Digital Preservation in MBP

Huang Chen
Director of Digital Resources R&D Center,
Deputy-Director of South Technical Center of CADAL
Zhejiang Uni. Libraries
2006.08.16

CADAL
Administrative Center for
China-US Million Book
Digital Library Project

Topic

- Over view
- Storage Construction
- Preservation
  - Format
  - Exchange & Migration
  - Backup
- Next Step

CADAL

Administrative Center for
China-US Million Book
Digital Library Project

Over View

- The China-US Million Book Digital Library Project (MBP) is a cooperated project of universities and institutes in China and USA, with funding from the Ministry of Education of China (MOE) and National Science Foundation of USA (NSF).
- The objective of this project is to create a free-to-read, searchable collection of one million books, with a half in the Chinese language and the other half in the English language, available to everyone over the Internet.

CADAL
Administrative Center for China-US Million Book Digital Library Project

Over View

- MBP in China is led by Zhejiang University and Chinese Academy of Sciences
- and is jointly implemented by Peking Uni., Tsinghua Uni., Fudan Uni., Nanjing Uni., Shanghai Jiaotong Uni., Xi'an Jiaotong Uni., Wuhan Uni., Huazhong Uni. of Science and Technology, Zhongshan Uni., Jilin Uni., Sichun Uni. and Beijing Normal Uni.

Over View

- 16 Centers in China
- 15 Centers in India
- 1 Center in Egypt
- Planned :
  - Australia
  - and Europe

CADAL

Administrative Center for China-US Million Book Digital Library Project

Digital Process Center

MBP

Digital Process Center

MBP

Digital Process Center

MBP

Over View

- About 1,040,000 books scanned in China
  - 941,134 volumes of Chinese books and materials
  - 98,206 volumes of English books
  - About 700,000+ accessible on the web
  - Uses 50TB of storage

CADAL
Administrative Center for China-US Million Book Digital Library Project

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。

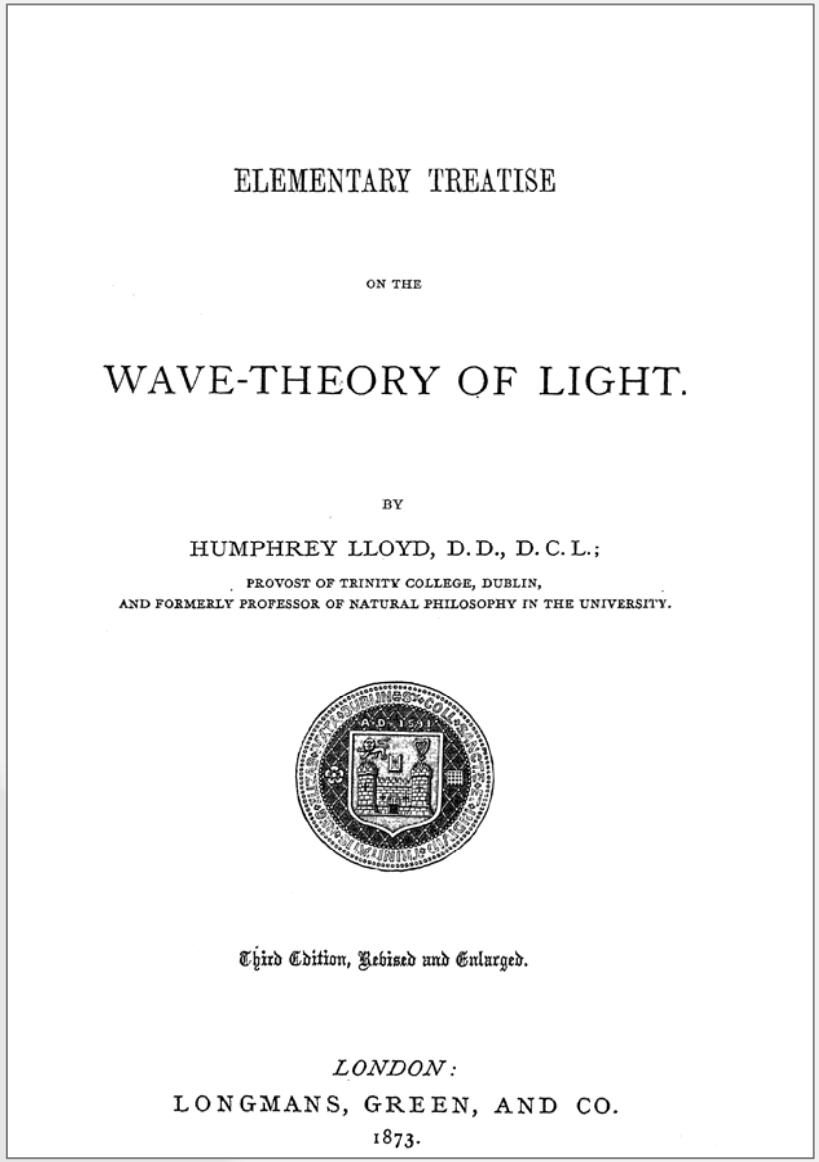
一 問題的癥結所在

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[illegible]是『別體[illegible]認爲周宣[illegible]篆，《史籀[illegible]時代、通行[illegible]爲『籀文』[illegible]被採用在[illegible]玄想千載以[illegible]先生說：[illegible]部字書，[illegible]生了一些[illegible]字形產生[illegible]三，如果[illegible]也應視爲[illegible]

我們[illegible]科學的關

Title	Elementary Treatise on the Wave-Theory of Light	
Author	Humphery Lloyd, D.D, D.C.L	
Language	English	
Subject	Physics	
Publisher	Longmans, Green & Co	
Year	1873	
Abstract	This book deals with the various aspects of the wave theory of light. It is a critical work which contains an analytical discussion of the most recent researches in Optics. It presents a clear and connected view of the subject.	

MBP


ELEMENTARY TREATISE

ON THE

WAVE-THEORY OF LIGHT.

BY

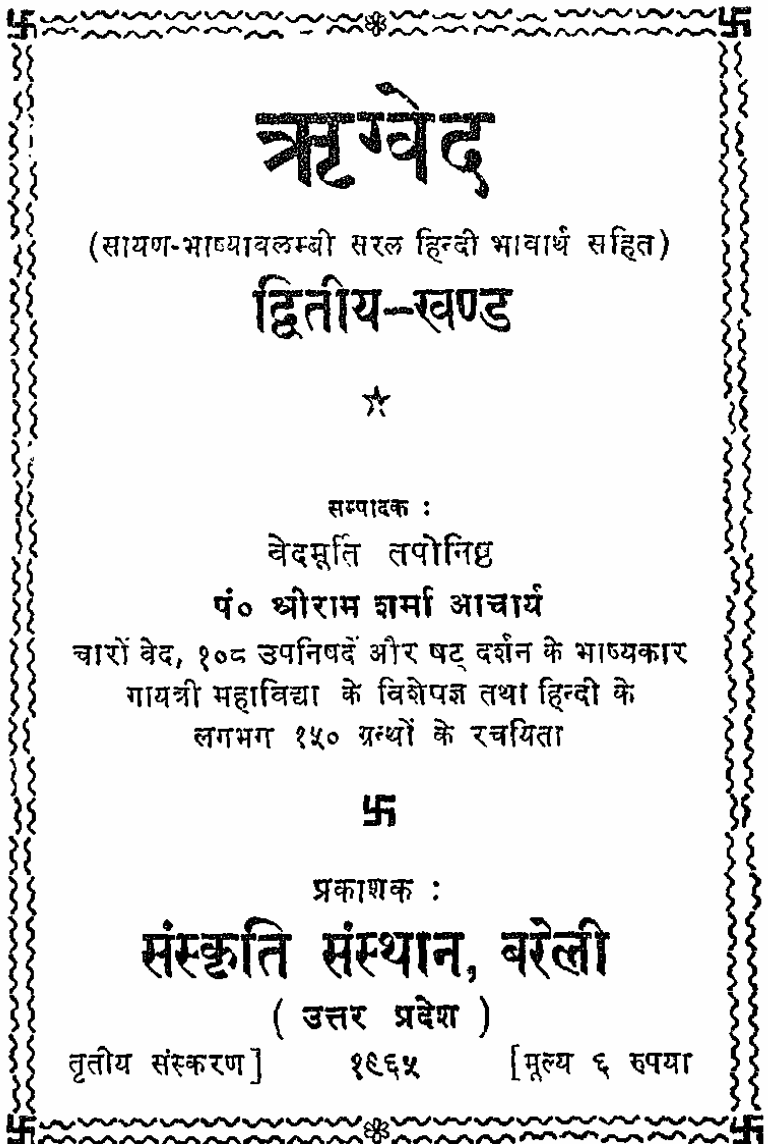
HUMPHREY LLOYD, D.D., D.C.L.;

PROVOST OF TRINITY COLLEGE, DUBLIN,
AND FORMERLY PROFESSOR OF NATURAL PHILOSOPHY IN THE UNIVERSITY.

Third Edition, Revised and Enlarged.

LONDON:
LONGMANS, GREEN, AND CO.
1873.

Title	Rig Veda
Author	Pandit Sriram Sharma Acharya
Language	Sanskrit
Subject	Philosophy
Publisher	Sanskriti Sansthan Bareli
Year	
Abstract	Rig Veda is the oldest of the Vedas. The Rig Veda is the oldest book in Sanskrit or any Indo-European language. Many great Yogis and scholars who have understood the astronomical references in the hymns, date the Rig Veda as before 4000 B.C., perhaps as early as 12,000. Modern western scholars date it around 1500 B.C., though recent archaeological finds in India (like Dwaraka) now appear to require a much earlier date



ऋग्वेद

(सायण-भाष्यावलम्बी सरल हिन्दी भावार्थ सहित)

द्वितीय-खण्ड

☆

सम्पादक :

वेदमूर्ति तपोनिष्ठ

पं० श्रीराम शर्मा आचार्य

चारों वेद, १०८ उपनिषदें और षट् दर्शन के भाष्यकार
गायत्री महाविद्या के विशेषज्ञ तथा हिन्दी के
लगभग १५० ग्रन्थों के रचयिता

卐

प्रकाशक :

संस्कृति संस्थान, बरेली

(उत्तर प्रदेश)

तृतीय संस्करण] १९६५ [मूल्य ६ रुपया

Overview

MBP

- Web Service Available since Sep. 2004
- From Jan. to Jun.2006
  - Web total Hits:19,358,590;
  - Daily Hits:122,710;
  - Total Visitors:151,440；
  - Total Bandwidth 7534.6GB，equal to 1000 Vol./Day（45.5GB）。
- http://www.cadal.cn

CADAL

Administrative Center for China-US Million Book Digital Library Project

Topic

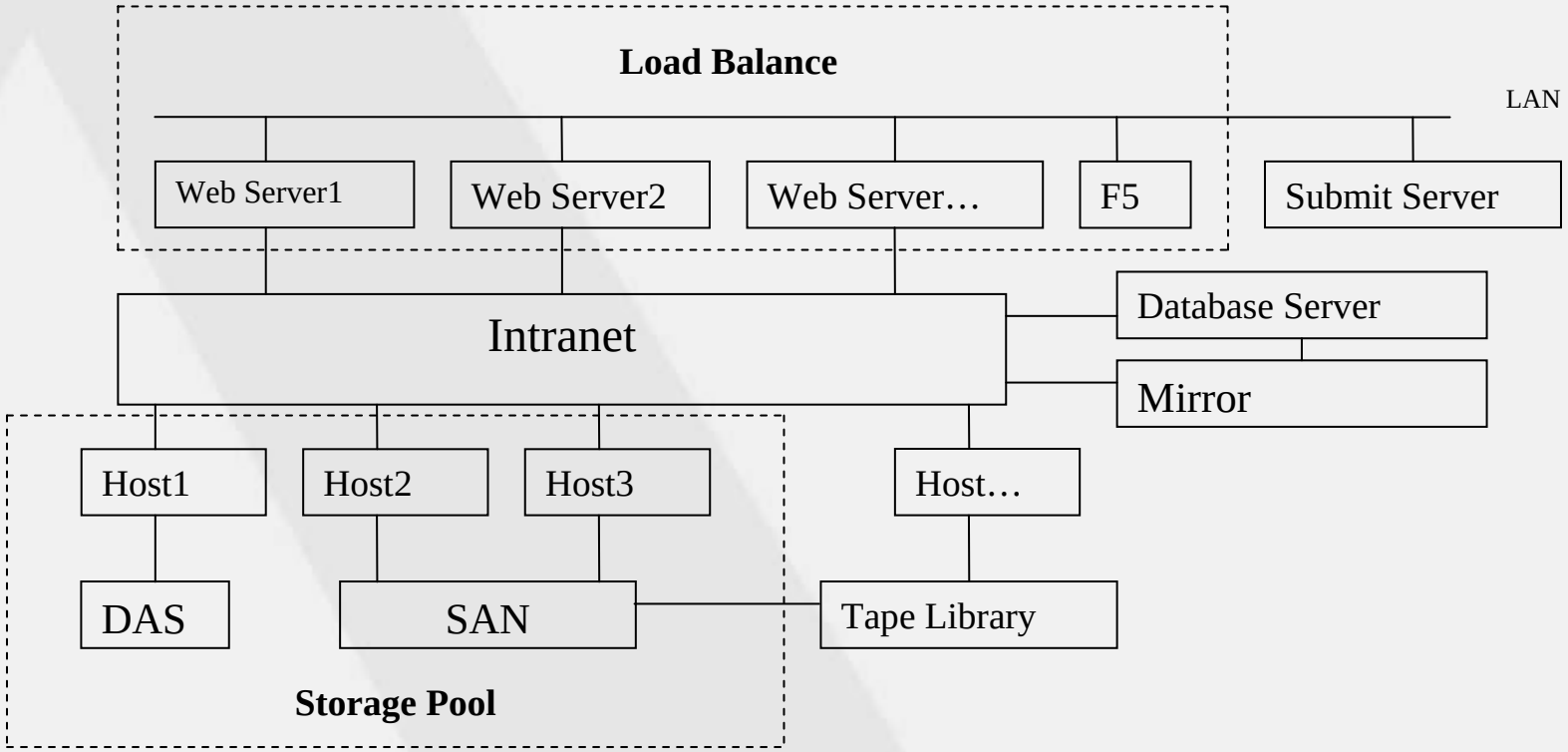
- Over view
- Storage Construction
- Preservation
  - Format
  - Exchange & Migration
  - Backup
- Next Step

CADAL
Administrative Center for China-US Million Book Digital Library Project

Storage Construction

MBP

System Figure



CADAL

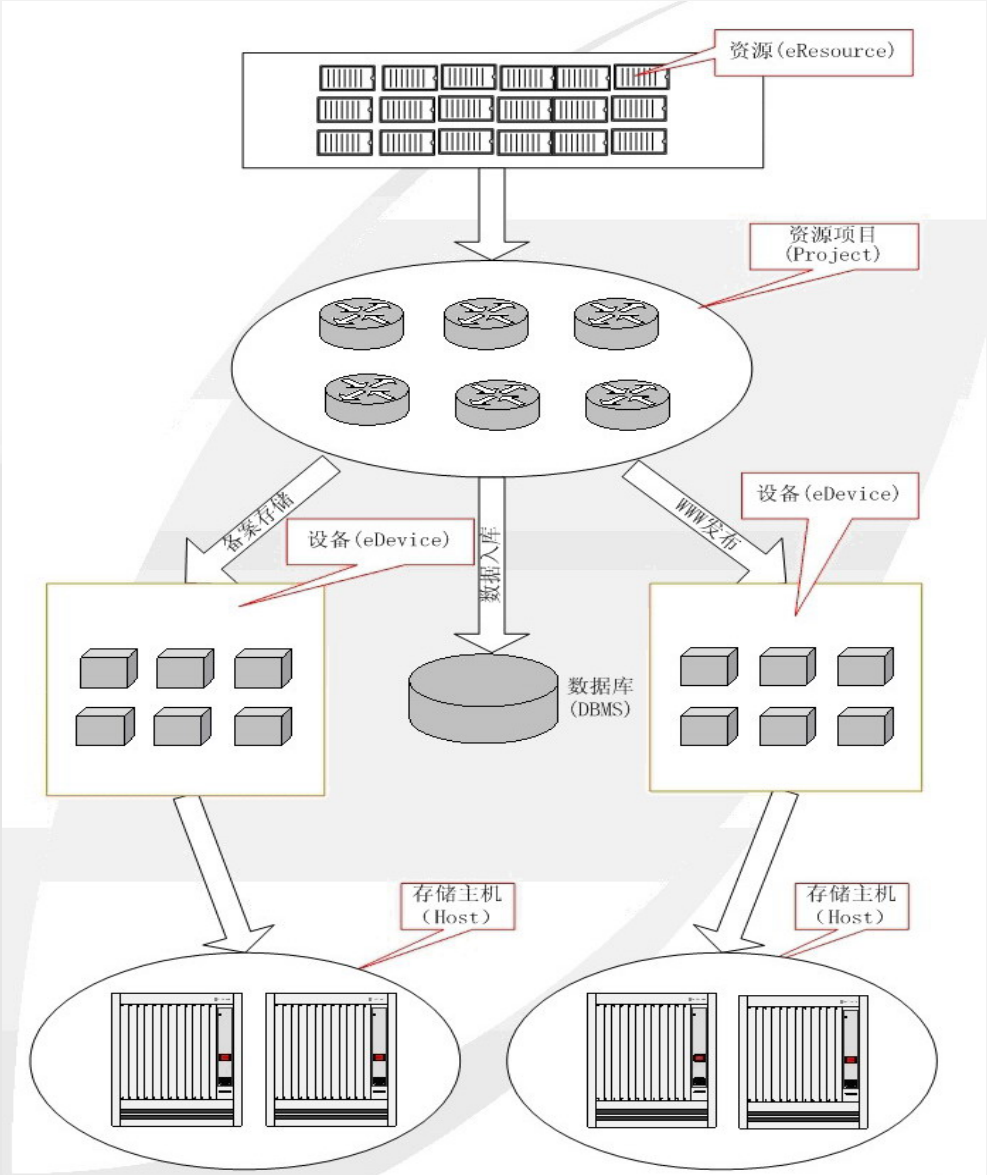
Administrative Center for China-US Million Book Digital Library Project

Storage Construction

MBP

Storage Layer

- eResource
  - The smallest storage unit, include document, image, audio, video, software, etc.
- Project (Collection)
  - To classify resource, for example, dissertation, rare books, etc.。
- eDevice
  - As a LUN come from a storage, like “E:” is a volume from NAS1 which belongs to DELL 220S, “F:” is a volume from SAN which belongs to IBM DS8100.



CADAL

Administrative Center for China-US Million Book Digital Library Project

Resource Expression

MBP

- Absolute Path
  - **eResource:\\eID.ProjectAlias.eDeviceAlias.HostAlias**
- Publish Path
  - **eResource:\\eID.ProjectAlias.virtualpath.HostAlias**
- **Note:**
  - **“eID” is the ID of resources, Sample: 06000001**
  - **“ProjectAlias” is a alias of the resource project**
  - **“eDeviceAlias” is a alias of the device, it points the absolute path of the storage device**
  - **“VirtualPath” is virtual path of WEB publish**
  - **“HostAlias” is a alias name of a host (Storage Server, Web Server, etc.)**

CADAL

Administrative Center for China-US Million Book Digital Library Project

Storage Construction

MBP

PetaBox In IA

- 100 Tb per rack,80 machines
- 150k in hardware, $1,500/TB
- http://www.archive.org/web/petabox.php

CADAL

Administrative Center for China-US Million Book Digital Library Project

Data Center

MBP

Data Center

MBP

Topic

- Over view
- Storage Construction
- Preservation
  - Format
  - Exchange & Migration
  - Backup
- Next Step

CADAL
Administrative Center for China-US Million Book Digital Library Project

Preservation

MBP

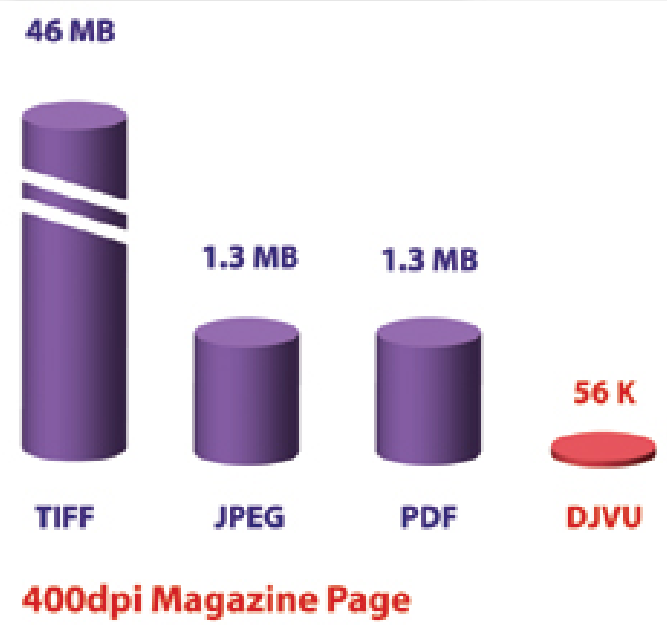
Format

- What format we need?
  - Storage
  - Representation
- Character
  - Smaller size
  - High resolution
  - Perfectly revert

CADAL

Administrative Center for China-US Million Book Digital Library Project

Format



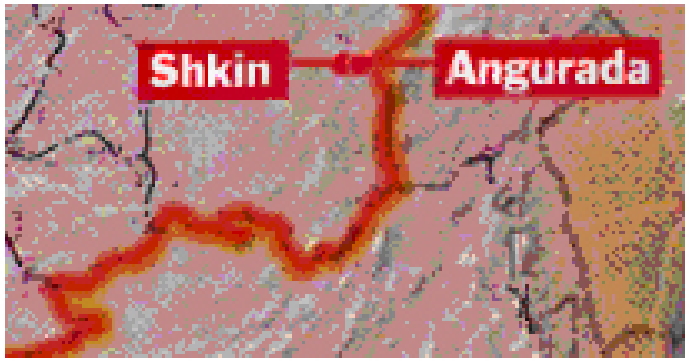
Small File Size

DjVu document images are 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.

Format

High resolution & Perfectly revert

DjVu: 104K



Very small file size AND high image quality and readable text.

Competing Technology: 137K

Very small file size BUT low image quality and unreadable text.

Preservation

MBP

Format

- eBook Directory
  - oTIFFs (TIFF)
  - pTIFFs (DjVu)
  - Metadata (XML)
  - Navigation (XML)
- Free Plug-in

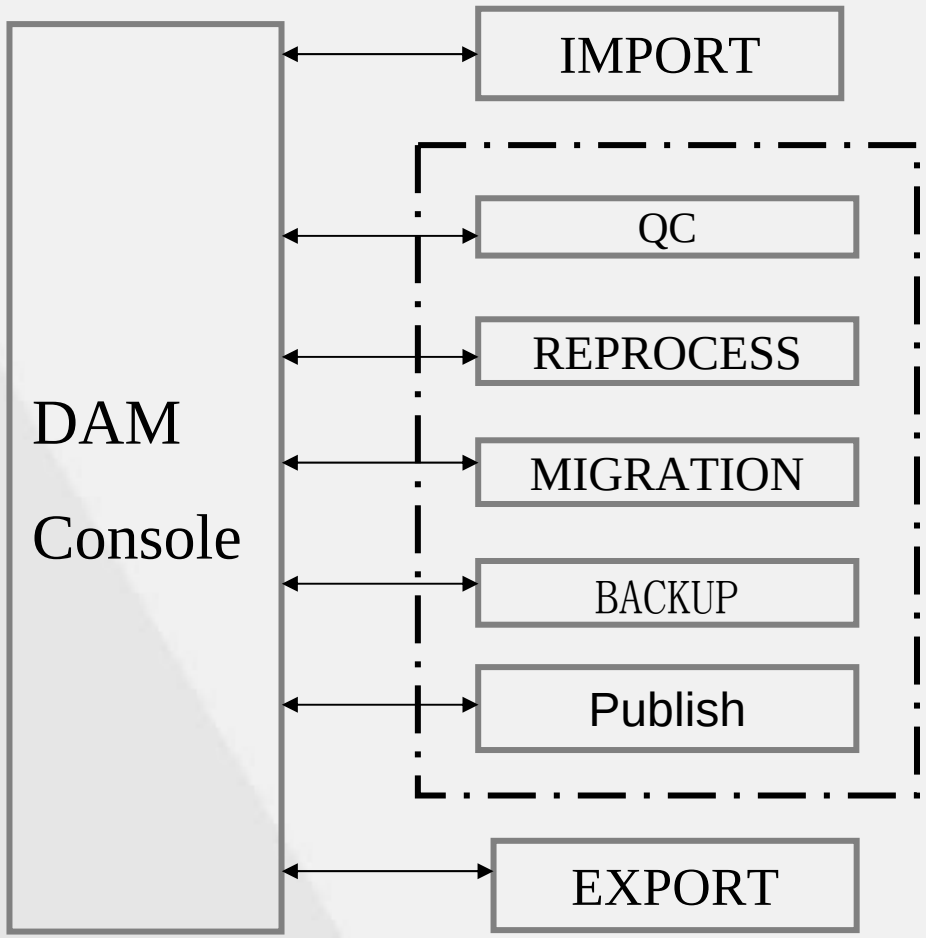
CADAL

Administrative Center for China-US Million Book Digital Library Project

Preservation

- Exchange
  - FTP
  - DVD
  - FTP (ISO with XML)
- Migration

Exchange & Migration



CADAL
Administrative Center for China-US Million Book Digital Library Project

Preservation

MBP

Backup

- Near line
- Offline

SATA Beast

- **42 x Serial ATA drive box**
- **Up to 21Tb in 4U**
- **Redundant RAID controllers**
- **Two 760W PSU's dual fans**
- **700 MB/s Sustained RAID 5 reads (DUAL contr / 4 x FC)**
- **520 MB/s Sustained RAID 5 Writes (DUAL contr / 4 x FC)**
- **RAID 6**

CADAL

Administrative Center for China-US Million Book Digital Library Project

Topic

- Over view
- Storage Construction
- Preservation
  - Format
  - Exchange & Migration
  - Backup
- Next Step

CADAL
Administrative Center for China-US Million Book Digital Library Project

Next Step

- Offline Backup
  - Tape Library
- Remote Backup
  - Mirror Site
  - Run time backup
- Auto Migration
- Union Storage Management

CADAL

Administrative Center for
China-US Million Book
Digital Library Project

Thanks

CADAL

Administrative Center for
China-US Million Book
Digital Library Project