OpenRefine: or How I Learned to Stop Worrying and Love Data Transformation

Kara Long
Metadata and Catalog Librarian
Baylor University

Darryl Stuhr
Associate Director, Digital Projects
Baylor University

Agenda:

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

Spencer American Popular Sheet Music Project

SOUVENIR EDITION
Little Annie Rooney
The Only Authorized Edition
This Photo of Mary Pickford As She Appears
In The Photo Play "LITTLE ANNIE ROONEY"
Is Used With Her Permission.
WITH UKULELE ARRANGEMENT
M. WITMARK & SONS New York

Spencer American Popular Sheet Music Project

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

Spencer American Popular Sheet Music Project

CONTENTdm®

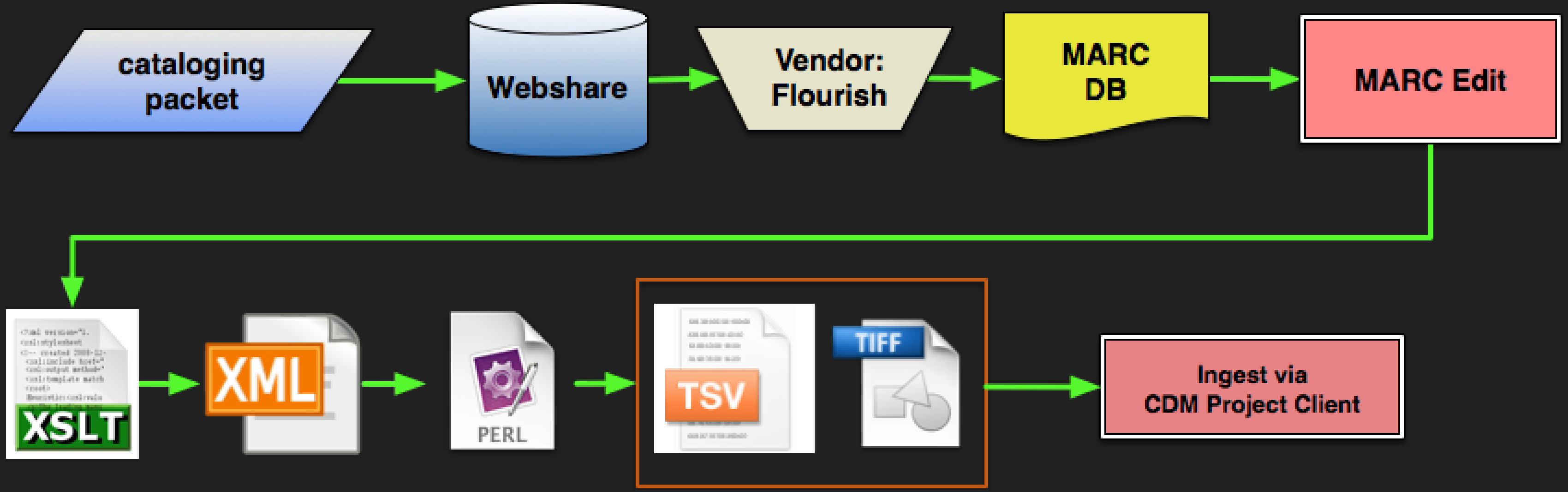
□ IN-HOUSE

☑ OUTSOURCE

Flourish:

Music-Contract-Cataloging-

Spencer American Popular Sheet Music Project



cataloging packet

Webshare

Vendor: Flourish

MARC DB

MARC Edit

XSLT

XML

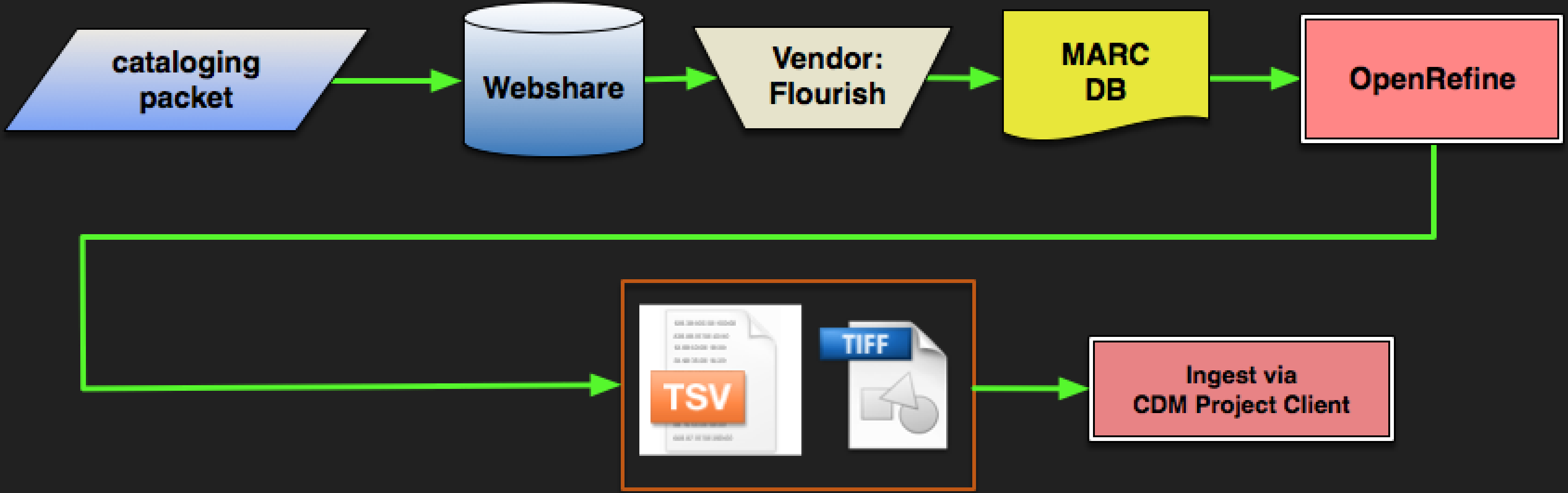
PERL

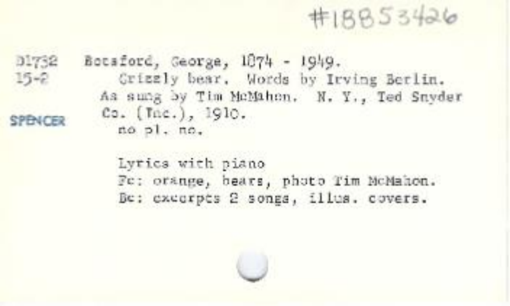
TSV

TIFF

Ingest via CDM Project Client

Spencer American Popular Sheet Music Project



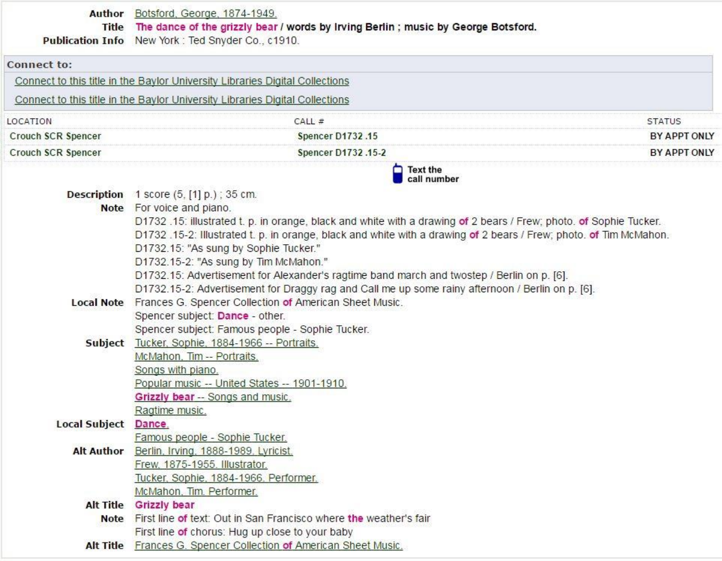


#18853426

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

SPENCER

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



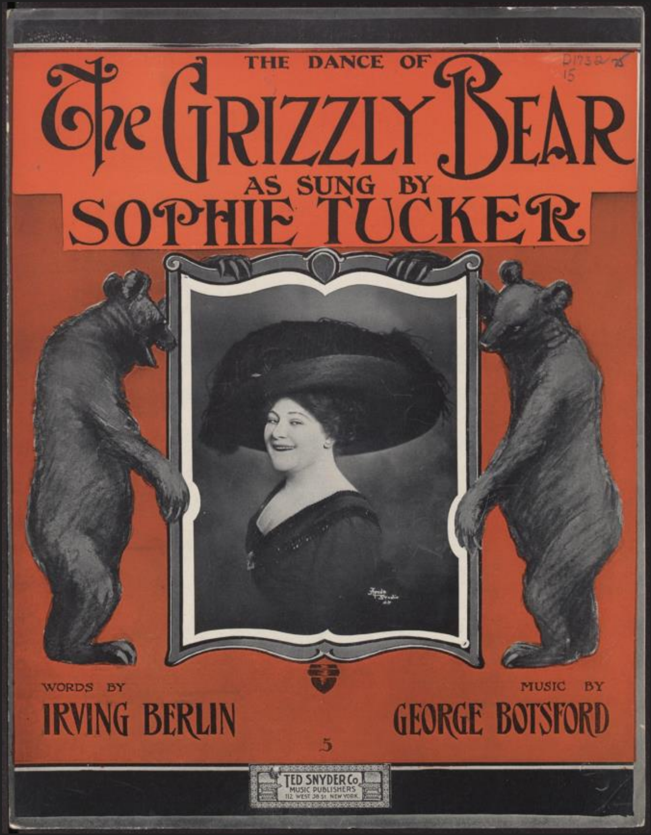
Author Botsford, George, 1874-1949.
Title The dance of the grizzly bear / words by Irving Berlin ; music by George Botsford.
Publication Info New York : Ted Snyder Co., c1910.

Connect to:
Connect to this title in the Baylor University Libraries Digital Collections
Connect to this title in the Baylor University Libraries Digital Collections

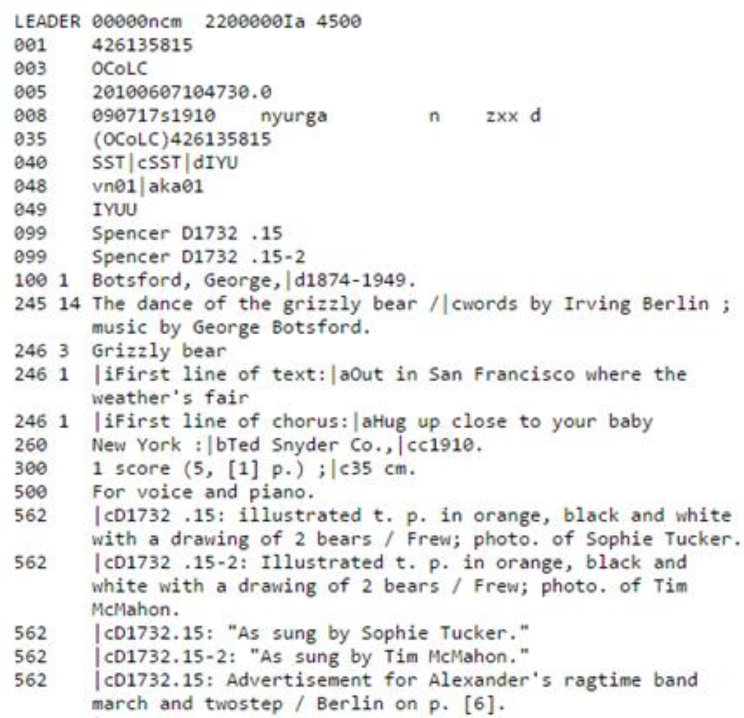
LOCATION	CALL #	STATUS
Crouch SCR Spencer	Spencer D1732 .15	BY APPT ONLY
Crouch SCR Spencer	Spencer D1732 .15-2	BY APPT ONLY

Text the call number

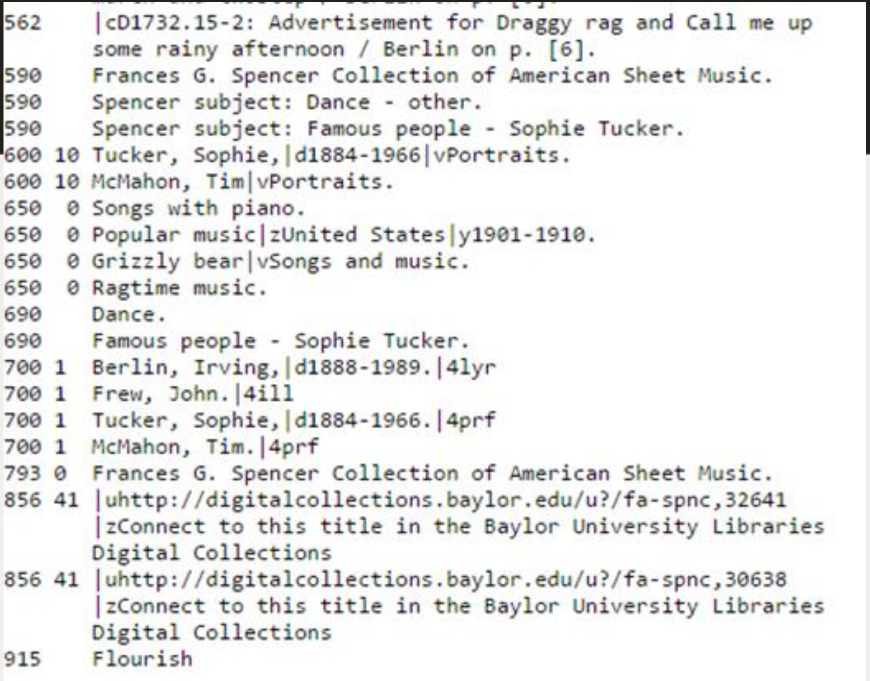
Description 1 score (5, [1] p.) ; 35 cm.
Note For voice and piano.
D1732 .15: illustrated t. p. in orange, black and white with a drawing of 2 bears / Frew; photo. of Sophie Tucker.
D1732 .15-2: Illustrated t. p. in orange, black and white with a drawing of 2 bears / Frew; photo. of Tim McMahon.
D1732.15: "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 Draggy 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, 1875-1955. 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

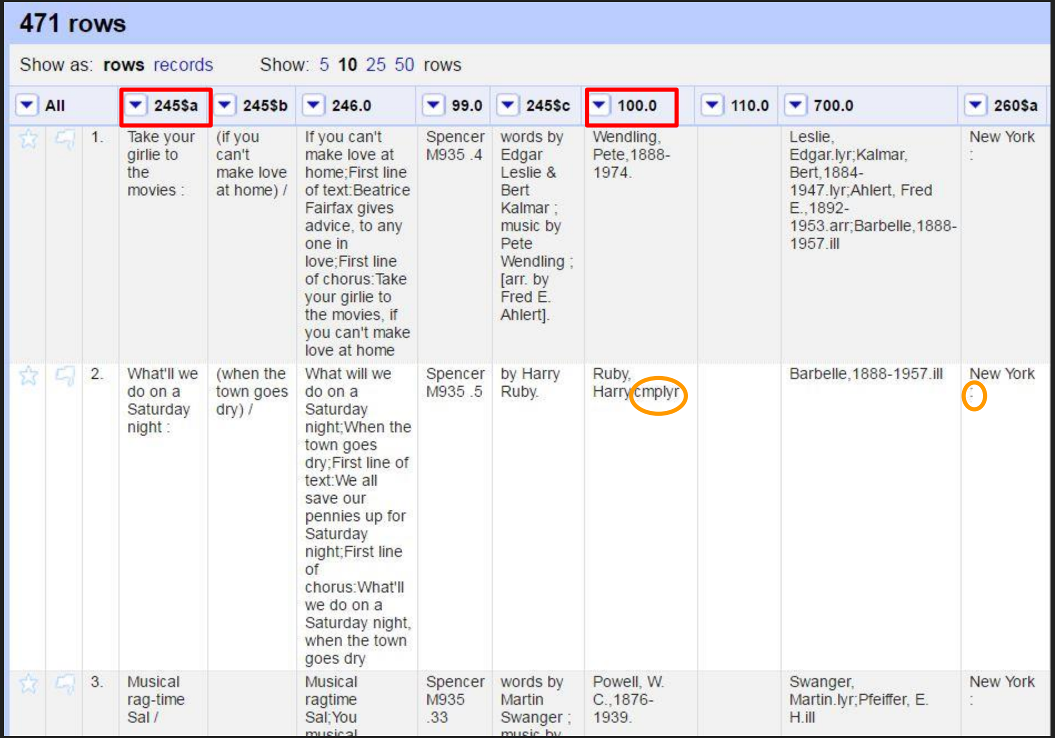


```
LEADER 00000ncm  2200000Ia 4500
001    426135815
003    OCoLC
005    20100607104730.0
008    090717s1910    nyurga         n    zxx d
035    (OCoLC)426135815
040    SST|cSST|dIYU
048    vn01|aka01
049    IYUU
099    Spencer D1732 .15
099    Spencer D1732 .15-2
100 1  Botsford, George,|d1874-1949.
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
```

Gentlemen, you can't fight in here.
This is the War Room!



471 rows

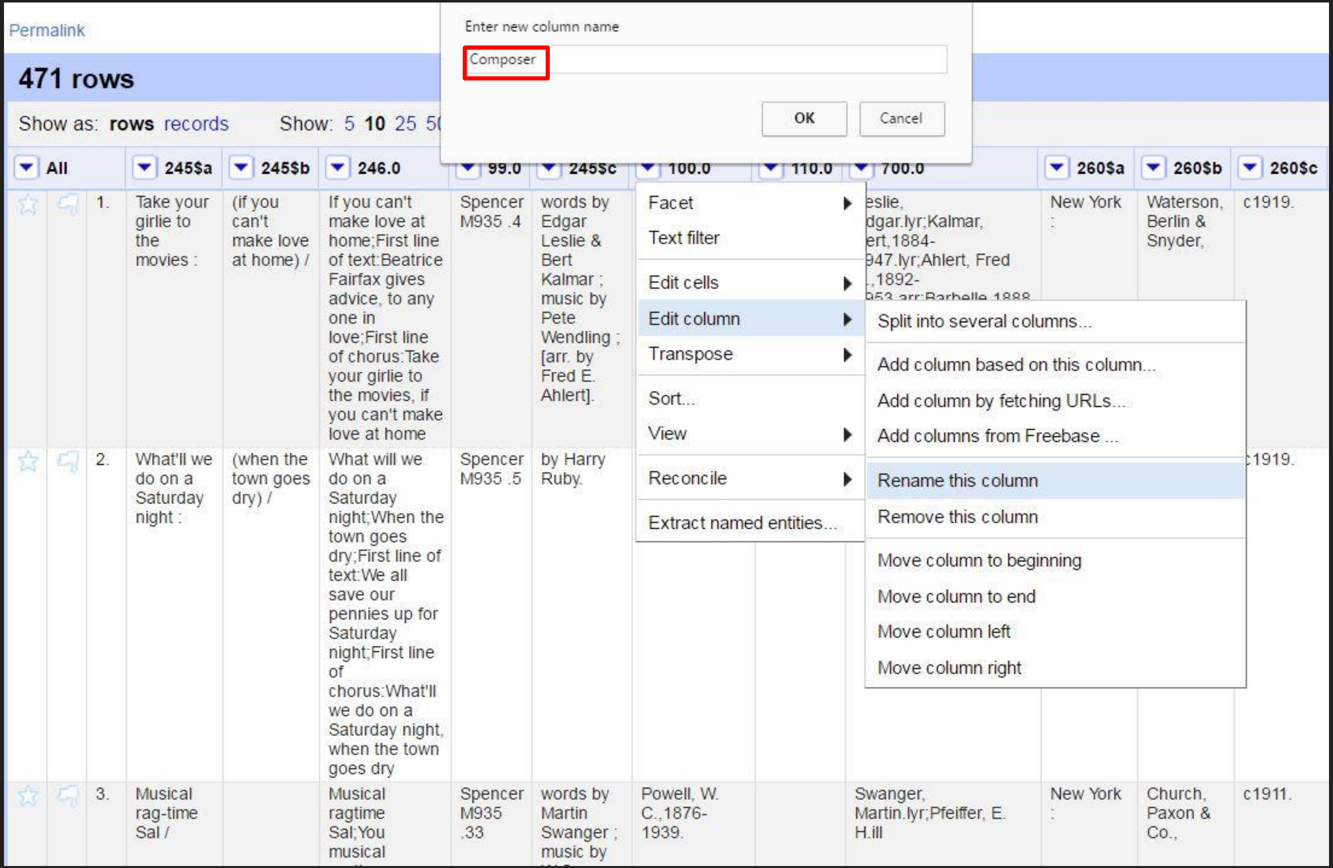
Show as: **rows** records   Show: 5 **10** 25 50 rows

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.cmplyr		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 :

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



Composer

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].			eslie, dgar.lyr;Kalmar, ert,1884-947.lyr;Ahlert, Fred .,1892-953.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.						:1919.
☆		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 :	Church, Paxon & Co.,	c1911.

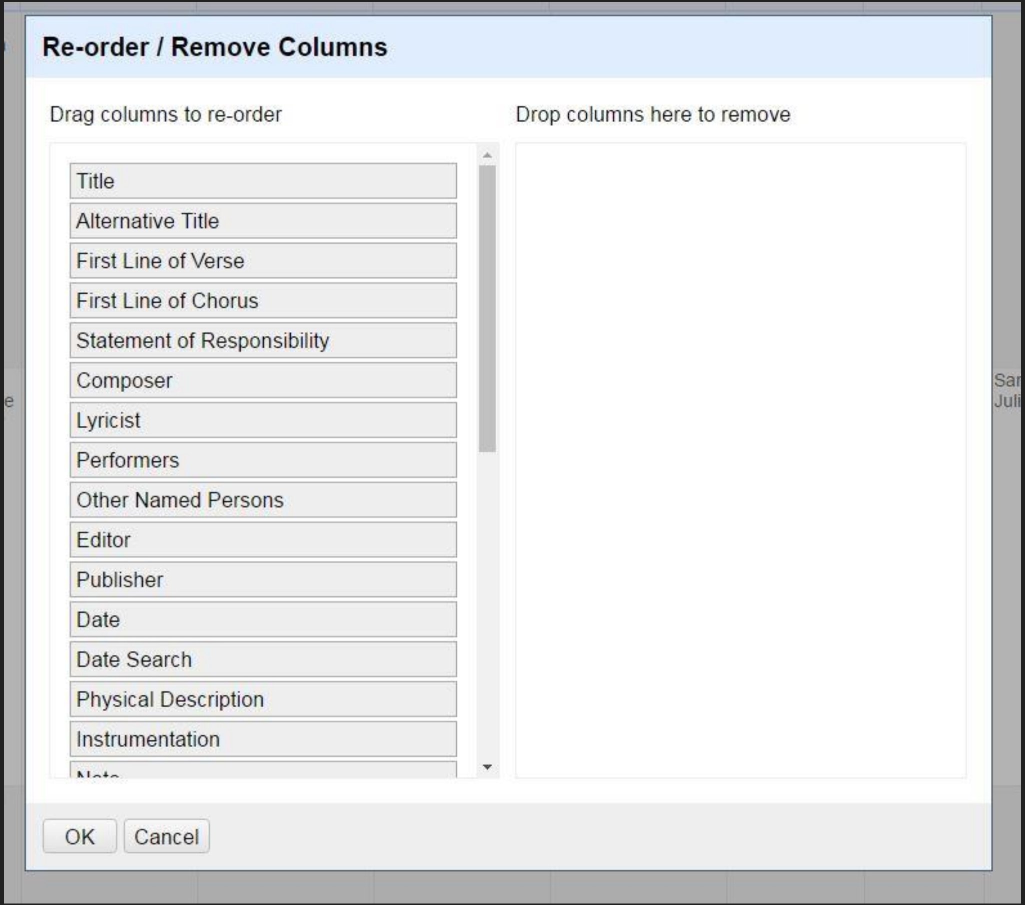
Columns are the primary units of interaction. The drop down menu of functions at the column level allows us to rename, reorder, or transform columns.

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

MARC 100 → Composer

Re-ordering columns

Columns must also exactly match the order that the corresponding fields appear in our CONTENTdm collection. Once all the fields have been re-named, they can be re-ordered under the All columns menu.

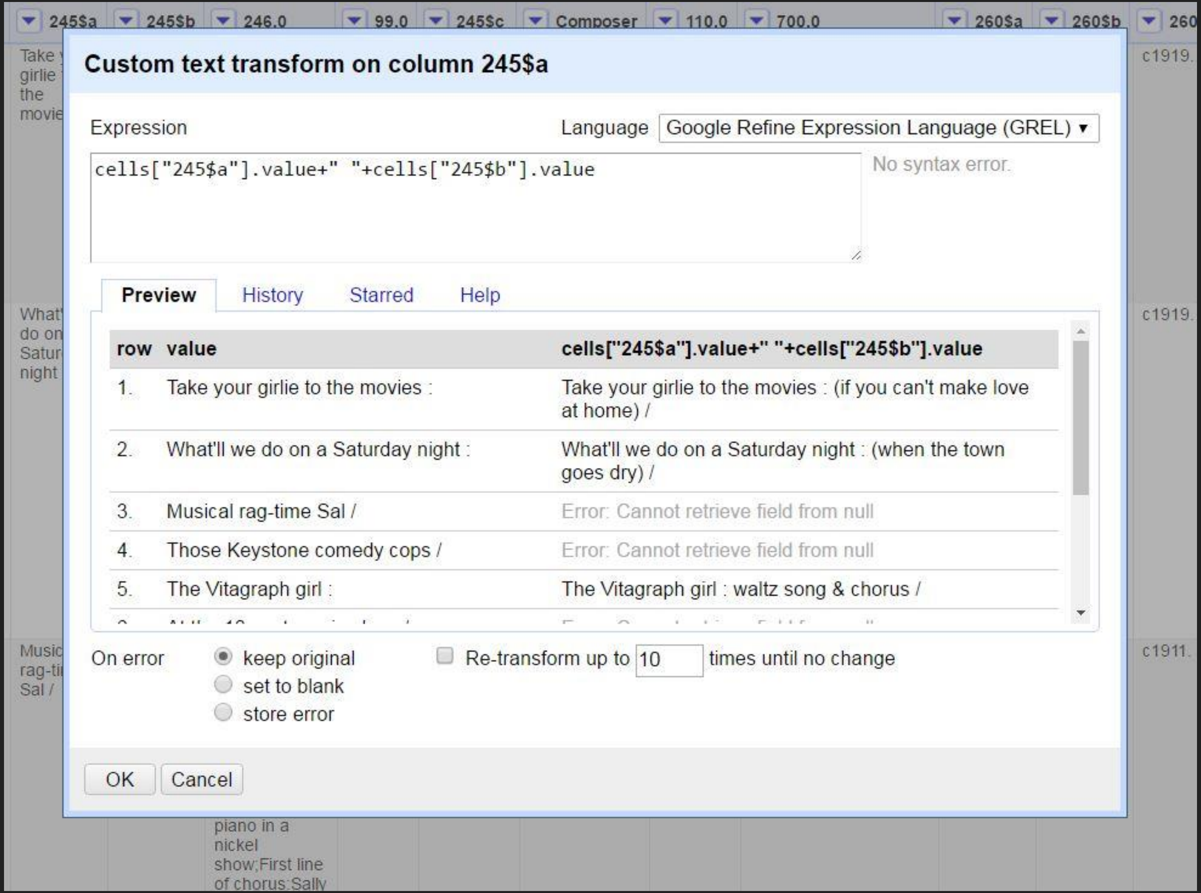


Re-order / Remove Columns

Drag columns to re-order

- Title
- Alternative Title
- First Line of Verse
- First Line of Chorus
- Statement of Responsibility
- Composer
- Lyricist
- Performers
- Other Named Persons
- Editor
- Publisher
- Date
- Date Search
- Physical Description
- Instrumentation
- Note

Drop columns here to remove



Custom text transform on column 245$a

Expression

Language: Google Refine Expression Language (GREL) ▾

```
cells["245$a"].value+" "+cells["245$b"].value
```

No syntax error.

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 /

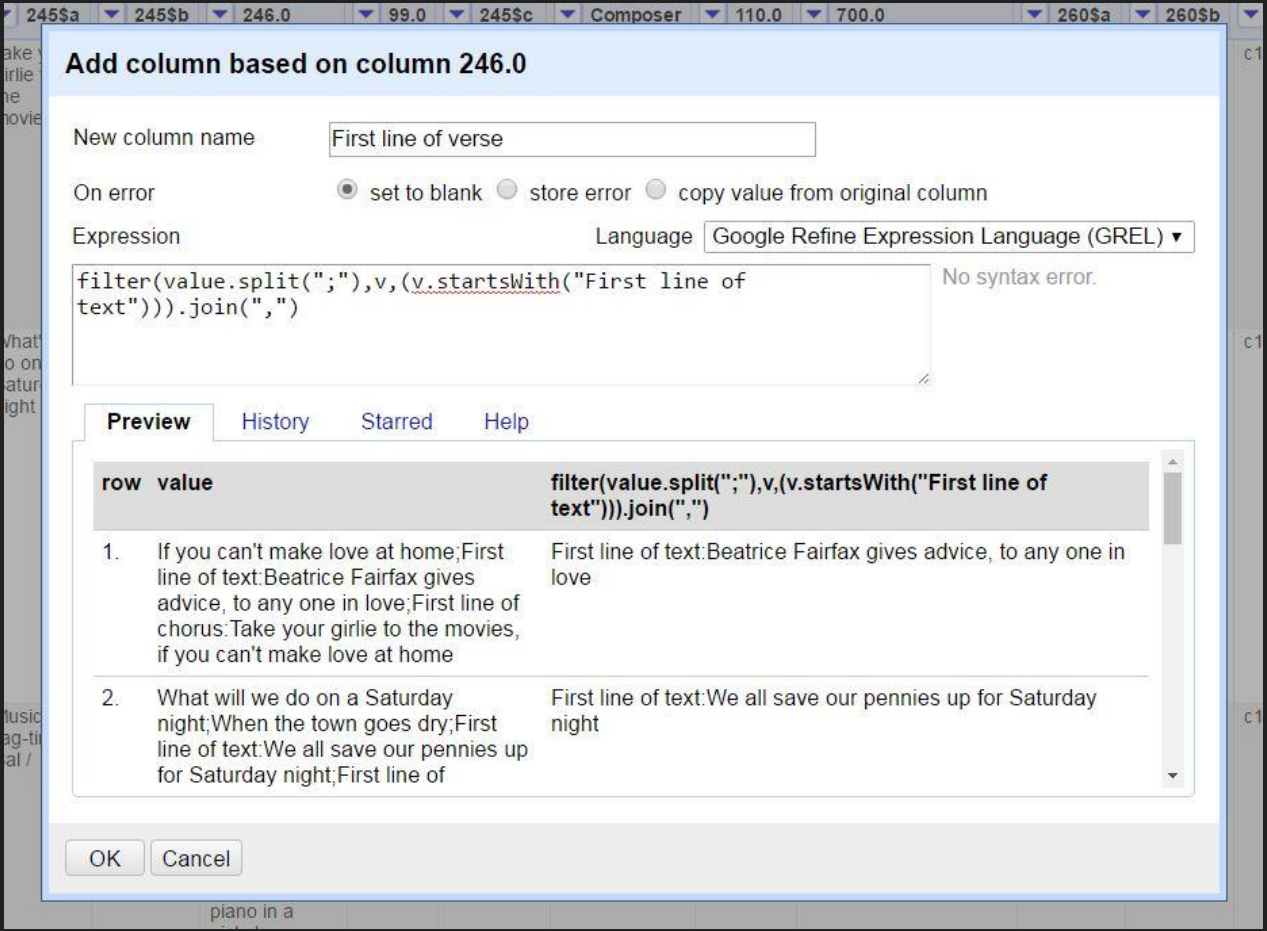
On error: ● keep original ○ set to blank ○ store error

□ Re-transform up to 10 times until no change

Joining Values

Transform data with Google Refine Expression Language (GREL)

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



Add column based on column 246.0

New column name: First line of verse

On error: ● set to blank ○ store error ○ copy value from original column

Expression — Language: Google Refine Expression Language (GREL) ▾

```
filter(value.split(";"),v,(v.startsWith("First line of
text"))).join(",")
```

No syntax error.

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

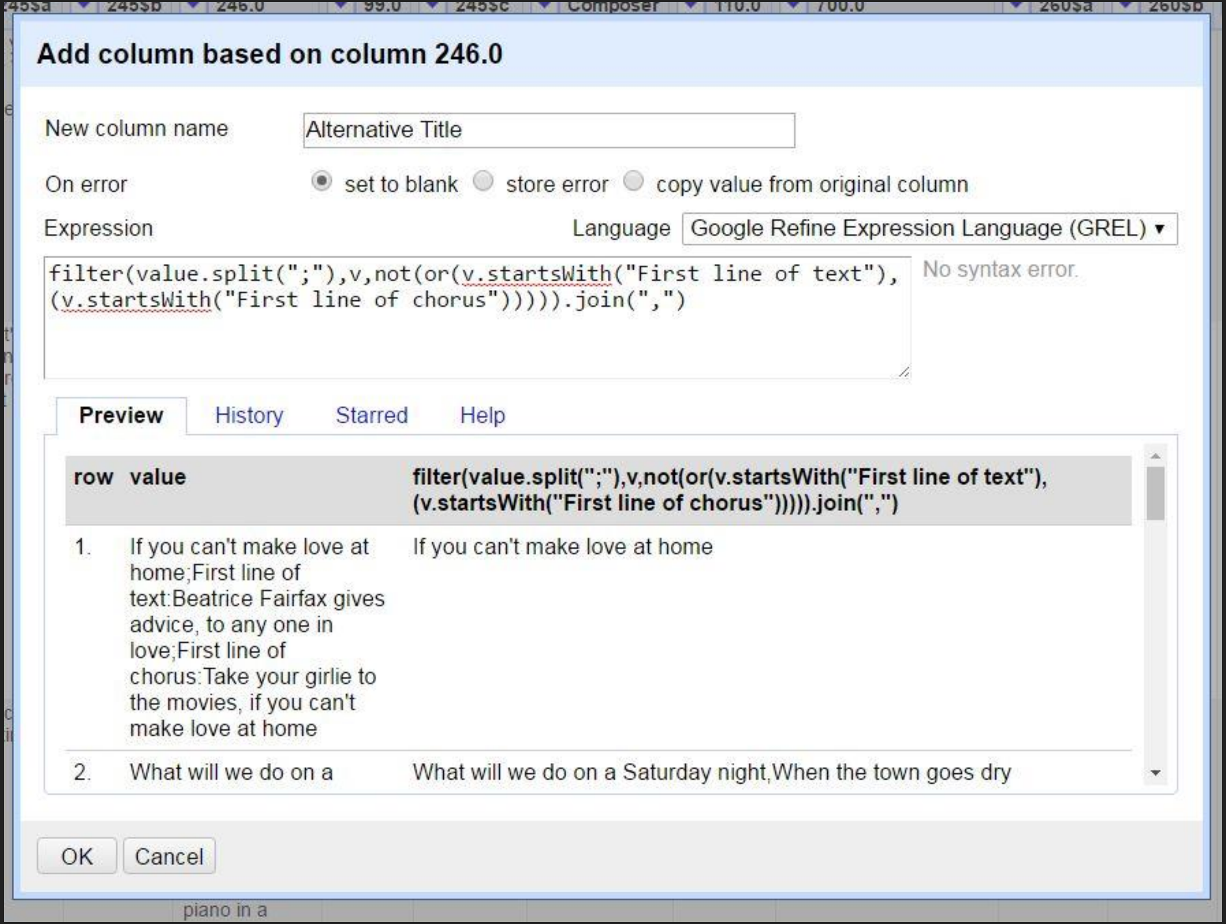
Splitting values

The 246 must be split into two or three fields:

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

Splitting values

Know your data!



Add column based on column 246.0

New column name: Alternative Title

On error: ● set to blank ○ store error ○ copy value from original column

Expression

Language: Google Refine Expression Language (GREL) ▾

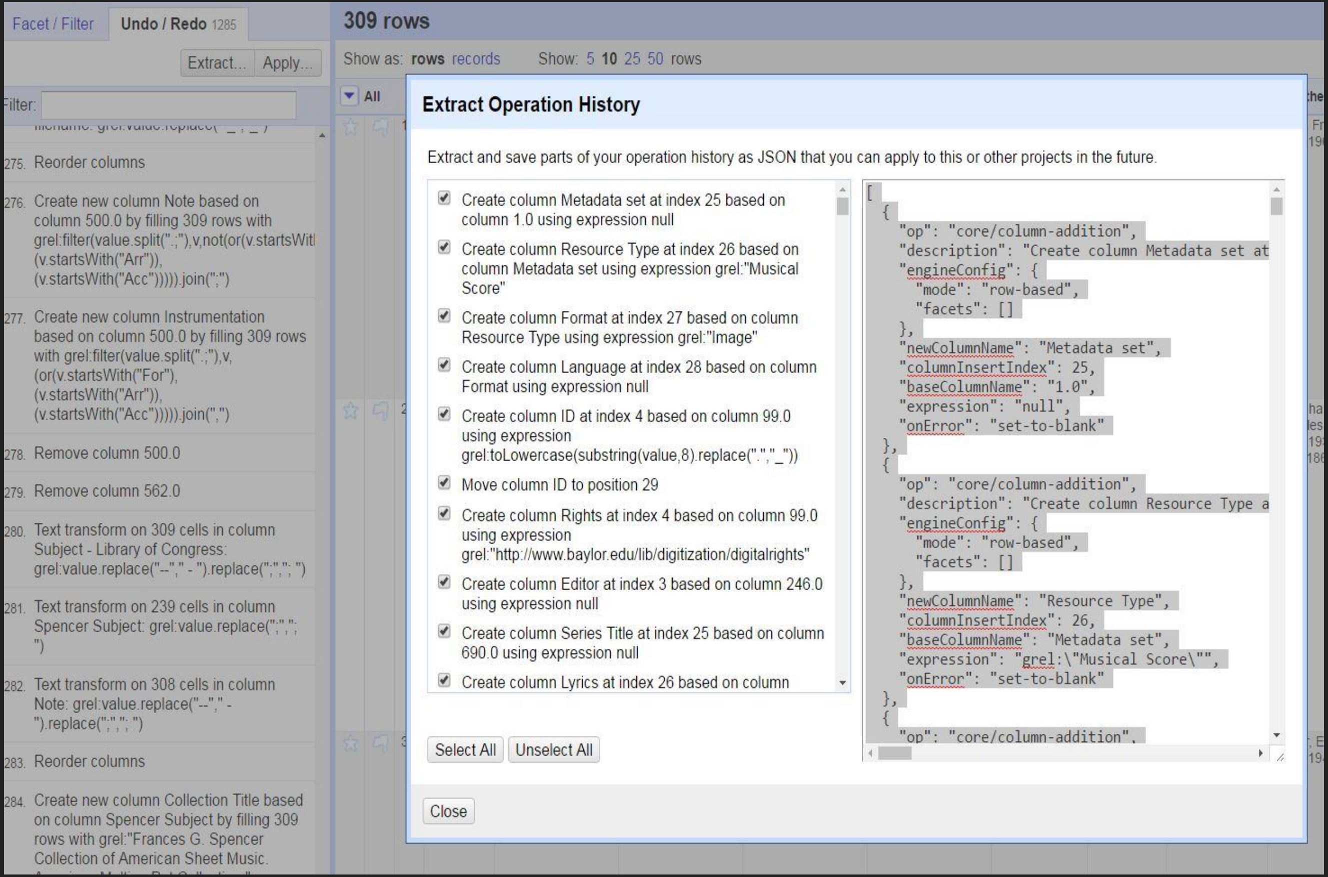
```
filter(value.split(";"),v,not(or(v.startsWith("First line of text"),
(v.startsWith("First line of chorus"))))).join(",")
```

No syntax error.

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 one 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

Extract and save operation history

Apply this to new data sets that need the same kind of clean up



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.startsWit (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

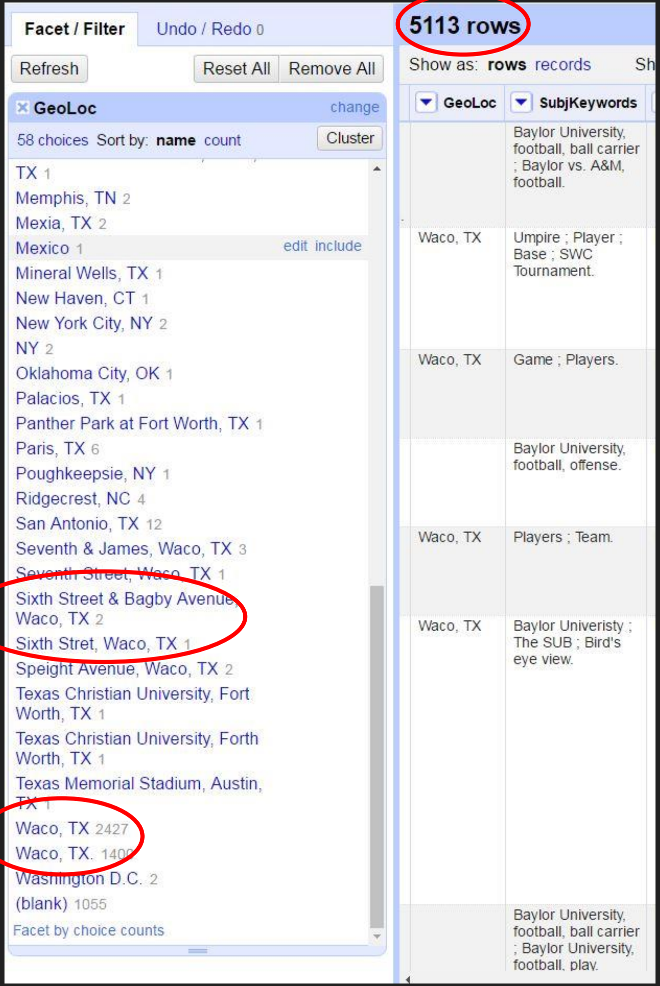
284. Create new column Collection Title based on column Spencer Subject by filling 309 rows with grel:"Frances G. Spencer Collection of American Sheet Music.

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

```
[
  {
    "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",
```



TX	1
Memphis, TN	2
Mexia, TX	2
Mexico	1
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.	1400
Washington D.C.	2
(blank)	1055

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 Univeristy ; The SUB ; Bird's eye view.
	Baylor University, football, ball carrier ; Baylor University, football, play.

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

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.