


# Fearless

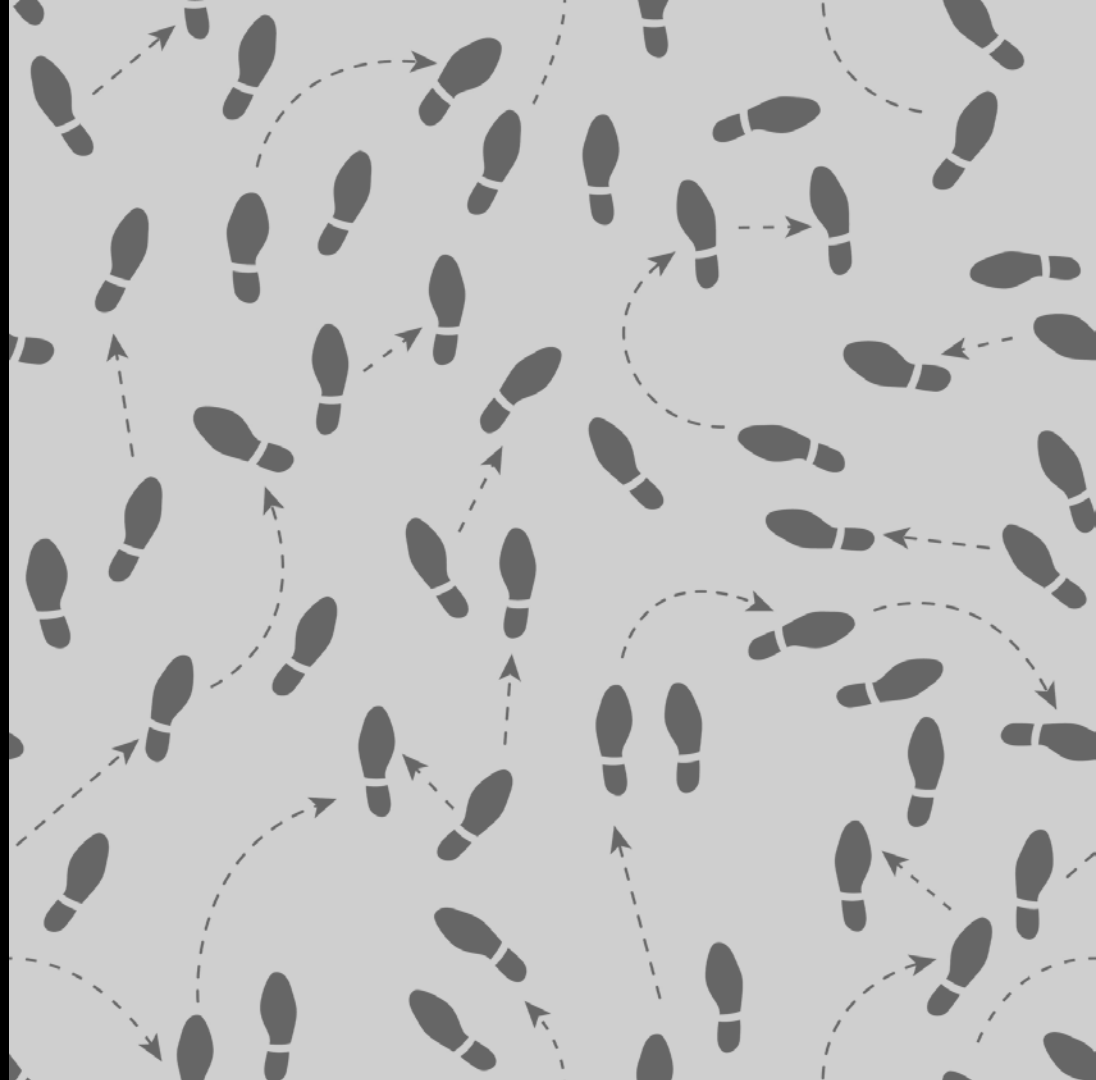
## Transformation: Applying OpenRefine to Digital Collections

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 OLD OAKEN BUCKET.



*Written by*  
**JARVIS WOODWORTH, ESQ.**  
 Adapted to a favorite  
**SCOTCH AIR.**

Published by **L. BRADLEE**, 704 Washington St.

LINCOLN & MITCHELL'S  
 HIS EXTRAVAGANZA AS PRODUCED  
 AT THE GRAND OPERA HOUSE, CHICAGO.  
**ABES**  
**TOY**  
**LAND**  
 AND LYRICS BY  
 GLEN MACDONOUGH  
 BY  
 TOR HERBERT

[illegible]

**LITTLE IRISH ROSE**



# Franklin D. Roosevelt

March  
(see the January issue)

DEDICATED TO THE  
32<sup>nd</sup> PRESIDENT OF  
THE UNITED STATES  
OF AMERICA, . . .

WILLIAM H. WOODIN

DYING CAESAR



Harry A. Smith  
PLATE ---- BY  
Hugo Riesenfeld



## A painting of a woman in 18th-century dress standing on a wooden bridge, holding a flag, with a mountain landscape in the background. The woman is wearing a light blue long-sleeved dress with a full skirt and a red shawl draped over her shoulders. She is holding a flag with a yellow field and a red cross. The background shows a mountain landscape with a small house and a river.





# 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



TEXAS STATE LIBRARY  
AND  
ARCHIVES COMMISSION



CONTENTdm®

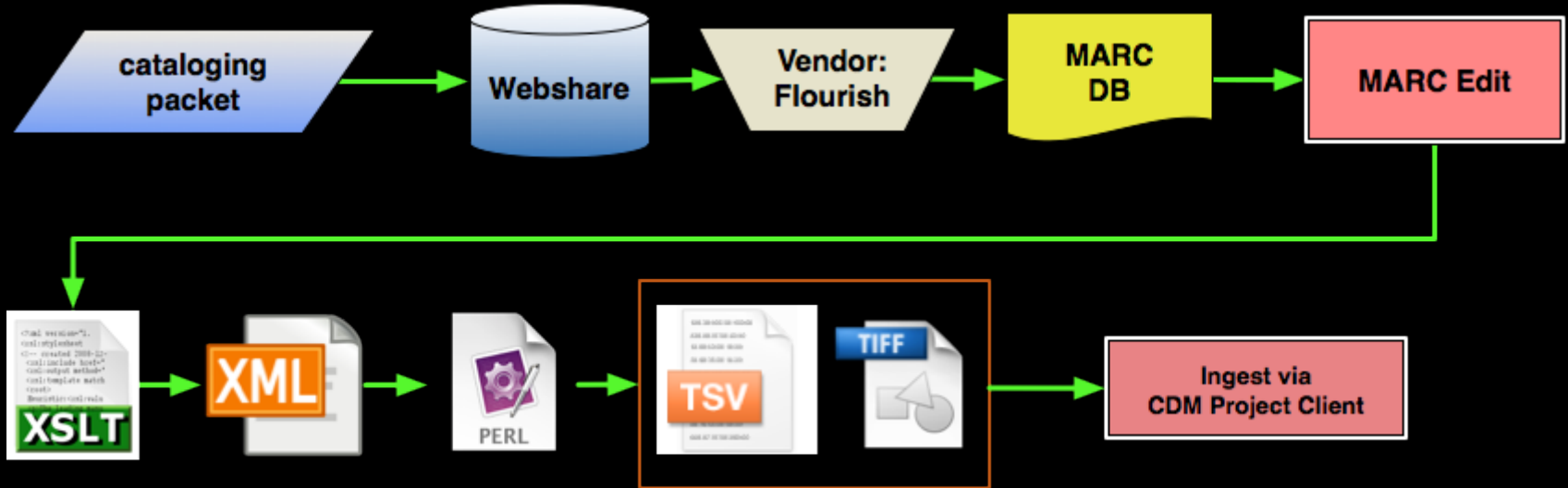
Adding Partners



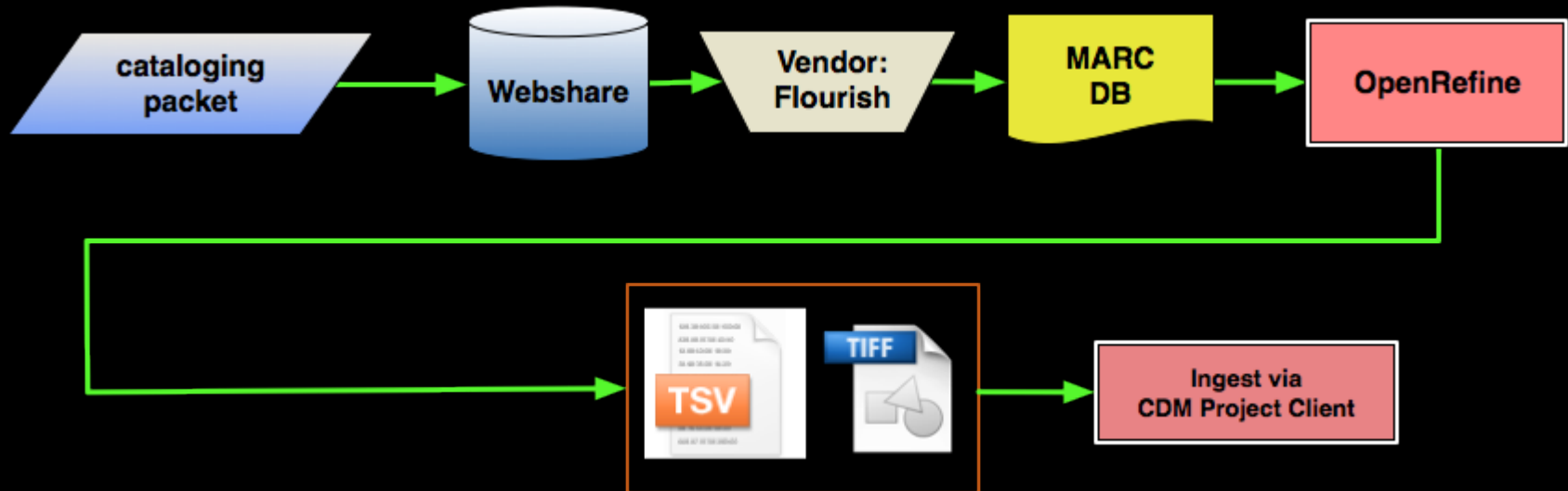
**flourish:**

Music-Contract-Cataloging

# the Workflow



# the Workflow Revised



#18853426

D1732 Botsford, George, 1874 - 1949.  
15-2 Grizzly bear. Words by Irving Berlin.  
As sung by Tim McMahon. N. Y., Ted Snyder  
Co. (Inc.), 1910.  
SPENCER no pl. no.

Lyrics with piano  
Pc: orange, bears, photo Tim McMahon.  
Bc: excerpts 2 songs, illus. covers.

Author	Botsford, George, 1874-1949.
Title	The dance of the grizzly bear / words by Irving Berlin ; music by George Botsford.
Publication Info	New York : Ted Snyder Co., c1910.

Connect to:

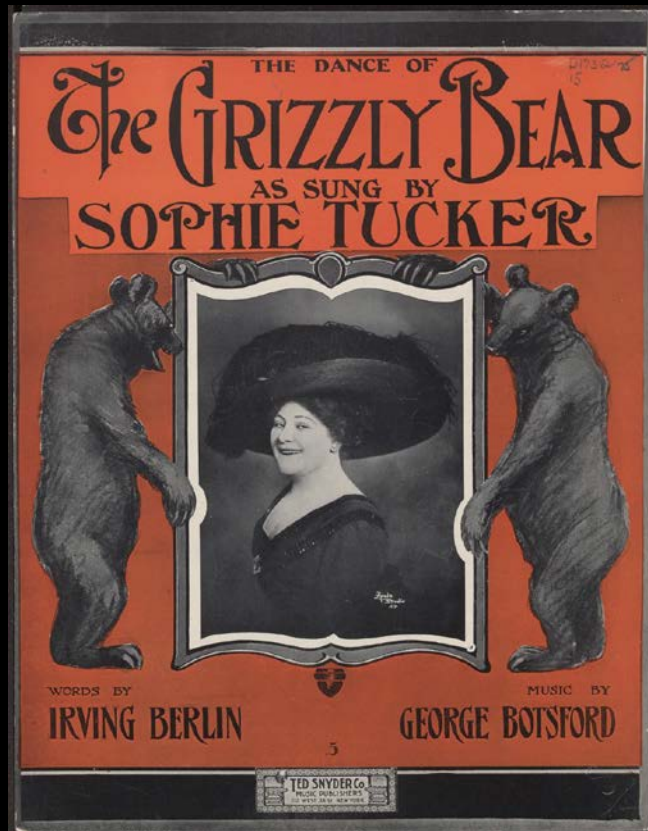
[Connect to this title in the Baylor University Libraries Digital Collections](#)

[Connect to this title in the Baylor University Libraries Digital Collections](#)

LOCATION	CALL #	STATUS
Crouch SCR Spencer	Spencer D1732 .15	BY APPT ONLY
Crouch SCR Spencer	Spencer D1732 .15-2	BY APPT ONLY

[Text the call number](#)

Description 1 score (5, [1] p.) : 35 cm.  
Note For voice and piano.  
D1732 .15: illustrated t. p. in orange, black and white with a drawing of 2 bears / Frew; photo. of Sophie Tucker.  
D1732 .15-2: illustrated t. p. in orange, black and white with a drawing of 2 bears / Frew; photo. of Tim McMahon.  
D1732 .15: "As sung by Sophie Tucker."  
D1732 .15-2: "As sung by Tim McMahon."  
D1732 .15: Advertisement for Alexander's ragtime band march and twostep / Berlin on p. [6].  
D1732 .15-2: Advertisement for Craggy rag and Call me up some rainy afternoon / Berlin on p. [6].  
Local Note Frances G. Spencer Collection of American Sheet Music.  
Spencer subject: **Dance** - other.  
Spencer subject: Famous people - Sophie Tucker.  
Subject Tucker, Sophie, 1884-1966 -- Portraits.  
McMahon, Tim -- Portraits.  
Songs with piano.  
Popular music -- United States -- 1901-1910.  
Grizzly bear -- Songs and music.  
Ragtime music.  
Local Subject **Dance**.  
Famous people - Sophie Tucker.  
Alt Author Berlin, Irving, 1888-1989. Lyricist.  
Frew, 1876-1966. Illustrator.  
Tucker, Sophie, 1884-1966. Performer.  
McMahon, Tim. Performer.  
Alt Title **Grizzly bear**  
Note First line of text: Out in San Francisco where the weather's fair  
First line of chorus: Hug up close to your baby  
Alt Title Frances G. Spencer Collection of American Sheet Music.





# MARC

```


LEADER 00000ncm 2200000Ia 4500
001 426135815
003 OCoLC
005 20100607104730.0
008 090717s1910 nyurga n zxx d
035 (OCoLC)426135815
040 SST|cSST|dIYU
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 :|bTed Snyder Co.,|cc1910.
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 twostep / Berlin on p. [6].

```

```

562 |cD1732.15-2: Advertisement for Draggy rag and Call me up
some rainy afternoon / Berlin on p. [6].
590 Frances G. Spencer Collection of American Sheet Music.
590 Spencer subject: Dance - other.
590 Spencer subject: Famous people - Sophie Tucker.
600 10 Tucker, Sophie,|d1884-1966|vPortraits.
600 10 McMahon, Tim|vPortraits.
650 0 Songs with piano.
650 0 Popular music|zUnited States|y1901-1910.
650 0 Grizzly bear|vSongs and music.
650 0 Ragtime music.
690 Dance.
690 Famous people - Sophie Tucker.
700 1 Berlin, Irving,|d1888-1989.|4lyr
700 1 Frew, John.|4ill
700 1 Tucker, Sophie,|d1884-1966.|4prf
700 1 McMahon, Tim.|4prf
793 0 Frances G. Spencer Collection of American Sheet Music.
856 41 |uhttp://digitalcollections.baylor.edu/u/?fa-spnc,32641
|zConnect to this title in the Baylor University Libraries
Digital Collections
856 41 |uhttp://digitalcollections.baylor.edu/u/?fa-spnc,30638
|zConnect to this title in the Baylor University Libraries
Digital Collections
915 Flourish

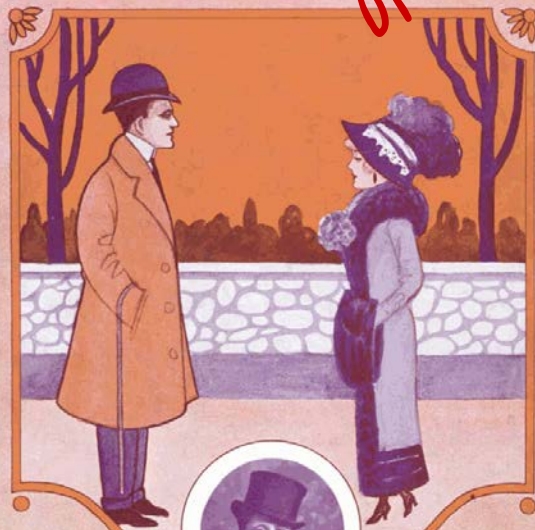
```

<b>Title</b>	The dance of the grizzly bear
<b>Alternative Title</b>	Grizzly bear
<b>First line of verse</b>	Out in San Francisco where the weather's fair
<b>First line of chorus</b>	Hug up close to your baby
<b>Statement of Responsibility</b>	words by Irving Berlin ; music by George Botsford.
<b>Composer</b>	Botsford, George, 1874-1949
<b>Lyricist</b>	Berlin, Irving, 1888-1989
<b>Performers</b>	Tucker, Sophie, 1884-1966
<b>Publisher</b>	New York : Ted Snyder Co.
<b>Date</b>	1910
<b>Physical Description</b>	1 score (5, [1] p.) ; 35 cm.
<b>Instrumentation</b>	For voice and piano.
<b>Note</b>	D1732 .15: "As sung by Sophie Tucker." ; D1732 .15: Advertisement for Alexander's ragtime band march and twostep / Berlin on p. [6].
<b>Cover Art Description</b>	D1732 .15: Illustrated t. p. in orange, black and white with a drawing of 2 bears / Frew; photo. of Sophie Tucker.
<b>Subject - Library of Congress</b>	Tucker, Sophie, 1884-1966 -- Portraits. ; McMahon, Tim -- Portraits. ; Songs with piano. ; Popular music -- United States -- 1901-1910. ; Grizzly bear -- Songs and music. ; Ragtime music.
<b>Spencer Subject</b>	Dance. Famous people - Sophie Tucker.
<b>Collection Title</b>	Frances G. Spencer Collection of American Sheet Music. American Melting Pot Collection.
<b>OCLC</b>	426135815
<b>Call No.</b>	Spencer D1732 .15
<b>Rights</b>	<a href="http://www.baylor.edu/lib/digitization/digitalrights">http://www.baylor.edu/lib/digitization/digitalrights</a> 
<b>Metadata set</b>	2009_11.bib.xml Wed Jul 7 11:54:26 2010
<b>Resource Type</b>	Musical Score
<b>Format</b>	Image
<b>Language</b>	No Linguistic Content
<b>ID</b>	d1732_15
<b>Custodian</b>	Baylor University, Crouch Fine Arts Library

# I'M AFRAID PRETTY MAID

## I'M AFRAID

BY  
IRVING BERLIN



TED SNYDER &  
WATKINSON BERLIN & SNYDER ©  
PUBLISHED BY THE  
MUSIC COMPANY OF AMERICA

GENE  
LOAN

of Data Loss!

# OpenRefine

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



# OpenRefine

## Creating a new project

471 rows

Show as: **rows** records      Show: 5 10 25 50 rows

▼ All	▼ 245\$a	▼ 245\$b	▼ 246.0	▼ 99.0	▼ 245\$c	▼ 100.0	▼ 110.0	▼ 700.0	▼ 260\$a
☆	1.	Take your girlie to the movies :	(if you can't make love at home) /	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	Spencer M935 .4	words by Edgar Leslie & Bert Kalmar ; music by Pete Wendling ; [arr. by Fred E. Ahlert].	Wendling, Pete,1888-1974.	Leslie, Edgar.lyr;Kalmar, Bert,1884-1947.lyr;Ahlert, Fred E.,1892-1953.arr;Barbelle,1888-1957.ill	New York :
☆	2.	What'll we do on a Saturday night :	(when the town goes dry) /	What will we do on a Saturday night;When the town goes dry;First line of text:We all save our pennies up for Saturday night;First line of chorus:What'll we do on a Saturday night, when the town goes dry	Spencer M935 .5	by Harry Ruby.	Ruby, Harry. emplyr	Barbelle,1888-1957.ill	New York :
☆	3.	Musical rag-time Sal /		Musical ragtime Sal;You musical	Spencer M935 .33	words by Martin Swanger ; music by	Powell, W. C.,1876-1939.	Swanger, Martin.lyr;Pfeiffer, E. H.ill	New York :

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



Permalink

**471 rows**

Show as: **rows** records Show: 5 10 25 50

	All	245\$a	245\$b	246.0	99.0	245\$c	100.0	110.0	700.0	260\$a	260\$b	260\$c
1.	Take your girlie to the movies :	(if you can't make love at home) /	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	Spencer M935 .4	words by Edgar Leslie & Bert Kalmar ; music by Pete Wendling ; [arr. by Fred E. Ahlert].	Facet	Leslie, Edgar;Kalmar, Bert,1884-1947;Ahlert, Fred E.,1892-1963;arr.Barbelle,1888	New York :	Waterson, Berlin & Snyder,	c1919.		
2.	What'll we do on a Saturday night :	(when the town goes dry) /	What will we do on a Saturday night;When the town goes dry;First line of text:We all save our pennies up for Saturday night;First line of chorus:What'll we do on a Saturday night, when the town goes dry	Spencer M935 .5	by Harry Ruby.	Reconcile				c1919.		
3.	Musical rag-time Sal /		Musical ragtime Sal;You musical	Spencer M935 .33	words by Martin Swanger ; music by	Extract named entities...	Powell, W. C.,1876-1939.	Swanger, Martin;Pfeiffer, E. H.III	New York :	Church, Paxson & Co.,	c1911.	

Enter new column name

Composer

OK Cancel

# OpenRefine

## Renaming columns

Columns are the primary units of interaction.

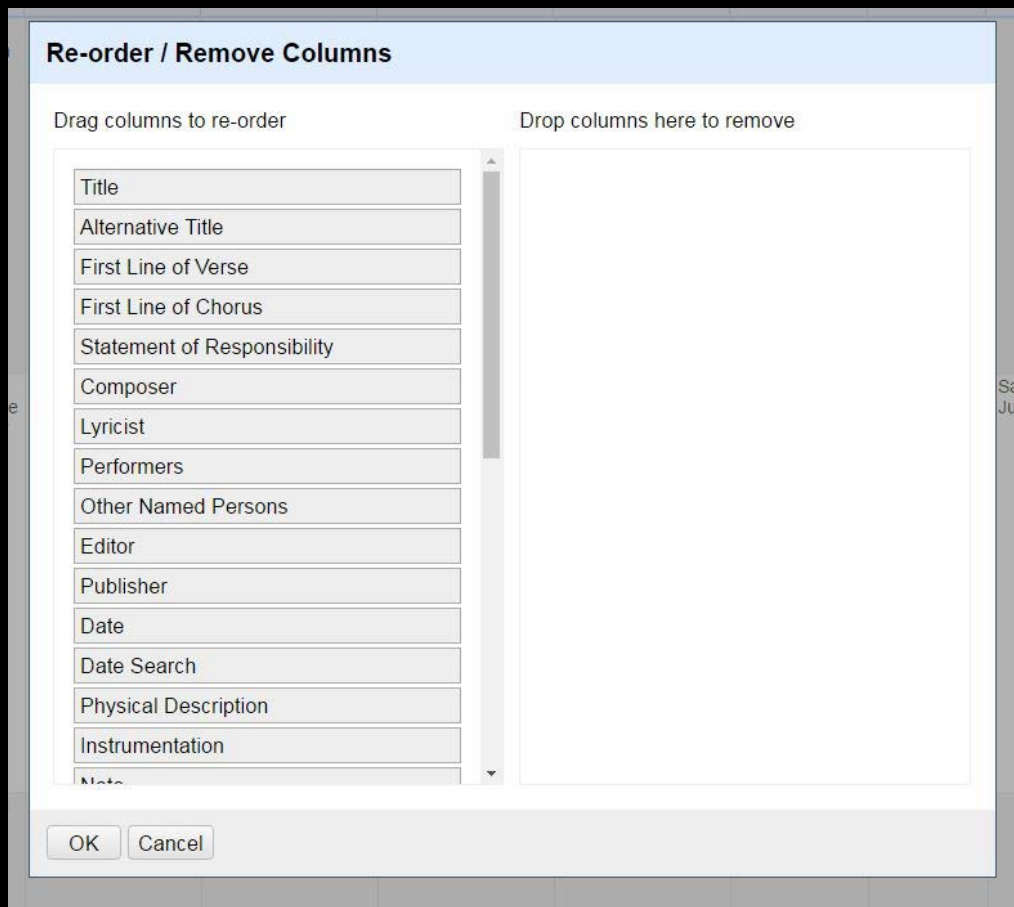
Column names must exactly match our CDM field names in order for upload the metadata.

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.



# OpenRefine

## Joining Values

Transform data  
with Google Refine  
Expression  
Language (GREL)

Joining the 245\$a  
and 245\$b to  
create the Title  
field

Custom text transform on column 245\$a

Expression: `cells["245$a"].value+" "+cells["245$b"].value` Language: Google Refine Expression Language (GREL) No syntax error.

Preview History Starred Help

row	value	cells["245\$a"].value+" "+cells["245\$b"].value
1.	Take your girlie to the movies :	Take your girlie to the movies : (if you can't make love at home) /
2.	What'll we do on a Saturday night :	What'll we do on a Saturday night : (when the town goes dry) /
3.	Musical rag-time Sal /	Error: Cannot retrieve field from null
4.	Those Keystone comedy cops /	Error: Cannot retrieve field from null
5.	The Vitagraph girl :	The Vitagraph girl : waltz song & chorus /

On error: ☒ keep original ☐ set to blank ☐ store error ☐ Re-transform up to 10 times until no change

OK Cancel

245\$a 245\$b 246.0 99.0 245\$c Composer 110.0 700.0 260\$a 260\$b

### Add column based on column 246.0

New column name

On error ☒ set to blank ☐ store error ☐ copy value from original column

Expression  Language

No syntax error.

**Preview** History Starred Help

row	value	filter(value.split(";"),v,(v.startsWith("First line of text"))).join(",")
1.	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	First line of text:Beatrice Fairfax gives advice, to any one in love
2.	What will we do on a Saturday night;When the town goes dry;First line of text:We all save our pennies up for Saturday night;First line of	First line of text:We all save our pennies up for Saturday night

OK Cancel

# OpenRefine

## Splitting values

The 246 must be split into two or three fields:

- Alternative Title
- First line of verse
- First line of chorus

45Sa 245B 246.0 99.0 245C Composer 110.0 700.0 260Sa 260B

### Add column based on column 246.0

New column name

On error ☒ set to blank ☐ store error ☐ copy value from original column

Expression  Language  No syntax error.

**Preview** History Starred Help

row	value	filter(value.split(";"),v,not(or(v.startsWith("First line of text"),(v.startsWith("First line of chorus"))))).join(",")
1.	If you can't make love at home;First line of text:Beatrice Fairfax gives advice, to any in love;First line of chorus:Take your girlie to the movies, if you can't make love at home	If you can't make love at home
2.	What will we do on a	What will we do on a Saturday night,When the town goes dry

OK Cancel

piano in a

OpenRefine

Splitting values

Know your data!



Facet / Filter. Undo / Redo 1285 309 rows

Extract... Apply... Show as: rows records Show: 5 10 25 50 rows

Filter:

275. Reorder columns

276. Create new column Note based on column 500.0 by filling 309 rows with grel.filter(value.split(","),v.not(or(v.startsWith("Arr"),(v.startsWith("Acc")))))join(",")

277. Create new column Instrumentation based on column 500.0 by filling 309 rows with grel.filter(value.split(","),v,(or(v.startsWith("For"),(v.startsWith("Arr"),(v.startsWith("Acc")))))join(",")

278. Remove column 500.0

279. Remove column 562.0

280. Text transform on 309 cells in column Subject - Library of Congress: grel.value.replace("-", " ").replace(","," ")

281. Text transform on 239 cells in column Spencer Subject: grel.value.replace(","," ")

282. Text transform on 308 cells in column Note: grel.value.replace("-", " ").replace(","," ")

283. Reorder columns

284. Create new column Collection Title based on column Spencer Subject by filling 309 rows with grel:"Frances G. Spencer Collection of American Sheet Music."

### Extract Operation History

Extract and save parts of your operation history as JSON that you can apply to this or other projects in the future.

- ☒ Create column Metadata set at index 25 based on column 1.0 using expression null
- ☒ Create column Resource Type at index 26 based on column Metadata set using expression grel:"Musical Score"
- ☒ Create column Format at index 27 based on column Resource Type using expression grel:"Image"
- ☒ Create column Language at index 28 based on column Format using expression null
- ☒ Create column ID at index 4 based on column 99.0 using expression grel.toLowerCase(substring(value,8).replace(".", "\_"))
- ☒ Move column ID to position 29
- ☒ Create column Rights at index 4 based on column 99.0 using expression grel:"http://www.baylor.edu/lib/digitization/digitalrights"
- ☒ Create column Editor at index 3 based on column 246.0 using expression null
- ☒ Create column Series Title at index 25 based on column 690.0 using expression null
- ☒ Create column Lyrics at index 26 based on column

Select All Unselect All

Close

```
[{"op": "core/column-addition", "description": "Create column Metadata set at", "engineConfig": {"mode": "row-based", "facets": []}, "newColumnName": "Metadata set", "columnInsertIndex": 25, "baseColumnName": "1.0", "expression": "null", "onError": "set-to-blank"}, {"op": "core/column-addition", "description": "Create column Resource Type a", "engineConfig": {"mode": "row-based", "facets": []}, "newColumnName": "Resource Type", "columnInsertIndex": 26, "baseColumnName": "Metadata set", "expression": "grel:\\\"Musical Score\\\"", "onError": "set-to-blank"}, {"op": "core/column-addition"}
```

# OpenRefine

## Extract and save operation history

Facet / Filter Undo / Redo 0 **5113 rows**

Refresh Reset All Remove All Show as: rows records Sh

**GeoLoc** change

58 choices Sort by: name count Cluster

TX 1  
 Memphis, TN 2  
 Mexia, TX 2  
 Mexico 1 edit include  
 Mineral Wells, TX 1  
 New Haven, CT 1  
 New York City, NY 2  
 NY 2  
 Oklahoma City, OK 1  
 Palacios, TX 1  
 Panther Park at Fort Worth, TX 1  
 Paris, TX 6  
 Poughkeepsie, NY 1  
 Ridgecrest, NC 4  
 San Antonio, TX 12  
 Seventh & James, Waco, TX 3  
 Seventh Street, Waco, TX 1  
 Sixth Street & Bagby Avenue, Waco, TX 2  
 Sixth Stret, Waco, TX 1  
 Speight Avenue, Waco, TX 2  
 Texas Christian University, Fort Worth, TX 1  
 Texas Christian University, Forth Worth, TX 1  
 Texas Memorial Stadium, Austin, TX 1  
 Waco, TX 2427  
 Waco, TX, 140  
 Washington D.C. 2  
 (blank) 1055

Facet by choice counts

GeoLoc	SubjKeywords
	Baylor University, football, ball carrier ; Baylor vs. A&M, football.
Waco, TX	Umpire ; Player ; Base ; SWC Tournament.
Waco, TX	Game ; Players.
	Baylor University, football, offense.
Waco, TX	Players ; Team.
Waco, TX	Baylor University ; The SUB ; Bird's eye view.
	Baylor University, football, ball carrier ; Baylor University, football, play.

# OpenRefine

## Identifying clean up in existing CONTENTdm collections

- Text faceting
- Custom text facets
- Identifying duplicates

# Invaluable Resources

[http:// openrefine.org/](http://openrefine.org/)

[http:// freeyourmetadata.org/](http://freeyourmetadata.org/)

[https:// github.com/ OpenRefine/ OpenRefine/ wiki/ GREL-Functions](https://github.com/OpenRefine/OpenRefine/wiki/GREL-Functions)

Verborgh, Ruben, and Max De Wilde. *Using OpenRefine*. Birmingham: PACKT Publishing, 2013.

Van Hooland, Seth, and Ruben Verborgh. *Linked Data for Libraries, Archives, and Museums: How to Clean, Link, and Publish Your Metadata*. Chicago: Neal-Schuman, 2014.