CADAL and the Literature of Republic of China

Meng Zhan[1] YanQing Xiao[2]

[1]Wuhan University Library, Wuhan Hubei, China 430072

[2]School of Library and Information Science, Wuhan University, Wuhan Hubei, China 430072

Abstract — **This paper introduce the CADAL project and the plan of establishing the digital resources of the literature of Republic of China; analysis in detail the characteristic of the literature of Republic of China; description the methods of process management and quality control in the process of digitized literature of Republic of China.**

Keyword — **CADAL Literature of Republic of China, Digital preservation**

I. ABOUT CADAL

CADAL (China-America Digital Academic Library) project is a collaborative project between computer scientists in China and the USA. The goal of this project is to establish digital resources of millions books, which will be use in the field of education and scientific research and supply strong support to teaching and scientific research in higher education. This project can promote the share of digital resources between China and USA [1]. The project is funded by both China and the United States, the US partner has invested equivalent of 10 million U.S. dollars of software and hardware systems to support the project [2]. In September 2002, Chinese State Development Planning Commission, Ministry of Education and Ministry of Finance issued the document of "Certain opinions on the strengthening of the "211 Project" in the "10th Five-Year Plan"". It has listed CADAL to be an important component of the public service system of "211 Project" and invested 80 million RMB to support this project. Throughout the whole process of CADAL project, all the libraries that have participated in this project have invested the equipments and human resources to support the project. Actually the total investment cost both of China government and libraries is more than 100 million RMB.

In CADAL project, the type of resources includes rare book, literature of Republic of China, modern Chinese books, Chinese journals, English books, dissertations and multimedia resources, etc. The project plans to construct a more comprehensive full-text database about Literature of Republic of China with total copies of more than one hundred thousands [1]. Until now, it has digitized more than 100 thousands books and more than 110 thousands journal issues of Republic of China, these literature are the collection in mainly university libraries, such as Peking University, Tsinghua University, Fudan University, Wuhan University, Nanjing University, Jilin University, Zhejiang University etc. All of these universities are the key universities in the historical period of Republic of China, and therefore their collection has significant academic value.

II. ABOUT THE LITERATURE OF REPUBLIC OF CHINA

A. Republic of China and its Literature

The times of Republic of China is from 1911 Xinhai Revolution to 1949 the foundation of People's Republic of China, lasting 38 years. It is a quite special historical era and the most complicated period in Chinese history. There were the fiercest collision between the Chinese and west civilization in this stage, for example the Chinese culture and the West culture exchanged; new ideology challenged the old, etc. Therefore a number of academic works and historical resources had been generated [4]. On the other hand, the publication industry also appeared prosperity, he new material and technology had been used in the published filed; the publishing house, newspaper office and libraries had been set up rapidly. As the reason of above the documents that published in this historical times not only had higher value in history and literature[3], but also created some innovative improvements in the history of printing, for example papermaking technology, printing technology and bounding technology etc. It is reported that more than 100,000 books had been published in this period[4]. In addition, there were a great deal of journals; newspapers, magazine, brochures, all political parties publications, government documents and civil publications emerged. These literatures are the part of Chinese culture.

According to the "Total Bibliography for the Republic of China", the literature published between 1911 and 1949 belong to the literature of Republic of China. From this definition we can see that the publish time is the only proviso of defining the literature of Republic of China. As a matter of fact, we have verified the uniqueness of the proviso in the practice of CADAL. Because the times of Republic of China is a transition period from ancient China to modern China, so the literature of in this time have more mixed situation in papermaking, version and bounding, for example the paper including "rice paper" that made in ancient traditional handwork and the paper

that with higher acidity made by new technique; it has both manuscripts and printed ones; in binding mode, it has both traditional thread binding and modern machine binding; in the literature contents, it has both classical Chinese and colloquial words. The mixture features of classical literature and modern literature is the very historical mark of the transitional era when the Chinese society came into the modern times from the ancient times.

B. Characteristics of the Literature of Republic of China

During the construction of CADAL project, we have widely encountered various types of literature of Republic of China and also have collected and organized all this kind of literature, they are include books, journals, and thesis. These resources are stored in the library, school reference rooms and archives of Wuhan University. In the process of digitizing the literature, we found that the literature of Republic of China has some characteristics that are greatly different from the ancient books and modern literature, mainly manifested as below:

B.1. Characteristics of the paper

The paper of the literature of Republic of China is always appearing yellow color, hard and fragile. We have digitized more than 17,000 volumes of books and journals, after finished the digitization there are 70% of which needs repairing. Compared with the ancient books, which stored in the same environment, the breakage rate of the literature of Republic of China is much higher. What causes this situation? According to the research results of National Library of China's project "Investigation and analysis of the acidity and preservation situation in paper quality of collection", the paper material in that era is usually mechanical groundwood paper and acidic chemosynthesis paper. This kind of paper is of strong acidity poor quality and short storage period[6] , so the paper of literature of Republic of China is hard and fragile, even the slightest touch will break it . But ancient books paper were made up by phloem or fibre and used the ancient traditional handcrafts. This kind of paper is usually neutral or alkaline, so it has the proverb of "paper long-lived millennium". For example, books in Song and Yuan Dynasties are nearly a thousand years from now, while taking out a rare book you can still read it by one page to another[7]. This research result shows that the keep time of newspaper of the Republic of China is 50 to 100 years, and the book's is about 100 to 200 years. According to the above results, the newspaper in this era is approaching naturally disappearing.

B.2. Characteristics of binding and layout

In binding mode, the ancient books are all in form of traditional thread binding modern books and journals all in mechanical binding, but in the Republic of China, western mechanical papermaking technology and printing art began to be widely used in China, so book binding began to transit from traditional thread binding to mechanical binding. In the early stage of Republic of China, literature written on the Xuan paper (rice paper) maintained the traditional handcrafts, which are called thread-bound edition of Republic of China. In the later stage, nearly all the literature used mechanical binding, so the coexistence of both binding modes is a special feature of the literature in that period. The pages of the literature at that time are easy to loosen when viewed, because in the early stage of the Republic of China, the binding of those documents is in low quality as a result of rough mechanical binding technology and cheesy material used for bookbinding.

In the digitized scanning, although we have used non-contact scanning equipment to protect the literature, there are still 70% of them been damaged to different extent. The damage rate of scattered sheets reaching up to 20%, with 6 books couldn't be repaired at all. According to our survey, other participate libraries of which in the CADAL project have the similar problems.

B.3. Characteristics of publishing

Document is the carrier of thought, culture and events in every historical period. In the special period of Republic of China, it had been in an unparalleled developing period in document publishing. The document has following characteristics in publishing:

The first is diversification of publishing forms. There are both documents issued by formal publishing houses, such as bookstores and printing houses, and documents issued by informal ones, such as books issued by groups, political parties and governments in border regions, separate edition written by famous people, privately printed books and handwritten copies. These informal publications account for a quite large proportion.

The second is diversification of version forms. In Republic of China, a variety of publishing technology were used to publications, such as mimeographing, letterpress printing, handwritten, block printing, lithographic printing and copperplate printing, etc. The diversification of version forms is predominant in the early stage of Republic of China. In the later stage, most of the publications are letterpress printed.

B.4. Characteristics of content

Republic of China is a special era, in which new systems replaced the old, Chinese and west cultures exchanged, lots of collision and creation occurred in ideological and cultural fields. It is another contention of numerous schools of thoughts since the period of the Spring and Autumn Annals of the warring States. Books of different political stand, academic viewpoints and ideological inclination are published and disseminated in various ways. The content has both maintenance of

feudal domination and promotion of democratic revolution; both proposition of new culture reform and maintenance of ancient cultural tradition. The most typical is the New Culture Movement advocated by Lu Xun, thought of new-democracy advocated by Sun zhongshan, communist ideology advocated by the revolutionaries represented by Mao Zedong. Books with these progressive thinking were widely spread in society, brought ideological inspiration to the people. In addition, publications on behalf of Guo Mindang government's political and ideological guidelines, publications on new science and technologies and publications on west culture accounted for a large proportion of the literature at that time. Summing up the above, we can see that the content of literature in the Republic of China is the contention of numerous schools of thoughts, gathering all the thoughts in politics, ideology, culture, science and technology and even in people's social lives from feudal society to socialist society, factually recorded the history in that era.

III.   ANALYSIS OF THE CURRENT SITUATION OF LITERATURE OF THE REPUBLIC OF CHINA (IN THE CHINESE MAINLAND ONLY)

A. Distributing status in Chinese Mainland

The collection of the literature of the Republic of China is widely distributed in the China Mainland, covering archives, libraries, historical materials house, literary history research centers and so on. In the quantity of the collection, The Second Historical Archives of China is the place where the literature is mostly collected[3], followed by National Library of China and those provincial public libraries and archives. University libraries in China also have an abundant collection of this kind of literature. These universities were built before 1949, such as Peking University, Nanjing University, Fudan University, Wuhan University. Besides the literatures published during the time of Republic of China, those universities also preserve the thesis written by students at that time. The thesis, also are informal publication and mostly written by handwritten, are of great importance for historical and academic research.

B. Devastation

The acidification of the literature is so serious that many of those documents have been in the condition of devastation. To make things worse, some even have no intensity of the paper and will scatter at a touch, which prevents readers from reading. A case in point is for the National Library of China, 90% of its 670 thousands collections published in the period of the Republic of China is being destroyed up to medium degree. Among them, 100% collections in the early of the Republic of China are in dilapidation[6]. Other libraries have the similar problems. For example, more than 90%

literatures of all the 160 thousands in the Jilin public library are damaged[8], and at least 60% of all the 700 thousands for Nanjing public library[9], more than 50% of 170 thousands for Chongqing public library[10]. As for the Wuhan public library, more than 50 thousands literatures published in the Republic of China are discarded as useless[11]. It has also been reported that at least half of the literatures in the university libraries that taking part in the CADAL project are broken. So we can come to a conclusion that it is now urgent for us to take all the measures to protect our literatures in the Republic of China.

C. Reading

Not all the literatures of Republic of China are open for the public to read freely as for some historical reasons. Some of them are open for the common people to read as "New Good Edition" or modern literatures; some others are open only for researchers as interior documents. The so-called "new good edition" was brought up by Mr Zhao Wanli who was the director of the department of ancient books at National Library of China in 1953, and the criterion for this new good edition is the collection's value in thought, history and art[12]. Although the literature in accordance with new good edition involves the whole period of the Republic of China, all libraries consider the criterion of new good edition in terms of the thought and the revolution. This leads to only parts of the literature coming to the new good edition reading room for readers. For those libraries that have no new good edition reading rooms also put those literature in the modern literature reading room for publicly reading according to the criterion of the new good edition.

IV.   DIGITIZATION PRACTICE

CADAL considered digitizing the literature of Republic of China as an important task of the project, which can not only protect the content permanently, but also enlarge usage forms of the literature. Fourteen of the total sixteen participate libraries of CADAL have digitized their literature. Under the unified planning of CADAL management center, all participate libraries had made thorough programs on resources selection, standards implementation and quality control of the process, thus explored experience in digital preservation and rational use of the literature. Next, we will make some concrete analysis of the above elements, taking the resources processing in library of Wuhan University as an example.

A. Necessity for digital preservation

A.1. Demand analysis

The demand of digitizing the literature of Republic of China includes content of two aspects that are the digital preservation and patrons' networked use. Our library has about 15 thousands volumes of this kind literature, 85%

of them are open to patrons in the reading room, since there are more humanity subjects in our university, the literature is always of great needs to the faculty and students. But for the reasons of paper and binding, they are too severely damaged to serve the patrons. So we need to do the digitization urgently, when finished, the original books will be locked for closed preservation and the digital copy will be open to the patrons. In addition, according to the current state of our university, the literature of Republic of China is collected not only in the library but also in school reference rooms and archives, so centralized digitization and establishing a networked sharing database will be beneficial to the faculty and students.

A.2. Resource increment

In the library circle, the general understanding of resource increment is improving use value of literature. As for the literature of Republic of China, if it is digitized and served online, which can provide information retrieval based on networked database, patrons can also read the full text online when needed. This mode has largely changed the original one, in which the patrons must go to the library for full text. So it would be more convenient for patrons to use the literature and also the efficiency has been greatly increased. According to the related policies made by CADAL management center, every participated library can share the literature of the entire project. That means our digital collection can provide service to patrons from other participate libraries and our patrons can use their collections too. It can be seen that participation in CADAL can not only increases the use value but also enlarge the use extent of literature of Republic of China.

From the above two aspects, we believe the digitization is imperative.

B. Criterions for resource selection

According to the characteristics of the literature, we established the following selection criterions:

Range of resource selection: all collections of literature of Republic of China stored in the libraries, school reference rooms and archives. The library manages the registration work.

Resource screening criterions: according to the "Main Catalog of Republic of China", all the documents published between 1911 and 1949, whether in any binding form, with strong academic theme or touching upon politics can be regarded as the "New good edition". This is a criteria CADAL doesn't include.

Although the criterions have been formulated, we found it difficult to control in the actual practice. For one thing, the year of book published is clear, but the concrete month is difficult to control for lack of basis. For another, there is a contradiction between the criterions of "New good edition" and truly reflect history.

C. Application of the digital standards

To make sure that the digitization is smoothly carried on, CADAL management center has made a series of standards, such as "Draft Metadata standards for Digital Text", "Books of Republic of China Description Standards ", "Journal Description Standards ", "Draft Digital Text Processing Flow Standards", "Data Submit Standards". These standards are the guidelines of the project, which are set up to regulate and control all links concerning the digitization process. In addition, CADAL adopted the Open Ebook Standards for production and preservation of digital resources. For example, the standard scanning resolution rate is 600 dpi, metadata adopts Dublin Core standards, catalog navigation uses XML document, image files are preserved in TIFF and DjVu formats. CADAL adopted the universal standards, which are much different from other domestic ebook projects, which always need special reader's program and low display precision. It is not only convenient for patron's reading, but also conducive to the long-term preservation and further development and utilization of the digital resources.

CADAL is the first cross-library cooperate project in domestic university libraries, which needs a series of regulation and standards in resource processing, metadata indexing and information retrieval to ensure quality. But in the practice we realized that, the standards made by the management center couldn't be well operated, so we made some revisions to the metadata standards on the basis of following CADAL standard frame to complete this international cooperative project with high quality. For example, we added the field of "essentialness and repetition", "qualifier" etc; refined the specification of Title field (How to do with the series, muti-language books); made specific provision on "space-time scales", "Punctuation". These revisions not only follow the standard frame of CADAL, but also reflect the special nature of the literature of the Republic of China .By doing this, the description of metadata is more practicable and feasible, and at the same time, the quality of the digitalized literature in our library is guaranteed, providing a solid foundation for data exchanging and sharing in future.

D. The control factor in digitizing processes

Quality control and the process management in the digitizing process are of important content for CADAL project quality. To be efficient and high quality to complete the task of CADAL project we have strengthened the control of digitizing process and data quality, and groped the management method that is used in the cooperation between the different departments in library, and between the library and data processing company. This organic coordination divides into working process management and processing quality control:

D.I. Management of Working process

In our library the digitizing process of the literature of Republic of China has divided into these work flow: catalog produce - check multiple copy (compare with other libraries in CADAL group) - resource preparing and send processing - scanning and metadata produce (professional processing company take the responsibility) - checking quality of scanning and catalog - metadata inspection - the quality selective examination – data submitting. We have established the management standard for each work flow, every process with working sheet to control the document processed flow, and we also have established the rule for the number of woke sheet and the number of document. In the digital processing, we can inquire the processing state of each kind of document at any time, can know and grasp the rhythm of processing work, and can make entire processing working specification and efficiently. Standard management of working process not only have guaranteed data quality, but also have guaranteed the safety of document.

D.2. Quality control of digitizing

Scanning quality control: the professional scanning company undertakes the scanning work for our library. In the contract that signs with company, we have put the standards of scanning process and data quality into the contract as an attachment file, so on scanning quality has guarantee in legal. In the beginning of scanning, we were training company's people and done a test with a few documents to evaluate the company's technology. After entering batch processing, company carries out data according to this standard to scan and deliver, the librarian check each page's quality, if have some disqualification data it will be return company to process again. Since being on contract clause and in actual process to control the quality of digitizing, therefore the clarity of the digital document of the Republic of China in our library is very high, even some is clearer than paper document.

Metadata quality control: Because of the metadata quality directly affects search quality of retrieval system, therefore, we emphasize the quality of metadata especially, and have put in plenty of manpower and energy to guarantee the quality of data. First, we have revised the metadata standard of the CADAL in expanding some sub-field and making clear the minutia of major field content, for example making the "keyword" field as essential field of it should given three keywords. The process of describing metadata was divided into 2 steps, first the company draw the metadata from the document according to standard (keywords do not make requirement), second librarian inspect the metadata and add keywords. For control the quality of the metadata we have established the rule for inspecting metadata, except for the requiring of technology, we have set a rule to deduct points for error rate of metadata in the rule. When the deducting points were to reach certain value, we should deduct the processing fee of company. Under the double control in the requiring of technical and economic, the metadata quality has gotten guarantee.

V.CONCLUSION

Implement and the establishment of CADAL project have realized really the effect and purpose that the China-U.S.A. scientist of two countries imagines originally. The accomplishment of this project construction has realized the network spread and the digital preservation for human cultural bequest, and also establishment a sharing platform of the cultural resources between China and U.S.A. It is of important protected for the literature of the Republic of China. Therefore CADAL project not only has the current practical meaning on sharing the global resource, but also has profound historical meaning in protecting human cultural resources. This project will be future generations place focus attention.

Because of the time is very tightly, therefore, there is not enough time to consider the different kind of problem and detail about the project, for example, the collecting range for each type of resources hasn't to plan carefully, so the digital document is absents integrality in each type of resource. Additionally the quality of data processing and metadata describing has been required in standard and criterion but in fact, there were not checking and evaluating the quality of digital resources when each library submit the data. So it caused the different level in entire data on scanning quality and metadata quality. This situation will reflect on the search quality in retrieve system. CADAL project management center also recognized this problem and are carrying out concentration with effective method for metadata quality to rectify, they will plan to make efficient method to control the quality in future.

VI.   REFERENCE:

[1] Chen Haiying. Zhu Haikang, A summarization on the China-America digital academic library. The journal of academic library 2005(1) p3
[2] CADAL project management central website http: //www. Caudal. Net /
[3] Wang Xiangfen. To development and the social demand of the literature of the Republic of China, Library work and study 2005(4) p63
[4] Li Fang. The work of protecting and collecting the literature of the Republic of China. Library construction 2005 (2) p109
[5] Zhao Jihai China-America digital library plan (CADAL)
[6] JianHe. The literature of the Republic of China is faced with to disappear nearly. Beijing Archive 2005 (5) p10.
[7] Xinhua net Jiangsu channel 2006-04-08
[8] In Jilin province There are one hundred and sixty thousands literatures of the Republic of China will damage serious (the retrieval of September 3,2006)

http://www.hsm.com.cn/node2/node116/node1486/node1487/userobject6ai225774.html

[9] Books and newspaper one turn will be broken, the literature of the Republic of China is close to scrap danger (the retrieval of September 3,2006)

http://www.zjcnt.com/Templete/print_detail.php?article_id=55341&article_type

[10] In Chongqing Public library there are half of literature of the Republic of China having destroyed (the retrieval of September 3,2006)

http://www.libnet.sh.cn/yjdd/list.asp?id=1872

[11] In Wuhan public library there are fifty thousands literatures of the Republic of China is close to scrap. (the retrieval of September 3,2006)

http://61.183.175.92:8080/publish/sylm_1/whyw_3/2006-04-1284978.html

[12] Zhao ChangHai, The chart in new good editor book of library. Document, information & knowledge, 2004 (3) p22

Author brief introduction:

Zhan Meng, Wuhan university library, assistant professor, research field: Library computer management. Email: mzhan@lib.whu.edu.cn

Xiao Yanqin, Library school of Wuhan university, master, research field: information management. Email: yqxiao@lib.whu.edu.cn