

# A Web-Based Resource Discovery Architecture for Digital Archive Systems

Shian-Hua Lin

Department of CSIE

National Chi Nan University

E-Mail: [shlin@csie.ncnu.edu.tw](mailto:shlin@csie.ncnu.edu.tw)

Jan-Ming Ho and Pei-Xian Kou

Institute of Information Science

Academia Sinica

E-Mail: {[hoho](mailto:hoho@iis.sinica.edu.tw),[reno](mailto:reno@iis.sinica.edu.tw)@iis.sinica.edu.tw}

## Abstract

In past several years, researchers, archival organizations and institutes invested efforts in constructing digital archives. Nowadays, there are rich digital archives distributed in many archival systems. However, to retrieve archival contents from these systems, users must be familiar to the system's URL address, the user interface and functions provided by the system, since no single Web site (like the portal site) was provided to users in the digital archive environment. Therefore, digital archives are still hard to be accessed by general users. In this paper, based on our system "Knowledge Portal (KP)", we propose the resource discovery architecture to facilitate the interoperability among diverse digital archive systems or other information systems. In the architecture, we follow the Dublin Core (DC) Metadata Element Set as the standard of exchanging the information. Using DC fields, diverse metadata of archives are integrated into the common format with single retrieval interface. Organizing catalogues to construct a global view for diverse contents is also considered in this paper. Based on the OAI framework, the OAI-PMH is implemented and integrated with the directory engine of KP to provide an integrated view of archives. Since KP is also a platform of Web portal site, resources available from the Web can also be managed, indexed and retrieved in the environment. That is we achieve a We-based portal as the discovering architecture not only for digital archives but also for the public Web resources. Based on the system, Web resources and digital archives can be seamlessly integrated in the same portal and users are able to access both resources simultaneously.

## 1. Introduction

In the past years, many people's views of digital libraries (DLs) are based on the foundations of traditional libraries. How to build digital archives is the key concern of DLs. Therefore, most studies focus on the construction of digital archive (DA) systems. However, DLs are more than digital collections (archives) [13]. That is the creation of large amount of digital collections merely facilitates managers and content experts to store and manage archives. These valuable contents are not popular to us, the general users. Obviously, the benefits of DLs will not be appreciated unless they are easy to use effectively [7]. Therefore, we need a single portal system to drive DA systems toward DL systems that facilitate users to access diverse archives from various DA systems on the Internet. In other words, a Web DL system is the integration of a Web portal site and various DA systems. Searching and browsing are the basic and major using methods on the Internet. That is the basic service of the portal is providing searching and browsing integrated contents of many DA systems.

A fundamental issue for integrating DLs is *interoperability* [13][14]: the capability of exchanging and sharing contents, queries, and services. From a slightly different perspective, interoperability in DLs is the ability to generate a single (virtual) view on many different library components without sacrificing autonomy [1]. To keep the autonomy of each DA system, we use the OAI architecture [12] as the basic framework of the portal. Based on the OAI, we integrate contents of several DA systems with our information system, Knowledge Portal (KP) [8][9]. KP was carried out for collecting, organizing, indexing and searching various documents on the Internet. It provides two basic services: searching and browsing for Web users. We extend interface modules of KP to facilitate the integration with DA systems and KP is therefore the single portal to integrate DA systems as the DL Web portal.

In the following sections, we first introduce related studies of the paper. Then we describe the overview of our system, KP. Based on the OAI framework, we illustrate how to carry out the OAI-PMH in KP to provide the OAI-based integration of digital archives. Finally, we describe the conclusion and future work of the paper.

## 2. Related Work

In the DESIRE project [4], the topology of metadata formats is categorized into three bands [3]:

- Band-one includes relatively unstructured data that are automatically extracted by Web search engines and portal sites. The metadata do not support the function of searching by fields.
- Band-two includes structured data that contain full enough description to support field searching. Typically, these data are simple enough to be created by non-specialist users, or not to require significant discipline-specific knowledge. Dublin Core is one of the examples [5][6][21]. Such metadata can be applied to discovering or harvesting services with a little manual effort.
- Band-three includes full descriptive formats that are associated with research or scholarly activity and require specialist knowledge to create and maintain. Since specialists tend to use rich and complex formats to maintain their contents, archives of DL systems are usually maintained in this format.

Intuitively, simpler metadata are easy to be automatically processed by the computer software. However, complex metadata, such as the band-three metadata, are hard to be processed without efforts of specialists. Consequently, the band-two metadata is the compromise of two metadata. In this paper, we use Dublin Core as the bridge to integrate contents of diverse DL systems to facilitate a uniform searching environment.

In the portal of DL systems, we also need provide the browsing service to navigating the virtual (integrated) of different DL contents. Open Archive Initiative – Protocol for Metadata Harvesting (OAI-PMH) [12][11] is a widely used protocol to exchange directories and metadata of different DL systems. In the OAI architecture, two classes of participants are defined to facilitate the data exchange. The *data provider* is combined with DL systems to support the OAI-PMH, it likes the Web server supports the HyperText Transfer Protocol (HTTP) [17]. The *service provider* submits OAI-PMH requests to the data provider to harvest the information. It looks like a Web crawler that follows HTTP to grab documents from the Web server. The communication between data and service providers is based on the HTTP. The requirement for metadata interoperability is addressed by requiring that OAI data providers supply metadata in a common format [10] – the Dublin Core Metadata Element Set [6]. Therefore, the common metadata of the portal is the Dublin Core fields. A record of the OAI framework is an XML [18] encoded byte stream that serves as a packaging mechanism for harvested metadata.

### 3. The Overview of KP

Knowledge Portal (KP) is basically an integrated directory and search engine for Web resources. Currently, KP provides the following services:

- Directory management and resource classification. It provides hierarchical management for resources and learns the classification knowledge from the supervised hierarchy to support automatic classification.
- Search engine and query analysis. It is a full-text and keyword search engine. By extracting keywords and analyzing query logs, it also supports the data mining of keyword (concept) associations.

The system architecture is shown in Figure 1. The *Crawler* is modularized to handle different protocols such as HTTP, NNTP, POP3, IMAP, OAI-PMH, network file systems (LAN documents), etc. The *Parse* deals with various formats of resources grabbed by the Crawler, such as HTML [18], XML, Office, PDF, E-Mail (eml) document resources, etc. To support the index and search of Dublin Core fields, the *Index Engine* is carried out to build metadata (15 DC fields) and full-text indices based on the Dublin Core mapping pre-defined in the Parser or the user-defined mapping while crawling the XML or OAI metadata. The *Directory/Search Engine* supports the browsing of directories and the searching of Web resources. The directory can be manually added or imported via XML data files or OAI Harvesters. The Learning System learns the classification knowledge and mines concept associations [8][9].

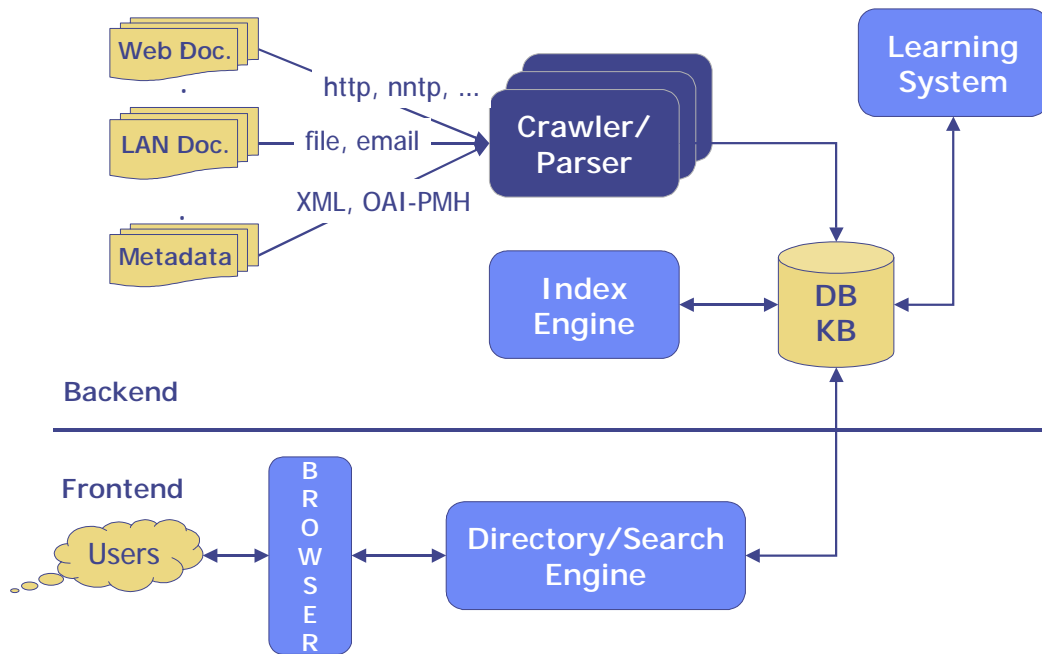


Figure 1. KP's system architecture.

### 4. Using KP as the Portal of DL Systems

As we described previously, there are two classes of participants in the OAI environments: Data and Service Providers. The Data Providers in the interface connecting DL systems and the Internet. It handles the OAI-PMH requests and retrieves metadata corresponding to these requests. The Service Provider (the OAI Harvester) issues OAI-PMH requests to grab metadata of diverse resources. The

service provider is the interface to request the OAI metadata and it is therefore implemented in the KP's Crawler. Consequently, KP can be a DL portal by extending the OAI Harvester in the KP's Crawler and Parser. The DL portal based on KP and the OAI framework is shown in Figure 2.

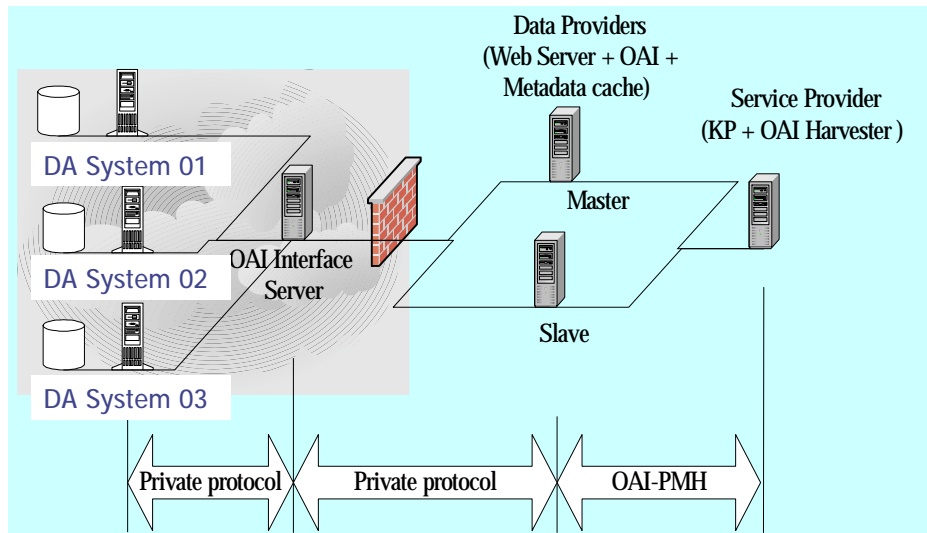


Figure 2. A DL portal based on KP and OAI.

Intuitively, the resource discovering processes include the following three steps [2]:

1. Selecting which library (repository) to look in;
2. Selecting which documents (resources) within a library to look at; and
3. Selecting fragments of data (text, numeric data, images) from within a document.

We illustrate how KP applies the OAI harvester to achieve the discovering processes and map resources to the portal.

#### 4.1. Selecting Repository

In the OAI framework, metadata records are disseminated from *repositories*, which are network accessible servers of data providers supporting the set of OAI protocol requests. The URL is employed to uniquely identify the repository. That is selecting the repository is the same with discovering a site on the Web. Therefore, an OAI repository containing a hierarchical metadata is mapped to a directory of KP by filling the repository's URL. Repositories can be added by manual efforts or by checking DA systems automatically (if DA systems support the OAI-PMH). Then, KP's OAI harvester connects to the repository based on the HTTP and OAI-PMH.

Abstractly, repositories contain items, and each metadata record harvested from a repository corresponds to an item. To organize items according to hierarchical concepts, sets are used to hierarchically group items in the OAI. Each repository may define a hierarchical organization of sets that can have several top-level nodes, each of which is a set. Thus, KP's OAI harvester tries to send OAI command "ListSets" to map OAI sets into the KP's directory hierarchy after it connected to a repository.

#### 4.2. Selecting Resources

The OAI framework provides the selective harvesting mechanism in two types:

- Y Date-based: According to the date stamp define as “the date of creation, deletion, or latest date of modification” of an item (record), harvesting requests (OAI “GetRecord” command) may correspondingly contain a date range for partially harvesting metadata.
- Y Set-based: Sets may be used in harvesting items of specific groups. By sending the “ListSets” command to the repository, KP first gets metadata of sets and get records of interested sets according to the set’s metadata. In other words, KP can crawl partial hierarchies (sets) of the repository based on the selective harvesting of the OAI.

### **4.3. Selecting Fragments of Data**

The nature of an item - for example, what type of metadata is actually stored in the item, what type is derived on the fly, and whether the item includes the “full content” described by the metadata - is outside the scope of the OAI protocol [10]. That is OAI records only contain the metadata information. The selection of data fragments usually indicates the content described by the metadata. It depends on the application of portal. For example, using KP as a DL portal concerns about searching and browsing services. In the searching service, search engines only know how to process and index text contents (documents). The content type is usually specified in the metadata and search engines know how to interpret the content type and dispatch crawlers to grab the full-text content. That is KP achieves the function of selecting data fragments according to the content type of the record’s metadata.

## **5. Conclusion and Future Work**

In this paper, we extend our portal site platform, KP, to adopt the OAI framework and achieve a DL portal environment for integrating digital archives and Web resources. It’s a beginning of driving DA systems toward an integrated and easily used DL system. Based the OAI framework, it becomes feasible to integrate diverse archives and provide the union searching and browsing services. However, the OAI framework is a passive system since the modification of a record is never notified by the DA system. The case is similar with that of the portal site which does not know when is a Web site modified. To integrated with our DA systems efficiently, we must extend the OAI Data Provider to be able to notify Service Providers the modification. Based on HTTP, SOAP [20], XML, and the flexibility from the Web service, implementing Data and Service Providers as Web service components to achieve an active system is our future work.

Currently, the mapping and integration of diverse archives are based on Dublin Core Metadata Element since the mapping without semantic loss needs efforts of specialists. Studies on the automatic semantic fusion of archives are ongoing. In [16], Schatz illustrated the generations of information retrieval in DLs: from syntax (text search) to structure (document search) to semantic (concept search). He also predicted, by 2010, the visions will be realized, with concept search enabling semantic retrieval across large collections. The research trend is how to reduce the effort of specialist while mapping various archives into the portal. In the future, based on cooperative experiences with specialists, we will try to develop procedures and software systems to reduce the needed manual effort while integrating digital archives.

## **6. Reference**

- [1] Birmingham, B., etc., “EU-NSF Digital Library Working Group on Interoperability between Digital Libraries,” <http://www.iei.pi.cnr.it/DELOS/NSF/interop.htm>.

- [2] Buckland, M., "Selecting Libraries, Selecting Documents, Selecting Data," International Symposium on Research, Development and Practice in Digital Libraries, 1997 (ISDL'97).
- [3] DESIRE Project, "A review of metadata: a survey of current resource description formats," <http://www.ukoln.ac.uk/metadata/desire/overview/overview.pdf>.
- [4] DESIRE Project, "DESIRE Information Gateways Handbook," <http://www.desire.org/handbook/handbook.pdf>.
- [5] Dublin Core, "Dublin Core Metadata Initiative," <http://dublincore.org/>.
- [6] Dublin Core Metadata Element Set, Version 1.1: Reference Description, <http://dublincore.org/documents/1999/07/02/dces/>.
- [7] Fox, E. A. and Sornil, O., "Modern Information Retrieval – Chapter 15: Digital Libraries," Addison Wesley, Edited by Baeza-Yates, R. and Riberiro-Neto, B., 1999.
- [8] Lin, S. H., Chen, M. C., Ho, J. M., and Huang, Y. M., "ACIRD: Intelligent Internet Documents Organization and Retrieval," IEEE Transactions on Knowledge and Data Engineering, 14(3), May/June, 2002.
- [9] Lin, S. H., Shih, C. S., Chen, M. C., Ho, J. M., Ko, M. T. and Huang, Y. M., "Extracting Classification Knowledge of Internet Documents: A Semantics Approach," Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98), August, 1998.
- [10] Logoze, C. and Van de Sompel, Herbert, "The Open Archive Initiative: Building a Low-Barrier Interoperability Framework," International Joint Conference on Digital Library, 2001.
- [11] OAI, "Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting," <http://www.openarchives.org/OAI/2.0/guidelines.htm>.
- [12] OAI, "The Open Archives Initiative Protocol for Metadata Harvesting," <http://www.openarchives.org/OAI/openarchivesprotocol.htm>.
- [13] Paepcke, A., Chang, C.-C. K., Garcia-Molina, H., and Winograd, T., "Interoperability for Digital Libraries Worldwide," Communications of the ACM, 41(4), Apr. 1998, pp. 33-43.
- [14] Paepcke, A., Cousins, S. B., Garcia-Molina, H., Hassan, S. W., and Ketchpel, S. P., etc., "Using Distributed Objects for Digital Library Interoperability," IEEE Computer, 29(5):61-68, May 1996.
- [15] Powell, A and Apps Ann, "Encoding OpenURLs in Dublin Core Metadata," <http://www.ariadne.ac.uk/issue27/metadata/>.
- [16] Schatz, B. R., "Information Retrieval in Digital Libraries: Bringing Search to the Net", Science, 275:327-334, 17 January 1997.
- [17] W3C HTTP, "Hypertext Transfer Protocol," <http://www.w3.org/Protocols/>.
- [18] W3C HTML, "HyperText Markup Language," <http://www.w3.org/MarkUp/>.
- [19] W3C XML, "Extensible Markup Language," <http://www.w3.org/XML/>.
- [20] W3C SOAP, "XML Protocol Working Group," <http://www.w3.org/2000/xp/Group/>.
- [21] Weibel, S., Godby, J., Miller, E., and Daniel, R., "OCLC/NCSA Metadata Workshop Report," <http://dublincore.org/workshops/dc1/report.shtml>.