Converting Millennium ILS Bibliographic records into Dublin-Core XML format for DSpace

PNC 2009 Annual Conference and Joint Meetings
Taipei, Taiwan

Alan Ng
Hong Kong University Libraries
Introduction
HKU Libraries

• established in 1912
• the oldest academic library in HK
• main library and 6 branches
HKU Libraries

- 2.84M total physical volumes
- 49K print periodical titles
- 80K electronic periodical titles
- 1.90M e-book
HKU Libraries

• Millennium ILS from Innovative Interface Inc.

• hosting the HKALL union catalog for 8 university libraries in HK
HKU Scholars Hub

- collects intellectual output of HKU for fulltext open access
- http://hub.hku.hk/
HKU Scholars Hub

• uses DSpace (version 1.5)
• OAI-compliant
• implements DCMI
HKU Scholars Hub

• 25300+ records (as of 2009 June)
• Articles
• Conference paper
• Postgraduate thesis and others
• 1.6M download (as of 2009 June)
HKU Scholars Hub

- some records originate from the OPAC
- HKU postgraduate thesis
- Digital editions from HKU Press
- Bibliographic MARC fields are mapped to DC XML data
<table>
<thead>
<tr>
<th>MARC</th>
<th>DC Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>identifier -- other</td>
<td></td>
</tr>
<tr>
<td>008</td>
<td>language</td>
<td></td>
</tr>
<tr>
<td>020</td>
<td>identifier -- isbn</td>
<td></td>
</tr>
<tr>
<td>022</td>
<td>identifier -- issn</td>
<td></td>
</tr>
<tr>
<td>050</td>
<td>subject -- lcc</td>
<td></td>
</tr>
<tr>
<td>092</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>110</td>
<td>a</td>
<td>contributor -- author</td>
</tr>
<tr>
<td>245</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>260</td>
<td>b</td>
<td>publisher</td>
</tr>
<tr>
<td>260</td>
<td>c</td>
<td>date -- issued</td>
</tr>
<tr>
<td>300</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>490</td>
<td>a</td>
<td>relation -- ispartofseries</td>
</tr>
<tr>
<td>5XX</td>
<td>description</td>
<td></td>
</tr>
<tr>
<td>650</td>
<td>subject -- lcsh</td>
<td></td>
</tr>
<tr>
<td>710</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>856</td>
<td>u</td>
<td>identifier</td>
</tr>
</tbody>
</table>
A record in OPAC

http://library.hku.hk/record=b4200627
Same record in Hub

http://hub.hku.hk/handle/123456789/55513
Automated batch processing
Incentives

• needs to convert 100+ records at a time
• tedious, easy to make mistake manually
• time consuming
Automated approach

• efficiency
• accuracy
• eliminate duplicated effort of data entry
• easier quality control of converted data
Perl programming

- free of charge
- easy to program
- powerful in handling plain text in MARC
- runs on any computer platform
- needs a persistent URL syntax to locate a particular record on OPAC
Perl programming

- reads in a list of bibliographic record numbers
- captures the MARC records on OPAC real time one by one via HTTP
- regards the returned HTML as plain text
A dissertation submitted in partial fulfilment of the requirements for the Bachelor of Science (Speech and Hearing Sciences), The University of Hong Kong, June 30, 2007.

thesis (B.Sc)--University of Hong Kong, 2007.

Also available in print.


Includes bibliographical references (p. 27-30).

Also available in print.


A dissertation submitted in partial fulfilment of the requirements for the Bachelor of Science (Speech and Hearing Sciences), The University of Hong Kong, June 30, 2007.

Thesis (B.Sc.)--University of Hong Kong, 2007.

Cochlear implants

University of Hong Kong.

Production of stop consonants by Cantonese-speaking cochlear implant users / Yuen, Wing-Ka, Charlotte.

Click to view the E-thesis via HKU Scholars Hub

Speech and Hearing Sciences

996 y [cip] 1126 [DHKU Hub] [eth] [go] [tt] a6178

http://library.hku.hk/search~S6?/.b4200627/.b4200627/1%2C1%2C1%2CB/marc~b4200627
Perl programming

• extracts the essential MARC fields using Regular Expression
• constructs the DC fields according to the mapping table
• converts 100+ records in a couple of minutes
<dublin_core>
  <devalue element="contributor" qualifier="author">Yuen, Wing-ka, Charlotte.</devalue>
  <devalue element="contributor" qualifier="other">University of Hong Kong.</devalue>
  <devalue element="date" qualifier="issued">2007</devalue>
  <devalue element="description">
    A dissertation submitted in partial fulfilment of the requirements for the Bachelor of Science (Speech and Hearing Sciences), The University of Hong Kong, June 30, 2007.
  </devalue>
  <devalue element="description">Also available in print.</devalue>
  <devalue element="description">Thesis (B.Sc)–University of Hong Kong, 2007.</devalue>
  <devalue element="format" qualifier="extent">32 p. : ill.</devalue>
  <devalue element="format" qualifier="mimetype">application/pdf</devalue>
  <devalue element="identifier" qualifier="hkul">http://library.hku.hk/record=b4200627</devalue>
  <devalue element="identifier">http://lookup.lib.hku.hk/lookup/bib/B42006272</devalue>
  <devalue element="language">eng</devalue>
  <devalue element="publisher">University of Hong Kong (Pokfulam Road, Hong Kong)</devalue>
  <devalue element="rights">
    The author retains all proprietary rights, such as patent rights and the right to use in future works.
  </devalue>
  <devalue element="subject" qualifier="lczh">Cochlear implants</devalue>
  <devalue element="title">
    Production of stop consonants by Cantonese-speaking cochlear implant users
  </devalue>
  <devalue element="type">UG_Thesis</devalue>
</dublin_core>
Running Perl program

- runs natively on Unix, Linux and Mac OS X
- needs Perl interpreter on Windows
- download ActivePerl
Running the program on Mac OS X

```
-bash-3.2$
-bash-3.2$
-bash-3.2$
-bash-3.2$ ./Bib2DC.pl
This program will convert INNOPAC Bib record into Dublin Core format for DSpace
Usage:
   1) Bib2DC.pl
      - display this help

   2) Bib2DC.pl [bibnum list file] [bib collection type] [content type]
      e.g. bib2dc.pl sample_list.txt bsc 01

List of bib collection type:
bsc - BSc Theses
cupem - CUPEN Student Reports
ssrc - Social Sciences Research Centre
chp - Community Health Projects
hkup - Digital Editions from HKU Press
gen - general MARC record

List of bib content type:
01  - PG_Thesis
02  - PG_ThesisAbstract
03  - PG_ThesisTOC
04  - UG_Thesis
05  - Article
06  - ConferencePaper
07  - Text
08  - Postprint
09  - Abstract
10  - MovingImage
11  - Presentation
12  - Image
13  - Patent
14  - NewspaperArticle
15  - Book
16  - Journal
17  - Preprint
18  - BookChapter
19  - LearningObject
20  - StudentProject
21  - Findingaid

-bash-3.2$
-bash-3.2$
```
Demo
Recap

- uses existing MARC records for DSpace
- uses Perl program for fast batch converting
- retrieves MARC in real time via HTTP
- works with any OPAC with persistent URL
- source codes is free for sharing
Thank You !!

My contact : lbalan@hku.hk