9th International Symposium on Electronic Theses and Dissertations
IXe Symposium international sur les thèses et mémoires électroniques
June 7 – 10 Juin, 2006, Quebec City / Québec, Canada

ETD2006 Conference

Developing an E-theses Collection in the CityU Institutional Repository

Philip Wong
City University of Hong Kong Library
lbphilip@cityu.edu.hk

ABSTRACT

This paper presents the background and process the CityU Library builds the E-theses collection in the institutional repository. Like many institutions, E-theses is not a new product in the CityU Library. Also like many institutions, the demand for a creating a new platform goes up when the limited functions and weak interface of the old system no longer meet the needs of users. Recently, the Library developed the university's IR. It is found that IR is an ideal platform for housing the E-theses collection. IR provides a centralized place for preserving and distributing the quality student works. It can expose the academic works to the Internet via OAI-PMH. The visibility gained will help the Library to obtain approval from authors to digitize and distribute their old theses. From the perspective of building IR, adding E-theses to IR gives a boost to its content and will help to promote IR to the academic communities of the university.

Keywords: ETD, institutional repository, digital libraries, open access, DSpace

1. INTRODUCTION

Like many institutions, E-theses or ETD (Electronic Theses and Dissertations) is not something new but exists for a while in the City University of Hong Kong (CityU) Library. The Library began its digitization project of theses and dissertation in 2002 and launched the existing E-theses collection in 2004. The interface of the existing E-theses collection consists of browse lists created from the theses records downloaded from the online library catalog. Metadata details and links to the abstract and full-text remain in the catalog records. One drawback of distributing E-theses through online library catalog is that the searching capability is limited by the vendor software. Besides, maintaining browse pages is a labor intensive job. In 2005 the Library began to study and develop the university's institutional repository. It was found that the enhanced searching and browsing features of the repository, and its compliancy to the OAI-PMH protocol make it the ideal place to house a new E-theses collection. It can also help to solve a long term accessing problem. Due to some unclear access right policies in the past, a large volume of old theses cannot be opened for academic exchange or to the public. By publicizing their research outputs in the repository, authors of past theses are given a chance to see the value of information sharing and would be willing to grant the right to digitize their works for wider access. From the perspective of building IR, adding thousands of E-theses records to IR will give a boost to the content. The increased usage and collections will help the growth and promotion.

2. CITY UNIVERSITY AND CITYU LIBRARY

City University of Hong Kong (CityU) is one of the government funded universities in Hong Kong. She was found in 1984 as the City Polytechnic of Hong Kong and upgraded to a self-accrediting university in 1994. There are three faculties, four schools, one division and one community college. Courses offered cover subjects in Humanities, Social Sciences, Science, Engineering, Law and Creative Media. Degrees conferred include doctoral, master, bachelor and associate degrees. In the academic year of 2005/06, there are 16,000 full-time students and 9,500 part-time students. CityU is one of the top ranked institutions in the world in research and teaching. The University is also a pioneer in integrating information technology into teaching, learning, and research.

The CityU Library was established in the same year of the University. In 2004/05 the Library has reached a collection of more than 844,200 volumes of books and 179,800 volumes of bound periodicals. It is holding 3,800 print serials titles and subscribes to over 50,000 E-book titles. The integrated library system is INNOPAC. The citation linking software is SFX. There are 19 professional staff and about 90 supporting staff. CityU Library is one of the busiest libraries in the region. In regular semester time, the Library is opened until 1:00 am. The number of entrance in 2004/05 reached 1.85 millions. The high usage and limited space have created a strong demand for electronic resources and the digital collections.

3. EXISTING E-THESES COLLECTION

3.1 Features Of Existing E-theses

As pointed out by many, the provision of E-theses has significant advantages for students, faculty, staff, and the institution as a whole: students can learn more about IT skills and copyright in the process of creating the thesis; electronic copies are likely to be read and be cited more widely; scholarly outputs of the institution can be more widely distributed and the research profile of the whole institution can be promoted (RGU Library, 2006). To unlock this scholarly asset, there has been international and national effort to promote the creation and sharing of e-theses.[1]

CityU's theses and dissertation collection includes papers dated back to 1990. As of 2006 there are 1200 research postgraduate theses (Ph.D. and M.Phil.) and 1300 taught postgraduate dissertations of other degree types. The Library began the digitization project of theses in 2002. In 2003, the School of Graduate Studies required the submission of electronic copy of Ph.D. and M.Phil. theses along with the print copy. In line with this new electronic submission requirement, the Library launched the first and existing E-theses collection.

The existing E-theses collection is built around the online library catalog. MARC records are created for the print copies when they are deposited to the Library. Records can be searched by author, title, subject and keyword. Online links to the abstract and full-text are provided by adding tag 856. To add browsing capability, records are extracted from the online catalog and sorted by titles, degrees and departments. Most users of E-theses start from the browse lists. [Figure 1]

[1] Examples include NTLTD (Networked Digital Library of Theses and Dissertations): http://www.ndltd.org/ and JISC FAIR (Focus on Access to Institutional Resources) Program: http://www.jisc.ac.uk/index.cfm?name=programme_fair

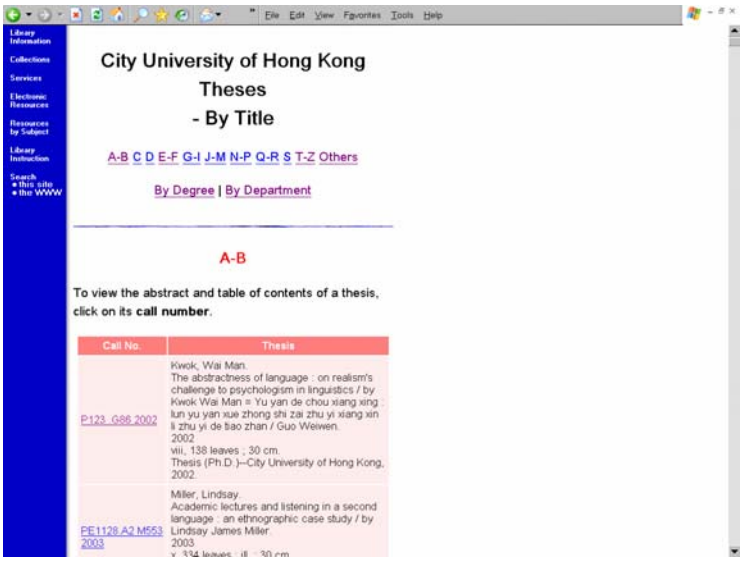

Figure 1. Browsing by title in the existing E-theses collection

3.2 Need For A New Platform

While the browse lists are helpful, features like keyword searching in abstract are still missing. Constructing the browse lists on a regular basis also creates heavy work load on the technical support team. There are other needs to look for a new platform for the E-theses. A good E-theses collection is more than a subset of online catalog records. It is part of the research strength of the institution. It should be made more accessible to the users. Its design and interface should be in line with other digital collections .

4. INSTITUTIONAL REPOSITORY

4.1 Why Created IR

The Library began the CityU IR project in 2004. One motive of building the IR is to support the strategic goal of enhancing knowledge management in the campus. As proposed in the CityU Information Services Strategic Plan 2005-2010 (Office of the CIO, 2005), in the next five years the University will strive to provide the access to scholarly information in electronic and print format in a timely manner. One way to realize the goal is to effectively capture, preserve and disseminate the scholarly outputs of the University.

Building an institutional repository brings other benefits. IR can (i) provide a new publishing paradigm for the scholars and (ii) increase the visibility and prestige of the institution (Crow, 2002). For the first one, IR provides a digital platform for publishing preprints, post-prints, working papers, technical reports, conference papers and other kinds of intellectual outputs. The publishing process also helps to secure intellectual property by establishing priority registration of ideas. For the second one, the strength of the scholarship of an institution can be showcased in the repository. By exposing its content to the Internet via OAI-PMH protocol,[2] the metadata of the documents in IR can be searched globally (Gibbons 2004a).

[2] Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH): http://www.openarchives.org/

4.2 Systems Evaluation

In 2004, the Library formed the IR task force. Members came from Reference, Circulation, Law and Systems. The task force studied and compared common open-source systems. A number of good references have been provided by Crow (2002b), Gibbons (2004b, 2004c), and the Budapest OAI Guide (2004). Combining with local needs, the Library summarized that the CityU IR should meet the following functional requirements.

1. It can support Chinese.
2. It should be OAI-PMH compliant to allow harvesting.
3. It should provide an administrative interface for data input and statistics reports.
4. Its metadata set is flexible enough to describe existing and future digital collections.
5. It allows customization in interface and source.
6. It can be set up and supported by the library with minimal cost.

After careful evaluation, the Library chose DSpace as the IR platform as it fulfilled all the requirements. DSpace is an open source software jointly developed by MIT and Hewlett-Packard. It runs on Tomcat web server, PostgreSQL database system, JSP and Java.[3]

4.3 Technical Configurations and CJK Searching Issues

CityU IR runs on DSpace. As of May 2006, the DSpace version used is 1.3.2. It is built on Linux Enterprise running Tomcat 5 Standalone. The physical server is an Intel Pentium 4 2.8 GHz with 1 GB RAM.

DSpace is built on Java. It can support Unicode. However, the original Java source in DSpace does not tokenize the Chinese characters correctly. As a result, searching of Chinese fails. To solve the problem, the Java class enabling the correct tokenization of Chinese characters is downloaded from the Web before software compilation.[4] It is hoped that in the future release of the software the correct Java class will be included in the source code package.

4.4 Current Collections in IR

CityU IR was launched in February 2006. The home page is http://dspace.cityu.edu.hk/. It forms part of the CityU Library Digital Initiatives (http://www.cityu.edu.hk/lib/digital/).

Current collections in the IR have an emphasis on student works. They include the followings.

Outstanding Academic Papers by Students (OAPS)
This is a collection of outstanding academic papers written by students. The scope includes term papers, term projects, case studies, reports and publications of different courses and levels. Papers are selected by academic departments.

Student Works with External Awards
This is a collection of awards won by CityU students participated in external contests and competitions. Many awards were results of keen competitions organized by prestigious societies and

[3] DSpace Home: http://www.dspace.org/.
[4] The Java class DSTokenizer.java that can tokenize Chinese characters correctly is available from the Apache Jakarta Lucene Sandbox: http://jakarta.apache.org/lucene/docs/lucene-sandbox/.

organizations. The award winning works are in a variety of format including research papers, photographs, audio clippings and video clippings.

Undergraduate Final Year Projects
This is a collection of final-year projects undertaken by undergraduate students for fulfilling their degree requirements. The project reports are selected and submitted to the Library by the departments.

Theses and Dissertations
This is the new E-theses collection. It contains digitized theses and dissertations of CityU students.

Other collections under development include research outputs by academic departments.

5. E-THESES IN CITYU IR

5.1 Why Migrated E-theses To IR

When it comes to select a platform for E-theses, there can be three options:

a. Develop one's own system.
b. Modify from an E-thesis specific software.
c. Build from a generic software platform.

CityU rejected option (a) in the early stage because of the lack of technical support and of the feeling that it is not wise to re-invent the wheel.

Software in option (b) usually contains a submission module. More sophisticated software include a workplace for the advisor to supervise, accept or reject the submission. Thesis-specific software also includes specific metadata set, for example, an "advisor" field, a "degree" field, or a "defense date" field. A frequently used software of this type is Virginia Tech's open-source software ETD-db.[5]

Software in option (c) is not designed specifically for theses. Instead it is a generic solution for archiving and distributing miscellaneous collections. Good examples are DSpace and Eprints.[6]

When choosing a platform for E-theses, one can also ask the following questions.

1. Does it need to include an online submission module? Is a workplace for supervising and approval necessary? How can the submission module be integrated into the existing practice and regulations of the graduate school?

2. Is a thesis-specific metadata set necessary? Would it be an hindrance or a convenience for information interchange? Should the specific metadata be conformed to some standard like the ETD-MS (ETD Metadata Standard) of NTLTD?[7]

[5] ETD-db Home: http://scholar.lib.vt.edu/ETD-db/
[6] Eprints Home: http://www.eprints.org/
[7] Electronic Theses and Dissertations - Metadata Standard (ETD-MS): http://www.ndltd.org/standards/metadata/current.html

Jones (2003) ran a detail study comparing ETD-db and DSpace. He concluded that a direct comparison is hard as each package is driven by different motivations. However, they chose DSpace in the Thesis Alive! project, [8] as DSpace is a better supported package in terms of data structure, security, administration and future expansion. What can be added is, the fact that the ETD-MS schema supported by ETD-db is not totally OAI compliant may create an obstacle in metadata harvesting.

CityU Library chose to migrate E-theses to CityU IR for the following reasons.

1. DSpace IR is OAI-PMH complaint. It is ready to expose the records to harvesting engines in the Internet.
2. By publicizing the theses in the Internet, more authors of the past would be likely to grant the access rights to digitize and distribute their theses.
3. E-theses in CityU emphasizes on post-submission workflow, there is no demand for paper authoring and supervision. If such needs arise, submission add-on like Tapir can be evaluated.
4. Record is assigned with persistent identifier (CNRI handle), this provides long term reference to the materials.
5. Grouping theses together with other student works under IR enables a one-stop search for all quality student works. This provides a unified platform to showcase the academic achievements of students.
6. Strategically, by adding thousands of thesis records to the content, it provide a fast start-up to the growth of IR, and to gain recognition from other academic communities of the university.

5.2 Metadata In E-theses

As OAI-PMH only supports the Dublin Core metadata schema, non-DC elements are avoided when defining new metadata. If needed, new metadata elements are added as qualified DC elements.[9] Table 1 lists the metadata elements in the E-theses in IR.

[8] Thesis Alive! at Edinburgh University Library: http://www.thesesalive.ac.uk/

[9] However, since only unqualified DC elements are supported by OAI-PMH, qualified DC will be flattened to unqualified DC when being harvested, as a result, useful field information may become lost.

Table 1. CityU E-theses Metadata

Unqualified DC	Qualified DC	Meaning
title		title in English
title	alternative	title in Pinyin romanization
title	alternative	title in Chinese
contributor	author	author name in English
contributor	author	author name in Chinese
contributor	department	department
date	issued	publication date
publisher		publisher
subject		subject heading
description		physical and bibliographic description
description	degree	degree
description	abstract	abstract
identifier	catalog	link to record in online catalog
identifier	uri	CNRI handle

5.3 Migration Process and Current State

As online catalog records for theses already exist, there is no need to re-input records in DSpace. The migration process includes the following steps: (i) MARC records are exported from the catalog. (ii) A conversion program will map and convert the MARC tags to Dublin Core elements in XML format. (iii) Chinese characters are converted from internal diacritic codes (EACC) to Unicode/UTF-8 for storing into DSpace.[10] (iv) New metadata fields will be added. These include a "department" field to describe the academic department where the degree was earned, a "degree" field and an "abstract" field. (v) The Dublin Core XML records will be imported into DSpace using the import utility.

It turns out that adding the abstract field is the most time consuming job as it involves running OCR on over 2,500 previously scanned PDF files. It is expected that the OCR job will be finished in June 2006 and the migration work be completed in July.

5.4 Hierarchy Of E-theses Collections In IR

The E-theses in DSpace are organized in a hierarchy of community, sub-communities and collections. The top community is Electronic Theses and Dissertations. Under the top community, there are sub-communities created for departments. Under each sub-community, there are collections created for degrees (Ph.D., M.Phil., M.A., M.Sc. etc). [Figure 2]

[10] In MARC 21 specification, Chinese characters are encoded in the EACC (East Asian Coded Character) set.

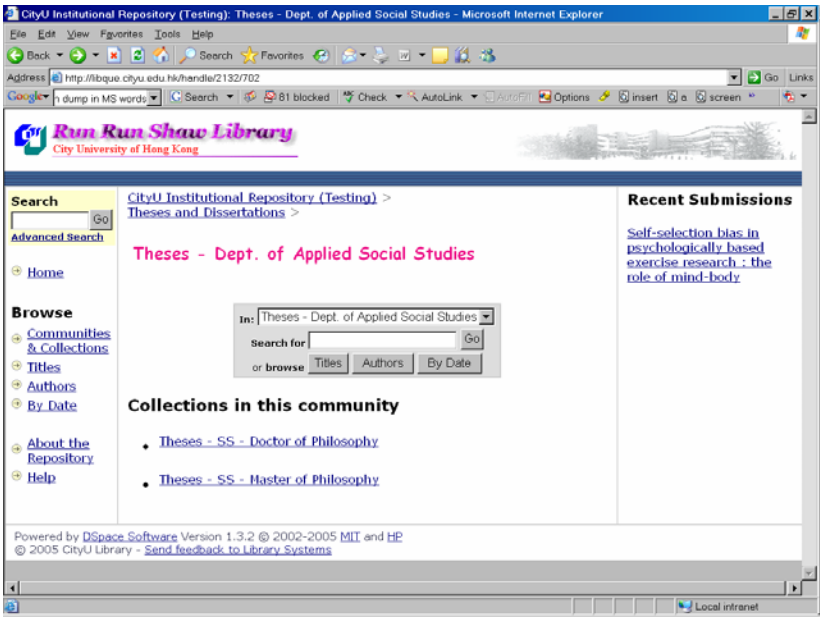

Figure 2. Community, sub-communities and collections in new E-theses

The nice thing about DSpace is that one can have an item appears in more than one collection. If a new community is created for an academic department, a thesis or dissertation can be mapped back to its degree-granting department. By providing cross mapping, a user can search a thesis item by starting from either the owning community of the thesis or from the departmental community. [Figure 3].

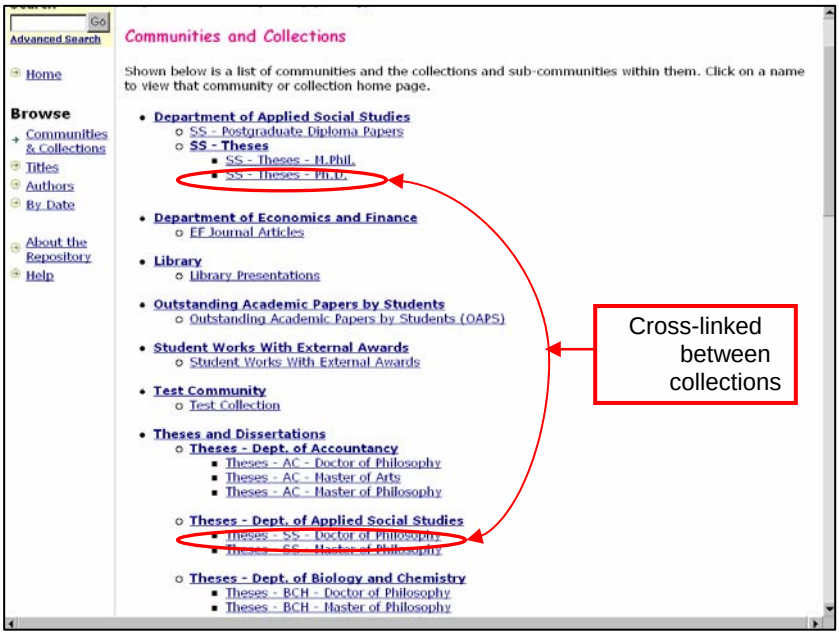

Figure 3. Theses items mapped between collections.

5.5 Access Rights

As most existing theses were submitted before the electronic submission requirements issued in 2004, due to unclear access right policies in the past, a number of theses can not be opened for academic exchange or for the public. To obtain permission from authors, in 2005 the Library sent letters to alumni to ask for approval. The response rate was not satisfactory. About one third had returned. The Library will spend more effort to contact the alumni again. By providing a publishing platform in IR, it is likely that more authors will agree to grant rights to digitize and distribute their theses.

5.6 Future Enhancements

Submission of theses and dissertations are currently managed by the School of Graduate Studies for research postgraduate degrees and by individual departments for taught postgraduate degrees. Electronic copies in PDF format are passed to the Library. To automate the submission process, an online submission module can be added. Such module can provide file uploading function and simple metadata input interface for the academic units. A more advanced submission module will include a supervised authority facility allowing the supervisors to observe the ongoing work and to approve or reject the project online. Adoption of such advanced module requires changes in the submission policy in the academic units. At current there is no such demand in CityU. An example of the advanced submission module is the Tapir Project developed by the University of Edinburgh. [11] (Andrew 2004)

Another possible feature to be added is full-text searching. However, as the electronic requirement is new and most of the old theses are digitized from print copy, full-text searching for all can not be provided yet.

6. CONCLUSION

The Library is in the process of migrating the E-theses collection from the old browsing interface to the CityU IR developed in DSpace. There are many advantages of relocating E-theses to the new home. Records can be harvested by Internet engines easily via OAI. Research outputs of students will be more visible. Searching and display of records are enhanced. Authors of the old theses will be likely to grant access rights to the Library once their theses are published in a more visible platform. All quality student works can be searched in a single platform. Lastly, by adding E-theses into IR, it provides a boost to the content of IR. The increased usage and collections will help to promote IR to other academic communities of the University.

7. REFERENCES

Andrew, T. (2004). Theses Alive! : an E-theses management system for the UK. Available at: http://hdl.handle.net/1842/423 (accessed 23 May 2006).

Budapest OAI Guide (2004). Budapest Open Access Initiative: A Guide to Institutional Repository Software. Available at http://www.soros.org/openaccess/software/ (accessed 23 May 2006).

Crow, R. (2002). The case for institutional repositories: a SPARC position paper, *ARL Bimonthly Report, 223*. Available at: http://www.arl.org/sparc/IR/ir.html (accessed 23 May, 2006).

[11] Tapir (Theses Alive Plug-in for Institutional Repositories) for DSpace, developed by Edinburgh University Library. Home: http://www.thesesalive.ac.uk/dsp_home.shtml

Crow, R. (2002b). SPARC institutional repository checklist & resource guide. Available at: http://www.arl.org/sparc/IR/IR_Guide.html (accessed 23 May, 2006).

Gibbons, S. (2004a). Benefits of an institutional repository. *Library Technology Reports. 40*(4), pp. 11-16.

Gibbons, S. (2004b). Features and functionalities. *Library Technology Reports. 40*(4), pp. 31-40.

Gibbons, S. (2004c). Institutional repository system overviews. *Library Technology Reports*. 40(4), 41-53.

Jones, R. (2004). DSpace vs. ETD-db: choosing software to manage electronic theses and dissertations. *Ariadne, 38*. Available at: http://www.ariadne.ac.uk/issue38/jones/ (accessed 23 May 2006).

Office of the CIO (2005). *Information Services Strategic Plan 2005-2010, Office of Chief Information Officer, City University of Hong Kong*.
Available at: http://www.cityu.edu.hk/cio/cist/issp05/issp05.html (accessed 23 May, 2006).

Open Society Institute. (2004). Budapest Open Access Initiative: a guide to institutional repository software, v. 3.0. Available at http://www.soros.org/openaccess/software/ (accessed 20 May, 2006).

RGU (Robert Gordon University) Library. *Why ETDs?*
Available at: http://www2.rgu.ac.uk/library/etds/why.html (accessed 23 May, 2006).

ABOUT THE AUTHOR

Philip Wong is the Systems Librarian at the City University of Hong Kong Library, where he oversees the library systems and digital library projects. Before joining CityU he worked for the Hong Kong University of Science and Technology Library. He holds a MS in Library Science from the University of North Carolina at Chapel Hill and a MS in Computer Science from the State University of New York at Buffalo. His research interest are in digital libraries and archives, and Chinese information processing in library.