

NewBooks

How to find out what's new in the stacks
(using metadata extracted from the Library Catalog)

by Keith Jenkins

Presented to the Metadata Working Group of the Cornell University Library, 2007-10-19

The Problem:

We wanted to see what new books were being added to the Mann Library collections.

And the catalog didn't support this.

The **Original** Strategy

- Extract new monograph records from the catalog
 - Monthly
- Create a web interface
 - Organize by LC classification
 - Display basic book information
 - Link to the full catalog record for the title

The **Current** Strategy

- Extract new monograph records from the catalog **for all CUL libraries**
 - Monthly
- Create a web interface
 - Organize by LC classification
 - Display basic book information
 - Link to the full catalog record for the title
 - **Filter by Library, Language, or LC Class**
 - **Archive previous months**

Extracting the metadata

- Export MARC records from the catalog
- Convert MARC to MARCXML
- Convert MARCXML to an efficient format for filtering and display

Exporting MARC records

- Use VgerSelect to export all new books cataloged in the last month
 - Cataloged in September, 2007
 - Monograph
 - Publication date 2004 or later
- Export bibliographic AND holdings records (since we want to display location, call number)

VgerSelect

- Initial query:
 - Date Updated \geq 2007 SEP 1
 - Record type = a (Language)
 - Bibliographic level = m (Monograph)

- ~ 25,000 records

VgerSelect

- Filtering criteria:
 - 008/date >= 2004
 - 948 1x \$a starts with "200709"
- ~6,000 records, 9MB each month

VgerSelect

```
20071002 19:52:22 Job began
20071002 20:16:09 Job ended
          00:23:47 Elapsed time for job

20071002 19:52:22 Search (SQL step) began
20071002 19:52:22 Search (SQL step) ended
          00:00:00 Elapsed time for search step

20071002 19:52:22 Examination of records began
20071002 20:16:09 Examination of records ended
          00:23:47 Elapsed time for examination
          0.0536304871 Average time (seconds) to examine each record
          18.6461107218 Records examined per second
          1,118.7666433076 Records examined per minute
```

Initial SQL statement:

```
SELECT DISTINCT T1.BIB_ID FROM CORNELLDB.BIB_MASTER T1,
CORNELLDB.BIB_HISTORY T2, CORNELLDB.BIB_TEXT T3 WHERE
(T2.ACTION_DATE >= TO_DATE('2007 09 01', 'YYYY MM DD') AND
T2.ACTION_TYPE_ID <> 1) AND T3.BEGIN_PUB_DATE >= '2004' AND
T3.BIB_FORMAT = 'am' AND T1.BIB_ID = T2.BIB_ID AND T1.BIB_ID
= T3.BIB_ID
```

Filtering criteria:

```
select/reject by tag/ind/sfd code plus text
Tag: 948
948/a like 200709
```

```
0 records to be read
26608 records actually read
20703 rejected by 'filtering' characteristics
```

```
5905 accepted for inclusion in output file
```

```
5905 MARC-format records written to output file:
```

```
C:\newbooks\marc\vgersele.CUL_200709.20071002.195222.marc.mrc
```

```
6050 subsidiary holdings records written to output file:
```

```
C:\newbooks\marc\vgersele.CUL_200709.20071002.195222.marc.mrc
```

marc2xml.php

- Convert MARC fields 001, 004, 005, 008, 245, 260, 300, 852, 880 to XML
 - 001 and 004 : IDs for linking bibl and holdings
 - 005 : timestamp to distinguish new titles and new holdings
 - 008 : language code
 - 245 : title
 - 260 : publication info
 - 300 : physical description (size, pages)
 - 852 : holdings location and call number
 - 880 : title in CJK scripts

Metadata

- We end up with the following elements:
 - BibID
 - Title
 - Description
 - Language
 - Location and Call number
- Metadata is stored as a serialized PHP array
 - compact file (9MB MARC becomes 2MB)
 - quick access

Web Interface (PHP)

- Can filter by any of the following:
 - Date (monthly data back to June 2006)
 - Location (library)
 - Language (only common ones are listed)
 - Classification (first 1-2 letters)
- Displays are bookmark-able
- RSS feeds for any filter combination

<http://supportingcast.mannlib.cornell.edu/newbooks/>

Dynamic Tables of Contents

- Uses first 1-2 letters of the call number
- Looks at number of books in each classification
 - Group by first letter of call number
 - If $n > 20$, split according to second letter
(but only if the splinter group has > 4 items)

Integration into Mann Website

- Worked with Nick Cappadona
- Added "New Books" page in CommonSpot
 - Keith supplied a <div> to insert into the page
 - Nick inserted into the page, adding styling

<http://mannlib.cornell.edu/collections/development/newbooks.cfm>

Some limitations

- Monographs only
- Monthly update schedule

Contact

Please refer all questions and suggestions to

Keith Jenkins
kgj2@cornell.edu