Information Integration on the Web: Homework 5

Due on February 20, 2015

Prof. Ambite & Knoblock

Tushar Tiwari

Problem 1

Briefly describe the selected fields. [15 points]

Solution

display

The display field states the manner in which the painting is displayed. Whether it is displayed within a frame or just a canvas or a panel or just a sheet.

acquired_id

The acquired_id is a sort of transaction id for the acquisition of the painting.

acquired_from

The acquired_from is the description of from whom was the painting acquired and by what means.

artist2 - last,first name/link/birth year and death year

In the case when the painting has only one author then artist1 - (...) fields are filled in with artist2 - (...) fields empty. But if a painting has multiple authors then there are multiple rows for that many authors and the author2 - (...) fields are filled in and the author1 - (...) fields empty.

bibliography

Bibliography is the list of references used to describe the background of the painting.

Problem 2

Explain the cleaning operations performed for each field. [50 points]

Solution

display

The display field had values like framed, Framed, Frame and unspecified, (blank). So these values were combined. To combine these values first create a text facet on the display column. Then I manually edited in the facets. Changed Frame, framed to Framed and changed unspecified to (blank)

acquired_from

The acquired_from field was missing data due to a faulty pull of data. While scraping the data, to separate acquired_id from acquired_from a split on '('was done. However that did not work out well as there was a '('in the text of acquired_from which was now copied into acquired_id. So we basically had to strip acquired_id of this text which belonged to acquired_from and so we append the value of acquired_id to acquired_from. To filter out such records do a filter on acquired_id with 'by exchange'. Then do a transform on acquired_from.

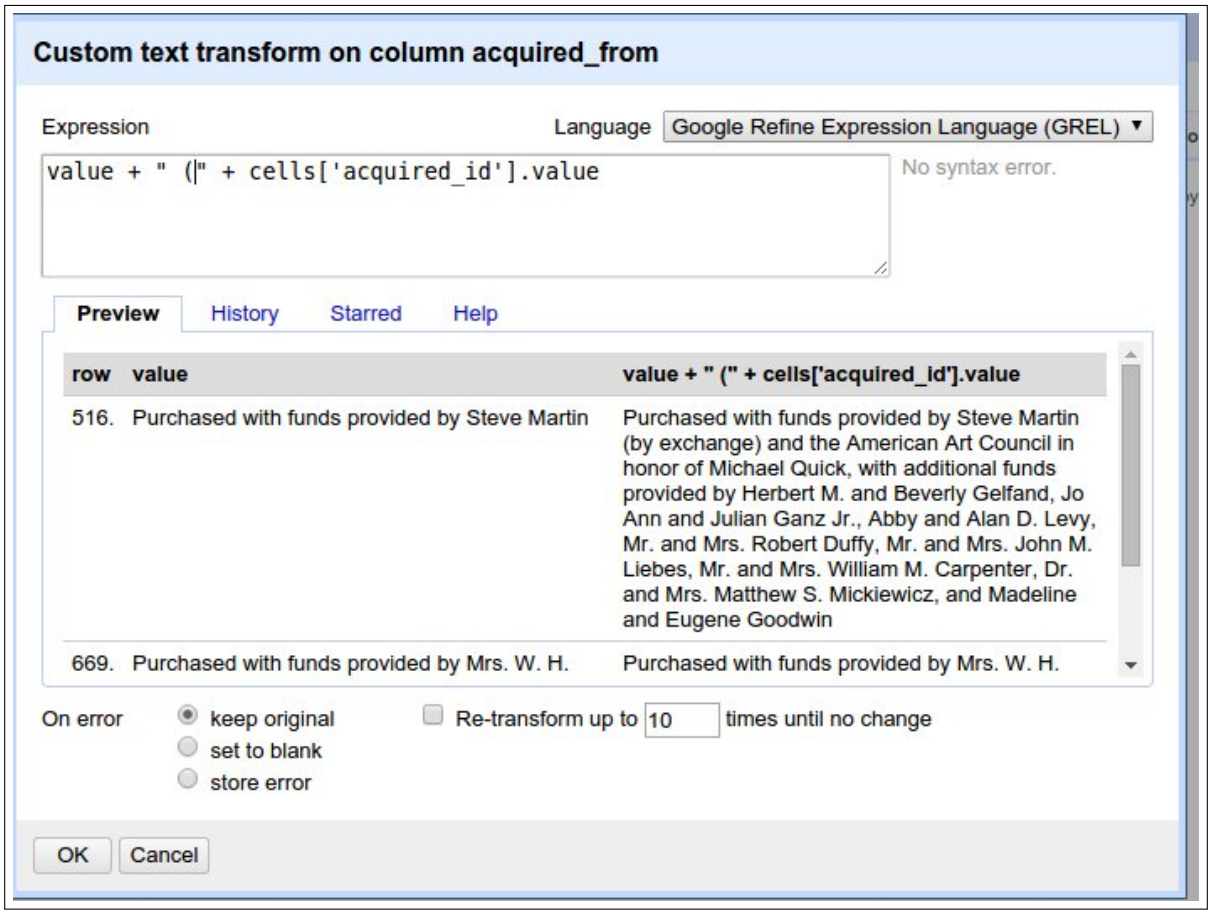


Figure 1: Transform acquired_from column

acquired_id

Since this data is not available for the rows whose acquired_from was transformed previously(luckily only 2), we manually add the data.

artist2 - (...)

So these fields have been described before already. Quick note, after creating the project I have renamed all these columns from __anonymous__ - artist - __anonymous__ - (...) to artist2 - (...). Basically we want multiple rows of the painting for multiple authors which is already provided by openrefine while importing. Only problem the artist data exists in the artist2 fields. So all I have to do is copy every artist2 - (...) field to corresponding artist - (...) field.

- I create a customized facet called facet by blank on artist - name - last and get all true records. That is, all records for whom artist - name - last is blank.
- Then on those records I create another customized blank facet on artist2 - name - last is false which fetched all records with some value other than blank for artist2 - name - last. Note that all records for which some value of artist2 exists artist will be blank by design.
- Now transform cells of artist - name - last by copying the artist2 - name - last.
- Repeat for all artist2 - (...) cells
- After copying all columns delete all the artist2 columns.

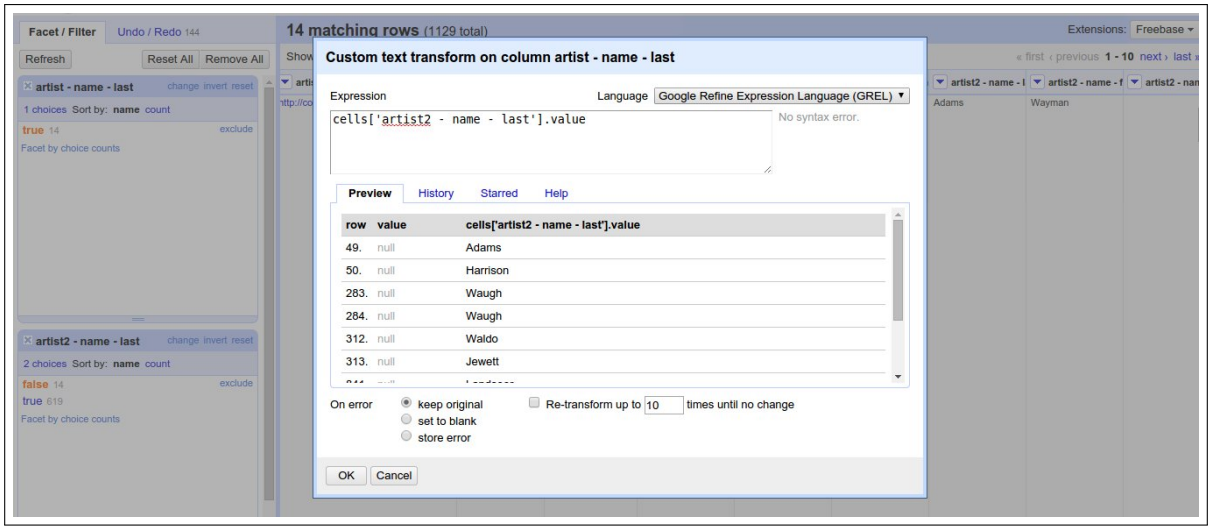


Figure 2: Transform artist - name - last column

bibliography

For every item in the bibliography section of the painting, openrefine creates a new record for that painting. So we want to combine those different records that contain the different items of the bibliography section. To do this we shall

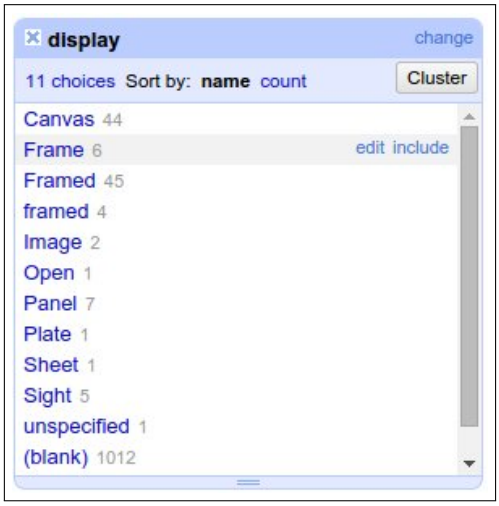
- Blank down on the link column if not already blank downed. If multiple records have similar links then only the topmost record will have the link. Make sure that the sort is on the links before blanking down.
- Move link column to beginning.
- Now in bibliography goto edit cells >> Join multivalued cell. For the separator choose a unique separator such that a split can be performed at a later time on this data (I choose 'abcdefgh'). This action will push up and append all the different values of bibliography for that record into the topmost record and all lower records' bibliography value becomes null.
- All the extra rows created due to bibliography can be deleted because they will not contain any information for any column.
- Then do a secure fill down i.e. transform the column by doing row.record.cells["bibliography"].value[0]. This basically copies down the value of bibliography to same paintings only. A fill down will copy unless a new value is encountered.
- Finally transform the cell values by doing a forEachIndex(value.split('abcdefgh'), i ,val, (i+1) + ". " + val).join(" "). This will nicely add a numbered ordering to the bibliography items.

Problem 3

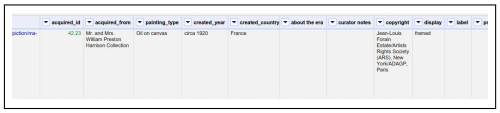
Show one example of the cleaning done for each field. Show the full record before and after cleaning.. [15 points]

Solution

display



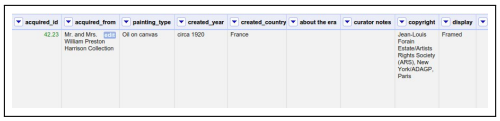
(a) The facets



(b) A record

Figure 3: Before cleaning display

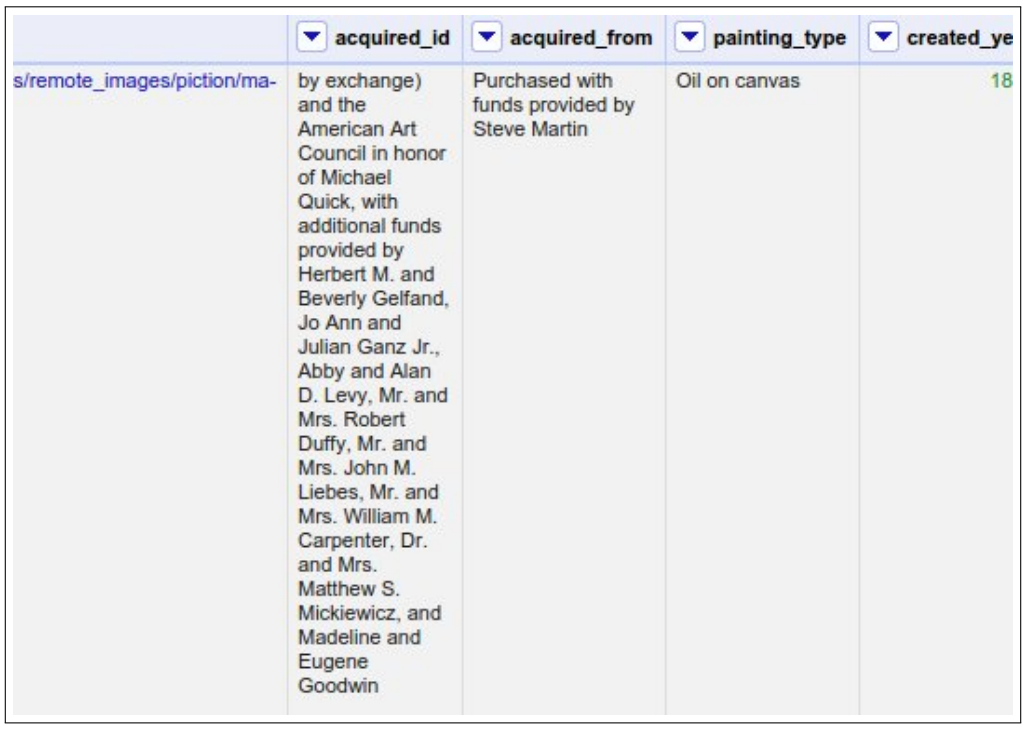
(a) The facets



(b) A record

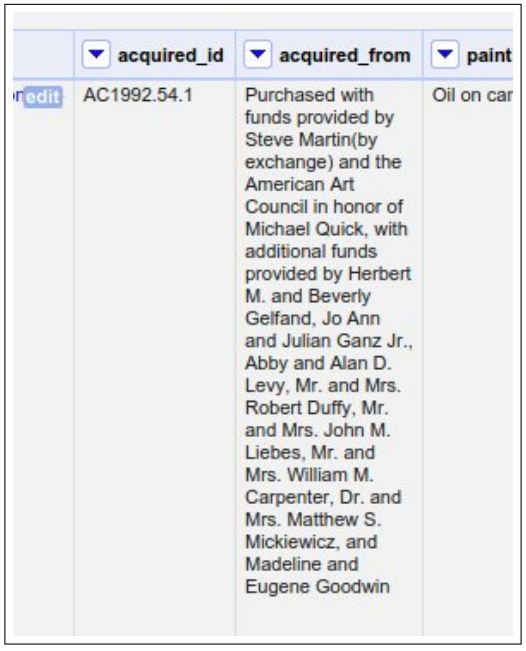
Figure 4: After cleaning display

acquired_from & acquired_id



	acquired_id	acquired_from	painting_type	created_ye
s/remote_images/piction/ma-	by exchange) and the American Art Council in honor of Michael Quick, with additional funds provided by Herbert M. and Beverly Gelfand, Jo Ann and Julian Ganz Jr., Abby and Alan D. Levy, Mr. and Mrs. Robert Duffy, Mr. and Mrs. John M. Liebes, Mr. and Mrs. William M. Carpenter, Dr. and Mrs. Matthew S. Mickiewicz, and Madeline and Eugene Goodwin	Purchased with funds provided by Steve Martin	Oil on canvas	18

Figure 5: Before cleaning



	acquired_id	acquired_from	paint
edit	AC1992.54.1	Purchased with funds provided by Steve Martin(by exchange) and the American Art Council in honor of Michael Quick, with additional funds provided by Herbert M. and Beverly Gelfand, Jo Ann and Julian Ganz Jr., Abby and Alan D. Levy, Mr. and Mrs. Robert Duffy, Mr. and Mrs. John M. Liebes, Mr. and Mrs. William M. Carpenter, Dr. and Mrs. Matthew S. Mickiewicz, and Madeline and Eugene Goodwin	Oil on car

Figure 6: After cleaning

artist2 - (...)

(a) 1st artist of some painting

(b) 2nd artist of same painting

Figure 7: Before cleaning

(a) 1st artist of some painting

(b) 2nd artist of same painting

Figure 8: After cleaning

Figure 9: A filter on the artists of the particular painting returns rightly two entries

bibliography

(a) 1st item of bibliography

(b) 3rd item of bibliography

(c) 3rd item of bibliography

Figure 10: Before Cleaning

(a) 1st author's bibliography

(b) 2nd author's bibliography

Figure 11: After Cleaning