Mass digitization of the collections of the academic libraries in China

By Prof. Jihai Zhao

Zhejiang University Libraries, Hangzhou, China

Email: jhzhao@lib.zju.edu.cn

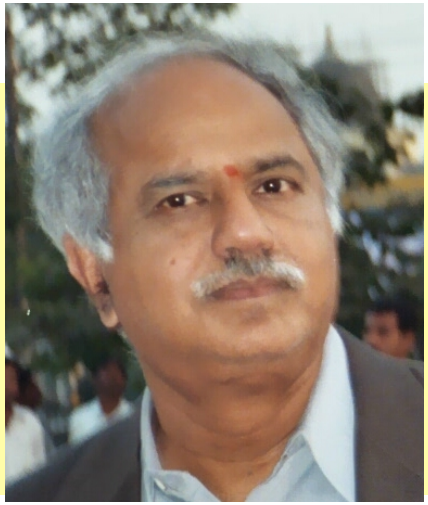
Outline

- Introduction
- Content Selection
- Data Production
- Metadata Creation
- Digital Preservation and Access
- Current Status
- Next Step
- Conclusion

1. Introduction

- The Million Book Digital Library Project (MBP) is one of mass digitization projects around the world, jointly carried out by universities and institutes in China, USA and India, with funding from the Ministry of Education of China (MOE), the National Science Foundation of USA (NSF) and Indian Government. The project was initiated by Dr. Raj Reddy, Professor of Carnegie Mellon University in 2000.

"Attempt to understand & solve the technical, economic, & social policy issues of providing online access to all creative works of the human race."



Dr. Raj Reddy

- In Dec. 2000, the *Memorandum of Understanding on the China-US Million Book Digital Library Project* was signed.
- A short name of the project: CADAL (China-America Digital Academic Library).
- In 2002, the Ministry of Education of China (MOE) decided to invest RMB 70 million yuan to scan 1 million books.

- The CADAL Project is led by Zhejiang University and Chinese Academy of Sciences as the PIs, and is jointly implemented by other 14 key universities in the China, including Peking Univ., Tsinghua Univ., Fudan Univ., Nanjing Univ. etc.
- By Aug. 2006, 1 million books had been scanned, and the first phase (2002-2006) project was completed. The second phase project (2008-2011) will be initiated soon.

2. Content Selection

- The content to be scanned was selected in the collections of the libraries in the partner universities and institutes of the project.
- The Chinese Language materials selected to be scanned include
  - ancient books published before 1911, which are out-of copyright,
  - books and journals published since 1911, which may be in the public domain or still in the copyright protection period, and
  - dissertations of the partner institutions.

- The selection of English books was mostly carried out in USA, and it focused on non-copyrighted materials and copyrighted ones of receiving permissions for digitization and Internet access.
- English books were also selected for scanning in the academic libraries in China.

- How to avoid duplication of scanning books.
  - 4 partners responsible for check duplication for 4 types of books and materials, respectively.

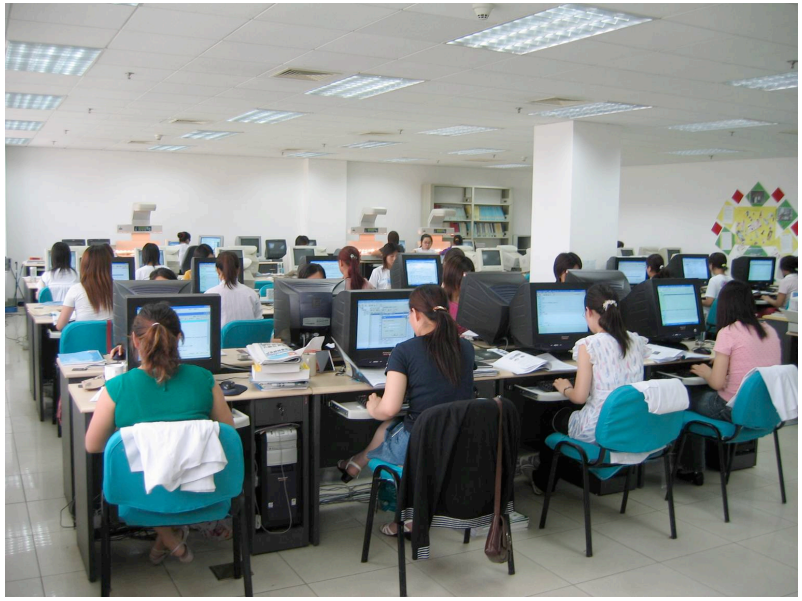
3. Data Production

- 16 Chinese partners established scanning centers in their libraries. They are
  - Chinese Academy of Sciences
  - Fudan University
  - Nanjing University
  - Peking University
  - Tsinghua University
  - Zhejiang University

A scanning center at Zhejiang University

Data Production (cont.)

- Beijing Normal University
- Huazhong University of Science & Technology
- Jilin University
- Shanghai Jiaotong University
- Sichuan University
- Sun Yat-sen University
- Wuhan University
- Xi'an Jiaotong University
- Renmin University
- China Agricultural University



A scanning center at the Chinese Academy of Science Library

Data Production (cont.)

We established a scanning center in Shenzhen Free Trade Zone for digitizing the English books shipped from US and Hong Kong.

Scanning center in Shenzhen Free Trade Zone


Distribution of scanning centers in China for the Million Book Project
Changchun
Beijing
Xi'an
Nanjing
Shanghai
Chengdu
Wuhan
Hangzhou
Guangzhou
Shenzhen

Data Production (cont.)

- The scanning approach is to digitize the documents at the archival quality of 600 dpi. The resolution of the images is high enough so as to permit printing as legible as the original pages.

- Quality Control (QC)
  - QC in scan and data production
  - QC after data submitted to the South Center in Zhejiang University (15 staff) and the North Center in the Academy of Sciences (7 staff).

Chinese OCR

- 130 PC servers are equipped in the CADAL OCR center, conducting the character recognition. More than 4000 books (about one million pages) can be recognized every day.

4. Metadata Creation

- For Chinese books, a defined metadata standard (Edocument Metadata, Version 2.0) has been released, which combines DC with CNMARC. The software named "MetaCreator" has been developed by Zhejiang University Libraries and applied in the scanning centers in the partners.
- For English books, OCLC provides the project partners with MARC records at no charge.

5. Digital Preservation and Access

- CADAL has produced approximately 300 million pages. The database houses both an image file and a text file at about 50-60 megabytes per book. Creating and managing such a vast information base poses many technological challenges and provides a fertile test bed for innovative research in many areas.

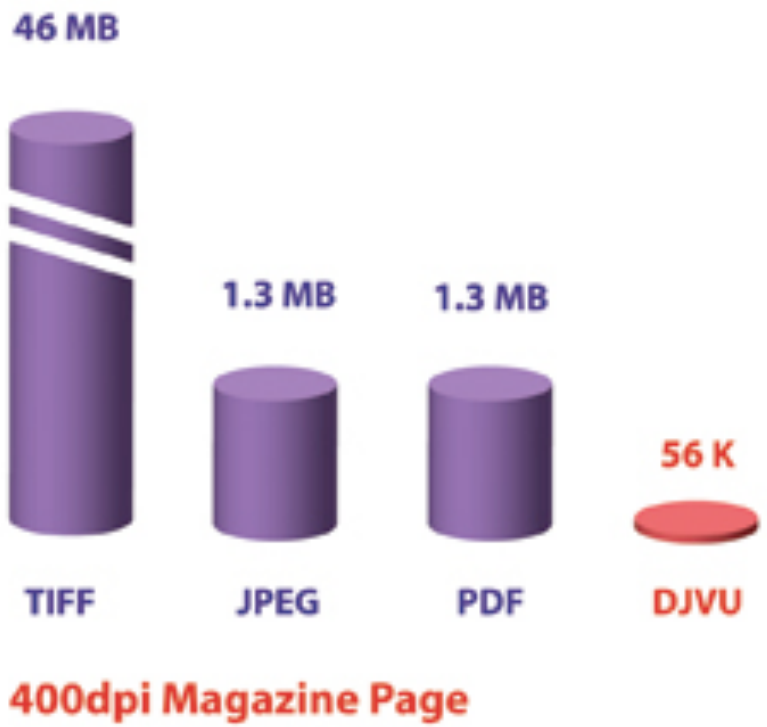
Digital Preservation and Access (cont.)

- Mirroring the database in several places in China can not only provide fast access, as the network speeds at the various nodes would be different, but also ensure security and long-term preservation.

- The images of the CADAL digital books are used for viewing, while the texts produced by OCR are used for searching. To speed up the viewing of the images, we have tested and applied the DjVu as the publishing format. DjVu technology is a highly sophisticated imaging language developed at AT&T Labs. Conventional image-viewing software decompresses images in their entirety before displaying them. DjVu technology, however, keeps the image in memory in a compact form and decodes only the area displayed on the screen in real time as the user views the image. As a result, the initial view of the page loads very quickly, and the visual quality progressively improves as more bits arrive.

Small File Size

DjVu document images are one of the smallest in the industry, up to 1,000 times smaller than TIFF files, and anywhere from 10 to 100 times smaller than JPEGs or PDFs depending on how these JPEGs or PDFs were created.



monetary policy "accommodates" a change in fiscal policy; that is to say, the supply of money and credit is expanded just enough to avoid any increase in interest rates. (The effect of relaxing this assumption is explained later.)

If the government increased its purchases of goods and services by $10 billion, private income before tax would initially rise by the same amount. Tax revenues would be $2.5 billion higher, and private disposable income would rise $7.5 billion, of which consumers and business would spend $6 billion. This additional spending would generate another increase in income, with $1.5 billion going to taxes and the remaining $4.5 billion to consumers and business. Of the latter amount, consumers and business would spend $3.6 billion, which would generate still another round of rising income and spending, and so on. The total increase in GNP (including the initial $10 billion of government purchases) would amount to $25 billion ($10 + 6 + 3.6 + \cdots$). This is a multiplier of 2.5 times the original increase in spending.

Consider what would happen if, instead of increasing its purchases, the government reduced tax rates by the equivalent of $10 billion. Consumers and business would again spend 80 percent of the higher after-tax incomes, or $8 billion. This would generate the same amount of additional private income, of which consumers and business would receive $6 billion and spend $4.8 billion, and so on. The total increase in GNP would be $20 billion ($8 + 4.8 + \cdots$), or two times the original tax cut. The difference between the multipliers in the two illustrations reflects the differences in first-round effects of the expenditure and tax changes: in this round output is raised by the entire amount of an increase in purchases but by only 80 percent of a tax reduction. (The first-round effects of an increase in transfer payments—say for social security,

文史之阨無代無之爲古書延一線之脈作續
之湯俾國史不至無徵弘文藉以不朽者賴有
古之士刊布不絕傳鈔未已耳姚士粦序尚白
祕笈云吾郡未嘗無藏書家卒無有以藏書聞
蓋知以祕惜爲藏不知以傳布同好爲藏耳何
祕惜則箱橐中有不可知之秦劫傳布則毫楮
有遞相傳之神理此傳不傳之分不可不察者
所謂不可傳布之說有四大抵先正立言有一
怒而百世與者則子孫爲門戶計而不敢傳鬬
炫博樂於我知人不知則寶祕自好而不肯傳
軸相假無復補壞刊謬而獨踵還癡一諺則虞
鈔而不樂傳舊刻精整或手書姸妙則懼翻摹

籀文考述

籀文是與西周文字有直接關係、並且迄今仍然懸而未決的文字學和書法史上的大問題。由於《史籀篇》是中國歷史上第一部字書。秦始皇書同文字時以它爲底本改作小篆，遂使古往今來的有關研究著述都不免有所涉及。因循舊例，筆者亦將所獲管見陳述如次。

一 問題的癥結所在

按照文獻記載，《史籀篇》爲最古的字書，其字爲『籀文』，又名『大篆』。《漢書·藝文志》小學列《史籀》十五篇，自注云：『周宣王太史，作大篆十五篇，建武時亡六篇矣。』《說文解字敘》稱：『及周宣王太史籀，著大篆十五篇，與古文或異。』是『籀』爲人名，官職爲太史。近人王國維以籀有『誦讀』之義，遂以『史籀』爲『太史籀書』之省略，後人取爲篇名。①或以『史籀』連讀爲人名，即《漢書·古今人表》所載春秋戰國間之『史留』。②此爲問題之一。許慎著《說文解字》的宗旨是『今敘篆文，合以古、籀』，據東漢尚存的《史籀篇》九篇列出與小篆寫法相異的籀文二百二十五個，表明它多數與小篆同形。王國維由此懷疑《史籀篇》非周宣王時所作，乃是春秋戰國間的秦人作品，爲『西土文字』，王說的金是『別體認爲周宣篆，《史時代、通爲『籀文被採用在玄想千載先生說：部字書，生了一些字形產生三，如果也應視爲

我們

科學的關

Digital Preservation and Access (cont.)

- A portal has been established at Zhejiang University Libraries, and the users of the partners are allowed to search and view the digitized collections in the public domain of the project freely within the IP addresses of the participating institutes:

  http://www.cadal.zju.edu.cn

高等学校中英文图书数字化国际合作计划
CHINA-US MILLION BOOK DIGITAL LIBRARY PROJECT

中文版

Home | Services | Help | About CADAL | Login

Searches: Quick | Advanced | Image | Video | Calligraphy

Search Browse

☑ Ancient ☑ Minguo ☐ Journal ☐ Modern ☐ Dissertation ☐ Painting ☐ Video ☐ English

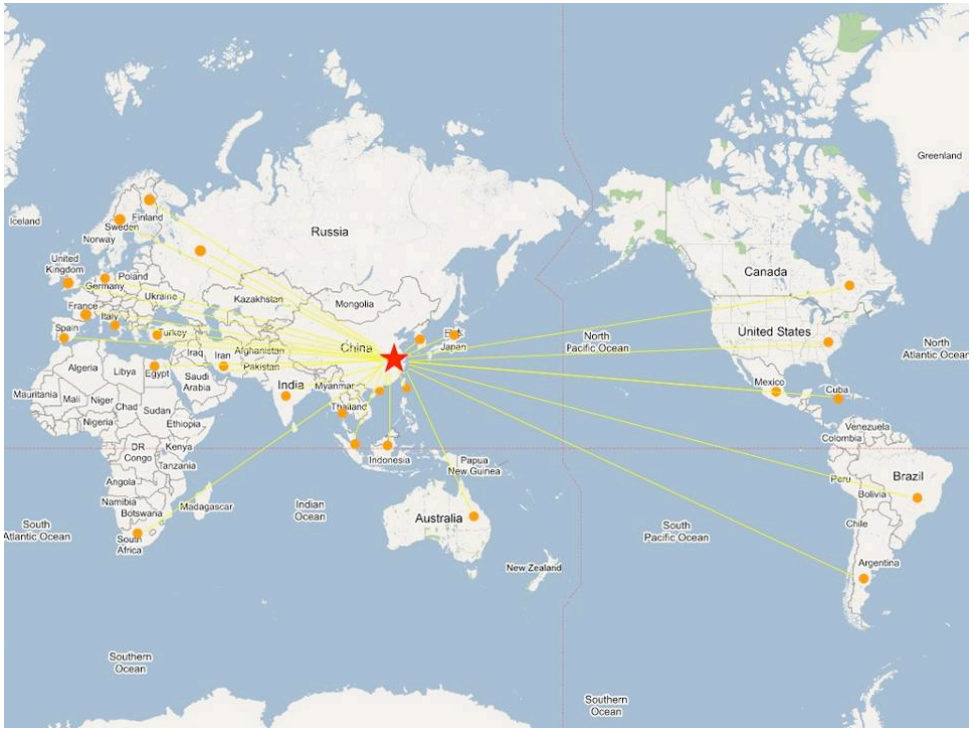
Login / Register

CADAL Center | CALIS | Zhejiang University

DCD Laboratory | CADAL Team | Help

Digital Preservation and Access (cont.)

Users around the world have access to the digital books, covering 71 nations and regions, and 3,681 books or 149,672 pages were read per day averagely for the past year.



6. Current Status

- Till now, 1.41 million items of Chinese and English books and other materials have been scanned in the Chinese partners.

Current Status (cont.)

- The digitized materials include
  - **207 thousand** Chinese ancient books,
  - **315 thousand** Chinese books and journals published during 1911-1949 (*Min Guo*),
  - **460 thousand** Chinese books published after the year 1949,
  - **190 thousand** dissertations and theses, and
  - **237 thousand** English books

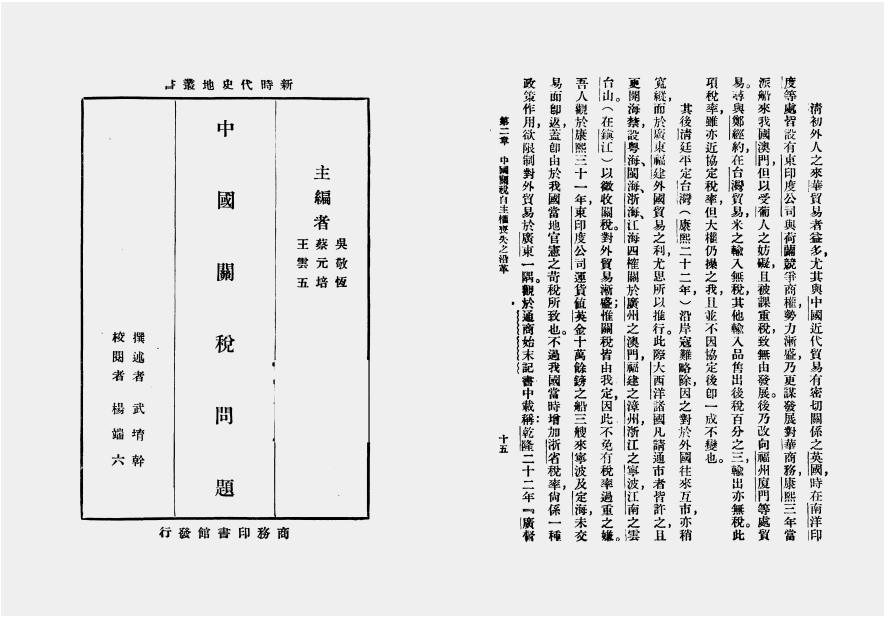
7. Next Step

- Scanning the second million books.
- Digital preservation. Storage and backup.
- Access. CADAL portal improvement.
- Cooperation. develop new partners, technology and personnel exchange…

8. Conclusion

- Mass digitization is a good means to preserve the fragile ancient and old books for the academic libraries, and to provide digital service to the public as well as the academic communities.

Min Guo books Before scanning



新時代史地叢書

中國關稅問題

主編者 吳敬恆 蔡元培 王雲五

撰述者 武堉幹

校閱者 楊端六

商務印書館發行

清初外人之來華貿易者益多，尤其與中國近代貿易有密切關係之英國，時在南洋印度等處皆設有東印度公司與荷蘭競爭商權，勢力漸盛，乃更謀發展對華商務，康熙三年當派船來我國澳門，但以受葡人之妨礙，且被課重稅，致無由發展。後乃改向福州廈門等處貿易，尋與鄭經約，在台灣貿易，米之輸入無稅，其他輸入品舊出後稅百分之三，輸出亦無稅，此項稅率，雖亦近協定稅率，但大都仍機之我，且並不因協定後即一成不變也。

其後清廷平定台灣（康熙二十二年）沿岸寬鬆略除，因之對於外國往來互市，亦稍寬緩，而於廣東輻湊外國貿易之利，尤思所以推行。此際大西洋諸國凡請通市者皆許之，且更開海禁，設粵海、閩海、浙海、江海四權關於廣州之澳門，福建之漳州，浙江之寧波，江南之雲台山。（在鎮江）以徵收關稅。對外貿易漸盛；惟關稅皆由我定，因此不免有稅率過重之嫌。吾人觀於康熙三十一年，東印度公司運貨值英金十萬餘鎊之船三艘來寧波及定海，未交易而即返，蓋即由於我國當地官憲之苛稅所致也。不過我國當時增加浙省稅率，尚係一種政策作用，欲限制對外貿易於廣東一隅。觀於通商始末記書中載稱乾隆二十二年『廣督

第二章 中國關稅自主權喪失之證據 十五

Scanned pages of a *Min Guo* book

After scanning, the *Min Guo* books are patched up and bound up in the libraries, so that the state of books is improved for long preservation.

Book patching and binding after scanning

Patched and bound *Min Guo* books

- We are preparing the second phase of the project (2008-2011) , and hope to scan next 1 million books and more materials in other formats, such as audios, videos, newspapers, microforms, paintings, photos, and so on. We hope CADAL project will make more and more great contributions in knowledge dissemination, academic communications and digital preservation of cultural treasures, beneficial for this and next generations.

Thank you very much!

October 30, 2008 in Singapore