A Holistic Approach to Purpose-driven Repositories of Digital Content: Issues and Challenges

Samson Soong, Ph.D.
University Librarian
The Hong Kong University of Science and Technology

Abstract: During the past few years there has been a remarkable development of institution-based digital repositories. In the development of digital repositories with different purposes and dissimilar contents, libraries are being challenged by multiple notions of easy access, reliable authentication, steady growth, adequate support, and long-term sustainability. The aim of this paper is to share the HKUST Library’s experience in developing, implementing, and maintaining purpose-driven digital repositories with a view to provide users with effective access to digital information contents needed for teaching, research, or other purposes.

Table of Contents
1. Concise overview of the HKUST Library and its digital initiatives
2. Purposes and features of two HKUST digital repositories
3. Key differences and similarities between the two repositories
4. Holistic approach to access, growth, support, sustainability issues
5. Conclusions

*******************************************************************

There has been an explosion in studies in recent years examining how ubiquitous information technology is in almost everything we do in libraries. What is abundantly clear is that digital technology is increasingly the conduit for most information transfer and delivery these days and is responsible for how we access, retrieve, retain, manipulate, recall, and preserve scholarly and other valuable information.

This paper seeks to share and reflect upon the experiences of the latest technology applications at the Hong Kong University of Science and Technology Library and to offer some insights about how innovative library digital initiatives and information access strategies are being applied to achieve new service goals and purposes.

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

1. Concise overview of the HKUST Library and its digital initiatives

The Hong Kong University of Science and Technology (HKUST) is a relatively young university. It was founded in 1991. In 14 short years, the University has achieved international distinction for its educational excellence, research strength, and institutional flexibility. The University was ranked among the top 50 research universities in the world by The Times of the UK in 2004. HKUST’s long-term strategic goals call for strong library resources and services in support of its research programs as well as its curricula in a wide range of subject areas. For the HKUST Library, this requires continued development of its collections, enhancement of library services, and emphasis on the effective use of state-of-the-art technology and advanced information delivery systems.

Actually, since its founding in 1991, the HKUST Library has been a leader and pioneer in all aspects of information technology development and implementation. In 1991, it implemented the first library online catalog in Hong Kong with full Chinese capability. In that year, it also set up the first large-scale campus-wide CD-ROM Network in an Asian academic library. The HKUST Library helped design and implement the first course reserve full-text image management system in Asia in 1993. It established the first Library Web Server in Hong Kong in 1995. During the period of 1997 to 2000, the Library managed a regional mirror site for the Ovid database system. At the same time, it participated in a consultancy service for the Open University of Hong Kong’s Electronic Library Project (1997-2001). In 2001, the HKUST Library implemented the first native XML-based digital library system in Hong Kong. A year later, in 2002, it developed an XML Name Access Control Repository System as a global prototype. The Library also established the first digital Institutional Repository in Hong Kong in 2003. And during the last year, 2004, it implemented the first Web-based digital university archives in Hong Kong.

My paper focuses on our two purpose-driven digital repositories, namely, the Digital Institutional Repositories and the Digital University Archives. These two repositories have generated a large number of unique issues and challenges during and after their implementation.

2. Purposes and features of two HKUST digital repositories
The HKUST Institutional Repository (http://repository.ust.hk/) was created in May 2003 to collect, make available, and preserve in digital format the scholarly output created by the University community. It provides an interface which is easy for self-archiving by faculty and organizes the archived documents in a logical, easily retrieved fashion. The Repository is powered by an open-source system DSpace developed at MIT\(^1\) which complies with the Open Archives Initiative (OAI) protocol; thus allowing articles and documents in the Repository to be easily discovered by web search engines, services, and indexing tools. OAI, as many of you know, is a collaborative effort to develop interoperability mechanisms that facilitate access to distributed digital content in the academic environment. OAI provides the framework which makes it easy to identify and index the content in distributed repositories. Documents in the HKUST Institutional Repository (IR) are indexed in OAIster (http://oaister.umd.umich.edu/o/oaister/) where over 5.7 million documents from more than 520 institutions can be found. The CNRI Handle System is used to assign persistent identifiers to all material posted to the repository. These identifiers are resolvable in perpetuity, and will remain valid even if content migrated to a new system. This allows documents in this Repository to be properly and effectively cited in other research.

The HKUST Digital University Archives (http://archives.ust.hk/dspace/) functions as the digital institutional memory for the University, and is the official online repository for university records that have permanent value. Our Digital University Archives (UA) started initially in 1998 after the Library staff spent a year to customize BASIS, a commercial software package acquired to meet the functional requirements of this digital project. During the first 5 to 6 years, via HKUST's Intranet, members of the University could search archived materials by keyword or phrase in the full-text, the title of the document, and/or by the date of the event. After a lengthy and rigorous fine-tuning of the DSpace software, we successfully changed our digital university archives to a DSpace-based repository system in October 2004. This made possible Internet access to all our University publications, e.g. Academic Calendar, HKUST Newsletter, Alumni News and Genesis, dating back to their first issues. This expansion of access, from Intranet to Internet, enables not only our staff and students, but also the general public and our alumni, to access University publications from any corner of the world. The largest body of materials in the University Archives is administrative records from different offices in the University. However the Digital UA also has newspaper clippings about HKUST and its faculty, students, and staff; photographs, sound recordings, and videotapes documenting the various events and activities relating to the University; and records
of some student and staff associations.

3. **Key differences and similarities between the two repositories**

There are a good number of commonalities and differences between these two purpose-driven digital repositories we have established at HKUST. As a result, different sets of issues and challenges are involved in implementing and sustaining these two systems.

Let us first look at the commonalities or similar attributes of these two repositories. Both repositories are powered by the open-source system DSpace with some modification and enhancements. Both repositories are institution-based. As you know, there are departmental or subject-oriented repositories. They perform a similar role to institutional repositories, but only for a particular subject area or for a specific academic department. The two digital repositories I am discussing here are both university-wide systems.

Both repositories contain full-text documents or publications. They don’t simply provide abstracts, citations, or links to licensed or subscribed information resources. Both repositories are meant to be permanent or everlasting and will be maintained for the long term. Both systems necessitate various established policies for archiving, retaining, withdrawing, etc. in order to regulate their activities and to meet their intended objectives. And, both repositories can be accessed remotely, basically from anywhere in the world.

While there are a good number of similar aspects or attributes, there are also major differences between these two repositories. First of all, the Digital IR is a repository of research output of the university, while the Digital UA contains largely administrative contents.

Our Digital Institutional Repository is an open access repository of HKUST scholarly output. It provides a means for researchers at HKUST to make their research results available to anyone, anywhere, at any time. This applies to published journal articles which the publishers allow them to archive in the repository, as well as to other types of scholarly output such as conference papers, theses, or technical reports. Our IR is not intended for open access publishing, nor providing a way to bypass the peer-review and publication process.² It is simply a means to make HKUST research results freely and effectively available online to the global
The Digital University Archives, however, is largely a repository of administrative and historical records of the University, as well as various publications and items documenting events and activities relating to the University. The Digital UA provides online access services to faculty, staff, students, alumni, and the general public to non-restricted records and university publications, but not all materials in the University Archives.

Documents in the Digital IR are organized by “communities” based on the academic departments or research units at the University. The materials in the Digital UA are organized by “offices of records”, both administrative and academic units of the University.

The submission of documents to the Digital IR is voluntary on the part of HKUST scholars or authors. But I need to qualify this “voluntary” submission a bit. At this stage, many submissions are proxy submission in which the Library staff obtain permission from scholars and do the submissions on their behalf. The submission of materials to the Digital UA is somewhat obligatory after our initial requests, although there is no official regulation or requirement yet for the originating offices to submit their administrative records. While some offices or units at the University send their records to the University Archives as a routine activity, many other units do not. We frequently send reminder to these offices and acquire new record series from them. Therefore, content recruitment or promotion is needed for both repositories, but significantly more so for the Digital IR than for the Digital UA during the current stage.

For all documents in the Digital IR, we need to obtain permission from the authors in order to place their documents in a supposedly open access system. As part of the IR submission form, the scholar is asked to indicate that he or she agrees to a non-exclusive distribution license. Public access to the University publications and other non-restricted materials in the Digital UA, however, require no special access permission from the originating offices. But access to the administrative records in the Digital UA is governed and limited by university regulations.

Presently, there is a considerable difference in the database growth rate of these two repositories. The Digital IR grows relatively slower than the Digital UA. But we believe that, 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.

As I pointed out before, the Digital IR is powered by DSpace which complies with the Open Archives Initiative (OAI) protocol; thus allowing archived articles or documents to be easily discovered by web search engines, services, and indexing tools. OAI compliance is considered essential for this open access repository, but such compliance is not necessary for the Digital UA due to the very different nature of the materials in the UA.

Another major difference between the two repositories is their inclusion or exclusion of materials submitted. In the case of IR, we generally include what has been submitted by the HKUST scholars. But we do not digitize all items submitted to the University Archives. For Digital UA, our current priority gives to university publications and to archival documents that are more frequently used.

4. Holistic approach to access, growth, support, sustainability issues

The commonalities and differences between the two digital repositories which I have discussed have generated many different issues and challenges for us. Our experiences in implementing these repositories suggest that those libraries attempting to implement digital repositories with different purposes and dissimilar contents will face very different, and sometimes unique, issues and challenges.

One of the primary challenges for digital repositories is how to properly manage access to information that may be sensitive or restricted by copyright or regulations. Addressing this question requires the attention of systems staff who design software for controlling digital content use and misuse, and project leaders who play an important role in ensuring a balance between protecting sensitive documents and enabling easy access to non-restricted materials. For instance, while a large body of materials in our Digital UA is available for public access, many of the administrative records are not currently available for public use. Therefore, we have made necessary enhancements to the open-source DSpace System so that after a user has logged in, only the record groups and series which he or she is authorized to access will display.
The challenges faced by IR or any open access repository, however, are somewhat different. Powered by DSpace, the HKUST Digital IR is fully Open Archive Initiative compliant. The Open Archive Initiative (OAI) has helped promote faculty self-archiving and interoperable standards for data file sharing. A major result of OAI is its Protocol for Metadata Harvesting (OAI-PMH)\(^5\). The DSpace system supports OAI-PMH as a data provider. A 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 bitstreams. While we at HKUST have not restricted access to any documents in our IR, but other DSpace sites have as they have included in their repositories files or documents that are not openly accessible. The system or software a library has selected to support a digital repository affects what that library can or wants to do. Modifications or enhancements to the software will have to be made when added access control is necessary.

Another major implementation issue or challenge is related to the sustained growth of digital repositories. In the case of our Digital University Archives, “selection” is a very important component or factor in its growth. At HKUST, we in a way consider that as a "collection development" function in building the Digital UA. On a day to day basis, University Archives staff must exercise their judgment, not only in what we include in the Digital UA but also in what we keep permanently in our archives collection and what to return to the originating offices or departments. Typically, only documents of historical and enduring value will be kept by us. A set of retention and disposition guidelines for different types of administrative records has been developed to help provide guiding principles to originating units or departments on what and when to send their materials or records to the University Archives.

In the case of Digital IR, the growth issue is very different. As we know, publishers increasingly are allowing 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 Project Romeo (www.sherpa.ac.uk/romeo.php), Thomson Scientific estimated that publishers now allow over half of all scholarly articles to be archived by their authors. Although there are already over 2,000 documents posted by the faculty at HKUST to our Digital IR, the number of published articles in the Repository is far less than half of the articles they have published in scholarly journals. This is partially due to their lack
of awareness of the Repository or a lack of incentive for these researchers to do so. Since the repository allows them to share and showcase their research output, we have hoped that competition in having more scholarly documents of their own in the Repository will eventually be intense amongst HKUST faculty and researchers. Overcoming such an initial barrier actually is a critical need for all IR developers. The key issue here is trying to turn on the willingness of faculty to put their scholarly output into institutional repository which eventually leads to a firm commitment on their part to automatically archive and preserve the important contents over time.

A more holistic approach to developing purpose-driven digital repositories needs to take many other issues and challenges into consideration. In addition to the issues I have already discussed, digital repository developers ought to think about the associated administrative responsibility, long-term retention and sustainability, technological suitability, system security, and procedural accountability, relying on a shared understanding across the necessary range of stakeholders of what is to be archived and how it will be done.

Administrative responsibility extends to meeting appropriate digital archiving standards, having backup and recovery procedures, and assuring the security of the digital contents. From the very beginning, there should be a fundamental commitment to implementing appropriate standards and best practices that affect its operations, particularly those that directly influence its viability and sustainability. Libraries choosing to develop digital repositories should also be committed 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 should certainly assure the security of the digital contents. Special attention should be given to processes that address data integrity to avoid loss of data, detect changes in data, and restore lost or corrupted data. The repository should continually review its policies and procedures to ensure that appropriate growth can occur. Staffing levels and expertise should also be appropriate to the work to be undertaken to support the repository. And all repository practices and policies should be documented and made available to the users if they are relevant to the use of the repository.

5. **Conclusions**

Digital repositories represent new ways of organizing information content and are taking shape in a variety of experimental forms. 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 purpose-driven digital repositories should be a continuation of our time-honored functions of acquiring, organizing, and making available the resources needed by our users. Some of the approaches we have used or are exploring very much reflect our traditional roles in selecting, evaluating, and providing access to information content. After all, our mission as librarians is to help users find information relevant and critical to their work.

While digital repositories may be implemented and maintained by the library, they should not be viewed narrowly as library projects. Instead, they often require close or creative collaboration with other groups and units on the campus to meet intended needs and to achieve their full potential. To facilitate information access, we should collaborate actively with faculty, researchers, and other staff in administrative departments in our own institution to help disseminate their scholarly output or materials emanating from various units if such information can and should be made accessible. Digital repositories never are complete on starting day. Like a garden, they must continue to receive attention and cultivation.

The role and the impact of purpose-driven digital repositories, particularly institutional repositories, to the research community are increasingly being studied or explored by individuals and organizations. For instance, 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. With the results of these types of studies, we will certainly know more about the evolving role and the net impact of IRs and other types of digital repositories.

Purpose-driven digital repositories, if properly implemented, can help us expand our information access strategy to support the information needs of our users. An international network of distributed digital repositories will help provide universal electronic access to important and unique information resources at academic and research institutions. Such a distributed network will facilitate global access to digital collections from multiple institutions without assembling those collections in one place and help aggregate virtually open access materials in all subject areas. But a prerequisite for such a wonderful network of distributed digital repositories is the good groundwork done by all libraries using a holistic and responsible approach to their implementation and sustainability.
Reference Notes

1 “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.

2 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.

3 The University Archives, housed in the HKUST Library, is the official repository 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.

4 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.”


6 The seven participating institutions are: Australian National University, California Institute of Technology, Cornell University, Monash University, NASA Langley, Max Planck Society and
For instance, the Digital Library Federation 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.