ETDs at HKU: From 0 to 13,000 in 5 Years!

David Palmer
Technical Services Support Team Leader, The University of Hong Kong

Keywords: Electronic thesis, retrospective conversion, born-digital

ABSTRACT

The University of Hong Kong began its ETD program in 2000 after convincing the HKU Senate of the merits of the initiative. We have received 1,000+ born digital theses. Through a series of fortuitous events, all retrospective print theses will soon be digitized. This has ramifications for inter-library loan, and physical shelving. Along the way we learned several strategies for dealing with copyright issues, and access. We have stopped our abstract submission to UMI, and now rely on OAIster, Yahoo, Google, and others to provide information discovery & retrieval to potential users.

1. BACKGROUND

The University of Hong Kong (HKU), established in 1912, is the oldest in Hong Kong. It now has 12,000 students. The first recorded thesis was done in 1928, though all previous to 1941 were lost during the occupation of WWII. The Library now holds 13,000 thesis titles in PhD, MPhil and other degrees. We currently receive about 950 printed theses every year.

In 1999, in order to make our theses more visible, the Library created a new database, “Hong Kong University Theses Online” (HKUTO). USMARC21 records were extracted from the catalogue and uploaded into an Oracle database. It provided searching in Roman and Chinese on author, title, subject (LCSH), as well as on degree and program. Because of copyright concerns, we limited our digitizing to abstracts and tables-of-contents in image-only PDF.

Approximately at this same time, Ed Fox of Virginia Tech visited HKU to evangelize electronic theses and the NDLTD.

2. BORN-DIGITAL FULLTEXT

As a member of HKU's Academic Council for IT in Education (ACITE), I suggested and received agreement that we make a proposal that HKU begin a program for fulltext electronic theses. In November 2000, this proposal was adopted by the HKU Senate for the research degrees of PhD and MPhil students entering from the January 2001 intake. It is a mandatory requirement but allows for students to "opt-out" if they wish. After this time, most of our ten faculties then adopted this program for their other non-research degrees. We received our first two born-digital theses in November 2001. We have now received 1,200 electronic theses under this program.

3. ACCESS TO HKU THESES

UMI. In 2000 faculty suggested that the University begin sending theses to UMI in order that our theses become more visible and used. However the question of who would pay for these submissions, and the difference in requirements between UMI and HKU for thesis format was problematic. Therefore the University chose instead to join the NDLTD, believing in time that this route would eventually provide a universal search engine on global theses and prove more economical than UMI. In this regard, using VTOAI OAI-PMH2 PERL Implementation, we made our HKUTO data OAI compliant in 2002. Our thesis data was some of the first harvested by VTLS' Networked Digital Library of Theses & Dissertations, and the University of Michigan's OAIster.

However the several initiatives of NDLTD for federated searching and union thesis catalogues proved slow in materializing. Therefore in 2003 the Library and the Graduate School decided to begin a pilot project for one year and submit abstracts for PhD and MPhil students to UMI. The Library agreed to pay for these submissions for the length of the pilot.

At the end of this pilot in January 2005, we had submitted 425 abstracts to UMI. We evaluated and decided that we would not continue these submissions for the following reasons.

1. Collecting and processing of the UMI applications by the Library proved tedious and time consuming. Many eligible students did not submit as there were no

penalties for not doing so. Asking the students to submit online directly to UMI would still require an authority at the University to approve the applications and verify that they were indeed HKU graduates.

2. The Library funding for these submissions was cut.
3. While paying for abstract only submission to UMI was cheaper than fulltext submission, it was found that the University still wanted fulltext online theses, and thus must produce them itself. No cost saving was achieved.
4. The record in UMI provided abstract only, with no link to the online fulltext available for worldwide access at HKU. Remote users would thus ignore our theses, or assume wrongly that the only access was through offline application to our inter-library loan office.
5. Several popular search engines began to index HKUTO. Unlike UMI with HKU abstracts only, they also provide links to our fulltext online theses.

Google et al. In March 2004 Yahoo announced their Content Acquisition Program (Sherman 2004) and began harvesting OAI data from OAIster and other OAI data providers. HKU theses are now thus discoverable and accessible from Yahoo. Google Scholar has released a crawler upon our HKUTO, which will soon make HKU theses discoverable and accessible from there also. Other search engines covering HKUTO are, OCLC's XTCat NDLTD Union Catalog, and Elsevier's Scirus.

Acknowledgement Page. In May 2004 the Graduate School sought legal counsel on our provision of worldwide access to the HKU electronic theses. We received advice that users should be asked to explicitly state their agreement to terms of usage for the thesis, in order to remove the University from risk. We therefore added an intervening acknowledgement page which asks the user to agree to

1. No commercial use of this thesis or any part of its contents is allowed.
2. No uses are allowed except those for the purposes of scholarship or research
3. I agree to use this thesis under the terms of the Hong Kong SAR Copyright Ordinance.

They also must input their email address, to which we will then send a URL of the fulltext thesis. An unforeseen benefit is that we are now able to count downloads in terms of titles downloaded instead of number of files (eg. five files associated with one title).

Usage Statistics. While usage of the print thesis collection as remained steady,

usage of the online theses has climbed to five times that of the print. Fiscal year to date at end of May 2005 showed 51 inter-library loans and 14,819 checkouts of print titles. During this same period there were 68,299 downloads of electronic titles.

4. RETROSPECTIVE CONVERSION

HKU holds the policy that thesis authors hold the copyright. For many years the idea of digitizing retrospectively the entire collection seemed impossible, with the main obstacle being the need to contact each author individually to gain his or her permission. In 2004 this all changed.

HKU Departmental Libraries. Due to the distributed nature of the HKU faculties and their governance over the thesis producing programs, it was only in 2004 that the Faculty of Social Sciences joined the HKU program of mandatory e-theses for their non-research thesis programs. An unexpected discovery was that for the previous two years, the Department of Psychology had been scanning their non-research print theses into image-only PDF. As do most institutions in Hong Kong, The University of Hong Kong suffers from a lack of space. The constraint of space for storage meant that there was no longer room to continue to receive and store theses. The Library quickly received these 90 titles and included them in HKUTO. In the future Psychology will retain only a copy of the e-thesis and return or discard the print. Most of the HKU Departmental libraries find themselves faced with similar space problems. The Department of Earth Sciences passed to us all of their non-research theses and gave us permission for digitizing. We expect that lack of space will soon force other departments to do the same.

Individual Permission. With this development, retrospective conversion no longer seemed impossible. We gathered postal and email addresses from the Alumni Affairs Office and the Registry and did a mass mailing in December 2004, from which we are still receiving responses. We have received so far, 100s of "return-to-sender" and bounce-backs, 60 negative responses, and 1,600 positive responses. Our letter included this paragraph,

> "If we don't hear anything from you we will assume that you approve. However, should you decide at a future time that you do not want your thesis included, simply notify us in writing and we will immediately remove it from the database."

Many authors still retained machine readable files and sent them to us for conversion. Many had questions about copyright. In most cases an explanation that placing the thesis online would actually reduce the chance of plagiarism allayed their worries. Since the autumn of 2004, HKU has used Turnitin for the prevention of plagiarism. We requested them to send their crawler on HKUTO, and thus make the detection of plagiarism on HKU theses much easier.

With approximately 11,000 bound theses to scan, it would have taken our 4-man Reformatting Team several years to complete. Fortunately the synchronicity of events provided a much better alternative.

Million Book Project (MPB). Dr. Raj Reddy of Carnegie Mellon University founded the Universal Library with partners in India and China. In 2003 HKU joined the MBP, which was the first project of the Universal Library. The group of 19 Universities in China for the Universal Library became known as the China - America Digital Academic Library (CADAL). Besides their Chinese language titles, CADAL agreed to supply the MBP with 100,000 digitized English language titles. Working with the guideline that these titles should be ones that were published before 1923, they discovered that they did not have nearly enough such titles in their collection. They therefore asked HKU to aid in the supply of these titles.

We believe that HKU will eventually ship 60,000 titles to CADAL for digitization. These will include pre-1923 imprints, rare books, special collections, HK Government publications, and all of the print-only HKU thesis collection. The first shipment of 15,000 will go out in June 2005. Costs for digitization and shipment will be covered by the MBP.

An Observation. The Google Print project, CMU's Universal Library, and others will quickly place millions of titles online. Most of these titles will be ones that are held by many libraries worldwide. Inevitably much redundancy will occur. Therefore, an individual library's unique worth in this endeavour is providing unique titles, such as the Hong Kong government publications and the HKU thesis collection.

5. CONCLUSION

After all titles are digitized, print copies will be moved to our off-site storage to free valuable shelving in the Main Library. We expect that this will act as an incentive for

the departmental libraries to stop collecting print copies from their students. Our inter-library loan will stop lending HKU theses, and refer the user of any such requests to our HKUTO repository.

Submission of e-theses at HKU is now an accepted practice. The two remaining faculties without the e-thesis requirement are expected to soon adopt this requirement. The number of e-theses at HKU has reached a critical mass. HKU is a member of the Digital Dissertation Consortium hosted by the Academia Sinica Computer Center of Taiwan and thus can access the 38,000 UMI fulltext titles mounted therein. HKU Library hosts a mirror site for the China National Knowledge Infrastructure (CNKI), among which are 50,000 fulltext thesis volumes done in China. HKU's entire collection of 13,000 thesis titles, minus a few refusals, will be online at the end of 2005. With these numbers, HKU expects to very soon celebrate access to one million e-book (monograph) titles.

6. REFERENCES

Academia Sinica Computer Center. (2002-2005). *Digital Dissertation Consortium*. Retrieved May 22, 2005 from http://pqdd.sinica.edu.tw/twdaoeng/basic.html (requires login).

Carnegie Mellon University. (n.d.). *Universal Library*. Retrieved May 22, 2005 from http://www.ulib.org/html/index.html

China – America Digital Academic Library. (2004). *China – US Million Book Digital Project*. Retreived May 22, 2005 from http://www.cadal.net/

Digital Library Research Laboratory, Virginia Tech. (2002). *VTOAI OAI-PMH2 PERL Implementation*. Retrieved May 22, 2005, from http://www.dlib.vt.edu/projects/OAI/software/vtoai/vtoai.html

Elsevier Ltd. *Scirus*. (2005). Retrieved May 22, 2005 from http://www.scirus.com/srsapp/

Internet Archive. (n.d.). *Million Book Project*. Retrieved May 22, 2005 from http://www.archive.org/details/millionbooks

iParadigms, LLC. (1998-2005). *Turnitin*. Retrieved May 22, 2005 from http://www.turnitin.com/

Libraries, The University of Hong Kong. (1999-2005). *HKUTO*. Retrieved May 22, 2005, from http://sunzi.lib.hku.hk/hkuto/

Qinghua Tongfang Guangpan Gufen Youxian Gongsi. (1999-2005). *China National Knowledge Infrastructure (CNKI)*. Retrieved May 22, 2005, from http://cjn.lib.hku.hk/ (requires login).

Sherman, C. (2004). Yahoo Announces Content Acquisition Program. *SearchEngineWatch*, 2 Mar 2004. Retrieved May 22, 2005 from http://searchenginewatch.com/searchday/article.php/3320071

University of Michigan Digital Library Production Service. (2002-2005). *OAIster*. Retreived May 22, 2005, from http://oaister.umdl.umich.edu/o/oaister/

VTLS, Inc. (2001-2005). *Networked Digital Library of Theses & Dissertations*. Retrieved May 22, 2005, from http://zippo.vtls.com/cgi-bin/ndltd/chameleon