The Emerald Research Register for this journal is available at
www.emeraldinsight.com/researchregister

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0264-0473.htm

Building an open archive union catalog for digital archives

Shien-Chiang Yu
Department of Information and Communications, Shin-Hsin University, Taipei, Taiwan, Republic of China, and
Hsueh-hua Chen and Huai-wen Chang
Department of Library and Information Science, National Taiwan University, Taipei, Taiwan, Republic of China

Abstract

Purpose – In January 2002, the National Science Council of Taiwan launched a National Digital Archives Program (NDAP) and has proceeded with the implementation of a system related to the open archives initiative (OAI) framework. This paper aims to introduce the protocol and the prototype system of the project.

Design/methodology/approach – A general review of the project.

Findings – The OAI interoperability framework has received much attention from scholars of library and information sciences. In Europe and North America, many academic organizations and universities have undertaken theoretical studies, system design, and implementation of the OAI framework. In January 2002, the National Science Council of Taiwan launched a NDAP, and the institutional project of the National Taiwan University is one of its institutional projects. Now, the project has proceeded with the implementation of a system related to the OAI framework.

Originality/value – Provides information of value to information professionals.

Keywords Digital libraries, Taiwan

Paper type General review

Introduction

The lack of interoperability is one of the significant issues that digital libraries (DLs) currently face. The inability to federate, filter and provide value-added services for remote content limits DLs to covering local holdings. One of the reasons is that each DL is aimed at the needs of a particular community (Suleman and Fox, 2001). The open archive initiative (OAI) is one major effort to address technical interoperability among distributed archives (Liu *et al.*, 2001). In essence it supports a system of interconnected components, where each component is a DL. OAI was born in the meeting of Universal Pre-print Service that took place in October 1999 in Santa Fe with Paul Ginsparg, Rick Luce, and Herbert Van de Sompel. The OAI referred to the Harvest system (Bowman *et al.*, 1995). The motivation for this creation arose because different databases and systems were not interoperable. Therefore, related data or data from different fields of science were stored in different locations and were not integrated, which made the flow of data imperfect. Representatives participating in the meeting regarded it as necessary to develop an interoperable standard for academic electronic pre-print and related digital archives. Thus, OAI was established (Ginsparg *et al.*, 1999). And in January 2001, OAI announced the open archives initiative protocol for metadata harvesting (OAI-PMH) to provide a feasible solution for the interoperability of network resources (Sompel and Lagoze, 2000).

The Electronic Library
Vol. 23 No. 4, 2005
pp. 410-418

0264-0473
DOI 10.1108/02640470510611472

In Europe and North America, many academic organizations and universities have undertaken theoretical studies, system design, and implementation of the OAI framework. In January 2002, the National Science Council (NSC) of Taiwan launched a National Digital Archives Program (NDAP) (www.ndap.org.tw/). This is a major policy of the Taiwan government concerning digital content resulting from the proposal to create a knowledge based economy. Many universities and research organizations participate in this program, and the institutional project of National Taiwan University is one of its institutional projects. With seven sub-projects, this project is urgently attempting to build an interoperable mechanism to share and conserve all valuable collections, retrieve the digital collections of these content holders via a union interface, and allow the general public to access the digital collections. For these reasons, this project will utilize the OAI-PMH to carry out related studies and implement a union catalog system. This paper will introduce the concept of OAI, and discuss experiences in planning and building the union catalog of the institutional project of National Taiwan University.

The concept of the OAI-PMH

OAI after a period of testing announced the OAI-PMH in January 2001 (OAI v1.0). In July 2001 the revised version 1.1 was issued and OAI-PMH 2.0 was the latest and formal version which was published in June 2002 (Sompel and Lagoze, 2001a, b; Lagoze and Sompel, 2002). OAI-PMH 1.0 introduced the unqualified Dublin Core element set as a baseline for metadata interoperability. It focuses on facilitating the discovery of document-like objects. OAI-PMH 1.1 was a revision of the 1.0 specification taking account of changes to the emerging XML Schema specification. Both v1.0 and 1.1 were experimental in nature. OAI-PMH 2.0 is a stable protocol, and no longer experimental. Once again the focus of the protocol expanded; now it was said to be concerned with the recurrent exchange of metadata about resources between systems. OAI has already submitted the OAI-PMH to the World Wide Web Committee (W3C), hoping it will become the international standard for metadata sharing. Through its independent platform and mutual operation, it can provide and promote the efficient dissemination of content.

The intentions of the OAI-PMH include exposing and harvesting, which are defined in the OAI protocol as:

(1) defining a data provider which can expose its metadata through the HTTP-based protocol; and

(2) defining a mechanism for metadata harvesting from repositories.

According to the different tasks in the OAI organization, there are two groups – data providers and service providers. Participants must in terms of their types of service, register for one of two roles (Figure 1).

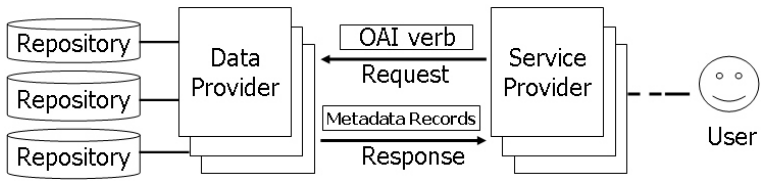

Figure 1.
The roles of data provider and service provider

- *Data provider*. Maintains one or more repositories (web servers) that support the OAI-PMH as a means of exposing metadata.
- *Service provider*. Issues the OAI-PMH requests to data providers and uses the metadata harvesting from data providers as a basis for building value-added services.

The primary purpose of the OAI-PMH is incremental bulk transfer of metadata (harvesting). There is no remote search facility. Instead, a provider of services acquires data from a data provider, stores and processes it locally, and then supplies services to users based on that data (Suleman and Fox, 2003). Besides the data and service providers, it includes server components to manipulate the data.

- *Set*. A set is an optional construct for grouping items for the purpose of selective harvesting. Repositories may organize items into sets. Set organization may be flat, i.e. a simple list or hierarchical. Multiple hierarchies with distinct, independent top-level nodes are allowed.
- *Record*. The OAI framework defines a record, which is an XML-encoded byte stream that serves as a packaging mechanism for harvested metadata.

Because each institution has its respective digital archive system with individual search interface, data structure, communication protocol, management policy and so on, there is an inability to federate, communicate and share data with each other transparently. For the purpose of interoperability the establishment and achievement of a union catalog will be the key, to providing the user with the ability to search all of the collected records from a single search interface. OAI-PMH provides the minimum complexity, but maximum convenience to fulfill interoperability, thus keeping the balance between functional enhancement and developmental simplification. This was the reason why this project used OAI-PMH. Some of the advantages of adopting the OAI-PMH interoperability framework are listed below.

(1) *It provides a new model for scholarly communication*. The OAI develops and promotes interoperability solutions that aim to facilitate the efficient dissemination of content. Furthermore, by adopting the metadata harvesting method it can cover a variety of media formats, data types and contents, etc.

(2) *It is easy to implement*. The OAI-PMH were designed to be very simple and efficient. By avoiding complexity, enabling an existing information repository to function as an OAI-compliant data provider is a relatively simple process (Breeding, 2002). Protocol requests and responses, defined in the OAI-PMH only include six verbs (OAI 2.0):

   - GetRecord: to retrieve an individual metadata record from a repository;
   - Identify: to retrieve information about a description of archive repository-standards and protocols implemented;
   - ListIdentifiers: to retrieve record identifiers, optionally corresponding to a specified set or data range;
   - ListMetadataFormats: to retrieve the supported metadata formats available from a repository;

- ListRecords: to harvest records corresponding to a specified metadata format from a repository;
- ListSets: to retrieve the sets and subsets from a repository.

(3) *It is open.* Everyone can apply the framework of OAI-PMH to build various of data provider or service provider.

(4) *It adopts the web standard.* The OAI-PMH uses current standards wherever applicable on the internet. All data that is transferred in response to a request is encoded in an XML format defined using XML Schema and transmitted on HTTP. Taking advantage of these standards lets the OAI solve problems such as crossing platforms etc.

Institutional project of National Taiwan University

The Institutional Project of National Taiwan University is one of the NDAP's institutional projects. There are seven institutions participating in the project, including the NTU Library, Departments of Botany, Entomology, Geosciences, Anthropology and Zoology and the Computer and Information Networking Center. Its primary research scope contains the following:

- to understand the history and features of collections;
- to study various metadata formats both domestically and internationally;
- to understand relations among the metadata, the database and the system framework; and
- to understand the information demand and retrieval behavior of potential users

During the first few years, the main task of the project has been to develop a management system capable of handling various types of metadata and to implement it in each institution (Yu *et al.*, 2003), but not to integrate it among these institutions. The major reason for this is each institution has a specific application domain. These metadata records and digital objects (e.g. image, picture, and voice) were edited or made by each institution, and different science domains have relations of differing frequencies, such as the geographic influence of ecological distribution (geosciences and zoology). Therefore, the following project is set to build a union catalog to integrate metadata records that harvest from various digital collection institutions. As with a digital collection portal, users, especially researchers, do not individually query databases and can directly fetch all of the related data.

The NDAP addressed the idea of creating a union catalog of National Digital Archives. A union catalog can be built based on two models, a collective union catalog or a distributed virtual union catalog. The former has the advantage of offering better search results, but has the disadvantage of a high construction cost. The advantage of adopting a virtual union catalog is the low construction cost, but it offers poor search results. In order to maintain the advantages of both models and avoid their drawbacks, the project is designed to adopt the OAI-PMH framework, with the program office playing the role of a service provider. National Taiwan University, Academia Sinica, the National Palace Museum, and the National Museum of Natural Science will participate in the initial phase to form the OAI test-bed team. They will build the union catalog of the national digital archives with the OAI-PMH to automatically harvest

metadata from each repository periodically. When the test-bed team achieves its goal, more repositories will participate in the system.

Implementing the union catalog

The digital collection institutions should share resources with each other and provide users with a transparent access channel. To achieve this, the institutional project of National Taiwan University has made use of the OAI-PMH to create an interoperable system. Using this protocol, it will facilitate communication between service providers and data providers, and the data of digital collections can keep its original metadata structures or Dublin Core format. Besides, users can search and access resources conveniently through the OAI-based system.

Figure 2 shows the initial system structure of the National Taiwan University, according to OAI-PMH. The operations of each component are described below.

Data provider

This converts the metadata records from data providers to XML format in batches, and maps them to the Dublin Core metadata. The OAI-PMH only allows for the retrieval of identifiers and associated records from the remote repository. The data provider could base the response on current machine load or limit the frequency at which requests will be serviced. The records required for harvesting from the data provider to service provider consists of two parts, the identifier and datestamp, and the XML metadata record in the request format. Delivery data must be encapsulated by this format.

Service provider

This provides for storing the responded metadata records from data providers and recording the related attributes, including the update time, the original system number, and the source of data, etc. It defines contents by index arguments to establish index files which facilitate the function of searching. It also creates and maintains the interfaces for user authorization and system administration and supplies web services to search and browse via internet.

Database

Handles the recording and administrating of the metadata records and low resolution digital objects which are harvested from data providers. It also has the index tables for function of information retrieval and the parameters for system administration. It also provides the mapping arguments of metadata and the definitions of XML schema.

The final objective in this union catalog project is depicted in Figure 3. Because of the OAI-PMH based on data and service providers having two intentions the exposing

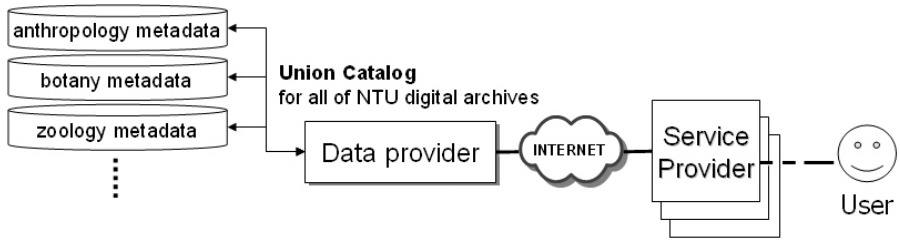

Figure 2.
The main structure of the NTU system

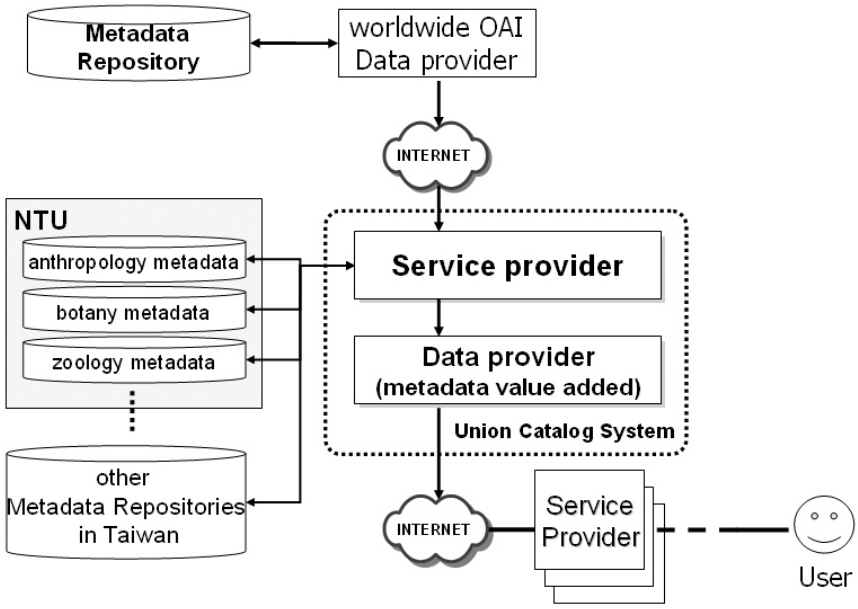

Figure 3.
The final system structure of the NTU project

and harvesting of metadata, the system should combine both of them not only through integrating the metadata records from each repository in the project but also through the OAI service provider to harvest the metadata records from worldwide OAI data providers. By integrating the two resources, the system can play the role of a metadata portal and provide users with a variety of metadata records from different data providers.

To sum up what has been analyzed above, before developing the system, several actions must be performed first:

- collecting the depth of data from each data provider, including the structure, syntax, semantics of the metadata and the relationships between the metadata and digital objects, and determining the extent and range of data which the repositories allowing to provide;
- sorting the metadata mapping tables for converting formats among metadata records, and establishing the XML schema definition of the metadata;
- defining the access points of the metadata scheme in order to allow the system to access similar fields among different metadata; and
- creating the authority control of the metadata, which allows the correlation of one record field with another.

Issues and solutions

There are several issues when implementing the system.

(1) *Metadata.* Most data providers (institution in NTU project) can provide complete records, but a few data providers, because of the consideration of property rights and access restrictions, can only expose partial metadata

elements. Consequently, there are different data fields or format in the metadata types between data providers and service providers. For this reason, we must design the mapping solution when integrating and harvesting these metadata elements from these data providers.

(2) *Digital object.* Digital collections not only include a descriptive metadata record, but also contain images, pictures, or voices, which are called digital objects. It is not easy to imbed the digital objects in XML, and the issue of property rights leads some data providers to not want to share their digital objects or to provide them while reducing the resolution.

(3) *System platform.* Each repository uses a different system platform, which may be Microsoft Windows Server or Linux. This creates diversity in application software, database and web servers.

For the convenience and cost of implementation, we have revised the procedure for integrating records from each repository as in Figure 4. The system functions of the repositories include transforming the original metadata to digital library metadata, filtering and recombining the incomplete elements and reducing the resolutions of digital objects, etc. Then, the union catalog system can harvest these value-added data via OAI service provider, and process indexes for use. When users find what they need, they can depend on the identification numbers of records and link to the repositories of the original digital collection institution. The repositories have the right to decide whether or not to provide detailed metadata records and contents.

The system is based on the OAI-PMH version 2.0. Java is the language for the development of software. And it adopts the SQL Server 2000 of Microsoft to build the database. Our experiences of implementation are discussed below.

(1) Since OAI aims to promote interoperability, DC metadata has been adopted as a lowest common-denominator metadata format which all data providers should support.

(2) Taking copyright issues into consideration, some data are not appropriate to be made public. Therefore, there must be a procedure for filtering data on the side

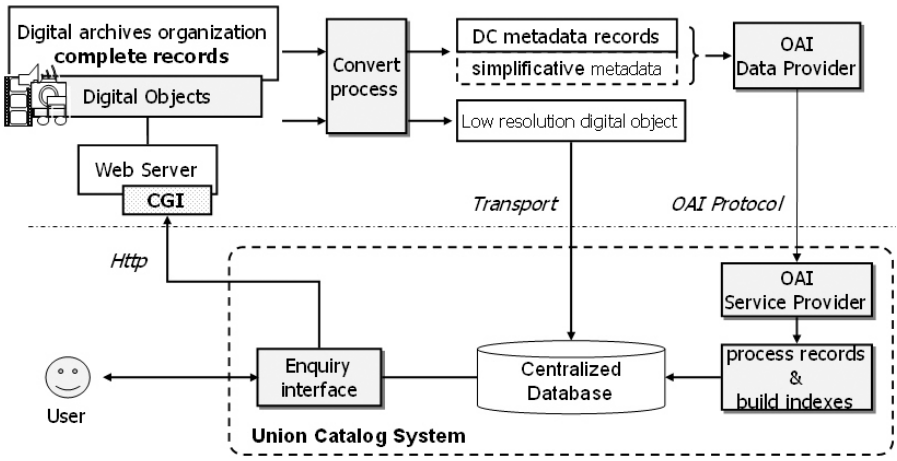

Figure 4.
The procedure of records integrated

of data providers. Based on the above need, the repositories should build not only the OAI related programs but also processing programs involving data change and transformation.

(3) Because the OAI-PMH only concerns the harvesting of metadata, there is no mechanism for handling digital object types (e.g. image, picture, voice), which are included in the metadata record. Therefore, all other document formats and contents need the assistance of other application programs, and the system should provide accurate full text or multimedia linking, display or functioning. The service provider not only has to provide the integrated searching function but also has to display the simple contents of digital objects, such as thumbnails or partial contents of the multimedia files. Overall, the service provider must take advantage of the OAI-PMH to harvest metadata from data providers, and also needs to upload the simple information about digital objects.

Conclusion

This paper has introduced the concept of the open architecture initiative and discussed experiences in planning and building the union catalog of the Institutional Project of National Taiwan University. The sharing mechanism of open archives is important to DLs, as it allows the sharing of with each other and provides users with a transparent access channel. The OAI-PMH is an open standard to expose and harvest and service and data providers can communicate with each other more easily and the repositories can keep their original metadata or Dublin Core format. Besides, through the simple and standardizing process to achieve the goal of sharing, using and value-adding, users can search and access resources conveniently. The OAI framework is not intended to replace other approaches, such as Z39.50, but to provide easily implemented and deployed alternative solutions. It will provide a low-barrier interoperability model for service providers and data providers to communicate more easily, and allow users to access information correctly and quickly.

References

Bowman, C.M. *et al.*, (1995), "The harvest information discovery and access system", *Computer Networks and ISDN Systems*, Vol. 28 Nos 1/2, pp. 119-25.

Breeding, M. (2002), "The emergence of the open archives initiative", *Information Today*, Vol. 19 No. 4, pp. 46-7.

Ginsparg, P., Luce, R. and Sompel, H.V. (1999), "Call for participation in the UPS initiative aimed at the further promotion of author self-archived solutions", available at: www.openarchives.org/ups-invitation-ori.htm

Lagoze, C. and Sompel, H.V. (2002), "The open archives initiative protocol for metadata harvesting, v.2.0", available at: www.openarchives.org/OAI/openarchivesprotocol.html

Liu, X. *et al.*, (2001), "Arc-an OAI service provider for digital library federation", *D-Lib Magazine*, Vol. 7 No. 4, available at: www.dlib.org/dlib/april01/liu/04liu.html

Sompel, H.V. and Lagoze, C. (2000), "The Santa Fe convention of the open archives initiative", *D-Lib Magazine*, Vol. 6 No. 2, available at: www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html

Sompel, H.V. and Lagoze, C. (2001a), "The open archives initiative protocol for metadata harvesting, v.1.0", available at: www.openarchives.org/OAI/1.0/openarchivesprotocol.htm

Sompel, H.V. and Lagoze, C. (2001b), "The open archives initiative protocol for metadata harvesting, v.1.1", available at: www.openarchives.org/OAI/1.1/openarchivesprotocol.htm

Suleman, H. and Fox, E.A. (2001), "A framework for building open digital libraries", *D-Lib Magazine*, Vol. 7 No. 12, available at www.dlib.org/dlib/december01/suleman/12suleman.html

Suleman, H. and Fox, E.A. (2003), "Leveraging OAI harvesting to disseminate theses", *Library Hi Tech*, Vol. 21 No. 2, pp. 219-27.

Yu, S.C., Lu, K.Y. and Chen, R.S. (2003), "Metadata management system: design and implement", *The Electronic Library*, Vol. 21 No. 2, pp. 154-64.

(Shien-Chiang Yu received an MS degree in Library and Information Science from Fu-Jen Catholic University and a PhD degree in Information Management from the National Chiao-Tung University in 1997 and 2003, respectively. He is an associate professor and joined the Department of Information & Communications, Shin-Hsin University in 2003. His research interests are: spatial database, electronic commerce, information retrieval, and metadata. E-mail: ysc@cc.shu.edu.tw

Hsueh-hua Chen received an EdD degree in Higher Education and MEd in Educational Media and Librarianship from the University of Georgia, and a BA in Library Science from the National Taiwan University. She is currently a professor in the Department of Library and Information Science at the National Taiwan University. She was also a PI of the National Science Council-funded Digital Museum Initiative project (1998-2000). She is the author of more than 40 articles covering digital libraries, metadata, information organization, knowledge management and serves on the editorial board of many library and information science related journals. Currently, Dr Chen has been heavily involved in fostering digital library and knowledge management search and education in Taiwan, and has continued to receive research grant from NSC for National Digital Archives Program and many research projects. E-mail: sherry@ccms.ntu.edu.tw

Huai-wen Chang received a BEd degree and majored in Library and Information Science from the National Taiwan Normal University. She is a graduate student of the Department of Library and Information Science at National Taiwan University. Her research interest are in digital libraries and interoperability, digital preservation, and distributed retrieval protocols. E-mail: huaiwen@mail.lis.ntu.edu.tw)