Building and Sustaining Digital Repositories in Support of Global Information Access and Collaboration

Samson Soong
University Librarian, the Hong Kong University of Science and Technology
Email: soong@ust.hk

Keywords（關鍵詞）: Digital Repositories; Digital Information Access; Digital Archiving and Preservation; Global System of Distributed Repositories; Alternative Repository Models

【摘要】

近年來大學數位化資源庫的發展引人注目，這些數位化資源庫展現了全球合作的新模式--圖書館不再僅僅是為其本地的讀者服務，而是置身於信息存取和保存的全球合作網路之中。在建構效用不同、內容各異的數位化資源庫的過程中，圖書館要應付種種難題，包括：易於存取、認證可靠、成長穩定、支援充足、以及運作持久等。本文旨在介紹香港科技大學圖書館在籌建、實施與維護數位化資源庫，使其使用者能夠有效地存取教學、研究或其他活動所需要的數位化信息內容，並努力增進其在全球分散式資源庫網路中的持久性和互動性之經驗。

【Abstract】

Institution-based repositories of digital content have developed remarkably in recent years. These digital repositories represent an emerging framework of global cooperation in which libraries move beyond serving their local communities and participate in a global network for cooperative information access and preservation.

Developing digital repositories with different purposes and dissimilar contents, challenges librarians with issues of easy access, reliable authentication, steady growth, adequate support, and long-term sustainability. This article intends to share the HKUST Library's experience in developing, implementing, and maintaining purpose-driven digital repositories with the goal of providing users effective access to digital content for teaching, research, or other purposes; while at the same time promoting its sustainability and interoperability in a global network of distributed repositories.

INTRODUCTION

In recent years, many research and academic libraries have established repositories of digital information content, usually Open Archive Initiative (OAI)-compliant.[1] These digital repositories are designed to facilitate archiving, discovery, and retrieval of important information resources. OAI-compliant repositories will increasingly be the conduits through which important digital information is archived, discovered, and delivered. Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH) [2] facilitates a new global distributed model for storage and retrieval of digital information content. [3]

This article seeks to share and reflect upon the experiences of implementing two institution-based repositories at the Hong Kong University of Science and Technology Library and to offer some insights into how best to implement such digital projects and information access strategies to achieve long-term service goals and purposes.

Let me begin with some background information on the HKUST Library and its digital initiatives over the years.

HKUST LIBRARY AND ITS DIGITAL INITIATIVES

Founded in 1991, the Hong Kong University of Science and Technology (HKUST) is a relatively new university. In 15 short years, the University has achieved international distinction for educational excellence, research strength, and institutional flexibility. The University was ranked one of the top 50 research universities in the world by *THES, the Times Higher Education Supplement* (London) in both 2004 and 2005. [4] The University's strategic goals call for strong library resources and services in support of its many research programs and teaching in a wide range of subject areas. This requires HKUST Library to continue to develop its collections, enhance its services, and emphasize the effective use of state-of-the-art technology and advanced information delivery systems.

Since its founding, the HKUST Library has been a leader and pioneer in many aspects of information technology development and implementation. Below is a brief outline of its achievements.

1991	Implemented the first online library catalog in Hong Kong with full Chinese capability.
1991	Set up the first large-scale campus-wide CD-ROM network in an Asian academic library
1993	Designed and implemented the first course reserve full-text image management system in Asia.
1995	First library web server in Hong Kong
1997-2000	Managed a regional mirror site for the Ovid database system
1997-2001	Participated in a consultancy service for the Open University of Hong Kong's Electronic Library Project
2001	Implemented the first native XML-based digital library system in Hong Kong
2002	Developed an XML Name Access Control Repository System as a global prototype
2003	Established the first digital institutional repository in Hong Kong
2004	Implemented the first Web-based digital university archives in Hong Kong

This article centers on HKUST Library's two purpose-driven digital repositories: the digital HKUST Institutional Repository and the Digital University Archives. Implementing these two repositories generated a large number of unique issues and challenges. Although the repositories are up and running, some challenges remain, while the HKUST librarians continue to face new issues in developing and maintaining them.

TWO PURPOSE-DRIVEN REPOSITORIES: THEIR PURPOSES AND FEATURES

Created in May 2003, the HKUST Digital Institutional Repository (IR) (http://repository.ust.hk/) collects, makes available, and preserves the scholarly output of the University community in digital format. It provides a user-friendly interface for self-archiving by faculty and researchers. It organizes the archived digital documents in a logical, easily retrieved fashion. Powered by DSpace, an open-source OAI-compliant software developed at MIT [5],

the IR's archived articles and documents are easily discovered by web search engines, digital discovery services, and electronic indexing tools. For example, the IR is indexed by OAIster (http://oaister.umdl.umich.edu/o/oaister/) where over 6.5 million documents from more than 600 institutions can be found, as well as by Google Scholar (http://scholar.google.com/). The IR uses the CNRI Handle System to assign persistent identifiers to all material posted to the repository. These identifiers are resolvable in perpetuity, and will remain valid even if content migrates to a new digital system. This allows scholarly documents in this repository to be properly and effectively cited in other research work.

As the digital institutional memory for the University, the HKUST Digital University Archives (http://archives.ust.hk/dspace/) is the official online repository for university records that have permanent value. The Digital UA provides online access to faculty, staff, students, alumni, and the general public to non-restricted records and university publications, but not all materials in the University Archives [6]. It took considerable work and ingenuity to create.

In 1997, HKUST Library bought commercial software, BASIS, which staff members then spent a year customizing to meet the requirements of a digital archive. From 1998 to 2004, members of the University could search archived materials by keyword or phrase in the full-text or the title of the document, and/or by the date of the event or publication (using an HKUST Intranet). However, the BASIS software was not completely satisfactory. In October 2004, after new time-consuming software modifications, we successfully changed this digital university archives to a DSpace-based repository system.

The new repository system made possible Internet access to all the public University publications (for example, *The Academic Calendar, HKUST Newsletter, The Alumni News* and *Genesis*) back to their first issues. Going from an Intranet based system to an Internet based one enabled HKUST staff, students, and alumni, as well as the general public to access archived university publications from any corner of the world. The Digital UA also contains newspaper clippings about HKUST and its faculty, staff, and students; photographs, sound and video recordings documenting the various activities and events of the University; and important records of some student and staff associations. The modified DSpace software allows wider access and provides confidential security for the largest body of materials in the UA: administrative records from different offices of the University.

SIMILARITIES BETWEEN TWO REPOSITORIES

The IR and the Digital UA have a large number of similarities. Thus, librarians at HKUST faced (and continue to face) some similar issues and challenges in implementing and supporting these two systems.

- Powered by the open-source software DSpace, both the Digital UA and the IR systems required modification and enhancements.
- Both repositories are institution-based university-wide systems (not departmental or subject-oriented as other repositories can be).
- Both repositories contain only full-text documents or publications (unlike other so-called repositories which merely provide abstracts, citations, or links to other databases or licensed information resources).
- Both the IR and the Digital UA are meant to be permanent, to be maintained for the long term.
- Both repositories require established policies for archiving, retaining, withdrawing, etc. in order to regulate their activities and to meet their intended objectives.
- Both repositories can be accessed remotely from anywhere in the world.

DIFFERENT MANAGEMENT ISSUES AND CHALLENGES

Issues of access and preservation created different management issues and challenges for both repositories. Based on our experience, libraries attempting to simultaneously develop digital repositories with different purposes and dissimilar contents must have a clear vision of the different needs and policies for each one.

Clear Vision of Content

The IR is a repository of research output of the university; while the Digital UA mostly contains the university's administrative records and contents. The fundamental difference in content means that systems enhancements, policies and procedures for the two repositories are quite different. For instance, the IR needs a policy decision on revised versions of a research paper, i.e. whether to keep the first and revised versions in the repository or only the latest version, while the Digital UA does not require such a decision. The Digital UA requires retention policies for different types of administrative records, but the IR usually keeps its records and contents permanently, for another instance.

The content of the IR required us to have a clear sense of what the IR was: an open access repository of HKUST scholarly output. It is not intended for open access publishing, nor does it exist to provide a way to bypass peer-review of a referred publication process.[7] The IR is a means to make HKUST research results freely and effectively available online to the global research community.

The Digital University Archives is a repository of administrative and historical records of the University, as well as various publications and items documenting important events and activities relating to the University. The archival content forced us to be extremely concerned with issues of preservation.

Organizing Principles

Documents in the digital IR are organized by "communities" based on the academic departments or research groups at the University. The materials in the digital UA are organized by "record groups" and "record series" based on the originating administrative and academic units of the University, following the traditional archival principal of provenance. The concept of "communities" in the IR is therefore more flexible. You can make certain collections or documents more noticeable to either a specified group of researcher and users within the University community or to the public. Currently, each academic department at HKUST has a "community" on the IR but more specific communities can be added. For instance, a "molecular biology community" or a "synthetic biology community" can be considered when necessary, in addition to the existing "biology community".

Access vs. Security & Choosing the Right Software

A major challenge for digital repositories managers is how to control access to information that may be sensitive or restricted by copyright or regulations. Project leaders must maintain a balance between protecting sensitive documents and enabling easy access to non-restricted materials and direct IT or systems staff to design mechanisms for controlling content use or misuse. Both repositories required the project leaders to look at issues of access and security, but the different content and purpose of each repository emphasized a different side of the balance. The system or software a library selects to support a digital repository will affect what that library can or want to do. One of the reasons HKUST Library chose DSpace was because we could modify the software when added access control was necessary.

By using the DSpace system (which is OAI-PMH compliant), the library can expose Dublin Core metadata for every document or item in its DSpace-based repository. For material that is restricted to local access, the item metadata is exposed to OAI harvesters but the system will enforce the restriction when a user requests the associated document or article. While some DSpace-based IRs at other

universities have restricted access to certain groups of documents in their databases, we at HKUST have not restricted access to any documents in our IR. We did use DSpace to restrict access to certain files or documents in the digital UA.

OAI compliance allows the archived articles and documents in digital repositories discovered easily by web search engines, digital discovery services, and indexing tools. OAI compliance, however, is more important for the IR than the Digital UA due to the difference in their content and purposes.

HKUST's Digital UA contains a large body of material available for public access, alongside many administrative records currently unavailable for public use. Library Systems staff enhanced the open-source DSpace software to display only public records as a default. Any other document requires a login and password. The system then displays the record groups and publications that are authorized for that particular login. Users never see even the headings of archived records to which they have no right to access.

A good number of faculty members are already in the practice of making their full-text publications available either on their own website or departmental website. To convince them to deposit their publications in the IR, we point out to them that the software they use for their website is not OAI-compliant, thus the contents on their website are less publicly accessible. We also outline other benefits of the institutional repository at our website,http://library.ust.hk/info/db/repository.html#benefits.

Collection Development (Inclusion & Exclusion)

The IR includes whatever has been submitted by HKUST scholars and researchers. In contrast, the Library does not digitize all items submitted to the University Archives. The chief priority for the digital UA is to digitize the most frequently sought university publications and documents.

On a day to day basis, University Archives staff must exercise their judgment in what we include in the Digital UA, in what we keep permanently in our archives collection, and in what we return to the originating offices or departments. Typically, only documents of historical and enduring value will be kept. A set of retention and disposition guidelines for different types of administrative records has been developed to provide guiding principles to originating units or departments on what to send and when to send it to the University Archives. Archives staff has also created a retention schedule for different administrative records.

Content Recruitment (Submission)

HKUST scholars or authors submit documents to the IR voluntarily; but this "voluntary" submission needs to be qualified. At this stage, many submissions are proxy submissions in which the Library staff obtain permission from scholars and do the submissions on their behalf.

In contrast, submission to the Digital UA has become customary with university administrative offices after our initial requests. Although the originating offices are not yet officially required to submit their administrative records, some offices or units at the University send their records to the University Archives as a routine activity. Other units do not. UA staff members often send reminders to previously non-contributing offices and obtain new record series. Thus, while both repositories need to recruit content, at this stage, librarians must be more active in recruiting content for the IR than for the Digital UA.

Permission

Library staff members need to obtain permission from the authors in order to place their documents in the IR. On the submission form, the scholar is asked to indicate that he or she agrees to a non-exclusive distribution license. [8] In contrast, no special permission is required for public access to the University publications and other non-restricted materials in the UA. However, access to the administrative records

in the Digital UA is governed and limited by university regulations.

Our IR collects different types of material, including published peer-reviewed journal articles for which we have experienced more difficulties in persuading faculty or research staff to deposit them. While a large number of them are supportive of the open access cause, they are often concerned about copyright issues. Therefore, we supply a link from our IR website to that of the Romeo/SHERPA Project (http://www.sherpa.ac.uk/romeo.php?all=yes) which provides information on publisher copyright and self-archiving policies, and we also include a summary of self-archiving policies of major publishers.(http://library.ust.hk/info/db/repository.html#intellectual_property). In addition, we check publishers' copyright and self-archiving policies pertaining to each of the individual published articles we add to the repository.

Growth Rate

Presently, there is a considerable difference in the database growth rate of these two repositories. The IR has grown and continues to grow more slowly than the Digital UA. This is partially a result of laws and customs relating to intellectual property.

Publishers, however, increasingly allow researchers to archive their own published content into institutional repositories. Based on average articles published in the journals of the *2003 Journal Citation Report* and publishers listed as part of the Romeo/SHERPA Project(www.sherpa.ac.uk/romeo.php), Thomson Scientific estimated that publishers now allow over half of all scholarly articles to be self-archived by their authors. [9]

There are over 2,100 documents in the IR posted by HKUST faculty members. This is far less than half the number of articles that the same faculty members have published in journals. Why is there such a discrepancy between what is possible and what has been achieved? The main reasons are because faculty members lack awareness of the Repository, and because the faculty members and other HKUST researchers lack incentive to post in the IR. Overcoming such an initial barrier is a critical task for all IR developers. The key is to find a way to make faculty members willing and eager to put their scholarly output into institutional repository. This, hopefully, will lead to a firm commitment on their part to automatically archive and preserve the important contents over time.

In order to develop such eagerness and desire, HKUST librarians have worked hard to increase faculty members' awareness of the IR. These methods include giving information sessions at different academic departments and finding "faculty champions" who submit material themselves and encourage their colleagues to do so. The Library Administration has also worked with the general University Administration to make submitting documents for the IR part of their regular reporting of research output. In addition, the IR has a "Top 20" list. In the beginning years of the IR, author's who made it into the monthly top 20 were informed this by email. By doing so, the Library hoped to generate pride, enthusiasm, and perhaps even a sense of competition among faculty members and researchers at HKUST to share and showcase their research output.

Once we have built up confidence in the user groups in using the repositories, the growth of both repositories will be steady and sustained. Thus, at the current stage, we try to promote both repositories to their respective potential users. The "marketing strategies" we have used help to a certain extent, but we will need to work harder to reach a wider audience.

SUSTAINABILITY AND RELIABILITY OF DIGITAL REPOSITORIES

A responsible and holistic approach to developing and sustaining digital repositories needs

to take many issues and challenges into consideration. Ideally, digital repositories of scholarly and other important digital content will come to act as "trusted repositories" to ensure reliable access to their content over time.[10] In addition to the issues discussed earlier, digital repository developers should consider the associated administrative responsibility, long-term retention and sustainability, technological suitability, system security, and procedural accountability. Repository administers rely on a consensus "across the necessary range of stakeholders of what is to be archived and how it will be done". [11] Administrative responsibility extends to meeting appropriate digital archiving standards, having backup and recovery procedures, and assuring the security of the digital contents. [12]

From the very beginning, repository administrators should commit to implementing appropriate standards and best practices, particularly those that directly influence repository viability and sustainability. Libraries choosing to develop digital repositories should also commit to the long-term retention, management of, and access to digital assets on behalf of depositors and users. The software and hardware used in the operation of a digital repository must assure the security of the digital contents. Administrators should give "special attention to processes that address data integrity to avoid loss of data, to detect any inadvertent changes in data, and to restore lost or corrupted data". [13]

The repository managers or administrators should continually review its policies and procedures to ensure that appropriate growth can occur. Staffing levels and expertise need to be appropriate to the work to be undertaken to support the repository. All repository practices and policies relevant to the use of the repository, including those related to rights management, should be documented and made available to the users. Repository managers or administrators need to consider other long-term relevant issues such as inter-working between digital repositories, especially the prospective for common services and interoperability among different types of repositories.[14] It is also vitally important that libraries with digital depositories work to support and enhance such distributed digital repository architectures, regionally and globally.

CONCLUSIONS

Individuals and organizations are increasingly exploring and studying the digital repositories' role in and impact on Academia (particularly institutional repositories). Recently, seven research institutions joined with Thomson Scientific in a *Web Citation Index* pilot project to explore the proper relationship between *ISI Web of Knowledge*, *Web of Science*, and the world of digital institutional repositories.[15] The findings of this and similar studies in the future will give is greater knowledge of the evolving role and the net impact of IRs and other types of digital repositories.

Digital repositories take shape in a variety of experimental forms and represent new ways of organizing information. They vary in the types of content, in their intended purposes, as well as in their relationship to content creators and users. Despite these differences, the development and promotion of digital repositories will continue librarians' time-honored labor/duty of acquiring, organizing, and making available the resources needed by our users. At HKUST the approaches we have used or are exploring reflect our traditional roles in selecting, evaluating, and providing access to information content. This fulfills our mission to help users find information relevant and critical to their work.

Although digital repositories may be implemented and maintained by the library, we should not view them narrowly as "library projects". To help a repository reach full potential, librarians need to collaborate actively with faculty, researchers, and other staff in administrative departments at our institutions to help disseminate their scholarly output or materials emanating from various

units (if such information is to be made accessible). Digital repositories are never complete on starting day; like gardens, we must cultivate them with attention and vigor.

Digital repositories, if properly implemented, can help us expand our information access strategy to support the information needs of our users. A reliable international network of distributed digital repositories will help provide universal electronic access to important and unique information resources at academic and research institutions. With the help of Google Scholar, OAIster, Elsevier's Scirus, [16] and other emerging search engines, such a distributed network will facilitate global access to digital content from multiple institutions without assembling such content in one place. Networked institutional repositories will help to aggregate virtually open access materials in all subject areas.[17] A prerequisite for such a fantastic network of distributed digital repositories is the good groundwork done individually by all libraries using a holistic and responsible approach to their initial implementation and long-term sustainability.

NOTES

[1] The Open Archives Initiative works to promote faculty self-archiving and interoperable standards for file sharing. OAI protocol (OAI-PMH) is a collaborative effort to develop interoperability mechanisms that facilitate access to distributed digital content in the academic environment. It provides the framework to make it easy to identify, index, and access the content in distributed repositories.

[2] Van de Sompet, H. and Lagoze, C., "The Santa Fe Convention of the Open Archives Initiatives", *D-Lib Magazine*, Vol. 6 No. 2, Feb. 2000 available at www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html .

[3] As David W. Lewis says, "New standards like the Open Archives Initiative's Protocol for Metadata Harvesting make possible a new distributed model for the storage and retrieval of documents. This model will deconstruct the journal as it is currently known into individual articles in much the same way that music file sharing has deconstructed the record album. Repositories of e-prints, images, and other documents will use the developing standards to make these items available to the world." in "The Innovator's Dilemma: Disruptive Change and Academic Libraries", *Library Administration & Management, Chicago*, Spring 2004, Vol 18, Issue 2, p. 70.

[4] World University Rankings, *Times Higher Education Supplement*, London, http://www.thes.co.uk/statistics/international_comparisons/2005/top_unis.aspx?window_type=popup.

[5] "DSpace is a specialized type of digital asset management or content management system: it manages and distributes digital items, made up of digital files (or "bitstreams") and allows for the creation, indexing, and searching of associated metadata to locate and retrieve the items. It is designed to support the long-term preservation of the digital material stored in the repository." MIT DSpace website.

[6] The University Archives, housed in the HKUST Library, is the official depository for the records of the Hong Kong University of Science and Technology. Archives staff members evaluate, preserve, organize, and provide access to documentary sources dating from the founding of the university to the present. These historical records exist in many formats, including handwritten, typewritten, and printed documents; audiovisual materials; university publications; and ephemera.

[7] As James Pringle, Vice President of Product Development of Thomson Scientific, points out, "Open access (OA) publishing is growing in importance, and, in parallel, the role of institutional repositories (IRs) has come to the forefront of discussion within the library community. The two are intertwined but not

synonymous, and different motivations are driving the growth of each." "Partnering helps institutional Repositories thrive", *KnowledgeLink Newsletter*, Thomson Scientific, Feb. 2005.

[8] The non-exclusive distribution license states, "I am submitting this collection of files and associated bibliographic metadata for inclusion in the HKUST Institutional Repository. I hereby grant to The Hong Kong University of Science and Technology (HKUST) the irrevocable, non-exclusive royalty free right to reproduce, distribute, display, and perform this work in any format including electronic formats throughout the world for educational, research and scientific non-profit uses during the full term of copyright including renewals and extensions via the HKUST Institutional Repository mechanisms maintained by the HKUST Library. I also hereby grant to HKUST the non-exclusive right to sub-license these rights to others should the University forego the ability to maintain distribution. I warrant that I have the copyright to make this grant to HKUST unencumbered and complete."

[9] See Pringle, James, "Partnering helps institutional repositories thrive", *KnowledgeLink Newsletter*, Thomson Scientific, Feb. 2005.

[10] For the issues related to "trusted repositories", see "Trusted repositories: Attributes and Responsibilities, an RLG/OCLC Report" at http://www.rlg.org/en/pdfs/repositories.pdf.
There is a similar concept, "digital repository certification", a project being developed by the National Archives and Records Administration and Research Library Group, which "helps to identify digital repositories capable of reliably storing, migrating, and providing access to digital collections" and tries to "define certification requirements, to delineate a process for certification, and to identify a certifying body (or bodies) that can implementthe process". See http://www.rlg.org/en/page.php?Page_ID=580.

[11] Trusted repositories: Attributes and Responsibilities, an RLG/OCLC Report, p. 13.

[12] *Ibid*, p.13.

[13] *Ibid*, p.14-15.

[14] *Ibid*, p.14.

[15] The seven participating institutions are: Australian National University, California Institute of Technology, Cornell University, Monash University, NASA Langley, Max Planck Society and the University of Rochester (see James Pringle, "Partnering Helps Institutional Repositories Thrive", *Knowledgelink Newsletter*, Thomson Scientific, Feb. 2005)

[16] Google Scholar offers a major resource for scientists and academic researchers and maximizes the opportunities offered by open access journals and open repositories. "Well-behaved repositories (without XML or UTF-8 errors)" are updated by OAIster on a regular basis. Scirus is a comprehensive science-specific search engine available on the Internet, linking to more than 167 million indexed scientific pages and documents.

[17] The Digital Library Federation, for instance, has announced recently its intent to create a collaborative digital library that will provide universal electronic access to collections in multiple research institutions. The collaborative library—called the Distributed Open Digital Library (DODL)—is intended to provide global access to collections from multiple institutions without assembling those collections in one place. The DODL will begin by aggregating members' collections of public-domain materials in the humanities and social sciences, will develop an extensive finding service for these collections, and will incorporate numerous other service features to facilitate use of the collections by scholars, teachers, students, and the public. A collections development working group will begin planning content development, and a technical working group will start devising an enabling infrastructure for sharing that content.