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.

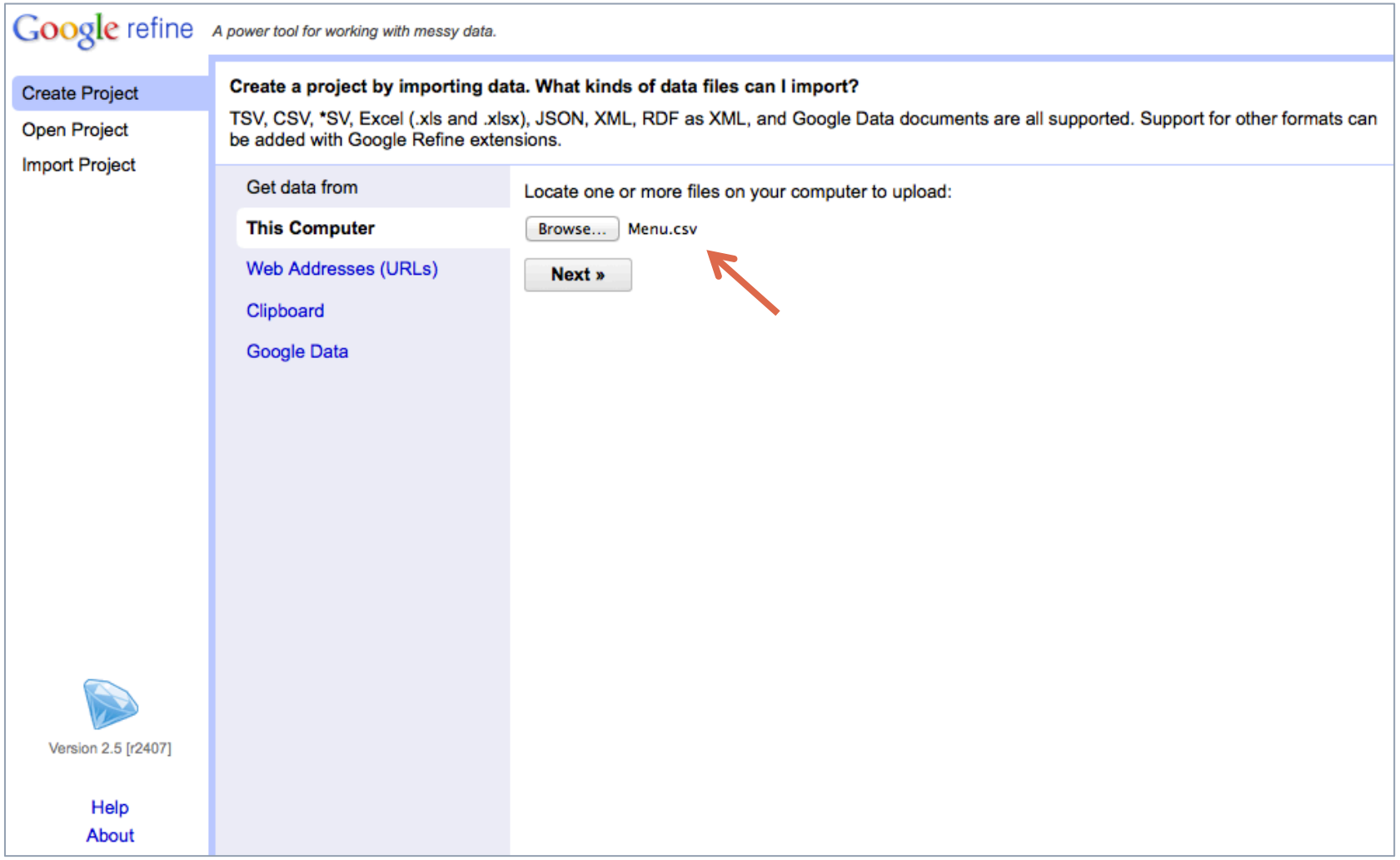
OpenRefine looks like a ***spreadsheet***, but it acts like a ***database***.

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

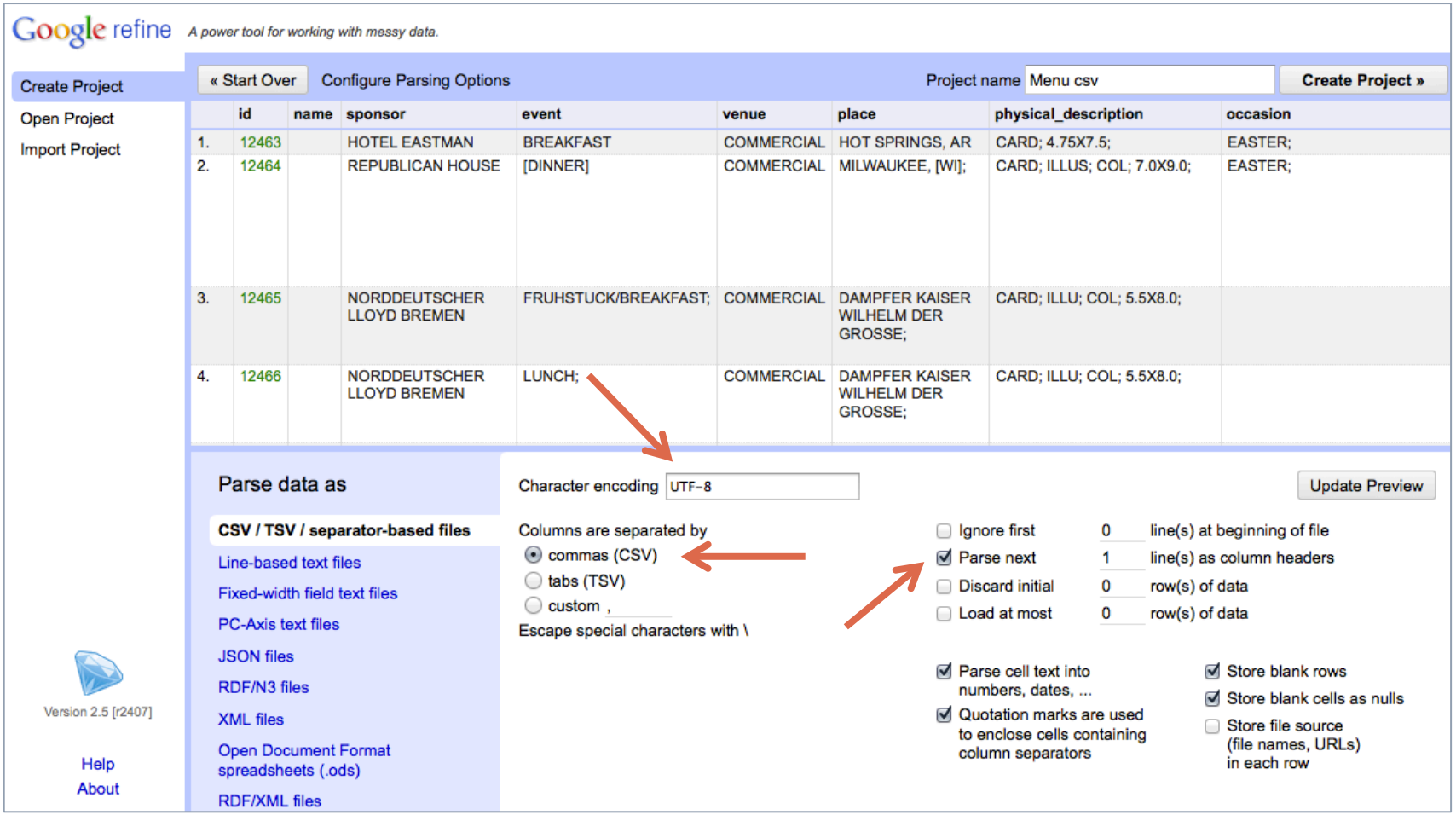
Today we will cover….

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

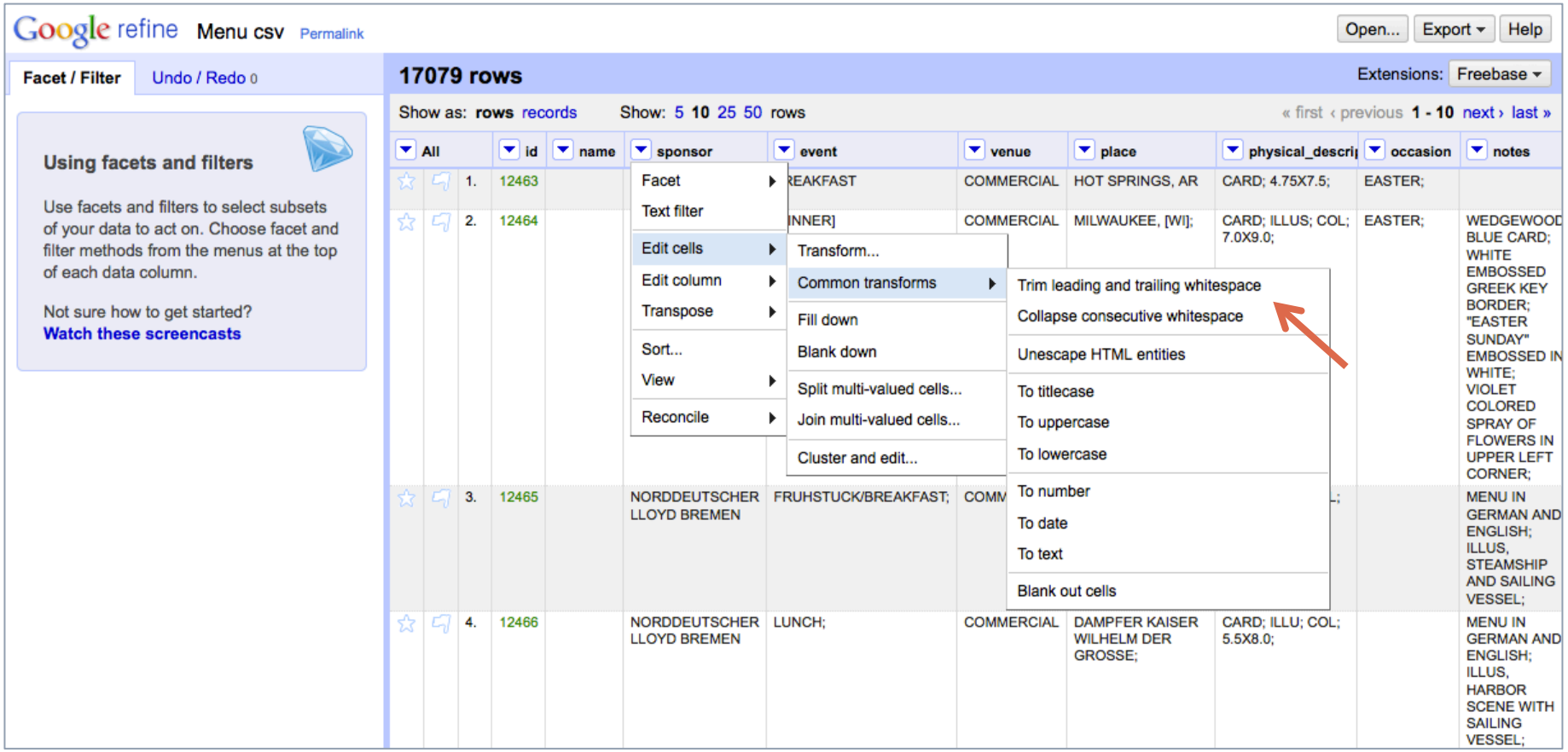
Creating a Project



Creating a Project



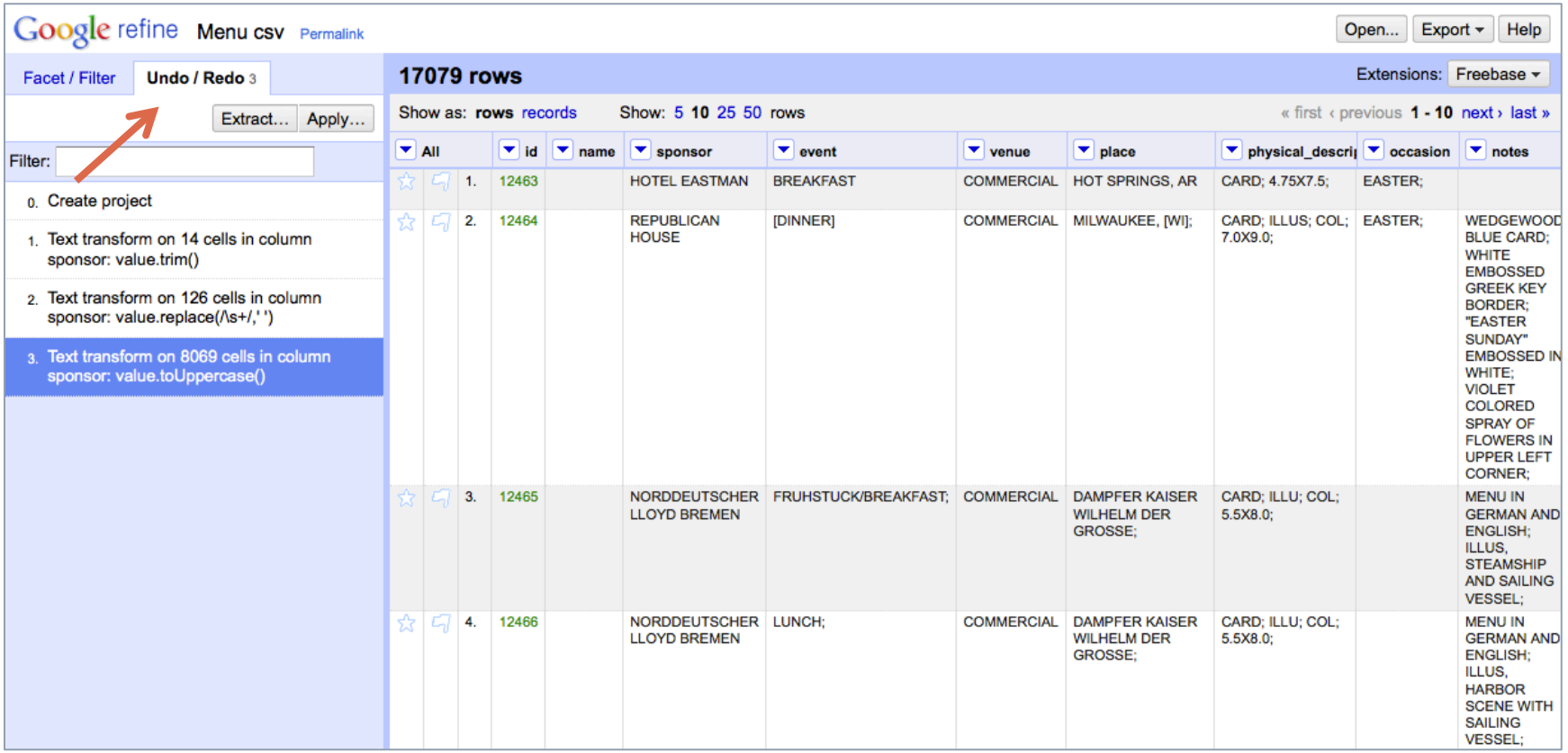
Basic Normalization



Other ones to try:

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

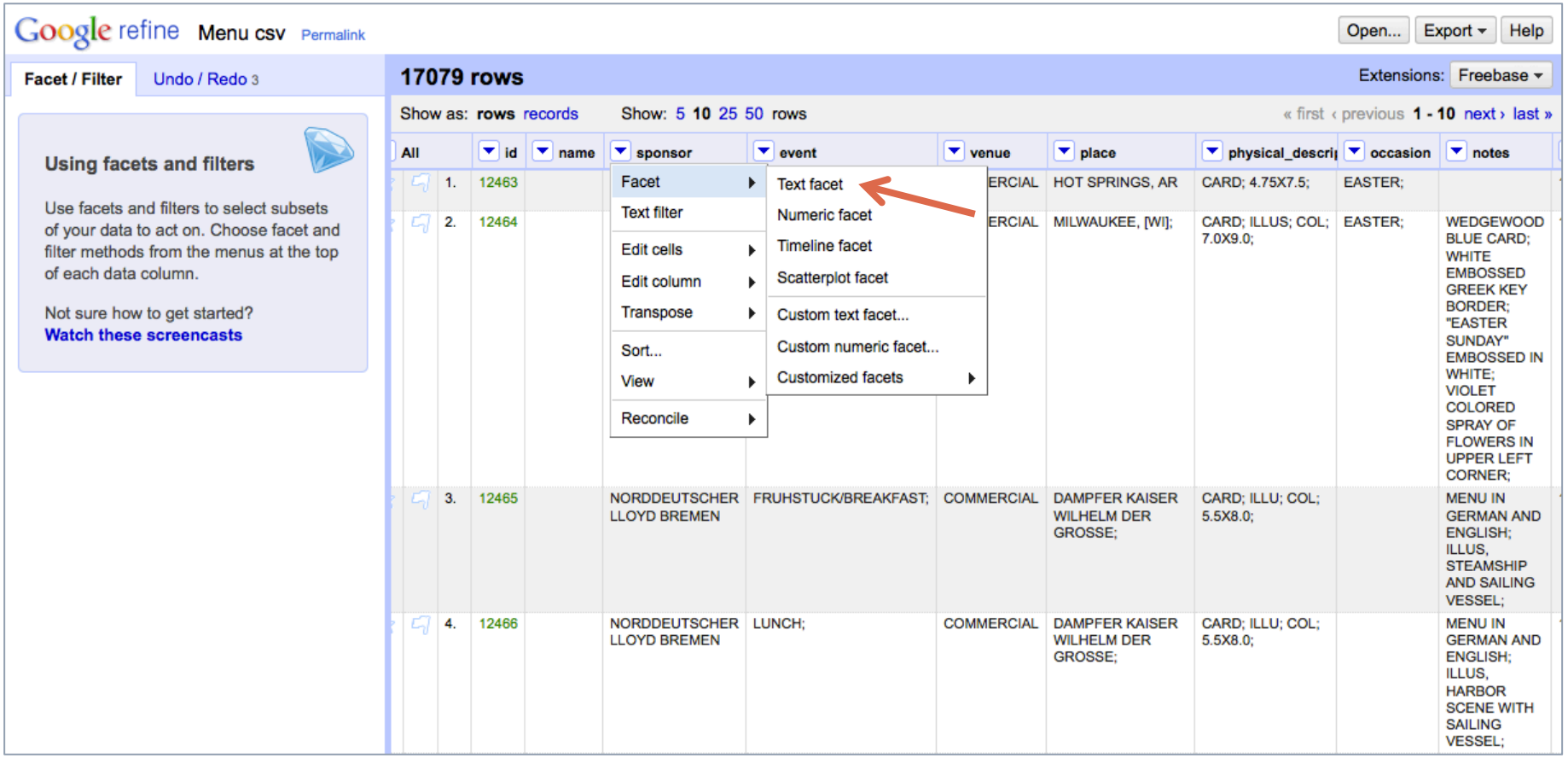
A Note on Tracking Project History



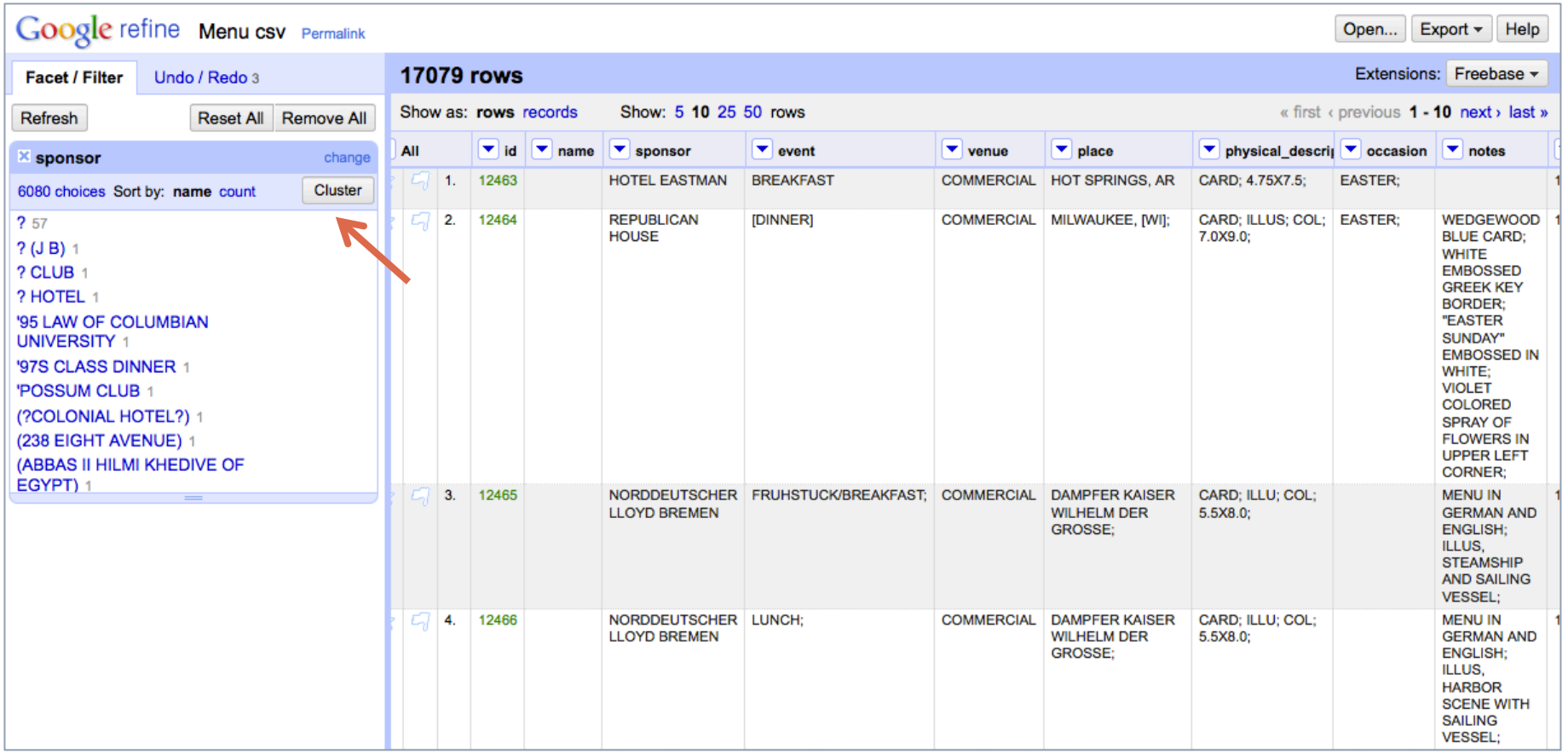
Don’t be afraid to make mistakes!

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

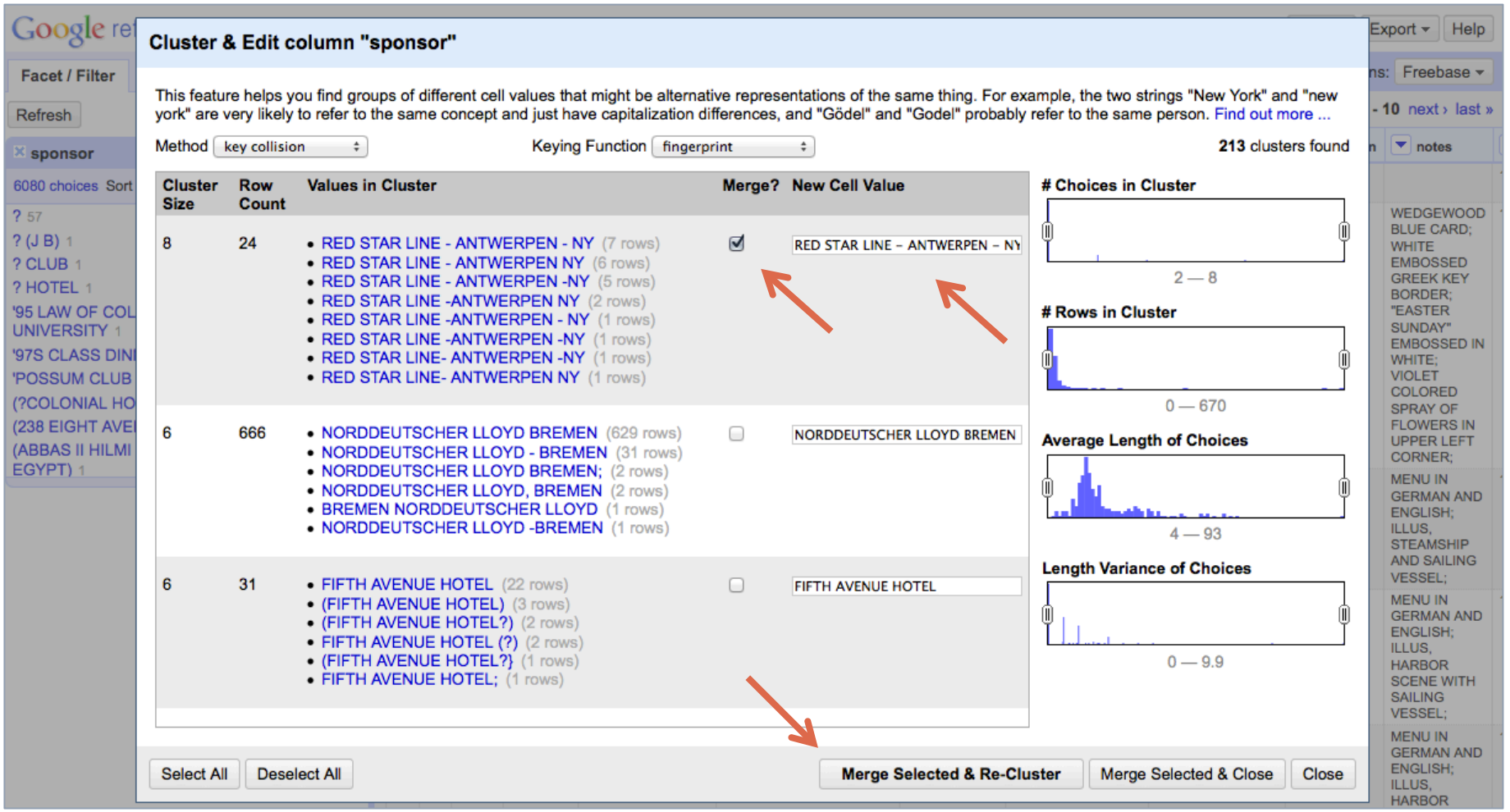
Faceting and Clustering



Faceting and Clustering



Faceting and Clustering



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

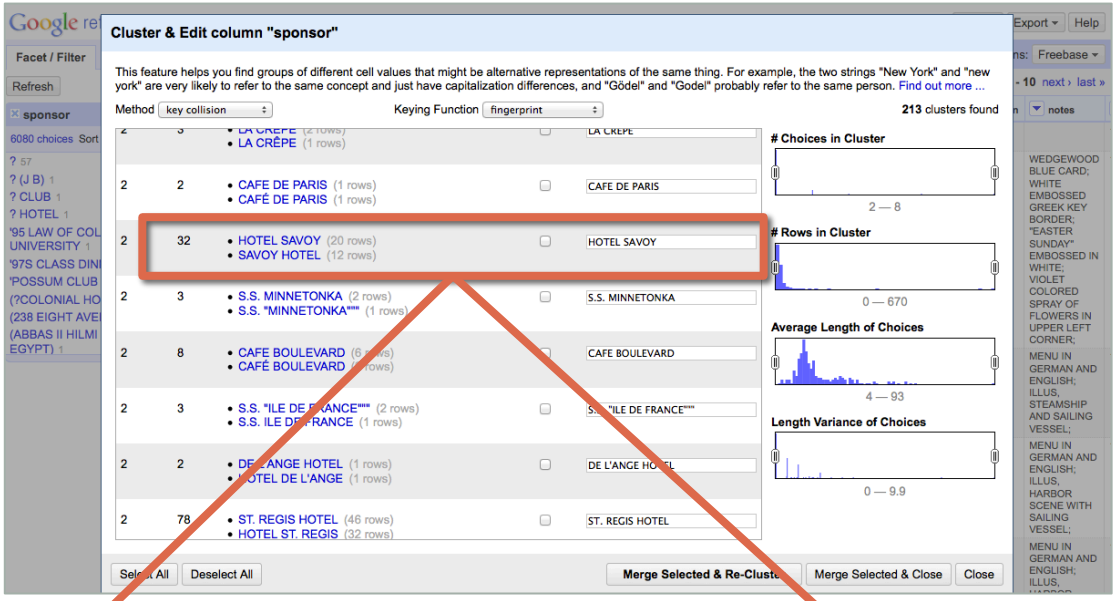
This can be computationally intensive – and can sometimes crash – so use with caution or patience on large data sets

Be careful when clustering

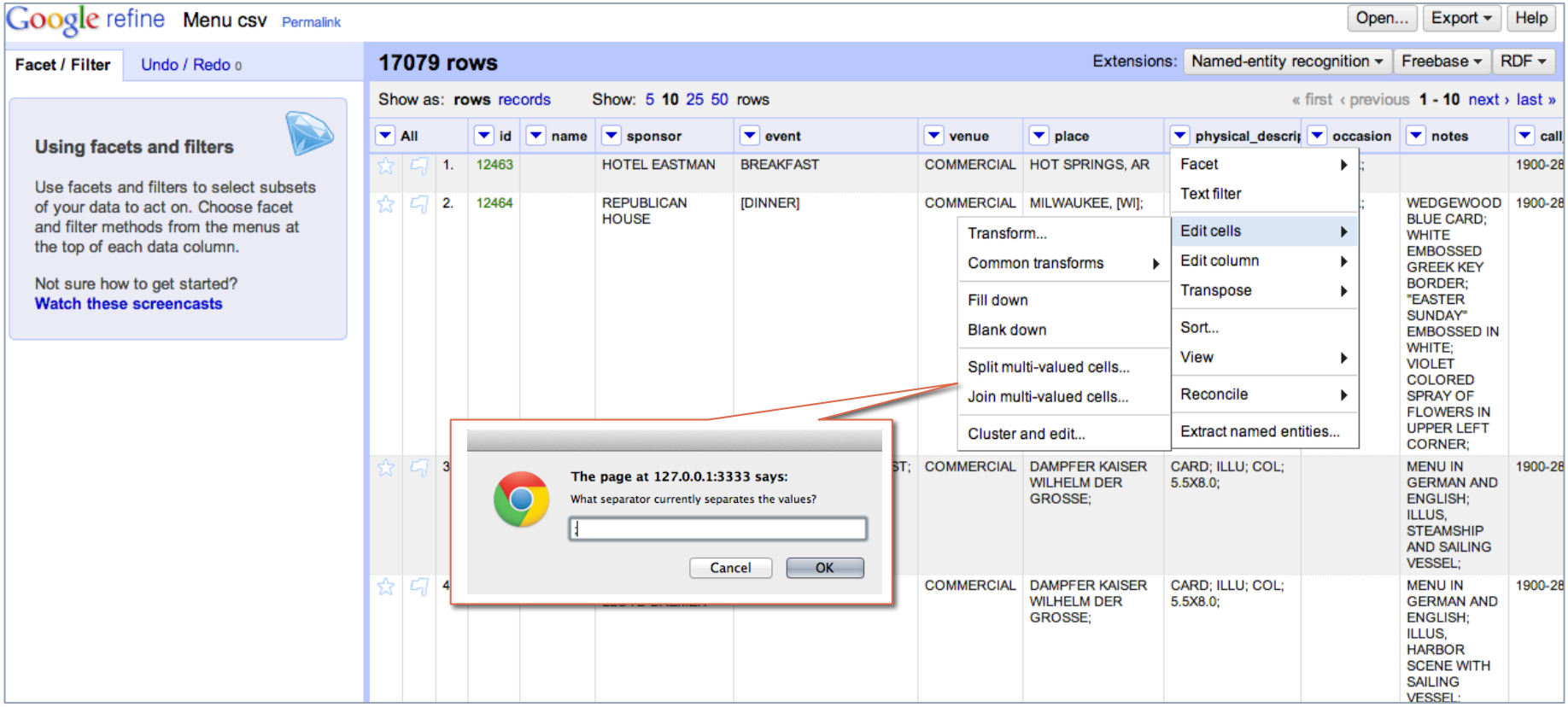
This is part of that important “human-only” work we’ve talked about

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

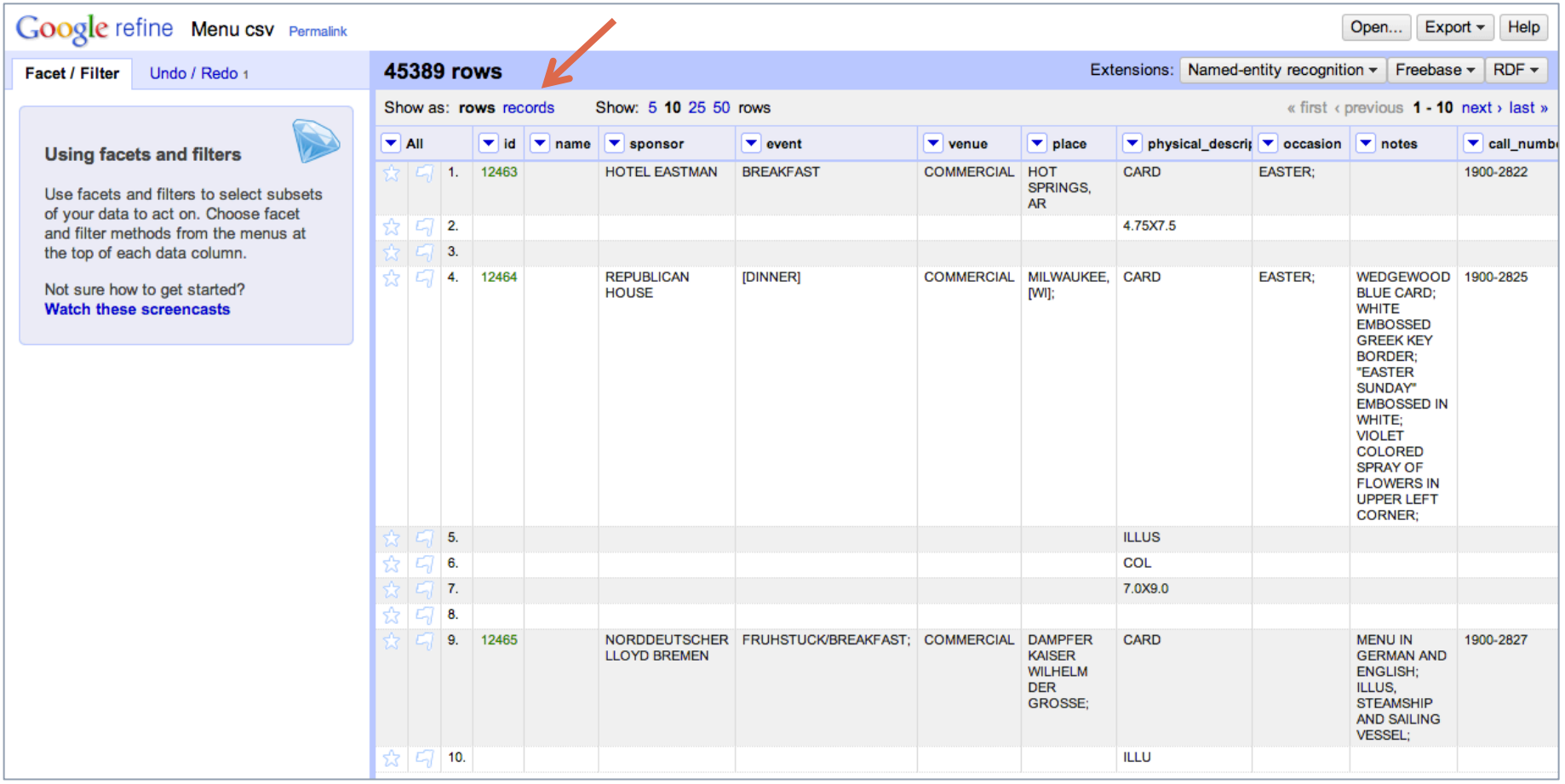
Savoy Hotel
Strand
London WC2R 0EU
United Kingdom



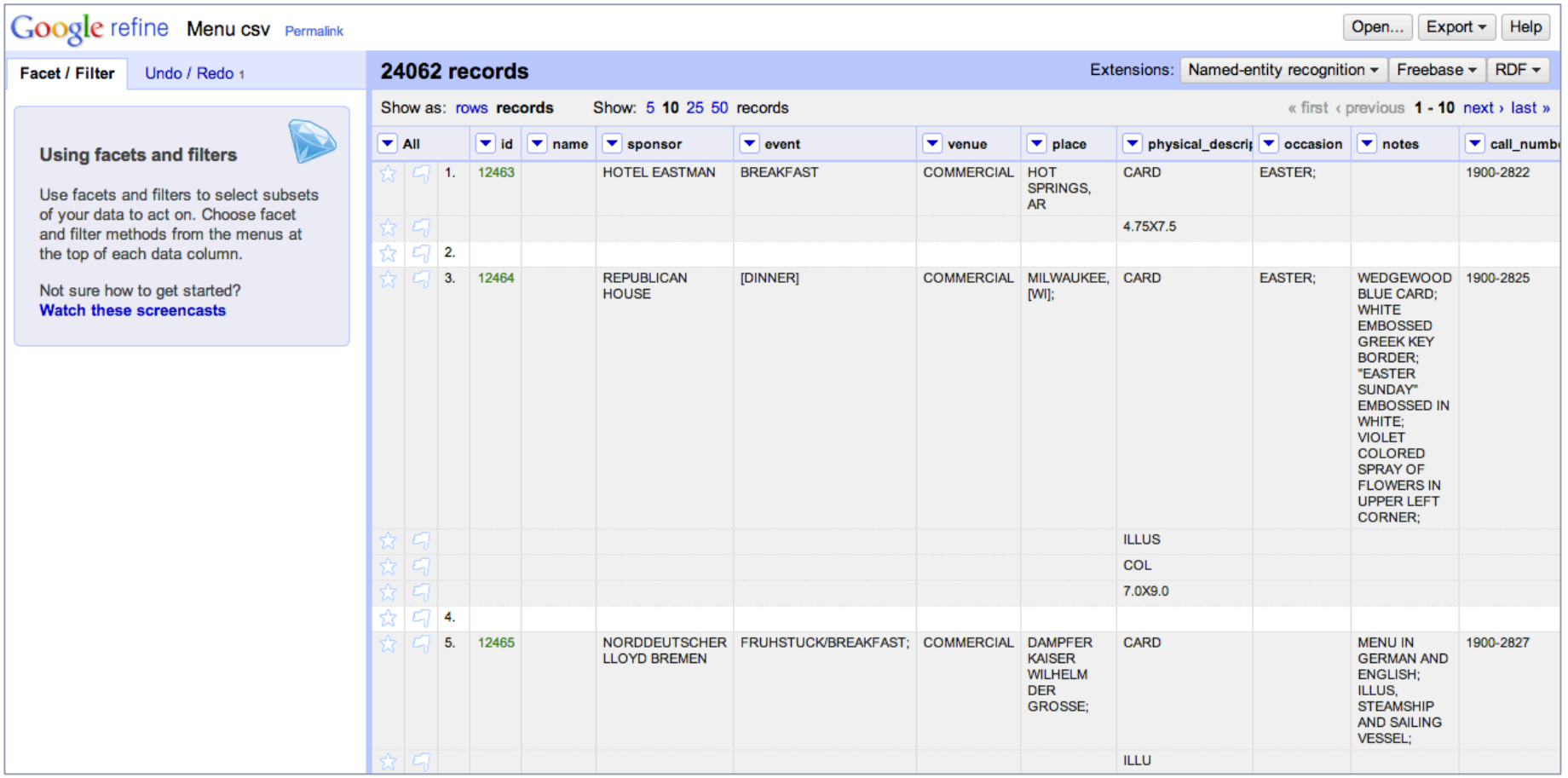
Records vs. Rows



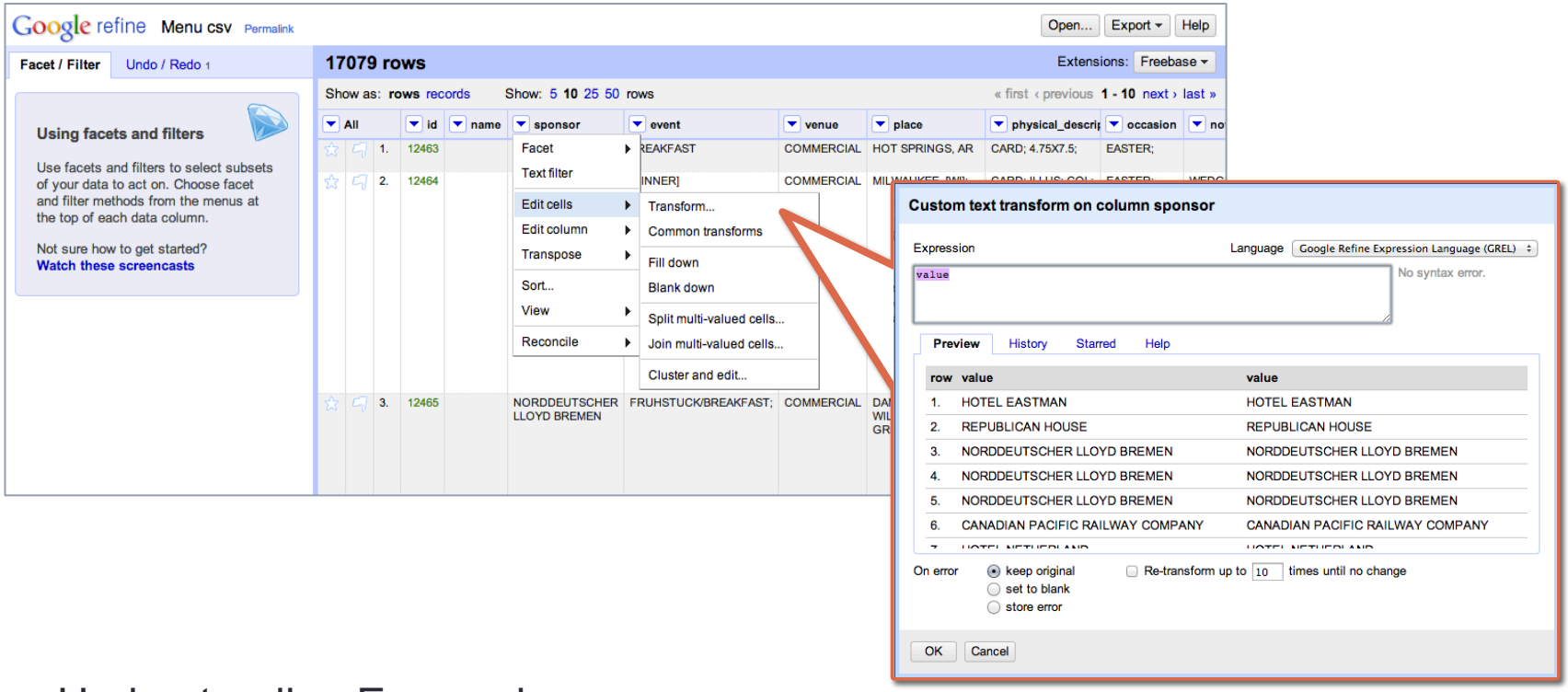
Multiple Values, Multiple Rows



One Record, Multiple Rows



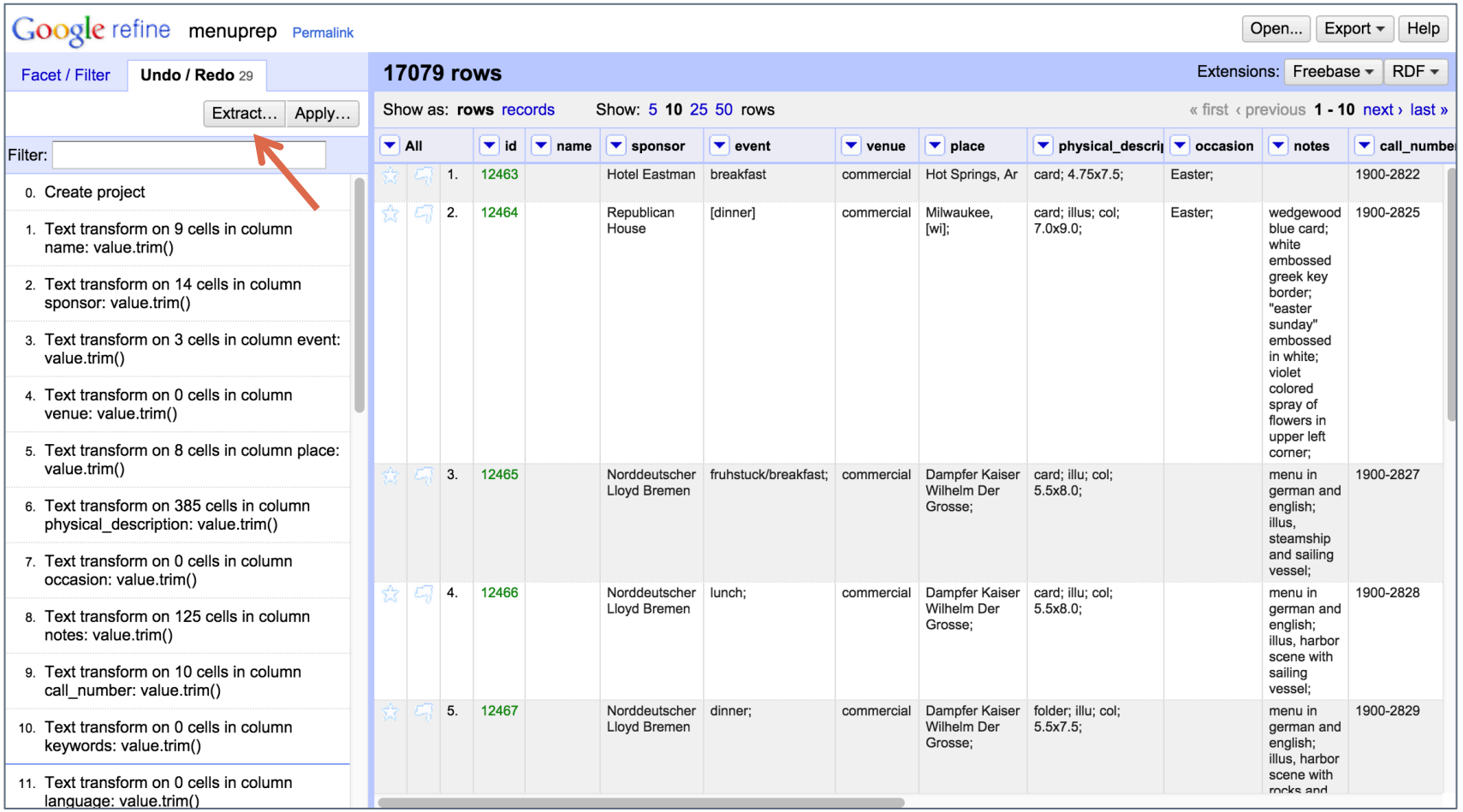
Advanced Transformations



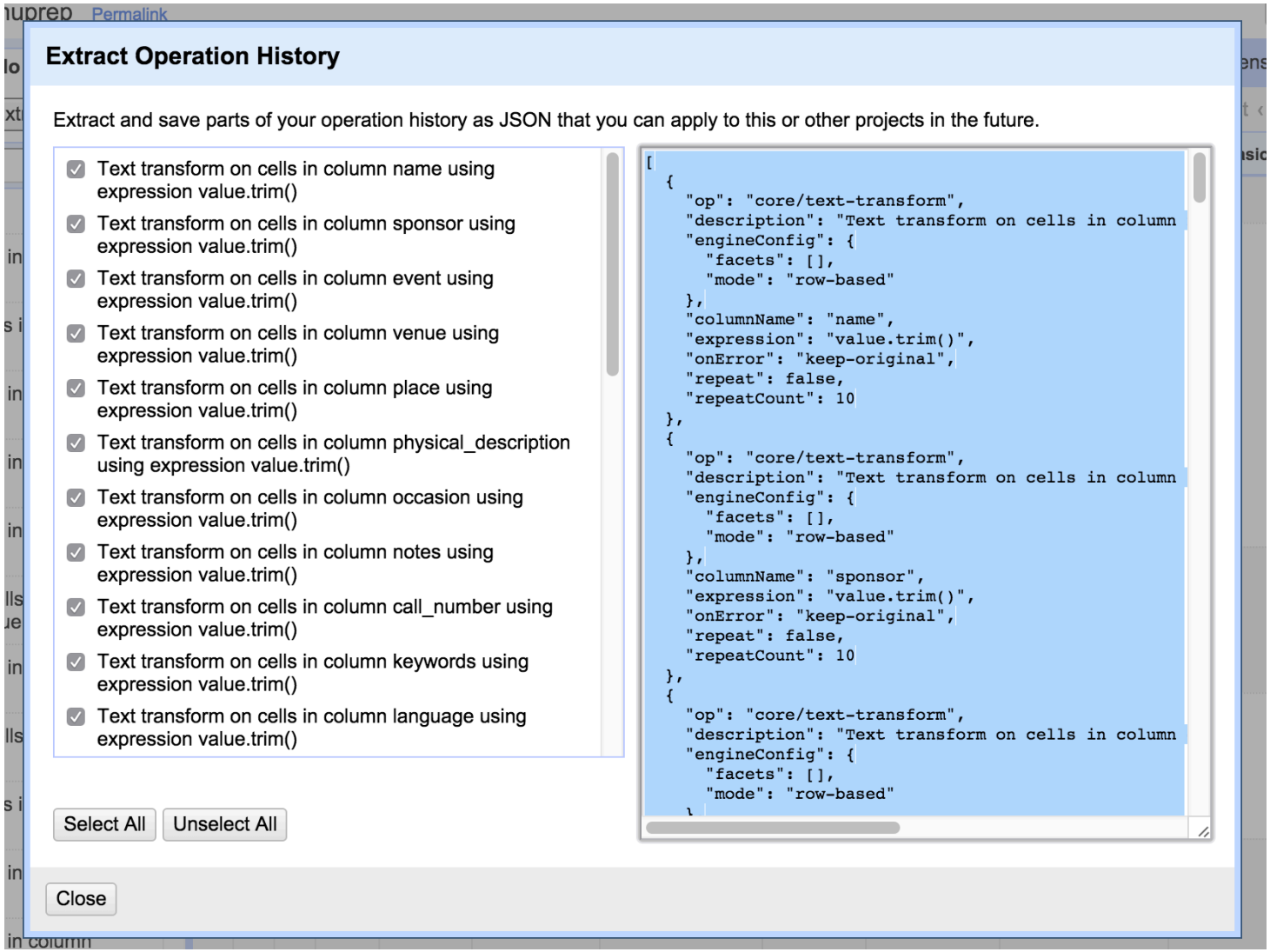
Understanding Expressions:
https://github.com/OpenRefine/OpenRefine/wiki/Understanding-Expressions

GREL Cheat Sheet:
http://arcadiafalcone.net/GoogleRefineCheatSheets.pdf

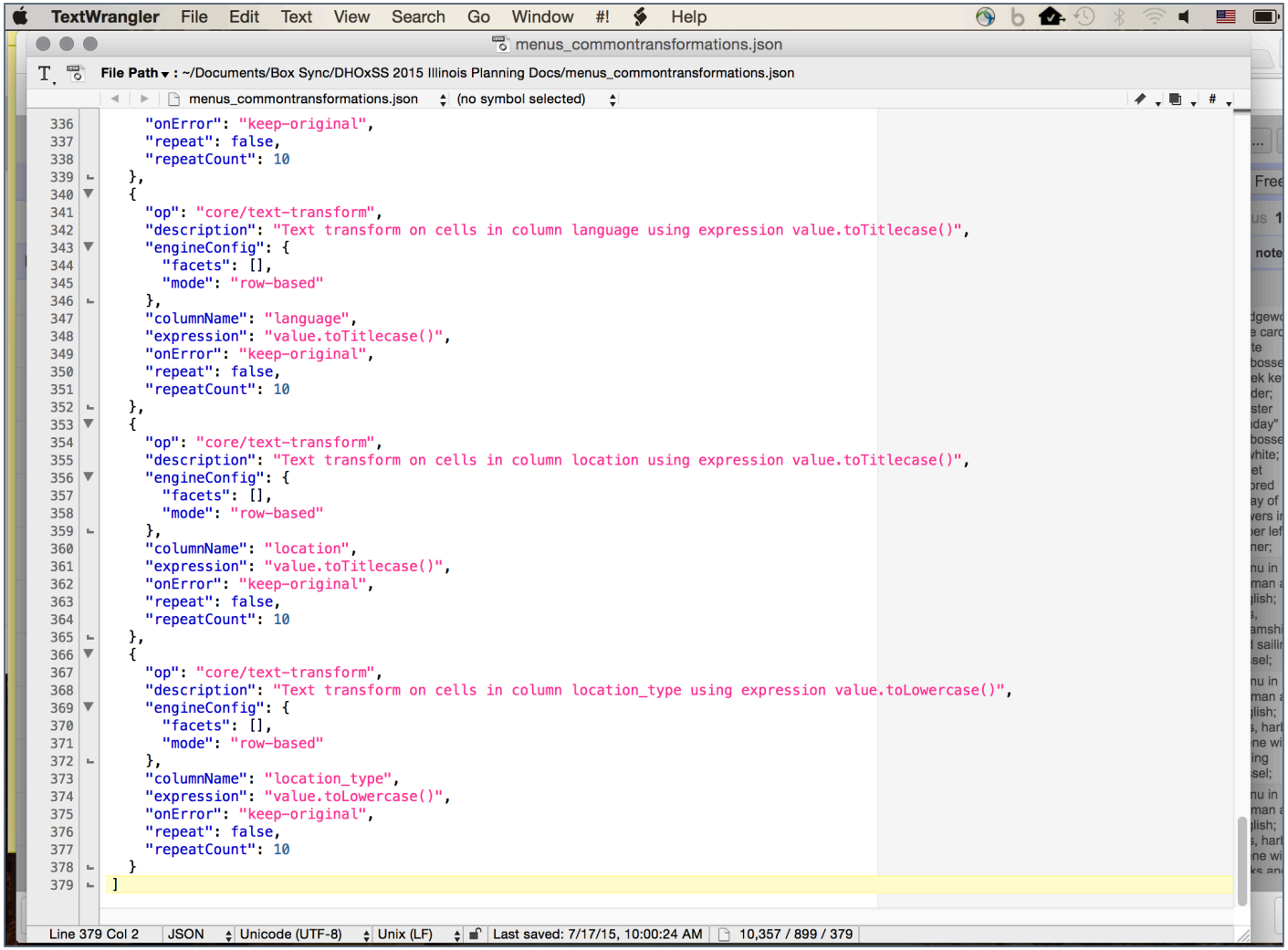
Extracting History as Provenance



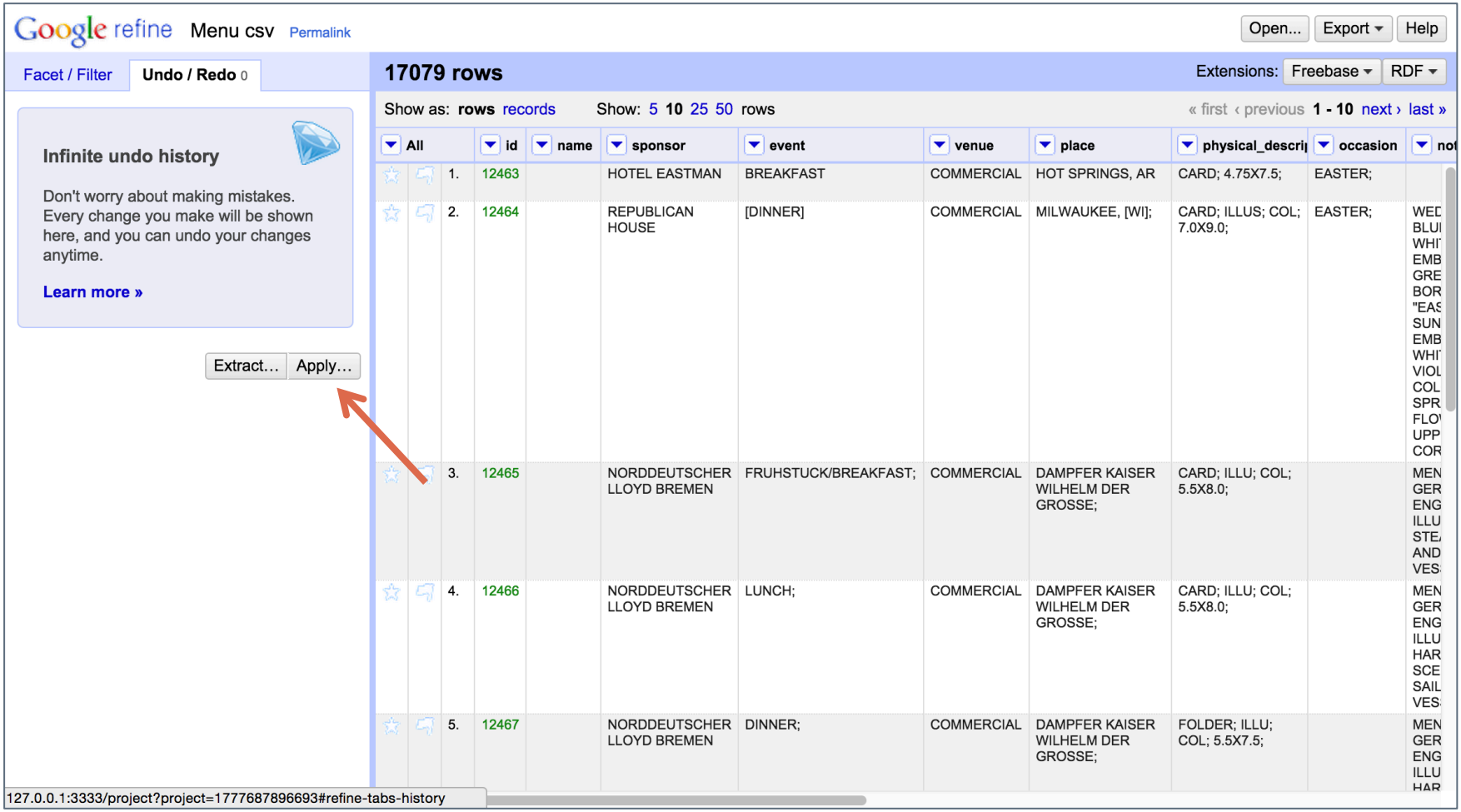
Extracting History as Provenance



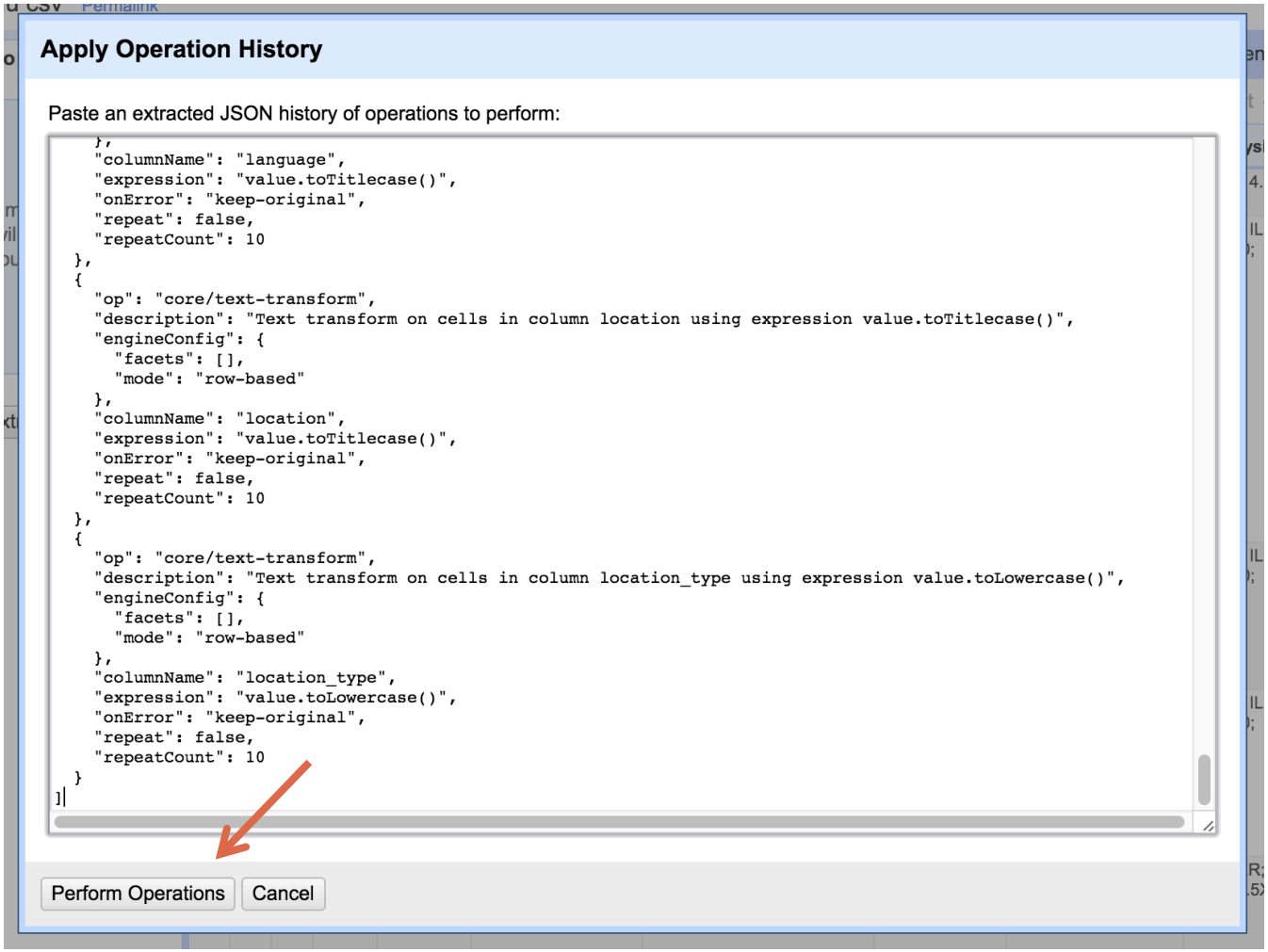
Extracting History as Provenance



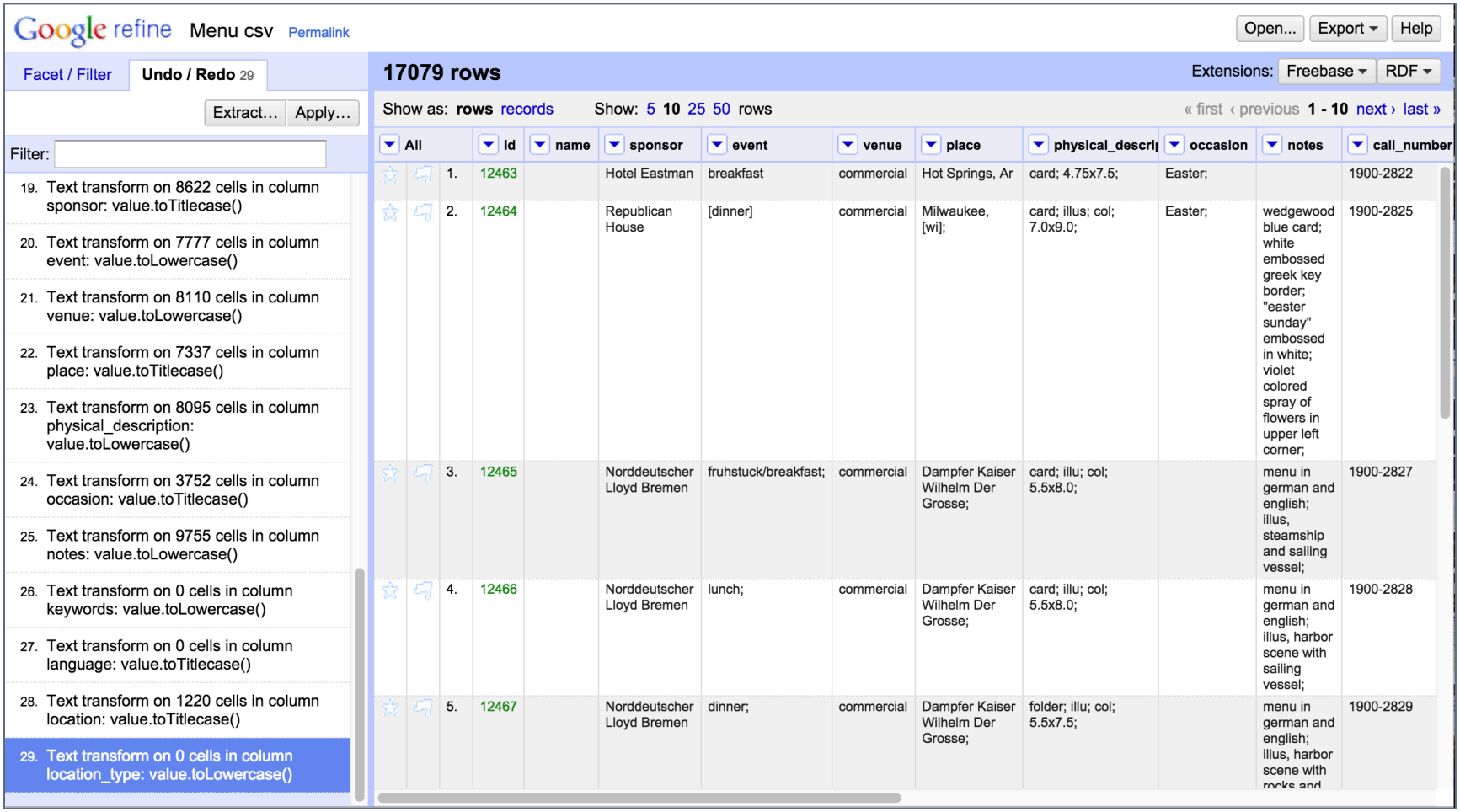
Applying JSON to New Project



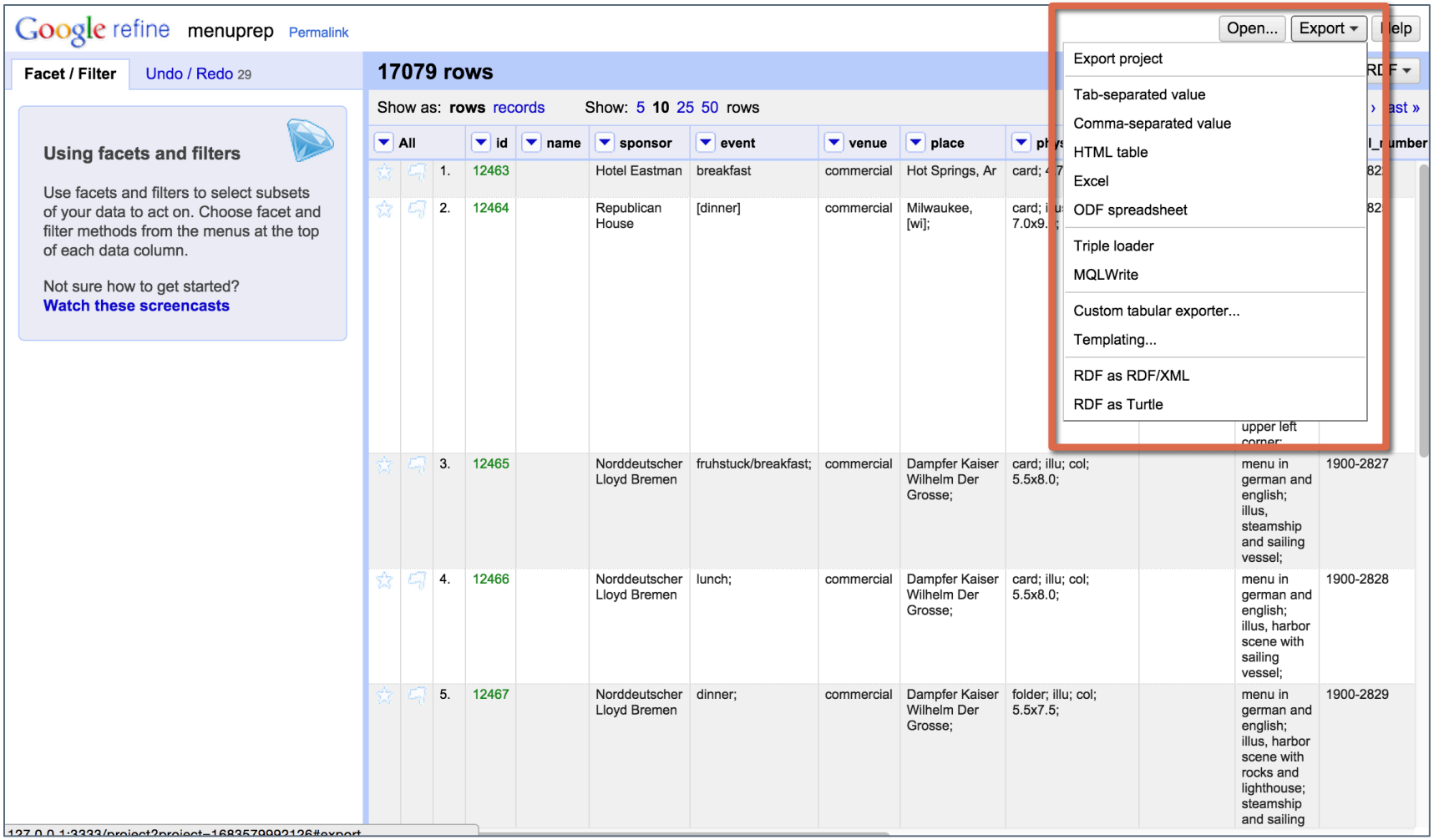
Applying JSON to New Project



Applying JSON to New Project



Exporting your Project



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 a 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.

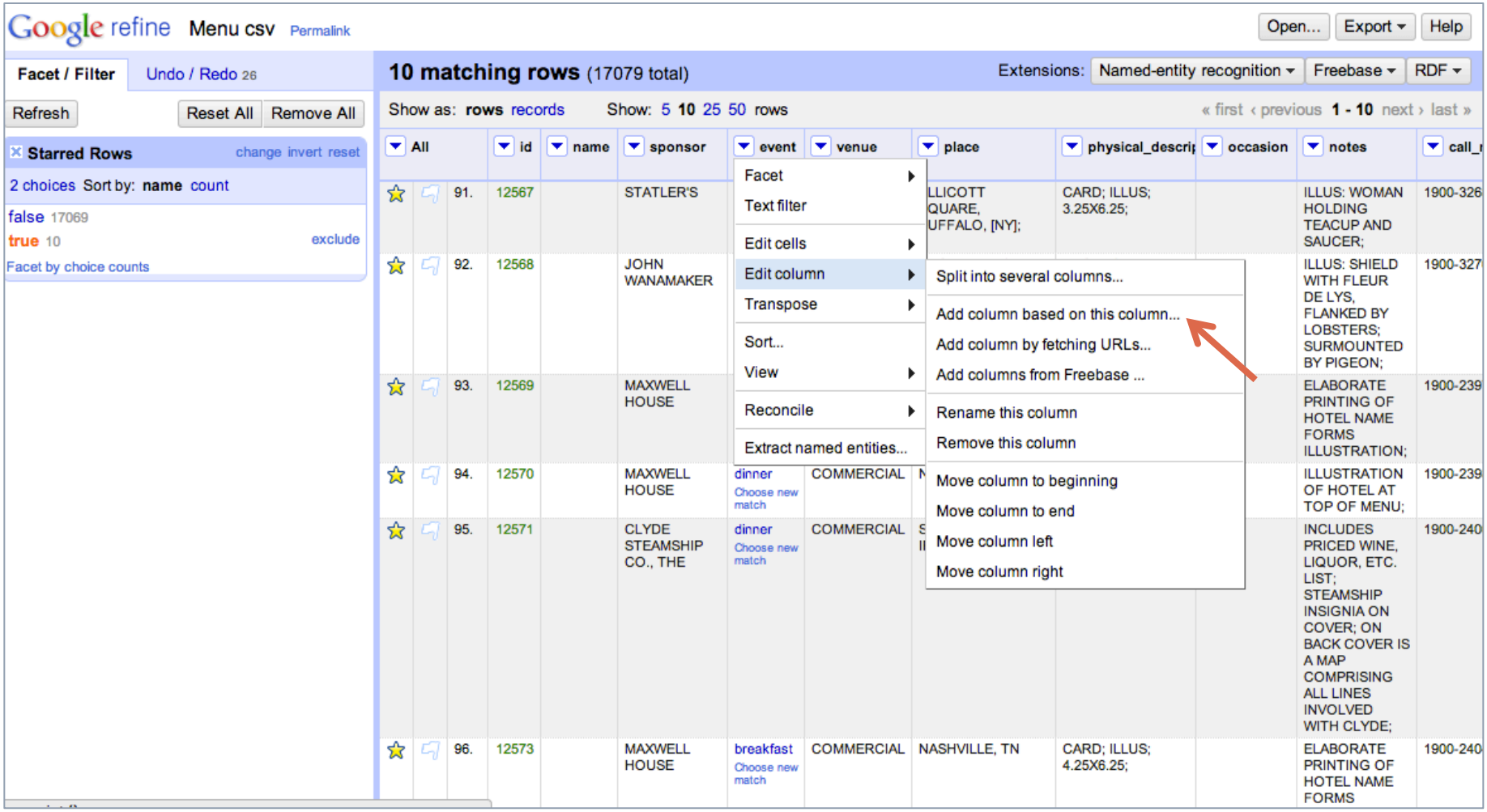
Why do we need it?

- 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/

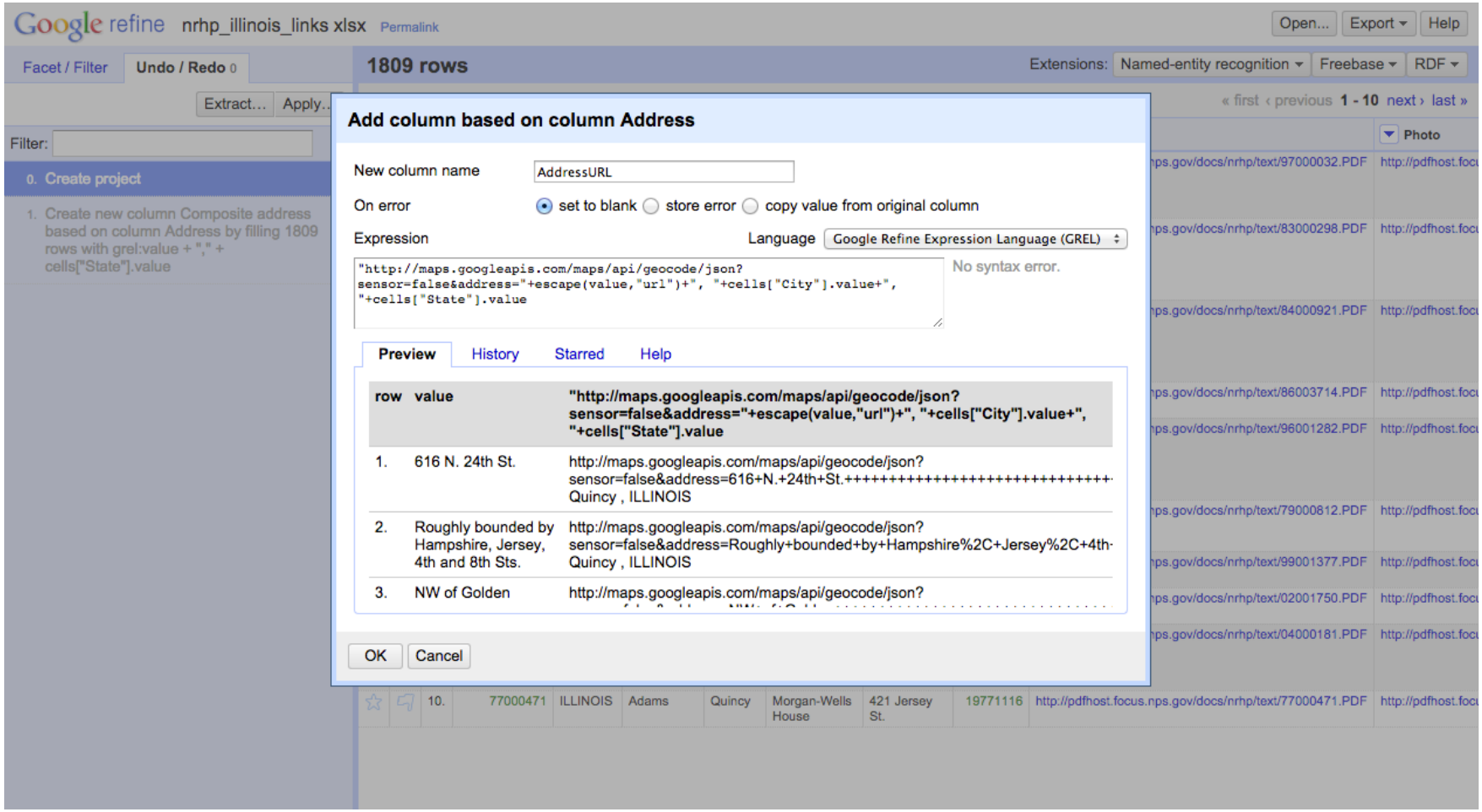
How can OpenRefine help?

- 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

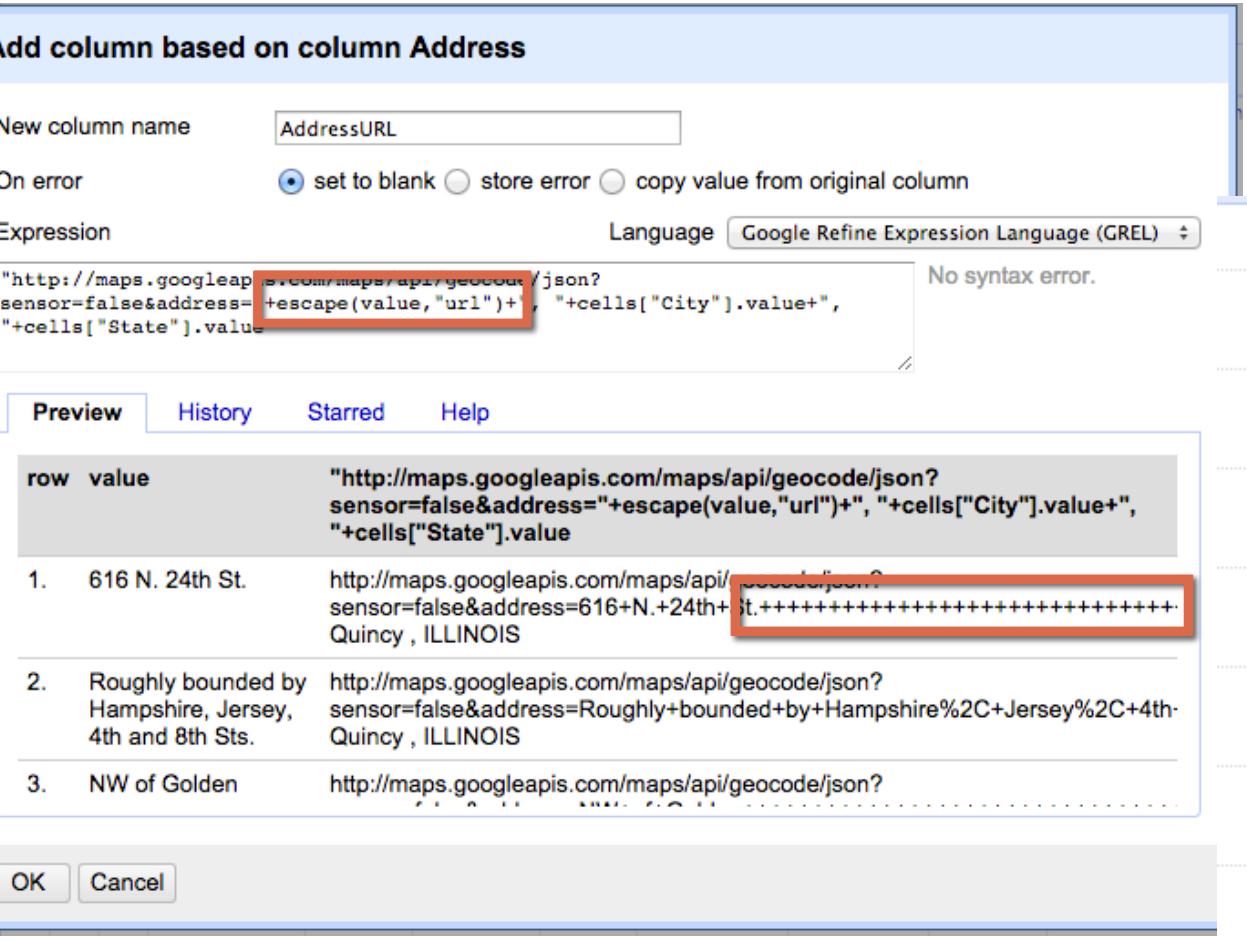
Create a Column for New URLs



Wait a second….

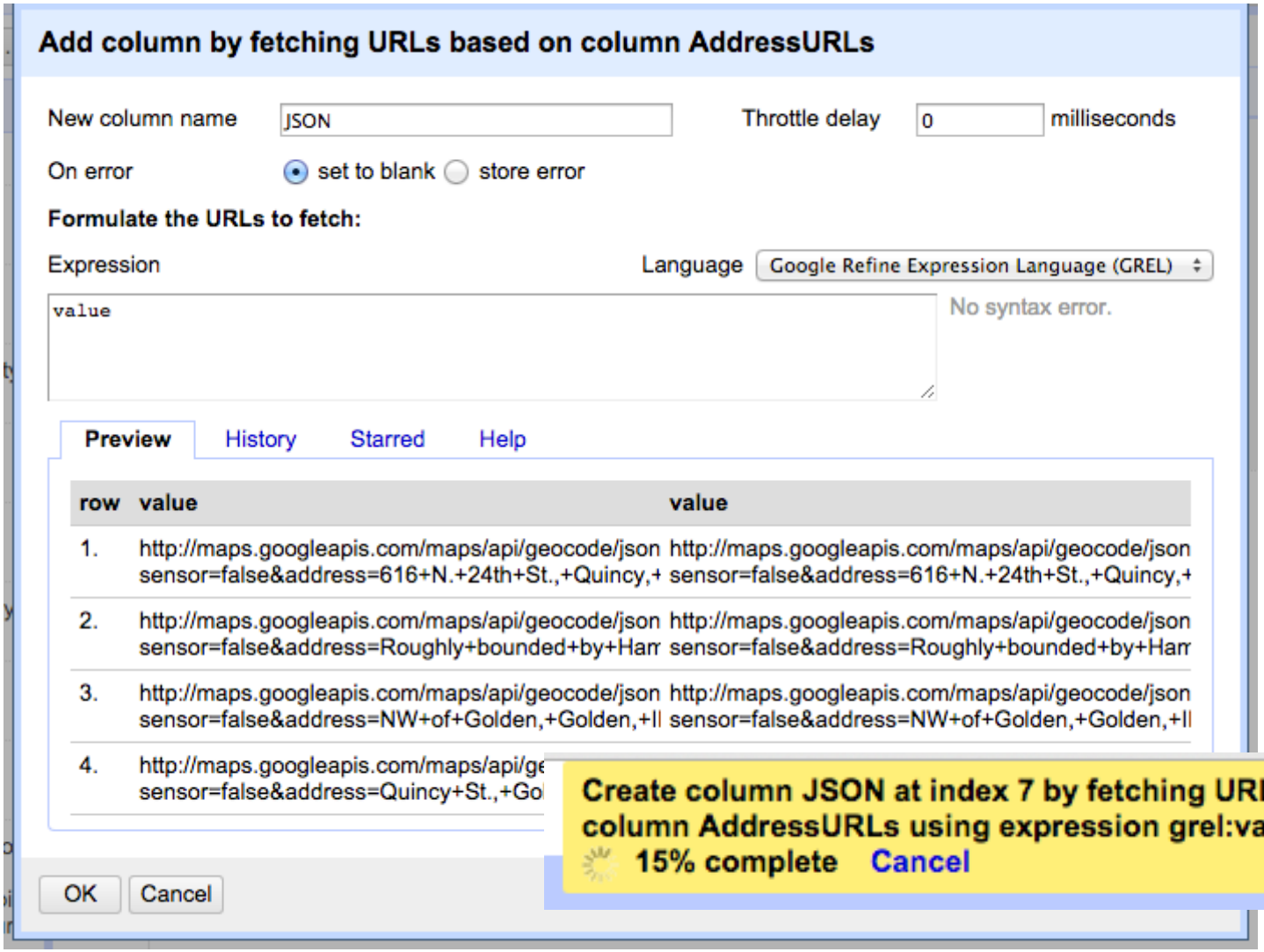


Remember: clean your data!

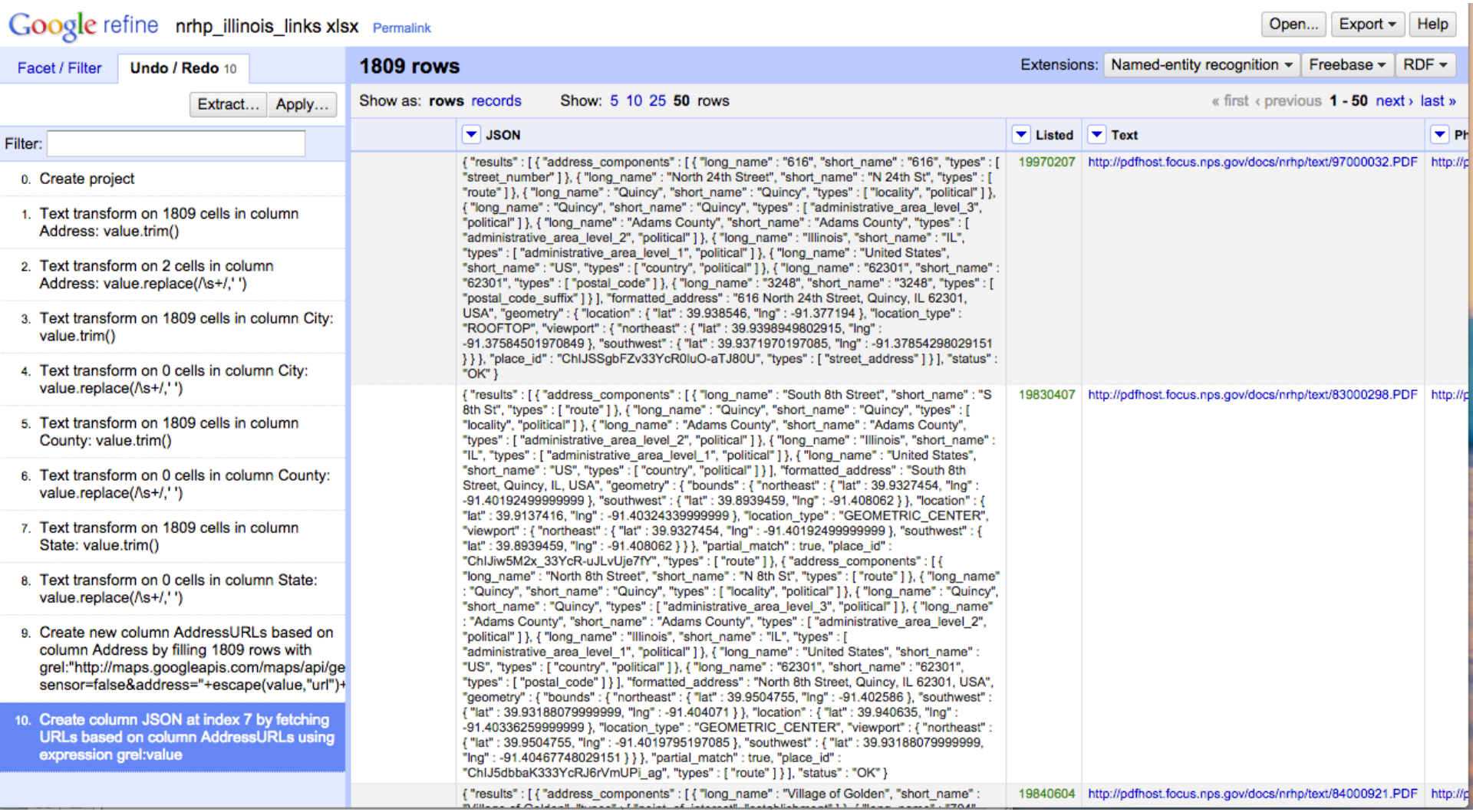


0. Create project
1. Text transform on 1809 cells in column Address: value.trim()
2. Text transform on 2 cells in column Address: value.replace(/\s+/,' ')
3. Text transform on 1809 cells in column City: value.trim()
4. Text transform on 0 cells in column City: value.replace(/\s+/,' ')
5. Text transform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

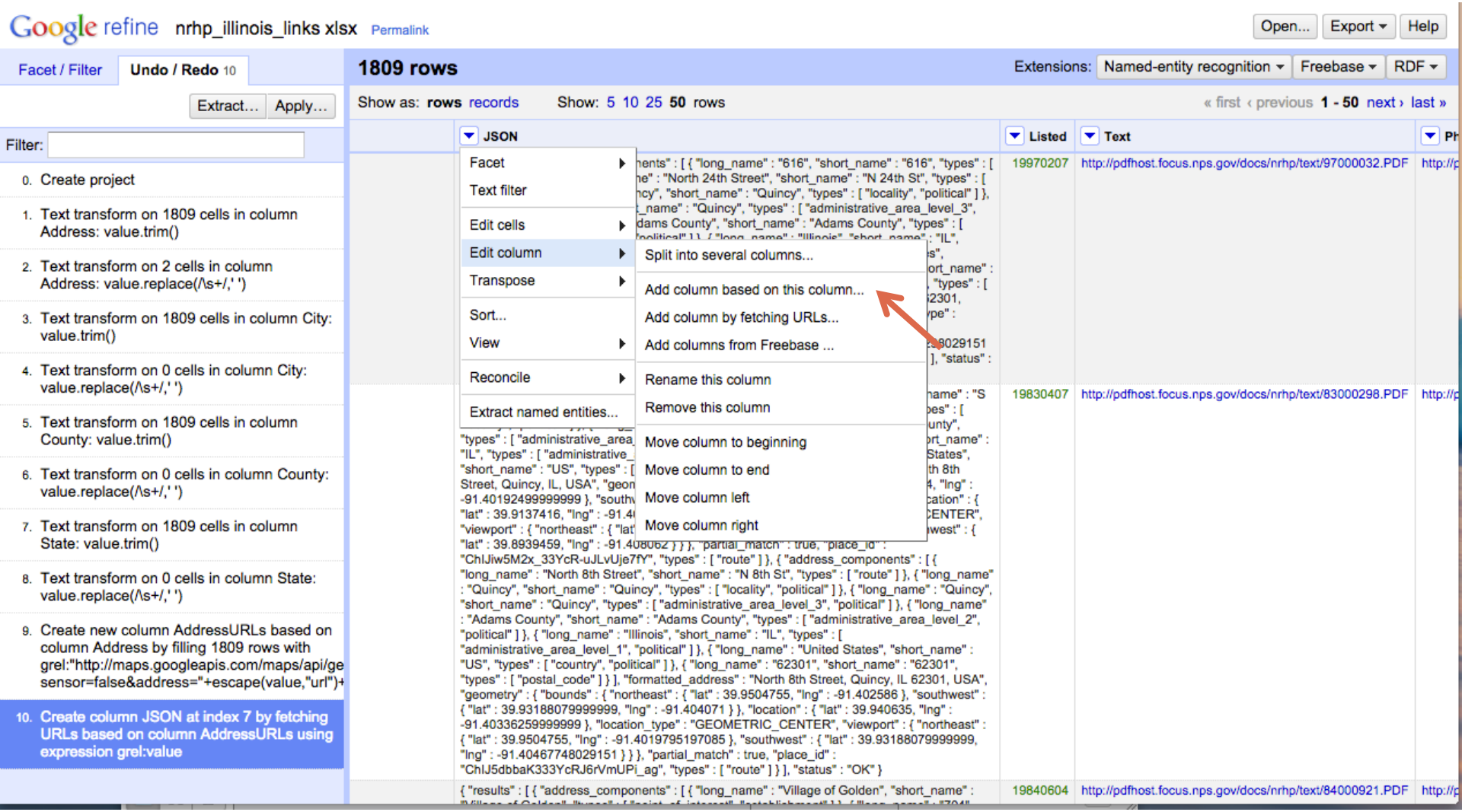
Fetching URLs



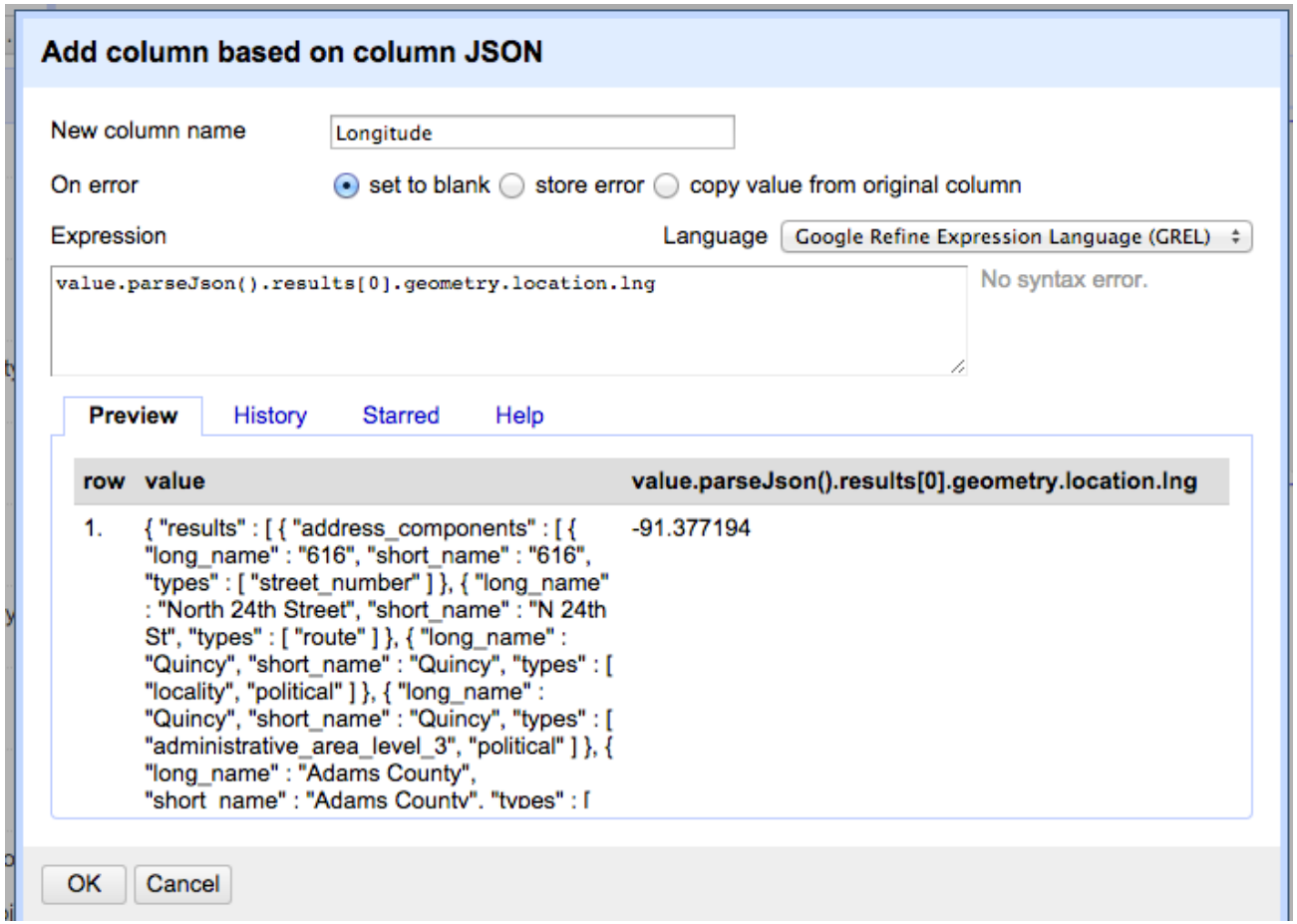
Results: JSON!



Parsing JSON

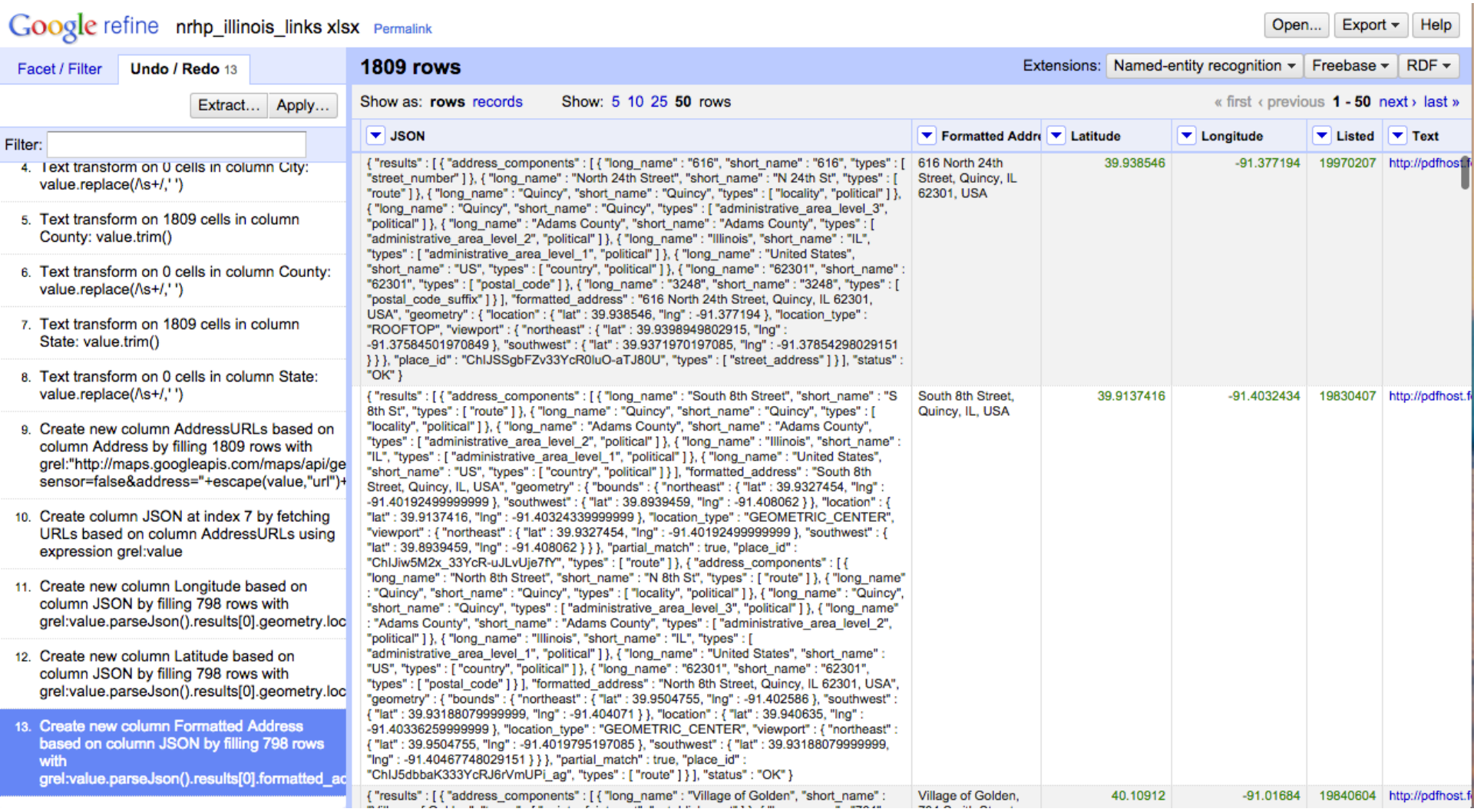


Parsing JSON



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

We're ready to make maps!



FROM RESTAURANTS TO ROAD TRIPS

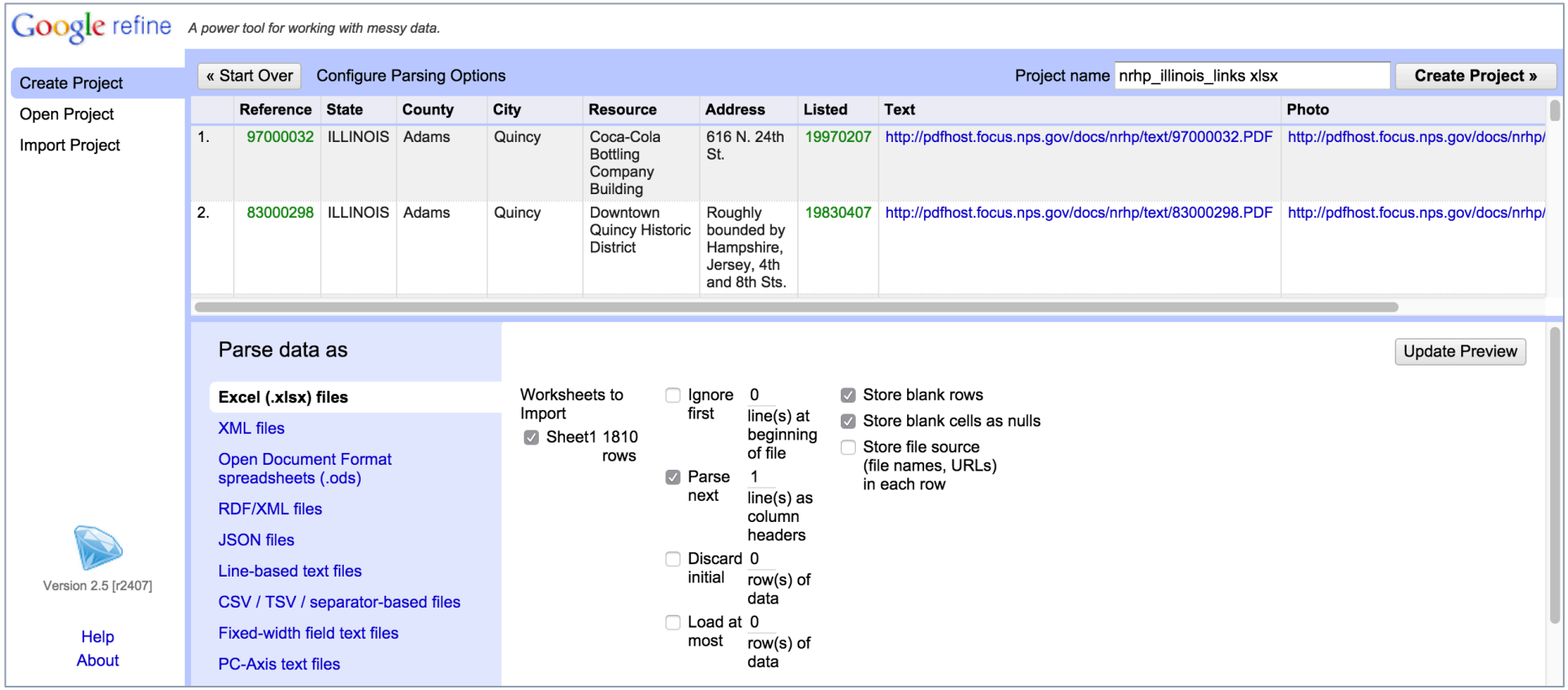
The National Register of Historic Places

Reconciling with Linked Data

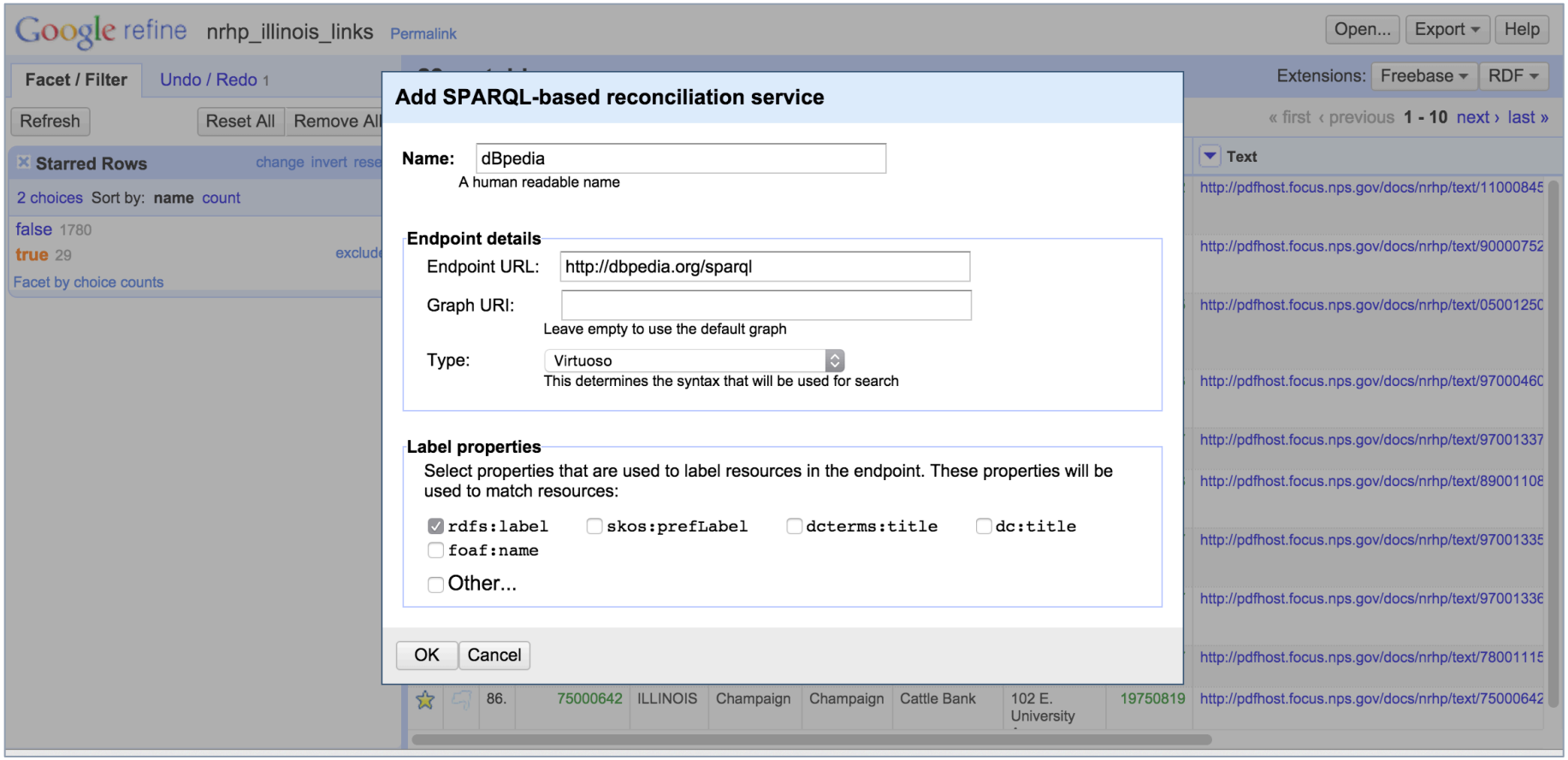
What you'll need:

- 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

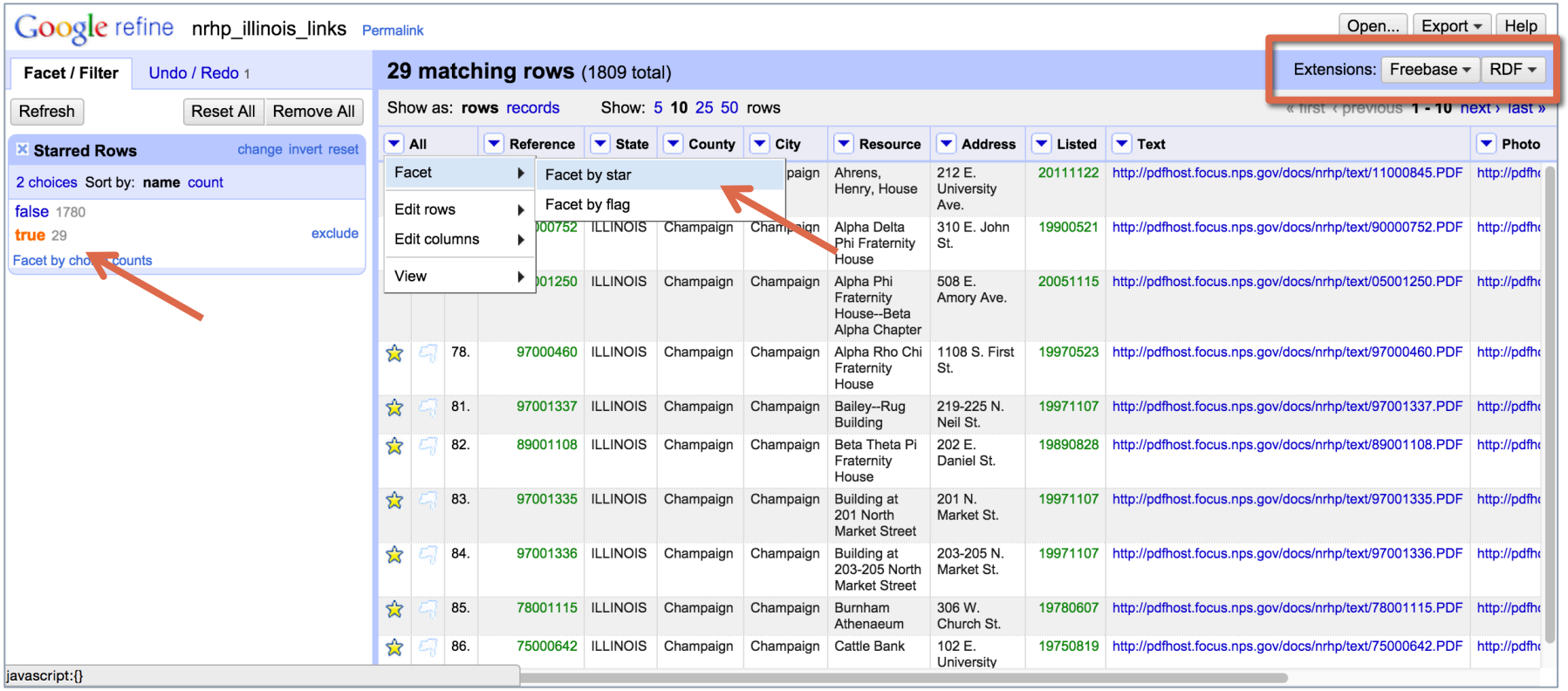


Add a Reconciliation Service



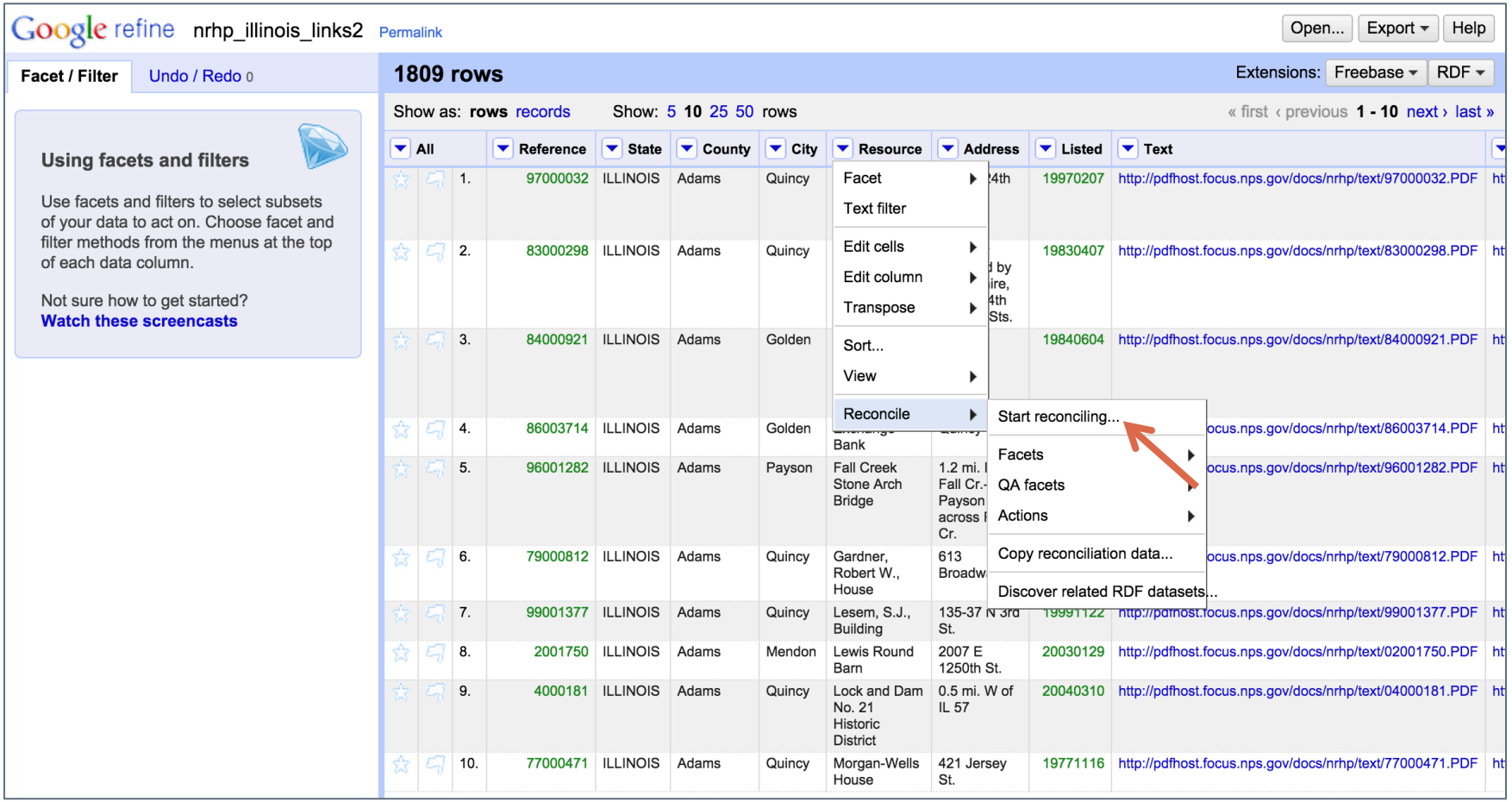
DBpedia extracts structured information from Wikipedia and makes it available as a SPARQL endpoint.

Start with Subset of Data

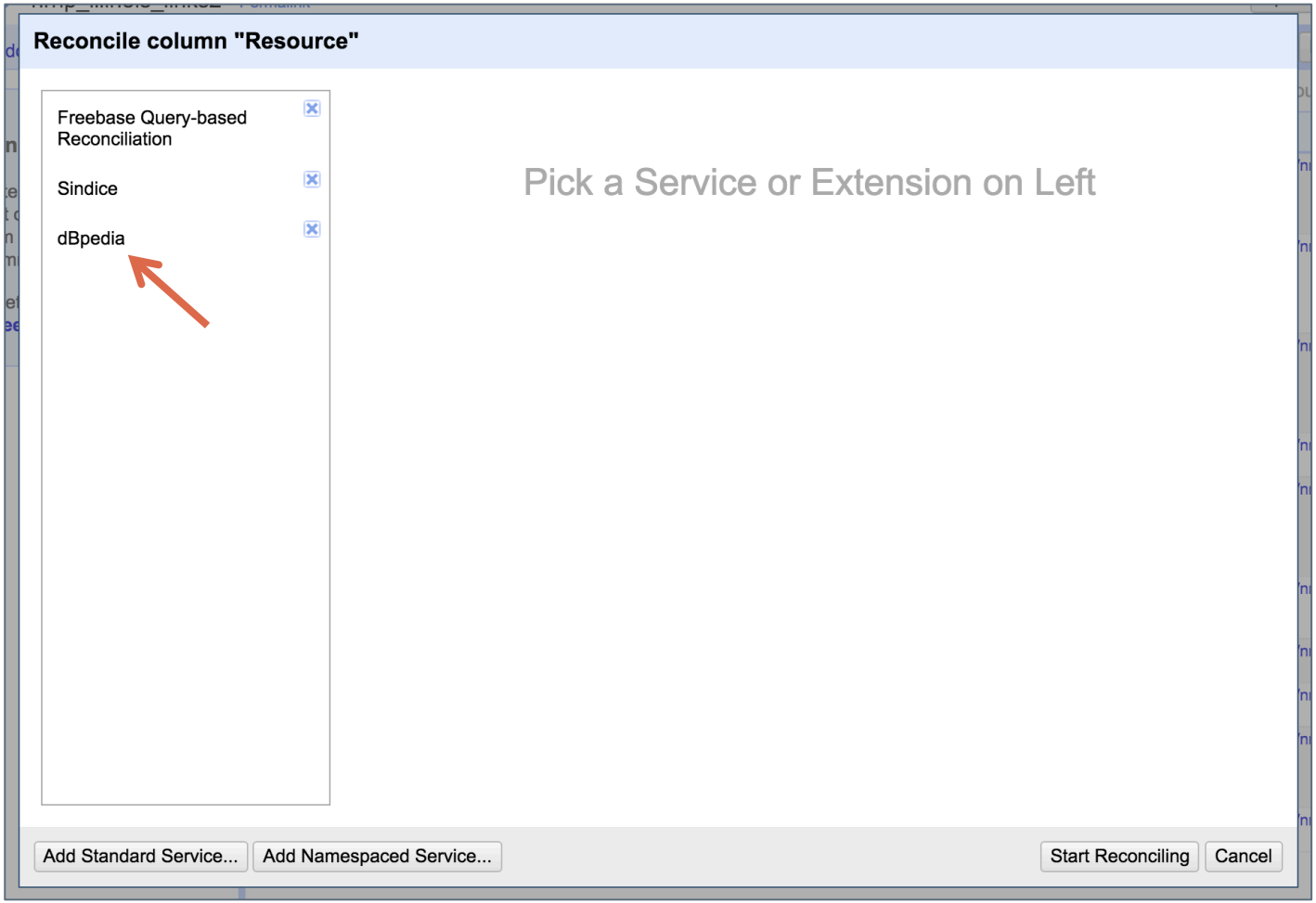


Star a small set of rows for for an initial test set, facet by star, and select “true”.

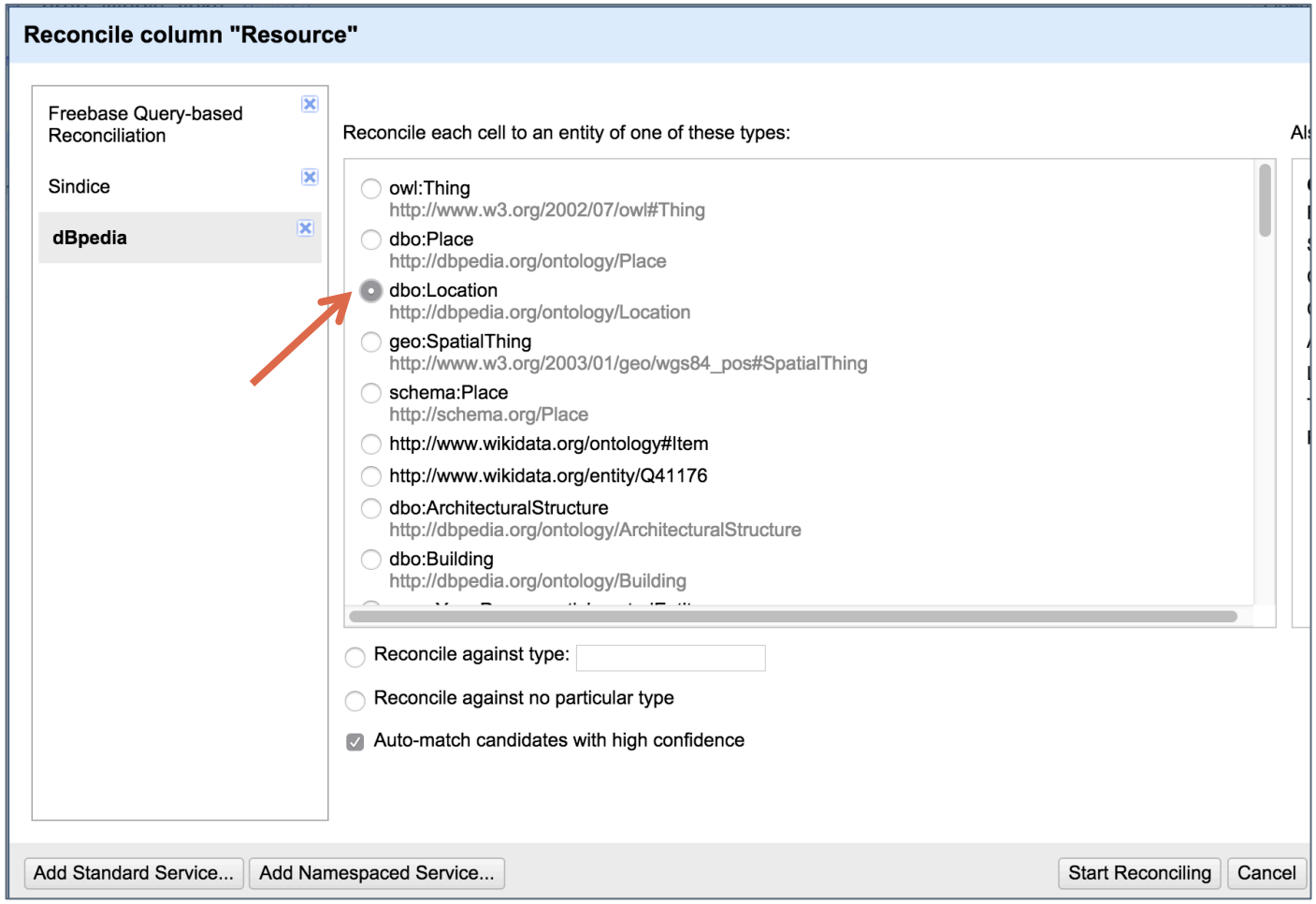
Choose which Column to Reconcile



Choose your Reconciliation Service

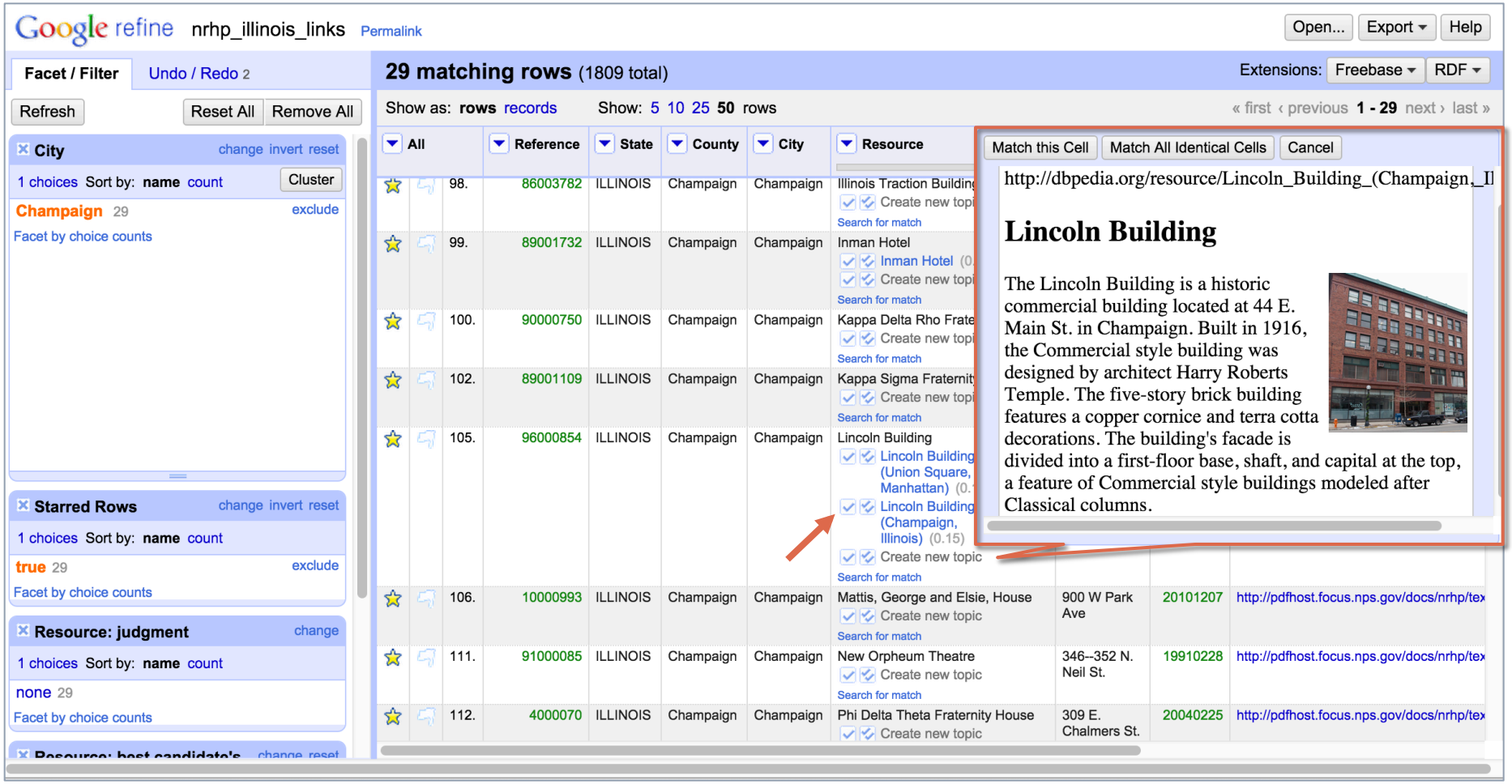


Choose a Type of Reconciliation

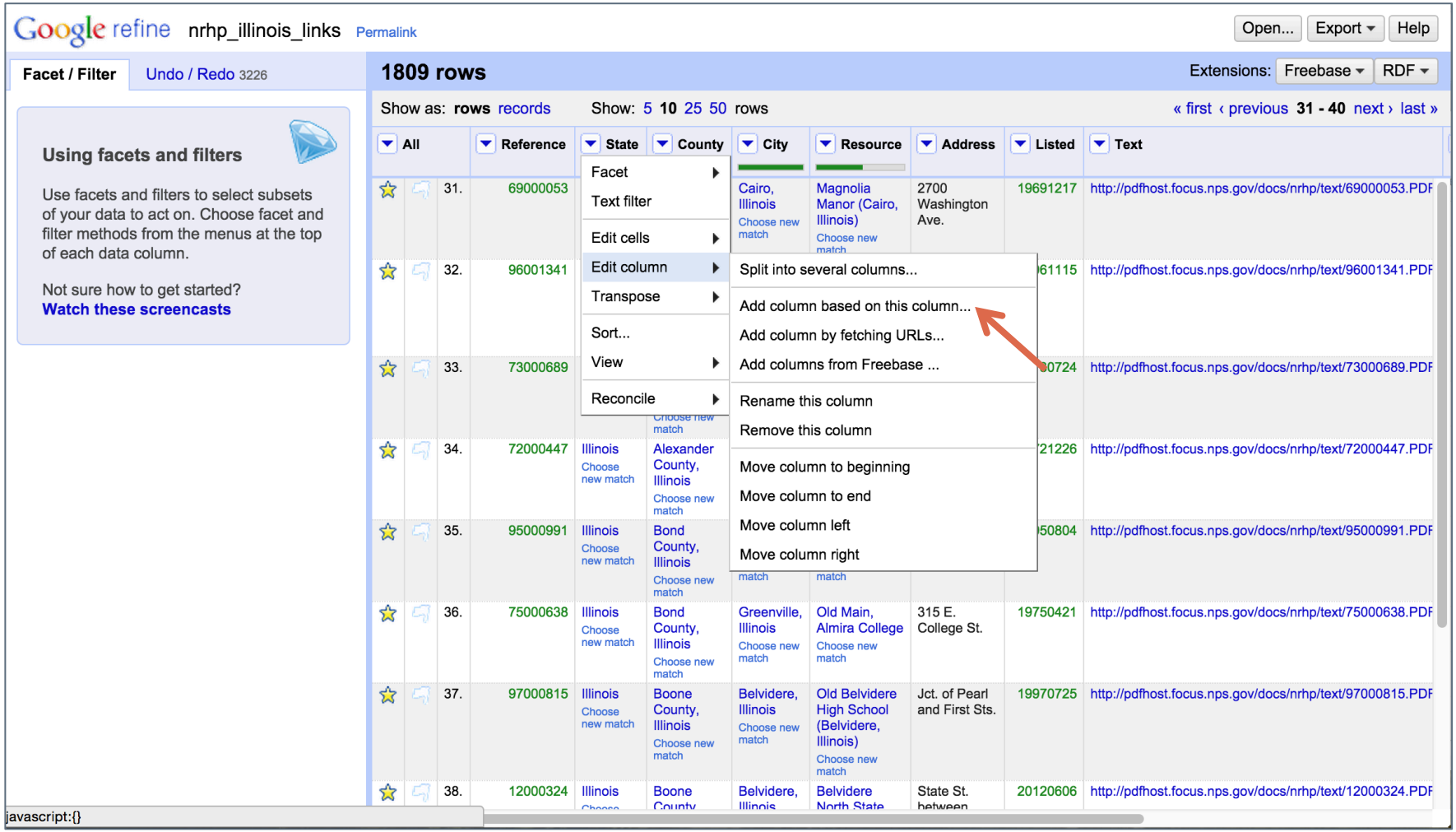


If you aren't sure which name space or class to use, you can usually find additional information about these ontologies online.

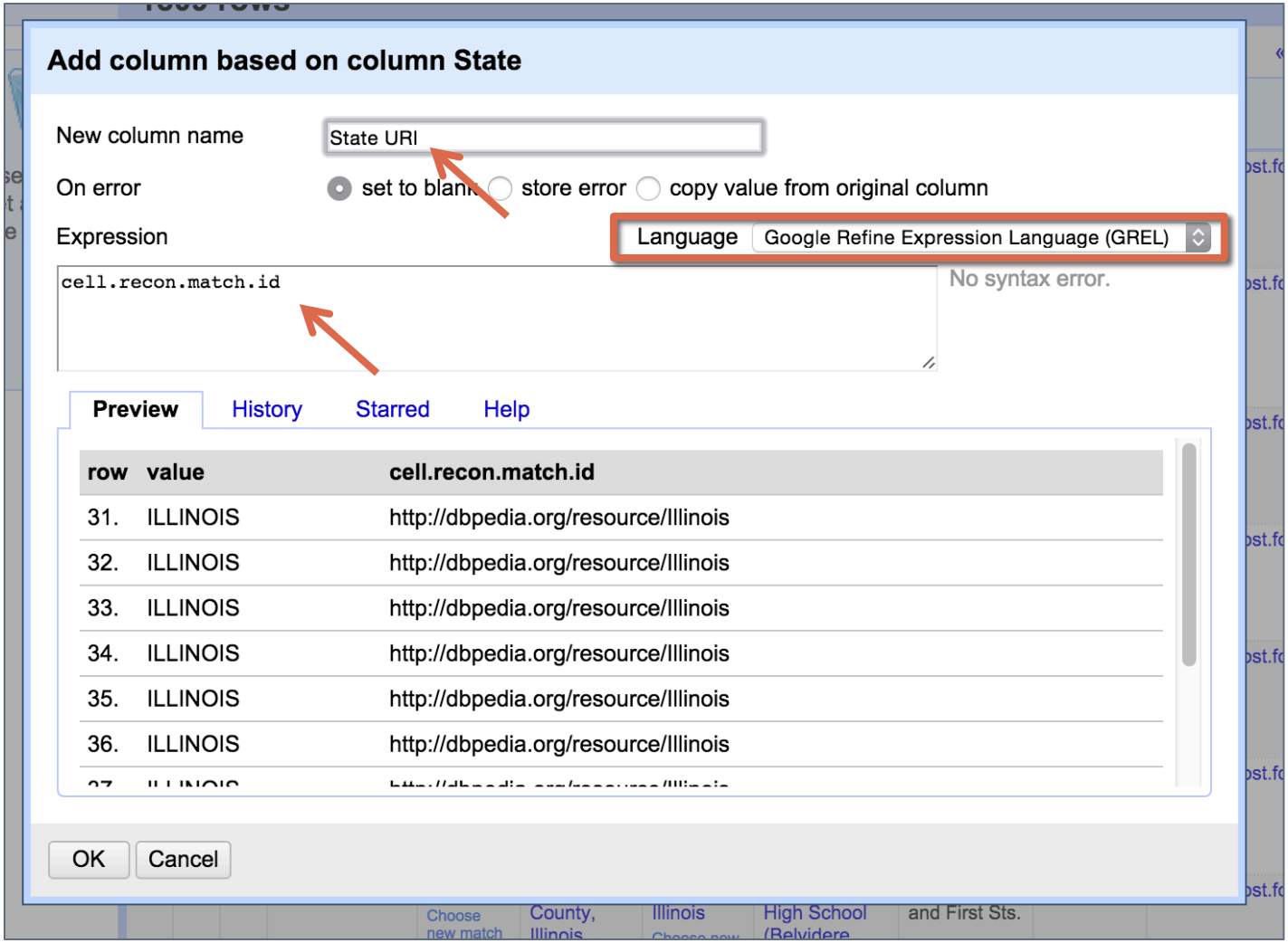
Review Results



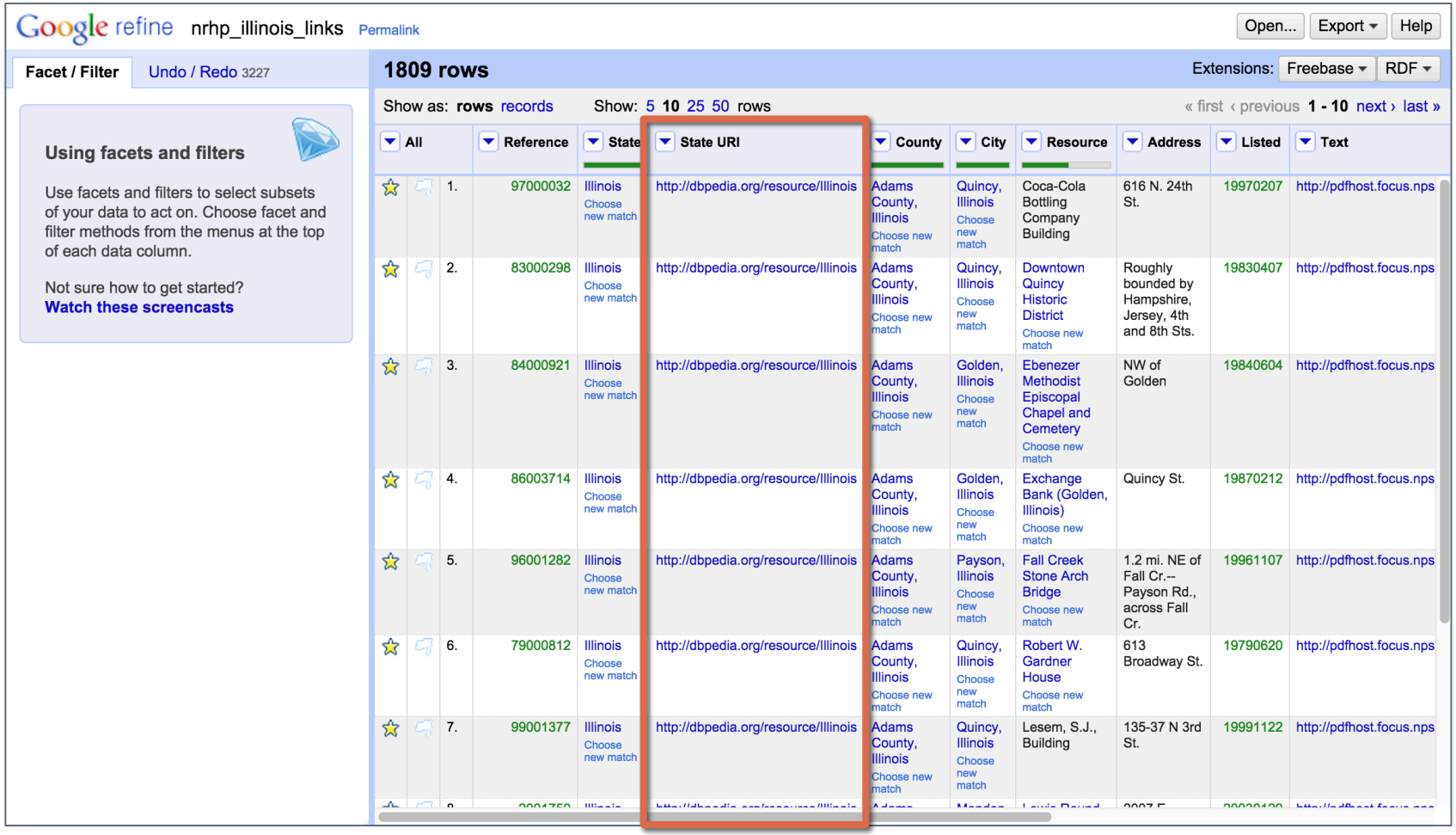
Create a Column for New URLs



Create a Column for New URLs

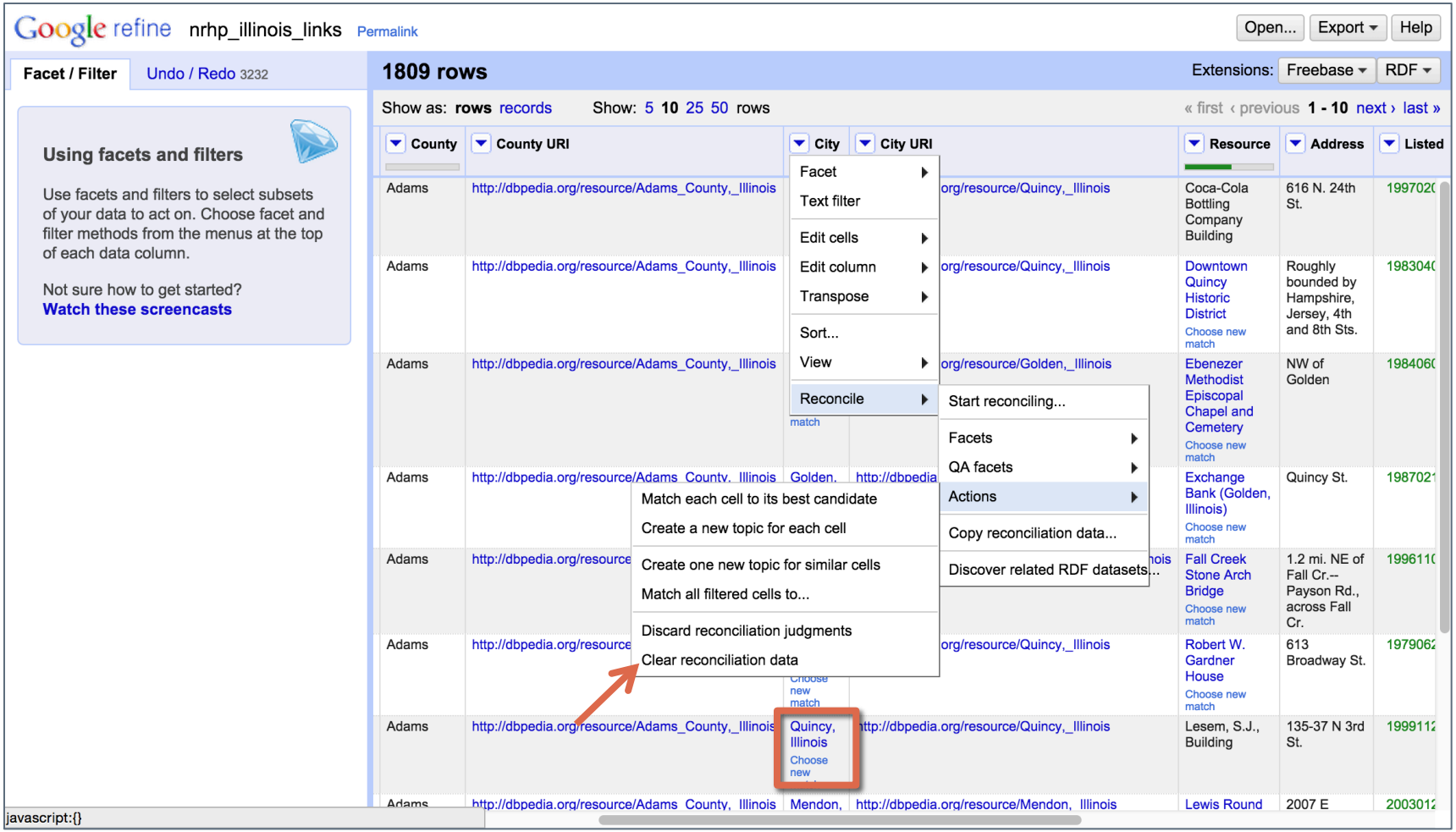


Review URI Column

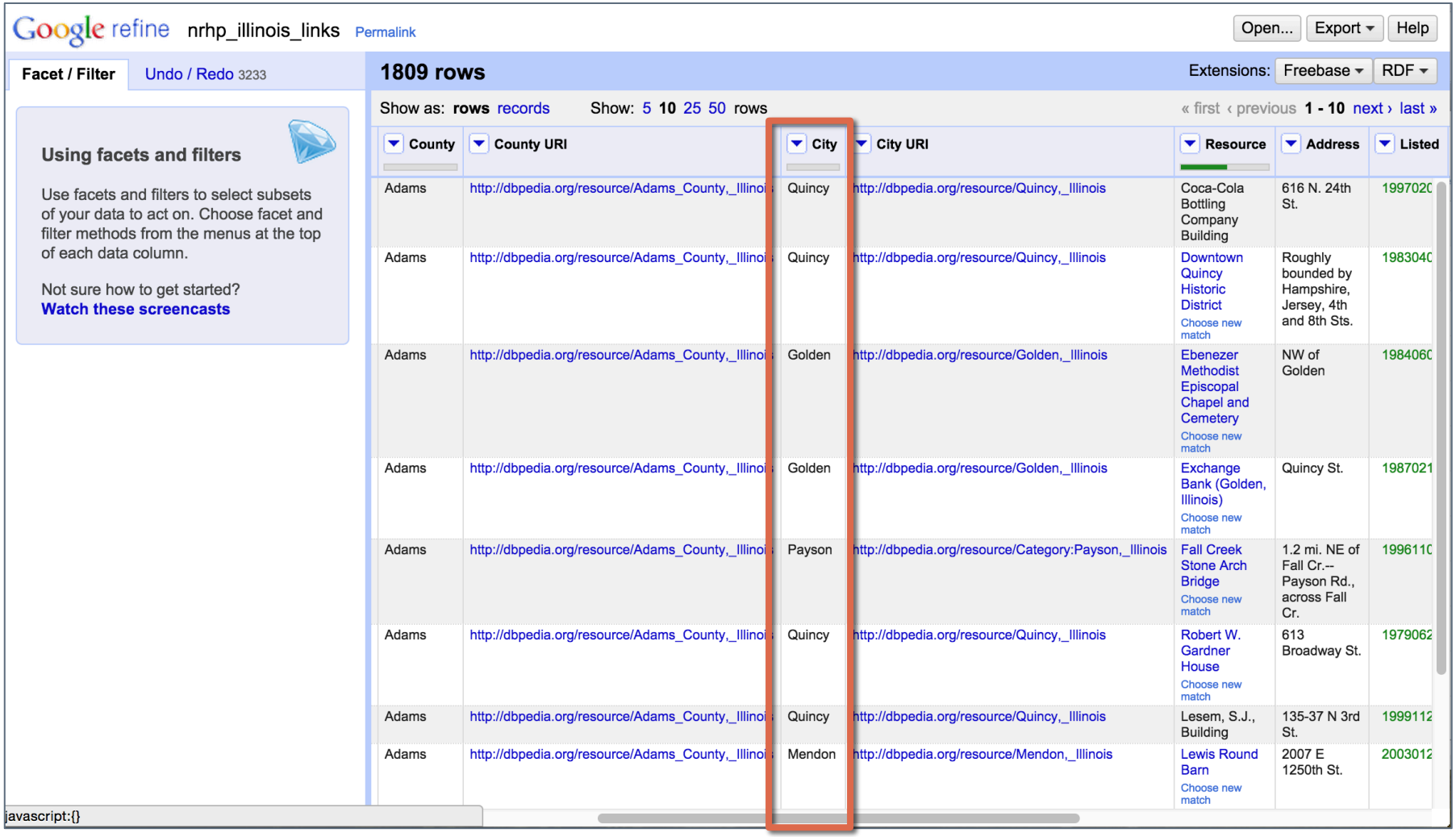


Each event cell now has a unique identifier and a path to discover related information from other data sets.

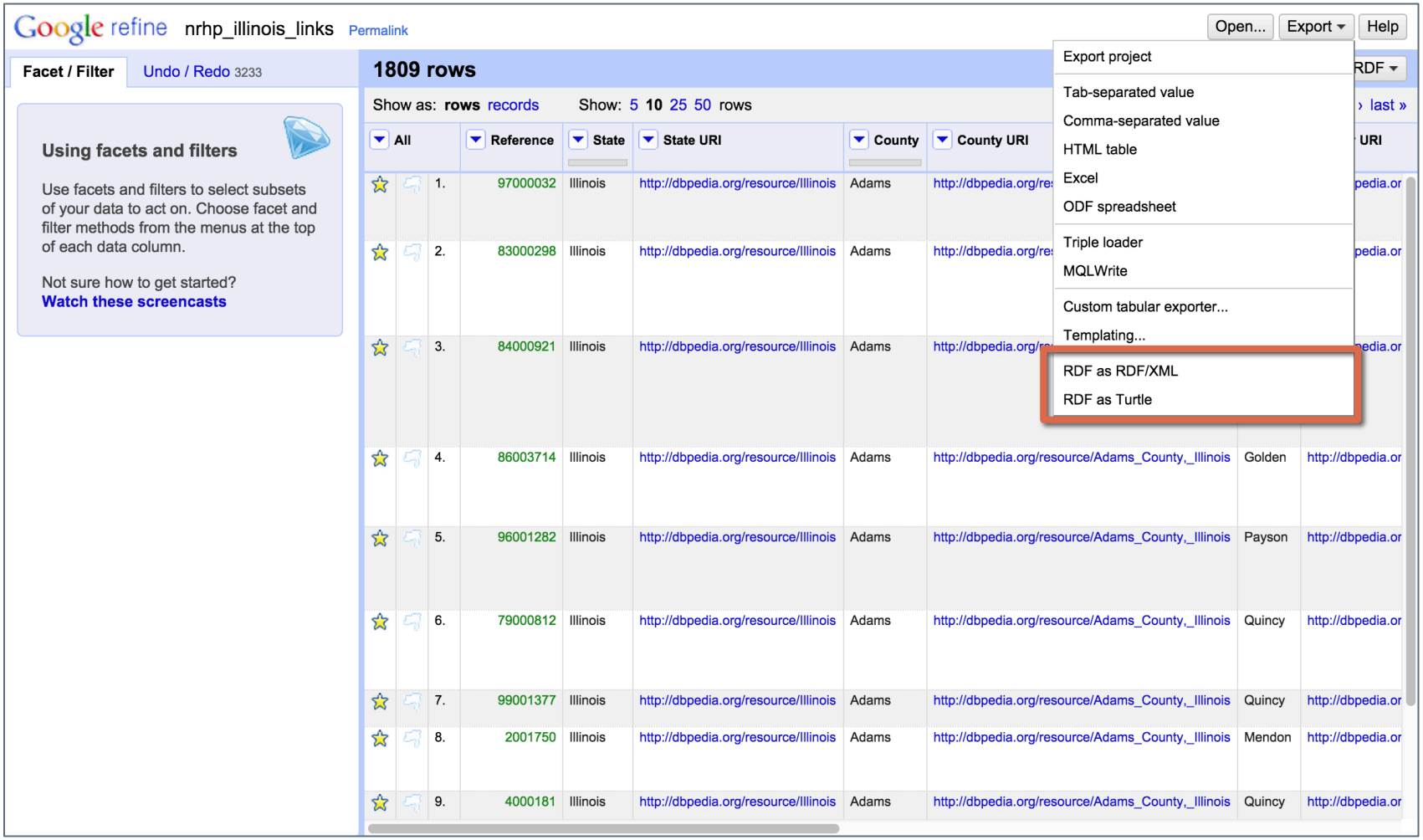
Clear Reconciliation Data



Clear Reconciliation Data

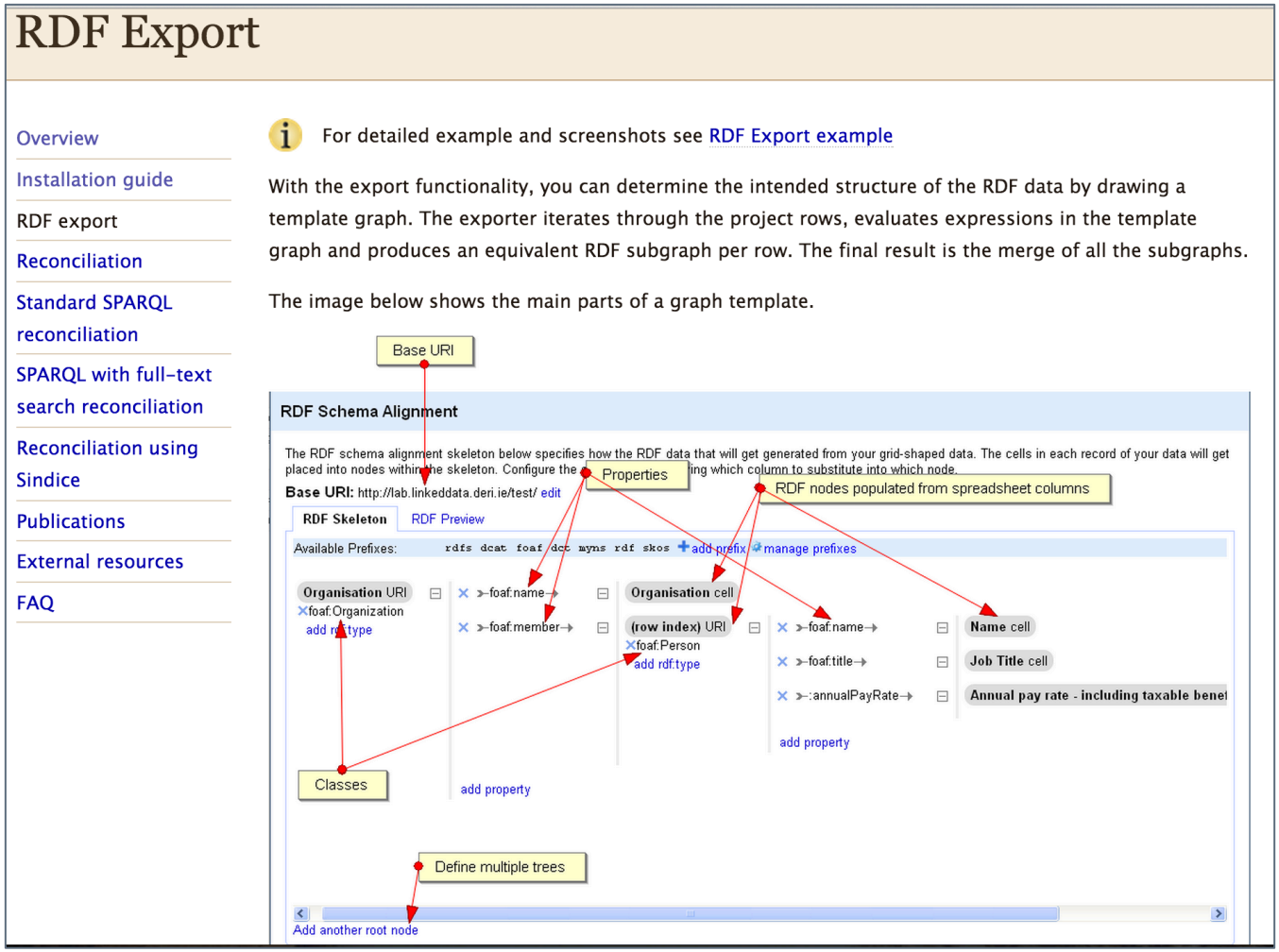


Exporting



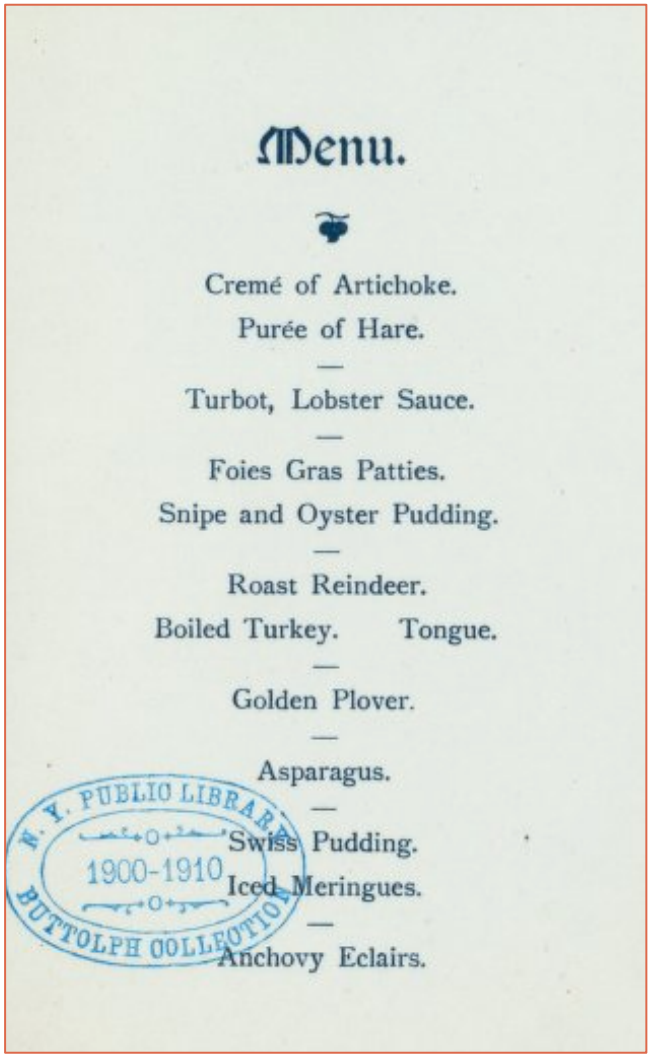
… not so fast!

Caveat on Exporting to RDF



http://refine.deri.ie/rdfExportDocs

Discussion and Questions



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