面向数字图书馆的海量信息管理体系结构研究*

邢春晓+, 曾 春, 李 超, 周立柱

(清华大学 计算机科学与技术系,北京 100084)

A Study on Architecture of Massive Information Management for Digital Library

XING Chun-Xiao+, ZENG Chun, LI Chao, ZHOU Li-Zhu

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

+ Corresponding author: Phn: +86-10-62789150, E-mail: xingcx@mail.tsinghua.edu.cn, http://dbgroup.tsinghua.edu.cn/xingcx

Received 2002-09-19; Accepted 2002-12-24

Xing CX, Zeng C, Li C, Zhou LZ. A study on architecture of massive information management for digital library. *Journal of Software*, 2004,15(1):76~85.

http://www.jos.org.cn/1000-9825/15/76.htm

Abstract: This paper investigates the challenging issues and technologies in managing very large digital contents and collections, and gives an overview of the works and enabling technologies in the related areas. Based on the analysis and comparison of the related work, a novel architecture of massive information management for digital library is designed. The key components and core services are described in detail. Finally, a case study THADL (Tsinghua University architecture digital library) that complies with the architectural framework is presented.

Key words: digital library; architecture; massive information management; interoperability; metadata

摘 要: 分析了数据密集型应用的特点,讨论了管理海量数字资源面临的技术挑战和关键问题,并综述了支持高性能数据密集型应用的相关工作,包括标准、技术和应用系统.在分析和比较相关工作的基础上,设计了一个新型的面向海量信息管理的数字图书馆体系结构,并描述了其中的关键功能组件和核心服务模块.最后,给出了一个遵循该体系结构设计和实现的应用实例——清华大学建筑数字图书馆.

关键词: 数字图书馆;体系结构;海量信息管理;互操作;元数据

中图法分类号: TP311 **文献标识码**: A

* Supported by the National Natural Science Foundation of China under Grant No.60221120146 (国家自然科学基金); the National Grand Fundamental Research 973 Program of China under Grant No.G1999032704 (国家重点基础研究发展规划项目(973))

XING Chun-Xiao was born in 1967. He received his B.S. degree from Beijing University of Aeronautics and Astronautics in 1990, got his Ph.D. degree from Northwestern Polytechnical University in 1999, and finished his postdoctor work at Department of Computer Science and Technology, Tsinghua University in 2002. Now he is a professor and director at the WEB and Software Technology Research Center, Tsinghua University. He is the member of IEEE. His research interests include database, massive storage management, distributed multimedia systems, and digital library. **ZEGN Chun** was born in 1976. He is a Ph.D. candidate in Computer Science at the Tsinghua University. His research interests are database, personalization service, and digital library. **LI Chao** was born in 1978. She is a Ph.D. candidate in the Department of Computer Science and Technology, Tsinghua University. Her research interests include database, massive storage systems, and digital library. **ZHOU Li-Zhu** was born in 1947. He is a professor and doctoral supervisor at the Department of Computer Science and Technology, Tsinghua University. His current research areas are database, knowledge base, and digital library.

1 Introduction

In the recorded history of human being, the printed materials used to play a dominant role in the preservation and pervasion of human information and knowledge. However, with the rapid development of technologies in computer, communication, multimedia and storage, this role is giving away to the digital resources in the new era. The explosive growth of information in digital forms has posed challenges not only to traditional archives and their information providers, but also to organizations in the government, commercial and non-profit sectors. According to the latest report by Lyman and Varian[1], the world's total yearly production of print, film, optical, and magnetic content would require roughly 1.5 billion gigabytes of storage which is roughly 250 megabytes for every person on the earth. Printed documents of all kinds comprise only 0.003% of the total. Magnetic storage is by far the largest medium for storing information and is the most rapidly growing section, with a shipped hard drive capacity doubling every year. The types of digital resources are diverse. They include digital texts, documents, scientific data, images, animation, video, audio etc. The applications of the digital resources are quite broad, including DL (digital library), movie/video center, other public media (television, broadcast, newspaper, etc.), museum, and national or cooperative information center. At the same time the information highway, which is represented by Internet, has been an important tool of the pervasion of digital resources. The governments, companies, groups, research institutes, non-government organizations, education institutes all over the world put massive information on the Web.

1.1 Technology challenges and key issues

These massive digital resources present many challenging issues in data management technology area. The following are some examples. (1) Data model. Traditional data model theories are only applicable to structured data, but not for the massive digital resources of various types and they are mostly semi-structured or unstructured. Thus, new data models are demanded. (2) System architecture. Traditional database management systems are designed for business data processing featured by concurrent, short, and update transactions. Therefore transaction management and concurrent control remains as the center of system architecture. The architecture is not suitable for the management of digital resources as classical transaction concept is becoming less important in these resources. We need to pursue novel and universal frameworks for massive digital resources management. (3) Massive information storage. The volume of digital data resources is counted by terabytes or petabytes. Traditional storage devices using SCSI cannot work for efficient storage, online migration and persistent archive of such massive digital resources. So the research of multi-level storage systems, SAN (Storage Area Networks) and other technology are inevitable. (4) Organization and interoperation of massive information. To the various types of media, building the data organization model for the exchanging and sharing of information source, and developing the international standards for the metadata and digital object are both very important. The catalog and organization of digital resources based on the metadata standards is required. (5) Query processing. In traditional database systems, queries are expressed in query language such as SQL, but in the query and search of massive digital resources, many new mechanisms should be used, such as keyword search, full-text search, similarity query, and content-based multimedia retrieval. How to integrate the query methods (including SQL, OQL, and different XML query languages, e.g., XQL, XML-QL, XML-GL) efficiently to build an efficient and flexible query processing method has not been satisfactorily solved yet.

To solve the problems mentioned above will remain as a major goal to researchers in the next few years. To fulfill this end, we present a novel architecture for massive information management of digital resources in this paper. This architecture is intended to meet the requirements of managing digital resources characterized by distributed, dynamic, massive and heterogeneous properties. The other parts of this paper are organized as follows:

In Section 2 we review the related works and enabling technologies for supporting high performance data-intensive applications. In Section 3, we propose a novel architecture of massive information management for digital resource management, and briefly describe its key functional components and service components. In section 4, a case study based on THADL (Tsinghua University Architecture Digital Library) is introduced. The conclusion and the future work are given in Section 5.

2 Overview of the Related Work

The IEEE STD 610.12[2] defines architecture as the structure of components, their relationships, and the principles and guidelines governing their design and evolution over time. A wealth of previous work has addressed the research and development of architecture for digital library. Arms[3,4] described the key concepts in the architecture of a digital library, and presented an architecture for information in digital libraries, such as digital objects, metadata, handles system, and repositories. Baldonado[5] provided an infrastructure that affords interoperability among heterogeneous, autonomous digital library services based on an extensible metadata architecture. Liu[6] discussed the requirements of digital library applications based on the OAI (open archives initiative) and designed a scalable and reliable infrastructure by using the HTTP proxy, cache, gateway, and web service concepts. Lagoze[7] described the core components of the architecture for the NSDL (national science, mathematics, engineering, and technology education digital library). The design for a technical and organizational infrastructure has been formulated based on the first phase of the interoperability infrastructure, including the metadata repository, search and discovery services, rights management services, and user interface portal facilities. Wactlar[8] described the architecture of informedia system which provides a full-content search and retrieval of video and audio media, and implemented a fully automated process to enable daily content capture, information extraction and storage in on-line archives. We will give an overview of the related standards and enabling technologies as follows.

2.1 OAIS

In the past few years, a number of organizations and projects have adopted a reference model developed by the Consultative Committee for Space Data Systems called the OAIS (open archival information systems)[9] Reference Model, which has been adopted as ISO 14721:2002. Despite its origins in the space data community and its initial application to satellite and GIS data, the OAIS model has attracted a widespread interest in the international library and archive communities. The OAIS model is designed as a conceptual architecture framework for the understanding and increased awareness of archival concepts needed for a long term digital information preservation and access. It describes different long term preservation strategies and techniques and compares the data models of digital information preserved by archives for discussing how data models and the underlying information may change over time. The reference model addresses a full range of archival information preservation functions including ingest, archival storage, data management, access, and dissemination. However, the OAIS model is oriented to a situation where a particular type of data is closely aligned with a designated community and where the archives tends to form around the intersection of homogeneous data resources and users with very similar data needs. That this is not the situation in archives and research libraries has been cited as the major challenge to extending OAIS as a generic model. The OAIS composite of functional entities is showed in Fig.1.

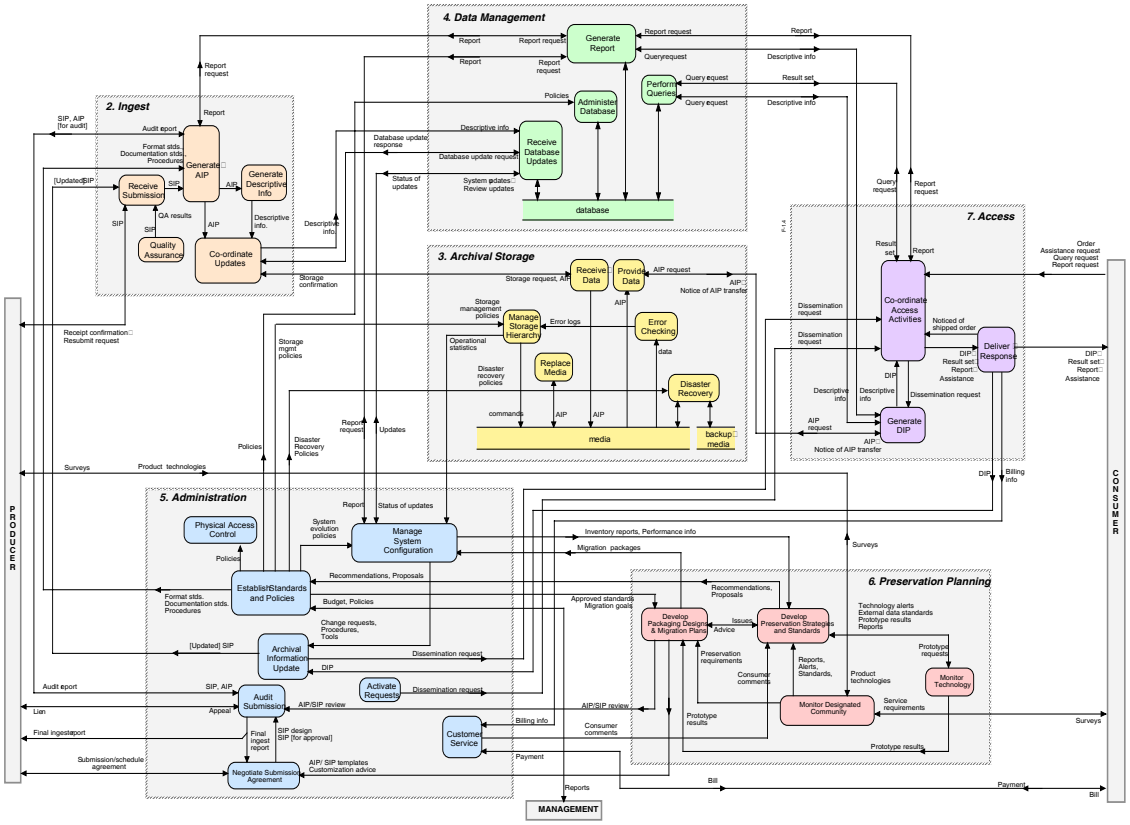

Fig.1 OAIS: Composite of functional entities

2.2 MSSRM

The IEEE Reference Model for OSSI (open storage systems interconnection)[10], previously known as the MSSRM (mass storage system reference model) Version 5, provides the framework for a series of standards for application and user interfaces to open storage systems. The OSSI Model identifies the high-level abstractions that underlie modern storage systems. It defines common terminology and concepts that allow the architectures of existing and future systems to be described and compared. The OSSI Model provides a conceptual and functional framework within which independent teams of experts may proceed with detailed OSSI interface definitions. But the OSSI model is not for the implementation specification, only a conceptual and functional framework. Its typical application is HPSS (high performance storage system). HPSS[11] is software that provides hierarchical storage management and services for very large storage environments. HPSS provides a scalable parallel storage system for highly parallel computers as well as traditional supercomputers and workstation clusters. HPSS is designed to use network-connected (as well as directly connected) storage devices to achieve high transfer rates. The design is based on IEEE Mass Storage System Reference Model, version 5. HPSS hopes to represent future scalability requirements that are very demanding in terms of total storage capacity, file sizes, data rates, number of objects stored, and number of users. HPSS is part of an open, distributed environment based on The Open Group's DCE (distributed computing environment) products that form the infrastructure of HPSS. HPSS is the result of a collaborative effort by leading US Government supercomputer laboratories and industry to address very real, very urgent high-end storage requirements. Scalability is in several dimensions: data transfer rate, storage size, number of name space objects, size of objects, and geographical distribution. Although developed to scale for order of magnitude

improvements, HPSS is a general-purpose storage system. However the architecture of HPSS still needs to be modified for the massive information management in digital library.

2.3 Digital library formal model——5S model

Digital libraries are complex information systems and therefore demand formal foundations lest development efforts diverge and interoperability suffers. Reference [12] proposed the fundamental abstractions of 5S (streams, structures, spaces, scenarios, and societies), which contribute to define digital libraries rigorously and usefully. Streams are sequences of abstract items used to describe static and dynamic content. Structures can be defined as labeled directed graphs, which impose organization. Spaces are sets of abstract items and operations on those sets that obey certain rules. Scenarios consist of sequences of events or actions that modify states of a computation in order to accomplish a functional requirement. Societies comprehend entities and the relationships between and among them. Together these abstractions relate and unify concepts, among others, of digital objects, metadata, collections, and services required to formalize and elucidate digital libraries. The formal model shows it can clearly and formally define a minimal digital library. But the formal model still needs to be improved and revised in digital library development.

2.4 Network storage technologies

The development of the Internet and the emergence of the data-intensive applications lead to the dramatically increasing demand for massive storage. The traditional storage method DAS (directly attached storage) can no longer meet the requirements of the massive storage information management to the storage subsystems, i.e. scalability, reliability, security, high-availability, and efficiency & virtualization of the management. By introducing the concept of networks, networked storage makes storage independent of servers or even communication networks and becomes the powerful substitution of the traditional storage method. There are three main networked storage methods nowadays: NAS (network attached storage), FC-SAN (storage area network based on fiber channel) and IP-SAN (storage area network based on IP storage technologies such as iSCSI).

NAS is developed based on the mature technologies, such as technologies of hard disk, RAID, network communication protocol, light-weighted file system firming, etc. As there are many good characteristics of NAS: low cost, easy to install, easy to employ, easy to manage, easy to scale and good reliability etc., NAS can be used to quickly solve the problem of relatively large-scale data sharing across the different platforms. However, NAS not only moves the bottleneck from between the servers and the storage devices to the communication networks, but also can merely be distributed in a limited physical scope. So NAS is not suitable for the application concerning massive data access, remote backup, and disaster recovery.

Different from NAS, FC-SAN[13] (see Fig.2) involves many new technologies, such as technologies of FC (Fiber Channel) hard disk, new protocols, HBAs (Host Bus Adapters), FC switches, FC hubs and bridges, etc. The management software of FC-SAN is complex, and its cost of FC-SAN is very expensive, including the hardware and software purchasing cost, and the installation and maintenance service cost. Besides, there are still some problems of interoperability and compatibility, as there is no a uniform standard in the SAN industry till now. But, FC-SAN has many advantages that NAS could not be compared with, i.e., high-speed data transmission, great scalability and flexibility, high-availability, centralized management, massive data access, LAN-Free Backup, remote mirror, and disaster recovery. So FC-SAN still takes up the high-end market of the networked storage.

If NAS is suitable for small-scale organizations and FC-SAN is suitable for large-scale organizations, then IP-SAN[14] (see Fig.3) is suitable for medium-scale organizations. From the aspect of technology, the protocol and iSCSI adapters are fire-new, but the IP switches, hubs and network management are all mature. Because of our familiarity with the deployment, configuration and management of IP network and the relatively low price of IP

infrastructure compared with that of FC, IP-SAN could be undertaken by such a kind of medium-scale organizations who have the infrastructure of Ethernet already and require the block-level instead of file-level data storage I/O. IP-SAN almost owns all the advantages of FC-SAN, but the performance of its products is not as excellent as FC-SAN now. We can expect its performance improvement in the future.

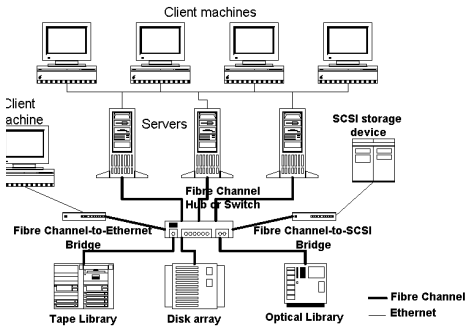

Fig.2 FC-SAN

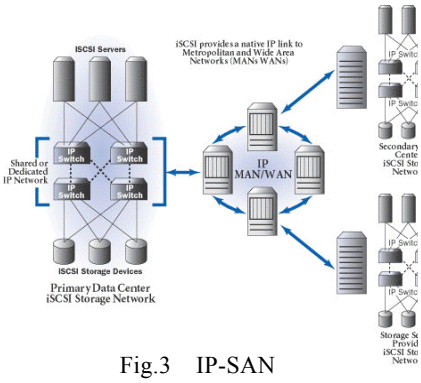

Fig.3 IP-SAN

3 Architecture of Massive Information Management for Digital Library

Until recently, there has been no common approach for architecture development and use in large-scale digital libraries construction. All kinds of libraries, research institutions and universities traditionally developed their DL architectures using techniques, vocabularies, and presentation schemes that suit their unique needs and purposes. In this section, we propose a new architecture of massive information management for digital library based on analyzing and researching the related works. We design the architecture by complying with related standards and making use of enabling technologies. The motivation is providing a common framework and software platform for constructing the large-scale digital library.

3.1 Design principles

The following set of principles for building architectures are critical to the objectives of the guidance. (1) Standardization. We will comply with related international and national standards in all layers of the architecture. (2) Componentization. Because it is widely accepted as a good software engineering practice, most modern programming environments adopt some form of component models. (3) Reusability. The technical infrastructure that provides functional specifications, paradigms of object-oriented programming, and component model has been particularly effective strategies for producing reusable and powerful software. (4) Scalability. The scalability of the proposed architecture is achieved by adopting a progressive and multi-layer framework, and providing the mechanisms for scaling services appropriately based on the service-demand and resource-availability. (5) Interoperability. Interoperability among heterogeneous systems and collections is a central theme. The different systems and collections have a wide variety of data types, metadata standards, protocols, authentication schemes, and business models. The goal of interoperability is to build coherent services for users by integrating components that are technically different and managed by different organizations. This requires agreements to cooperate at three levels: technical, content and organizational levels. (6) Architectures should be relatable, comparable, and scalable across DLs. This principle requires the use of common terms and definitions, such as metadata, digital object, repository, collection, and so on. This principle also requires that a common set of architectural building blocks is used as the basis for architecture descriptions. (7) Architectures should be adaptable, reliable, maintainable and testable. The ultimate achievement of this principle will reduce overall software costs, improve product and service

quality, help organizations to thrive, or even survive in our current and future turbulent times.

3.2 Architecture design

We design a multi-layer architecture to support the service and management in DL (digital library) based on complying with a set of design principles above. The architecture is a more OSI-like reference model, but it focus on the massive information service, management, and archival (see Fig.4).

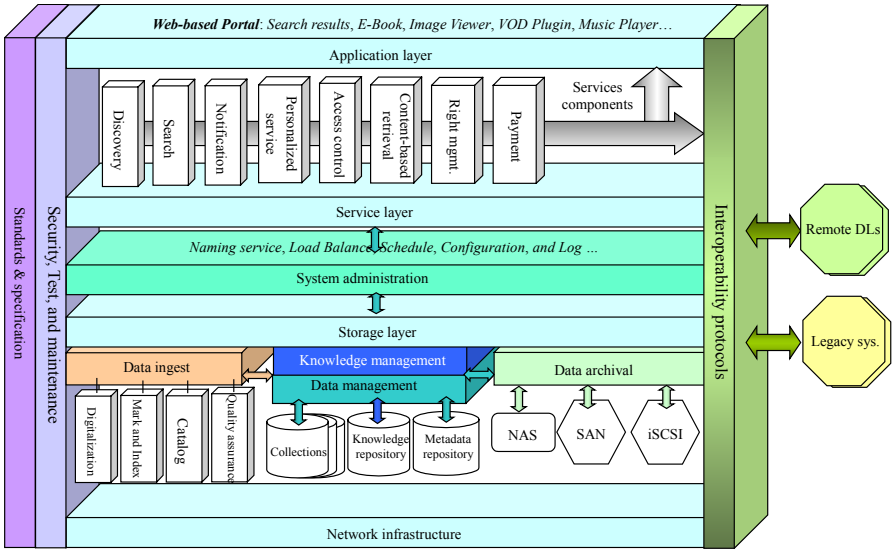

Fig.4 An architecture of massive information management for DLs

(1) Network Infrastructure provides the fundamental interconnecting and communicating services and functions. (2) Storage layer consists of data ingest, data management, knowledge management, and data archival. (3) System administration layer provides the services and functions for the overall operation of the digital library system. Administration functions include naming service, load balance, schedule policy, and configuration management of system hardware and software. It also provides system engineering functions to monitor and improve system operations and performance. It is also responsible for establishing and maintaining each layer's standards and policies. (4) Service layer provides major service components including discovery, search, content-based retrieval, personalized service, notification, access control, right management, and payment. (5) Application layer uses Web-based user interface to provide all kinds of related services to users friendly by using E-book, image viewer, VOD plugin, and music player. The Web-based Portals provides an access to the digital library's collections and is optimized for different users and different purposes effectively. (6) Interoperability protocols: The vertical interoperability protocols is responsible for not only the internal information communication of digital library, but also the remote digital libraries and legacy system. The technologies of middleware, agent, and distributed object will be used to establish all kinds of interoperable protocols. Nowadays, metadata-based interoperable protocol such as Dienst, SDLIP, and OAI has been designed and used Z39.50. We are trying to reach a compromise between a full-scale, all encompassing search middleware design such as Z39.50, OAI and a web-based search engine for designing a lightweight efficient digital library interoperability protocol by analyzing previous search middleware designs and enabling technologies. (7) Security, test, and maintenance: ensure that the architecture is open, robust, maintainable and testable. It will reduce overall software costs and improve system performance and quality. (8) Standards and specifications: In order to ensure system openness, scalability, and interoperability, we will comply with all kinds of related standards and specifications to design the system

architecture, including: Metadata: Dublin Core 1.1, USMARC, CNMARC, XML/RDF; Distributed object and middleware standards: CORBA, J2EE, .NET; interoperable protocol: Z39.50, OAI, SDLIP; Networking storage: NAS, SAN, IP-SAN; Mass storage reference model: OAIS, MSSRM; Web service protocol: UDDI, WSDL; Data exchanging and encoding: SGML, HTML, and XML.

3.3 Key functional components

(1) Data ingest. This component provides the services and functions to accept resource digitalizing, marking, indexing, and cataloging from producers, and prepares the contents for data management and knowledge management. Data ingest function also performs quality assurance on metadata and digital objects, which complies with the related data formatting and documentation standards. Meanwhile Data Ingest also gathers and catalogs new metadata and digital objects from remote DLs and WWW by using interoperability protocols.

(2) Data management. The component provides the services and functions for storing, maintaining, and accessing both metadata repositories and collections from data ingest. Data Management functions include administering the metadata database, multimedia databases and file systems. It will maintain schema, view definitions and referential integrity, and perform database updates (loading new descriptive information or administrative data). The central storage of metadata is provided by data ingest and metadata is gathered from other DLs by interoperability protocols. By using output interfaces between layers, it will provide data services to service layer, such as search and browse services. Meanwhile, object server (such as video server) that is managed by data management provides all kinds of multimedia services such as VOD. Data is typically stored as digital objects in file systems or as blobs in object-relational or object-oriented databases. Descriptive metadata is typically stored as attributes in metadata databases that complies with metadata standards, such as Dublin Core, USMARC, CNMARC. The metadata attributes use the XML (eXtensible markup language) to label the information content and a DTD (document type definition) to build a semi-structured representation of the information.

(3) Knowledge management. Knowledge is represented as sets of relationships of domain concepts that are expressed as rules used within inference engines. This component provides mechanisms for managing all of these types of relationships. Every discipline deals with multiple concept spaces that describe relationships between physical variables, data sets, collections, applications, and domain knowledge. These concept spaces provide a hierarchy of levels of the implied knowledge based on ontology. Ontology[15] is the key technology used to describe the semantics of information exchange, which is defined as specifications of a shared conceptualization of a particular domain. They provide a shared and common understanding of the domain that can be communicated across people and application systems, and thus facilitate knowledge sharing and reuse. The concepts and their relationships as knowledge rules are stored in knowledge repository. The components will provide services for the manipulation of specific applications.

(4) Data archival. This component provides the services and functions for the storage, maintenance and retrieval of metadata, digital objects, and knowledge rule for the long-term preservation. The functions include receiving the data need to be archived from data ingest and data management components, and adding them to the permanent storage, managing the storage hierarchy, refreshing the media on which archive holdings are stored, performing the routine and special error checking, and providing the data migration, replication, backup, and disaster recovery capabilities. The system will adopt different networked storage methods (such as NAS, FC-SAN and IP-SAN) according to different system scale and application requirements. The data archival standard OAIS reference model will be used in the architecture for the understanding and increased awareness of archival concepts needed for the long term digital information preservation and access. Meanwhile, we extend the model and make it suitable for archival requirements of digital library.

3.4 Core services components

(1) Discovery and search: The components are to provide fundamental capabilities for locating and finding resources and collections among the distributed digital libraries. The challenge is to encourage end-user resource discovery and information use in a variety of formats, from a number of local and remote sources, and in a seamlessly integrated way. The architecture uses metadata for resource discovery, and uses a keyword indexing search-engine for resource discovery as a complementary method. The standards and specifications for resource discovery include Z39.50, OAI, and SDARTS. The key techniques include multi-agent and middleware technologies that are based on metadata including USMARC, CNMARC, and Dublin Core Element Set 1.0.

(2) Content-Based retrieval: The components are to provide fundamental capabilities for query multimedia resources and collections, including text, image, video and music. The content-based image retrieval allows for image queries based on image examples, feature specifications, and primitive text-based search. Content-based video analysis, automatic video index, summarization, and relevant feedback are used for the video retrieval.

(3) Personalized service and notification: To quickly and easily gather useful information and knowledge to alleviate information overload problem, it has therefore become necessary to provide users with active and adaptive service mechanisms that automatically extract only relevant incoming documents. The component is able to provide the users with a personalized filtering and notification service based on user modeling and profile learning. We design the component that relies on many established techniques applied in IR (information retrieval), IF (information filtering), user modeling and machine learning, etc.

(4) Access control: Issues of intellectual property need to restrict access to objects or services in the digital library. Access management services are intended to provide a single mechanism to be shared by many systems. There are two key functions of the service: authentication and user profile management. Authentication service requests information from the user and indicates to applications that the identity provided is likely to be accurate. Authentication is currently implemented using Password/ID, restricted IP, and CA (certifying authority).

(5) Right management and payment: The first-generation of DRM (digital rights management) focuses on security and encryption as a means of solving the issue of unauthorized copying. That is, lock the content and limit its distribution to only those who pay. The second-generation of DRM covers the description, identification, trading, protection, monitoring and tracking of all forms of right usages over both tangible and intangible assets, including management of right holder relationships. The DRM component defines the roles and behavior of a number of cooperating and interoperating modules under the three areas of IP (intellectual property): Asset creation, management, and usage. Ideally, these modules would be engineered as components to enable systems to be built in a modular fashion. However, this implies a set of common and standard interfaces/protocols between the modules that does not yet exist. As DRM matures, the industry will move towards such standardization, for example, XrML[16], PDRL (open digital rights language)[17], and OEBF[18].

4 Case Study: THADL

Tsinghua University Architecture Digital Library (THADL) is developed as a prototype system, which started formally since March 2000. THADL maintains a balance between technology-focused research and content-based research. The project team brings together three research groups from different disciplines including computer science, architecture science, and library information management, and represents a substantial cooperation among computer researchers, librarians, and subject specialists. The large amount of multimedia materials in THADL repositories include papers, journals, photographs, manuscripts, drawings/blueprints, animation, video, and audio on Chinese ancient architecture. We have finished following the goals of THADL. By designing and developing

THADL prototype, we analyze and summarize the previous architecture (see Section 3). Meanwhile the architecture also guides the improvement and extension of system functions and services. The detail of THADL has already been described in Ref.[19]. The main research contents of THADL are as follows: (1) Exploring the efficient methods and technologies by analyzing, designing, and evaluating the THADL prototype system to pave the way for constructing a future large-scale digital library; (2) Building a Chinese architecture digital library that provides an intelligent, interactive, and collaborative learning environment on the Internet rather than the static digital resource repositories; (3) Presenting an efficient method to digitalize, index, and preserve most kinds of materials for Chinese architecture study; (4) Establishing metadata specifications and standards and their related issues for Chinese architecture science; (5) Supporting friendly, active, and personalized services for different users including students, scholars, librarians, and ordinary users on the Internet. However, THADL is a medium but comprehensive prototype system, and we have a long way to go for researching, testing, and analyzing whether our proposed architecture is suitable for the large scale digital library application such as the China Digital Library Project.

5 Conclusion and Future Work

In this paper we discuss the challenging issues and technologies in managing very large digital contents and collections, and give an overview of the related works and enabling technologies for supporting high performance data-intensive applications. We design a novel architecture of massive information management for digital library, and describe the key components and core services. Finally, the case study——THADL (Tsinghua University Architecture Digital Library) is given according to the architectural framework.

In the future work, we will study and develop software middleware for massive storage management, XML based search engine, and multilingual full-text search. We will design and implement a lightweight interoperable protocol based on the tailed Z39.50 protocol for supporting the large-scale distributed heterogeneous digital resources integration.

References:

[1] Lyman P, Varian H. How much information. 2000. University of California. http://www.sims.berkeley.edu/research/projects/how-much-info/how-much-info.pdf
[2] IEEE Standard Glossary of Software Engineering Terminology. http://standards.ieee.org/reading/ieee/std_public/description/se/610.12-1990_desc.html
[3] Arms WY. Key concepts in the architecture of the digital library. D-Lib Magazine, 1995. http://www.dlib.org/dlib/July95/07arms.html
[4] Arms WY, Blanchi C, Overly EA. An architecture for information in digital libraries. D-Lib Magazine, 1997. http://www.dlib.org/dlib/february97/cnri/02arms1.html
[5] Baldonado M, Chang CK, Gravano L, Paepcke A. The Stanford digital library metadata architecture. International Journal on Digital Libraries, 1997,1(2):108~121.
[6] Liu XM, Brody T, Harnad S, Carr L, Maly K, Zubair M, Nelson M. A scalable architecture for harvest-based digital libraries: The ODU/Southampton experiments. D-Lib Magazine, 2002,8(11). http://www.dlib.org/dlib/november02/liu/11liu.html
[7] Lagoze C, Hoehn W, Millman D. Core services in the architecture of the national digital library for science education (NSDL). In: Proc. of the 2nd ACM/IEEE-CS Joint Conf. on Digital Library. Portland: ACM Press, 2002. 58~65.
[8] Wactlar H. Multi-Document summarization and visualization in the informedia digital video library. In: Proc. of the 12th New Information Technology Conf. Beijing: Tsinghua University Press, 2001. 323~332.
[9] ISO 14721: 2002. Reference model for an open archival information system (OAIS), 2002. http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf
[10] Mass storage system reference model. IEEE Project 1244. http://www.ssswg.org/MSSRM.html
[11] HPSS: high performance storage system. http://www.sdsc.edu/hpss/hpss1.html
[12] Gonçalves MA, Fox EA, Watson LT, Kipp NA. Streams, structures, spaces, scenarios, societies (5S): A formal model for digital libraries. Technical Report, TR-01-12, Virginia Tech, 2001.
[13] Thornburgh RH, Schoenborn BJ. Storage Area Networks: Designing and Implementing a Mass Storage System. Prentice Hall, 2000.
[14] Clark T. IP SANS: An Introduction to iSCSI, iFCP, and FCIP Protocols for Storage Area Networks. Addison-Wesley, 2001.
[15] Staab S, Studer R, Schnurr HP, Sure Y. Knowledge processes and ontologies. IEEE Intelligent Systems, 2001,16(1):26~34.
[16] XrML 2.0. http://www.xrml.org/
[17] Open Digital Rights Language. http://odrl.net/
[18] OeBF-the Open eBook Forum. http://www.openebook.org/
[19] Xing CX, Wu KH, Luo DY, Zhou LZ, Liu GL, Qin YG. THADL: A digital library for Chinese ancient architecture study. In: Proc. of the 12th Int'l. Conf. on New Information Technology. Beijing: Tsinghua University Press, 2001. 373~382.