Fearless Transformation: Applying OpenRefine to Digital Collections

Kara Long
Catalog and Metadata Librarian
Baylor University Libraries
@thekaralong
Agenda:

- The Spencer project
- Re-examining the workflow
- Implementing OpenRefine
- Other uses for OpenRefine in digital collections
the Beginning

Project started in 1999

TexTreasures Grant to digitize 1,000 pieces

Descriptive metadata loaded into ILS

Static HTML was programmatically generated and placed on server
Adding Partners

CONTENTdm®

flourish:
Music-Contract-Cataloging
the Workflow

cataloging packet → Webshare → Vendor: Flourish → MARC DB → MARC Edit

XSLT → XML → PERL → TSV → TIFF → Ingest via CDM Project Client
the Workflow Revised

cataloging packet -> Webshare -> Vendor: Flourish -> MARC DB -> OpenRefine

TSV, TIFF -> Ingest via CDM Project Client
Lyrics with piano
Arr. orange, hearts, photo Sophie Tucker.
Set: excerpts 3 songs, titles, covers.
MARC

LEADER 00000ncm 22000001a 4500
001 426135815
003 OCoLC
008 20100607104730.0
008 009717s1910 n zxx d
035 (OCoLC)426135815
040 SST|cSST|dYU
048 vn01|aka01
049 IYUU
099 Spencer D1732 .15
099 Spencer D1732 .15-2
100 1 Botsford, George,|d1874-1949.
245 14 The dance of the grizzly bear /|cwords by Irving Berlin ;
music by George Botsford.
246 3 Grizzly bear
246 1 |ifirst line of text:|aOut in San Francisco where the
weather's fair
246 1 |ifirst line of chorus:|aHug up close to your baby
260 New York :|bHug up close to your baby
300 1 score (5, [1 p.]) ;|c35 cm.
500 For voice and piano.
562 |cD1732 .15: illustrated t. p. in orange, black and white
with a drawing of 2 bears / Frew; photo. of Sophie Tucker.
562 |cD1732 .15-2: illustrated t. p. in orange, black and white
with a drawing of 2 bears / Frew; photo. of Tim
McMahon.
562 |cD1732.15: "As sung by Sophie Tucker."
562 |cD1732.15-2: "As sung by Tim McMahon."
562 |cD1732.15: Advertisement for Alexander's ragtime band
March and two-step / Berlin on p. [6].
915 Flourish
The dance of the grizzly bear

Out in San Francisco where the weather's fair.

Hug up close to your baby

words by Irving Berlin; music by George Botsford.

Botsford, George, 1874-1949

Berlin, Irving, 1888-1989

Tucker, Sophie, 1884-1966

New York: Ted Snyder Co.

1910

1 score (5, [1] p.); 35 cm.

For voice and piano.

D1732.15: "As sung by Sophie Tucker."; D1732.15: Advertisement for Alexander's ragtime band march and twostep / Berlin on p. [6].

D1732.15: Illustrated e. p. in orange, black and white with a drawing of 2 bears / Frew; photo. of Sophie Tucker.

Tucker, Sophie, 1884-1966 -- Portraits; McMahon, Tim -- Portraits; Songs with piano; Popular music -- United States -- 1901-1910; Grizzly bear -- Songs and music; Ragtime music.

Dance; Famous people - Sophie Tucker.


426135615

Spencer D1732.15

http://www.baylor.edu/lib/digitization/digitalextc

2009_11.bib.xml Wed Jul 7 11:54:26 2010

Musical Score

Image

No Linguistic Content

d1732.15

Baylor University, Crouch Fine Arts Library
OpenRefine

- Interactive Data Transformation tool (IDT)
- Open source
- Runs locally
- Interactive like a spreadsheet
  - but more powerful
- Programmable like a database
  - but more exploratory
Creating a new project

- Rename and reorder MARC fields
- Join values
- Split values
- Re-format dates
- Remove unnecessary punctuation, delimiters, etc.
- Add new fields for the digital collection
Renaming columns

Columns are the primary units of interaction.

Column names must exactly match our CDM field names in order for upload the metadata.

MARC 100 → Composer
OpenRefine

Re-ordering columns

Once all the fields have been re-named, they can be re-ordered under the All columns menu.
Joining the 245$a and 245$b to create the Title field.
Splitting values

The 246 must be split into two or three fields:

- Alternative Title
- First line of verse
- First line of chorus
Add column based on column 246.0

New column name: Alternative Title

On error: set to blank

Expression: filter(value.split(";"),v,not(or(v.startsWith("First line of text"),(v.startsWith("First line of chorus")))),).join("","")

Preview:

<table>
<thead>
<tr>
<th>row</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>If you can't make love at home; First line of text: Beatrice Fairfax gives advice, to any one in love; First line of chorus: Take your girlie to the movies; if you can't make love at home</td>
</tr>
<tr>
<td>2.</td>
<td>What will we do on a Saturday night, When the town goes dry</td>
</tr>
</tbody>
</table>
OpenRefine

Extract and save operation history
Identifying clean up in existing CONTENTdm collections

- Text faceting
- Custom text facets
- Identifying duplicates
Invaluable Resources

http://openrefine.org/

http://freeyourmetadata.org/

https://github.com/OpenRefine/OpenRefine/wiki/GREL-Functions
