Research on Application of Digital Literature Archives Management Based on XML Database System

Yi Yang
The Archives of Jilin Jianzhu University
Changchun
China
yangydei@163.com

Journal of Digital Information Management

ABSTRACT: *Document files are history that records that human understanding and transformation of the world that formed directly in the social practice, is a precious historical and cultural wealth. International advanced archives service work are positioning on the basis of socialization, based on a highly efficient and reliable digital platform, to realize centralized management of the national archives of the and the social sharing. However, judging from the domestic situation, comprehensive archives management system is still blank, for documents and archives, the most important and most valuable social resources are only just beginning. This research purpose of this subject is to build a documents digitization management platform that can comprehensive collect and manage social archives document resources, using uniform standards and with high scalability.*

Subject Categories and Descriptors
H.2.1 [Logical Design]: Platform design; **H.2.7 [Database Administration]:** Digital management

General Terms: Architectural Design, Design Management System

Keywords: XML Database, Documents Files, Platform Design

Received: 4 June 2015, Revised 3 July 2015, Accepted 10 July 2015

1. Introduction

Document archive resources are inexhaustible knowledge of human society that the international community attaches great importance to the information resources construction and mining. Today's Internet and database technology continues to mature, form various types of specialized social system and data analysis system through data warehousing, data mining, knowledge management and other high-tech application, will plays an important role in the development of science and economy[1].Currently, international advanced national literature of using service work are positioning on the basis of socialization, and from the perspective of the status quo of domestic, integrated of literature archives management system is also blank currently, existing literature management system mainly is library management system, for documents and archives, the most important and most valuable social resources are only just beginning, not mentioned there has technology mature breakthrough, existing literature management system is mainly using traditional RDBMS as data storage technology [2], do basic digital work for this unit-oriented literature data, and establish business layer over it, the support for preparation of the structured/semi-structured documents, distributed storage and information is very limited. In cases of domestic, we also found some projects that use XML DB technology. Some of these commercial projects have been successfully implemented, but its products are not very well expand the depth and breadth of technology, it failed to perform advantages of XML DB such as in dealing with structured data, nor extend them towards the direction of integrated interactive platforms, raise the level of applications in the field [3]. Meanwhile, there are also some cases use a NXD (native XML database) technology to applied to some small technical documents and Archives Centre pilot, are exploring ways to XML database in the field of digital documentation in-depth reference and try to address the many challenging problems of its [4].

This paper based on the actual situation of domestic archives management and the difficulty, on the basis of the comparative analysis of heterogeneous data management and integration technology at home and abroad [5, 6], put forward the archives integrated management method that take XML database system as the main technical route, and then studied the NXD and XED the two mainstream XML database technology, based on the current social development subject, build all kinds of special knowledge base, providing social service environment, and for the purpose of serving the public, fully tapped the value of archives information.

2. Overview of XED and NXD- Related Technologies

2.1 NXD Technology

NXD (Native XML Database) is called the native XML database, also known as plain XML or source XML database, is designed according to the characteristics of XML data, used to store and manage XML documents in the database. Working with XML data in a natural way, like rows in the table in a relational database as basic logical units of storage, XML documents is the basic logical storage unit of NXD [7].A standard NXD, shall conform to the following characteristics:

(1) For the data in an XML document, define the logical model of XML documents, and according to the model to store and retrieve documents. Such models should at least include elements, attributes, PCDATA and document order.

(2) Like rows in the table in a relational database as basic logical units of storage, NXDBMS uses XML documents as the basic logical storage unit of NXD

(3) Does not require any particular underlying physical storage model, which could be based on a relational, hierarchical, or object-oriented database, or use the index file, compressed file such a specialized storage format.

2.2 XED Technology

XED (an XML Enabled database) is on the basis of the original database to extend XML support module, complete the format conversion and transmission between XML data and database [8]. In order to support some of the W3C XML standard operation, XED providing some new primitives on the basis of relational database (such as Oracle 9 ir2 adds some data packets to manipulate XML data, etc.), and optimize the XML processing module.

XED extension at the function structure of relational database provides high performance XML storage and retrieval technology. It not only offers a complete, efficient and sophisticated all the function provide by relational database, it also provides XML processing function associated with a native XML database.

Because XED database was extended on the basis of the relational database, different databases are different in function and principle, often a complete XED provides the following basic functions:

(1) Schemas the standard W3C XML data into the database management.

(2) Provide a native XML data type for conservation and management of XML document.

(3) Provide a set of methods and operators that allowed operating on XML content.

(4) Provide a standard method to access and update XML, such as XPath.

(5) Provides several standards API develop structures for programmatic access and manipulation.

(6) Provides for XML-specific memory management and optimizations

(7) Provides standard database functionality, such as transaction processing control, data integrity, reliability, security and scalability.

2.3 Comparison of NXD and XED Technology

Both NXD and XED can manage document-centric documents, also can manage a data-centric documents. But emphasis and technical advantages of the two are different, generally speaking, XED excel in management of data-centric documents, NXD better in document-centric documents management. Main technology comparison of NXD and XED are shown as table 1:

3. Platform Architecture and Technical Design

3.1 XML database Platform Selection

XML database management system is the platform's core technical support, targeted to choose an appropriate XML database management system is one of the key technical problem of architectural design [9].According to the data digitization and management requirements of literature archives management platform, the platform oriented object is the social public, therefore, maturity, reliability and performance of database management system is the most important factors for platform to choose from XED and NXD. We set the following reference index as a criterion for evaluation, to select database management systems:

(1) The efficiency structured and semi-structured data storage and query

(2) XML query methods and properties

(3) Reliability and stability of database management system

(4) The friendly development and using an interface

(5) Good maintainability

	NXD data base	XED data base
Data storage	Description of semi-structured data with hierarchical structure, the elements can be nested	Structured data that describes the rules, two dimensional tables, each of which is no longer subject to the basic data item
Seqentialization	Orderly, adjust order between element, attributes, PDDATA are not allowed to	The order of the fields can be changed
Data exchange capabilities	strong	Middleware or additional software is required to support
Other	Element names can be operational	Allows for data manipulation

Table 1. Main technology comparison of NXD and XED

Comparison direction	Compare projects	Oracle	Tamino
Data storage	Structured XML storage	Relatively excellent	excellent
	Semi-structured XML storage	common	excellent
Data query method	XML Fragment queries	excellent	excellent
	support XPath	yes	yes
	support XQuery	11g support	yes
Database performance	Index performance	strong	common
	Transaction capabilities	strong	common
	Performance of large amounts of data	strong	unknown
reliability	Database cluster	yes	no
others	Development interface	excellent	excellent
	Related tools	more	less
	maintainability	excellent	common

Table 2. Database features comparison

During the process of comparison, we choose Oracle database that has leading-edge technology in XED, and Tamino Server in matured commercialization NXD database Software AG as primary comparators, the comparative results are shown in table 2:

Combined with the actual analysis of the above data, we have simulated testing the Oracle XML DB data, Oracle can better performances of completed 500~800 million XML records' storage and query, to meet the platform's functionality and performance requirements. Combining these conditions into account we finally determine using Oracle as a underlying database platform.

3.2 System Architecture Design of Platform

According to the requirements analysis and selection of XML databases, we present overall design for the literature archives digitization management platform, platform architecture as shown in Figure 2.1:

(1) Software application platform uses J2EE framework for development, using component-based development technology to development. Platform uses MVC (Model View Controller) pattern, to separate the input, processing

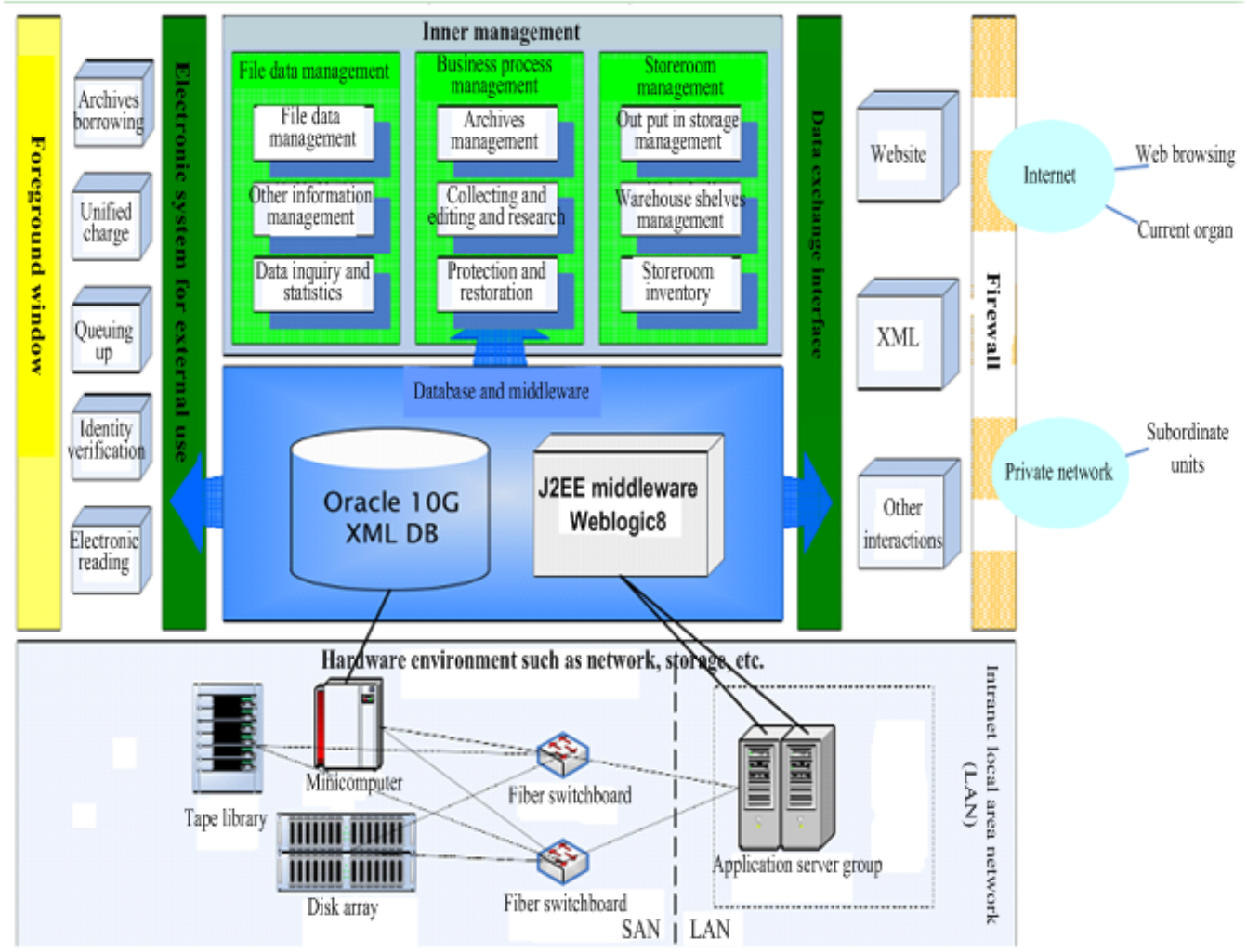

Figure 2.1 Platform architecture

and output of platform, according to the Model, View, Controller.

(2) Database and Middleware

Platforms using Oracle XML database as the core database software. Taking the whole platform's data load situation into account and database performance's requirements for processing, the database server uses minicomputer and installs SUN SOLARIS 9.0, database management software using Oracle 10G for Solaris to provides data storage services for the whole platform.

(3) Platform Storage

Platforms uses SAN architecture, provide network storage space through disk arrays. Platform takes fully redundant design, support for snapshot and clone technology to achieve a higher data security, can satisfy the need for future expansion of the storage capacity.

(4) Internal LAN

Network architecture is made of two layers of the star structure; core adopts high-performance gigabit switches, support gigabit exchange structure. Core switch is one o fthe most important switching equipment, the use of double machine method to eliminate single point failure of the core equipment. Access layer switches, with functions of high speed forward, support gigabit linked, through dual gigabit link linked to the two core switches, to eliminate the link-single point of failure.

(5) The Operating System

Platform based on different server, select different operating system. In order to provide better database performance, database server using SUN SOLARIS 9.0 version. Other services such as application servers, file servers and other PC server equipment all used Microsoft Windows Server 2003, the Enterprise Edition operating system, to ensure the stability, reliability and good fault tolerance of the platform.

3.3 Quality Design of Platform

(1) Platform Scalability Design

Platform adopts object-oriented component-based design pattern, in order to adapt to changing business processes, improving the maintainability and scalability of the project after the implementation, platform business logic and application functionality, and vertical and horizontal divided into components. In terms of development, we use a three-tier architecture, presentation, application logic, and data tiers, the client does not have application logic (thin client),

all application logic are in the middle tier, the middle-tier uses application server and middleware technologies, ensuring scalability and extensibility of the system. While platform uses open host system and network systems to ensure expansion of hardware systems, upgrades convenient, simple, and improved scalability.

(2) Security Design of Platform

Platform's design redundancy and disaster preparedness strategy for the storage, in response to disasters such as equipment failure. According to the different serious degree of disaster, the platform can be immediately recovered system and data from a single point of failure or manually. In order to realize the software security requirements of platform, the design realized the following considerations, respectively is: identity, authorization, data confidentiality and no denial behavior.

4. Operation and Evaluation of the Platform

4.1 Platform Tests

In the acceptance process of platform, do the third-party testing of professional bodies, in addition to a full range of functional tests; especially do the strict performance tests of the system, in order to ensure the stability and reliability of the platform. In performance tests, we mainly aimed at the platform concurrency and information retrieval response time. Testing tools for concurrency is Spirent's Avalanche220 network application-level simulation and performance testing system and related software; Information performance test tools is HP's Mercury Load Runner performance testing software.

(1) Concurrency Testing

Test uses Avalanche 220 to simulate single users and 100 concurrent users to access platforms. In the test, the step size is set to 10, each step size in 5 seconds by10 user access and lasts 10 seconds until 100 user location. Keep 100 concurrent users access 10 seconds, and then decreasing user to zero. The test results are shown in table 3.

Concurrency users (unit: a)	1	100
Concurrency users (unit: a)	0.019	0.372

Table 3. Result of concurrency testing

(2) The Response Time Test

We use Mercury Load Runner software test platform for various search conditions response time indicator, test result (part) as shown in table 4.

Platform validated the third-party testing, performance stability, able to withstand high pressures, and be able to guarantee a high response time in the case of large amount of data. Meanwhile, software interface specifications, good Chinese compliance, meet the requirements of software registration test and usage of platform.

Search condition	Record number (item)	Response time (seconds)
D4.7_blank	818455	14.337
D4.7_ File number:1999-	48012	6.344
D1.12_The County number: Sichuan Fu syndrome (92)	1199	1.442
Marriage records _blank	282704	12.208
Marriage record _woman's name Lin Haidong	1	1.82

Table 4. result of response time test

4.2 Platform Operation Situation

After acceptance of the platform, in the archives of Pudong New District, it's running in Shanghai pudong new archives, and fully used in the operational work of the various departments within the archives in Shanghai Museum up, meanwhile, the data exchange interface has been formed between unicom with external systems and subordinate unit.

So far, the platform database has the files of archives records of more than 2 million, more than 12 million archive data and relevant data, in addition, there are thousands of audio-visual archives information data, and so on characteristics information. In the case of the current file data, the platform for all files in the data are combined query response time within 15 s, common query response time can keep within 5 s.

5. Conclusions

In order to change currently situation that domestic literature archives lack efficient concentrated management [10], this paper research development status of XML database technology both at home and abroad, proposed solution that established a concentrated platform that achieve literature Archives Digital management through XML technology, and used Oracle XED database technology, develop a digital management platform that adapt to domestic literature archives integrated management, achieved the full collection and management of mass social literature resources. This platform's development and implement can solve the the problems that difficulty concentrating problems and resource management of domestic dossier, and difficulty to provide public services, and offer constructive ideas for future

electronic archives data center.

References

[1] Li, Cuiping., (2011). Present situation, problems and development of digital archives construction in colleges and universities. *Cultural and Educational Information*, 07:182-183.

[2] Lv, Bangzhen., Hu, Ying (2010). Digitization management strategies of Minority archives in Yunnan Province. *Archival Science Communications*, 02. 45-48.

[3] Weidong, Yang., Bole, Shi (2009). Overview of XML workflow management research. *Computer Research and Development,* (10) 11-12.

[4] Ma, Xiaoxing (2012). Overview of distributed Web server technology. *Journal of Computer Science*, (1) 7-12.

[5] Liao, Xiaoping., Wang, Zhijian., Liu, Shan (2008). Improvement of filtering algorithm based on XML in publish/subscribe model system. *Computer Development and Application,* (12) 7-8.

[6] Cai, Junren., Yu, Jianjia (2010). An improved algorithm of XML data stream queries based on YFilter. J*ournal of Fuzhou University* (natural science Edition), (06) 10-12.

[7] Yin, Guisheng., Shen, Jie., Xie, Xiaoqin (2011) Improve of XML data filtering algorithm in the automatic machine. *Journal of Harbin Engineering University*, 45, 46 (03).

[8] Li, Jing., Zhuang, Chengsan. (2010). Use XML to the database query. *Journal of Computer Applications*, (10) 28 and 29, .

[9] Liu, Wansuo., Guan, Shuiwen(2010) Using Web database to develop Internet-based distance teaching system. *Distance Education in China*, (12) 7-12 .

[10] Cong, Yu., Jagadish, H. V. (2008). XML schema refinement through redundancydetection and normalization. *The VLDB Journal.* (2) .