Migrate ALL THE THINGS:

Large Scale Library Data Manipulations with Perl

D Ruth Bavousett
Houston.PM
January 9, 2014

A Little Background

- Library automation biz is US$2bil/annual, for support alone
- All “bigger” libraries are on something already
- Data methods inconsistent between vendors
- Migrations often painful, lossy, expensive

Typical library data

- Bibliographic records (usually MARC—more on that in a bit)
- Item inventory records
- Patron biographic/demographic data
- Patron borrowing records
- Fines/Fees
- Requests/Holds
- Acquisitions records
- Serial publication information

MARC?

- "MAchine-Readable Cataloging"
- Developed by Library of Congress in early 1970s for magtape transfer of bibliographic descriptive data
- Maximum record size 99,999 bytes. (but note that some Unicode chars take two bytes, and limit includes "directory" information in header)
- Format has evolved over time—"tags" and "subfields"

A MARC Record

```
000  01355cam a2200277 a 450
001  1513231
005  20040316104619.0
008  890309t18761875ctua 000 1 eng
035  __ |9 (DLC) 89120664
050  00 |a PS1306 |b .A1 1876b
082  00 |a 813/.4 |2 20
100  1_ |a Twain, Mark, |d 1835-1910.
240  10 |a Adventures of Tom Sawyer. |k Selections
245  14 |a The adventures of Tom Sawyer / |c by Mark Twain.
260  __ |a Hartford, Conn. : |b American Pub. Co., |c 1876, c1875.
300  __ |a [102] p. : |b ill. ; |c 23 cm.
500  __ |a LC copy bound in blue cloth in a four-fold folder |5 DLC
500  __ |a Source: Gift of Leonard Kebler, Jan. 19, 1948. |5 DLC
650  _0 |a Sawyer, Tom (Fictitious character) |x Fiction.
651  _0 |a Mississippi River Valley |x Fiction.
650  _0 |a Runaway children |v Fiction.
650  _0 |a Male friendship |v Fiction.
651  _0 |a Missouri |x Fiction.
650  _0 |a Boys |x Fiction.
655  _7 |a Adventure stories.
655  _7 |a Humorous stories.
```

Whither MARC?

- MARC is useful for lots of “things”, but intangibles like Internet resources—not so much
- Hard limit to record size
- Complicated, arcane
- MARC-Must-Die movement

Typical migration work

- Small-town public library
  - ~100,000 MARCs, 120K items
  - ~10K patrons
  - ~4K current and overdue items out
  - ~6K-10K fine/fee records
  - ~200 holds

Project process

- Kickoff meeting
- Library extracts data, puts in Dropbox
- Examine data, do test migration, capturing chain of command-line scripts
    - Possible punch-list of errors, fine-tuning
- Library examines test data, configures system, trains
- Go-live—re-extract data, migrate, install
    - Possible punch-list of errors

Data Manipulation Process

- Ensure that items and patrons have unique ID (usually a barcode!)
- Splice MARC with embedded item records in Koha format
- Ensure that other data uses unique IDs as match point.
- Tidy up data—misspellings in item type codes, borrower city/state, etc. (use a map file, for repeatability)

Wanna see some code?

Q&A

Thanks for being here!