Implementing an Institutional Repository for Digital Archive Communities: Experiences from National Taiwan University

Chiung-min Tsai
Department of Library and Information Science,
National Taiwan University
Tel: +886 2 3366 2377
Fax: +886 2 2364 9512
tsaibu@ntu.edu.tw

Jieh Hsiang
Department of Computer Science and Information Engineering,
National Taiwan University
Tel: +886 2 3366 2281
Fax: +886 2 2363 4344
jhsiang@ntu.edu.tw

Hsueh-hua Chen
Department of Library and Information Science,
National Taiwan University
Tel: +886 2 3366 3259
Fax: +886 2 2364 5659
sherry@ntu.edu.tw

Abstract

This paper presents an empirical study of expanding and extending DSpace digital repository system for an academic institution. National Taiwan University created a portal web site, the Digital Archives Resource Centre (DARC), to provide the digital archives communities with preservation base for their rich research materials, and to provide the public with a helpful information retrieval service. Several modifications and extensions for the DSpace system are made in order to integrate various database resources with different formats among university departments. In this paper, we will present our empirical case in which adjustments of the DSpace system are made in areas such as data submission, metadata mapping, digital rights management, user interface and visualization. Our discussion will be focused on the useful applications in which digital archives communities benefit from the adapted DSpace system in disseminating their contents and long-term preservation within NTU campus. We will also discuss the ways in which the users benefit from the new system for searching and using digital archives.

Keywords:
Metadata, Dublin Core, DSpace, Institutional Repository, Digital Archives,

1. Introduction

National Taiwan University (NTU) launched its digital archives initiative as one of the eight institutional projects of National Digital Archive Program (NDAP) in 2002. The goal of NDAP sets on providing digital preservation for the nation's cultural heritages at leading content holders and providers, and utilizing these digital objects to promote the digital content industry in Taiwan [10]. As an essential contributor of NDAP, National Taiwan University created a portal service, Digital Archive Resource Center (DARC), to broaden public access to its valuable heritage holdings.

By gathering the resources from six digital archive communities in NTU, DARC provides an easy access to the university's most valuable digital archive collections. In additions, there are other services such as online backup system with disaster recovery, OAI-PMH service provider and data provider, as well as digital rights and licensing mechanism available in DARC. To accomplish these, we design workflows and provide a set of tools to help the university's digital archive communities integrating their digital contents into the institutional repository. There are two main purposes: first, the digital contents can be used openly for educational purpose; second, they can be further applied in research by faculties inside and outside the university.

According to the practical needs and considering the organizational and technical issues, we set up a Dspace-based prototype and used digital objects from NDAP for experiments. The diversity of these data actually helps us to improve our service in various aspects. After months of testing and modification, now we smoothly run a customized CJK-compatible version of DSpace system, which have successfully integrated about 60,000 records from the six content holders of National Taiwan University.

The purpose of this paper describes the implementation details of DARC, the digital archives' institutional repository in NTU. In section 2, we briefly summarize the historical development of DSpace. Section 3 provides our core applications of DSpace on system architecture, data ingestion, metadata mapping and standardization, and user interfaces, especially the Chinese enhancement. We conclude the paper with the review of strategies, following by a description of future work.

2. Background

A university-base institutional repository is often referred as “a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members [8].” Such institutional repository is also critical to developing, managing, leveraging digital content and bringing greater value of the institution’s digital outputs [4].

As mentioned, DARC is set to provide a uniform interface for the general public access. The major challenge that DARC faces is the lack of interoperability among heterogeneous collections [5]. The digitization of these collections involves a wide variety of content and data types which reflect the features of diverse research fields in the university. DSpace is a standard-based system that implements both the Open Archival Information Systems (OAIS) reference model and the Open Archives Initiative’s Protocol for Metadata Harvesting (OAI-PMH). The DSpace system was also designed to support interoperation with both other DSpace installations and other OAI-compliant archives [13]. The Dspace information model is built around the idea of “communities” that contain one or more “collections” of digital “items”, each of which can have one or more associated digital files, such as images. “Communities” can be defined as schools, departments, centers and programs in the university. In the case of DARC, there are six digital archive communities involved to offer accessibility to their digital objects.

The digital archive communities usually set up websites to present their digital contents in various formats including graphics, texts, video or audio clips. Less experienced users, however, could be unable to find the desired data items due to the unfamiliar interfaces and different ways of organization in each web site. A feasible approach is to integrate those data into a portal site and save the users’ time to examine each site one by one. This is why many universities require an integrated platform to show academic research results online in a uniform display [12]. Those concerns lead us to establish an institutional repository. Institutional repositories are burgeoning services in digital library systems [9]. They are currently being promoted for many kinds of uses around the world. Although there are a lot of surveys and evaluation papers about institutional repositories, we actually lack a real application to demonstrate the power of them in Taiwan.

The DSpace digital repository system provides us a test-bed for our prototype system. DSpace is a developed by HP Labs and MIT Libraries. It captures, stores, indexes, preserves and redistributes an organization's research material in digital formats. There are some other digital repository systems such as EPrints developed by University of Southampton, Fedora jointly developed by University of Virginia and Cornell University, and Greenstone developed by New Zealand Digital Library Project at the University of Waikato. The major

reason we chose to adopt Dspace system is because it contains several stable built-in modules such as OAI-PMH support, indexing and retrieval, and content management. As open source software, DSpace can be easily modified to develop new services upon it [12]. At present DSpace most closely fits our goals of an institutional repository, though, a common and standardized access interface could be deployed across the diverse repository and archival systems [3,6,14].

When implementing DSpace as the institutional repository for digital archive communities in NTU, there were three major issues: the interoperation of varied digital collections, metadata mapping and preservation, and language barriers. The following section describes our efforts to overcome those challenges.

3. DARC: a DSpace-based IR Model Implementation

Our application is based on the 2003 version of DSpace 1.1.1. The system runs on Fedora Linux, and comprises other open source middleware. However, the localized and customized systems were programmed by our team to enhance DSpace performance. All codes are written in Java programming language and are meant to be shared as open source. Other pieces of technological stack include database management systems (PostgreSQL 7.3.x, JDBC), web servers and user interface specifications (Apache, Tomcat, Open SSL/mod_ssl; Java 1.3/1.4, JSP1.2, Servlet 1.3), persistent identifiers (CNRI Handle System 5), indexing/searching (Lucene 1.3) and several useful tools. The followings describe the core elements of DSpace system in the construction of NTU institutional repository, DARC.

3.1 System Architecture

The digital archives in DARC are composed of a large and messy collection of systems, many of which will be distributed across the internet and controlled by a large number of independent players. Therefore, the task of building an institution's digital archives infrastructure is more likely to integrate those systems. Taking into account of interoperability among the participants, we chose DSpace as the platform for the DARC integrated environment, with several enhanced and modified functions. Due to the heterogeneous collections and vast research materials, our digital items usually composed of several digital formats, such as images, texts, as well as multimedia. The DSpace, allowing several files associated to one item, closely fit our purpose. We also implement the OAI

framework together with the handling system. Therefore, contents and metadata are loaded into and replicated at DARC data center.

Figure 1 illustrates the basic system architecture of DARC and also how this integrated system works. Individual digital archive collectors are likely to house significant local collections themselves. Managing large numbers of digital objects over time is a nontrivial task. A generalized metadata repository could be one of the key components of the DARC system. Its purpose is to provide not only a robust service to store, manage, protect, and serve heterogeneous records, but also information and facilities for the preservation of those objects [1].

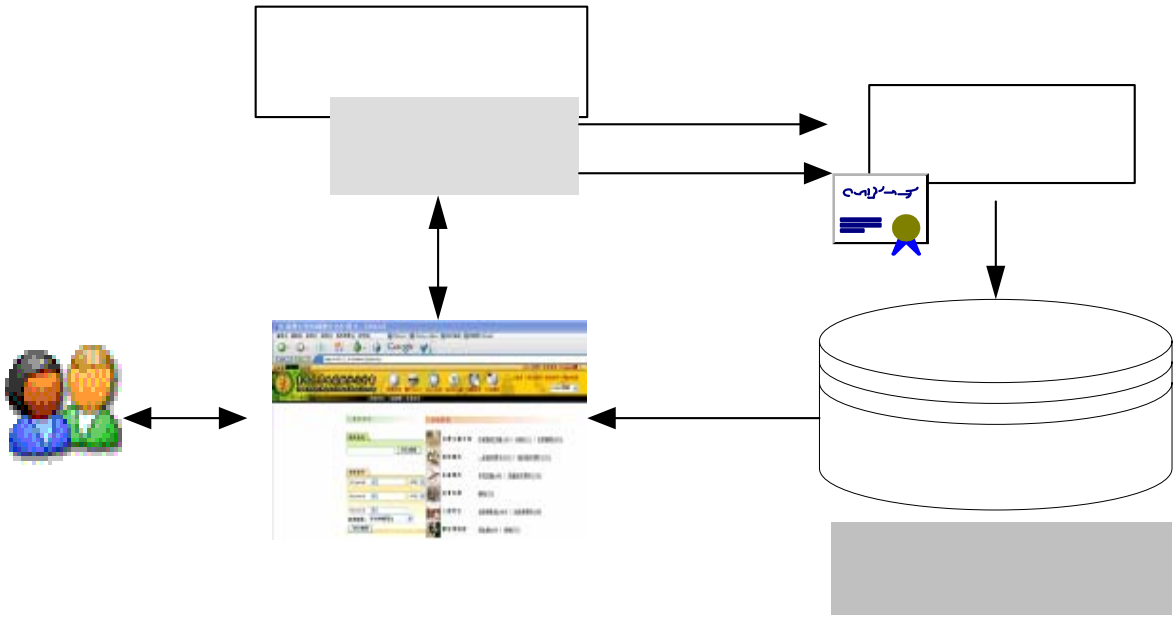


Figure 1. The integrated system framework of DARC

DSpace system is built in 3 layers: storage layer, business logic layer and application layer. Our modifications are actually across these three layers in our implementation. The major enhancements include data ingestion, metadata mapping, metadata preservation, full text search, browse, inter-catalog search, post-classification and Chinese enhancement. The remainder of this section gives brief descriptions to each component.

3.2 Data Ingestion

For each content holder, the instruments and workflow of digitalization are all unique. Therefore, digital archive communities have their own archiving system and use them to create their digital content. To minimize the extra work for submitting items to DARC, digital items are not directly submitted via DSpace Web Submit UI. We apply Batch Item Importer, which is adapted from DSpace, for ingesting data (as shown in Figure 2).

In DSpace, Item Importer stipulates that each item must be placed in separated subdirectory. Each item subdirectory contains a file for the item's metadata and several files, such as images, texts, audios and videos, which make up the item. The metadata must conform to the DSpace simple archive format. There are actually only three data columns in DSpace. Since most digital collections are existed and created differently, it is difficult and unnecessary to ask the data creator prepare their digital items for submission using same format. DARC adopted an easier approach: the data are uploaded in their raw format, and we provide easy APIs to transform them to DSpace simple archive format at server side.

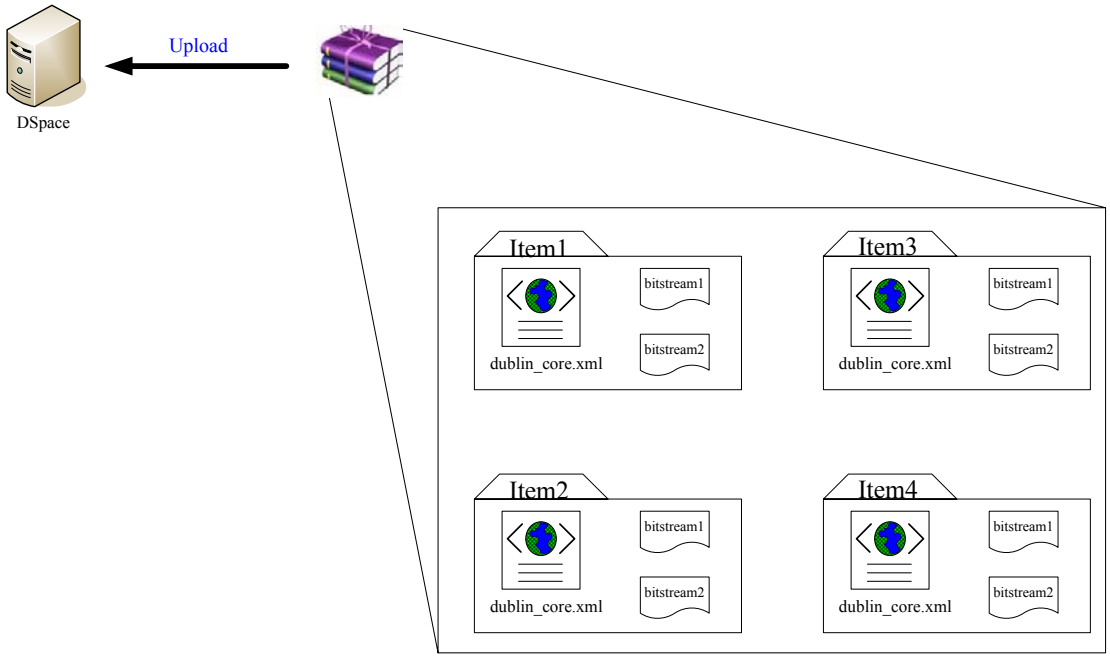


Figure 2. Uploading data with Batch Item Importer

There are two ways to upload items in batches to DARC, through the web interface or ftp. The web interface, however, is less practical because large item batches spend long time to transfer and cause timeouts easily. Updating is done in a similar way. This ingesting process is not done at one stage because the programs need to be instructed how to dispatch files into each item subdirectory, and the metadata mappings described in 3.4 have to be set

appropriately. However, the extra works that all the communities afford are still minimal: the only time-consuming work is to upload the items; and the metadata mapping to Dublin Core is often ready-made.

3.3 Chinese Enhancement

DSpace is designed for English users. It is not well function when translated into Chinese applications. However, it is the most important issue for us to make sure system Chinese compatible since there are collections only recorded in Chinese.

We applied UTF-8 encoding to enable the representation, distribution, storage and search of Chinese data. UTF-8 is a compromise character encoding that can be as compact as ASCII (if the file is just plain English text) but can also contain any unicode characters (with some increase in file size). With UTD-8, we are able to send “http requests” after fixing the bugs of coding in Tomcat. The UTF-8 encoding also allow us, using 1-gram for Lucene, to function support searching in Chinese, like Google can do. In Figure 3, our original Chinese interface and a later but more user-friendly interface are both dsiplayed. With the Chinese enhancement interface, the DARC are brought to a broad herizon. As our localization of the textual elements is in progress, those applications are available and fully shared for internationalization support of the web user interface evolvement.[1]

3.4 Metadata Mapping

Because digital objects and metadata come from many different sources, there is a major inability for the needs of long-term preservation and integration. DSpace uses only one qualified version of the Dublin Core schema, throughout the entire system. System administrators may edit the default Dublin Core registry through the DSpace administration interface to conform to other demands. However, there are more than ten collections in DARC. Each collection has its domain specific metadata format. We may unceasingly add new Dublin Core elements with qualifiers to cope with each metadata format, but it is not the best solution. Instead, some extensions have been developed and added to the original workflow. The idea is simple: we use Dublin Core schema for data search and keep the original metadata for display.

---

[1] http://wiki.dspace.org/I18nSupport?action=AttachFile&do=get&target=Messages_cht.properties

The strategy to deal with DARC metadata issues is then divided into two operations: one is to accept all existing metadata formats from collections; while the other is to provide converting tools for mapping them into Dublin Core. The specification of these operations is described as followings.

(a)

(b)

Figure 3. Chinese Enhancement of cross-catalog search and browse: (a) Original DSpace Interface, (b) User-friendly Interface.

Metadata harvesting: DARC offers two options, OAI-PMH and FTP, to enable digital archives communities submitting their metadata to the repository. The main purpose is to preserve the original data as much as possible regardless of the metadata formats.

Metadata mapping and conversion: It is not required for data providers to offer any extra metadata format. DARC provides applications for mapping original metadata into Dublin Core Schema, which is needed for object management and service development.

3.5 Preservation of Original Metadata

There are two types of metadata recorded in the repository. The original item-level metadata records, which are harvested from the content providers, are supposed to supply sufficient information for general public. In contrast, the standardized item-level metadata records, which are formed from mapping the original ones into Dublin Core schema, are offered for data manipulations such as search, browsing, classification and so on.

When importing the digital items in batches, the first step is to export the original metadata of each item from its source to an XML file. Our sources might be an Excel file, an Access file, a MySQL database, or even a Microsoft SQL server. We have designed a web application to read “tables” and “fields” from those sources, so users can set how the fields mapping into the Dublin Core on the web. Later on, we found that it is important to provide such a tool. The mappings process between the original metadata format and the Dublin Core definitely need to involve domain experts who usually know few about databases. Besides, for better presentation we allow the users to reorder the fields, hide some fields, and rename the fields (because database field names often contain unintelligible acronyms, Most DARC users want to see meaningful field names in Chinese). In order to provide context for item-level metadata harvested from different sources, we plan to create collection-level metadata for the collections in the next stage.

3.6 Indexing and Searching

Both the XML file of the original metadata and the XML file cross-walked to Dublin Core in DSpace simple archive format are generated for each item. The later will be consumed by the DSpace and added to the system database in order to build indexes for searching. The XML file of original metadata is stored as a normal bitstream attached to the item. When displaying an item in DARC, however, it can easily get the corresponding XML file to show more detail contents about it on the web page. Preserving original metadata also enhances the DARC in functionalities, such as building a classification tree or generating statistical results. These improvements are demonstrated in Figure 4.

DARC also offer an integrated and cross-catalog search and navigation. Specific catalogs frequently provide better services to users, as small catalogs are easier to use than large ones. Moreover, topical or format segregation can simplify the finding materials for those users who know precisely their demands. In order to deal with the increasing complexity of our integrated catalog, however, there is a need for a better means of organizing and explaining of what data are available. DARC is designated to provide more organized and coherent resources for users to navigate the more comprehensive interfaces and services. To deliver consistent information, all metadata records are mapped to Dublin Core either by the human effort or with locally-created conversion tools. For improving data retrieval, the DARC research team has developed the post-classification function that shows the total amount of records and their distribution in related catalogs when submitting a search request. Users may discover what else collections they are interested in. Figure 5 shows the full display of search results, including classification and record counts, detail descriptive metadata, as well as full images and texts.

3.7 User Interface

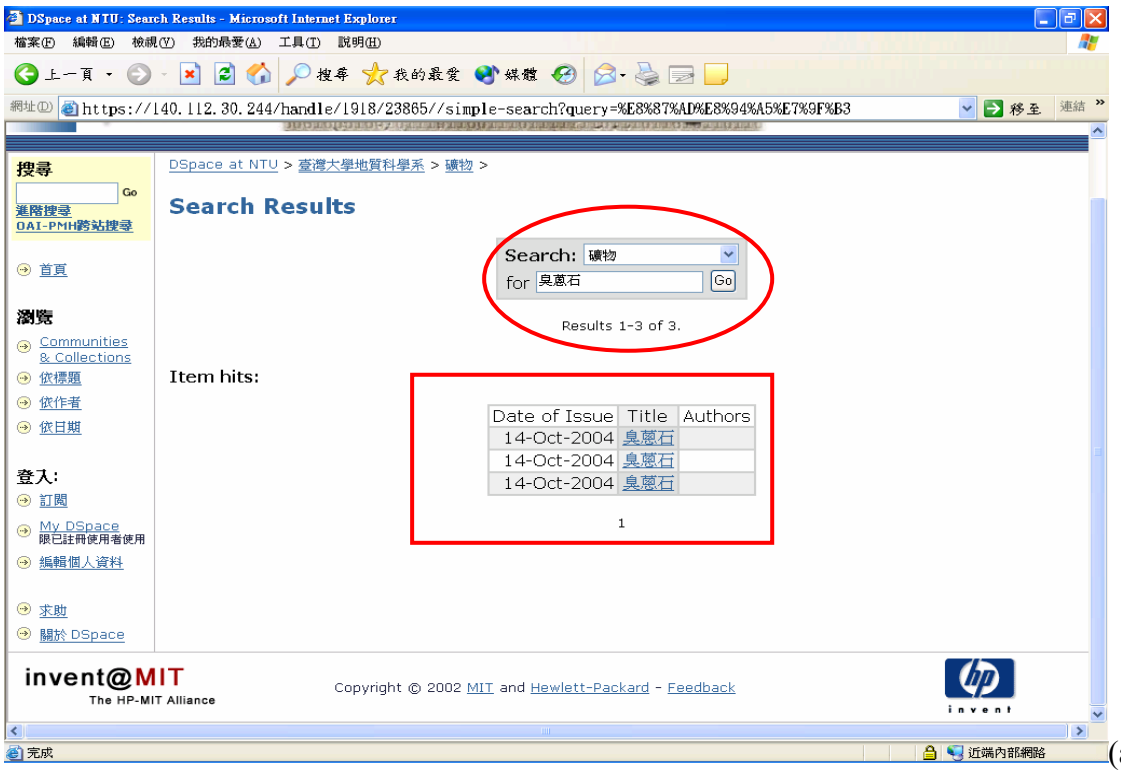
It is a big challenge to design the so-called user friendly interface, which will lead to whether DARC can be used smoothly and conveniently. According our research and experiences, we make significant modifications to the DARC user interface, especially concerning the search and display of results:

Browsing and searching: Both functions are located in one page to keep the interface as simple as possible.

Post-classification: The search results can be regrouped according to the classification for users to narrow down the search. Figure 6 shows the result of keyword search with those functions.

The search results display will include both basic and detailed information (as shown in Figure 5). There are different metadata elements for different subjects. The detailed information will present the original item-level records and provide the link to the original sources of digital objects.

Several services are added, including citation export, search within text, Google search, and hot list.



(a)

(b)

Figure 4. Display of data request from (a)Dspace interface, and from (b)DARC interface, which has added basic information and thumbnail

Figure 5. The display of search results: (a) browsing classification, (b) descriptive metadata (c) images and texts

Figure 6. The display of search results: the detailed browsing information with post classification.

4. Conclusion

Increasing rich digital content is placing greater demands on National Taiwan University, which is responsible for the creation, storage, management and preservation of that content. Owing to the heterogeneity of research nature, data type, storage media and operation system, the integration of individual content provider appears to be a tough task [7]. We totally agree that the metadata is important, when it really took us a lot of time to develop APIs for data mapping and conversion. Then building the classification system to include a variety of subjects was another big challenge. However, the generation of these infrastructures brought us the success of the academic digital archives production.

This paper addresses the experience of DARC, our implementation of DSpace system for the digital archive communities, as an institutional repository. We have developed a number of improved applications and utilities including data storage, metadata mapping, index and search, user-friendly interface and Chinese enhancement. DARC can be found online with collections available from the all our participants and several other services [11]. NDAP is now in the last year of its first phrase of funding from NSC, Taiwan. The second phrase will

be run in next year and through 2011. The phase 2 development plan for DARC is prioritized to first focus on the extensibility. In the next stage, we will:

make digital resources more accessible and useful through DARC services;
develop personalized services and create the usages and applications of digital resources; and
build the digital licensing platform and improve the licensing process of our digital content.

Acknowledgement

The work described in this paper was sponsored by the National Science Council, Taiwan, under the NSC grant number NSC-94-2422-H-002-007.

References

1. W. Arms, D. Hillman, C. Lagoze, D. Krafft, R. Marisa, J. Saylor, C. Terrizzi, and H. Van de Sompel, "A Spectrum of Interoperability: The Site for Science Prototype for the NSDL," D-Lib Magazine, 2002, vol. 8, no. 1. http://www.dlib.org/dlib/january02/arms/01arms.html
2. M.R. Barton and M. M. Waters, "Creating an Institutional Repository: LEADIRS Workbook," 2004, http://dspace.org/implement/leadirs.pdf
3. J. Bekaert and H. Van de Sompel, "Access Interfaces for Open Archival Information Systems Based on the OAI-PMH and the Open URL Framework for Context-Sensitive Services," 2005, http://www.citebase.org/cgi-bin/citations?id=oai:arXiv.org:cs/0509090
4. E. Blythe and V. Chachra, "The Value Proposition in Institutional Repositories." EDUCAUSE Review, 2005, vol. 40, no. 5, pp. 76-77.
5. H.H. Chen, H. W. Chang and C. M. Tsai, "Building Interoperability for Heterogeneous Collections: A Case Study of National Taiwan University Digital Archives Resource Center." The Proceeding of Workshop on the Science of the Artificial 2005, Huailan, Taiwan, December 7-9, 2005, pp.143-151.
6. Digital Library Federation, "Digital Library Content and Course Management Systems: Issues of Interoperation," 2004, http://www.diglib.org/pubs/cmsdl0407/cmsdl0407.pdf
7. C. Lagoze, S. Payette, E. Shin and C. Wilper, " Fedora: an Architecture for Complex Objects and Their Relationships," 2005, http://www.citebase.org/cgi-bin/citations?id=oai:arXiv.org:cs/0501012
8. E. Lee, "Building Interoperability for United Kingdom Historic Environment Information Resources," D-Lib Magazine, 2005, vol. 11, no. 6. http://www.dlib.org/dlib/june05/lee/06lee.html
9. C.A. Lynch, "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age," ARL, 2003, no. 226, pp. 1-7. http://www.arl.org/newsltr/226/ir.html

10. C.M. Morris, “Telling Great Stories: An NSDL Content and Communications System for Aggregation, Display, and Distribution of News and Features,” 2005, http://www.citebase.org/cgi-bin/citations?id=oai:arXiv.org:cs/0509094
11. National Science Council (NSC), National Digital Archives Programs: Introduction, 2004, http://www.ndap.org.tw/1_intro_en/introduction.php
12. National Taiwan University (NTU), National Taiwan University Digital Archives Resource Center, http://www.darc.ntu.edu.tw
13. W. R. Reilly, R. Wolfe and M. Smith, “MIT’s CWSpace Project: Packaging Metadata for Archiving Educational Content in DSpace.” International Journal on Digital Libraries, 2006, vol. 6, no. 2, pp.139-147.
14. M. Smith, M. Barton, M. Bass, M. Branschofsky, G. MacClellan, D. Stuve, R. Tansley and J.H. Walker, “DSpace: An Open Source Dynamic Digital Repository,” D-Lib Magazine, 2003, vol. 9, no. 1, http://www.dlib.org/dlib/january03/smith/01smith.html
15. I.H. Witten, D. Bainbridge, R. Tansley, C. Huang and L. Don, “StoneD: A Bridge between Greenstone and DSpace.” D-Lib Magazine, 2005, vol. 11, no. 9.. http://www.dlib.org/dlib/september05/witten/09witten.html