Using Islandora Digital Repository for Developing the CUHK ETD Collection

Presented by
Jeff LIU, Louisa LAM
Research Support & Digital Initiatives
The Chinese University of Hong Kong Library
August 7, 2017
About The Chinese University of Hong Kong (CUHK)

- Established in 1963
- A comprehensive research university in Hong Kong comprising of 9 Colleges, 8 Faculties and a Graduate School
- Undergraduates: 16,583 (2016)
- Postgraduates:
  - Research Postgraduates: 2,531 (2016)
  - Others: 1,033 (2016)
- PhD and MPhil awarded in 2015/16: 619
The Journey of ETD Collection at CUHK

2011
- Print theses at 9,968 vols. from 1967 to 2011 available at the Library
- Doctoral theses submitted since 1997 available online in ProQuest’s “Dissertations & Theses@CUHK”

2012
- Decided to create an online ETD Collection that supports open access to provide a one-stop platform to all CUHK theses and dissertations
- Embarked a mass digitization program to scan all print theses

2014
- Launched ETD on Dec 15, 2014 on a local stand-alone platform
- More than 9,000 volumes with over 1 million digital images

2016
- Migrated the collection to a new Digital Repository built on Islandora on Sep 1, 2016
Why Such the Move?
Features of the ETD Collection

- Includes both English and Chinese materials
  - Of all 13,629 titles, about 2,000 titles are in Chinese
    - Computer generated / typewritten / handwritten
    - Horizontal / vertical display of Chinese texts
- OCR of all English and Chinese theses
  - OCR accuracy rate:
    - Body: English (96%), Chinese (95%)
    - TOC and abstract: 99%
    - To enhance the OCR accuracy rate of theses in handwritten Chinese, a total of 438 titles in 490 volumes (169,496 images) were typed out
- Searchable PDF for whole document, TOC and abstract
明代王門諸子之良知學

http://repository.lib.cuhk.edu.hk/en/item/cuhk-321360
上海商務印書館與近代知識文化的傳播和塑造
(1897 至 1949)
--- 從書籍出版史角度考察

李家駒

歷史系
哲學博士論文

© 香港中文大學
2001 年 12 月

(此論文之版權屬香港中文大學所有。任何人士如欲發表論文之全部或任何部分，須先獲得研究院院長同意。)
Chinese Theses (submitted before 1990s)

- Chinese texts written by hand and displayed vertically
- OCR text can be found and located correctly in the document
Platform Development Considerations

- Different web sites with different user interfaces were created for different projects.
Fragmented Web Sites created development limitations and unsatisfactory user experience
Development Limitations (1)

- Different web interfaces for browsing and searching cause confusion to users (No single theme and under CUHK Library branding)
- Inconsistent look and feel
- Unable to perform one-stop search across all collections
- No standardized metadata schema across different projects
- Difficult to maintain each system in case of any breakdown
Development Limitations (2)

- Software and System Architecture
  - Tamino XML Database offers:
    - Good support of CJK handling including:
      - Character/phrase search
      - Simplified / Traditional Chinese conversion indexing and searching
    - But it is not very popular
      - No application frontend
      - Costly to maintain
      - Not OAI-PMH compliant for discovery
Interim Plan Before Developing A One-Stop Platform

Develop a single interface for all digital collections

Tamino XML server

Search result displayed separately on each DBs in their original layout
Why an Interim Solution?

- Take a shorter time for implementation; there is a need to launch ETD soon
- An unified web interface to provide better user experience
- Thematic design allowed for individual project
- Still use Tamino XML server to provide CJK support needed by ETD
- Can enable OAI-PMH for open access and system interoperability
- Staff mindset not yet ready for an open source collaborative development
The new local platform was launched in December 2014.
Features of the Stand-alone Platform (1)

- ETD together with other digitization projects are listed in an unified web interface, yet allowing search in specific fields required by ETD
- Built on Drupal CMS using iframe to display the search results and browse pages
Features of the Stand-alone Platform (2)

- Browse and advanced search by
  - Title
  - Author
  - Dept / Division
  - Degree
  - Year
- Basic View / Metadata view (in Dublin Core (DC))
- Full-text display of
  - Abstract
  - Table of contents
  - Body thesis
Development
Needs for a
New Digital
Repository (1)

- The interim solution is a quick fix only - a long-term solution must be developed to address the limitations of the platform
  - No complete standardized look and feel - frontend and backend still separated
  - Old technology still in use
    - Search and retrieval system built on the existing Tamino XML server
    - PDF files are housed at another server, full text search is not by default due to system capabilities
  - Still no cross-search across all digital collections
● The stand-alone platform displayed metadata in DC by cumbersome workflow
  ○ Each metadata load is using MARC → Excel → XML
  ■ Over 30 fields from the Innovative Millennium exported and converted in Excel manually
  ■ Data are loaded to another stand-alone platform
  ■ Time-consuming and error-prone
  ○ Incomplete capture of metadata from the original source; not as easy as a digital repository to implement standard metadata schema (e.g. METS, MODS, etc)
Development Needs for a New Digital Repository (3)

- OAI-PMH hosted at a separate platform that creates data sync issues - the data is still in Tamino, not migrated to the platform
- Every UI and metadata update need to submit a request to the technical team
- No modern features offered by a standard digital repository such as facets, RSS, social media tools
- Not flexible enough for fast-growing collections like ETD and large-scale implementation affecting the migration of other legacy digital collections
- A new team has been established since 2015 to prepare staff for the mindset change
Adoption of Islandora as the Second Generation ETD Platform

A new approach is needed to serve the ultimate aim: open access, open source, open scholarship

Tamino XML server (to be phased out)

Migrate Data to Islandora

Users
Better user experience for one single interface with frontend and backend support

Islandora
Integrated search result retrieval in one modern repository system

Library Search/Archive Space / Other internal and external Library applications
Brief Introduction of Islandora

Islandora is an open-source software framework designed to help institutions and organizations and their audiences collaboratively manage, and discover digital assets using a best-practices framework.

Islandora was originally developed by the University of Prince Edward Island's Robertson Library, but is now implemented and contributed to by an ever-growing international community.

Text source: Islandora website: http://islandora.ca/about
Image Source: Islandora Twitter https://pbs.twimg.com/media/B9_LnIUCAAEOAOV.png
Newly Launched ETD Collection on the CUHK Digital Repository in September 2016

http://repository.lib.cuhk.edu.hk/en

http://repository.lib.cuhk.edu.hk/en/collection/etd
Strength of Islandora for ETD (1)

- A truly one-stop platform with backend database and frontend application
- Powerful Chinese Japanese Korean (CJK) support on Solr, including the Simplified Traditional Chinese support
- Built-in OAI-PMH protocol that truly supports open access and discovery
- Front-end application built on Drupal allows standardization as well as flexibility in interface design; staff has gained experience with Drupal
- Drupal allows internationalization user interface that fits for university website multi-lingual requirement
- Supports search across all digital collections
Strength of Islandora for ETD (2)

- Supports a new semi-auto process to transform MODS from MARC in library catalogue on Millennium
- Linked data possibility (with the newly launched Alma)
- More efficient to allow update of records - no need to rely on technical team
  - Preparation of ETD data for ingestion took more than 2 weeks, but the migration process is so straightforward that it took less than 3 working days!
- Well-structured database allows easy management
- Standardization & Sustainability- easy to migrate to another platform in future if necessary
Strength of Islandora for ETD (3)

- Use Islandora Scholar Solution Pack for display and the Islandora Solr Facet Pages module for browsing
- Make full use of collection’s potential: auto-extraction of PDFs OCRed/embedded full text from the original PDF
- Open source instead of a proprietary system; strong user community supports quick development
- Use case should be similar for different libraries thus save time to reinvent the wheel
ETD on Islandora
- How it is better than the old platform (1)

- Browse by Title, Author, Dept / Division, Degree, Year is still kept
ETD on Islandora
- How it is better than the old platform (2)

- Default search of full text enabled - previously need to select it to avoid excessive loading on the system
- Search of and retrieval from all digital collections to yield more results
- Enhanced user experience with facet browsing that was not available before
ETD on Islandora
- How it is better than the old platform (3)
  - More flexible for restricted IP Access embargo
  - Embargo can be set for a specific title or by batches to restrict access from CUHK network only
Phase front accelerator effects in optical branching waveguides

ETD on Islandora
- How it is better than the old platform (4)
  - All the information available at the Library Catalogue record can be reused for displaying and indexing in the system
  - Much better than the previous metadata view
ETD on Islandora
- How it is better than the old platform (5)

- The Islandora compound solution pack allows grouping of multiple volumes for better navigation; the previous handling to merge the pdf was undesirable and misleading to users for their physical format
- New and efficient workflow for the backend data loading
Current Status and Achievements

- **Record Count as at July 2017**: 13,629 titles with 13,688 volumes in 2,067,478 digital images
- **The top most popular collection amongst all the 13 digital collections in the Repository**: A total of 500,945 access recorded as at July 2017 from the launch date
- **Open access and harvestable**: Harvested by many open access repositories such as NDLTD, HKIR, OpenDOAR, BASE, Core@UK, etc
Top Ten Most-Accessed ETD since Launch

As at July 2017

<table>
<thead>
<tr>
<th>Title</th>
<th>Access Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. 國民黨地方政權的軍事化與統治危機</td>
<td>23,972</td>
</tr>
<tr>
<td>2. 黃碧雲小說中的「暴力美學」研究</td>
<td>22,694</td>
</tr>
<tr>
<td>3. 香港西貢及其鄰近地區歌謡研究</td>
<td>18,586</td>
</tr>
<tr>
<td>4. 在真實與想像之間</td>
<td>13,034</td>
</tr>
<tr>
<td>5. 神的正義與神的旨意間的張力</td>
<td>9,887</td>
</tr>
<tr>
<td>6. Studies of low energy ion bombardment of cubic boron nitride (111) surfaces by reflection electron energy loss spectroscopy</td>
<td>9,785</td>
</tr>
<tr>
<td>7. 蘇童小說研究</td>
<td>7,964</td>
</tr>
<tr>
<td>8. 面對貧窮</td>
<td>7,399</td>
</tr>
<tr>
<td>9. The acquisition of metaphorical expressions, idioms, and proverbs by Chinese learners of English</td>
<td>4,906</td>
</tr>
<tr>
<td>10. 香港地水南音初探</td>
<td>3,699</td>
</tr>
</tbody>
</table>
Lessons Learnt (1)

- Timing and platform development for ETD are all affected by the historical development of a library, yet, librarians should have insight and courage to make the change.
- It is always important to add value to the content to enhance its use and reuse - ORC, full-text indexing and searching, full metadata provision are all helpful to users.
- It is no longer sustainable to develop a stand-alone platform for ETD.
- Essential to develop a platform that is standardized and meet international best practice.
Lessons Learnt (2)

- Standardization ensures ease of maintenance, migration and sustainability
- Open access, open source and open scholarship are the key of success
- Management support is essential for the continuous development of the platform for ETD and minimization of staff resistance to change
- A modernized digital repository platform offers an opportunity for the Library to reposition itself and re-think its roles in supporting e-research
Thank You

Q & A

Contact us by:

Twitter:
@cuhklibresearch
@lamlouisa
@jeffliuhk

Email:
louisalam@lib.cuhk.edu.hk
jeffliu@lib.cuhk.edu.hk