Architecture for Personal Digital Library

Shihn-Yuarn Chen, Chia-Ning Chang
Dept. of Computer and Information Science
National Chiao Tung University
1001 Ta Hsueh Rd., Hsinchu City
Taiwan

Ming-Jin Hwang, Hao-Ren Ke
University Library
National Chiao Tung University
1001 Ta Hsueh Rd., Hsinchu City
Taiwan

Wei-Pang Yang
Dept. of Information Management
National Dong-Hwa University
1, Sec. 2, Da Hsueh Rd., Shou-Feng, Hualien
Taiwan
Dept. of Computer and Information Science
National Chiao Tung University
1001 Ta Hsueh Rd., Hsinchu City
Taiwan

Abstract: - The research of digital library and content management is in progress for many years, and many great results have been achieved. Besides, many value-added applications have been built from the dot com mania, and they really improve the Internet environment and the convenience of our lives. However, there is still lack of a better way to preserve, manage, search and share personal digital content. In this article, a personal digital library architecture is proposed to solve the issues of the search mechanism of traditional file systems, and work as a foundation of value-added applications for content sharing.

Key-Words: - XML, Personal Digital Library, Metadata, Value-Added Applications

1 Introduction

In the past years, many efforts have been invested in the research of digital library, and many achievements are gained. Owing to this, people can acquire a great deal of integrated knowledge from, to name a few, e-books, e-journals, and on-line training courses.

However, with the blooming growth of equipments for capturing digital content, such as digital cameras and digital recorders, massive digital content floods into everyone's daily life and stuffs into hard drives and compact discs. With the traditional file system and search mechanism, it is hard to discover one specific digital datum, organize digital content, and establish the relationships between pieces of digital content.

Besides, students may take notes on lecture handouts or their notebooks, and then digitize these paper materials for exchange or long-term preservation. They also own a lot of digital content during their education, such as computer programs, reports, and video clips. Researchers may search materials on the Internet, digital libraries, or books, and then many web pages, journal or conference papers, and reading notes are stored. Finding out a report or a conference paper related to a specific topic usually takes much time. Furthermore, a report may relate to multiple research topics, it is also a issue to arrange the report in a proper directory.

Obviously, the traditional file system is not enough to solve the above issues, including preservation, cataloging, organization, and search of personal digital content. Thus, architecture for personal digital library is proposed. With this architecture, users can manage their digital content in a better manner, establish relationships for better search mechanism, construct new content, and share their content and knowledge with others.

2 Related Research Domain

2.1 Content Management

Content management is the processes and workflows involved in collecting, organizing, managing, and publishing information resources, and usually realized by a content management system (CMS). Many CMSs have been developed, including Zope [15] and DSpace [16]. In addition to collecting, organizing, managing, and publishing, many progressive researches of content management focus on revision, indexing, searching and access control.

2.2 Digital Library

Governments, organizations and universities exert many efforts on the digital library domain. Some researches and projects [4][5][6][7][8] attempt to achieve better utilization of library by extending the traditional library to the electronic library. Others [1][2][3] adopt many technologies like data mining, information retrieval, image processing, personalized services, and text summarization, to integrate information, discover knowledge, provide better user interface, etc.

2.3 Internet Value-Added Services

From the age of dot com mania, people get online because of the attraction of various value-added services, such as portals, Web Email, online shops. Although there is a decline, new services and improved services turn up, such as online albums, blogs.

These value-added services [17][18][19][20] help people to easily access Internet, preserve their digital content, and share their digital content with others.

3 Design of Personal Digital Library

In addition to the functions of the traditional file system, a personal digital library should provide functions similar to a digital library, a content management system, a portal, and a file distribution service. The following subsections explicate the design of a personal digital library, including preservation, cataloging, organization, search, packaging, distribution, authoring, version control, legacy system integration, and open APIs.

3.1 Digital Content Preservation and Cataloging

Digital content preservation and management is the basis of a digital library, as well as a personal digital library. The main task of preservation in a traditional digital library is digitizing, include the metadata of an art-work, the content of an article, the image and virtual reality video of a sculpture. However, for a personal digital library, the material is limited to personal digital content, such as documents, digital photos, bookmarks, and video clips, but even then preservation and catalogue is necessary.

Different to a traditional digital library, a personal digital library should allow the user to define his own categories or provide some predefined and recommended categories. Digital content can be placed into a “File Space” and assigned to one or more categories, instead of the traditional file system.

3.2 Organization, Metadata and Search Mechanism

In the traditional file system, a hierarchical folder structure and symbolic links are used to store files. A file can be put into a proper folder. For example, one user can put a photo of the White House into a folder named “USA”. The user can also use symbolic links to establish the relationship between a file and others; for example, the White House is a building, so a symbolic link of the photo of the White House can be created in the “Building” folder.

If the user wants to create an album disc of “buildings”, he/she can write all the files in the “Building” folder into the disc. However, the photo of the White House is not included, because only the symbolic link exists in the “Building” folder. In other words, a symbolic link can make up the lack of relationship in the traditional file system, but it is not enough to solve the lack of relationship in semantic level. Similar examples are too numerous to enumerate; for example, a pdf file of a journal paper may relate to a research project and a course material.

Thus, a personal digital library should allow a user to assign a piece of his digital content to one or more categories. Besides, the implicit metadata of a piece of digital content, such as the EXIF

information of a digital photo, should be automatically extracted when it is placed into the "File Space", and other types of metadata, such as the subject of a digital photo, can be added into a personal digital library by the owner.

In the traditional file system, the user can only search the file name or the available textual information by keyword terms. That is like searching a needle in a haystack, and the search result is often dissatisfactory. In a personal digital library, the search mechanism can employ the metadata and categories assigned by users in addition to the filename, and the user can receive more precise and better results.

3.3 Packaging

As the traditional digital library, effective utilization and distribution of digital content makes a personal digital library valuable. Packaging, distribution and value-added applications of digital content are related issues, and the issue of packaging is discussed in this subsection.

A personal digital library should provide two operating environments, web environment and console mode. In web environment, a user can access his/her digital content with rich value-added applications everywhere, and console mode would provide he/she a brief view of his/her digital content.

There is a scenario that a user may need to use the content offline; for example, a user may want to transfer some documents to his mobile device (ex. PDA) for reading while taking a train.

Thus, a personal digital library should provide "packaging mechanism" for packing some digital content and related metadata in a package. A package can be downloaded by the owner or accessed by other people if proper access control policies and DRM (digital right management) policies are satisfied. With metadata, the user who receives a package can perform metadata search instead of keyword search, and get better search results.

3.4 Distribution

A user may want to share his/her digital content to others. The distribution mechanism of a personal digital library is a little different to nowadays distribution mechanisms.

FTP and P2P are the most popular distribution methods over Internet. FTP and some P2P applications use a centralized model which all materials are collected on a centralized server. Other P2P applications use some distributed architectures; a client can get the shared material from multiple providers. A user can download digital content packages from a personal digital library, and then distributes them by these methods.

Besides, a user can distribute his/her digital content in a personal digital library, after he/she configures the access control policies. An access control policy contains users, operations, and materials. The users can perform operations on appointed materials if an access control policy exists.

The detail architecture of distribution would be discussed in subsection 4.3 .

3.5 Authoring and Version Control

The content owner sometimes has to modify the content in a personal digital library, and these authoring or modifying activities may happen online or offline.

For convenience, a personal digital library should provide simple editors online, such as tools for image resizing and a WYSIWYG HTML editor. If the owner wants to author or modify the content offline, he/she can use preferred tools, such as Adobe Photoshop and Macromedia Dreamweaver. After authoring or modifying, the owner can place the new or modified content to the personal digital library and provide new metadata for it.

A personal digital library should provide version control to handle the modifications of existent content made by a user. With this mechanism, the content owner can preserve the historical information of a piece of content, and can access the previous versions when he/she needs.

Besides, it would be better for a personal digital library to provide a mechanism for producing different media types of a piece of content. For example, a user may submit a large resolution picture into the personal digital library and request it to be used on a PDA or a mobile phone later; therefore the personal digital library should store the resized picture as well. Another scenario is that a user wants to distribute a HTML file as a GSM short message to his friend, thus the personal library should send the text content retrieved from the original HTML file.

3.6 Legacy Systems Integration

There are many working services in every universities and organizations. However, most of these services are not well-integrated, and users have to remember many accounts and passwords and get used to different user interfaces. A personal digital library should integrate the legacy systems and provide a unique interface for users to access the legacy systems.

Traditionally, forwarding the query to the legacy systems and re-arranging the result page is in

common use. However, in this manner, programmers usually spend too much effort on re-arranging and checking the interface changes of legacy systems.

To reduce the task of programmers and increase the usability of legacy systems, XML format message passing is a good choice. SOAP and Web Services are two technologies providing XML message passing. For legacy systems, a new function to receive and generate XML format message is necessary. For a personal digital library, a function to send request XML format message and parsing the returned XML format message should be provided.

3.7 Value-Added Applications Development

From the dot com mania, many value-added applications are released, such as portal [17], search engine [18], blog [19] and album [19]. They really enrich our daily life and improve the Internet development.

Besides packaging and distribution, value-added applications are also necessary to utilize the digital content of a personal digital library. In a personal digital library, pieces of digital content and their metadata are stored and can be used to develop various kinds of value-added applications for distribution, publishing, authoring, etc.

To support this, a personal digital library should provide rich APIs and utilities. In addition, a personal digital library can integrate multiple storages of different content owners to construct a virtual community digital library. Of course, studies on the social network and personal interests are also valuable.

4 Architecture of Personal Digital Library

4.1 Personal Digital Library Framework

Fig. 1 shows the software architecture of the Personal Digital Library (PDL, in short) framework. The PDL framework contains three layers: storage layer, middleware layer and utility layer.

Storage layer contains the File Space, Metadata Repository and Policy Repository. File Space is used to store the digital content from one user and some encoding is used if necessary. Metadata Repository stores the metadata of digital content and indexing information, and Policy Repository stores the DRM information and access control policies.

Middleware layer contains Information Retriever and File Retriever. File Retriever communicates with utility layer, accesses the File Space and performs encoding and decoding of the digital content.

Information Retriever contains two components: Repository Retriever and XMF (XML-based Metadata Management and Manipulation Framework) [9]. Each is used to receive the request from utility layer and access the Metadata Repository and Policy Repository. The difference of these two components is that the XMF is used in console mode (off-line) and the Repository Retriever is used in web environment.

XMF takes responsibility of metadata and policies retrieval. XMF can perform the keyword search and relationship management of metadata of digital content by using FCM (Filter Constraint Module) and RCM (Relation Constraint Module). XML-based metadata and policies can be managed by MOM (Metadata Object Module). For further details about XMF, please refer to [9].

In the utility layer, the personal digital library system designer should provide utilities and APIs for developing other value-added applications for users. In the presentation layer, a user interface wrapper is essential. It wraps the user interface of applications according to the devices and Internet access conditions.

4.2 Architecture of the Personal Digital Library Environment

Users can use the Personal Digital Library (PDL) through web environment and console application environment.

Fig. 2 shows the web environment architecture of the PDL. The PDL server handles all the tasks of retrieving, querying and maintaining digital content, metadata and policies, and it also provides an interface for the user. Besides, PDL framework provides APIs such that other applications can be developed. Therefore, the user can use the browser to access the PDL or other applications based on the PDL, such as web-HD, albums, and blogs.

In addition, the PDL server can integrate legacy services, such as traditional digital libraries and e-learning systems. This integration task is done by the XML format message passing. The PDL server would send XML format request messages to legacy systems for requesting information and data related to the user. Then the results would also be returned in XML format. Thus, some modifications should be taken on legacy services to wrap the original result to XML format message.

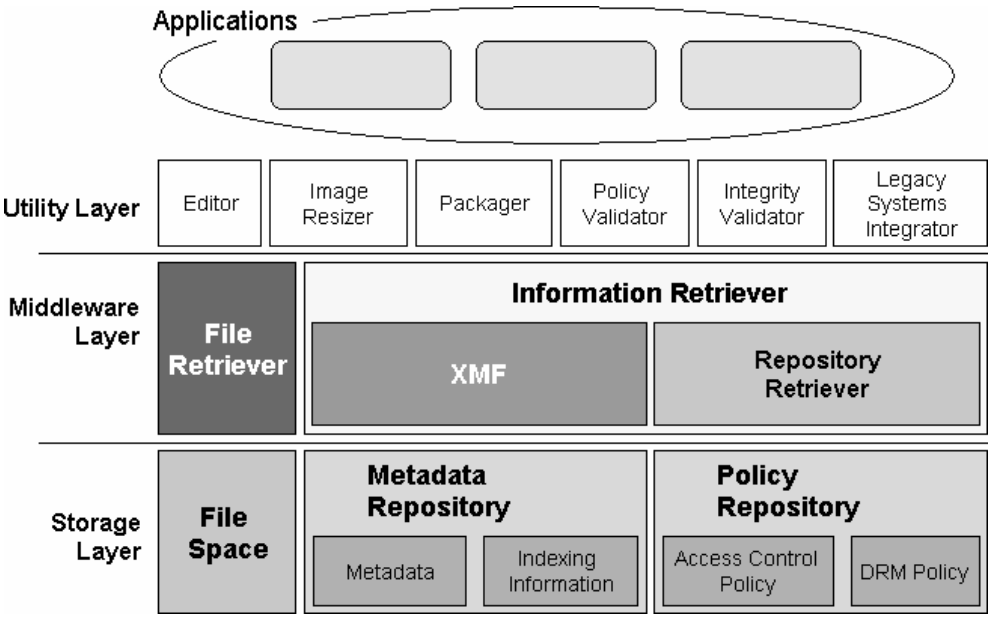

Fig. 1: Software architecture of Personal Digital Library framework.

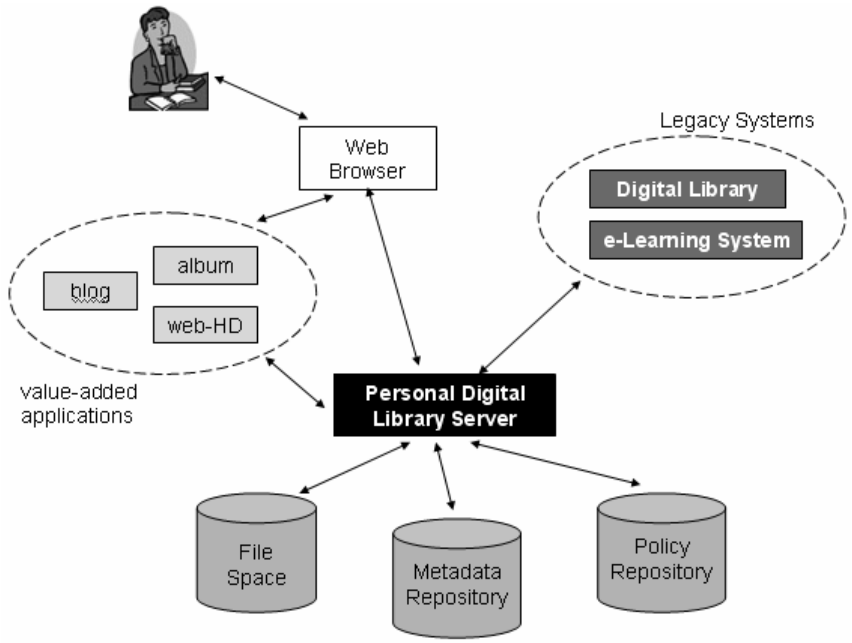

Fig. 2: The web environment of the Personal Digital Library

4.3 Distribution in the Personal Digital Library

To distribute digital content, there are two common models, client-server model and peer-to-peer model.

Client-server model centralizes the storage layer, File Space, Metadata Repository and Policy Repository. The digital content owner can set the access control and DRM policies on the centralized server for content sharing. Then, the other users who are policy-satisfied can perform a search in the owner's Metadata Repository and access the target digital content. This centralized client-server model, reduces the risk of instable storage on PC and increase the availability of digital content.

Peer-to-peer model lets the storage layer to be managed on the computer of the digital content owner. In this model, the task of managing Metadata Repository and Policy Repository is handled by the owner and the File Space is placed the owner's computer. This design can reduce the loading of centralize server (even no centralized server) and reduce the response time of access of the personal digital library.

Besides, a hybrid model may comprehend the advantages of client-server and peer-to-peer models,

and avoid the disadvantages. This hybrid model leaves the File Space on the centralized server to increase the availability, places the metadata and policy repositories on the owner's computer to reduce the response time and server loading, and backups metadata and policy repositories on the centralized server to increase the robustness of storage. However, the load of network communication will increase in this model.

In addition to the web environment, the user can also use the PDL in console mode. The user can download the package of his/her digital content and related metadata from the web environment and store it in the console. The PDL console application provides similar functions as the web environment, except the value-added applications and legacy systems integration. The PDL console application provides a brief view of the digital content and the metadata, and is helpful for reading, especially on a mobile device.

5 Conclusion and Future Work

An architecture of a personal digital library that provides digital content preservation, metadata annotation, searching, distribution, authoring and packaging is proposed in this paper. Based on this architecture, many value-added applications can be built, and digital content can be taken, accessed, distributed and managed by the owner anywhere and anytime. Besides, multiple personal digital libraries can be integrated via this architecture to achieve a virtual community digital library.

In the future, the whole architecture and some applications based on it will be implemented, and some research topics, such as searching efficiency, indexing, security, and data mining will be proceeded.

References:

[1] G. Amati, C. Carpineto, G. Roman, Comparing weighting models for monolingual information retrieval, *CLEF 2003*, Trondheim, Norway, 2003.

[2] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker; Query by Image and Video Content: The QBIC system, *IEEE Computer*, pp. 23-32, vol. 28, issue 9, 1995.

[3] S. Belongie, C. Carson, H. Greenspan, and J. Malik; Color- and texture-based image segmentation using EM and its application to content-based image retrieval, *Proceedings of the International Conference on Computer Vision (ICCV'98)*, 1998, pp. 675-682.

[4] Edward A. Fox; Digital Libraries, *IEEE Computer*, Volume 26, Issue 11, 1993, pp. 79-81

[5] Chung-Sheng Li, and Stone, H.S., Digital Library Using Next Generation Internet, *IEEE Communications Magazine*, Volume 37, Issue 1, 1999, pp. 70 – 71

[6] CORPORATE The Stanford Digital Libraries Group, The Stanford Digital Library Project, *Communications of the ACM*, Volume 38, Issue 4, 1995, pp. 59 – 60.

[7] Robert Wilensky, UC Berkeley's Digital Library Project, *Communications of the ACM*, Volume 38, Issue 4, 1995, pp. 60.

[8] Laurie Crum, University of Michigan Digital Library Project, *Communications of the ACM*, Volume 38, Issue 4, 1995, pp. 63-64.

[9] Shihn-Yuarn Chen, Hao-Ren Ke, and Wei-Pang Yang, Heterogeneous Metadata Management and Manipulation using an XML-based Framework, *Int. Computer Symposium*, Dec. 15-17, 2004, Taipei, Taiwan, pp. 9-14

[10] Shien-Chiang Yu, Kun-Yung Lu, Ruey-Shun Chen, Metadata management system: design and implementation, *The Eletronic Library*, Vol 21, Num 2 2003, pp. 154-164

[11] Ruey-Shun Chen, Shien-Chiang Yu, Developing an XML framework for metadata system, *Proceedings of the 1st international symposium on Information and communication technologies*, 2003, pp. 267 – 272

[12] Daniel Higgins, Chad Berkley, Matthew B. Jones, Managing Heterogeneous Ecological Data Using Morpho, *Proceedings of the 14th International Conference on Scientific and Statistical Database Management*, 2002, pp. 69-76

[13] Josep M. Ribó, Xavier Franch, A Multi-version Algorithm for Cooperative Edition of Hierarchically-Structured Documents, *Proceedings of 7th International Workshop on Groupware*, 6-8 Sept. 2001 pp. 154 – 163.

[14] Gr'egory Cob'ena, Serge Abiteboul, Am'elie Marian, Detecting Changes in XML Documents, *Proceedings of the 18th International Conference on Data Engineering*, 2002, pp. 41-52.

[15] Zope.org, http://www.zope.org/

[16] DSpace.org, http://www.dspace.org/

[17] Yahoo!, http://www.yahoo.com/

[18] Google, http://www.google.com/

[19] Blogger, http://www.blogger.com/

[20] Flickr!, http://www.flickr.com/