image should fall under the category of preservation images, or 300dpi full color TIFF files, so that they may be reduced to lower specifications in the future. Furthermore, some project units use copies for key-in; the ratio of copies should be based on the font size and clearness of text in the original book.

(II) Digital File Naming Principles

Once digital images are scanned, they must be named for management and search purposes.

Notices of file naming to ensure filenames are readable in different operating systems include:

1. Use lowercase letters and numbers in filenames.
2. Avoid using special characters, such as % / ? # * -

Besides general principles, additional naming rules are required to show the type of media and different characteristics of digitization objects. Filenames for digital images of scanned books include three layers – book code, volume number, and page number. In which the page number is the filename with the file extension of .tif. Please refer to local literature image coding principles of TELDAP.⁴

e.g. aaaaa000zzzzzzzz.tif
 aaaaa=book code;
 000=volume number;
 zzzzzzzz=page number.

1. Layer 1: Book code

Variable length: Defined by project units; numbers are recommended.

2. Layer 2: Volume number

Fixed length: three digits.

3. Layer 3: Page number

(1) The length of the filename is 8 bytes and coded according to the page number of the book.

e.g. Page 1 → 00000001.tif

(2) The page number of the cover is fixed at c0000001.jpg. If the

⁴Taiwan e-Learning and Digital Archives Program, 10-4 Local Document Image Coding Principles, “Technology Collection” 2007 Version. <http://www2.ndap.org.tw/eBook08/showContent.php?PK=158>. Search: September 2010.

book was originally a paperback book but processed into hardback book, then the cover should still be of the original paperback book.

- (3) Pages in between the cover and the first page of the main text, including the preface and table of contents, should be indicated by the letter “a” in the first byte of the filename, e.g. a0000001.tif, a0000002.tif.....
- (4) Pages after the main text, such as appendices, tables and charts, and references, should be indicated by the letter “b” in the first byte of the filename, e.g. b0000001.tif, b0000002.tif.....
- (5) Blank pages or advertisements that are numbered in the book should still be scanned and named according to their order in the book.
- (6) Blank pages that are not numbered do not need to be scanned.
- (7) If there are inserts that are not numbered, then a “_” is added after the filename of the previous page and numbered according to its order. For example, if there are 2 inserts between pages 86 and 87 that were not numbered they should be named “000086_1.tif” and “000086_2.tif.”
- (8) If the left and right pages of a book are separately numbered, pages on the left should be indicated with a lowercase L and pages on the right should be indicated with a lowercase R, e.g. page 133 on the left → l0000133.tif, page 12 on the right → r0000012.tif.
- (9) If the main text has two sets of page numbers, e.g. each chapter is separately numbered starting from 1, so pages have two sets of page numbers – the page of the book and the page of the chapter, then the page number of the book should be used.
- (10) When there are any questions regarding the page number, consult with the responsible staff before scanning.

(III) Manual Input Rules

Project units must establish input results when implementing manual input. Clear cataloging formats are required, not only for the text, but also symbols, images, tables, annotations, paragraphs, page number, columns, proofreading marks, blank characters, blank lines, missing characters.....etc.

1. Page number, column: Each column should start with half width English letters and digits in the form pxxxxn, xxxx represents four digits, and n may be a (upper column), b (middle column) or c (lower column).
2. Preface and title: No spaces should be left in front.
3. Author and translator: Leave four full width spaces in front.
4. Main text: No spaces should be in front.
5. Interlinear notes: Place between a set of half width ().
e.g. is input as: 十一月 (二段)
6. Double line interlinear notes: Also placed between a set of half width (), notice the direction of the text.

十
一
月
二
段

e.g. 望江南 (三寶三段送佛一段)

7. Blank line: Leave a blank line.
8. Space: Enter a full width space.
9. Circles: Enter a “。” at the position of the circle.

望
江
南
三
送
佛
實
一
段

e.g. 身所居。二自受用土。自受

10. Proofreading tags: Use digits and a set of []

身
所
居
。
二
自
受
用
土
。
自
受

e.g. 相[01]把成陰陽。

11. Special symbols: Use full width symbols.

相把成陰陽

e.g. 相把成陰陽 is input as: 有一○為千▲洪州黃還。無□

12. Images are indicated with 【圖】.

有一▲還, 洪州無, 為千

e.g. 有一▲還, 洪州無, 為千 is input as: 【圖】第七末那識轉平等性智

13. Missing characters: Use assembled character components or glyph expressions (introduced in the following section) if possible; if characters are blurry or hard to express, use a full width ●.⁵

Every type of literature has its own typesetting, writing style or wording, so suitable manual input rules should be established accordingly. However, full-text input should generally follow the basic principles below:

1. Key-in text according to the book. Personnel should not guess any unclear areas and consult a professional.
2. Break line whenever the text in the book uses a new line.
3. Most ancient books don't have punctuation marks, so sentences are segmented during input but new punctuation marks are not added.

(IV) Missing Character System

Due to region, time or other factors, multiple forms of the same character appeared (e.g. “眾” and “□”) that cannot all be listed straightforward. Thus, once current computer exchange codes are used to handle Buddhist or Taoist scriptures, the issue of missing characters constantly appears. A fundamental and practical solution for such numerous but rarely used missing characters is to propose an effective encoding method based on component assembly rules of Chinese characters.⁶




⁵Chinese Buddhist Electronic Text Association, “Zokuzokyo Tripitaka Input Rules and Examples,” “Taisho Tripitaka Text Format,” <http://www.cbeta.org/onepage/note.htm>. Search: October 2010.

⁶Missing Chinese Characters and Input of Devanagari, Pali and Tibetan Alphabets Input” by Chuang Te-Ming, “Information Management for Buddhist Libraries,” Issue 14, June 1998.













At present, there are two standards for missing characters in Taiwan; one of them is the widely used Chinese Character Component Searching System of Academia Sinica.⁷ Basic units of Chinese characters are called components and used to assemble other characters. For example, “日” and “京” are components of “景”, “景” and “頁” are components of “顛”, and “顛” is a component of “灑”.

Components also have layers, for example “顛” can be taken apart into “景” and “頁”, in which “景” can be further broken down to “日” and “京”. The most common ways to take apart Chinese characters include horizontally joint (𠂇), vertically joint (𠂈) and contain (△). Therefore, “顛” is equal to “景” 𠂇 “頁”, “景” is equal to “日” 𠂈 “京”, “圍” is equal to “口” △ “韋”. In addition, a number of symbols were created to make input more convenient, indicating how identical components are arranged, e.g. “兢” with two horizontally joined “克” is equal to “〇〇克”, “𠂇” with two vertically joined “戈” is equal to “〇戈”, “轟” with three “車” in the shape of a triangle is equal to “〇車”, “燄” with four “火” is equal to “〇〇火”. Characters that cannot be expressed using this method can be represented with “?” (Table 2-3).

Table 2-3 Academia Sinica Chinese Character Component Assembly Rules

Type	Meaning	Symbol	Description	Example
Symbol for breaking characters apart	Horizontally joint		When components are arranged from left to right	灑, 順
	Vertically joint		When components are arranged from top to bottom	含, 義
	Contain		When components are arranged from out to in	圍, 魁, 連

⁷National Digital Archives Program, “Components and Rules for Assembled Chinese Characters,” “Technology Collection,” 2002.

Type	Meaning	Symbol	Description	Example
Symbol convenience			Two identical components vertically joined	炎
			Three identical components vertically joined	
			Two identical components horizontally joined	朋, 林, 孖
			Three identical components horizontally joined	
			Three identical components arranged in the shape of a triangle	焱, 轟, 磊
			Four identical components horizontally joined	
			Four identical components vertically joined	燄
			Four identical components arranged in the shape of a square	
Other	Start indication		When a character is broken down to more than two components, these two icons indicate the start and end of the character	 片戶
	End indication			
	Missing character indication		Used in the place of missing character components	

The other is the unique component assembly rules of the Chinese Buddhist Electronic Text Association (CBETA)⁸ developed based on the system of Academia Sinica. In contrast to the Chinese character components used by Academia Sinica, CBETA uses the BIG5 system as basic units and

⁸ Chinese Buddhist Electronic Text Association, Search: October: 2010, <http://www.cbeta.org/data-format/rare-rule.htm>.

therefore does not need to create its own characters. This eliminates the need for users to separately install any programs or image files.

The system of CBETA uses four mathematical symbols and ten symbols in total, seven of which are “*”, “/”, “@”, “-”, “+”, “(” and “)” and used to express the relative position of components; “?” indicates that a certain component cannot be expressed; in addition, two half-width symbols “[” and “] ” indicate the start and end of the character (Table 2-4).

Table 2-4 Character Component Assembly Rules of CBETA

Symbol	Description	Example
*	Horizontally joint	明 = 日 * 月
/	Vertically joint	音 = 立 / 日
@	Contain	因 = 口 @ 大 or 閒 = 門 @ 月
-	A certain component is removed	青 = 請 - 言
+	A certain component is removed and replaced with another component	閒 = 間 - 日 + 月
?	Indicates a special component that cannot be represented	背 = (? * 匕) / 月
()	Calculation separation symbol	繞 = 組 - 且 + ((土 / (土 * 土)) / 兀)
[]	Text separation symbol	羅 [目 * 侯] 羅母耶輸陀羅比丘尼

Of the two character component assembly rules described above, the former is mainly adopted by government agencies and databases developed by Academia Sinica, and is the earliest character component system developed in Taiwan; the later is used by CBETA alone and simplifies the complexity of Chinese character components, helping key-in personnel to easily assemble missing characters. Furthermore, the government also developed a national standard Chinese interchange code – the “CNS11643 Chinese Standard Interchange Code (<http://www.cns11643.gov.tw/AIDB/welcome.do>)” developed by the Electronic Data Processing Center, Directorate-General of Budget, Accounting and Statistics to resolve the issue of insufficient Chinese characters and self-made characters on the PC. However, besides government agencies and household registration offices, this standard is rarely used by the general public.

(V) Generalized Markup Language

Markup is the annotation of documents to record different information that allow computers to process the documents. To prevent self created markup systems from affecting the interoperability of data exchange, the international society has long begun to establish a common international standard. The first markup language – SGML (Standard Generalized Markup Language) was invented in 1986. SGML defines how to describe rules for a set of markup tags, but was not popular due to its great complexity. A well known example of markup language in widespread use today is HTML (Hypertext Markup Language), an instance of SGML. The simple and easy-to-use syntax of HTML led to its popularity following the rise of the internet. Different languages, cultures and operating platforms are able to communicate via HTML, a standard and common language, bringing information exchange in a global village to unprecedented speeds and extents.

However, the weakness of HTML, which is also its strength, is gradually appearing; HTML can no longer satisfy emerging needs on the internet. SGML is powerful enough but too complex, HTML is simple but not powerful enough. In the light of this, markup language specialists designed a powerful but not too complicated markup language suitable for the internet – XML (eXtensible Markup Language).

Markups can be divided into two categories, “typesetting or format display” markup and “data structure or content” markup. For example, HTML (Hypertext Markup Language) might be used as follows:

Discussion on Buddhist Data Digitization Technology – In the Case of the < b >Chinese Buddhist Electronic Text Association< /b >.

Here, < b >.....< /b > indicates “Chinese Buddhist Electronic Text Association” should be in bold letters. This is the first type of markup for “format.” Markup for “content” might be in the follow form:

Records of the Grand Historian

< byline type="Author">Sima Qian< /byline >

Here, < byline >.....</ byline > indicates that the author of Records of the Grand Historian is Sima Qian. By separating “display format” from

content, computers are able to “understand” contents.⁹

With a common markup language – XML, markup format will be the same. The same markup language and format can be used to define different tag names; for example, there are the following methods to mark a paragraph:

1. < p >.....< /p >
2. < para >.....< /para >
3. < 段落 >.....< /段落 >

These markups all conform to XML standards, but will cause problems in information exchange, and require a markup conversion procedure. This issue can be resolved if there is a unified format for markups.

In addition, the TEI (Text Encoding Initiative), which had existed since SGML, studied different Western literatures and compiled a tag set in hopes of driving electronic literature sharing and exchanges. Since the TEI tag set is based on literature, it better corresponds to contents and structure of literature compared with SGML, HTML or XML. For example, there are specific tags for complete bibliographic information, relationship between literature and its source, version and language used, satisfying document markup requirements. This guideline will introduce the TEI and operating rules in chapter five – metadata establishment.

⁹“Markup Language Applications” by Chou Pang-Shen, “Information Management for Buddhist Libraries,” Issue 24, December 2000.



Four. Object Digitization Procedures

Object digitization procedures refer to scanning, input and proofreading operations in the digitization process. These three operations can be considered the backbone or foundation of full-text digitization. In fact, books can be considered digitized after they are scanned, input and proofread. However, electronic full-text produced from these three operations are raw materials that still need to go through tagging operations to gain value for future applications and research.

I. Scanning

Considerations when deciding whether to purchase a scanner for digitization or outsourcing scanning operations include the size of the collection, project funding and cost effectiveness. If project units decide to scan books themselves, selecting a scanner with “automatic paper feeding function” and “automatic numbering and file saving function” will save labor cost. In the case of the Institute of Modern History, Academia Sinica, book scanning operations are as shown in the flowchart below:¹⁰

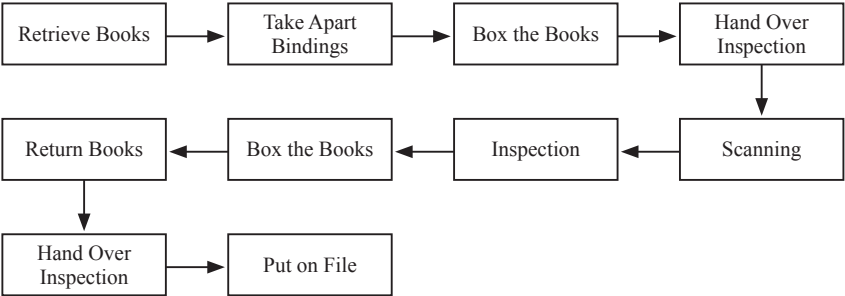


Fig.4-1 Book Scanning Flowchart of the Institute of Modern History, Academia Sinica

(I) Retrieve Books

Retrieve books on the booklist for scanning personnel or the contractor to copy and scan.

¹⁰National Digital Archives Program, “1-9 Digital Archiving Workplace Regulations: In the Case of Diplomatic Files,” “Technology Collection.” Search: September 2010, website: http://www.ndap.org.tw/2_techreport/files/174.pdf.

(II) Take Apart the Bindings

Take apart the bindings of books or copies and cut along the junction of the edges of every two pages for scanning.

(III) Box the Books

Place the book in a paper box and label the title and volume number for identification.

(IV) Hand Over Inspection

Staff members should check that the book title and quantity are correct and sign the form (Table 4-1) before allowing scanning personnel or the contractor to retrieve books.

Table 4-1 Form for Retrieving Book

Title	Quantity	Date Retrieved	Signature of Borrower	Date Returned	Signature of Responsible Staff	Notes

(V) Scanning

1. Scanning procedures are as follows:

- (1) Scanning.
- (2) Sample inspection of scanning quality – Whether if there are lines or if the image is slanted or unclear.
- (3) After scanning a book, check if any pages were left out.
- (4) Name files according to file naming principles.
- (5) Sample inspection of page number.
- (6) File conversion.
- (7) Burn CD.
- (8) Browse files to see if any are missing or unreadable.
- (9) Put on file.
- (10) Clean the scanner.

2. Notices of Scanning

- (1) In principle one book should be handled by one person. Scanning personnel may only move on to the next book after the previous book has been completely scanned. Scanning different books at the same time or assigning two or more scanning personnel to the same book is strictly prohibited.
- (2) The scanner should be configured to output high-end images, or 300dpi TIFF full color images. Image files used for manual input can be converted into TIFF-g4 black and white images, which are clear and small files.

(VI) Inspection

After scanning books, compile a file list with filename, file size and number of entries for subsequent quality and quantity inspections.

(VII) Box, Check and Return Books

When scanning personnel or the contractor returns books, they should restore it to its original state and arrange them in their original order.

II. Input

There are two ways to input text, one is to manually key-in text (the old fashioned way) and the other is to use OCR (Optical Character Recognition) software. The latter saves time and effort because it can automatically analyze characters on an image file and convert them into a text file. However, current OCR systems are only able to recognize machine printed text, and are still unable to accurately recognize handwritten copies of ancient books and records or text printed using woodblocks. Therefore, books and records unsuitable for OCR is manually input character by character, sentence by sentence; this is currently the most commonly used input method by full-text digitization projects in Taiwan.

In addition, there is still another way to produce text files, that is to collect electronic files on the internet, and then adjust its format into the project unit's text format. Normally, only history and religious books have existing electronic text files, especially Buddhist scriptures, because Buddhists believe that copying scriptures can accumulate charitable and pious deeds. For this reason, numerous believers have voluntarily keyed-in text online to accumulate

charitable and pious deeds and benefit the preaching of Buddhist dharma. Currently, only the CBETA has collected Buddhist electronic text from the internet for reference during proofreading. In the following section we will introduce operating procedures and notices on manual input and OCR.

(I) Manual Input

Manual input is an extremely time-consuming task as books need to be keyed-in character by character into an electronic file. Unless the number of books planned for digitization is extremely small and can be handled by existing manpower of project units, most full-text digitization projects outsource this operation. Even with continuous rise of labor cost today, outsourcing manual input is still the best option for minimizing labor requirements of project units.

Markets that provide manual Chinese input services in Asia besides Taiwan include Mainland China and India. Taiwan's manual input service providers offer the advantages of high quality and effective communications, but are relatively expensive; Mainland China offers relatively low cost, but the quality of service providers is inconsistent and might require strenuous communication, or complete training; India offers low labor cost and high quality at the same time, but input language is limited to European or Indian languages. Therefore, most full-text digitization projects in Taiwan choose to engage in long-term cooperation with a reputable domestic contractor.

Project units should use scanned images or copies for input, and follow input rules previous established. There are no restriction as to what software should be used, any word processing software should suffice, e.g. Notepad++, Gedit, Hanshu 漢書98, or Windows notebook. However, keyed-in text files should only contain text and named according to the book number. This is to prevent interference and constraints of file format when the files are applied for different purposes in the future.

In the case of the CBETA, the project uses 漢書2000 for input, and requires that the correct annotation information be provided, demanding that in addition to the main text, the page and column that the text is located (At the upper left of Fig.4-2 p0001a indicates the first column of the first page) and number of lines should also be recorded to retain the distinguishing layout of Buddhist scriptures.

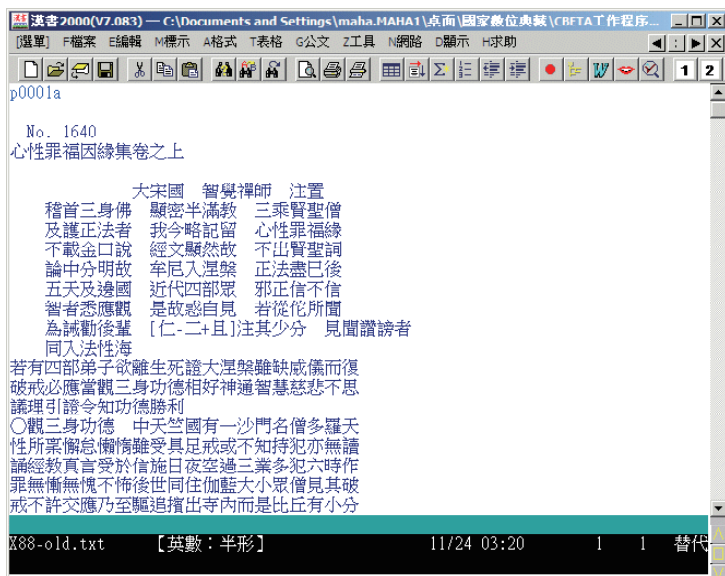


Fig. 4-2 Electronic Text File of the CBETA Manually Input by a Contractor

(II) OCR System

OCR (Optical Character Recognition) uses an optical input device, such as a scanner or digital camera to acquire text data on printed or handwritten documents, and then uses various recognition algorithms to identify and analyze characteristics of each character, transforming the image into machine readable text, e.g. American National Standard Code for Information Interchange (ASCII code) or Unicode, after which it can be imported into a database for users to search.

From the perspective of OCR, Chinese characters are far more difficult to recognize than European and American words. Due to the massive number of Chinese characters and their complex structure and variants, OCR of Chinese characters only entered practical use in the past few years. Concerning the research and development of OCR technology, Taiwan has the 丹青中英日文文件辨識系統, 蒙恬認識王專業系統, and 全景軟體; Mainland China has TH-OCR and 北京漢王.¹¹ For a comparison of different OCR systems, refer to the

¹¹ “Journals and Newspapers Digitization Procedures Guideline” by Li Pei-Ying and Cheng Wan-Ju, Taipei City: Taiwan Digital Archives Expansion Project, April 2009, pages 33-44.

“Journals and Newspapers Digitization Procedures Guideline” (<http://content.ndap.org.tw/index/?p=1000&page=5>) published by this project in 2009.

OCR results are easily affected by the clarity of image files, thus a number of preparations and adjustments must be made for OCR:

1. File Conversion

Larger contrast between text and background benefits OCR, so color image must be converted into black and white or gray scale images to remove excessive colors for OCR, thus increasing its accuracy.

2. Remove Specks

Some books have phonic symbols or marks not belonging to the text itself (Fig.4-3, checks and dashes between lines). Specks must first be removed using image processing software to produce a new clearer image before OCR can be applied.

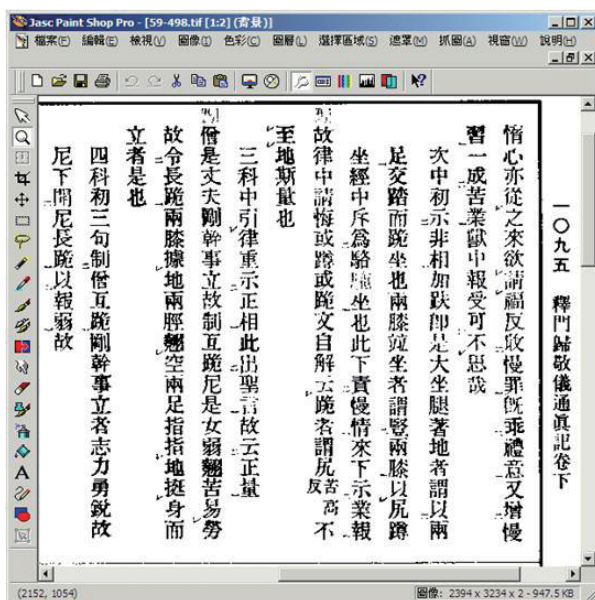


Fig.4-3 Scanned Image File of the CBETA with Phonics Symbols and Specks

3. Import into the OCR System

Import the converted file into the selected OCR system (Fig.4-4, Fig.4-5) and execute the OCR function to produce a text file.

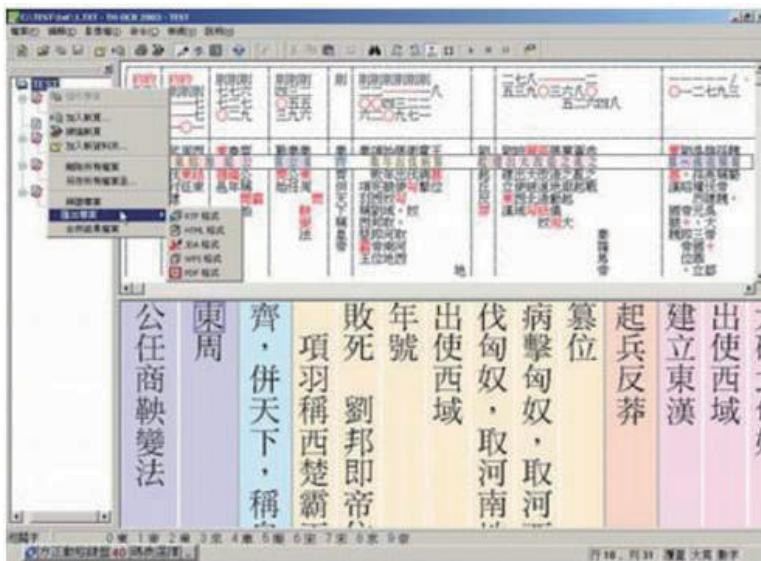


Fig.4-4 TH-OCR System Interface

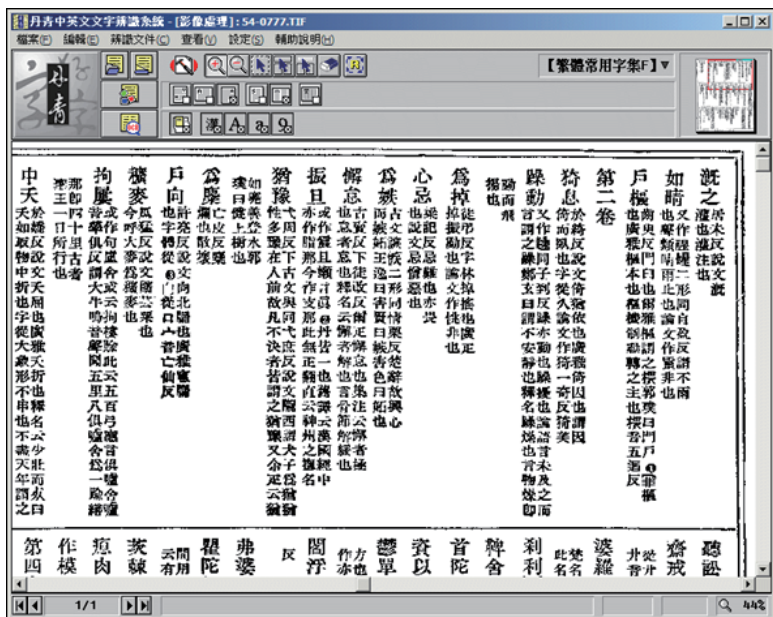


Fig.4-5 丹青 OCR System Interface

4. String Replacement

Due to the complex structure of Chinese characters and the large number of similar characters, OCR will make mistakes recognizing characters. The correct character can be determined based on the string of characters it is in, so if a list of common mistakes (Fig.4-6) can be compiled, it will allow the rapid replacement of incorrect character strings of OCR. This will make proofreading more convenient, and increase the accuracy of OCR generated text files to roughly 90%.

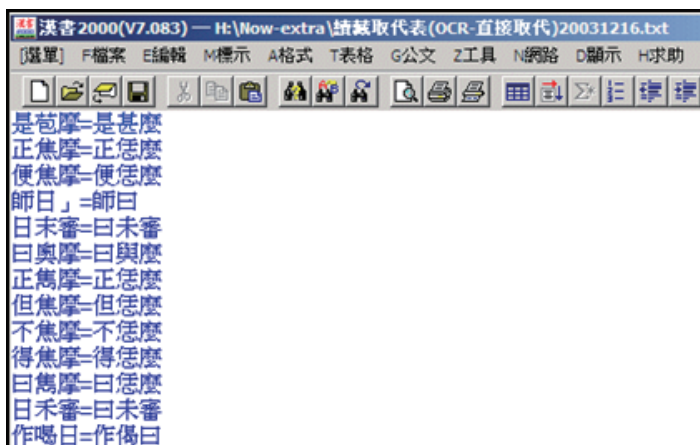


Fig.4-6 CBETA List of Common Incorrect Character Strings

III. Proofreading

According to studies and statistics, one person can type roughly 48 thousand characters in a workday with an error rate of 0.4~0.5%. In other words, the accuracy of manual input is roughly 99.6%; depending on the clearness of the source text, accuracy of OCR ranges from 90% (ancient books that contain relatively more symbols, character variants and missing characters) to 96% (typeset modern books). Since the accuracy of text concerns the quality of digitization results, project units should try to lower the error rate as much as possible. Most project units currently set the target error rate at 0.01%, or only one wrong character is allowed out of every ten thousand characters. The only way to achieve this goal is to dedicate great effort into proofreading.

Current proofreading methods include manual proofreading, volunteer proofreading and file comparison; the latter two are products of the internet and technology. Using manual proofreading together with volunteer proofreading and file comparison will improve the accuracy of text, and produce high quality electronic full-text even closer to the original.

However, increasing the accuracy of text is not limited to correcting characters, but also involves missing characters, character variants and taboo characters that result from Chinese character changes and computer systems. Therefore, the final section focuses on solutions to missing characters, character variants and taboo characters.

(I) Manual Proofreading

1. First Proofreading

The most conventional approach is manually proofreading character by character, page by page. Although this requires high cost of labor and time, it has the lowest technology barrier, and is still the first option for most project units when it comes to proofreading. The first proofreading is usually carried out by the contractor or key-in personnel, who should proofread the text after key-in and immediately correct any mistakes.

2. Second Proofreading

After text files that have been proofread once are sent back to the project unit, project personnel should immediately conduct a second proofreading. Unlike the first proofreading, the second proofreading is a random inspection. If text files meet requirements, they are burned on an optical disk for backup; if they do not meet requirements, they are rejected for the contractor or scanning personnel to make corrections and arrange for second inspection. At present, most full-text digitization projects in Taiwan set the standard at an error rate of at most 0.01%.

(II) Volunteer Proofreading

Due to the unique characteristics of certain books or unique advantages of certain project units, sometimes large groups of volunteers can be utilized to proofread electronic text. For example, Buddhists believe that copying scriptures will allow them to accumulate merit. Therefore CBETA utilized an “online proofreading” mechanism to find some nine hundred volunteers, and assigned the task of proofreading one page to each volunteer; procedures for volunteer proofreading are as follows:

1. Submit an application on the website of CBETA (<http://www.cbeta.org/index.htm>).
2. Retrieve a text file and image file of the scripture.
3. Proofread the text file by comparing it with the image file.
4. Send the text file back after proofreading it.

In a full-text digitization project of Beijing, a superior department directly assigned professors of related fields in numerous universities to proofread text files. Gathering together large numbers of scholars to assist with digitization will not only accelerate work efficiency, but also improve the quality of digitization results.

So how are the results of volunteer proofreading? According to statistics of the CBETA, Buddhist electronic text files generated using OCR had an accuracy of only 90%, but accuracy reached 98% after volunteer proofreading.

(III) File Comparison

In the conventional approach, even if a text file is proofread four times or ten times, there might still be incorrect characters. In the light of this, the CBETA designed a file comparison program (Fig.4-7) that compares two text files containing the same content but were produced using different input methods (e.g. manual input and OCR), and outputs a separate file containing differences between the two files. This saves time looking for mistakes one character at a time, and allows proofreading personnel to directly find mistakes and correct them. If other project units wish to use this method, they can ask their information department to design a similar program to increase proofreading efficiency.

(IV) Chinese Character Variants

Character variants are different characters with the same sound and meaning of the standard character, e.g. variants of “體” on the Ministry of Education’s “Dictionary of Chinese Character Variants” (<http://140.111.1.40/>) include “体 and others (see Fig.4-8). Character variants were created for different reasons in different time periods and regions, but should be viewed as the same word. However, sometimes the usage of Chinese character variants is determined by the context and differences in syntax and semantics cannot be ruled out. Therefore, Chinese character

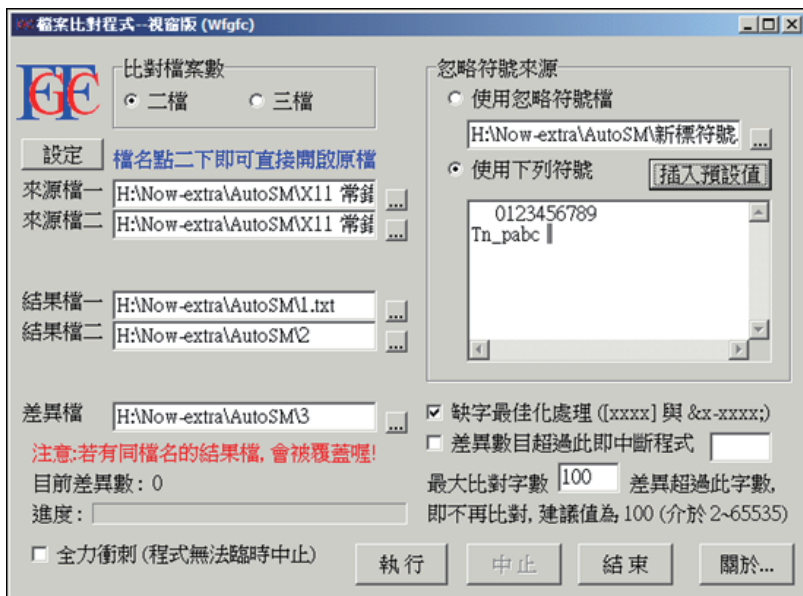


Fig.4-7 Interface of the CBETA's File Comparison Program

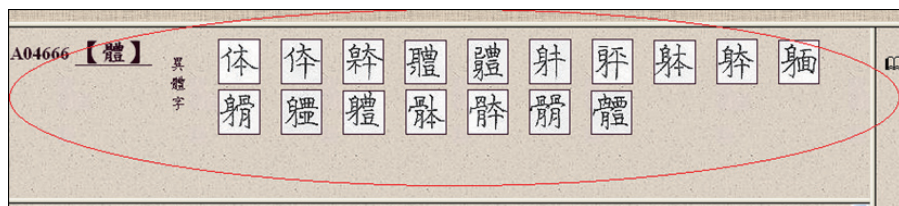


Fig.4-8 Dictionary of Chinese Character Variants of the Ministry of Education

variants are often difficult to handle.¹²

Although it is difficult to determine character variants, most full-text digitization projects currently still use the principle of retaining the original character during input. Rules for handling character variants of the Chinese Classics Full-Text Database Team of Academia Sinica, which created

¹²“UNICODE List of Variant Forms of Chinese Characters and Dictionary of Chinese Character Variants” by Yuen Kuo-Hua, “National Digital Archives Program Newsletter,” Volume 3 Issue 8, 2004. Search: September 2010, Website: http://www2.ndap.org.tw/newsletter06/news/read_news.php?nid=217.

Scripta Sinica, are as follows:

1. If the character variant has the same meaning and usage of the standard character, then it is input as the standard character.
2. If the meaning and usage of a character variant cannot be determined from the context, then the original appearance of the character is retained.
3. There are still character variants with the same sound and meaning, and the only difference is how a certain component is written. For example, “福” (U+798F) and “福” (U+FA1B) have in principle the same radical “示”, which is written differently. For the convenience of browsing, only the standard form “福” is used for input, and the variant is separately indicated. If the meaning of the character variant cannot be ascertained, then the original appearance is retained.

(V) Missing Characters

During early periods, most full-text digitization projects in Taiwan adopted BIG5 because international standard codes were inadequate for Chinese characters, driving the development of a coding system to make up for missing characters. However, since the appearance of the more powerful Unicode, new projects now directly use Unicode and permanent units of old projects are also gradually converting from BIG5 to Unicode. This is because some missing characters in BIG5 were coded in Unicode, thus reducing the number of missing characters.

The most common approach to handling missing characters is still to “create characters for missing characters,” meaning that when a Chinese character is not listed in Unicode, it is sent to the Chinese Document Processing Lab, Academia Sinica’s dedicated unit for processing missing characters, which then uses its Chinese Character Component Searching System to create an image file of the character, and then codes the image with its coding method; this process is shown in the flowchart below:¹³

¹³Taiwan Historica missing character processing procedures.

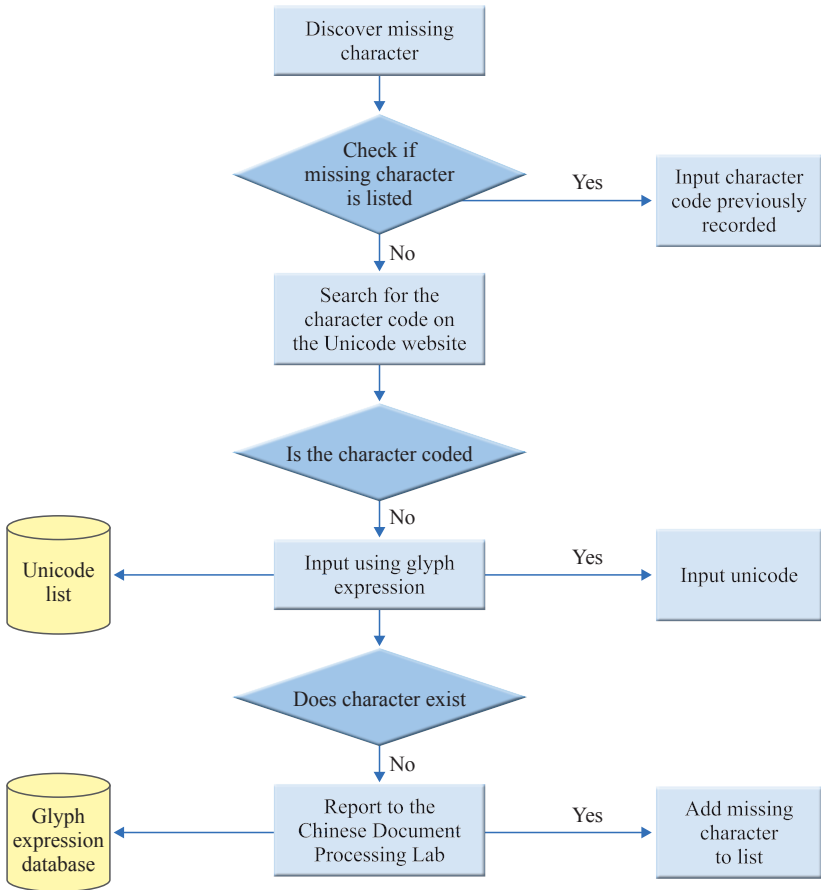


Fig.4-9 Missing Character Processing Procedures

To search for a missing character on the website of Unicode (Fig.4-10) (<http://www.unicode.org/charts/unihanrsindex.html>), first click on the radical-stroke count of the character, and then search for the character based on the remaining strokes to see if the missing character was included in Unicode. For more information on the Chinese Character Component Searching System of the Chinese Document Processing Lab of the Institute for Information Science, Academia Sinica, visit <http://www.sinica.edu.tw/~cdp/>.

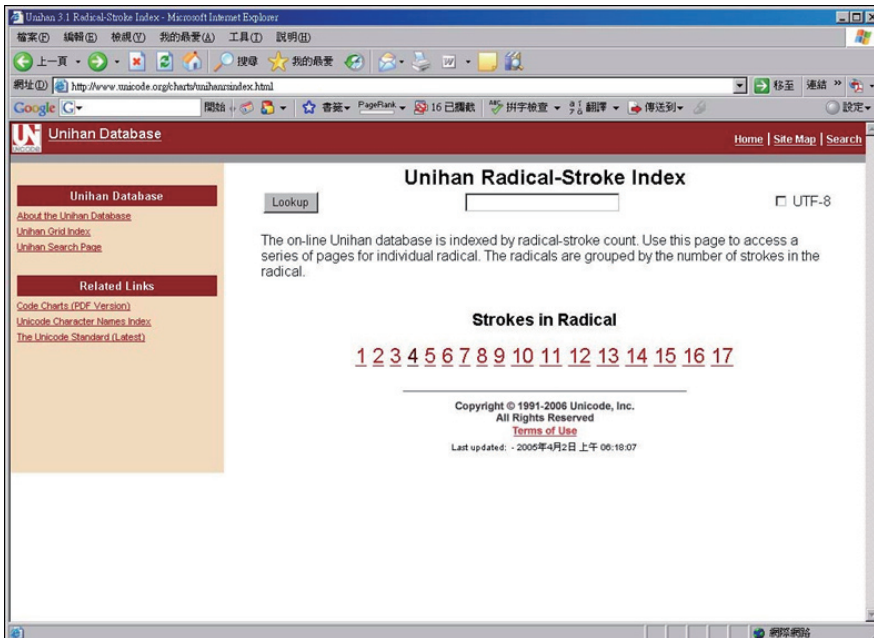


Fig.4-10 Unihan Database

With regards to the issue of missing characters, full-text digitization projects today are gradually turning away from creating new characters, and they are not waiting for Unicode to include newly discovered missing characters either. Instead, they are approaching the issue from markup. When a missing character appears in text, tags and text are used to describe the usage and meaning of the missing character, as well as how components can be assembled to create the missing character. This approach saves time and effort creating characters, users can read the descriptions with ease, and researchers can follow the tags to gain source information and metadata of the missing character.

(VI) Apply New Punctuation Marks

The new punctuation marks we use today were established by Hu Shih and Ma Yu-Tsao et al. after the May Fourth Movement in 1919 based on old punctuation marks and adopting Western punctuation marks. Before this, most ancient Chinese books and records only used periods and pauses

(° and 丶) to indicate pauses and the end of a sentence. Question marks and exclamation marks did not exist then, so many readers today find pre modern texts hard to understand.

Therefore, besides recreating the exact same text in electronic form, full-text digitization projects can also add new punctuation to increase the research and dissemination value of the electronic text. However, this task will require experts and researchers to add punctuation and then proofread the results, again a time consuming task.

After the documents are keyed in, proofread, and character variants and missing characters are processed, the result can be called a complete electronic text. This is where most full-text digitization projects end. Yet, in an age that emphasizes digitization, resource circulation and value-added applications, such electronic text is not only inadequate, but also difficult to share due to different input and proofreading methods. Therefore, it is often necessary to markup the text files using a markup language that conforms to international standards, so as to achieve resource sharing, knowledge circulation and drive academic research. Only then will electronic text gain higher value for research, application and sharing.

The TEI (Text Encoding Initiative) is the leading international markup standard for the encoding of text, independent of language, era or genre. TEI is expressed in XML and TEI files contain Metadata on various aspects of a text in the “teiHeader”. It is therefore treated here in the following chapter on metadata.



Five. Metadata Establishment

I. Metadata Exclusive for Corpora – TEI

The TEI (Text Encoding Initiative, <http://www.tei-c.org/>) is an international interdisciplinary standard that has been widely used by libraries, museums, publishers, and individual scholars to present texts of various literatures and in the field of linguistics for online teaching and access. TEI is expressed in XML since version P4 (2002). Its latest version TEI P5, released in 2006 and updated incrementally under this version number, added a module for missing characters and a completely new management and customization system expressed in a dedicated TEI module (ODD).

The structure of a TEI document can be divided into two parts: the teiHeader and the actual text. The Header is similar to the copyright page of a book, and not only records the source, author, and publisher of the print or manuscript original, but also the name, identity, year and purpose of the digital text and the markup.

II. TEI Core Elements

TEI is a continuously evolving standard. Its development is overseen by the TEI Consortium an association of stakeholders in the Digital Humanities. TEI differs from other XML standards in that it is designed to allow for a high degree of customization. As the Humanities are dealing with complex texts it is necessary to give projects great freedom in what to encode about them. The TEI schema can therefore be customized in regular ways that document each customization.

In view of the central importance of TEI for the Digital Humanities, TELDAP in conjunction with Dharma Drum Buddhist College, has decided to localize the standard. The resulting Chinese version of all element and attribute definitions, the ROMA interface, and parts of the guidelines were published in 2009 as Bingenheimer, Marcus 馬德偉 (ed.): *TEI shiyong zhinan - yunyong TEI chuli zhongwen wenxian* TEI使用指南——運用TEI處理中文文獻 [Chinese TEI – A guide to using TEI with Chinese texts], Taipei: Taiwan E-learning and Digital Archive Program 數位典藏與數位學習國家型科技計畫. 384pgs. ISBN:978-986-01-8092-3. The translations and localized examples are also available through the TEI website, where the localization stands to make a great difference for the acceptance of the standard in the Chinese speaking world.

(I) List of Elements Described

The following list shows all the elements defined for the TEI Lite schema, with a brief description of each, and a link to its full specification in the Appendix.

- <abbr> contains an abbreviation of any sort.
- <add> contains letters, words, or phrases inserted in the text by an author, scribe, annotator, or corrector.
- <address> contains a postal or other address, for example of a publisher, an organization, or an individual.
- <addrLine> contains one line of a postal or other address.
- <anchor> attaches an identifier to a point within a text, whether or not it corresponds with a textual element.
- <argument> a formal list or prose description of the topics addressed by a subdivision of a text.
- <author> in a bibliographic reference, contains the name of the author(s), personal or corporate, of a work; the primary statement of responsibility for any bibliographic item.
- <authority> supplies the name of a person or other agency responsible for making an electronic file available, other than a publisher or distributor.
- <availability> supplies information about the availability of a text, for example any restrictions on its use or distribution, its copyright status, etc.
- <back> contains any appendixes, etc. following the main part of a text.
- <bibl> contains a loosely-structured bibliographic citation of which the sub-components may or may not be explicitly tagged.
- <biblFull> contains a fully-structured bibliographic citation, in which all components of the TEI file description are present.
- <biblScope> defines the scope of a bibliographic reference, for example as a list of pagenumbers, or a named subdivision of a larger work.
- <body> contains the whole body of a single unitary text, excluding any front or back matter.
- <byline> contains the primary statement of responsibility given for a work on its title page or at the head or end of the work.

- <catDesc> describes some category within a taxonomy or text typology, either in the form of a brief prose description or in terms of the situational parameters used by the TEI formal textDesc.
- <category> contains an individual descriptive category, possibly nested within a superordinate category, within a user-defined taxonomy.
- <catRef> specifies one or more defined categories within some taxonomy or text typology.
- <cell> contains one cell of a table.
- <change> summarizes a particular change or correction made to a particular version of an electronic text which is shared between several researchers.
- <choice> groups a number of alternative encodings for the same point in a text.
- <cit> a quotation from some other document, together with a bibliographic reference to its source.
- <classCode> contains the classification code used for this text in some standard classification system.
- <classDecl> contains one or more taxonomies defining any classificatory codes used elsewhere in the text.
- <closer> groups together dateline, byline, salutation, and similar phrases appearing as a final group at the end of a division, especially of a letter.
- <code> contains literal code.
- <corr> contains the correct form of a passage apparently erroneous in the copy text.
- <creation> contains information about the creation of a text.
- <date> contains a date in any format.
- <dateline> contains a brief description of the place, date, time, etc. of production of a letter, newspaper story, or other work, prefixed or suffixed to it as a kind of heading or trailer.
- contains a letter, word or passage deleted, marked as deleted, or otherwise indicated as superfluous or spurious in the copy text by an author, scribe, annotator, or corrector.
- <distributor> supplies the name of a person or other agency responsible for the distribution of a text.

- <div> contains a subdivision of the front, body, or back of a text.
- <divGen> indicates the location at which a textual division generated automatically by a text-processing application is to appear.
- <docAuthor> contains the name of the author of the document, as given on the title page (often but not always contained in a byline).
- <docDate> contains the date of a document, as given (usually) on a title page.
- <docEdition> contains an edition statement as presented on a title page of a document.
- <docImprint> contains the imprint statement (place and date of publication, publisher name), as given (usually) at the foot of a title page.
- <docTitle> contains the title of a document, including all its constituents, as given on a title page.
- <edition> describes the particularities of one edition of a text.
- <editionStmnt> groups information relating to one edition of a text.
- <editor> secondary statement of responsibility for a bibliographic item, for example the name of an individual, institution or organization, (or of several such) acting as editor, compiler, translator, etc.
- <editorialDecl> provides details of editorial principles and practices applied during the encoding of a text.
- <eg> contains a single example demonstrating the use of an element or attribute.
- <emph> marks words or phrases which are stressed or emphasized for linguistic or rhetorical effect.
- <encodingDesc> documents the relationship between an electronic text and the source or sources from which it was derived.
- <epigraph> contains a quotation, anonymous or attributed, appearing at the start of a section or chapter, or on a title page.
- <extent> describes the approximate size of the electronic text as stored on some carrier medium, specified in any convenient units.
- <figure> indicates the location of a graphic, illustration, or figure.
- <fileDesc> contains a full bibliographic description of an electronic file.
- <foreign> identifies a word or phrase as belonging to some language

other than that of the surrounding text.

- `<formula>` contains a mathematical or other formula.
- `<front>` contains any prefatory matter (headers, title page, prefaces, dedications, etc.) found at the start of a document, before the main body.
- `<funder>` specifies the name of an individual, institution, or organization responsible for the funding of a project or text.
- `<gap>` indicates a point where material has been omitted in a transcription, whether for editorial reasons described in the TEI header, as part of sampling practice, or because the material is illegible or inaudible.
- `<gi>` contains the name (generic identifier) of an element.
- `<gloss>` identifies a phrase or word used to provide a gloss or definition for some other word or phrase.
- `<group>` contains the body of a composite text, grouping together a sequence of distinct texts (or groups of such texts) which are regarded as a unit for some purpose, for example the collected works of an author, a sequence of prose essays, etc.
- `<head>` contains any heading, for example, the title of a section, or the heading of a list or glossary.
- `<hi>` marks a word or phrase as graphically distinct from the surrounding text, for reasons concerning which no claim is made.
- `<ident>` identifies or names any kind of object
- `<idno>` supplies any standard or non-standard number used to identify a bibliographic item.
- `<imprint>` groups information relating to the publication or distribution of a bibliographic item.
- `<index>` marks a location to be indexed for whatever purpose.
- `<interp>` provides for an interpretative annotation which can be linked to a span of text.
- `<interpGrp>` collects together `interp` tags.
- `<item>` contains one component of a list.
- `<keywords>` contains a list of keywords or phrases identifying the topic or nature of a text.
- `<l>` contains a single, possibly incomplete, line of verse.

- <label> contains the label associated with an item in a list; in glossaries, marks the term being defined.
- <language> characterizes a single language or sublanguage used within a text.
- <langUsage> describes the languages, sublanguages, registers, dialects etc. represented within a text.
- <lb> marks the start of a new (typographic) line in some edition or version of a text.
- <lg> contains a group of verse lines functioning as a formal unit, e.g. a stanza, refrain, verse paragraph, etc.
- <list> contains any sequence of items organized as a list.
- <listBibl> contains a list of bibliographic citations of any kind.
- <mentioned> marks words or phrases mentioned, not used.
- <milestone> marks the boundary between sections of a text, as indicated by changes in a standard reference system.
- <name> contains a proper noun or noun phrase.
- <note> contains a note or annotation.
- <notesStmt> collects together any notes providing information about a text additional to that recorded in other parts of the bibliographic description.
- <num> contains a number, written in any form.
- <opener> groups together dateline, byline, salutation, and similar phrases appearing as a preliminary group at the start of a division, especially of a letter.
- <orig> contains the original form of a reading, for which a regularized form is given in an attribute value.
- <p> marks paragraphs in prose.
- <pb> marks the boundary between one page of a text and the next in a standard reference system.
- <principal> supplies the name of the principal researcher responsible for the creation of an electronic text.
- <profileDesc> provides a detailed description of non-bibliographic aspects of a text, specifically the languages and sublanguages used, the situation in which it was produced, the participants and their setting.
- <projectDesc> describes in detail the aim or purpose for which

an electronic file was encoded, together with any other relevant information concerning the process by which it was assembled or collected.

- <ptr> defines a pointer to another location in the current document in terms of one or more identifiable elements.
- <publicationStmt> groups information concerning the publication or distribution of an electronic or other text.
- <publisher> provides the name of the organization responsible for the publication or distribution of a bibliographic item.
- <pubPlace> contains the name of the place where a bibliographic item was published.
- <q> contains a quotation or apparent quotation — a representation of speech or thought marked as being quoted from someone else (whether in fact quoted or not); in narrative, the words are usually those of a character or speaker; in dictionaries, q may be used to mark real or contrived examples of usage.
- <ref> defines a reference to another location in the current document, in terms of one or more identifiable elements, possibly modified by additional text or comment.
- <refsDecl> specifies how canonical references are constructed for this text.
- <reg> contains a reading which has been regularized or normalized in some sense.
- <rendition> supplies information about the intended rendition of one or more elements.
- <resp> contains a phrase describing the nature of a person's intellectual responsibility.
- <respStmt> supplies a statement of responsibility for someone responsible for the intellectual content of a text, edition, recording, or series, where the specialized elements for authors, editors, etc. do not suffice or do not apply.
- <revisionDesc> summarizes the revision history for a file.
- <row> contains one row of a table.
- <rs> contains a general purpose name or referring string.
- <s> contains a sentence-like division of a text.

- <salute> contains a salutation or greeting prefixed to a foreword, dedicatory epistle, or other division of a text, or the salutation in the closing of a letter, preface, etc.
- <samplingDecl> contains a prose description of the rationale and methods used in sampling texts in the creation of a corpus or collection.
- <seg> contains any arbitrary phrase-level unit of text (including other seg elements).
- <seriesStmt> groups information about the series, if any, to which a publication belongs.
- <sic> contains text reproduced although apparently incorrect or inaccurate.
- <signed> contains the closing salutation, etc., appended to a foreword, dedicatory epistle, or other division of a text.
- <soCalled> contains a word or phrase for which the author or narrator indicates a disclaiming of responsibility, for example by the use of scare quotes or italics.
- <sourceDesc> supplies a description of the source text(s) from which an electronic text was derived or generated.
- <sp> An individual speech in a performance text, or a passage presented as such in a prose or verse text.
- <speaker> A specialized form of heading or label, giving the name of one or more speakers in a dramatic text or fragment.
- <sponsor> specifies the name of a sponsoring organization or institution.
- <stage> contains any kind of stage direction within a dramatic text or fragment.
- <table> contains text displayed in tabular form, in rows and columns.
- <taxonomy> defines a typology used to classify texts either implicitly, by means of a bibliographic citation, or explicitly by a structured taxonomy.
- <TEI> contains a single TEI-conformant document, comprising a TEI header and a text, either in isolation or as part of a `teiCorpus` element.
- <teiHeader> supplies the descriptive and declarative information making up an electronic title page prefixed to every TEI-conformant

text.

- <text> contains a single text of any kind, whether unitary or composite, for example a poem or drama, a collection of essays, a novel, a dictionary, or a corpus sample.
- <term> contains a single-word, multi-word, or symbolic designation which is regarded as a technical term.
- <textClass> groups information which describes the nature or topic of a text in terms of a standard classification scheme, thesaurus, etc.
- <time> contains a phrase defining a time of day in any format.
- <title> contains the title of a work, whether article, book, journal, or series, including any alternative titles or subtitles.
- <titlePage> contains the title page of a text, appearing within the front or back matter.
- <titlePart> contains a subsection or division of the title of a work, as indicated on a title page.
- <titleStmt> groups information about the title of a work and those responsible for its intellectual content.
- <trailer> contains a closing title or footer appearing at the end of a division of a text.
- <unclear> contains a word, phrase, or passage which cannot be transcribed with certainty because it is illegible or inaudible in the source.

(II) Appendixes

Appendix A Substantive changes from the P4 version

This revision of the TEI Lite schema conforms to the TEI P5 Guidelines, which makes a number of changes from the TEI P4 Guidelines underlying earlier versions of TEI Lite. The following brief list indicates some of the major changes which will be needed in existing TEI P4-conformant documents before they can be used with the new schema. A fuller list is in preparation for publication as a part of TEI P5: the items listed here relate specifically to changes in TEI Lite only.

- At P5, a TEI document must declare a namespace of <http://www.tei-c.org/ns/1.0>
- The attributes `id` and `lang` are replaced by the attributes `xml:id` and `xml:lang` respectively. Values for the latter attribute must conform to

RFC 3066

- The element <choice> must be used to wrap <reg> and <orig> if both are supplied. Similarly for <sic> and <corr>, and for <abbr> and <expan>.
- ‘numbered divs’ (<div0>, <div1>, etc.) are not supported in this version of TEI Lite
- all pointing and linking mechanisms now use the same W3C-defined mechanism: there is no longer any distinction between internal and external pointing elements
- the content model of <change> has changed significantly

Appendix B Formal specification

The TEI Lite is a pure subset of the TEI. All of the elements defined in it are taken from the following standard TEI modules: tei, core, header, textstructure, figures, linking, analysis, and tagdocs.

The following elements from those modules are excluded from the schema: <ab>, <alt>, <altGrp>, <altIdent>, <analytic>, <attDef>, <attList>, <attRef>, <biblItem>, <biblStruct>, <binaryObject>, <broadcast>, <c>, <cb>, <cl>, <classSpec>, <classes>, <content>, <correction>, <datatype>, <dateRange>, <defaultVal>, <desc>, <distinct>, <div0>, <div1>, <div2>, <div3>, <div4>, <div5>, <div6>, <div7>, <egXML>, <elementSpec>, <equipment>, <equiv>, <exemplum>, <fsdDecl>, <headItem>, <headLabel>, <hyphenation>, <imprimatur>, <interpretation>, <join>, <joinGrp>, <link>, <linkGrp>, <listRef>, <m>, <macroSpec>, <measure>, <meeting>, <memberOf>, <metDecl>, <metSym>, <moduleRef>, <moduleSpec>, <monogr>, <normalization>, <phr>, <postBox>, <postCode>, <quotation>, <quote>, <recording>, <recordingStmt>, <remarks>, <schemaSpec>, <scriptStmt>, <segmentation>, <series>, , <spanGrp>, <specDesc>, <specGrp>, <specGrpRef>, <specList>, <state>, <stdVals>, <street>, <stringVal>, <tag>, <timeRange>, <timeline>, <valDesc>, <valItem>, <valList>, <variantEncoding>, <w>, <when>

III. TEI Best Practices

(I) Best Practices for Markup Procedures

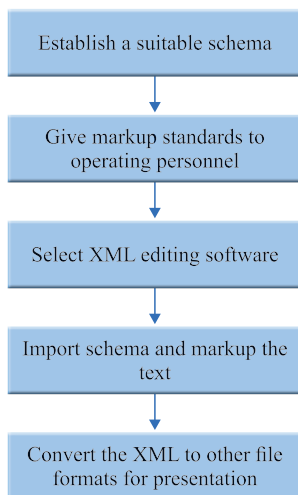


Fig.5-1 Markup Procedures

(II) Establish a Suitable Schema

TEI Markup Language consists of more than 400 elements with complex nesting rules, resulting in a relatively steep learning curve. However, most project units do not need to use all of the elements that TEI has to offer, and will select different subsets based on the genre of the text and the purpose of the markup. To assist with the creation of customized validators for individual projects, TEI maintains a tool called ROMA. Most users will want to use the online version of ROMA, as it can be used easily to create new and update existing schemas.

For a detailed introduction to ROMA refer to the TELDAP localization of TEI cited above.

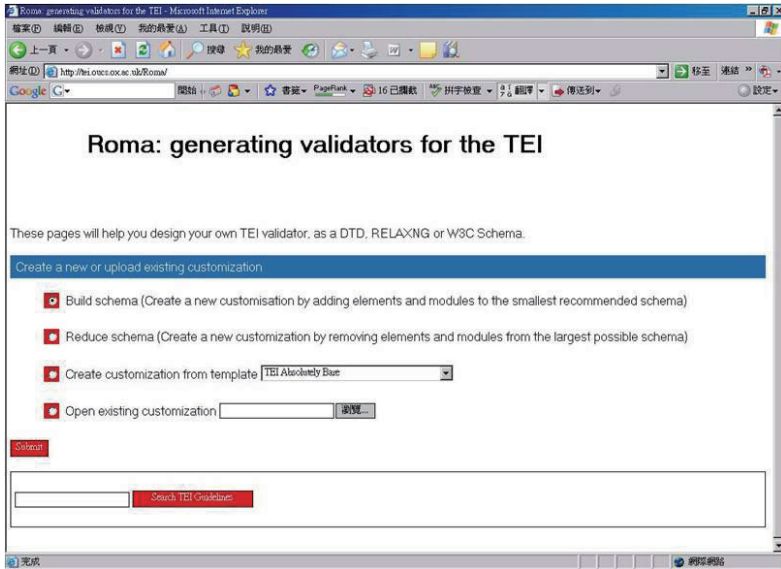


Fig.5-2 Homepage of Roma

Basically ROMA allows users to choose from a list of modules. Modules group elements according to function and genre.



Fig.5-3 Roma Modules Selection Interface

Based on the characteristics and complexity of a text, users only need to click on the link – add next to the module name to add the new module to the list of selected modules. To remove a selected module, users only need to click on the link – remove next to module names in the list of selected modules. These selected modules form a subset of TEI and are also called TEI Schema. To ensure that all customized TEI Schema conform to the TEI standard structure, which is required for international and inter-library exchange and circulation, this system adds the four main TEI structural modules – core, tei, header and textstructure into the list of selected modules, and does not allow users to remove them.

Besides selecting TEI modules, users can also modify element and attribute usage, create custom elements, and restrict attribute values.

TEI Roma: 製作TEI的文件模型檔

更改模組

重新開始 調整設定 語言 模組 新增元素 更改元素集 建立文件模型 建立說明檔 儲存設定檔 Sanity checker

back

以下元素列表所屬模組: textstructure

	包含	不包含	標籤名稱	描述	屬性
TEI	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	tei	(TEI document) contains a single TEI-conformant document, comprising a TEI header and a text, either in isolation or as part of a <code>teiCorpus</code> element.	更改屬性
argument	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	argument	A formal list or prose description of the topics addressed by a subdivision of a text.	更改屬性
back	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	back	(back matter) contains any appendices, etc. following the main part of a text.	更改屬性
body	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	body	(text body) contains the whole body of a single unitary text, excluding any front or back matter.	更改屬性
byline	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	byline	contains the primary statement of responsibility given for a work on its title page or at the head or end of the work.	更改屬性
closer	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	closer	groups together salutations, datelines, and similar phrases appearing as a final group at the end of a division, especially of a letter.	更改屬性
dateline	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	dateline	contains a brief description of the place, date, time, etc. of production of a letter, newspaper story, or other work, prefixed or suffixed to it as a kind of heading or trailer.	更改屬性
div	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	div	(text division) contains a subdivision of the front, body, or back of a text.	更改屬性
div1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	div1	(level-1 text division) contains a first-level	更改屬性

Fig.5-4 Tag set of the textstructure module

After selecting modules and defining the elements they contain, users download their TEI conformant customized validators the encoders are going to use when they validate their markup during the markup process. Currently ROMA offers to create validators in Relax NG (compact and XML syntax), W3C XML Schema, Schematron, and the legacy DTD.



Fig.5-5 Page for creating a TEI Schema



Fig.5-6 Page for creating TEI Documentation

(III) Markup Personnel

Markup is divided into hierarchal markup and content markup, the former marks the structure of text, such as title, author, paragraphs, lines.....etc., which are relatively standardized and do not require professional knowledge, so general key-in personnel should be able to handle them during text input.

The later involves determination and recognition of contents, such as missing characters, character variants, taboo characters or special annotations, and must be handled by a professional (e.g. project member or trained personnel) using a tool book, such as a dictionary.

(IV) Software

1. Notepad++

Notepad++ (<http://notepad-plus-plus.org/>) is a powerful, free open-source text editor. After installing the XML-Tools plugin it serves as a

convenient XML editor. Typically a text editor should:

- (1) Provide powerful text editing, regular expression search and replace, preview and print functions.
- (2) Provide hexadecimal editing functions.
- (3) Syntax highlighting for different languages.
- (4) Indent and parse XML, perform simple XSLT transforms. Oxygen

Oxygen (<http://www.oxygenxml.com/index.html>) is a full-fledged, dedicated commercial XML editor that allows programmers and encoders to make use of the full spectrum of XML-related technologies (Schema, XSLT, XQuery, XInclude, connection to databases, different views etc.). Oxygen ships with out of the box TEI support.

(V) Convert the TEI-XML to other formats for presentation.

TEI/XML is a master format for digital text. For presentation TEI data must be transformed into other formats, such as HTML, PDF, ODT, e-Pub, or even DOCX. Transformations from XML are often done with XSLT but can be achieved through most general purpose programming languages.



Six. Database and Other Applications

After documents are transformed into electronic full-text, they can be displayed and modified in various ways. Digitization frees documents from the constraints of print, and inspires new possibilities of knowledge production and analysis.

I. Database Establishment

Full-text databases are a basic application of electronic full-text and serve a similar purpose as digital libraries, which put entire collections on the A database should have the following functions:

(I) Full-Text Retrieval

Typical library or database text retrieval functions stop at title, author and key words. This approach uses books to search for text, but does not consider the need to use text to search for books. Full-text retrieval allows users to find books and authors when they only know a part of the text, and list documents with similar text for comparison, analysis or statistics.

(II) Hierarchal Retrieval

Some documents are part of a hierarchal relationship that is not displayed in the results of searches by title or full text. Database design and interface should show users the position of retrieved documents in their original structures.

(III) Authority Control

Authority control is mainly used for names, places, agencies and themes, enabling the disambiguation of items with different names or spellings.

(IV) Image Linkage

Adding a link to the source image file in the full-text will turn electronic full-text databases into electronic full-text and image databases, allowing users to see how the original book was arranged while reading its contents.

II. Produce Optical Disks

Optical disks serve two functions – preservation and circulation. For information that needs to be promoted and disseminated, optical disks offer the advantages of low cost, highly portable and easy to read, making them an indispensable tool independent from the internet. Among domestic full-

text digitization projects, the CBETA releases the latest version of Buddhist electronic text on optical disks (Fig.6-1) on an annual basis, making them available to even more people.



Fig.6-1 CBETA CDs that Contain Buddhist Electronic Texts

III. Develop Related Encyclopedias and Dictionaries

In the process of full-text digitization, missing characters, character variants, markup, and authority control entries and vocabulary recorded during proofreading can be developed into an online tool for providing information and knowledge.



Seven. Digital Rights Management

In this information age, the boundless internet makes digital resources easy to obtain, but also puts precious treasures and private assets at risk of illegal access. Although full-text is not as prone to being infringed for extravagant profits as digital images, quoting the words of others without indicating the source is still an act of plagiarism. To protect the copyright or intellectual property right of electronic full-text, common digital content protection methods are introduced below, and protection plans of several electronic full-text databases are also provided for your consideration.

I. Creative Commons Licensing

Creative Commons licenses are copyright right licenses that originated in the U.S. to publicly release a portion of access rights of works. In other words such licenses provide a simple and legally effective way for creators to declare the scope of which digital content may be freely used, while retaining some rights. This allows works to be freely accessed by users around the world while protecting certain rights that the copyright owner hopes to retain.

Creative Commons licenses published in 2002 were designed based on related laws of the U.S. Therefore, Creative Commons licenses for digital content must be adapted to laws of each state and region.¹⁴

As to actual licensing procedures, creators can release rights that they deem fit through a series of options, including “whether if the work must be attributed,” “whether if the work may be used for commercial purposes,” “whether if derivative works are allowed” and “whether if derivative works must be shared under the same conditions.” A variety of simple icons (Fig. 7-1) allow users to easily identify the type of license used, benefitting the circulation of the work.¹⁵

Creative Commons has four licensing terms:¹⁶

1. Attribution



¹⁴“Creative Commons Taiwan,” Search: September 2010, <http://creativecommons.org.tw/>.

¹⁵ Creative Commons Taiwan, <http://creativecommons.org.tw/>.

¹⁶“CC-Creative Commons Taiwan,” Search: December 2010, <http://creativecommons.org.tw/>.

This license lets others distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation. There are the following notices as to how to attribute a work:

If the work provides the following information, then the users must record the information in verbatim on each copy in a suitable form on the media or tool used.

- (1) Retain the complete copyright notice of the work;
- (2) Indicate the name or pen-name of the creator or licensor, or a third person designated by the creator or licensor in the copyright notice or licensing terms.
- (3) The name of the work is indicated;
- (4) The address of a work provided by the copyright owner or licensor.

2. No Derivative Works



This license allows for redistribution, commercial and non-commercial, as long as it is passed along unchanged and in whole.

3. Non-Commercial



This license lets others remix, tweak, and build upon your work non-commercially.

4. Share Alike



This license lets others remix, tweak, and build upon your work as long as they license their new creations under the identical terms.

At present, there are a number of Creative Commons licenses for copyright owners to choose from. In the case of Creative Commons

Taiwan, if a copyright owner chooses the license “Attribution-NonCommercial-NoDerivs 2.5 Taiwan,” then the user must specify the copyright owner of the original work to use it.

Unlike the traditional “all rights reserved” setting of most websites, which creates the issue of copyright infringement without legal licensing or reasonable use, Creative Commons offers a “some rights reserved” approach to copyright, allowing anyone to use digital content according to licensing terms. A portion of project results displayed on the portal of TELDAP is licensed using Creative Commons licenses. To license a work, visit Creative Commons Taiwan and follow instructions to choose a license, and then add the system generated code to your website to complete licensing.

*Creative Commons website: <http://creativecommons.org.tw/>

*Introduction to Creative Commons: http://creativecommons.org.tw/cc_intro_anime

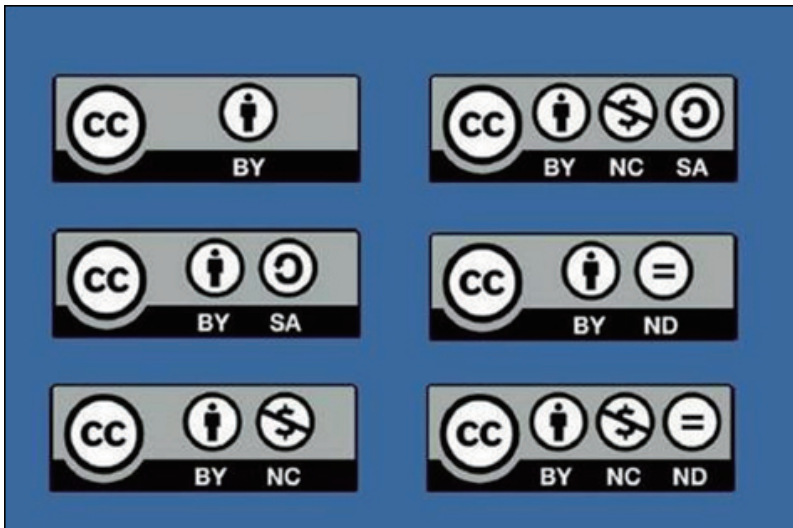


Fig.7-1 Six Licenses

II. Digital Rights Managements (DRM)

Digital Rights Management (DRM) is a digital content protection mechanism that integrates hardware with software for encrypting files. The Internet Data Center (IDC) defines DRM as “The chain of hardware and software services and technologies confining the use of digital content to authorized use and users and managing any consequences of that use throughout the entire life cycle of the content. DRM is one kind of content protection technology.”¹⁷

After digital content is created, DRM sets access rights to protect the rights of the creator and provider, e.g. the number of times the file can be read, saved, copied, forwarded and burned, the number of times the file can be played, whether if the file can be copied or printed, expiration date of the file, whether if a specific password is required to open the file, and if the file is read only. Sometimes DRM technology requires users to use specific software or hardware to open a file. Using the digital archives system of the Multimedia Center as an example, which was developed by the digital archiving technology development team of the Institute for Information Science, Academia Sinica, users must first acquire an account to login. When users click on a file they wish to download, the system will notify the user of the file’s access rights and authentication method. After completing the authentication process, users will be able to play the digital content in their computers, at which time the system will notify users of their access rights (e.g. number of times the file can be played, or if save as or print functions are denied).

III. Access Rights Restrictions

(I) Restricting User IP Addresses

Some libraries or museums manage the circulation of their resources by charging for access or restricting user IP addresses. This method is usually adopted when the library or museum holds a unique or precious collection, or the digital content involves copyright and licensing issues.

(II) Free Access

In contrast to digital content protection measures, such as user pays

¹⁷ Dahl Joshua and Kevorkian Susan, “Understanding DRM Systems”, An IDC Research White Paper, 2001.

and restricting user IP addresses, some institutes allow free access to their digital content based on the spirit of knowledge sharing and serving the public. Although users may freely browse and access full-text free of charge, they are required to specify the source when quoting or citing content, so as to show respect to the content provider.



Eight. Equipment and Cost Analysis

Digitization is an immense and costly endeavor that not only includes the tangible, such as equipment and software/hardware costs, workroom, utilities and maintenance, but also the intangible, such as personnel training cost, and knowledge and time of researchers for text segmentation.

In the process of full-text digitization, cost of intangible knowledge is hard to calculate. Therefore, this guideline only considers labor cost, equipment cost and outsourcing cost, providing options and possible costs associated with each option.

I. Equipment Selection Considerations

In this section we will explain considerations for selecting equipment required for full-text digitization work. Main equipment includes scanners and computer software and hardware.

(I) Scanner Selection

Considerations of scanner selection include the condition of physical objects and digital image specifications. Scanners currently in the market include desktop platform scanners, desktop automatic sheetfed scanners, desktop seamless scanners, and drum scanners. Some Chinese classics might be damaged as a result of their long history, so besides considering the functions of digitization equipment, project units should also consider the fragility of objects, and adopt digitization equipment and methods that cause the least harm to objects. If project units intend to implement full-text digitization themselves, we recommend using sheetfed scanners with automatic numbering and file saving functions, so as to save scanning time and simplify scanning work.

(II) Computer System Selection

Computer systems are divided into software and hardware.

1. Computer Hardware Selection

There are two types of computers in the market – the PC, which is for commercial purpose or general word processing, and the MAC, which is for graphics processing and publications. Since full-text digitization mainly involves word processing, e.g. input, proofreading and markup, a PC will suffice for digitization operations.

When selecting a PC, although most digitization work stops at word

processing, numerous windows might be operating at the same time, or personnel will need to browse through scanned image files, so individual components still need to be specially selected, for example:

- (1) Generally speaking, the more RAM the better.
- (2) High-end graphics cards are recommended for projects that require image processing to display more details. Low-end graphics cards or built-in graphics chips on motherboards generally suffice for word processing.
- (3) Hard disks used for storing files are the larger the better.

2. Computer Software Selection

Software required for different operations is introduced below.

- (1) Digital Image Processing: Adobe Photoshop is often used for image editing and processing (<http://www.adobe.com/tw/products/photoshop/>). However, the free open source tool GIMP (<http://www.gimp.org/>) is well sufficient for most tasks and should be used first on all computers, so institutions can limit the number of licenses they have to take out for Photoshop. “Free, open source tools first, commercial solutions later” is a general strategy to reduce costs and foster intelligent use of resources in a project.
- (2) Key-in: Generally speaking, any text editor can be used. We recommend Notepad++, or Hanshu wenshu shuli xitong 漢書文書處理系統,” a domestic product.
- (3) OCR: Please refer to “Journals and Newspapers Digitization Procedures Guideline” for an introduction to OCR systems, which includes a comparison of several brands. Most projects that are involved in full-text digitization recommend using “丹青文件辨識系統,” which can recognize Chinese, English and Japanese characters at the same time.
- (4) Markup software: For the markup stage encoders need a tool that does three things well: parses for well-formedness, auto-validates the markup against a schema on entry, and allows for transformations into other formats. There are dozens of adequate free XML editors available. The interface should be localized in a language the encoders are familiar with. XML Copy Editor or

Exchanger XML Lite are workable cross-platform solutions. For a full-fledged XML programming environment we recommend the Oxygen XML Editor which is reasonably priced and affords high capabilities.

- (5) Other: The main work of full-text digitization consists of input, proofreading and markup, and thus requires extremely high labor cost. Some projects or units will develop system software, such as image proofreading, file comparison, word processing, markup, file conversion...etc., to simplify work and improve work efficiency. This software can either be developed by the project unit's own information department or outsourced. However, if software development is outsourced, it should be in open source code for the convenience of future modification.

II. Cost Analysis

(I) Cost Components

Costs associated with digitization include three components: Material, labor and miscellaneous costs:

1. Material cost is the cost of consumables used for digitization work.
2. Labor cost is mainly the salaries of personnel.
3. Miscellaneous costs can be divided into direct costs and indirect costs:
 - (1) Direct costs include cost and amortization of equipment and cost of information software.
 - (2) Indirect costs include workroom renovation, rent, utilities and other.

This cost analysis provides a way to calculate cost, equipment selection and manpower allocation should be based on the project unit's budget constraints or purpose of digitization results. If digital content will be licensed in the future, then the estimated cost can serve as a basis for calculating licensing fee.

(II) Cost Estimation

1. Calculation Method:

There are two ways to calculate cost based on amortization of

equipment:

- (1) Calculating amortization of equipment based on service life
$$\frac{[\text{Labor cost (NTD)} + \text{Amortization of equipment (NTD)}]}{\text{Digital output (Number of images)}} = \text{Cost per image (NTD/Image)}$$
 - A. Labor cost is mainly salaries of personnel
 - B. Amortization of equipment = $(\text{Equipment cost} + \text{Software cost} - \text{Remaining Value}) / \text{Service life}$
- (2) Calculating amortization of equipment based on digital output
$$\frac{[\text{Labor cost (NTD)} / \text{Digital output (Number of images)}] + \{[\text{Equipment cost (NTD)} + \text{Software cost (NTD)}] / \text{Digital output (Number of images)}\}}{\text{Digital output (Number of images)}} = \text{Cost per image (NTD/Image)}$$

Furthermore, regular backup and replacement of storage media is another cost associated with the long-term preservation of digital content, and there is also the risk of data being damaged and human resource management issues that should be considered.



Nine. Outsourcing

The cost of equipment, manpower and time required for digitization work often affects the quality of digitization results. Project units should consider whether or not to outsource all or some digitization procedures based on actual work conditions. Outsourcing is defined as “to hand over a certain portion of an enterprise’s internal management functions along with associated assets to an external supplier or service provider for a certain period of time at a price agreed by both parties (proviso included).”¹⁸ From a management perspective, outsourcing is a viable option for digital archiving, provided that existing labor cost is not increased and good management operations are implemented.

In the case of Chinese full-text database digitization work, digitization of large quantities of books and paper documents involve scanning, key in and input. Quite a few project units have selected to outsource digitization work under efficiency and cost considerations. In the following section we will briefly introduce outsourcing procedures of the Chinese Classics Full-Text Database Team of the Institute of History and Philology, Academia Sinica.

I. Outsourcing Copying and Scanning

(I) Establish Copying and Scanning Standards:

Copies are provided to the contractor for key-in purposes; decide on the ratio based on the size and clearness of characters.

Images should be scanned according to “Fu Ssu Nien Library Standard Operating Procedures for Full Color Image Scanning and Assessment” on a ratio of 1:1, 300dpi, full color (black and white), and in TIFF format.

(II) Retrieve Books:

Retrieve books on the booklist for the contractor to print and scan. At present, there are three sources of books: 1. Purchased books; 2. Books provided by researchers; 3. Books collected by Academia Sinica.

If there are books on the booklist that are not allowed to be borrowed according to regulations of the repository, then reselect books for digitization.

¹⁸ “Digitization Procedures Guideline: Outsourcing Management” by Kao Chih-Tung, Chen Hsiu-Hua, Chen Mei-Chih and Lin Fang-Chih, Taipei: Taiwan Digital Archives Expansion Project, March 2009.

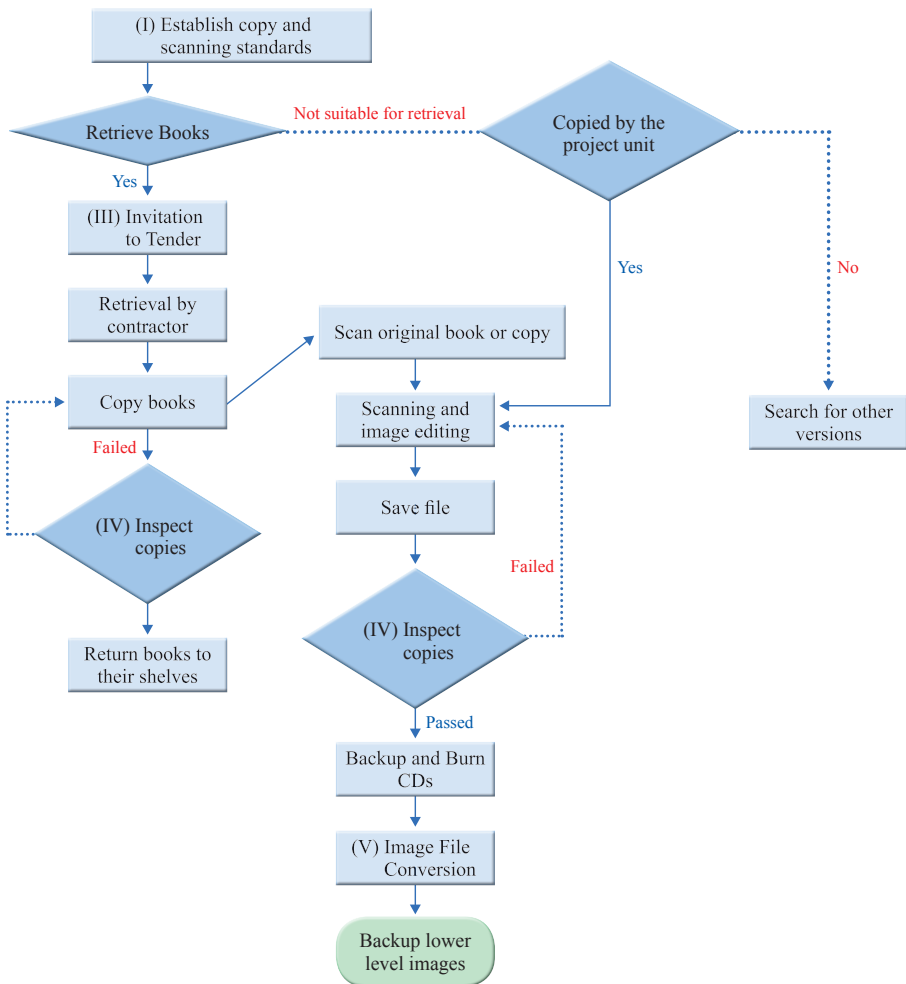


Fig.9-1 Copying and Scanning Outsourcing Flowchart Compiled by the Chinese Classics Full-Text Database Team

(III) Invitation to Tender:

Following tender regulations of the project units.

(IV) Assessment:

After copying documents, the contractor should send the copies and originals back to the project unit for assessment. After scanning documents,

the contractor must return the files to the workroom for assessment, and go on to image file conversion once the files pass the assessment.

(V) Image File Conversion:

Convert image files into lower level formats according to the full color image scanning and assessment operation standards. These files will be used for subsequent procedures. Images in lower level formats should also be backed up.

II. Outsourcing Key-in and Initial Markup

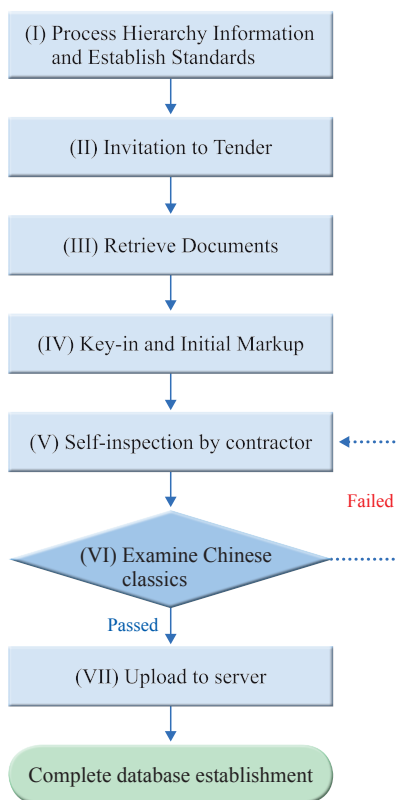


Fig.9-2 Key-in and Initial Markup Outsourcing Flowchart Compiled by the Chinese Classics Full-Text Database Team

(I) Process Hierarchy Information and Establish Standards:

This is a task for outsourcing; number the files and construct the basic hierarchical structure according to book style and requirements of researchers, provide descriptions and examples for input, and establish standards for the operating environment and restrictions.

(II) Invitation to Tender:

Decide on the number of characters to input based on the annual budget, and then invite contractors to bid according to the Government Procurement Act.

(III) Retrieve Documents:

After a contractor wins the bid, the contractor should retrieve documents from the workroom. Personnel at the workroom should explain notices of input to the contractor; provide a XML tagging program and “missing character utility,” demonstrate how to operate the programs, and provide information for the contractor to understand how to install and use the programs.

(IV) Key-in and Initial Markup:

1. Key-in: Input text according to Chinese book key-in principles (according to the original pattern in the book, unclear parts are not guessed or processed).
2. Initial Markup: Use text editor, such as Notepad++, for an initial markup of the document’s “level.”

(V) Inspection by the Contractor:

Files should only be returned if error rate is under 0.01%.

(VI) Sampling Inspection:

Personnel should randomly inspect files that are returned and upload them to the server for backup if they reach standards (error rate under 0.01%). If they do not reach standards, they should be rejected and the contractor should arrange for a second inspection after making corrections. According to assessment procedures of the Government Procurement Act, related units should send personnel to the work room to inspect digital files; if files do not meet requirements, require the contractor to correct errors within a deadline for a second inspection.

(VII) Establish a First Proofreading Database:

Upload the files and establish a first proofreading database for access by other personnel.

III. Outsourcing First and Second Proofreading

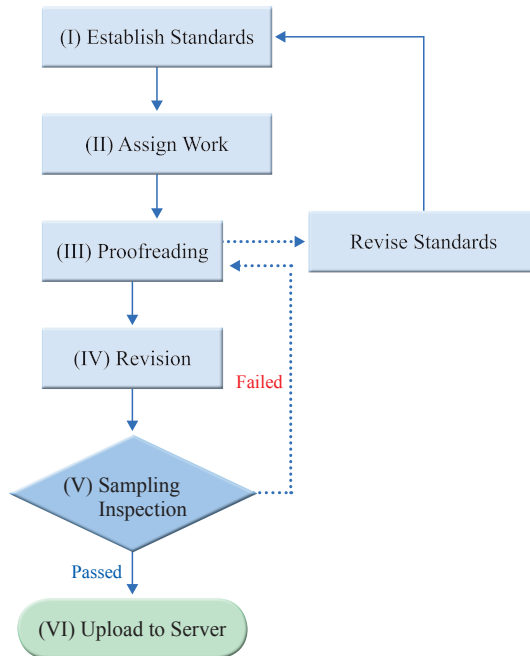


Fig.9-3 Proofreading Outsourcing Flowchart Compiled by the Chinese Classics Full-Text Database Team

(I) Establish Standards:

1. Establish rules for finding character variants, wrong characters, missing characters and taboo characters.
2. Establish sampling inspection standards.

(II) Work Assignment:

1. Evenly assign work according to the total quantity.
2. Assign work based on difficulty to personnel with the required

expertise.

(III) Proofreading:

Proofread the image file character by character and adjust standards for differences of each book based on actual conditions.

(IV) Revision:

Correct any mistakes that were identified.

(V) Sampling Inspection:

Sample the same number of characters from each book according to the standards previously established for personnel other than those responsible for proofreading to inspect. Proofread the book again if error rate exceeds the standard. If the book is too difficult to read (e.g. written in cursive style, running script or characters are hard to recognize), suitable proofreading personnel must be found, and use tool books or other versions of the book for proofreading.

(VI) Upload Files:

Upload the files to the server for backup.

The procedures above are the most common when outsourcing full-text digitization. However, since digitization is a long-term task, project units should consider their budget planning and tender issues when deciding whether or not to outsource digitization operations. Sometimes project units can implement digitization more efficiently and at lower cost. Therefore, when project units intend to use outsourcing, they should consider how work should be planned, establish specifications, select a suitable contractor, communicate and negotiate with the contractor, and assess digitization results.

For more details on outsourcing procedures, also see “Digitization Procedures Guideline: Outsourcing Management.”



Ten. Conclusions

In the past, Taiwan was a global leader in the field of Chinese classics full-text digitization in both quantity and quality. However, Mainland China is now gradually carrying out full-text digitization as well, and is producing a large amount of electronic full-text. In the light of this, domestic institutions that have long been dedicated to full-text digitization continue to invest a great amount of resources into producing high quality, high standard full-text databases. This “Chinese Classics Full-Text Database Digitization Procedures Guideline” focuses on metadata of document tags. As explained in previous chapters, marking up documents is the best way to increase the research value and application scope of electronic full-text. Also, among current markup systems, TEI is the most highly recommended by scholars, libraries and museums. We hope that all full-text digitization projects, ongoing or still being planned, will include TEI into their standard operating procedures, and benefit the future utilization of digitization results.

With the help of Mr. Tu Cheng-Min, convener of the Chinese Classics Full-Text Database Sub-group, TELDAP Taiwan Digital Archives Expansion Project has been devoted to the sharing of Chinese classics full-text digitization research and development results. Up to now the sub-group has offered numerous TEI workshops to training markup talents in Taiwan. Besides continuing to localize and translate TEI into Chinese, the sub-group also hopes to offer missing character related courses or lectures, as well as international and domestic conferences on full-text digitization research and technologies, making Taiwan’s Chinese classics full-text digitization environment better and more competitive.

Finally, this guideline would not have been possible without the experiences of pioneers in full-text digitization, who so generously shared them with us. Here we would like to specially thank Mr. Tu Cheng-Min, Convener of the Chinese Classics Full-Text Database Sub-group and also Vice President of the Dharma Drum Buddhist College, Director Wu Pao-Yuen of CBETA, and fellow colleagues of the Chinese Classics Full-Text Database Team of the Institute of History and Philology, Academia Sinica. We truly hope that this guideline, which contains full-text digitization experiences of the top projects in Taiwan, will attract more project units or personnel to enter the field of full-text digitization and successfully complete their work.



References

Books

王雅萍、林彥宏，《文書檔案數位化工作流程指南》，台北市：數位典藏拓展臺灣數位典藏計畫，2009年04月。

高朗軒、陳秀華，《書畫數位化工作流程指南》，台北市：數位典藏拓展臺灣數位典藏計畫，2009年04月。

李佩瑛、程婉如，《期刊報紙數位化工作流程指南》，台北市：數位典藏拓展臺灣數位典藏計畫，2009年04月。

高芷彤、林芳志、陳秀華、陳美智，《數位化工作流程指南：委外製作》，台北市：數位典藏拓展臺灣數位典藏計畫。2009年，4月。

馬德偉，《TEI 使用指南——運用 TEI 處理中文文獻》，台北市：數位典藏與數位學習國家型科技計畫。2009年。

JournalPapers

謝清俊、林晰，〈中央研究院古籍全文資料庫的發展概要〉，中央研究院資訊科學研究所文獻處理實驗室，1997年3月。

香光尼眾佛學院圖書館，〈佛教資料電子化（一）〉，《佛教圖書館館訊》，第十四期，1998年6月。

莊德明，〈漢字缺字處理與梵巴藏字母的輸入〉，《佛教圖書館館訊》，第十四期，1998年6月。

香光尼眾佛學院圖書館，〈佛教資料電子化（二）〉，《佛教圖書館館訊》，第十五期，1998年9月。

黃鴻珠，〈「觀前顧後」資料電子化的要訣之一〉，《佛教圖書館館訊》，第十五期，1998年9月。

香光尼眾佛學院圖書館，〈「佛教資料電子化研討會」實錄〉，《佛教圖書館館訊》，第十八/十九，1999年9月。

香光尼眾佛學院圖書館，〈電子佛典製作〉，《佛教圖書館館訊》，第二十四期，2000年12月。

香光尼眾佛學院圖書館，〈佛教知識組織管理研討會（二）〉，《佛教圖書館館訊漢籍全文數位化工作流程指南 91 館訊》，第三十二期，2002年12月。

中央研究院資訊科學研究所文獻處理實驗室，《漢字構形資料庫使用手冊》，台北市：中央研究院資訊科學研究所，2002年。

香光尼眾佛學院圖書館，〈佛教工具書編輯〉，《佛教圖書館館訊》，第三十五/三十六期，2003年12月。

香光尼眾佛學院圖書館，〈佛教文獻檢索與利用（二）〉，《佛教圖書館館訊》，第四十期，2004年12月。

數位典藏國家型科技計畫，《國家數位典藏通訊》，第三卷第八期，2004年。

國立臺灣大學典藏數位化計畫，《台灣文獻文物典藏數位化計畫》，2006年11月20日，<<http://140.112.113.4/project/default.asp>>。

莊德明，〈漢字資訊化的困境及因應：談如何建立漢字知識庫〉，第二屆漢字文化節學術研討會，2006年5月。

洪振洲、馬德偉、張伯雍、志賢、黃仁順，〈佛教位典藏與GIS技術應用經驗分享〉，Proceedings of International Conference of Digital Archives and Digital Humanities, Dec 1-2, 2009, pp. 141-166.

洪振洲、馬德偉 (Marcus BINGENHEIMER)、許智偉，〈台灣佛教文化數位典藏之發展 - Digital Archives for the Study of Taiwanese Buddhism.〉，2011，Vol. 49:1 (49卷1期)，pp. 103-133.

Dahl Joshua and Kevorkian Susan, "Understanding DRM Systems", An IDCResearch White Paper, 2001.

Internet Resources

數位典藏與數位學習國家型科技計畫，《技術彙編》2007年版。<http://www2.ndap.org.tw/eBook08/index.html>

教育部「異體字字典」網站。<http://140.111.1.40/>

中央研究院資訊科學研究所文獻處理實驗室。<http://www.sinica.edu.tw/~cdp/>
創用CC網站。<http://creativecommons.org.tw/>

The Unicode Consortium。<http://www.unicode.org/charts/unihanrsindex.html>

TEI (Text Encoding Initiative)。<http://www.tei-c.org/>

Notepad++。<http://notepad-plus-plus.org/>

Oxygen xml editor。<http://www.oxygenxml.com/index.html>

Chinese Classics Full-Text Database Digitization Procedures Guideline

Advisory Unit: National Science Council, Executive Yuan

Issuer: Simon C. Lin

Editor-in-Chief: Simon C. Lin

Executive Editors: Meng-Yin Wu, Yu-Ju Lin

Authors: Ya-Ping Wang, Hsiao-Lin Hsieh

Translator: Tai-Yu Chen

Reviewer: Marcus Bingenheimer

Publisher: International Collaboration and Promotion of Taiwan e-Learning
and Digital Archives Program

Address: No.128, Sec.2, Academia Rd., Nankang District, Taipei City, 115
Institute of Physics, Academia Sinica

Tel: +886-2-2789-8311

Fax: +886-2-2783-5434

Website: <http://collab.teldap.tw>

Email: teldap-collab@twgrid.org

Typesetting and Design: EVERGREEN INTERNATIONAL CORP.

All Rights Reserved, Not for Sale

本書譯自拓展臺灣數位典藏計畫出版之
數位化工作流程指南：漢籍全文