Macroanalysis of China's Digital Library

Liu Xiaobai, Sun Wei & Zhen Xihui,
National Library of China
PRC

In modern world, computers and networks develop quite quickly. The information sources of network will be the nucleus of networking. Moreover, digital library will become future's strategic network system. Hours, days, months or even years calculate almost all kinds of information life cycles. Whose vitality is more powerful? Digital library's life is measured by decades or even centuries.

What will digital library deal with? It must organize information sources in good order, meet users' demands efficiently, protect intellectual property and share global knowledge resources by using modern high technology towards the twenty-first century.

There is plenty of information in libraries, museums and archives in the world. Much information is the telling witness of the historical development of the human beings and the treasure of man's knowledge. It is a challenge for us to utilize it to serve man by better using modern computer and network technologies. It is well known that computer and network technologies are the important basis for making full use of all kinds of resources. However, they are updated very quickly and these tools still cannot completely resolve our problems. We must seek more rational methods to realize the architecture of digital library.

Digital library stands for a new basic infrastructure and knowledge environment. It will be built into super large-scale, expandable and interactive knowledge deposit through integration and using most newly computing, communicating and digitizing technologies. At present, nearly twenty countries and regions are establishing digital libraries. The first batch of models has appeared on Internet. For example, fifty resource repositories have provided services via Internet in "America Memory" Digital Library Project of the Library of Congress. It will come to a climax by the year 2000.

In the past few years, developed countries have gained rapid development in the technical research of digital library. Gratifying achievements have also been made in domestic technical research. The constructive preparation work has attained great support and help from the State Council, the Ministry of Culture, the Ministry of Science and Technology, the Ministry of Education, the Chinese Academy of Sciences, Beijing

Municipal Government, the General Bureau of Broadcast and Television, Tsinghua University, Peking University, Beijing Telecom, Beijing Cable TV Station and overseas Chinese.

Part One: Overall Conception of China's Digital Library

In the world, some European countries and the United States are making great efforts in the research and construction of digital library, and have achieved many fruits, among them are:

- Using Digital Object as the basic elements to set up data managerial systems. Digital Object, a new data architecture, will be used in the next-generation Internet;
- Dividing the organization of resources into metadata and object data; the search system is aimed at the metadata set; the protection of intellectual property is aimed at the object data set;
- Markuping multimedia resources with SGML/XML;
- Adopting object-oriented technology such as CORBA to be the basis of cross-platforms;
- Prioritizing funds in the construction and organization of featuring resources;
- Cooperating in national and global sphere and constructing various resources;
- Network adaptivity is the basis of the utilization of digital library;
- Meeting users' demands.

As early as in 1995, the National Library of China began to trace the research of digital library of developed countries, and declared scientific research projects to the Ministry of Culture and the State Planning Committee. And in 1999 the experimental environment was established. The key technology has been verified.

- Technical support for the data resources construction, including using SGML to markup resources;
- Metadata search systems;
- Support for the storage of multi-status digitized resources and search systems;
- High-speed library-wide network;
- Using public Internet and user-interfaced networks;
- Experimentation and practice of cooperative construction of resource repositories;
- Search object database cross-provinces by using public networks, schedule object database cross-regions;

1. Overall Tentative Plan for the Digital Library Project

The digital library project is a super large-scale information-oriented distributed system. In this project, large-scale software, network, computers, information organization and market-oriented users' operation are needed. Public networks will be used to realize high-speed connection of library resources and users' access. Mature technologies will be used to expand distributed computer systems with mass storage. Mature software products will also be used, including 863 Project of China. We will establish information resource repositories through multi-party cooperation to promote the applications of digital library. Inside the National Library of China, we will set up resource-making system with the capability of industrial production and producing various kinds of multimedia literature.

(1) Pattern Analysis of Mass Data

Because of the massive information, the important problem is what methods will be used to realize resource sharing in digital library. Currently, every country has the same ideas in the research of digital library. It takes a long time for simple full-text search to inquiry mass databases. Simple full-text search does not help at all for mass information with decades or even hundreds of TB. For example, the United States has carried out a test for an information system with only 1TB using simple full-text search. However, it took six hours to find the results that meet users' needs.

There are some methods in organizing information:

A. We can divide information into metadata and object data. All kinds of inquiry tools can take their advantages by using metadata sharing;
B. We can send inquiry terms to URL or URN, and then send the search results to users.

The advantage of the first method is that it can make full use of information classification and feature description to construct feature data of some information. These data are called metadata. Then we can share the metadata. The search system searches information that meets users' needs in the metadata repository.

For example, there are 7.5 million items of literature each year in the world. Suppose there are 200 pages in each item and 400 Chinese characters. If each Chinese character uses two bytes, these items need 7,500,000 x 200 x 400 x 2 = 1,200 G space.

If we use key words to search, it will become very difficult.

Suppose we use metadata to save information. If each item needs 50 fields, the average length of each field is 1,500 bytes. So these items need 7,500,000 x 1,500 = 10.25 G space. It is no doubt that it will be more accurate to search data than just using simple full-text search.

The examples above make us have a general idea of simple object full-text search and data search with object descriptions. Therefore, the first thing we must do is further deal with data and divide data into metadata and object data.

(2) Analysis and Judgements of Systems' Processing Capabilities

a. Analysis of Capabilities That Metadata and Search System Need

Everyday lots of readers will use metadata search system if we adopt metadata to search and share. For example, the daily access of a newspaper in Singapore is 2.5 million hits. In the man vs. Machine Fight of Chess held in America, the hourly access is 3 million. In the observation of Mars, there are millions of hits daily. Adopting metadata search and sharing are similar with those examples. The computer system must find the search results in the shortest time, and transfer these results to users. There is a lower ratio of the bytes occupied to the object data. Therefore, the processing capabilities of computers and network systems will be bulk concurrent parallel and small transmitting number.

b. Analysis of Capabilities That Object Data System Needs

There are mass object data in digital library and massive memory space is needed. When a user needs to search object data after metadata inquiry, the object data system must confirm the rights and liability and then send object data to users under controlled condition. The treatment of E-business and copyrights is one of the nucleus technologies. Take the collection of the National Library of China for example. Suppose 1 million items of it is put into object database, and there are 200 pages in each item, and each page needs 15K bytes for scanning. So the memory space will be about 1,000,000 x 200 x 15,000 = 3,000 G.

Thus, it can be seen that 1 million items of literature need about 3,000 G of memory space. So object data should be distributed in different spaces in order to use computer and network system efficiently.

To those large resource centers, because users will use object data system to find their own answers, the concurrent parallel hits will be a large number.

Therefore, computers and network systems will face large concurrent parallel and large transmit in the object data system of digital library.

Sending search terms to URL or URN and then send search results to users will simplify the issues. However, it needs standard search terms of all resource systems. Otherwise, there is no way to share search terms. The important issue is to know how many URLs or URNs have the capability of searching. For instance, in Resource Center A, if a user wants to search certain key words, it must send the search item to itself and to the existing URLs or URNs at the same time. After URLs or URNs search in their own databases, they send the results to the users of Resource Center A. If there are 100 resource centers, and 1 million users search certain key words in resource center, in the network there will be 100 x 1,000,000 = 100 million.

It is a large number for 100 million search items to transfer in networks.

Therefore, the second key problems is:

- Computers and network systems for search will face large concurrent parallel and small transfer;
- Computers and network systems for resource transfer will face large concurrent parallel and large transfer;
- The making of metadata and object data;
- Centralized ability to search system.

If we adopt the making standards of metadata and object data, it will be as simple as using HTML. All kinds of search systems can search metadata. Because the object system is a distributed model, the other essential issue is how to schedule object data to users. The organization of scheduling system is of vital importance to the nation-wide digital library project.

And the third key problems is:

- Scheduling systems of nation-wide digital library resources;
- Object data system;
- Relativity between metadata and object data;
- Strict mechanism for E-business.

The digital library project is a nation-wide one. We must avoid duplicate construction and waste of resources. In the over all conception of the digital library project, we must make full use of telecommunications, broadcast and television and other public network systems, aiming at connecting the state center of digital library to branch center. In addition, we will use telephone lines, cable TV, and broadband IP networks to make links between these centers and the households and families.

In 2-2 the bold lines represent the high-speed interconnection between the State Center and Branch Center, and the fine lines stands for network systems between users and centers. It may be telephone lines, cable TV and broadband IP.

Subsystem scheduling Subsystem Inquiry Users
Metadata Base
Object Database Data Processing of Subsystem
2-1 Architecture of Digital Library

User Group User Group
State Center of Digital Library
User Group User Group

Branch Center of Digital Library Branch Center of Digital Library

User Group User Group User Group User Group

2-2 Centers for Digital Library

2. State Center of Digital Library

(1) Main Responsibilities

a. Establishing and Amplifying Standards and Norms That Digital Library Uses

If there are international standards, we can apply them indiscriminately. If there are no such standards, we will formulate our own industrial standards in accordance with international practice. Standards that need to be formulated as soon as possible are:

- Metadata standards;
- Object data construction standards;
- Standards for scheduling codes and relative naming principles;
- Data navigating standards.

b. Coordinating and Standardizing the Construction of Resource Repositories

We should organize some resource repositories of digital library with coordination of the entire resource center according to the resources in each of the resource centers.

In order to avoid duplicate construction and waste of resources, we should coordinate resource construction. The limited funds must be used for the establishment of resource repositories with the greatest usage.

c. Establishing Metadata Sharing and Search System of the State Center

The metadata resource center should be able to hold mass metadata system by using parallel data technology and distributed computer system.

A search system with large concurrent parallel must be established in the light of metadata resource center so that users can search quickly in both the state center and the branch center.

d. Establishing a Nation-wide Scheduling System of Digital Library

The nation-wide object data scheduling system is established to forward the scheduling system in the relevant regions according to the scheduling codes that a user defines when he/she searches metadata. And it will then point to the object system.

e. Establishing Object Data System of the State Center of Digital Library

This system is centered on the feature collection of the Nation Library of China. It must be able to send object data to users who need them according to the valid scheduling system, intellectual property and users' protocols. And it should be capable of dealing with large concurrent parallel and bulk transfer.

f. Establishing Manifold High-speed Networks Connecting to the Branch Center of the Digital Library

We should use efficiently all possible public networks to establish high-speed networks connecting to the branch center, focusing on the rapid circulation of scheduling systems and metadata sharing systems.

g. Establishing Research & Development Center of Digital Library

At present, digital library technology is one of the fields with fast development in the world. China is a developing country. China's digital library should have its own characteristics of technology and operation to meet the urgent needs of the urban and the countryside. Therefore, the establishment of research and development center is a necessity.

The main tasks of the center are:

- Designing and establishing the overall technical architecture of China's digital library;
- Providing comprehensive and long-term technical support for the construction of China's digital library;
- Researching the technological development of digital library in the world;
- Providing technical equipment for China's digital library and carrying out

proposals of different periods;

- Examing and checking the cooperative items;
- Coordinating standards, protocols, scheduling and resource usage with international digital libraries.

h. Establishing Links with Satellite, Covering Digital Libraries in Remote Areas

i. Establishing All Kinds of Users' Line Attachment

We should use cable TV and television networks to line with users in Beijing area, aiming at the realization of connecting digital library with innumerable households and families.

j. Processing Center of Digitized Resources

We should make the original data such as books, magazines, newspapers, CDs, Video Tapes, Cassette Tapes and Microforms become data which can be used in digital library.

k. Law Centers

The responsibilities of the law center are:

- Protecting intellectual property;
- Resolving copyright disputes;
- Using network resources;
- Protecting users' legitimate rights and interests;
- Safeguarding the network resources.

(2) Relations between the State Center, Branch Centers, Resource Centers and Metadata, Object Data, Scheduling Systems, Search Systems

2-3 indicates the relations between the state center, branch centers and resource centers. It also indicates the relations between metadata, object data, scheduling systems and search systems.

State Center of Digital Library

Resource Center

Metadata Resource Search System

Resource Center Object Data Resource

Scheduling System Users' Interface

Branch Center of Digital Library Branch Center of Digital Library

Metadata Resource Search System Metadata Resource Search

System

Object Data Resource Object Data Resource

Scheduling System Users' Interface Scheduling System Users' Interface

Resource Center Resource Center Resource Center Resource Center

2-3 Sketching Map of Metadata, Object Data, Scheduling Systems and Search Systems

There is a logical relation between these elements. However, the physical linking relation relies on the public networks all over China.

- We should make connection to the state center and the branch centers by using high-speed networks of 100 Megabits, 1,000 Megabits or even higher;
- We should use the users' line attachment of public networks to connect the information sources in digital libraries of the country to innumerable households and families;
- We should also use satellite channel to make connection to the resource centers in remote areas.

3. Branch Center of Digital Library

The main responsibilities are:

(1) Implementing Standards of the Construction for the Orderly Development of Digital Library in Both Branch Centers and Regional Areas.

(2) Coordinating and Standardizing the Construction of Resource Repositories

We should organize some resource repositories with coordination of the entire resource center according to the resources in each of the resource centers.

In order to avoid duplicate construction and waste of resources, we should coordinate resource construction. The limited funds must be used for the establishment of resource repositories with the greatest usage.

(3) Establishing Metadata Sharing and Search System of the State Center

The mirror metadata resource center should be able to hold mass metadata system by using parallel data technology and distributed computer system.

A search system with large concurrent parallel must be established in the light of metadata resource center so that users can search quickly in both the state center and the branch center.

(4) Establishing a Nation-wide Scheduling System of Digital Library

The nation-wide object data scheduling system is established to forward the scheduling system in the relevant regions according to the scheduling codes that a user defines when he/she searches metadata. And it will then point to the object system.

(5) Establishing Object Data System of the Branch Center of Digital Library

These systems are centered on the feature collection of the libraries all over China. They must be able to send object data to users who need them according to the valid scheduling system, intellectual property and users' protocols. Moreover, they should be capable of dealing with large concurrent parallel and bulk transfer.

(6) Establishing Manifold High-speed Network Connecting to the Branch Center of Digital Library

We should use efficiently all possible public networks to establish high-speed networks connecting to the state center and the other branch centers, focusing on the rapid circulation of scheduling systems and metadata sharing systems.

(7) Establishing All Kinds of Users' Line Attachment

We should use cable TV and television networks to line with users in local areas, aiming at the realization of connecting digital library to innumerable households and families.

(8) Processing Center of Digitized Resources

We should make the original data such as books, magazines, newspapers, CDs, VideoTapes, Cassette Tapes and Microforms become data which can be used in digital library.

4. Resource Center of Digital Library

The resource center represents the places where information is put. The main responsibilities are:

(1) The metadata can be forwarded to the state center and the branch centers for sharing. And object data are stored locally.

(2) It is capable of sending object data to user group and searching through the state center and the branch centers.

(3) We should develop local users, and enter the resource centers, the state center and the branch centers through local public networks.

(4) We should make efforts in the resource construction, expand and enrich the resources of digital library according to the standards of digital library.

(5). We should flexibly form computer systems, memory systems and supporting systems.

Part Two: The Basis of China's Digital Library

It may take about 10 years to complete the general architecture of China's digital library project. The main task is to establish the state center based on the National Library of China (NLC) and branch centers based on regions and mass information.

The basic requirements of branch centers of digital library are:

- Enough local users who can use telephone lines, cable TV and broadband IP networks to link with branch centers;
- The norm resource repositories can be used for these local users and served for users in the networks progressively.
- Sound basis for E-business with security and liability;
- Fair-sized investment scope used for the establishment and construction of networks and computer systems.

It is very important to analyze the markets of the local information users so that we can better organize resource construction and develop new functions. The accurate analysis of markets is the essential issue for digital libraries, which assume sole responsibility for its profits and losses. Digital library can also be built in information collection places that meet the demands.

1. The National Library of China's Preparation for the Construction of the State Center of Digital Library

In 1995, the National Library of China began to survey and study the international digital libraries. On October 2, 1998 Vice-premier Li Lanqing visited the National Library of China and pointed out that the second construction would be built into a digital one. After his visit the National Library of China has made great efforts in preparation for the construction.

(1) Basic Construction of Manifold Information

With the help of the leaders of the State Council, leaders of the Ministry of Culture and Beijing Municipal Government, great progress has been made.

- In February 1999, the library-wide network of the National Library of China was completed. Servers and computers were linked with the network;
- In October 1997, the National Library of China was connected with Beijing Telecom. T1 channel has been put into use. It has a capability of using four T1 channels;
- In March 1999, the General Bureau of Broadcast and Television connect 1,000M fiber optics to the National Library of China. This channel is built based on Ethernet. Its speed is 100 M/1,000 M in the coming few years. It will reach higher afterwards. It is possible for the connection between digital library center and resource centers;
- In April 1999, Beijing Cable TV Station connected 100 M/1,000 M fiber optics to the National Library of China. It is possible for the seamless links with Beijing Cable TV Station and resource centers and branch centers in Beijing area;
- In February 1999, Beijing Telecom completed the connection between the Branch Library of the National Library of China and the State Council;
- In April 1999, the network between Tsinghua University, Peking University and the Chinese Academy of Sciences was put into use;
- In May 1999, Zhongshan Library of Guangdong Province was linked to the National Library of China through the network of China Telecom.

Broadband IP Multi-network

CERnet

Peking University

Beijing Cable TV Network Tsinghua University CSTnet

Chinese Academy of Sciences

National Library of China

Network for the Ministries CCPnet

Television Network the State Council

Branch Library of NLC

CSTInet

2-4 Links between the Network of the National Library of China and the Major Networks in China

(2) Processing Center of Digitized Literature

- In March 1999, a processing center of digitized literature was set up in the National Library of China. A three-month's experiment was taken in processing and management circulation. We have accumulated some experience of resource

processing;

(3) Demonstration System of Digital Library

- In March 1999, we finished the development and integration of the full set of software and hardware systems in the demonstration environment of China's digital library;
- From January - March 1999 the National Library of China finished the test of SGML markup system, and completed the five resource repositories that the experimental environment needs. It laid a solid foundation for the use of markup language and data standards;
- In March 1999, we finished the test of cross-platform, cross-repository search and object transfer control, and completed the transfer control between the experimental system of digital library and object data of other systems.
- In May 1999, we used bibliographic data as metadata and full text as object to carry out an experiment of search and transfer between the National Library of China and Zhongshan Library of Guangdong Province.

(4) National Key Projects

China's Experimental Digital Library was set as one of the essential scientific projects in China in September 1997. This project is carried out among six libraries of China. In addition, since 1996, the National Library of China has carried through research that is relative to the projects of digital library.

(5) State 863 Project

Knowledge Network - Systematic Project of Digital Library is listed as one of State 863 Projects. We implemented comprehensive research in general design, intelligent agency and systematic integration in cooperation with Shuguang Company.

(6) Standardization

We have started research work in the making of digital library's standards and industrial standards.

(7) Analysis System of Users' Feedback

We have analyzed the information usage in the network of the National Library of China. So it is easier for us to put emphasis on some data processing.

2. Basic Preparation for the Construction of China's Digital Library

In May 1999, seven people of the National Library of China went to East China and Zhujiang Delta and made comprehensive technical investigations.

(1) Development of Network

Telecommunications network and cable TV networks have been set up, and broadband IP network is in establishment.

Methods that users' insertion can adopt:

a. Full two-way telephone line insertion. Its speed can reach 56K. If some adjustment is made in telecommunication network, seamless transfer with 56K can be got.
b. Cable Modem. Two-way insertion with 2M bandwidth can be realized through cable TV networks.
c. Broadband insertion by using broadband IP technology.
d. Packed insertion through the download of cable Modem and the upload of telephone Modem.

The prior choice is a, b, and d.

(2) Coordination of Computer System and Network

The coordination of computer system and network is one of the key problems. There is a big difference in different areas. We should resolve the problems with consideration of local technical strength and investors.

(3) Possibility of Using State Scientific and Research System

We should make use of research fruits from research institute and universities, and seek combination with scientific and educational system to solve important difficult technical problems.

(4) Possibility of Industrial Processing

We all realize that it is helpful to scan images first and then to use full texts.

(5) Coordination of User's Interface

It is very important to teach those who are non-computer professionals to use digital library. If we can realize standardization of information categories and key words, it will become easier to browse resources in digital library by using mouse and remotes.

(6) Coordination of Machine Tops

Network and television can become integration by using machine top technology.

(7) Collection Places

If there are suitable industrial environment and standards of the making of data, we can process data in a large scale. Most collection places have no ideas about the features of collected resources. It brings heavy difficulty in the development of metadata feature search.

(8) Feedback from Users

ISP and ICP attach great importance to feedback from users. But to information collection places, valid analysis system of feedback has not been set up. So it is hard to know what kind of information users care about. It may bring setbacks to resource processing.

(9) Establishing Marketing Mechanism with Interests Sharing

We must coordinate with organizations and companies. We should help each other, make mutual benefits and mutual investment under socialist market economy. It is of help to the establishment and promotion of digital library.

3. Basis for the Cooperation in China's Digital Library Project

(1) We should make use of the research fruits. There are many achievements in State 863 Project. They are important scientific basis for the construction of digital library.

(2) Many technologies in Internet can be used in the digital library project.

(3) There are different kinds of information in information-collected place so that there is great prerequisite for mutual resource construction.

(4) There is a potential in the construction of digital library because of potential markets and large number of user groups. Close cooperation between user groups and communities with powerful economic strength will also help the project.

Therefore, "Uniform standards and plans, coordination and cooperation, implementation step by step and common construction" is the guiding policy in the construction of China's digital library.

Part Three: Key Problems that China's Digital Library Project Need to Resolve

Problems in engineering and technology are:

1. Property of Digital Library Project

Digital library is one project with ordered organization of resources that information collection places carry out across the country under distributed computer systems, various network environment and cooperative software. The information of digital library object has the characteristics of "uniform standards and plans, common construction". Small amount of funds for the construction comes from the government. The large part comes from investors and funds collected by local areas.

2. Technical Routes

In order to speed up the process of this project, mature technologies and research fruits especially from 863 Project will be applied. For those regions where there are the basis and certain number of users, we will construct digital library network first. According to the construction process of metadata and object data, computers systems and memory systems will be added and network will be adjusted constantly. The useless resources will be stored on the stack. While the useful resources will stored in the machines. The general design of computer systems, network systems and resource construction is flexible and expandable.

3. Digitization Process of Literature

The important issue is to transfer the traditional media to digitized collection with the advantage of saving space. What kind of forms will we use to store digitized collections? How will we manage the literature processing? We must resolve these problems.

4. Markuping Process of Object Data

After the digitization of literature, how to markup data and how to markup industrially are the essential issues on the speed of processing of digital library's resource construction.

5. Search Engine for Mass Data

With regard to digital library technology, we need search tools for mass data. If there is a mass metadata base with millions of entries and each entry with 50 fields and 1.5K bytes, a powerful search tool must be needed to meet users' search requirements.

6. Application of Cable TV Networks

How to send information to innumerable households and families through cable TV network is an essential issue for digital library project. Machine tops and cable Modem technology is of importance to those who are non-computer professionals.

7. Cooperation among Information Collection Places

The main function of digital library is to organize resources based on certain subjects. So we need ordered organization and cooperation to use resources, and refine subjects.

8. Cross-Regions Network

The network construction provides us with an expressway. And digital library provides with goods that are transported on this expressway. The two supplement and complement each other. There would be no digital library without networks and network would not help at all without information resources.

9. Virtuous Circle of Funds

An important issue of constructing digital library in certain area is how to get funds and how to get a virtuous circle of funds. It is a good idea that we cooperate with those companies with rich experience, definite user groups and powerful funds.

10. Laws

In the operation of digital library, we may come across issues relating to intellectual property. In order to guarantee the normal operation of digital library, we must learn to resolve these issues and protect intellectual property and copyrights.

(Liu Xiaobai: Director, Information and Network Department, the National Library of China
Sun Wei: General Engineer, Information and Network Department, the National Library of China
Zhen Xihui: General Manager, Beijing Modern Wenjin IT Research Center)