Practices and Open Problems of Document Digitization For Million Book Project

ZHENG Xiao-Hui

Tsinghua University Library, Beijing 100084, China

Abstract: **Tsinghua University Library participated in the China-America Digital Academic Library Project at the end of 2002 and finished 50000 E-books and Dissertations in three years. The digitization Center was established in March of 2003 to accommodate the retro-digitization trend. Three years practices shaped a semi-automation producing line, at the meanwhile, some issues were discussed for second phase plan consideration.**

Key words: **Digitization, workflow**

1. INTRODUCTION

Tsinghua University Library participated in the first phase CADAL Project (Another name China-US Millionbook project) at the end of 2002 and finished 50000 E-books and dissertation in Jul 2006. To facilitate the project smoothly forward and meet the growing demand for digitizing capabilities throughout Tsinghua Library, the Digitization Center was founded in March of 2003 as part of the Digital Library Research Division of the Tsinghua University Library. After 3 years practices, some experiences are shared here:

A .In house or out source

To implement CADAL project, sixteen digitization centers established at sixteen participated academic libraries. 12 centers chose outsourcing mode. Tsinghua University Digitization center finally decided in house digitization process after carefully investigation. Different mode has different advantages and disadvantages. The primary advantage for in house process is that the center can control over all procedures, handling of materials and quality of products. There is no need to send valuable or fragile originals off-site and no worry about working with a vendor who turns out to be incompetent, provides something other than what was required, or goes out of business. Especially digitizing special collections, including unique materials always present challenges for handling and security. More important, in house operation provides a foundation of experience that helps to create policies, cost analyses, standard making, and data transferring. Also, keeping the production line in house makes other small digitization projects smoothly forward in the whole flexible organization. On the other hand, in house operation exposed some problems such as staffing and management efficiency. Most of the libraries all agreed that there was a need for instructions especially in the area of software's selection, staff and workflow management, and quality control. Outsource is also a good choice if the library hope better quality output and to speed up the process in the short time and no need to keep the expertise in library. But some of libraries found that some vendors especially small ones could not meet the targeted deadlines. They must also be careful when choosing vendors, since the intellectual content and copyright belong to the universities and can be subject to abuse, especially some e-book publisher. It's difficult to constrain them not to expand their collection using MBP resources with government funding. Some library outsourced the scanning process but carried out in library, in this case, libraries need to provide

space and equipments for the vendors. Many libraries chose this kind of mode to finish the task..

B. Planning and Source Material Selection

50000 volumes was a big number for us to be digitized in three years, although Tsinghua University Library owned 2.5 million volumes collection totally. What kind of materials would be selected is the first problem we faced. Firstly, copyright was the place to start, if the materials was in the public domain, the work could proceed. CADAL's collection was classified as ancient book (Before 1911), Book published in Republic of China (1911-1949), Journal published in Republic of China (1911-1949), Modern Book (After 1949), Dissertation, Drawing, Video, and English Book. After discussion deeply with ancient book department and circulation department, we decided to digitize the traditional thread binding book first. Two reasons pushed to make the decision, Firstly, this part was easy to handle, and especially we could scan two pages once to speed up. Secondly, this part wasn't retrievable in the OPAC system, and necessary to be discovered urgently. CADAL's collection selection policy was that each center submitted the list to be digitized to Zhe jiang Univeristy Library, the administrator center of CADAL. After their de-duplication processing, the list returned from Zhe Jiang Univ. could be regarded as part of CADAL collection list. So the center which sends the list firstly has more chance to be permitted for digitization. After interviewing some students in humanity department, we quickly selected 40000 useful traditional books for the project, and other 10000 volumes were planed to digitize dissertation which part is unique collection and needn't to be de-duplicated.

C. Digitization Process

To finish a CADAL E-Book, several steps were necessary: Material preparation, Scanning, Image processing, Metadata creation and pachaging, quality control and data storage and backup. CADAL provided lots of related softwares for the whole process: Scanning tools as Quickscan, Finereader 7.0; Image processing tools: ImageProcess, Scanfix, Irfanview; Metadata creation and Encoding tool: OEBEditor; Recognition and image conversion tools: CADAL e-book creation package, DJVUPRO. Although these softwares can help us establishing the digitization line, the speed and efficiency still were not satisfactory in the image process phase. Different from other digitization centers, after selection and comparison, we selected the software named Bookshop to process the images. This software supported image batch cutting, cropping, clearing and de-skewing. The steps for CADAL image process are the following:

1. Scan two pages in one image with 600 dpi. (See Fig 1)
2. Rename the files with odd number to accommodate double pages in one scanning image naming using XnView software for batch process. The process is necessary for the third step's renaming.
3. Check each image for de-speckling and cropping by manual work.
4. Cut one image into two pages using Bookshop software and name each image in order.
5. De-skewing each pages batch using Bookshop. (See Fig 5)
6. Convert Tiff files into Djvu format using Djvu Pro software.

Tip: Turn up-down to scan the traditional thread books, because the page number is from right to left, when the page was cut into two pages using Bookshop software, the naming would be given from left to right.

Fig. 1. After the first step

Fig. 5. After 5 steps

D. Facility and Staff Management

The room 510 around 80 square meters is adjacent to the office of digital library research division. The facility was considered based on the room space and digitizing content: Three flatbed AVA3 AVISION scanners, two FB6000E AVISION flatbed scanners for the bookedge near the binding scanning, and a Minolta PS 7000 for face up scanning. These equipments were provided by CADAL administrator center. Data's preservation and the safety are quite important for special collection's retro-digitization. Considering the cost and access frequently in each step, we chose Dell 2850 as the server connected with 2T Dell 220S storage. The open infrastructure was easily to be expanded. Each client PC accessed data in Dell 2850 for image processing & packaging. Another NAS Conventive 5000 system connected with Dell's DAS system with 1000GB network card for final data's backup. Also, we prepared 12 200G hard disks off line backup system as the third copy. Totally the data space is round 7T. After consulting the experts from Peking University Library, Zhejiang University Library, and Library of Chinese Academy of Sciences, we recruited 10 working staff. 6 persons for scanning process in two time

slots, 3 persons for metadata and E-book making, 1 person for quality control. The image process was arranged in additional working time. Staff who are willing to get more money by working more one or two hours can take this part of job. The basic idea for the arrangement is that using the equipments sufficiently. So the scanning time was planned in 14 hours per day for two batches of staff. First batch is from 7:00-4:00 with 1 hour lunch time, the second batch is from 12:00-8:00 with 1 hour super time. The performance management of staff is a critical factor for the cost accounting. Each person need scan 400 pages per hour with 200 images.

II. OPEN PROBLEMS AND CONSIDERATIONS

The first phase CADAL was finished in Jul 2006, some problems and thoughts were described in follow for the second phase plan's consideration.

A. content Discovery

The digital content in the first phase CADAL Project are almost image-based. The content discovery relies on metadata description. The administrator center distributed several roughly metadata standards for each digitization center and each center filled in the metadata information based on their own understanding and practice. As a result, even same series can't be retrieved as a whole due to the different description for the collaboration digitization and metadata creation. More flexible and specific standards with more cataloging details and examples are needed in future.

B. Resource Selection

In the first phase, each center must finish the digitization job before the deadline. They have limited time to consider their whole collection digitization planning combined with the content in CADAL. Even more, some source materials were selected blindly thus the coverage of the million books was not systematically. As an example, although the source books published in Republic of China are around 200000 covered in CADAL, we don't know how many published in this period are not included and in which how many we have in each academic library. In a word, we need to make clear the academic library union collection and the percentage of the CADAL content share.

C. OCR Processing

One of the objectives of CADAL is to form a full-text searchable data house. Now each digitization center was distributed a suite of OCR toolkit. The kernel of OCR software was developed by Tsinghua University Electronic Engineering Department which represents the highest level of Chinese OCR. To achieve high performance and efficiency, the OCR kernel technologies such as character recognition and segmentation, layout analysis and understanding etc should be improved especially focused for CADAL ancient book resources. CADAL project may fund such technical-oriented project for better services.

D. Copyright Problem

Till now, almost 400,000 dissertations and modern books of CADAL collection are not open for Tsinghua University. Because this part is still in copyright protect domain. Getting the permission is a big task for administrator center. How to find a best solution need to be discussed.

E. Organization Structure

CADAL established sixteen digitization centers. Training task is quite heavy, especially in the beginning phase. For better support, two technical support centers were setup. Zhe Jiang University Library is responsible for the south area of China, while Tsinghua University Library is responsible for the north area of China. As I mentioned before, more than 10 centers chose outsourcing. Some center in south of China outsourced the job to the vendor in Beijing. The big disadvantage is that the digital

content will be hold by vendors. We can't control their abusive behavior. If we change the pattern, for example, to establish sixteen or more collection providers, and reduce the number of the digitization centers, the support task will be relatively easy, the metadata will be consistency and the collection can be controlled more effectively.