

HUMANITIES DATA PRAXIS

Elizabeth Wickes (Research Data Services, UIUC)

Eleanor Dickson (Scholarly Commons, UIUC)

Andrea Thomer (School of Information Sciences, UIUC)

10 March 2017

Humanities data — *the very idea*

Data?

\$#%@@@\$# !!

Data: some definitions

a collection of facts from which conclusions may be drawn
(wordnet.princeton.edu/perl/webwn)

Factual information (such as measurements and statistics) used as a basis for reasoning, discussion, or calculation.

(www.ic.gc.ca/eic/site/stco-levc.nsf/eng/h_qw00037e.html)

a collection of organised information, usually the result of experience, observation or experiment, ... may consist of numbers, words, or images, particularly as measurements or observations

(en.wikipedia.org/wiki/Data)

A collection of observations.

(www.fs.fed.us/r3/coconino/forest-resources/archaeology)

information, especially information organized for analysis
(www.handsontheland.org/lms/mod/glossary/view.php)

Data, our* definition

Data are *propositions*

- (i) systematically asserted . . .
- (ii) as evidence

*Dubin et al. 2009-2014, various Renear lectures

Data, our* definition

Data are *propositions*

- (i) systematically asserted . . .
- (ii) as evidence

In the sciences data can look like:

A series of observations, rendered as numbers, analyzed with statistics

In non-science, data can look like:

A collection of texts, rendered in words, interpreted qualitatively

In interdisciplinary work data can look like:

All of the above (and more)

*Dubin et al. 2009-2014, various Renear lectures

Your data

What kind of data do you have or work with?

How is it organized and encoded (written onto some sort of media)?

[audience participation]

Data, our* definition

Data are *propositions*

- (i) **systematically asserted. . .**
- (ii) **as evidence**

**When using computers to analyze any kind of data,
they need to be presented in a particularly systematic way –
otherwise the computer won't know what to do with it.**

That's where data cleaning/wrangling/curation comes in...

*Dubin et al. 2009-2014, various Renear lectures

Agenda – an entirely too brief overview of data cleaning tools and techniques

- Hands on: Spreadsheet activity
- Hands on: Tidy data in Excel
- [break]
- Follow along: Normalized data in Open Refine
- Follow along: Charting data in R

CLEANING DATA WITH OPENREFINE

Andrea Thomer

If you're following along -

Locate the file named “**Menu.csv**” and we’ll review the data together

Introducing OpenRefine

OpenRefine is a tool for *manipulating* and *cleaning* data.

OpenRefine looks like a *spreadsheet*,
but it acts like a *database*.

Use OpenRefine to *explore*, *clean*,
and *link* your data.

Today we will cover....

- ✓ Creating a New Project
- ✓ Basic Normalization
- ✓ Faceting and Clustering
- ✓ Records vs. Rows
- ✓ Advanced Transformations
 - (if we have time) (we probably won't)

Creating a Project

Google refine A power tool for working with messy data.

Create Project Open Project Import Project

Get data from

This Computer Locate one or more files on your computer to upload:
 Menu.csv

Web Addresses (URLs)

Clipboard

Google Data


Version 2.5 [r2407]

[Help](#) [About](#)



Creating a Project

Google refine A power tool for working with messy data.

Create Project « Start Over Configure Parsing Options Project name: Menu csv Create Project »

Open Project Import Project

	id	name	sponsor	event	venue	place	physical_description	occasion
1.	12463		HOTEL EASTMAN	BREAKFAST	COMMERCIAL	HOT SPRINGS, AR	CARD; 4.75X7.5;	EASTER;
2.	12464		REPUBLICAN HOUSE	[DINNER]	COMMERCIAL	MILWAUKEE, [WI];	CARD; ILLUS; COL; 7.0X9.0;	EASTER;
3.	12465		NORDDEUTSCHER LLOYD BREMEN	FRUHSTUCK/BREAKFAST;	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE;	CARD; ILLU; COL; 5.5X8.0;	
4.	12466		NORDDEUTSCHER LLOYD BREMEN	LUNCH;	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE;	CARD; ILLU; COL; 5.5X8.0;	

Parse data as

Character encoding: UTF-8 Update Preview

CSV / TSV / separator-based files

Line-based text files commas (CSV) tabs (TSV) custom , Ignore first 0 line(s) at beginning of file
Fixed-width field text files Parse next 1 line(s) as column headers
PC-Axis text files Discard initial 0 row(s) of data
JSON files Load at most 0 row(s) of data
RDF/N3 files
XML files
Open Document Format spreadsheets (.ods)
RDF/XML files

Escape special characters with \

Parse cell text into numbers, dates, ... Store blank rows
 Quotation marks are used to enclose cells containing column separators Store blank cells as nulls
 Load at most Store file source (file names, URLs) in each row

Version 2.5 [r2407] Help About

Basic Normalization

Google refine Menu csv Permalink

Facet / Filter Undo / Redo 0

17079 rows

Show as: rows records Show: 5 10 25 50 rows Extensions: [Freebase](#)

« first < previous 1 - 10 next > last »

All	id	name	sponsor	event	venue	place	physical_descri	occasion	notes
1.	12463		Facet	REAKFAST	COMMERCIAL	HOT SPRINGS, AR	CARD; 4.75X7.5;	EASTER;	
2.	12464		Text filter	[INNER]	COMMERCIAL	MILWAUKEE, [WI];	CARD; ILLUS; COL; 7.0X9.0;	EASTER;	WEDGEWOOD BLUE CARD; WHITE EMBOSSED GREEK KEY BORDER; "EASTER SUNDAY" EMBOSSED IN WHITE; VIOLET COLORED SPRAY OF FLOWERS IN UPPER LEFT CORNER;
			Edit cells	Transform...					
			Edit column	Common transforms	Trim leading and trailing whitespace				
			Transpose	Fill down	Collapse consecutive whitespace				
			Sort...	Blank down					
			View	Unescape HTML entities					
			Reconcile	To titlecase					
				To uppercase					
				To lowercase					
3.	12465		NORDDEUTSCHER LLOYD BREMEN	FRUHSTUCK/BREAKFAST;	COMM				MENU IN GERMAN AND ENGLISH; ILLUS, STEAMSHIP AND SAILING VESSEL;
4.	12466		NORDDEUTSCHER LLOYD BREMEN	LUNCH;	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE;	CARD; ILLU; COL; 5.5X8.0;		MENU IN GERMAN AND ENGLISH; ILLUS, HARBOR SCENE WITH SAILING VESSEL;

Using facets and filters 

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

A red arrow points to the "Collapse consecutive whitespace" option in the "Common transforms" submenu.

Other ones to try:

- Collapse consecutive whitespace
- Change to title/lower/upper case

A Note on Tracking Project History

The screenshot shows the Google Refine interface with a red arrow pointing to the "Undo / Redo" tab at the top left. The main area displays 17079 rows of data with various columns like id, name, sponsor, event, venue, place, physical_descri, occasion, and notes. The "notes" column contains detailed descriptions of the transformations applied to each row. The first few rows show transformations such as "Text transform on 14 cells in column sponsor: value.trim()", "Text transform on 126 cells in column sponsor: value.replace(/\s+,'')", and "Text transform on 8069 cells in column sponsor: value.toUpperCase()". The "notes" column for the third row, for example, includes "WEDGEWOOD BLUE CARD; WHITE EMBOSSED GREEK KEY BORDER; "EASTER SUNDAY" EMBOSSED IN WHITE; VIOLET COLORED SPRAY OF FLOWERS IN UPPER LEFT CORNER;".

id	name	sponsor	event	venue	place	physical_descri	occasion	notes
12463	HOTEL EASTMAN	BREAKFAST	COMMERCIAL	HOT SPRINGS, AR	CARD; 4.75X7.5;	EASTER;		
12464	REPUBLICAN HOUSE	[DINNER]	COMMERCIAL	MILWAUKEE, [WI];	CARD; ILLUS; COL; 7.0X9.0;	EASTER;		WEDGEWOOD BLUE CARD; WHITE EMBOSSED GREEK KEY BORDER; "EASTER SUNDAY" EMBOSSED IN WHITE; VIOLET COLORED SPRAY OF FLOWERS IN UPPER LEFT CORNER;
12465	NORDDEUTSCHER LLOYD BREMEN	FRUHSTUCK/BREAKFAST;	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE;	CARD; ILLU; COL; 5.5X8.0;			MENU IN GERMAN AND ENGLISH; ILLUS, STEAMSHIP AND SAILING VESSEL;
12466	NORDDEUTSCHER LLOYD BREMEN	LUNCH;	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE;	CARD; ILLU; COL; 5.5X8.0;			MENU IN GERMAN AND ENGLISH; ILLUS, HARBOR SCENE WITH SAILING VESSEL;

Don't be afraid to make mistakes!

...you can always use the Undo/Redo tab to navigate to earlier stages in your project.

Faceting and Clustering

Google refine Menu csv Permalink Open... Export Help

Facet / Filter Undo / Redo 3

Using facets and filters 

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

17079 rows Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

Extensions: Freebase ▾

All	<input type="checkbox"/> id	<input type="checkbox"/> name	<input type="checkbox"/> sponsor	<input type="checkbox"/> event	<input type="checkbox"/> venue	<input type="checkbox"/> place	<input type="checkbox"/> physical_descrip	<input type="checkbox"/> occasion	<input type="checkbox"/> notes	
1.	12463			Facet ▾	Text facet	ERCIAL	HOT SPRINGS, AR	CARD; 4.75X7.5;	EASTER;	
2.	12464				Numeric facet	ERCIAL	MILWAUKEE, [WI];	CARD; ILLUS; COL; 7.0X9.0;	EASTER;	WEDGEWOOD BLUE CARD; WHITE EMBOSSED GREEK KEY BORDER; "EASTER SUNDAY" EMBOSSED IN WHITE; VIOLET COLORED SPRAY OF FLOWERS IN UPPER LEFT CORNER;
3.	12465		NORDDEUTSCHER LLOYD BREMEN	FRUHSTUCK/BREAKFAST;	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE;	CARD; ILLU; COL; 5.5X8.0;			MENU IN GERMAN AND ENGLISH; ILLUS, STEAMSHIP AND SAILING VESSEL;
4.	12466		NORDDEUTSCHER LLOYD BREMEN	LUNCH;	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE;	CARD; ILLU; COL; 5.5X8.0;			MENU IN GERMAN AND ENGLISH; ILLUS, HARBOR SCENE WITH SAILING VESSEL;

A red arrow points to the "Text facet" option in the "Facet" submenu of the "event" column's context menu.

Faceting and Clustering

Google refine Menu csv Permalink

Facet / Filter Undo / Redo 3

Refresh Reset All Remove All

sponsor change
6080 choices Sort by: name count Cluster

17079 rows

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

All	id	name	sponsor	event	venue	place	physical_descri	occasion	notes
1.	12463		HOTEL EASTMAN	BREAKFAST	COMMERCIAL	HOT SPRINGS, AR	CARD; 4.75X7.5;	EASTER;	
2.	12464		REPUBLICAN HOUSE	[DINNER]	COMMERCIAL	MILWAUKEE, [WI];	CARD; ILLUS; COL; 7.0X9.0;	EASTER;	WEDGEWOOD BLUE CARD; WHITE EMBOSSED GREEK KEY BORDER; "EASTER SUNDAY" EMBOSSED IN WHITE; VIOLET COLORED SPRAY OF FLOWERS IN UPPER LEFT CORNER;
3.	12465		NORDDEUTSCHER LLOYD BREMEN	FRUHSTUCK/BREAKFAST;	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE;	CARD; ILLU; COL; 5.5X8.0;		MENU IN GERMAN AND ENGLISH; ILLUS, STEAMSHIP AND SAILING VESSEL;
4.	12466		NORDDEUTSCHER LLOYD BREMEN	LUNCH;	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE;	CARD; ILLU; COL; 5.5X8.0;		MENU IN GERMAN AND ENGLISH; ILLUS, HARBOR SCENE WITH SAILING VESSEL;

Faceting and Clustering

Cluster & Edit column "sponsor"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method key collision Keying Function fingerprint 213 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
8	24	<ul style="list-style-type: none">• RED STAR LINE - ANTWERPEN - NY (7 rows)• RED STAR LINE - ANTWERPEN NY (6 rows)• RED STAR LINE - ANTWERPEN -NY (5 rows)• RED STAR LINE -ANTWERPEN NY (2 rows)• RED STAR LINE -ANTWERPEN - NY (1 rows)• RED STAR LINE -ANTWERPEN -NY (1 rows)• RED STAR LINE- ANTWERPEN -NY (1 rows)• RED STAR LINE- ANTWERPEN NY (1 rows)	<input checked="" type="checkbox"/>	RED STAR LINE - ANTWERPEN - NY
6	666	<ul style="list-style-type: none">• NORDDEUTSCHER LLOYD BREMEN (629 rows)• NORDDEUTSCHER LLOYD - BREMEN (31 rows)• NORDDEUTSCHER LLOYD BREMEN; (2 rows)• NORDDEUTSCHER LLOYD, BREMEN (2 rows)• BREMEN NORDDEUTSCHER LLOYD (1 rows)• NORDDEUTSCHER LLOYD -BREMEN (1 rows)	<input type="checkbox"/>	NORDDEUTSCHER LLOYD BREMEN
6	31	<ul style="list-style-type: none">• FIFTH AVENUE HOTEL (22 rows)• (FIFTH AVENUE HOTEL) (3 rows)• (FIFTH AVENUE HOTEL?) (2 rows)• FIFTH AVENUE HOTEL (?) (2 rows)• (FIFTH AVENUE HOTEL?) (1 rows)• FIFTH AVENUE HOTEL; (1 rows)	<input type="checkbox"/>	FIFTH AVENUE HOTEL

Select All Deselect All Merge Selected & Re-Cluster Merge Selected & Close Close

Choices in Cluster

Rows in Cluster

Average Length of Choices

Length Variance of Choices

Export Help

ns: Firebase

- 10 next > last »

notes

WEDGEWOOD
BLUE CARD;
WHITE
EMBOSSED
GREEK KEY
BORDER;
"EASTER
SUNDAY"
EMBOSSED IN
WHITE;
VIOLET
COLORED
SPRAY OF
FLOWERS IN
UPPER LEFT
CORNER;
MENU IN
GERMAN AND
ENGLISH;
ILLUS,
STEAMSHIP
AND SAILING
VESSEL;
MENU IN
GERMAN AND
ENGLISH;
ILLUS,
HARBOR
SCENE WITH
SAILING
VESSEL;
MENU IN
GERMAN AND
ENGLISH;
ILLUS,
HARBOR

Kinds of Clustering

PAUSE before continuing

❖ Key collision (fastest, safest)

- Fingerprint, Ngram Fingerprint = defaults
 - Match normalized strings in different ways
- Metaphone = English pronunciation

❖ Nearest Neighbor

- PPM = Partial matching
- Levenshtein = edit distance

This can be computationally intensive – and can sometimes crash – so use with caution or patience on large data sets

Be careful when clustering

This is part of that important “human-only” work we’ve talked about

Hotel Savoy
59th St. & 5th Ave.
New York, New York

Savoy Hotel
Strand
London WC2R 0EU
United Kingdom

Google refiner

Facet / Filter

Refresh

sponsor

6080 choices Sort

? 57

? (J B) 1

? CLUB 1

? HOTEL 1

'95 LAW OF COLLEGE UNIVERSITY 1

'975 CLASS DINING

'POSSUM CLUB (?COLONIAL HOTEL)

(238 EIGHT AVENUE)

(ABbas II HILMI EGYPT) 1

Cluster & Edit column "sponsor"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method key collision

Keying Function fingerprint

213 clusters found

Choices in Cluster

Rows in Cluster

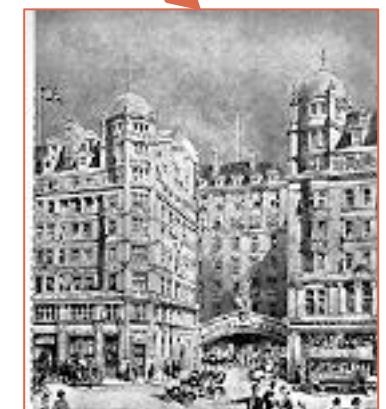
Average Length of Choices

Length Variance of Choices

WEDGEWOOD BLUE CARD; WHITE EMBOSSED GREEK KEY BORDER; "EASTER SUNDAY" EMBOSSED IN WHITE; VIOLET COLOURED SPRAY OF FLOWERS IN UPPER LEFT CORNER; MENU IN GERMAN AND ENGLISH; ILLUS STEAMSHIP AND SAILING VESSEL; MENU IN GERMAN AND ENGLISH; ILLUS HARBOR SCENE WITH SAILING VESSEL; MENU IN GERMAN AND ENGLISH; ILLUS

Method	Keying Function	Count	Value
key collision	fingerprint	213	LA CREPE
key collision	fingerprint	2	CAFE DE PARIS
key collision	fingerprint	2	HOTEL SAVOY
key collision	fingerprint	32	SAVOY HOTEL
key collision	fingerprint	2	S.S. MINNETONKA
key collision	fingerprint	2	CAFE BOULEVARD
key collision	fingerprint	2	S.S. "ILE DE FRANCE"
key collision	fingerprint	2	DE L'ANGE HOTEL
key collision	fingerprint	2	ST. REGIS HOTEL
key collision	fingerprint	78	ST. REGIS HOTEL
key collision	fingerprint	1	LA CREPE
key collision	fingerprint	1	CAFE DE PARIS
key collision	fingerprint	1	HOTEL SAVOY
key collision	fingerprint	1	S.S. MINNETONKA
key collision	fingerprint	1	CAFE BOULEVARD
key collision	fingerprint	1	S.S. "ILE DE FRANCE"
key collision	fingerprint	1	DE L'ANGE HOTEL
key collision	fingerprint	1	ST. REGIS HOTEL

Select All Deselect All Merge Selected & Re-Cluster Merge Selected & Close Close



Records vs. Rows

Google refine Menu csv Permalink Open... Export Help

Facet / Filter Undo / Redo 0 17079 rows Extensions: Named-entity recognition ▾ Freebase ▾ RDF ▾

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

Using facets and filters 

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

	All	id	name	sponsor	event	venue	place	physical_descri	occasion	notes	call
1.	12463		HOTEL EASTMAN	BREAKFAST		COMMERCIAL	HOT SPRINGS, AR	Facet			1900-28
2.	12464		REPUBLICAN HOUSE	[DINNER]		COMMERCIAL	MILWAUKEE, [WI];	Text filter			WEDGEWOOD BLUE CARD; WHITE EMBOSSED GREEK KEY BORDER; "EASTER SUNDAY" EMBOSSED IN WHITE; VIOLET COLORED SPRAY OF FLOWERS IN UPPER LEFT CORNER;
3.						COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE;	Transform...	Edit cells		1900-28
4.						COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE;	Common transforms	Edit column		
								Fill down	Transpose		
								Blank down			
								Sort...			
								Split multi-valued cells...			
								Join multi-valued cells...			
								Reconcile			
								Cluster and edit...			
								Extract named entities...			

The page at 127.0.0.1:3333 says:
What separator currently separates the values?

Cancel OK

Multiple Values, Multiple Rows

Google refine Menu csv Permalink

Facet / Filter Undo / Redo 1

Using facets and filters 

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

45389 rows 

Show as: [rows](#) [records](#) Show: [5](#) [10](#) [25](#) [50](#) rows

Extensions: [Named-entity recognition](#) [Freebase](#) [RDF](#)

	All	id	name	sponsor	event	venue	place	physical_descrip	occasion	notes	call_number
1.	12463		HOTEL EASTMAN	BREAKFAST		COMMERCIAL	HOT SPRINGS, AR	CARD	EASTER;		1900-2822
2.											4.75X7.5
3.											
4.	12464		REPUBLICAN HOUSE	[DINNER]		COMMERCIAL	MILWAUKEE, [WI];	CARD	EASTER;	WEDGEWOOD BLUE CARD; WHITE EMBOSSED GREEK KEY BORDER; "EASTER SUNDAY" EMBOSSED IN WHITE; VIOLET COLORED SPRAY OF FLOWERS IN UPPER LEFT CORNER;	1900-2825
5.								ILLUS			
6.								COL			
7.								7.0X9.0			
8.											
9.	12465		NORDDEUTSCHER LLOYD BREMEN	FRUHSTUCK/BREAKFAST;	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE;	CARD			MENU IN GERMAN AND ENGLISH; ILLUS, STEAMSHIP AND SAILING VESSEL;	1900-2827
10.								ILLU			

One Record, Multiple Rows

Google refine Menu csv Permalink Open... Export Help

Facet / Filter Undo / Redo 1

Using facets and filters 

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

24062 records Extensions: Named-entity recognition ▾ Freebase ▾ RDF ▾

Show as: rows records Show: 5 10 25 50 records « first < previous 1 - 10 next > last »

All	id	name	sponsor	event	venue	place	physical_descrip	occasion	notes	call_number
1.	12463	HOTEL EASTMAN	BREAKFAST		COMMERCIAL	HOT SPRINGS, AR	CARD	EASTER;		1900-2822
2.								4.75X7.5		
3.	12464	REPUBLICAN HOUSE	[DINNER]		COMMERCIAL	MILWAUKEE, [WI];	CARD	EASTER;	WEDGEWOOD BLUE CARD; WHITE EMBOSSED GREEK KEY BORDER; "EASTER SUNDAY" EMBOSSED IN WHITE; VIOLET COLORED SPRAY OF FLOWERS IN UPPER LEFT CORNER;	1900-2825
4.								ILLUS		
5.	12465	NORDDEUTSCHER LLOYD BREMEN	FRUHSTUCK/BREAKFAST;		COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE;	CARD		MENU IN GERMAN AND ENGLISH; ILLUS, STEAMSHIP AND SAILING VESSEL;	1900-2827
								ILLU		

Advanced Transformations

Using facets and filters

Facet / Filter Undo / Redo 1

17079 rows

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

Extensions: Freebase ▾

All id name sponsor event venue place physical_describ occasion no

1. 12463 REAKFAST COMMERCIAL HOT SPRINGS, AR CARD; 4.75X7.5; EASTER; WEDS

2. 12464 INNER] COMMERCIAL MILWAUKEE, WI CARD; ILLUS.COM EASTER; WEDS

3. 12465 NORDDEUTSCHER FRUHSTUCK/BREAKFAST; COMMERCIAL DA WIL GR

Edit cells ► Transform...
Edit column ► Common transforms
Transpose ► Fill down
Sort... Blank down
View Split multi-valued cells...
Reconcile ► Join multi-valued cells...
Cluster and edit...

Custom text transform on column sponsor

Expression Language Google Refine Expression Language (GREL)

value No syntax error.

Preview History Started Help

row	value	value
1.	HOTEL EASTMAN	HOTEL EASTMAN
2.	REPUBLICAN HOUSE	REPUBLICAN HOUSE
3.	NORDDEUTSCHER LLOYD BREMEN	NORDDEUTSCHER LLOYD BREMEN
4.	NORDDEUTSCHER LLOYD BREMEN	NORDDEUTSCHER LLOYD BREMEN
5.	NORDDEUTSCHER LLOYD BREMEN	NORDDEUTSCHER LLOYD BREMEN
6.	CANADIAN PACIFIC RAILWAY COMPANY	CANADIAN PACIFIC RAILWAY COMPANY
7.	HOTEL NETHERLAND	HOTEL NETHERLAND

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

OK Cancel

Understanding Expressions:

<https://github.com/OpenRefine/OpenRefine/wiki/Understanding-Expressions>

GREL Cheat Sheet:

<http://arcadiafalcone.net/GoogleRefineCheatSheets.pdf>

Extracting History as Provenance

Google refine menuprep Permalink Open... Export Help

Facet / Filter Undo / Redo 29 Extract... Apply...

17079 rows Show as: rows records Show: 5 10 25 50 rows Extensions: Firebase RDF

Filter:

0. Create project

1. Text transform on 9 cells in column name: value.trim()

2. Text transform on 14 cells in column sponsor: value.trim()

3. Text transform on 3 cells in column event: value.trim()

4. Text transform on 0 cells in column venue: value.trim()

5. Text transform on 8 cells in column place: value.trim()

6. Text transform on 385 cells in column physical_description: value.trim()

7. Text transform on 0 cells in column occasion: value.trim()

8. Text transform on 125 cells in column notes: value.trim()

9. Text transform on 10 cells in column call_number: value.trim()

10. Text transform on 0 cells in column keywords: value.trim()

11. Text transform on 0 cells in column language: value.trim()

1. 12463 Hotel Eastman breakfast commercial Hot Springs, Ar card; 4.75x7.5; Easter; 1900-2822

2. 12464 Republican House [dinner] commercial Milwaukee, [wi]; card; illus; col; 7.0x9.0; Easter; wedgewood blue card; white embossed greek key border; "easter sunday" embossed in white; violet colored spray of flowers in upper left corner; 1900-2825

3. 12465 Norddeutscher Lloyd Bremen fruhstück/breakfast; commercial Damper Kaiser Wilhelm Der Grosse; card; illu; col; 5.5x8.0; menu in german and english; illus, steamship and sailing vessel; 1900-2827

4. 12466 Norddeutscher Lloyd Bremen lunch; commercial Damper Kaiser Wilhelm Der Grosse; card; illu; col; 5.5x8.0; menu in german and english; illus, harbor scene with sailing vessel; 1900-2828

5. 12467 Norddeutscher Lloyd Bremen dinner; commercial Damper Kaiser Wilhelm Der Grosse; folder; illu; col; 5.5x7.5; menu in german and english; illus, harbor scene with rocks and 1900-2829

Extracting History as Provenance

luprep Permalink

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.

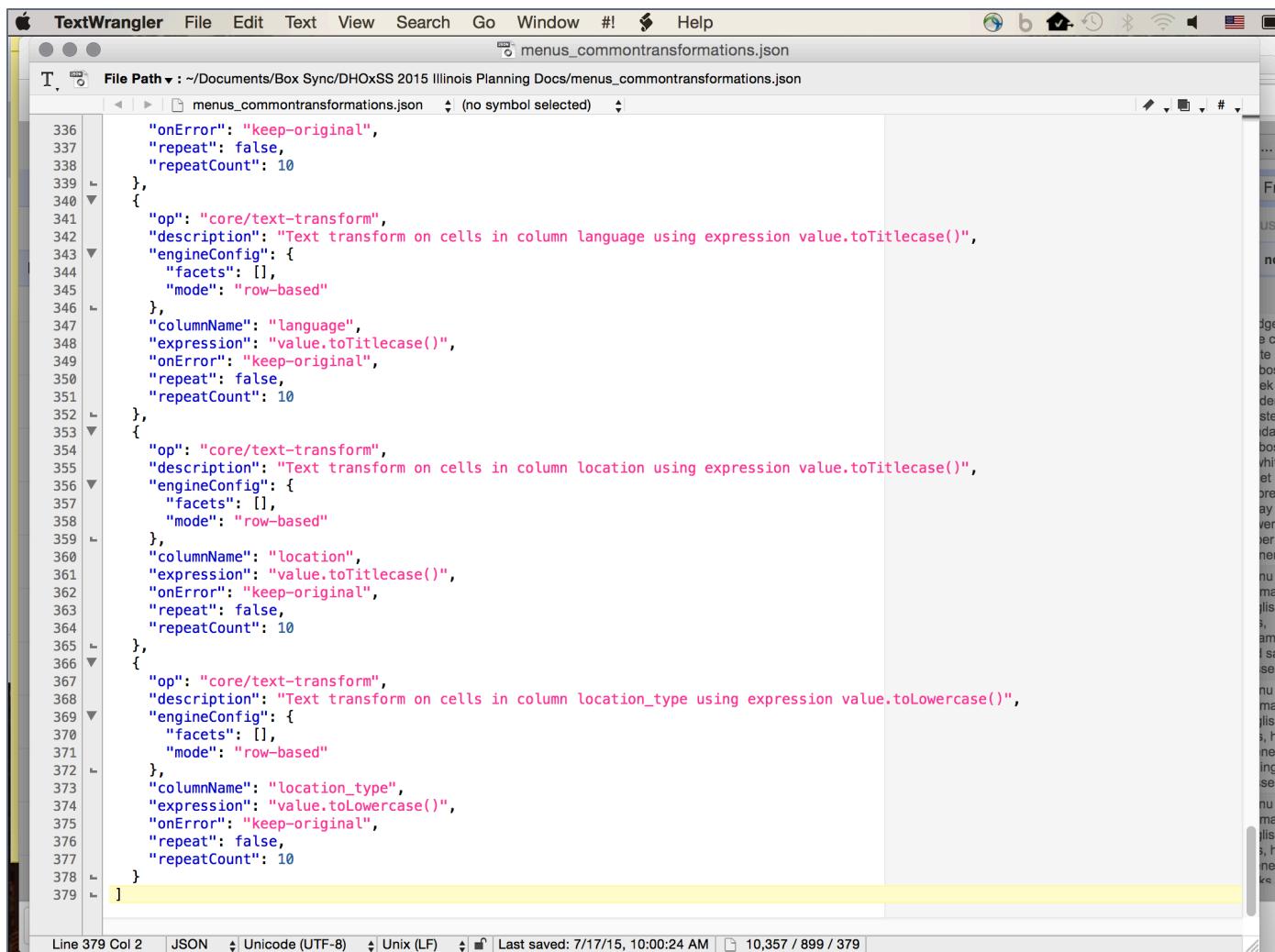
- Text transform on cells in column name using expression value.trim()
- Text transform on cells in column sponsor using expression value.trim()
- Text transform on cells in column event using expression value.trim()
- Text transform on cells in column venue using expression value.trim()
- Text transform on cells in column place using expression value.trim()
- Text transform on cells in column physical_description using expression value.trim()
- Text transform on cells in column occasion using expression value.trim()
- Text transform on cells in column notes using expression value.trim()
- Text transform on cells in column call_number using expression value.trim()
- Text transform on cells in column keywords using expression value.trim()
- Text transform on cells in column language using expression value.trim()

Select All Unselect All

Close

```
[{"op": "core/text-transform", "description": "Text transform on cells in column name using expression value.trim()", "engineConfig": {"facets": [], "mode": "row-based"}, "columnName": "name", "expression": "value.trim()", "onError": "keep-original", "repeat": false, "repeatCount": 10}, {"op": "core/text-transform", "description": "Text transform on cells in column sponsor using expression value.trim()", "engineConfig": {"facets": [], "mode": "row-based"}, "columnName": "sponsor", "expression": "value.trim()", "onError": "keep-original", "repeat": false, "repeatCount": 10}, {"op": "core/text-transform", "description": "Text transform on cells in column event using expression value.trim()", "engineConfig": {"facets": [], "mode": "row-based"}, "columnName": "event", "expression": "value.trim()", "onError": "keep-original", "repeat": false, "repeatCount": 10}, {"op": "core/text-transform", "description": "Text transform on cells in column venue using expression value.trim()", "engineConfig": {"facets": [], "mode": "row-based"}, "columnName": "venue", "expression": "value.trim()", "onError": "keep-original", "repeat": false, "repeatCount": 10}, {"op": "core/text-transform", "description": "Text transform on cells in column place using expression value.trim()", "engineConfig": {"facets": [], "mode": "row-based"}, "columnName": "place", "expression": "value.trim()", "onError": "keep-original", "repeat": false, "repeatCount": 10}, {"op": "core/text-transform", "description": "Text transform on cells in column physical_description using expression value.trim()", "engineConfig": {"facets": [], "mode": "row-based"}, "columnName": "physical_description", "expression": "value.trim()", "onError": "keep-original", "repeat": false, "repeatCount": 10}, {"op": "core/text-transform", "description": "Text transform on cells in column occasion using expression value.trim()", "engineConfig": {"facets": [], "mode": "row-based"}, "columnName": "occasion", "expression": "value.trim()", "onError": "keep-original", "repeat": false, "repeatCount": 10}, {"op": "core/text-transform", "description": "Text transform on cells in column notes using expression value.trim()", "engineConfig": {"facets": [], "mode": "row-based"}, "columnName": "notes", "expression": "value.trim()", "onError": "keep-original", "repeat": false, "repeatCount": 10}, {"op": "core/text-transform", "description": "Text transform on cells in column call_number using expression value.trim()", "engineConfig": {"facets": [], "mode": "row-based"}, "columnName": "call_number", "expression": "value.trim()", "onError": "keep-original", "repeat": false, "repeatCount": 10}, {"op": "core/text-transform", "description": "Text transform on cells in column keywords using expression value.trim()", "engineConfig": {"facets": [], "mode": "row-based"}, "columnName": "keywords", "expression": "value.trim()", "onError": "keep-original", "repeat": false, "repeatCount": 10}, {"op": "core/text-transform", "description": "Text transform on cells in column language using expression value.trim()", "engineConfig": {"facets": [], "mode": "row-based"}, "columnName": "language", "expression": "value.trim()", "onError": "keep-original", "repeat": false, "repeatCount": 10}]
```

Extracting History as Provenance



A screenshot of the TextWrangler application window. The title bar reads "TextWrangler" and the file path is "File Path : ~/Documents/Box Sync/DHOxSS 2015 Illinois Planning Docs/menus_commontransformations.json". The main editor area displays a JSON object with several nested objects and arrays. The code is color-coded, with numbers on the left indicating line numbers. The JSON structure is as follows:

```
336     "onError": "keep-original",
337     "repeat": false,
338     "repeatCount": 10
339   },
340   {
341     "op": "core/text-transform",
342     "description": "Text transform on cells in column language using expression value.toTitlecase()",
343     "engineConfig": {
344       "facets": [],
345       "mode": "row-based"
346     },
347     "columnName": "language",
348     "expression": "value.toTitlecase()",
349     "onError": "keep-original",
350     "repeat": false,
351     "repeatCount": 10
352   },
353   {
354     "op": "core/text-transform",
355     "description": "Text transform on cells in column location using expression value.toTitlecase()",
356     "engineConfig": {
357       "facets": [],
358       "mode": "row-based"
359     },
360     "columnName": "location",
361     "expression": "value.toTitlecase()",
362     "onError": "keep-original",
363     "repeat": false,
364     "repeatCount": 10
365   },
366   {
367     "op": "core/text-transform",
368     "description": "Text transform on cells in column location_type using expression value.toLowerCase()",
369     "engineConfig": {
370       "facets": [],
371       "mode": "row-based"
372     },
373     "columnName": "location_type",
374     "expression": "value.toLowerCase()",
375     "onError": "keep-original",
376     "repeat": false,
377     "repeatCount": 10
378   }
379 ]
```

The status bar at the bottom shows "Line 379 Col 2" and "Last saved: 7/17/15, 10:00:24 AM".

Applying JSON to New Project

Google refine Menu csv Permalink

Facet / Filter Undo / Redo 0

17079 rows

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

Infinite undo history 

Don't worry about making mistakes. Every change you make will be shown here, and you can undo your changes anytime.

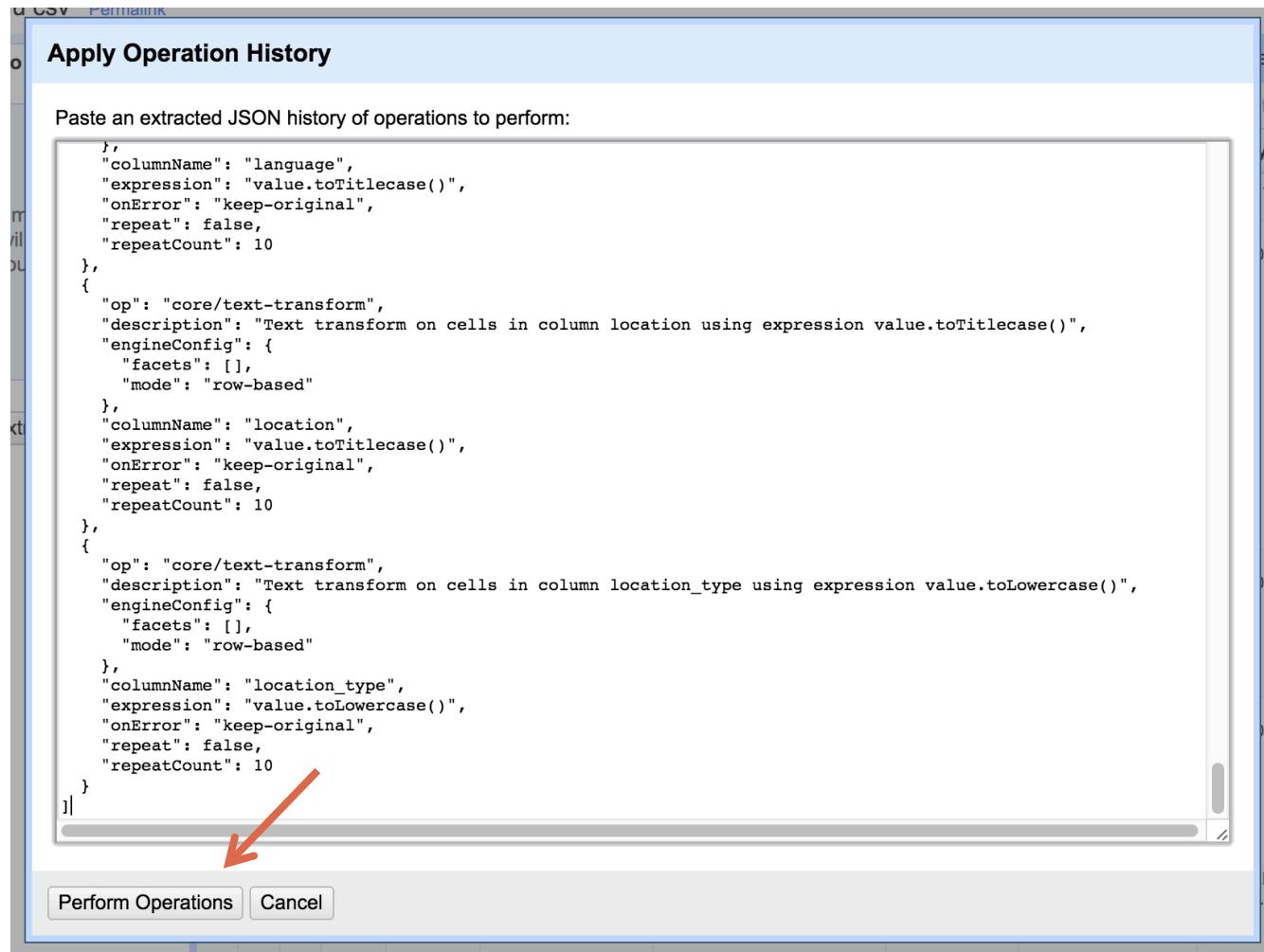
[Learn more »](#)

Extract... Apply... 

	All	id	name	sponsor	event	venue	place	physical_descrip	occasion	not
1.	12463		HOTEL EASTMAN	BREAKFAST	COMMERCIAL	HOT SPRINGS, AR	CARD; 4.75X7.5;	EASTER;		WED BLUI WHI EMB GRE BOR "EAS SUN EMB WHI VIOL COL SPR FLO UPP COR
2.	12464		REPUBLICAN HOUSE	[DINNER]	COMMERCIAL	MILWAUKEE, [WI];	CARD; ILLUS; COL; 7.0X9.0;	EASTER;		
3.	12465		NORDDEUTSCHER LLOYD BREMEN	FRUHSTUCK/BREAKFAST;	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE;	CARD; ILLU; COL; 5.5X8.0;			MEN GER ENG ILLU STE AND VES
4.	12466		NORDDEUTSCHER LLOYD BREMEN	LUNCH;	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE;	CARD; ILLU; COL; 5.5X8.0;			MEN GER ENG ILLU HAR SCE SAIL VES
5.	12467		NORDDEUTSCHER LLOYD BREMEN	DINNER;	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE;	FOLDER; ILLU; COL; 5.5X7.5;			MEN GER ENG ILLU HAR

127.0.0.1:3333/project?project=1777687896693#refine-tabs-history

Applying JSON to New Project



Applying JSON to New Project

17079 rows											Extensions: Freebase RDF			
Facet / Filter		Undo / Redo 29		Show as: rows records Show: 5 10 25 50 rows									« first < previous 1 - 10 next > last »	
				Extract...	Apply...									
Filter:		All	id	name	sponsor	event	venue	place	physical_descri	occasion	notes	call_number		
19.	Text transform on 8622 cells in column sponsor: value.toTitlecase()			1. 12463	Hotel Eastman	breakfast	commercial	Hot Springs, Ar	card; 4.75x7.5;	Easter;			1900-2822	
20.	Text transform on 7777 cells in column event: value.toLowerCase()			2. 12464	Republican House	[dinner]	commercial	Milwaukee, [wi];	card; illus; col; 7.0x9.0;	Easter;	wedgewood blue card; white embossed greek key border; "easter sunday" embossed in white; violet colored spray of flowers in upper left corner;		1900-2825	
21.	Text transform on 8110 cells in column venue: value.toLowerCase()													
22.	Text transform on 7337 cells in column place: value.toTitlecase()													
23.	Text transform on 8095 cells in column physical_description: value.toLowerCase()			3. 12465	Norddeutscher Lloyd Bremen	fruhstück/breakfast;	commercial	Dampfer Kaiser Wilhelm Der Grosse;	card; illu; col; 5.5x8.0;		menu in german and english; illus;		1900-2827	
24.	Text transform on 3752 cells in column occasion: value.toTitlecase()			4. 12466	Norddeutscher Lloyd Bremen	lunch;	commercial	Dampfer Kaiser Wilhelm Der Grosse;	card; illu; col; 5.5x8.0;		steamship and sailing vessel;		1900-2828	
25.	Text transform on 9755 cells in column notes: value.toLowerCase()													
26.	Text transform on 0 cells in column keywords: value.toLowerCase()													
27.	Text transform on 0 cells in column language: value.toTitlecase()													
28.	Text transform on 1220 cells in column location: value.toTitlecase()			5. 12467	Norddeutscher Lloyd Bremen	dinner;	commercial	Dampfer Kaiser Wilhelm Der Grosse;	folder; illu; col; 5.5x7.5;		menu in german and english; illus,		1900-2829	
29.	Text transform on 0 cells in column location_type: value.toLowerCase()										harbor scene with rocks and			

Exporting your Project

Google refine menuprep Permalink

Facet / Filter Undo / Redo 29

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

17079 rows

Show as: [rows](#) [records](#) Show: [5](#) [10](#) [25](#) [50](#) rows

			All	id	name	sponsor	event	venue	place	phys		
1.	12463				Hotel Eastman	breakfast		commercial	Hot Springs, Ar	card; illu;	menu in german and english; illus, steamship and sailing vessel;	1900-2827
2.	12464				Republican House	[dinner]		commercial	Milwaukee, [wi];	card; illu; col; 7.0x9.0;	menu in german and english; illus, harbor scene with sailing vessel;	1900-2828
3.	12465				Norddeutscher Lloyd Bremen	fruhstück/breakfast;		commercial	Dampfer Kaiser Wilhelm Der Grosse;	card; illu; col; 5.5x8.0;	menu in german and english; illus, steamship and sailing vessel;	1900-2829
4.	12466				Norddeutscher Lloyd Bremen	lunch;		commercial	Dampfer Kaiser Wilhelm Der Grosse;	card; illu; col; 5.5x8.0;	menu in german and english; illus, harbor scene with sailing vessel;	1900-2828
5.	12467				Norddeutscher Lloyd Bremen	dinner;		commercial	Dampfer Kaiser Wilhelm Der Grosse;	folder; illu; col; 5.5x7.5;	menu in german and english; illus, harbor scene with rocks and lighthouse; steamship and sailing	1900-2829

Export project
Tab-separated value
Comma-separated value
HTML table
Excel
ODF spreadsheet
Triple loader
MQLWrite
Custom tabular exporter...
Templating...
RDF as RDF/XML
RDF as Turtle

Further Resources

- *Main project page* (You can download the program here):
<http://openrefine.org/>
- *Github repository* (This is where the source code lives): <https://github.com/OpenRefine/OpenRefine>
- *Open Refine Documentation (AKA an extensive user manual)*: <https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users>
- *Additional Reconciliation Services*:
 - There are several listed here. Nomenklatura & Reconcile-csv may be of particular interest:
<https://github.com/OpenRefine/OpenRefine/wiki/Reconcilable-Data-Sources>
- *More on GREL*:
<https://github.com/OpenRefine/OpenRefine/wiki/GREL-String-Functions>

GEOCODING

Using OpenRefine

(based on a tutorial by Ahmad Assaf:

<http://ahmadassaf.com/blog/data-analysis/cleaning-geo-data-open-refine/>)

What's that?

- To **georeference** means to associate something with locations in physical space.
- **Geocoding**, or using a description of a location (e.g. a place name, address), to find geographic coordinates, is a kind of georeferencing

In other words, plotting strings like...

“10 feet west of the Thames”
“a mile north of Ann Arbor, Michigan”
“400 W. Elm St, Champaign, IL”

...on a map, to obtain more *precise* geographic coordinates.

Why do we need it?

- Converting geolocations from one format to another is notoriously difficult
 - (Which is why there are entire field of study concerned with geographic information systems)
- Having easily mappable (lat/long or UTM) coordinates makes informatics projects possible
 - For instance:
 - Pelagios, a digital map of the roman empire:
<http://pelagios-project.blogspot.com/2012/09/a-digital-map-of-roman-empire.html>
 - Google Ancient Places: <https://googleancientplaces.wordpress.com/>
 - Mapping the Republic of Letters: <http://republicofletters.stanford.edu/>

How can OpenRefine help?

- Converting coordinates can require some complicated wrangling of math, maps and datums
- Other services already do this well (Google!)
- With OpenRefine, we can call on those services to hugely simplify our workflows

Create a Column for New URLs

Google refine Menu csv Permalink

Facet / Filter Undo / Redo 26

Refresh Reset All Remove All

Starred Rows change invert reset

2 choices Sort by: name count
false 17069
true 10 exclude

Facet by choice counts

10 matching rows (17079 total)

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

	All		Id	name	sponsor	event	venue	place	physical_descri	occasion	notes	call_r							
91.	12567			STATLER'S				LLOCOTT QUARE, UFFALO, [NY];	CARD; ILLUS; 3.25X6.25;			ILLUS: WOMAN HOLDING TEACUP AND SAUCER;							
92.	12568			JOHN WANAMAKER		Edit column			Split into several columns...				ILLUS: SHIELD WITH FLEUR DE LYS, FLANKED BY LOBSTERS; SURMOUNTED BY PIGEON;						
93.	12569			MAXWELL HOUSE		Transpose			Add column based on this column...				ELABORATE PRINTING OF HOTEL NAME FORMS ILLUSTRATION;						
94.	12570			MAXWELL HOUSE	dinner Choose new match	COMMERCIAL			Sort...				Rename this column						
95.	12571			CLYDE STEAMSHIP CO., THE	dinner Choose new match	COMMERCIAL			View				Remove this column						
96.	12573			MAXWELL HOUSE	breakfast Choose new match	COMMERCIAL	NASHVILLE, TN	CARD; ILLUS; 4.25X6.25;					Move column to beginning						
													Move column to end						
													Move column left						
													Move column right						
													INCLUDES PRICED WINE, LIQUOR, ETC. LIST; STEAMSHIP INSIGNIA ON COVER; ON BACK COVER IS A MAP COMPRISED ALL LINES INVOLVED WITH CLYDE;						
													ELABORATE PRINTING OF HOTEL NAME FORMS						
													ELABORATE PRINTING OF HOTEL NAME FORMS						

Wait a second....

Google refine nrhp_ilinois_links.xlsx Permalink

Facet / Filter Undo / Redo 0 1809 rows Extensions: Named-entity recognition ▾ Freebase ▾ RDF ▾

Extract... Apply... Filter: Photo

« first < previous 1 - 10 next > last »

0. Create project

1. Create new column Composite address based on column Address by filling 1809 rows with gREL:value + ";" + cells["State"].value

Add column based on column Address

New column name: AddressURL

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

Expression: `"http://maps.googleapis.com/maps/api/geocode/json?sensor=false&address="+escape(value,"url")+", "+cells["City"].value+", "+cells["State"].value`

Language: Google Refine Expression Language (GREL) No syntax error.

Preview History Starred Help

row	value	AddressURL
1.	616 N. 24th St.	<code>"http://maps.googleapis.com/maps/api/geocode/json?sensor=false&address=616+N.+24th+St.+++++Quincy , ILLINOIS"</code>
2.	Roughly bounded by Hampshire, Jersey, 4th and 8th Sts.	<code>"http://maps.googleapis.com/maps/api/geocode/json?sensor=false&address=Roughly+bounded+by+Hampshire%2C+Jersey%2C+4th-Quincy , ILLINOIS"</code>
3.	NW of Golden	<code>"http://maps.googleapis.com/maps/api/geocode/json?sensor=false&address=NW+of+Golden"</code>

OK Cancel

10. 77000471 ILLINOIS Adams Quincy Morgan-Wells House 421 Jersey St. 19771116 http://pdfhost.focus.nps.gov/docs/nrhp/text/77000471.PDF

Remember: clean your data!

Add column based on column Address

New column name: AddressURL

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

Expression: `"http://maps.googleapis.com/maps/api/geocode/json?sensor=false&address=" + escape(value, "url") + ", " + cells["City"].value + ", " + cells["State"].value`

Language: Google Refine Expression Language (GREL)

No syntax error.

Preview:

row	value
1.	616 N. 24th St.
2.	Roughly bounded by Hampshire, Jersey, 4th and 8th Sts.
3.	NW of Golden

History Starred Help

OK Cancel

0. Create project
1. Text transform on 1809 cells in column Address: value.trim()
2. Text transform on 2 cells in column Address: value.replace(/\s+/, ')
3. Text transform on 1809 cells in column City: value.trim()
4. Text transform on 0 cells in column City: value.replace(/\s+/, ')
5. Text transform on 1809 cells in column County: value.trim()
6. Text transform on 0 cells in column County: value.replace(/\s+/, ')
7. Text transform on 1809 cells in column State: value.trim()
8. Text transform on 0 cells in column State: value.replace(/\s+/, ')

Fetching URLs

Add column by fetching URLs based on column AddressURLs

New column name Throttle delay milliseconds

On error set to blank store error

Formulate the URLs to fetch:

Expression Language Google Refine Expression Language (GREL) ▼

No syntax error.

Preview History Starred Help

row	value	value
1.	http://maps.googleapis.com/maps/api/geocode/json	http://maps.googleapis.com/maps/api/geocode/json
	sensor=false&address=616+N.+24th+St.,+Quincy,+IL	sensor=false&address=616+N.+24th+St.,+Quincy,+IL
2.	http://maps.googleapis.com/maps/api/geocode/json	http://maps.googleapis.com/maps/api/geocode/json
	sensor=false&address=Roughly+bounded+by+Harr	sensor=false&address=Roughly+bounded+by+Harr
3.	http://maps.googleapis.com/maps/api/geocode/json	http://maps.googleapis.com/maps/api/geocode/json
	sensor=false&address=NW+of+Golden,+Golden,+IL	sensor=false&address=NW+of+Golden,+Golden,+IL
4.	http://maps.googleapis.com/maps/api/geocode/json	
	sensor=false&address=Quincy+St.,+Golden,+IL	

Create column JSON at index 7 by fetching URLs based on column AddressURLs using expression grel:value
15% complete [Cancel](#)

OK Cancel

Results: JSON!

Google refine nrhp_ilinois_links.xlsx [Permalink](#)

Facet / Filter Undo / Redo 10 Extract... Apply...

1809 rows

Show as: **rows** records Show: 5 10 25 50 rows Extensions: [Named-entity recognition](#) [Freebase](#) [RDF](#)

« first < previous 1 - 50 next > last »

	JSON	Listed	Text	Ph
0. Create project	{ "results": [{ "address_components": [{ "long_name": "616", "short_name": "616", "types": ["street_number"] }, { "long_name": "North 24th Street", "short_name": "N 24th St", "types": ["route"] }, { "long_name": "Quincy", "short_name": "Quincy", "types": ["locality", "political"] }, { "long_name": "Quincy", "short_name": "Quincy", "types": ["administrative_area_level_3", "political"] }, { "long_name": "Adams County", "short_name": "Adams County", "types": ["administrative_area_level_2", "political"] }, { "long_name": "Illinois", "short_name": "IL", "types": ["administrative_area_level_1", "political"] }, { "long_name": "United States", "short_name": "US", "types": ["country", "political"] }, { "long_name": "62301", "short_name": "62301", "types": ["postal_code"] }, { "long_name": "3248", "short_name": "3248", "types": ["postal_code_suffix"] }], "formatted_address": "616 North 24th Street, Quincy, IL 62301, USA", "geometry": { "location": { "lat": 39.938546, "lng": -91.377194 }, "location_type": "ROOFTOP", "viewport": { "northeast": { "lat": 39.9398949802915, "lng": -91.37584501970849 }, "southwest": { "lat": 39.9371970197085, "lng": -91.3785429802915 } }, "place_id": "ChIJSSgbFzv33YcR0iuO-aTJ80U", "types": ["street_address"] }, "status": "OK" }	19970207	http://pdfhost.focus.nps.gov/docs/nrhp/text/97000032.PDF	http://p
1. Text transform on 1809 cells in column Address: value.trim()	{ "results": [{ "address_components": [{ "long_name": "South 8th Street", "short_name": "S 8th St", "types": ["route"] }, { "long_name": "Quincy", "short_name": "Quincy", "types": ["locality", "political"] }, { "long_name": "Adams County", "short_name": "Adams County", "types": ["administrative_area_level_2", "political"] }, { "long_name": "Illinois", "short_name": "IL", "types": ["administrative_area_level_1", "political"] }, { "long_name": "United States", "short_name": "US", "types": ["country", "political"] }, { "formatted_address": "South 8th Street, Quincy, IL, USA", "geometry": { "bounds": { "northeast": { "lat": 39.9327454, "lng": -91.40124999999999 }, "southwest": { "lat": 39.8939459, "lng": -91.408062 } }, "location": { "lat": 39.9137416, "lng": -91.40324339999999 }, "location_type": "GEOMETRIC_CENTER", "viewport": { "northeast": { "lat": 39.9327454, "lng": -91.40124999999999 }, "southwest": { "lat": 39.8939459, "lng": -91.408062 } }, "partial_match": true, "place_id": "ChIJlw5M2x_33YcR-uLvUje7fY", "types": ["route"] }, "address_components": [{ "long_name": "North 8th Street", "short_name": "N 8th St", "types": ["route"] }, { "long_name": "Quincy", "short_name": "Quincy", "types": ["locality", "political"] }, { "long_name": "Quincy", "short_name": "Quincy", "types": ["administrative_area_level_3", "political"] }, { "long_name": "Adams County", "short_name": "Adams County", "types": ["administrative_area_level_2", "political"] }, { "long_name": "Illinois", "short_name": "IL", "types": ["administrative_area_level_1", "political"] }, { "long_name": "United States", "short_name": "US", "types": ["country", "political"] }, { "formatted_address": "North 8th Street, Quincy, IL 62301, USA", "geometry": { "bounds": { "northeast": { "lat": 39.9504755, "lng": -91.402586 }, "southwest": { "lat": 39.93188079999999, "lng": -91.404071 } }, "location": { "lat": 39.940635, "lng": -91.40336259999999 }, "location_type": "GEOMETRIC_CENTER", "viewport": { "northeast": { "lat": 39.9504755, "lng": -91.4019795197085 }, "southwest": { "lat": 39.93188079999999, "lng": -91.40467748029151 } }, "partial_match": true, "place_id": "ChIJ5dbakK33YcRJ6rVmUPI_ag", "types": ["route"] }, "status": "OK" }	19830407	http://pdfhost.focus.nps.gov/docs/nrhp/text/83000298.PDF	http://p
2. Text transform on 2 cells in column Address: value.replace(/\s+/,'')	{ "results": [{ "address_components": [{ "long_name": "Village of Golden", "short_name": "Village of Golden", "types": ["locality", "political"] }, { "long_name": "Illinois", "short_name": "IL", "types": ["administrative_area_level_1", "political"] }, { "long_name": "United States", "short_name": "US", "types": ["country", "political"] }, { "formatted_address": "Village of Golden, IL, USA", "geometry": { "bounds": { "northeast": { "lat": 39.9504755, "lng": -91.402586 }, "southwest": { "lat": 39.93188079999999, "lng": -91.404071 } }, "location": { "lat": 39.940635, "lng": -91.40336259999999 }, "location_type": "GEOMETRIC_CENTER", "viewport": { "northeast": { "lat": 39.9504755, "lng": -91.4019795197085 }, "southwest": { "lat": 39.93188079999999, "lng": -91.40467748029151 } }, "partial_match": true, "place_id": "ChIJ5dbakK33YcRJ6rVmUPI_ag", "types": ["route"] }, "status": "OK" }	19840604	http://pdfhost.focus.nps.gov/docs/nrhp/text/84000921.PDF	http://p
3. Text transform on 1809 cells in column City: value.trim()				
4. Text transform on 0 cells in column City: value.replace(/\s+/,'')				
5. Text transform on 1809 cells in column County: value.trim()				
6. Text transform on 0 cells in column County: value.replace(/\s+/,'')				
7. Text transform on 1809 cells in column State: value.trim()				
8. Text transform on 0 cells in column State: value.replace(/\s+/,'')				
9. Create new column AddressURLs based on column Address by filling 1809 rows with grel:"http://maps.googleapis.com/maps/api/geocoder?sensor=false&address="+escape(value,"url")+				
10. Create column JSON at index 7 by fetching URLs based on column AddressURLs using expression grel:value				

Parsing JSON

Google refine nrhp_ilinois_links.xlsx [Permalink](#) [Open...](#) [Export](#) [Help](#)

Facet / Filter Undo / Redo 10 Extract... Apply...

1809 rows Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 50 next > last »

Filter:

	JSON	Listed	Text	Print
0. Create project				
1. Text transform on 1809 cells in column Address: value.trim()				
2. Text transform on 2 cells in column Address: value.replace(/\s+/,'')				
3. Text transform on 1809 cells in column City: value.trim()				
4. Text transform on 0 cells in column City: value.replace(/\s+/,'')				
5. Text transform on 1809 cells in column County: value.trim()				
6. Text transform on 0 cells in column County: value.replace(/\s+/,'')				
7. Text transform on 1809 cells in column State: value.trim()				
8. Text transform on 0 cells in column State: value.replace(/\s+/,'')				
9. Create new column AddressURLs based on column Address by filling 1809 rows with grel:"http://maps.googleapis.com/maps/api/geocoder?sensor=false&address="+escape(value,"url")				
10. Create column JSON at index 7 by fetching URLs based on column AddressURLs using expression grel:value	<ul style="list-style-type: none">FacetText filterEdit cellsEdit columnTransposeSort...ViewReconcileExtract named entities...	<ul style="list-style-type: none">▶ Add column based on this column...▶ Add column by fetching URLs...▶ Add columns from Freebase ...▶ Rename this column▶ Remove this column▶ Move column to beginning▶ Move column to end▶ Move column left▶ Move column right	<p>19970207 http://pdfhost.focus.nps.gov/docs/nrhp/text/9700032.PDF http://pdfhost.focus.nps.gov/docs/nrhp/text/9700032.PDF</p> <p>19830407 http://pdfhost.focus.nps.gov/docs/nrhp/text/83000298.PDF http://pdfhost.focus.nps.gov/docs/nrhp/text/83000298.PDF</p> <p>19840604 http://pdfhost.focus.nps.gov/docs/nrhp/text/84000921.PDF http://pdfhost.focus.nps.gov/docs/nrhp/text/84000921.PDF</p>	

Parsing JSON

Add column based on column JSON

New column name

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

Expression

```
value.parseJson().results[0].geometry.location.lng
```

No syntax error.

Preview History Starred Help

row	value	value.parseJson().results[0].geometry.location.lng
1.	{ "results" : [{ "address_components" : [{ "long_name" : "616", "short_name" : "616", "types" : ["street_number"] }, { "long_name" : "North 24th Street", "short_name" : "N 24th St", "types" : ["route"] }, { "long_name" : "Quincy", "short_name" : "Quincy", "types" : ["locality", "political"] }, { "long_name" : "Quincy", "short_name" : "Quincy", "types" : ["administrative_area_level_3", "political"] }, { "long_name" : "Adams County", "short_name" : "Adams County", "types" : ["administrative_area_level_2", "political"] }] }	-91.377194

OK Cancel

Or value.parseJson().results[0].geometry.location.lat
Or value.parseJson().results[0].formatted_address
Or...

We're ready to make maps!

Google refine nrhp_ilinois_links.xlsx [Permalink](#)

Facet / Filter Undo / Redo 13 Extract... Apply...

1809 rows Extensions: Named-entity recognition ▾ Freebase ▾ RDF ▾

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 50 next > last »

Filter:	JSON	Formatted Address	Latitude	Longitude	Listed	Text
4. Text transform on 0 cells in column City: value.replace(/\s+/,'')	{"results": [{"address_components": [{"long_name": "616", "short_name": "616", "types": ["street_number"]}, {"long_name": "North 24th Street", "short_name": "N 24th St", "types": ["route"]}], "long_name": "Quincy", "short_name": "Quincy", "types": ["locality", "political"]}, {"long_name": "Quincy", "short_name": "Quincy", "types": ["administrative_area_level_3", "political"]}, {"long_name": "Adams County", "short_name": "Adams County", "types": ["administrative_area_level_2", "political"]}, {"long_name": "Illinois", "short_name": "IL", "types": ["administrative_area_level_1", "political"]}, {"long_name": "United States", "short_name": "US", "types": ["country", "political"]}], "long_name": "62301", "short_name": "62301", "types": ["postal_code"]}, {"long_name": "3248", "short_name": "3248", "types": ["postal_code_suffix"]}], "formatted_address": "616 North 24th Street, Quincy, IL 62301, USA", "geometry": {"location": {"lat": 39.938546, "lng": -91.377194}, "location_type": "ROOFTOP", "viewport": {"northeast": {"lat": 39.939849802915, "lng": -91.37854501970849}, "southwest": {"lat": 39.9371970197085, "lng": -91.37854298029151}}, "place_id": "ChIJSSgbFZv33YcR0iuO-aTJ80U", "types": ["street_address"]}], "status": "OK"}	616 North 24th Street, Quincy, IL 62301, USA	39.938546	-91.377194	19970207	http://pdfhost.f...
5. Text transform on 1809 cells in column County: value.trim()	{"results": [{"address_components": [{"long_name": "South 8th Street", "short_name": "S 8th St", "types": ["route"]}], "long_name": "Quincy", "short_name": "Quincy", "types": ["locality", "political"]}, {"long_name": "Adams County", "short_name": "Adams County", "types": ["administrative_area_level_2", "political"]}, {"long_name": "Illinois", "short_name": "IL", "types": ["administrative_area_level_1", "political"]}, {"long_name": "United States", "short_name": "US", "types": ["country", "political"]}], "long_name": "South 8th Street, Quincy, IL, USA", "short_name": "South 8th Street, Quincy, IL, USA", "types": ["street_address"]}], "geometry": {"bounds": {"northeast": {"lat": 39.9327454, "lng": -91.4019249999999}, "southwest": {"lat": 39.8939459, "lng": -91.408062}}, "location": {"lat": 39.9137416, "lng": -91.4032433999999}, "location_type": "GEOMETRIC_CENTER", "viewport": {"northeast": {"lat": 39.9327454, "lng": -91.4019249999999}, "southwest": {"lat": 39.8939459, "lng": -91.408062}}}, "partial_match": true, "place_id": "ChIJw5M2x_33YcR-uJvUje7Y", "types": ["route"]}, {"address_components": [{"long_name": "North 8th Street", "short_name": "N 8th St", "types": ["route"]}], {"long_name": "Quincy", "short_name": "Quincy", "types": ["locality", "political"]}, {"long_name": "Quincy", "short_name": "Quincy", "types": ["administrative_area_level_3", "political"]}, {"long_name": "Adams County", "short_name": "Adams County", "types": ["administrative_area_level_2", "political"]}, {"long_name": "Illinois", "short_name": "IL", "types": ["administrative_area_level_1", "political"]}, {"long_name": "United States", "short_name": "US", "types": ["country", "political"]}], {"long_name": "62301", "short_name": "62301", "types": ["postal_code"]}], {"formatted_address": "North 8th Street, Quincy, IL 62301, USA", "geometry": {"bounds": {"northeast": {"lat": 39.9504755, "lng": -91.402586}, "southwest": {"lat": 39.9318807999999, "lng": -91.404071}}, "location": {"lat": 39.940635, "lng": -91.4033625999999}, "location_type": "GEOMETRIC_CENTER", "viewport": {"northeast": {"lat": 39.9504755, "lng": -91.4019795197085}, "southwest": {"lat": 39.9318807999999, "lng": -91.40467748029151}}}, "partial_match": true, "place_id": "ChIJ5dbak33YcRJ6rVmUPI_ag", "types": ["route"]}], "status": "OK"}	South 8th Street, Quincy, IL, USA	39.9137416	-91.4032434	19830407	http://pdfhost.f...
6. Text transform on 0 cells in column County: value.replace(/\s+/,'')	{"results": [{"address_components": [{"long_name": "Village of Golden", "short_name": "Village of Golden", "types": ["locality"]}], "long_name": "Village of Golden", "short_name": "Village of Golden", "types": ["political"]}], "long_name": "62301", "short_name": "62301", "types": ["postal_code"]}], {"formatted_address": "Village of Golden, IL 62301, USA", "geometry": {"bounds": {"northeast": {"lat": 40.10912, "lng": -91.01684}, "southwest": {"lat": 40.10912, "lng": -91.01684}}, "location": {"lat": 40.10912, "lng": -91.01684}, "location_type": "GEOMETRIC_CENTER", "viewport": {"northeast": {"lat": 40.10912, "lng": -91.01684}, "southwest": {"lat": 40.10912, "lng": -91.01684}}}, "partial_match": true, "place_id": "ChIJ5dbak33YcRJ6rVmUPI_ag", "types": ["route"]}], "status": "OK"}	Village of Golden,	40.10912	-91.01684	19840604	http://pdfhost.f...
7. Text transform on 1809 cells in column State: value.trim()						
8. Text transform on 0 cells in column State: value.replace(/\s+/,'')						
9. Create new column AddressURLs based on column Address by filling 1809 rows with grel:"http://maps.googleapis.com/maps/api/geocoder/json?sensor=false&address="+escape(value,"url")+						
10. Create column JSON at index 7 by fetching URLs based on column AddressURLs using expression grel:value						
11. Create new column Longitude based on column JSON by filling 798 rows with grel:value.parseJson().results[0].geometry.location.lng						
12. Create new column Latitude based on column JSON by filling 798 rows with grel:value.parseJson().results[0].geometry.location.lat						
13. Create new column Formatted Address based on column JSON by filling 798 rows with grel:value.parseJson().results[0].formatted_address						

FROM RESTAURANTS TO ROAD TRIPS

The National Register of Historic Places

Reconciling with Linked Data

What you'll need:

- A fairly clean data set
 - Navigate to http://bit.ly/DHOxSS_HumDat
 - Select nrhp_illinois_moundsites_links.xlsx in Monday > OpenRefine
- An RDF extension (<http://refine.deri.ie/>)
- One or more added reconciliation services
- A lot of patience

Create, Open, or Import a Project

Google refine A power tool for working with messy data.

Create Project « Start Over Configure Parsing Options Project name nrhp_illinois_links.xlsx Create Project »

Open Project Import Project

	Reference	State	County	City	Resource	Address	Listed	Text	Photo
1.	97000032	ILLINOIS	Adams	Quincy	Coca-Cola Bottling Company Building	616 N. 24th St.	19970207	http://pdfhost.focus.nps.gov/docs/nrhp/text/97000032.PDF	http://pdfhost.focus.nps.gov/docs/nrhp/
2.	83000298	ILLINOIS	Adams	Quincy	Downtown Quincy Historic District	Roughly bounded by Hampshire, Jersey, 4th and 8th Sts.	19830407	http://pdfhost.focus.nps.gov/docs/nrhp/text/83000298.PDF	http://pdfhost.focus.nps.gov/docs/nrhp/

Parse data as

Excel (.xlsx) files Update Preview

XML files

Open Document Format spreadsheets (.ods)

RDF/XML files

JSON files

Line-based text files

CSV / TSV / separator-based files

Fixed-width field text files

PC-Axis text files

Worksheets to Import
 Sheet1 1810 rows

Ignore first 0 line(s) at beginning of file Store blank rows
 Parse next 1 line(s) as column headers Store blank cells as nulls
 Discard 0 initial row(s) of data Store file source (file names, URLs) in each row
 Load at 0 most row(s) of data

Version 2.5 [r2407]

Help About

Add a Reconciliation Service

The screenshot shows the Google Refine interface with a modal dialog titled "Add SPARQL-based reconciliation service".

Name: dBpedia
A human readable name

Endpoint details

- Endpoint URL:
- Graph URI: Leave empty to use the default graph
- Type: Virtuoso
This determines the syntax that will be used for search

Label properties
Select properties that are used to label resources in the endpoint. These properties will be used to match resources:

- rdfs:label
- skos:prefLabel
- dcterms:title
- dc:title
- foaf:name
- Other...

Buttons: OK, Cancel

Background Data View:
nrhp_illinois_links Permalink
Facet / Filter Undo / Redo 1
Refresh Reset All Remove All
Starred Rows change invert reset
2 choices Sort by: name count
false 1780
true 29
Facet by choice counts
Extensions: Freebase ▾ RDF ▾
Text
http://pdfhost.focus.nps.gov/docs/nrhp/text/11000845
http://pdfhost.focus.nps.gov/docs/nrhp/text/90000752
http://pdfhost.focus.nps.gov/docs/nrhp/text/05001250
http://pdfhost.focus.nps.gov/docs/nrhp/text/97000460
http://pdfhost.focus.nps.gov/docs/nrhp/text/97001337
http://pdfhost.focus.nps.gov/docs/nrhp/text/89001108
http://pdfhost.focus.nps.gov/docs/nrhp/text/97001335
http://pdfhost.focus.nps.gov/docs/nrhp/text/97001336
http://pdfhost.focus.nps.gov/docs/nrhp/text/78001115
http://pdfhost.focus.nps.gov/docs/nrhp/text/75000642

DBpedia extracts structured information from Wikipedia and makes it available as a SPARQL endpoint.

Start with Subset of Data

The screenshot shows the Google Refine interface with the following details:

- Title Bar:** Google refine nrhp_ilinois_links Permalink
- Facet / Filter:** Shows "29 matching rows (1809 total)".
- Toolbar:** Open..., Export..., Help
- Extensions:** Freebase, RDF (highlighted with a red box).
- Facet by choice counts:** A red arrow points to this facet, indicating it's being used to select a subset of data.
- Table Headers:** All, Reference, State, County, City, Resource, Address, Listed, Text, Photo.
- Data Rows:** The first few rows are shown:

Facet	Reference	State	County	City	Resource	Address	Listed	Text	Photo	
Facet by star					Ahrens, Henry, House	212 E. University Ave.	20111122	http://pdfhost.focus.nps.gov/docs/nrhp/text/11000845.PDF	http://pdf...	
Facet by flag	100752	ILLINOIS	Champaign	Champaign	Alpha Delta Phi Fraternity House	310 E. John St.	19900521	http://pdfhost.focus.nps.gov/docs/nrhp/text/90000752.PDF	http://pdf...	
Facet by choice counts	001250	ILLINOIS	Champaign	Champaign	Alpha Phi Fraternity House--Beta Alpha Chapter	508 E. Amory Ave.	20051115	http://pdfhost.focus.nps.gov/docs/nrhp/text/05001250.PDF	http://pdf...	
	78.	97000460	ILLINOIS	Champaign	Champaign	Alpha Rho Chi Fraternity House	1108 S. First St.	19970523	http://pdfhost.focus.nps.gov/docs/nrhp/text/97000460.PDF	http://pdf...
	81.	97001337	ILLINOIS	Champaign	Champaign	Bailey--Rug Building	219-225 N. Neil St.	19971107	http://pdfhost.focus.nps.gov/docs/nrhp/text/97001337.PDF	http://pdf...
	82.	89001108	ILLINOIS	Champaign	Champaign	Beta Theta Pi Fraternity House	202 E. Daniel St.	19890828	http://pdfhost.focus.nps.gov/docs/nrhp/text/89001108.PDF	http://pdf...
	83.	97001335	ILLINOIS	Champaign	Champaign	Building at 201 North Market Street	201 N. Market St.	19971107	http://pdfhost.focus.nps.gov/docs/nrhp/text/97001335.PDF	http://pdf...
	84.	97001336	ILLINOIS	Champaign	Champaign	Building at 203-205 North Market Street	203-205 N. Market St.	19971107	http://pdfhost.focus.nps.gov/docs/nrhp/text/97001336.PDF	http://pdf...
	85.	78001115	ILLINOIS	Champaign	Champaign	Burnham Athenaeum	306 W. Church St.	19780607	http://pdfhost.focus.nps.gov/docs/nrhp/text/78001115.PDF	http://pdf...
	86.	75000642	ILLINOIS	Champaign	Champaign	Cattle Bank	102 E. University	19750819	http://pdfhost.focus.nps.gov/docs/nrhp/text/75000642.PDF	http://pdf...
- JavaScript Console:** javascript:{}

Star a small set of rows for an initial test set, facet by star, and select “true”.

Choose which Column to Reconcile

Google refine nrhp_illinois_links2 Permalink

Facet / Filter Undo / Redo 0

1809 rows

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

Extensions: [Freebase](#) [RDF](#)

« first < previous 1 - 10 next > last »

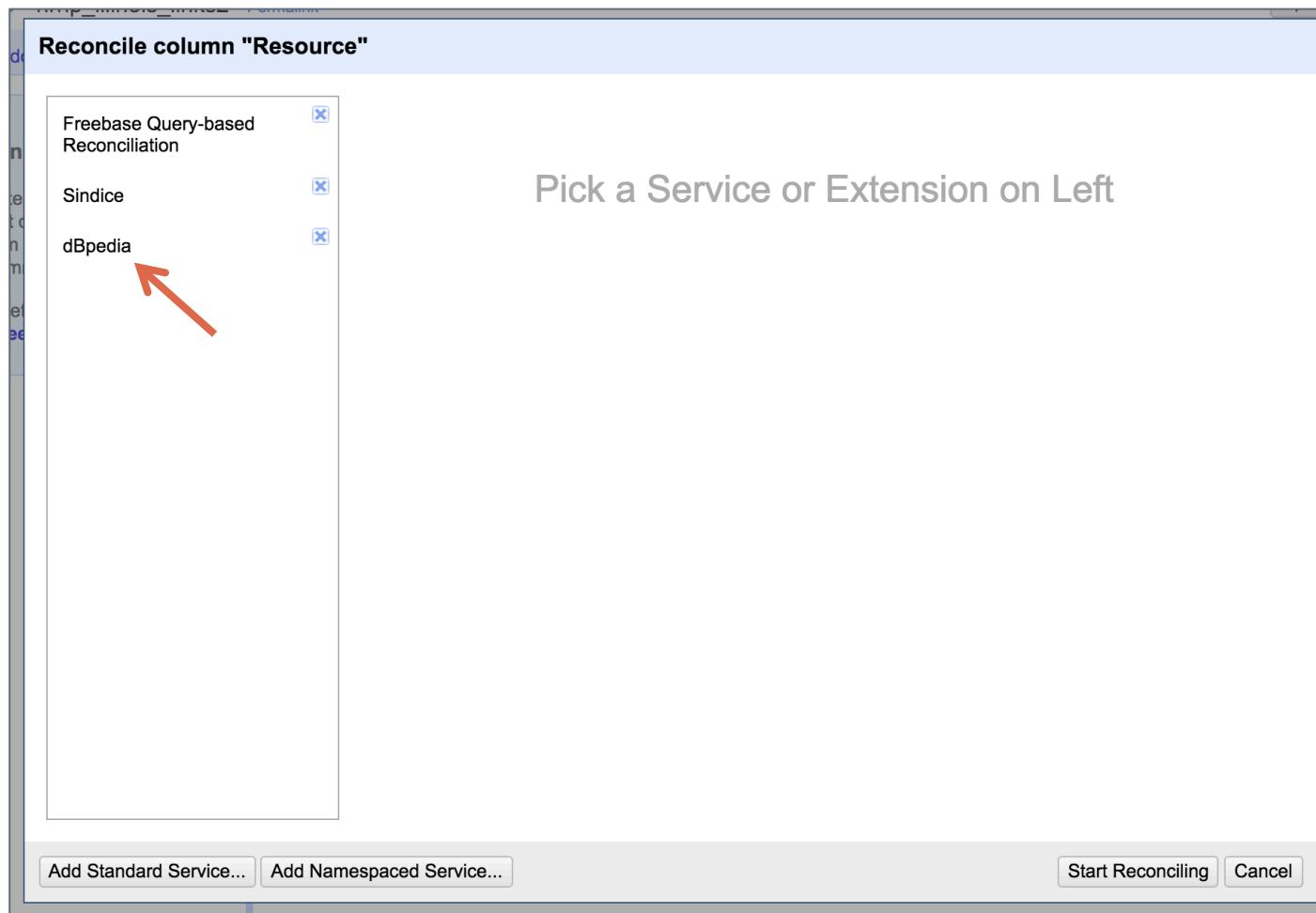
Using facets and filters 

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

1.	97000032	ILLINOIS	Adams	Quincy	Facet	4th	19970207	http://pdfhost.focus.nps.gov/docs/nrhp/text/97000032.PDF	ht										
2.	83000298	ILLINOIS	Adams	Quincy	Edit cells	by	19830407	http://pdfhost.focus.nps.gov/docs/nrhp/text/83000298.PDF	ht										
3.	84000921	ILLINOIS	Adams	Golden	Sort...		19840604	http://pdfhost.focus.nps.gov/docs/nrhp/text/84000921.PDF	ht										
4.	86003714	ILLINOIS	Adams	Golden	View														
5.	96001282	ILLINOIS	Adams	Payson	Reconcile	Start reconciling...													
6.	79000812	ILLINOIS	Adams	Quincy	Bank														
7.	99001377	ILLINOIS	Adams	Quincy	Fall Creek Stone Arch Bridge	1.2 mi. N Fall Cr. - Payson across I Cr.													
8.	2001750	ILLINOIS	Adams	Mendon	Gardner, Robert W., House	613 Broadway													
9.	4000181	ILLINOIS	Adams	Quincy	Lesem, S.J., Building	135-37 N 3rd St.	19991122	http://pdfhost.focus.nps.gov/docs/nrhp/text/99001377.PDF	ht										
10.	77000471	ILLINOIS	Adams	Quincy	Lewis Round Barn	2007 E 1250th St.	20030129	http://pdfhost.focus.nps.gov/docs/nrhp/text/02001750.PDF	ht										
					Lock and Dam No. 21 Historic District	0.5 mi. W of IL 57													
					Morgan-Wells House	421 Jersey St.	20040310	http://pdfhost.focus.nps.gov/docs/nrhp/text/04000181.PDF	ht										
							19771116	http://pdfhost.focus.nps.gov/docs/nrhp/text/77000471.PDF	ht										

Choose your Reconciliation Service



Choose a Type of Reconciliation

Reconcile column "Resource"

Freebase Query-based Reconciliation

Sindice

dBpedia

Reconcile each cell to an entity of one of these types:

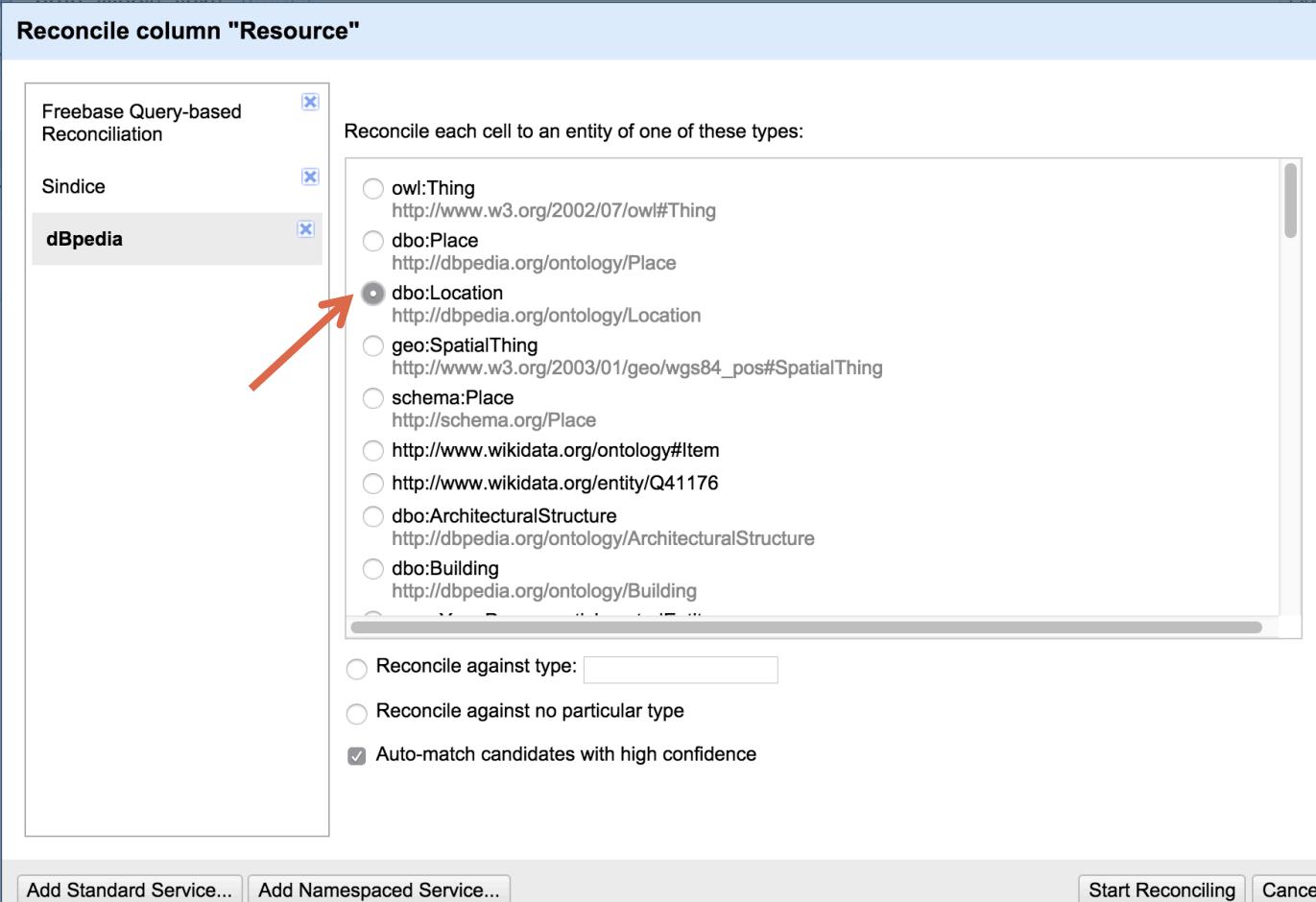
- owl:Thing
<http://www.w3.org/2002/07/owl#Thing>
- dbo:Place
<http://dbpedia.org/ontology/Place>
- dbo:Location
<http://dbpedia.org/ontology/Location>
- geo:SpatialThing
http://www.w3.org/2003/01/geo/wgs84_pos#SpatialThing
- schema:Place
<http://schema.org/Place>
- http://www.wikidata.org/ontology#Item
- http://www.wikidata.org/entity/Q41176
- dbo:ArchitecturalStructure
<http://dbpedia.org/ontology/ArchitecturalStructure>
- dbo:Building
<http://dbpedia.org/ontology/Building>

Reconcile against type:

Reconcile against no particular type

Auto-match candidates with high confidence

Add Standard Service... Add Namespaced Service... Start Reconciling Cancel



If you aren't sure which name space or class to use, you can usually find additional information about these ontologies online.

Review Results

Google refine nrhp_illinois_links Permalink Open... Export Help

Facet / Filter Undo / Redo 2

Refresh Reset All Remove All

City change invert reset
1 choices Sort by: name count Cluster
Champaign 29 exclude
Facet by choice counts

Starred Rows change invert reset
1 choices Sort by: name count
true 29 exclude
Facet by choice counts

Resource: judgment change
1 choices Sort by: name count
none 29
Facet by choice counts

Resource: best candidate's change reset

29 matching rows (1809 total)

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 29 next > last »

			Reference	State	County	City	Resource			
★	98.	86003782	ILLINOIS	Champaign	Champaign		Illinois Traction Building	<input checked="" type="checkbox"/>	Create new topic	
★	99.	89001732	ILLINOIS	Champaign	Champaign		Inman Hotel	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Inman Hotel (0)	
★	100.	90000750	ILLINOIS	Champaign	Champaign		Kappa Delta Rho Frater	<input checked="" type="checkbox"/>	Create new topic	
★	102.	89001109	ILLINOIS	Champaign	Champaign		Kappa Sigma Fraternit	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Create new topic	
★	105.	96000854	ILLINOIS	Champaign	Champaign		Lincoln Building	<input checked="" type="checkbox"/>		
							Lincoln Building (Union Square, Manhattan) (0.15)	<input checked="" type="checkbox"/>		
							Lincoln Building (Champaign, Illinois) (0.15)	<input checked="" type="checkbox"/>		
							<input checked="" type="checkbox"/>	Create new topic		
★	106.	10000993	ILLINOIS	Champaign	Champaign		Mattis, George and Elsie, House	<input checked="" type="checkbox"/>		
★	111.	91000085	ILLINOIS	Champaign	Champaign		New Orpheum Theatre	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Create new topic	
★	112.	4000070	ILLINOIS	Champaign	Champaign		Phi Delta Theta Fraternity House	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Create new topic	
							900 W Park Ave	20101207	http://pdfhost.focus.nps.gov/docs/nrhp/tx	
							346–352 N. Neil St.	19910228	http://pdfhost.focus.nps.gov/docs/nrhp/tx	
							309 E. Chalmers St.	20040225	http://pdfhost.focus.nps.gov/docs/nrhp/tx	

Match this Cell Match All Identical Cells Cancel

[http://dbpedia.org/resource/Lincoln_Building_\(Champaign,_Illinois\)](http://dbpedia.org/resource/Lincoln_Building_(Champaign,_Illinois))

Lincoln Building

The Lincoln Building is a historic commercial building located at 44 E. Main St. in Champaign. Built in 1916, the Commercial style building was designed by architect Harry Roberts Temple. The five-story brick building features a copper cornice and terra cotta decorations. The building's facade is divided into a first-floor base, shaft, and capital at the top, a feature of Commercial style buildings modeled after Classical columns.



Create a Column for New URLs

Google refine nrhp_illinois_links Permalink Open... Export Help

Facet / Filter Undo / Redo 3226

1809 rows Show as: rows records Show: 5 10 25 50 rows « first < previous 31 - 40 next > last »

Using facets and filters

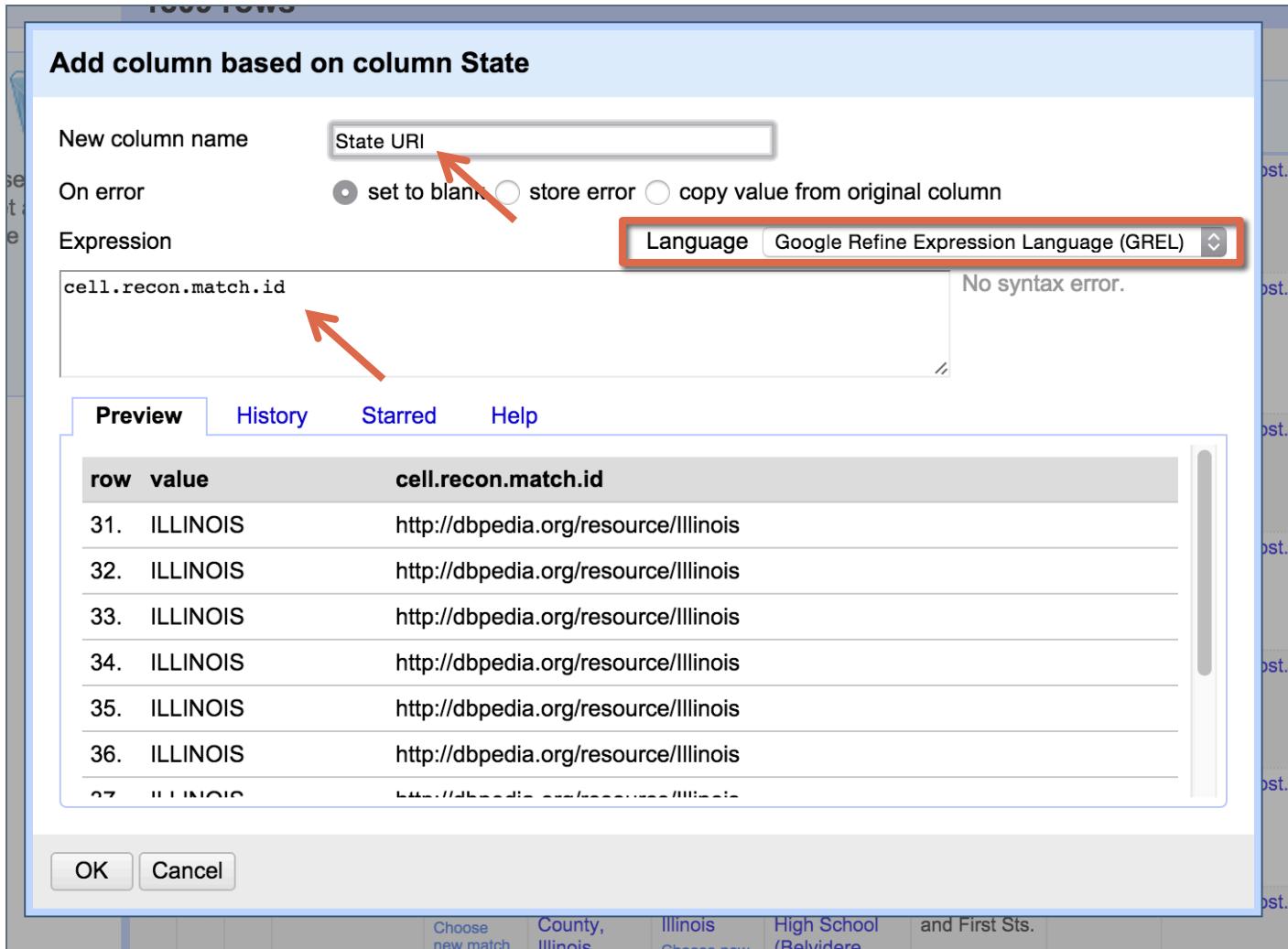
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

31.	69000053	Facet	Cairo, Illinois Choose new match	Magnolia Manor (Cairo, Illinois) Choose new match	2700 Washington Ave.	19691217	http://pdfhost.focus.nps.gov/docs/nrhp/text/69000053.PDF				
32.	96001341	Edit column	Split into several columns...	61115	http://pdfhost.focus.nps.gov/docs/nrhp/text/96001341.PDF						
33.	73000689	Transpose	Add column based on this column...	0724	http://pdfhost.focus.nps.gov/docs/nrhp/text/73000689.PDF						
34.	72000447	Sort...	Add column by fetching URLs...	21226	http://pdfhost.focus.nps.gov/docs/nrhp/text/72000447.PDF						
35.	95000991	View	Add columns from Freebase ...	50804	http://pdfhost.focus.nps.gov/docs/nrhp/text/95000991.PDF						
36.	75000638	Reconcile	Rename this column								
37.	97000815		Remove this column								
38.	12000324		Move column to beginning								
			Move column to end								
			Move column left								
			Move column right								
			match	match							
			Greenville, Illinois Choose new match	Old Main, Almira College Choose new match	315 E. College St.	19750421	http://pdfhost.focus.nps.gov/docs/nrhp/text/75000638.PDF				
			Boone County, Illinois Choose new match	Belvidere, Illinois Choose new match	Old Belvidere High School (Belvidere, Illinois) Choose new match	Jct. of Pearl and First Sts.	19970725	http://pdfhost.focus.nps.gov/docs/nrhp/text/97000815.PDF			
			Boone County, Illinois Choose new match	Belvidere, Illinois Choose new match	Belvidere North State	State St. between	20120606	http://pdfhost.focus.nps.gov/docs/nrhp/text/12000324.PDF			

javascript:{}

Create a Column for New URLs



The screenshot shows the 'Add column based on column State' dialog in Google Refine. The 'New column name' field is set to 'State URI'. Under 'On error', the 'set to blank' option is selected. The 'Expression' field contains the code 'cell.recon.match.id'. A red arrow points from the 'Language' button in the expression field to the 'Google Refine Expression Language (GREL)' dropdown. Another red arrow points from the 'cell.recon.match.id' expression to the preview table below. The preview table shows a list of rows where the value for 'cell.recon.match.id' is consistently 'http://dbpedia.org/resource/Illinois' for all entries labeled 'ILLINOIS' in the 'value' column.

row	value	cell.recon.match.id
31.	ILLINOIS	http://dbpedia.org/resource/Illinois
32.	ILLINOIS	http://dbpedia.org/resource/Illinois
33.	ILLINOIS	http://dbpedia.org/resource/Illinois
34.	ILLINOIS	http://dbpedia.org/resource/Illinois
35.	ILLINOIS	http://dbpedia.org/resource/Illinois
36.	ILLINOIS	http://dbpedia.org/resource/Illinois
37.	ILLINOIS	http://dbpedia.org/resource/Illinois

OK Cancel

Review URI Column

Google refine nrhp_illinois_links Permalink

Facet / Filter Undo / Redo 3227

1809 rows

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

Extensions: [Firebase](#) [RDF](#)

« first < previous 1 - 10 next > last »

Using facets and filters 

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

			All	Reference	State	State URI	County	City	Resource	Address	Listed	Text			
1.	97000032	Illinois	Choose new match			http://dbpedia.org/resource/Illinois	Adams County, Illinois	Quincy, Illinois	Coca-Cola Bottling Company Building	616 N. 24th St.	19970207	http://pdfhost.focus.nps			
2.	83000298	Illinois	Choose new match			http://dbpedia.org/resource/Illinois	Adams County, Illinois	Quincy, Illinois	Downtown Quincy Historic District	Roughly bounded by Hampshire, Jersey, 4th and 8th Sts.	19830407	http://pdfhost.focus.nps			
3.	84000921	Illinois	Choose new match			http://dbpedia.org/resource/Illinois	Adams County, Illinois	Golden, Illinois	Ebenezer Methodist Episcopal Chapel and Cemetery	NW of Golden	19840604	http://pdfhost.focus.nps			
4.	86003714	Illinois	Choose new match			http://dbpedia.org/resource/Illinois	Adams County, Illinois	Golden, Illinois	Exchange Bank (Golden, Illinois)	Quincy St.	19870212	http://pdfhost.focus.nps			
5.	96001282	Illinois	Choose new match			http://dbpedia.org/resource/Illinois	Adams County, Illinois	Payson, Illinois	Fall Creek Stone Arch Bridge	1.2 mi. NE of Fall Cr.—Payson Rd., across Fall Cr.	19961107	http://pdfhost.focus.nps			
6.	79000812	Illinois	Choose new match			http://dbpedia.org/resource/Illinois	Adams County, Illinois	Quincy, Illinois	Robert W. Gardner House	613 Broadway St.	19790620	http://pdfhost.focus.nps			
7.	99001377	Illinois	Choose new match			http://dbpedia.org/resource/Illinois	Adams County, Illinois	Quincy, Illinois	Lesem, S.J., Building	135-37 N 3rd St.	19991122	http://pdfhost.focus.nps			
8.	99001370	Illinois	Choose new match			http://dbpedia.org/resource/Illinois	Adams County, Illinois	Leavenworth, Illinois	2007 E. Leavenworth	200020420	http://pdfhost.focus.nps				

Each event cell now has a unique identifier and a path to discover related information from other data sets.

Clear Reconciliation Data

Google refine nrhp_ilinois_links Permalink

Facet / Filter Undo / Redo 3232

1809 rows

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

Extensions: [Freebase](#) [RDF](#)

Using facets and filters 

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

County	County URI	City	City URI	Resource	Address	Listed
Adams	http://dbpedia.org/resource/Adams_County,_Illinois	Facet	org/resource/Quincy,_Illinois	Coca-Cola Bottling Company Building	616 N. 24th St.	199702C
Adams	http://dbpedia.org/resource/Adams_County,_Illinois	Text filter	org/resource/Quincy,_Illinois	Downtown Quincy Historic District	Roughly bounded by Hampshire, Jersey, 4th and 8th Sts.	198304C
Adams	http://dbpedia.org/resource/Adams_County,_Illinois	Edit cells	org/resource/Golden,_Illinois	Ebenezer Methodist Episcopal Chapel and Cemetery	NW of Golden	198406C
Adams	http://dbpedia.org/resource/Adams_County,_Illinois	Edit column	Reconcile	Start reconciling...		
Adams	http://dbpedia.org/resource/Adams_County,_Illinois	Transpose	match	Facets		
Adams	http://dbpedia.org/resource/Adams_County,_Illinois	Sort...		QA facets		
Adams	http://dbpedia.org/resource/Adams_County,_Illinois	View		Actions		
Adams	http://dbpedia.org/resource/Adams_County,_Illinois	Reconcile	Match each cell to its best candidate	Copy reconciliation data...		
Adams	http://dbpedia.org/resource/Adams_County,_Illinois		Create a new topic for each cell	Discover related RDF datasets...		
Adams	http://dbpedia.org/resource/Adams_County,_Illinois		Create one new topic for similar cells		Fall Creek Stone Arch Bridge	1.2 mi. NE of Fall Cr.- Payson Rd., across Fall Cr.
Adams	http://dbpedia.org/resource/Adams_County,_Illinois		Match all filtered cells to...		Robert W. Gardner House	613 Broadway St.
Adams	http://dbpedia.org/resource/Adams_County,_Illinois		Discard reconciliation judgments		Lesem, S.J., Building	135-37 N 3rd St.
Adams	http://dbpedia.org/resource/Adams_County,_Illinois		Clear reconciliation data	Quincy, Illinois		199911C
Adams	http://dbpedia.org/resource/Adams_County,_Illinois			Choose new match		
Adams	http://dbpedia.org/resource/Adams_County,_Illinois				Lewis Round	2007 E
Adams	http://dbpedia.org/resource/Adams_County,_Illinois					200301C

Clear Reconciliation Data

Google refine nrhp_ilinois_links Permalink

Open... Export... Help

Facet / Filter Undo / Redo 3233

Using facets and filters 

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

1809 rows

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

County	County URI	City	City URI	Resource	Address	Listed
Adams	http://dbpedia.org/resource/Adams_County,_Illinois	Quincy	http://dbpedia.org/resource/Quincy,_Illinois	Coca-Cola Bottling Company Building	616 N. 24th St.	1997020
Adams	http://dbpedia.org/resource/Adams_County,_Illinois	Quincy	http://dbpedia.org/resource/Quincy,_Illinois	Downtown Quincy Historic District Choose new match	Roughly bounded by Hampshire, Jersey, 4th and 8th Sts.	1983040
Adams	http://dbpedia.org/resource/Adams_County,_Illinois	Golden	http://dbpedia.org/resource/Golden,_Illinois	Ebenezer Methodist Episcopal Chapel and Cemetery Choose new match	NW of Golden	1984060
Adams	http://dbpedia.org/resource/Adams_County,_Illinois	Golden	http://dbpedia.org/resource/Golden,_Illinois	Exchange Bank (Golden, Illinois) Choose new match	Quincy St.	1987021
Adams	http://dbpedia.org/resource/Adams_County,_Illinois	Payson	http://dbpedia.org/resource/Category:Payson,_Illinois	Fall Creek Stone Arch Bridge Choose new match	1.2 mi. NE of Fall Cr.—Payson Rd., across Fall Cr.	1996110
Adams	http://dbpedia.org/resource/Adams_County,_Illinois	Quincy	http://dbpedia.org/resource/Quincy,_Illinois	Robert W. Gardner House Choose new match	613 Broadway St.	1979062
Adams	http://dbpedia.org/resource/Adams_County,_Illinois	Quincy	http://dbpedia.org/resource/Quincy,_Illinois	Lesern, S.J., Building	135-37 N 3rd St.	1999112
Adams	http://dbpedia.org/resource/Adams_County,_Illinois	Mendon	http://dbpedia.org/resource/Mendon,_Illinois	Lewis Round Barn Choose new match	2007 E 1250th St.	2003012

javascript:{}

Exporting

Google refine nrhp_illinois_links Permalink

Facet / Filter Undo / Redo 3233

Using facets and filters 

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

1809 rows

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

1.	97000032	Illinois	http://dbpedia.org/resource/Illinois	Adams	http://dbpedia.org/resource/Adams_County,_Illinois		
2.	83000298	Illinois	http://dbpedia.org/resource/Illinois	Adams	http://dbpedia.org/resource/Adams_County,_Illinois		
3.	84000921	Illinois	http://dbpedia.org/resource/Illinois	Adams	http://dbpedia.org/resource/Adams_County,_Illinois		
4.	86003714	Illinois	http://dbpedia.org/resource/Illinois	Adams	http://dbpedia.org/resource/Adams_County,_Illinois	Golden	http://dbpedia.org/resource/Golden,_Illinois
5.	96001282	Illinois	http://dbpedia.org/resource/Illinois	Adams	http://dbpedia.org/resource/Adams_County,_Illinois	Payson	http://dbpedia.org/resource/Payson,_Illinois
6.	79000812	Illinois	http://dbpedia.org/resource/Illinois	Adams	http://dbpedia.org/resource/Adams_County,_Illinois	Quincy	http://dbpedia.org/resource/Quincy,_Illinois
7.	99001377	Illinois	http://dbpedia.org/resource/Illinois	Adams	http://dbpedia.org/resource/Adams_County,_Illinois	Quincy	http://dbpedia.org/resource/Quincy,_Illinois
8.	2001750	Illinois	http://dbpedia.org/resource/Illinois	Adams	http://dbpedia.org/resource/Adams_County,_Illinois	Mendon	http://dbpedia.org/resource/Mendon,_Illinois
9.	4000181	Illinois	http://dbpedia.org/resource/Illinois	Adams	http://dbpedia.org/resource/Adams_County,_Illinois	Quincy	http://dbpedia.org/resource/Quincy,_Illinois

Export project

RDF ▾

Tab-separated value

Comma-separated value

HTML table

Excel

ODF spreadsheet

Triple loader

MQLWrite

Custom tabular exporter...

Templating...

RDF as RDF/XML

RDF as Turtle

... not so fast!

Caveat on Exporting to RDF

RDF Export

Overview

Installation guide

RDF export

Standard SPARQL reconciliation

SPARQL with full-text search reconciliation

Reconciliation using Sindice

Publications

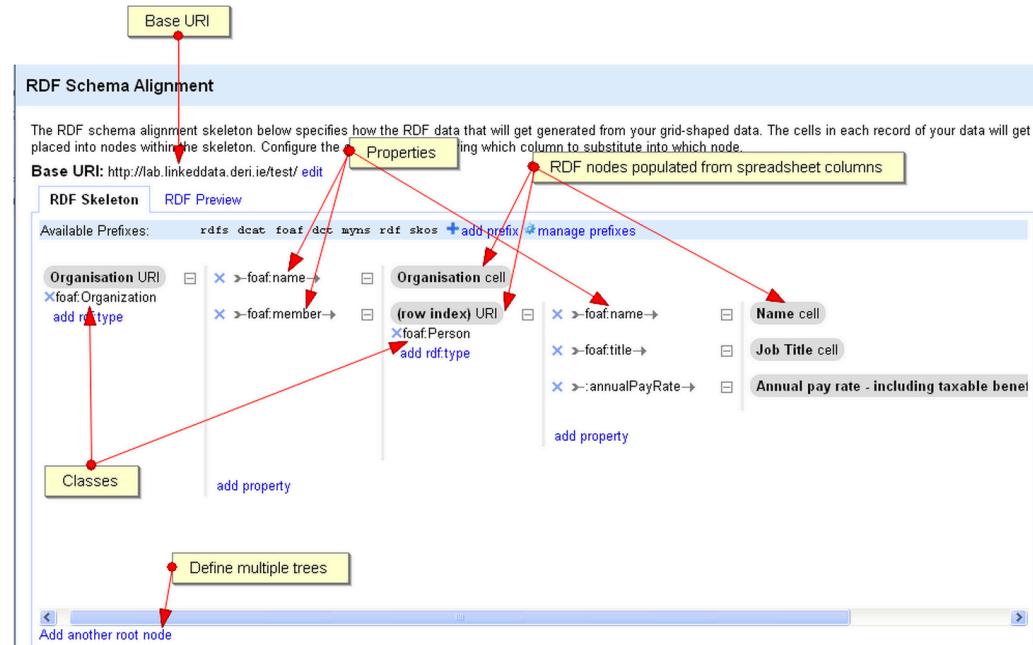
External resources

FAQ

 For detailed example and screenshots see [RDF Export example](#)

With the export functionality, you can determine the intended structure of the RDF data by drawing a template graph. The exporter iterates through the project rows, evaluates expressions in the template graph and produces an equivalent RDF subgraph per row. The final result is the merge of all the subgraphs.

The image below shows the main parts of a graph template.



<http://refine.deri.ie/rdfExportDocs>

Discussion and Questions

Menu.



Cremé of Artichoke.

Purée of Hare.

—
Turbot, Lobster Sauce.

—
Foies Gras Patties.

Snipe and Oyster Pudding.

—
Roast Reindeer.

Boiled Turkey. Tongue.

—
Golden Plover.

—
Asparagus.



—
Swiss Pudding.

Iced Meringues.

—
Anchovy Eclairs.

Menu

Hertford College, Oxford, UK
7 March 1907

Acknowledgments

The NYPL portion of this exercise was created at the Maryland Institute for Technology in the Humanities and developed through the Digital Humanities Data Curation Institute series, generously funded by the National Endowment of the Humanities.

Special thanks to Trevor Muñoz and Lydia Zvyaginsteva!

CIRSS
Center for Informatics Research in
Science and Scholarship

GRADUATE SCHOOL OF **LIBRARY AND
INFORMATION SCIENCE**
The iSchool at Illinois

