

# **Theory & Practice of Data Cleaning**

Introduction to OpenRefine

# A First Look at OpenRefine

- Creating a New Project
- Basic Normalization
- Different Facets (text, timeline, scatterplot)
- Clustering and Mass Edits
- Operation History: Provenance
- Separate videos:
  - Installing OpenRefine
  - Advanced Operations

# OpenRefine Overview

- OpenRefine is a power tool for data “wrangling”, specifically:
  - for getting an overview (exploring and “profiling”) data
  - for detecting and **cleaning** certain data errors
  - for transforming and linking data
- History:
  - Freebase Gridworks ... Google Refine ... OpenRefine

# Dataset Examples

- Working with two datasets:
  - USDA Directory of Farmers Markets
    - smaller, more curated (?) data
  - New York Public Library collection on historic restaurant menus
    - very “messy”, crowd-sourced data

# Example: USDA Farmers Market Data

USDA United States Department of Agriculture

About USDA Ask the Expert Contact Us En Español

Topics Programs and Services Newsroom Blog Site Map Glossary A-Z Index Advanced Search Help

You are here: Home / USDA Farmers Market

Promoting Local Food and Building Community

The USDA Farmers Market is the Department's own "living laboratory" for farmers market operations across the country. The market supports the local economy, increases marketing opportunities for farmers and small businesses, provides access to an assortment of local and regionally sourced products, and increases access to healthy, affordable food in the District of Columbia's Ward 2.

For 21 years the USDA Farmers Market has been brought to you by USDA's [Agricultural Marketing Service \(AMS\)](#), which supports farmers markets in communities across the country through grants, research, and technical assistance.

**About the USDA Farmers Market**

**Hours and Location**

Fridays, 9 a.m. to 2 p.m. (May 6 through October 28)  
Fridays, 4 p.m. to 7 p.m. (June 3 through September 30)  
Parking lot outside USDA Headquarters on the corner of Independence Avenue and 12th St, S.W., Washington, DC 20250.  
Nearest Metro: Smithsonian (Orange/Blue/Silver Line). For more public transportation options, see [www.wmata.com](#).

**2016 Day Market Vendors**

**Farmers and Growers:**

Apple Valley Orchards, Biglerville, PA  
C&T Produce, Fredericksburg, VA  
Diaz Berries and Fruits, Colonial Beach, VA  
King Mushrooms, Barclay, MD  
Little Wild Things City Farm, Washington, DC  
So Very Special, Frederick, MD

**Food Concessions:**

Bun'd Up  
Calvert Kettle Corn  
Dirty South Deli  
Eat 170  
Pinch  
Saison Wafel Bar

USDA Local Food Directories: Nation X Bertram

https://www.ams.usda.gov/local-food-directories/farmersmarkets

About AMS | News & Announcements | Careers | For Employees | Contact Us  
Advanced Search | A-Z Glossary & Index

Market News Rules & Regulations Grades & Standards Services Resources Selling Food to USDA

Home Stay connected:

## Local Food Directories: National Farmers Market Directory

The Farmers Market Directory lists markets that feature two or more farm vendors selling agricultural products directly to customers at a common, recurrent physical location. Maintained by the Agricultural Marketing Service, the Directory is designed to provide customers with convenient access to information about farmers market listings to include: market locations, directions, operating times, product offerings, accepted forms of payment, and more.

Visit our [Local Food Directories page](#) to find other operations offering locally grown products. If you are a market manager visit our [Local Food Directory Registration & Update page](#) to add or update a market listing. An [API](#) is available for developers to integrate this data into other applications.

Last update on November 10, 2016 11:57

Instructions

Search Near Products Available Payment Accepted Market Location Winter Markets State Contacts

Search near ZIP:  Distance: 5 miles   Map Markets

Info	MarketName	City	State	Website
<input type="checkbox"/>			All	
<input type="checkbox"/>	Caledonia Farmers Market Association - Danville	Danville	Vermont	<a href="#">View</a>
<input type="checkbox"/>	Stearns Homestead Farmers' Market	Parma	Ohio	<a href="#">View</a>
<input type="checkbox"/>	100 Mile Market	Kalamazoo	Michigan	<a href="#">View</a>
<input type="checkbox"/>	106 S. Main Street Farmers Market	Six Mile	South Carolina	<a href="#">View</a>
<input type="checkbox"/>	10th Street Community Farmers Market	Lamar	Missouri	<a href="#">View</a>
<input type="checkbox"/>	112st Madison Avenue	New York	New York	<a href="#">View</a>
<input type="checkbox"/>	12 South Farmers Market	Nashville	Tennessee	<a href="#">View</a>
<input type="checkbox"/>	125th Street Fresh Connect Farmers' Market	New York	New York	<a href="#">View</a>
<input type="checkbox"/>	12th & Brandywine Urban Farm Market	Wilmington	Delaware	<a href="#">View</a>
<input type="checkbox"/>	14th Street Farmers Market	Washington	District of Columbia	<a href="#">View</a>

Page 1 of 867 10 View 1 - 10 of 8,664

# Local Food Directories: National Farmers Market Directory

The Farmers Market Directory lists markets that feature two or more farm vendors selling agricultural products directly to customers at a common, recurrent physical location. Maintained by the Agricultural Marketing Service, the Directory is designed to provide customers with convenient access to information about farmers market listings to include: market locations, directions, operating times, product offerings, accepted forms of payment, and more.

Visit our [Local Food Directories page](#) to find other operations offering locally grown products. If you are a market manager visit our [Local Food Directory Registration & Update page](#) to add or update a market listing. An API is available for developers to integrate this data into other applications.

Last update on November 10, 2016 11:57

## Instructions

Search Near   Products Available   Payment Accepted   Market Location   Winter Markets   State Contacts

Search near ZIP:  Distance: 5 miles

 Map Markets

Info	MarketName	City	State	Website
<input type="checkbox"/>			All	
<input type="checkbox"/>	Caledonia Farmers Market Association - Danville	Danville	Vermont	<a href="#">View</a>
<input type="checkbox"/>	Stearns Homestead Farmers' Market	Parma	Ohio	<a href="#">View</a>
<input type="checkbox"/>	100 Mile Market	Kalamazoo	Michigan	<a href="#">View</a>
<input type="checkbox"/>	106 S. Main Street Farmers Market	Six Mile	South Carolina	<a href="#">View</a>
<input type="checkbox"/>	10th Street Community Farmers Market	Lamar	Missouri	
<input type="checkbox"/>	112st Madison Avenue	New York	New York	
<input type="checkbox"/>	12 South Farmers Market	Nashville	Tennessee	<a href="#">View</a>
<input type="checkbox"/>	125th Street Fresh Connect Farmers' Market	New York	New York	<a href="#">View</a>
<input type="checkbox"/>	12th & Brandywine Urban Farm Market	Wilmington	Delaware	
<input type="checkbox"/>	140th Street Farmers Market	New York	New York	<a href="#">View</a>

[Export to Excel](#) 

Page 1 of 867 | [<<](#) [>>](#) [10](#) | View 1 - 10 of 8,664

# OpenRefine: Create Project

The screenshot shows the OpenRefine web application running at `127.0.0.1:3333`. The left sidebar has a blue outline around the 'Create Project' option. The main content area is titled 'Create a project by importing data. What kinds of data files can I import?'. It lists supported formats: TSV, CSV, \*SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents. Below this, there's a section for 'Get data from' with options: 'This Computer' (selected), 'Web Addresses (URLs)', 'Clipboard', and 'Google Data'. A 'Choose Files' button and a message 'No file chosen' are visible. A 'Next »' button is at the bottom of this section. The bottom left of the sidebar shows a diamond icon and the text 'Version 2.6-rc.2 [TRUNK]'. At the very bottom of the sidebar are 'Help' and 'About' links.

# Importing Data ...

The screenshot shows the OpenRefine interface with a red arrow pointing to the "Project name" field in the top right corner. Another red arrow points to the "Parse data as" dropdown menu on the left, which is set to "CSV / TSV / separator-based files". A third red arrow points to the "Configure Parsing Options" section on the right, specifically highlighting the "Parse next 1 line(s) as column headers" checkbox.

A power tool for working with messy data.

Project name Export csv Create Project »

	FMID	MarketName	Website	Facebook	Twitter	Youtube
1.	1012063	Caledonia Farmers Market Association - Danville	<a href="https://sites.google.com/site/caledoniafarmersmarket/">https://sites.google.com/site/caledoniafarmersmarket/</a>	<a href="https://www.facebook.com/Danville.VT.Farmers.Market/">https://www.facebook.com/Danville.VT.Farmers.Market/</a>		
2.	1011871	Stearns Homestead Farmers' Market	<a href="http://StearnsHomestead.com">http://StearnsHomestead.com</a>			
3.	1011878	100 Mile Market	<a href="http://www.pfcmarkets.com">http://www.pfcmarkets.com</a>	<a href="https://www.facebook.com/100MileMarket/?fref=ts">https://www.facebook.com/100MileMarket/?fref=ts</a>		
4.	1009364	106 S. Main Street Farmers Market	<a href="http://thetownofsixmile.wordpress.com/">http://thetownofsixmile.wordpress.com/</a>			
5.	1010691	10th Street Community				

Parse data as

Character encoding

Update Preview

CSV / TSV / separator-based files

Line-based text files

Fixed-width field text files

PC-Axis text files

JSON files

MARC files

RDF/N3 files

XML files

Open Document Format spreadsheets (.ods)

Version 2.6-rc.2 [TRUNK]

Help About

Columns are separated by  commas (CSV)  tabs (TSV)  custom ,

Escape special characters with \

Ignore first 0 line(s) at beginning of file  
 Parse next 1 line(s) as column headers  
 Discard initial 0 row(s) of data  
 Load at most 0 row(s) of data

Parse cell text into numbers, dates, ...  
 Quotation marks are used to enclose cells containing column separators

Store blank rows  
 Store blank cells as nulls  
 Store file source (file names, URLs) in each row

# Voilà! 8664 rows imported ...

Screenshot of the OpenRefine interface showing the 'Farmers-Markets' project. A red arrow points to the status bar at the top center which displays '8664 rows'.

The interface includes a sidebar titled 'Using facets and filters' with a 'Watch these screencasts' link. The main area shows a table with 10 rows of data, each containing fields like FMID, MarketName, Website, Facebook, Twitter, and YouTube links. The table has a header row with dropdown menus for facet/filtering.

FMID	MarketName	Website	Facebook	Twitter	Youtube
1. 1012063	Caledonia Farmers Market Association - Danville	<a href="https://sites.google.com/site/caledoniafarmersmarket/">https://sites.google.com/site/caledoniafarmersmarket/</a>	<a href="https://www.facebook.com/Danville.VT.Farmers.Market">https://www.facebook.com/Danville.VT.Farmers.Market</a>		
2. 1011871	Stearns Homestead Farmers' Market	<a href="http://StearnsHomestead.com">http://StearnsHomestead.com</a>			
3. 1011878	100 Mile Market	<a href="http://www.pfcmarkets.com">http://www.pfcmarkets.com</a>	<a href="https://www.facebook.com/100MileMarket/?fref=ts">https://www.facebook.com/100MileMarket/?fref=ts</a>		
4. 1009364	106 S. Main Street Farmers Market	<a href="http://thetownofsixmile.wordpress.com/">http://thetownofsixmile.wordpress.com/</a>			
5. 1010691	10th Street Community Farmers Market				
6. 1002454	112st Madison Avenue				
7. 1011100	12 South Farmers Market	<a href="http://www.12southfarmersmarket.com">http://www.12southfarmersmarket.com</a>	12_South_Farmers_Market	@12southfrmsmkt	@
8. 1009845	125th Street Fresh Connect Farmers' Market	<a href="http://www.125thStreetFarmersMarket.com">http://www.125thStreetFarmersMarket.com</a>	<a href="https://www.facebook.com/125thStreetFarmersMarket">https://www.facebook.com/125thStreetFarmersMarket</a>	<a href="https://twitter.com/FarmMarket125th">https://twitter.com/FarmMarket125th</a>	Ins
9. 1005586	12th & Brandywine Urban Farm Market		<a href="https://www.facebook.com/pages/12th-Brandywine-Urban-Farm-Community-Garden/253769448091860">https://www.facebook.com/pages/12th-Brandywine-Urban-Farm-Community-Garden/253769448091860</a>		
10. 1008071	14&U Farmers' Market		<a href="https://www.facebook.com/14UFarmersMarket">https://www.facebook.com/14UFarmersMarket</a>	<a href="https://twitter.com/14UFarmersMkt">https://twitter.com/14UFarmersMkt</a>	

# The Text Facet “workhorse” ...

The screenshot shows the OpenRefine interface with the following details:

- Facet / Filter:** A facet for **MarketName** is displayed on the left, listing 8095 choices sorted by name and count. A red box highlights this facet area, and a red arrow points from the text "Now hit 'Cluster'" to the "Cluster" button at the bottom right of the facet panel.
- Table View:** The main area displays 8664 rows of data. The columns include:
  - MarketName:** Caledonia Farmers Market Association - Danville, Stearns Homestead Farmers' Market, 100 Mile Market, 106 S. Main Street Farmers Market, 10th Street Community Farmers Market, 112st Madison Avenue, 12 South Farmers Market, 125th Street Fresh Connect Farmers' Market, 12th & Brandywine Urban Farm Market, and 14&U Farmers' Market.
  - FMID:** 1012063, 1011871, 1011872, 1011873, 1011874, 1010691, 1002454, 1011100, 1009845, 1005586, and 1008071.
  - Website:** https://sites.google.com/site/caledoniafarmersmarket/, http://StearnsHomestead.com, http://www.pfcmarkets.com, http://thetownofsixmile.wordpress.com/, etc.
  - Facebook:** https://www.facebook.com/Danville.VT.Farmers.Market/, https://www.facebook.com/100MileMarket/?fref=ts, https://www.facebook.com/12SouthFarmersMarket, https://www.facebook.com/125thStreetFarmersMarket, https://www.facebook.com/pages/12th-Brandywine-Urban-Farm-Community-Garden/253769448091860, https://www.facebook.com/14UFarmersMarket, etc.
  - Twitter:** https://www.facebook.com/Danville.VT.Farmers.Market/, https://www.facebook.com/100MileMarket/?fref=ts, https://www.facebook.com/12SouthFarmersMarket, https://www.facebook.com/125thStreetFarmersMarket, https://www.facebook.com/pages/12th-Brandywine-Urban-Farm-Community-Garden/253769448091860, https://www.facebook.com/14UFarmersMarket, etc.
  - Youtube:** https://www.facebook.com/Danville.VT.Farmers.Market/, https://www.facebook.com/100MileMarket/?fref=ts, https://www.facebook.com/12SouthFarmersMarket, https://www.facebook.com/125thStreetFarmersMarket, https://www.facebook.com/pages/12th-Brandywine-Urban-Farm-Community-Garden/253769448091860, https://www.facebook.com/14UFarmersMarket, etc.
  - Other:** https://www.facebook.com/Danville.VT.Farmers.Market/, https://www.facebook.com/100MileMarket/?fref=ts, https://www.facebook.com/12SouthFarmersMarket, https://www.facebook.com/125thStreetFarmersMarket, https://www.facebook.com/pages/12th-Brandywine-Urban-Farm-Community-Garden/253769448091860, https://www.facebook.com/14UFarmersMarket, etc.
- Toolbar:** Includes buttons for Undo / Redo, Refresh, Reset All, Remove All, Open..., Export, and Help.
- Header:** Shows the project name "Farmers-Markets" and the URL "127.0.0.1:3333/project?project=1766664496116".
- Page Number:** "Bertram" is shown in the top right corner.

Now hit  
“Cluster” ...

# ... and the magic happens!

Farmers-Markets - OpenRefine

127.0.0.1:3333/project?project=1766664496116

Refine

Facet / Filter Undo / Redo

MarketName

8095 choices Sort by: name count

Caledonia Farmers Market Association - Danville 1  
Stearns Homestead Farmers' Market 1  
100 Mile Market 1  
106 S. Main Street Farmers Market 1  
10th Street Community Farmers Market 1  
112st Madison Avenue 1  
12 South Farmers Market 1  
125th Street Fresh Connect Farmers' Market 1

Cluster & Edit column "MarketName"

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 226 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
4	12	<ul style="list-style-type: none"><li>Main Street Farmers Market (9 rows)</li><li>MAIN STREET FARMERS MARKET (1 rows)</li><li>Main Street Farmer's Market (1 rows)</li><li>Main Street Farmers' Market (1 rows)</li></ul>	<input type="checkbox"/>	Main Street Farmers Market
4	5	<ul style="list-style-type: none"><li>Irvington Farmers Market (2 rows)</li><li>Irvington Farmer's Market (1 rows)</li><li>Irvington Farmers Market (1 rows)</li><li>Irvington Farmers' Market (1 rows)</li></ul>	<input type="checkbox"/>	Irvington Farmers Market
3	3	<ul style="list-style-type: none"><li>Wakefield Farmer's Market (1 rows)</li><li>Wakefield Farmers Market (1 rows)</li><li>Wakefield Farmers Market (1 rows)</li></ul>	<input type="checkbox"/>	Wakefield Farmer's Market
3	4	<ul style="list-style-type: none"><li>Columbus Farmers' Market (2 rows)</li><li>Columbus Farmers Market (1 rows)</li><li>columbus farmers market (1 rows)</li></ul>	<input type="checkbox"/>	Columbus Farmers' Market
3	3	<ul style="list-style-type: none"><li>WATERTOWN FARMERS MARKET (1 rows)</li><li>Watertown Farmers market (1 rows)</li><li>Watertown Farmers' Market (1 rows)</li></ul>	<input type="checkbox"/>	WATERTOWN FARMERS M
3	5	<ul style="list-style-type: none"><li>Rochester Downtown Farmers Market (3 rows)</li><li>Rochester Farmers Market (1 rows)</li><li>Rochester Farmers Market (1 rows)</li><li>Rochester Farmers Market (1 rows)</li></ul>	<input type="checkbox"/>	Rochester Downtown Farme

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

# Choices in Cluster

# Rows in Cluster

Average Length of Choices

Length Variance of Choices

Bertram

Extensions: 1 - 10 next last

Youtube Other

https://v http://ag type=m @12so rmMarket125th Instagram https://v UFarmersMkt

# ... select some (all!?) clusters and merge ...

Screenshot of the OpenRefine interface showing the "Cluster & Edit column 'MarketName'" tool.

The left sidebar shows facets for "MarketName" with 8095 choices, sorted by name count. Choices include "El Mercado Familiar", "Main Street Farmers Market", "Winter Farmers Market and Mea for Hope", etc.

The main panel displays a list of clusters found (226 total). Each cluster entry shows the cluster size, row count, values in the cluster, and options to merge them into a new cell value. A red arrow points to the "Select All" button at the bottom left of the cluster table.

Below the cluster table are four histograms analyzing the data:

- # Choices in Cluster: Range 2 - 4
- # Rows in Cluster: Range 2 - 34
- Average Length of Choices: Range 13 - 71
- Length Variance of Choices: Range 0 - 2.5

At the bottom right of the main panel, a red arrow points to the "Merge Selected & Re-Cluster" button.

Bottom navigation bar: Select All, Unselect All, Merge Selected & Re-Cluster, Merge Selected & Close, Close.

# ... resulting in a mass edit ...

The screenshot shows the OpenRefine interface for a project titled "Farmers-Markets". The main view displays a table with 8664 rows, filtered by the "MarketName" column. A red arrow points from the text ".. also reduced the choices from 8095 to 7846..." at the bottom to the facet panel on the left, which lists "7846 choices" for the "MarketName" facet. Another red arrow points from the text "Mass edit 659 cells in column MarketName" at the top to the status bar above the table, which also shows "Mass edit 659 cells in column MarketName". The table columns include FID, FMID, MarketName, Website, Facebook, Twitter, and YouTube.

FID	FMID	MarketName	Website	Facebook	Twitter	YouTube
1.	1012063	Caledonia Farmers Market Association - Danville	<a href="https://sites.google.com/site/caledoniafarmersmarket/">https://sites.google.com/site/caledoniafarmersmarket/</a>	<a href="https://www.facebook.com/Danville.VT.Farmers.Market/">https://www.facebook.com/Danville.VT.Farmers.Market/</a>		
2.	1011871	Stearns Homestead Farmers' Market	<a href="http://StearnsHomestead.com">http://StearnsHomestead.com</a>			
3.	1011878	100 Mile Market	<a href="http://www.pfcmarkets.com">http://www.pfcmarkets.com</a>	<a href="https://www.facebook.com/100MileMarket/?fref=ts">https://www.facebook.com/100MileMarket/?fref=ts</a>		<a href="https://www.youtube.com/watch?v=...">https://www.youtube.com/watch?v=...</a>
4.	1009364	106 S. Main Street Farmers Market	<a href="http://thetownofsixmile.wordpress.com/">http://thetownofsixmile.wordpress.com/</a>			
5.	1010691	10th Street Community Farmers Market				<a href="http://...">http://... type=m..."&gt;http://... type=m...</a>
6.	1002454	112st Madison Avenue				
7.	1011100	12 South Farmers Market	<a href="http://www.12southfarmersmarket.com">http://www.12southfarmersmarket.com</a>	12_South_Farmers_Market	@12southfrmsmkt	@12sou...
8.	1009845	125th Street Fresh Connect Farmers' Market	<a href="http://www.125thStreetFarmersMarket.com">http://www.125thStreetFarmersMarket.com</a>	<a href="https://www.facebook.com/125thStreetFarmersMarket">https://www.facebook.com/125thStreetFarmersMarket</a>	<a href="https://twitter.com/FarmMarket125th">https://twitter.com/FarmMarket125th</a>	Instagram
9.	1005586	12th & Brandywine Urban Farm Market		<a href="https://www.facebook.com/pages/12th-Brandywine-Urban-Farm-Community-Garden/253769448091860">https://www.facebook.com/pages/12th-Brandywine-Urban-Farm-Community-Garden/253769448091860</a>		<a href="https://www.youtube.com/watch?v=...">https://www.youtube.com/watch?v=...</a>
10.	1008071	14&U Farmers' Market		<a href="https://www.facebook.com/14UFarmersMarket">https://www.facebook.com/14UFarmersMarket</a>	<a href="https://twitter.com/14UFarmersMkt">https://twitter.com/14UFarmersMkt</a>	

*.. also reduced the choices from 8095 to 7846...*

# ... (in this case): Done with Normalization of MarketName column

Farmers-Markets - OpenRefine X Bertram

127.0.0.1:3333/project?project=1766664496116

Refine OPEN

Facet / Filter Undo / Redo

Refresh Reset A

MarketName

7846 choices Sort by: name count

Caledonia Farmers Market  
Association - Danville 1  
Stearns Homestead Farmers'  
Market 1  
100 Mile Market 1  
106 S. Main Street Farmers'  
Market 1  
10th Street Community Farmers'  
Market 1  
112st Madison Avenue 1  
12 South Farmers Market 1  
125th Street Fresh Connect  
Farmers' Market 1

Cluster & Edit column "MarketName"

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

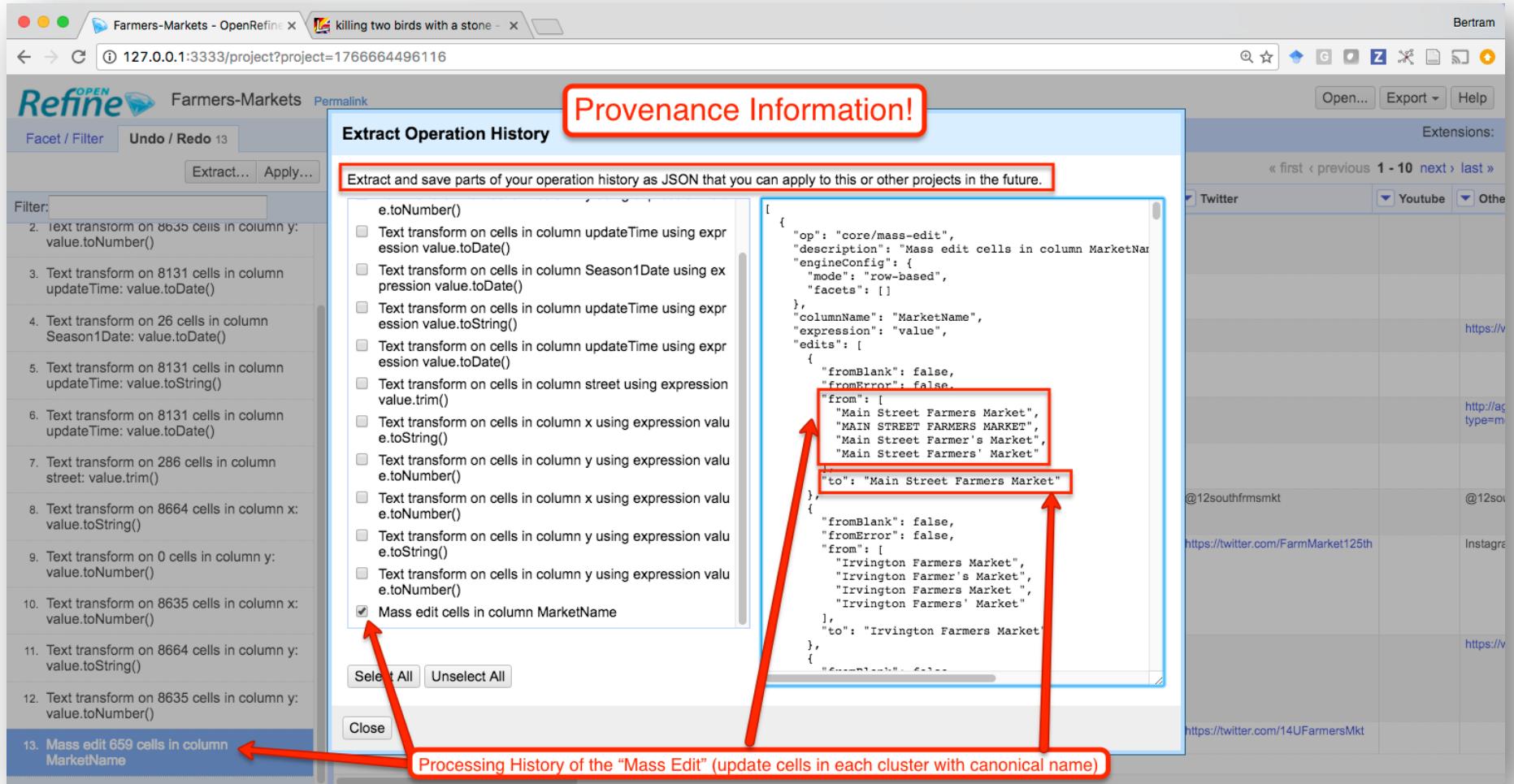
No clusters were found with the selected method

Try selecting another method above or changing its parameters

... or more precisely: all "clusters" have now size 1,  
i.e., we have normalized the names in the MarketName column!

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

# Undo/Redo: Operation History *(Provenance)*



# More Data Profiling: Timeline Facet

The screenshot shows the OpenRefine interface with the title "Farmers-Markets - OpenRefine x New Tab". The URL in the address bar is "127.0.0.1:3333/project?project=1766664496116". The main area displays 8664 rows of data with columns: Wine, Coffee, Beans, Fruits, Grains, Juices, Mushrooms, PetFood, Tofu, WildHarvested, and updateTime. A red arrow points from the "updateTime" column header to a context menu that is open over the data. The menu options include: Text facet, Numeric facet, Timeline facet (which is highlighted in blue), Scatterplot facet, Custom text facet..., Custom Numeric Facet..., and Customized facets. Another red arrow points from the "Timeline facet" option in the menu to the "Timeline facet" link in the "Extensions" sidebar.

# Timeline facet: hmm ... not working!?

The screenshot shows the OpenRefine interface for a project titled "Farmers-Markets". The main view displays 8664 rows of data across various columns labeled "N/A", "Coffee", "Beans", "Fruits", "Grains", "Juices", "Mushrooms", and "PetF".

In the "Facet / Filter" panel on the left, there is a facet for "updateTime" with the following settings:

- Value: "NaN-NaN-NaN" (highlighted with a red arrow)
- Format: "NaN:NaN:NaN — 18:00:00"
- Count: 8664 (Time) and 0 (Non-Time)
- Blank: 0
- Error: 0

The "Show as" dropdown is set to "rows" and the "Show" dropdown is set to "5 10 25 50 rows".

# Converting from String to Date!

The screenshot shows the OpenRefine interface for a project titled "Farmers-Markets". The main view displays 8664 rows of data across various columns like Wine, Coffee, Beans, Fruits, Grains, Juices, Mushrooms, PetFood, Tofu, and WildHarvested. A context menu is open over a cell in the WildHarvested column, listing various data manipulation options. Red arrows point to three specific items: "Edit cells" (the top item in the main menu), "Common transforms" (under the "Transform..." submenu), and "To date" (under the "Common transforms" submenu). The "To date" option is highlighted with a large red arrow pointing directly at it. The status bar at the bottom right indicates the current row number is 15.

Facet / Filter Undo / Redo 5

Refresh Reset All Remove All

updateTime change reset

NaN-NaN-NaN NaN:NaN:NaN — 18:00:00

Time Non-Time Blank Error  
0 8664 0 0

8664 rows

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

Wine Coffee Beans Fruits Grains Juices Mushrooms PetFood Tofu WildHarvested updateTime

Y Y N Y N Y N Y N

N N Y N N N Transform... Common transforms

- Trim leading and trailing whitespace
- Collapse consecutive whitespace
- Unescape HTML entities
- To titlecase
- To uppercase
- To lowercase
- To number
- To date**
- To text
- Blank out cells

Facet Text filter

Edit cells Edit column Transpose Sort... View Reconcile

8/2016 10:09 PM 9, 2016 15, 2016 1, 2013 28, 2014 Mar 1, 2012 May 1, 2015 Apr 7, 2014 Apr 3, 2014

# .. now we're in business!

Screenshot of the OpenRefine interface showing a project titled "Farmers-Markets".

The interface includes:

- Toolbar with "New Tab", "Bertram", and various icons.
- Address bar: `127.0.0.1:3333/project?project=1766664496116`.
- Header: "Refine OPEN Farmers-Markets Permalink".
- Facet / Filter panel:
  - Facet for "updateTime" showing a histogram from 2009-01-01 to 2016-04-07.
  - Counts for Time (8131), Non-Time (533), Blank (0), and Error (0).
- Text transform message: "Text transform on 8131 cells in column updateTime: value.toDate() Undo" with a yellow background and a red arrow pointing to the "Undo" button.
- Table view showing 8664 rows of data across 11 columns: Wine, Coffee, Beans, Fruits, Grains, Juices, Mushrooms, PetFood, Tofu, WildHarvested, and updateTime.
- Table header: "Show as: rows records Show: 5 10 25 50 rows".
- Table footer: "1 - 10 next > last »".
- Extensions menu: "Extensions: 1 - 10 next > last »".

Wine	Coffee	Beans	Fruits	Grains	Juices	Mushrooms	PetFood	Tofu	WildHarvested	updateTime
Y	Y	Y	N	Y	N	Y	N	N	6/28/2016 12:10:09 PM	
N	N	Y	N	N	N	Y	N	N	2016-04-09T00:00:00Z	
N	N	Y	N	N	N	N	N	N	2016-07-15T00:00:00Z	
									2013-01-01T00:00:00Z	
N	N	Y	N	N	N	N	N	N	2014-10-28T00:00:00Z	
N	N	N	N	N	N	N	N	N	2012-03-01T00:00:00Z	
Y	N	Y	N	Y	Y	Y	N	N	2015-05-01T00:00:00Z	
Y	N	Y	N	Y	N	N	N	N	2014-04-07T00:00:00Z	

# Exploring time slices: missing data ...

The screenshot shows the OpenRefine interface for a project titled "Farmers-Markets". The left sidebar displays a timeline facet for the column "updateTime", with a red arrow pointing to the date "2010-11-01". The timeline shows several data points, with the first one explicitly labeled "2010-11-01 15:40:12 — 08:36:36". Below the timeline, there are checkboxes for "Time", "No-Time", "Blank", and "Error", with "Time" checked. The main workspace shows a grid of 439 matching rows (out of 8664 total). A large red rectangle highlights a specific row in the grid. A red arrow points from the bottom of this highlighted row towards a callout box at the bottom left. The callout box contains the text: "Looking at 2010-2011 records only, data on market offerings is missing!". The right side of the interface shows various facets for columns like Wine, Coffee, Beans, Fruits, Grains, Juices, Mushrooms, PetFood, Tofu, WildHarvested, and updateTime, each with a dropdown menu icon.

Facet / Filter Undo / Redo 6

Refresh Reset All Remove All

Facet / Filter

updateTime change reset

2010-11-01 15:40:12 — 08:36:36

Time No-Time Blank Error

8131 53 0 0

439 matching rows (8664 total)

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

« first < previous 1 - 10 next > last »

Nine ▾ Coffee ▾ Beans ▾ Fruits ▾ Grains ▾ Juices ▾ Mushrooms ▾ PetFood ▾ Tofu ▾ WildHarvested ▾ updateTime ▾

2011-01-01T00:00:00Z  
2011-01-01T00:00:00Z

Extensions:

Open... Export Help

Looking at 2010-2011 records only, data on market offerings is missing!

# ... and slices with detailed data!

The screenshot shows the OpenRefine interface for a project titled "Farmers-Markets". The main area displays a grid of 616 matching rows (8664 total) with columns for various food items like Wine, Coffee, Beans, Fruits, Grains, Juices, Mushrooms, PetFood, Tofu, WildHarvested, and updateTime. A red box highlights the updateTime column, which lists dates from 2012-03-01 to 2012-05-25. A red arrow points from this column to a timeslice visualization on the left. The timeslice shows a timeline from 2012-01-01 to 2012-05-25, with a vertical bar indicating the current slice. A red callout box contains the text: "2012 timeslice has data about market offerings!". The interface includes standard OpenRefine navigation and filtering tools.

Wine	Coffee	Beans	Fruits	Grains	Juices	Mushrooms	PetFood	Tofu	WildHarvested	updateTime
N	N	N	N	N	N	N	N	N	N	2012-03-01T00:00:00Z
N	N	N	N	N	N	N	N	N	N	2012-05-18T00:00:00Z
N	N	N	N	N	N	N	N	N	N	2012-04-25T00:00:00Z
N	N	N	N	N	N	N	N	N	N	2012-05-03T00:00:00Z
N	N	N	N	N	N	N	N	N	N	2012-05-07T00:00:00Z
N	N	N	N	N	N	N	N	N	N	2012-05-24T00:00:00Z
N	N	N	N	N	N	N	N	N	N	2012-05-24T00:00:00Z
N	N	N	N	N	N	N	N	N	N	2012-05-24T00:00:00Z
N	N	N	N	N	N	N	N	N	N	2012-05-25T00:00:00Z

# Converting from String to Number ...

The screenshot shows the OpenRefine interface with a project titled "Farmers-Markets". The main view displays 8664 rows of data across various columns, including Time, Location, Credit, WIC, WICcash, SFMNP, SNAP, Organic, Bakedgoods, Cheese, Crafts, Flowers, Eggs, and Seafood.

A context menu is open over a row of data. The menu path is: **Edit cells** > **Common transforms** > **To number**. Red arrows point from the "already numeric" cell (-86.790709) and the "still string data" cell (-75.53446) to the "Edit cells" and "Common transforms" menu items respectively.

Annotations with red boxes and arrows:

- An annotation points to the cell **-86.790709** with the text "already numeric".
- An annotation points to the cell **-75.53446** with the text "still string data".

Code at the bottom left: `javascript:{}`

# ... from String to Number: Done!

The screenshot shows the OpenRefine interface with a project titled "Farmers-Markets". A yellow status bar at the top right indicates "Text transform on 8635 cells in column y: value.toNumber()". The main view displays a table with 8664 rows. A red box highlights the first two columns, "x" and "y", which now contain numeric values. Red arrows point from a callout box stating "Both x and y columns now have a numeric type!" to these columns. Another red arrow points from a callout box asking "What and where are the missing cells?" to the status bar. The table includes columns for Location, Credit, WIC, WiCash, SFMNP, SNAP, Organic, Bakedgoods, Cheese, Crafts, Flowers, Eggs, Seafood, and more.

x	y	Location	Credit	WIC	WiCash	SFMNP	SNAP	Organic	Bakedgoods	Cheese	Crafts	Flowers	Eggs	Seafood	...
-72.140305	44.411013			Y	Y	N	Y	N	Y	Y	Y	Y	Y	Y	N
-81.7285969	41.375118			Y	Y	N	Y	Y	-	Y	N	N	Y	Y	N
-85.574887	42.296024			Y	Y	N	Y	Y	N	Y	Y	N	Y	Y	N
-82.8187	34.8042			Y	N	N	N	N	N	-					
-94.2746191	37.495628			Y	N	N	N	N	-	Y	N	Y	N	Y	N
-73.9493	40.7939	Private business parking lot		N	N	Y	Y	N	-	Y	N	Y	Y	N	N
-86.790709	36.11837			Y	N	N	N	Y	Y	Y	Y	N	Y	Y	N
-73.9482477	40.8089533	Federal/State government building grounds		Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	N
-75.53446	39.742117	On a farm, from: a barn, a greenhouse, a tent, a stand, etc		N	N	N	N	Y	N	N	N	N	N	N	N
-77.0320505	38.9169984	Other		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N

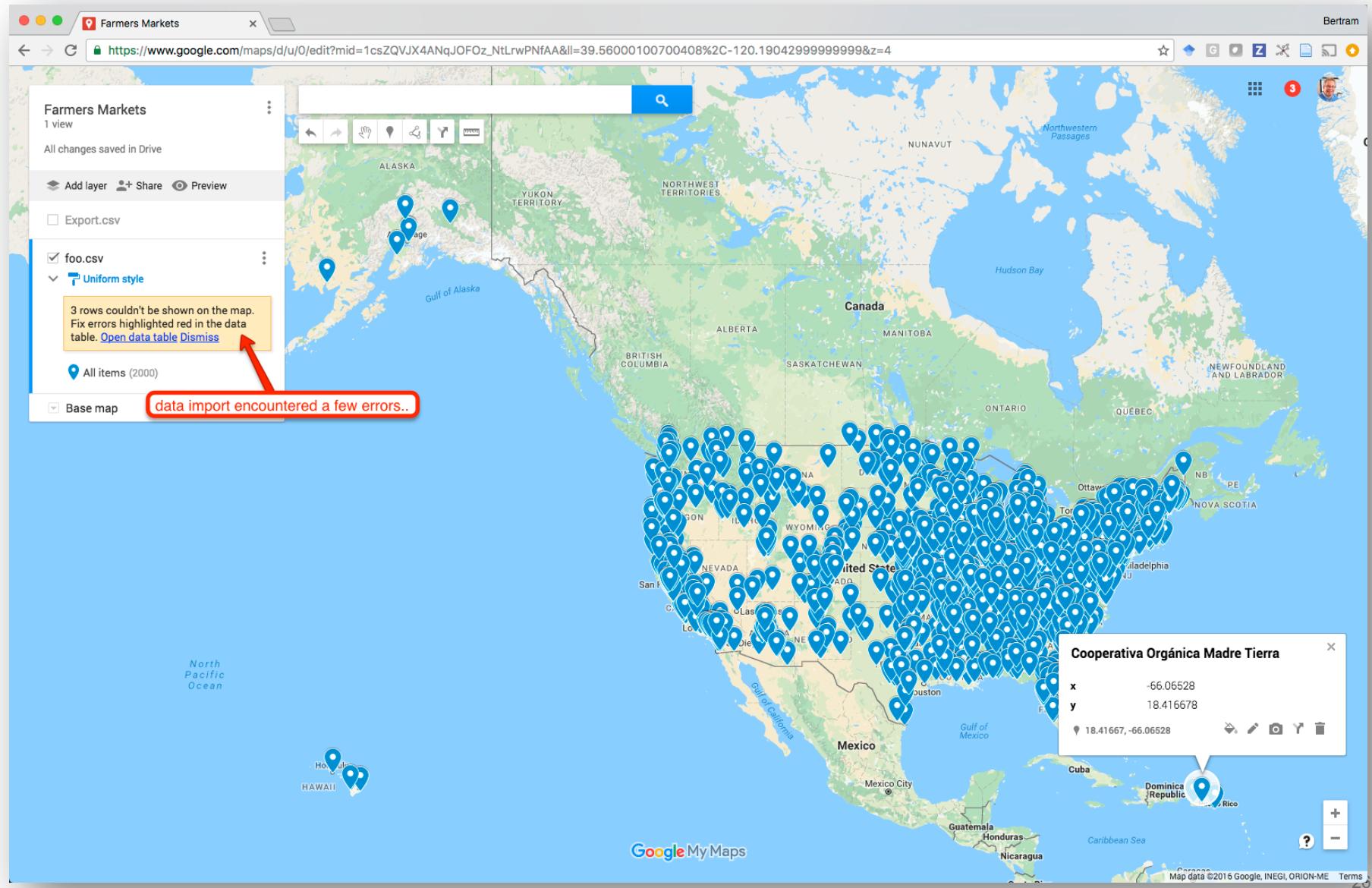
# The x,y (longitude,latitude) data lets us use a “gem”: Scatterplot Facet!

The screenshot shows the OpenRefine interface with the title "Farmers-Markets - OpenRefine" and the facet "Faceting · OpenRefine/OpenRe". The main area displays 8664 rows of data, with the first few rows visible:

Season4Time	x	y	Location	Credit	WIC	WICcash	SFMNP	SNAP	Organic	Bakedgo
	-72.140305						Y	N	Y	Y
	-81.7285969						Y	Y	-	Y
	-82.8187						Y	Y	N	Y
	-94.2746191						N	N	-	
	-73.9493	40.7939	Private business parking lot		N	N	Y	Y	N	-
	-86.790709	36.11837			Y	N	N	N	Y	Y
	-73.9482477	40.8089533	Federal/State government building grounds		Y	Y	N	Y	Y	Y
	-75.53446	39.742117	On a farm from: a barn, a	N	N	N	N	Y	N	N

The left sidebar shows a scatterplot of the data with a legend for "lin" and "log" scales, and a "reset" button. A red arrow points from the "Facet / Filter" section to the "Scatterplot facet" option in the context menu for the "y" column.

# Georeferenced data & (Google) maps!



Dealing with more messy and  
more complex data issues ...

... The NYPL Menus Project!

# Example: NYPL “Menu” Collection

A red arrow points to the URL bar, highlighting the website address: [menus.nypl.org](http://menus.nypl.org).

The website title is "What's on the menu?" with "Est. 2011" and a search bar.

The main navigation menu includes: Menus, Dishes, Data, Blog, About, and Help.

A call-to-action section encourages users to help transcribe historical restaurant menus:

- Help The New York Public Library improve a unique collection!
- We're transcribing our historical restaurant menus, dish by dish, so that they can be searched by what people were eating back in the day. It's a big job so we need your help! [Learn more](#).
- Connect: [menus@nypl.org](mailto:menus@nypl.org) | Twitter | Facebook

A central feature is the "Frutti di Mare!" section, which displays the latest transcribed menus:

Restaurant	Year	Dishes Transcribed
Bookbinders Sea Food House	1943	154 dishes
Legal Sea Foods	1998	103 dishes
Bill's Seafood Ship Café	1954	121 dishes
Fisherman's Grotto		6 dishes
The Great American Fish Company	1987	99 dishes
Rogano Restaurant & Sea Food Bar	1964	21 dishes

Below this, there are sections for "Help review", "Explore", and "Today's specials".

The "Help review" section shows menus that need review:

Restaurant	Year
Erie Railroad System	1939
Aerated Bread Company	1900
Norddeutscher Lloyd Bremen	1900

The "Explore" section shows a map of New York City with a highlighted area around Bryant Park and 40th Street, labeled "Map our Menus!".

The "Today's specials" section lists some dishes:

- Chicken Okra Soup
- Anisette
- Black & White
- Stewed Lobster In Port Wine
- Pigs Feet
- Gekochte & Gebratene



Whats on the menu? x Bertram

menus.nypl.org/data

A New York Public Library website Explore others! ⌂

NYPL Labs

# What's on the menu? Data

Search keyword(s) Go

Menus Dishes Data Blog About Help

## Data exports

There's a lot of data behind *What's on the Menu?*: a mix of simple bibliographic description of the menus (created by The New York Public Library) and the culinary and economic content of the menus themselves (transcribed by you). Now we're opening it up.

All data generated through *What's on the Menu?* is available in two ways:

### Spreadsheet Exports

On the 1st and 16th of every month, we'll post a complete export of all menu and dish data collected so far (menus, dishes, prices, and more).

Download the [latest data](#) export in CSV format (11/01/16).

### API

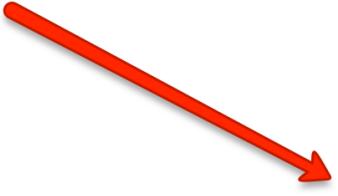
As the first project of NYPL Labs, we're happy to announce that Menus is also the first NYPL project to have a public API. In fact, we use this exact same API to build many of the features of this site.

You can learn how to use the API on our [Github](#) page, but you can get started now by [sending us an email](#) with the subject **API ACCESS** and a description of your project.

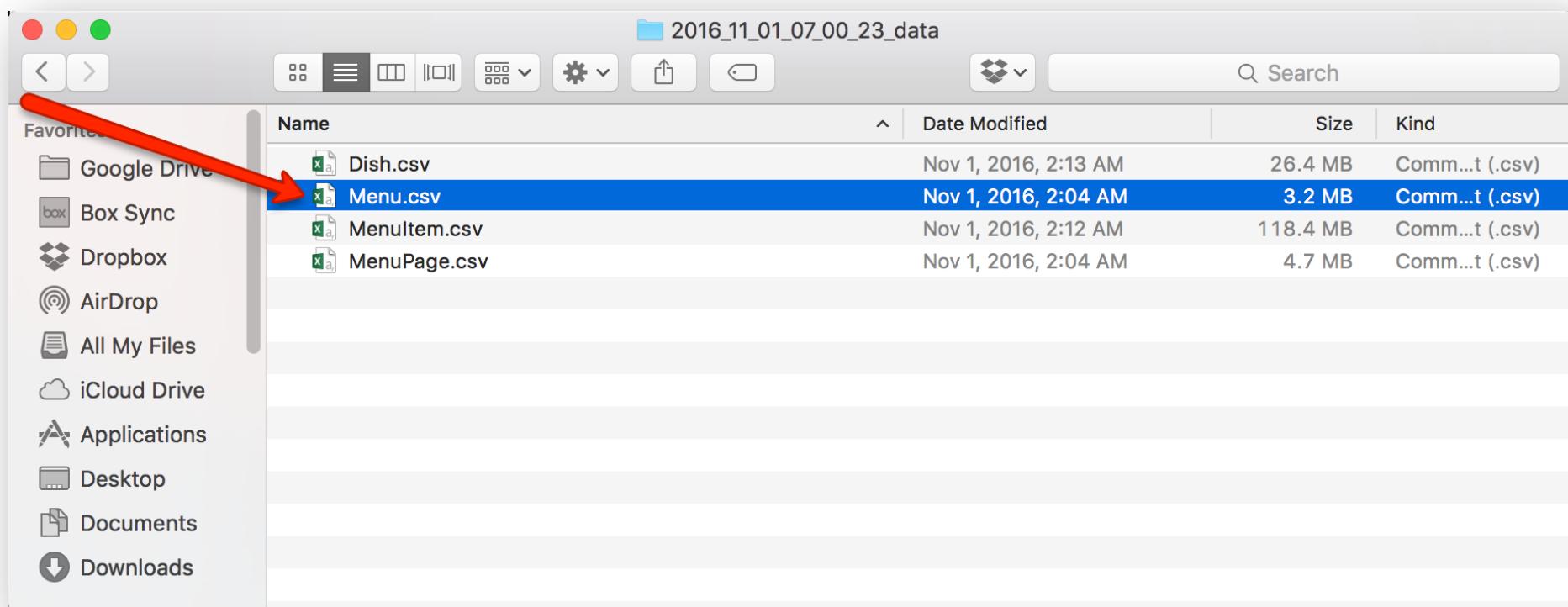
Also, feel free to add issues you've found using the API via our [Github issues](#) page.

#### What's an API?

No known copyright restrictions on this material. We ask that you credit The New York Public Library as source on any applications or publications.



# Unpacking and selection Menu.csv ...



# Complex Data Quality Issues: To fix or not to fix?

- Relying on volunteer transcription will often result in inconsistent data entry
- Even well-transcribed data is subject to challenges due to synonyms and spelling variants, etc.
- Also: entities change over time...

e.g., Childs' restaurant, originally launched by brothers Samuel and William Childs in 1889, grew to be one of the first national dining chains and dropped its apostrophe sometime after 1907.



vs.



*The same restaurant styled differently in 1907 (left) and 1916 (right)*

# Basic Normalization

Google refine Menu csv Permalink Open... Export Help

Facet / Filter Undo / Redo 0

17079 rows

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
1. 12463			Facet	BREAKFAST	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...			Trim leading and trailing whitespace		
			Edit column	Common transforms			Collapse consecutive whitespace		
			Transpose	Fill down					
			Sort...	Blank down					
			View	Unescape HTML entities					
			Reconcile	To titlecase					
				To uppercase					
				To lowercase					
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				MENU IN GERMAN AND ENGLISH; ILLUS, HARBOR SCENE WITH SAILING VESSEL;

# Faceting and Clustering

Google refine Menu csv Permalink Open... Export Help

Facet / Filter Undo / Redo 3

Refresh Reset All Remove All

**sponsor** change

6080 choices Sort by: name count Cluster

? 57  
?(J B) 1  
? CLUB 1  
? HOTEL 1  
'95 LAW OF COLUMBIAN UNIVERSITY 1  
'97S CLASS DINNER 1  
'POSSUM CLUB 1  
(?COLONIAL HOTEL?) 1  
(238 EIGHT AVENUE) 1  
(ABBAS II HILMI KHEDIVE OF EGYPT) 1

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

Google ref

Facet / Filter

Refresh

x sponsor

6080 choices Sort

? 57  
? (J B) 1  
? CLUB 1  
? HOTEL 1  
'95 LAW OF COL UNIVERSITY 1  
'97S CLASS DINI  
'POSSUM CLUB  
(?COLONIAL HO  
(238 EIGHT AVE  
(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

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

# Choices in Cluster

# Rows in Cluster

Average Length of Choices

Length Variance of Choices

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

Export Help

ns: Freebase

- 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

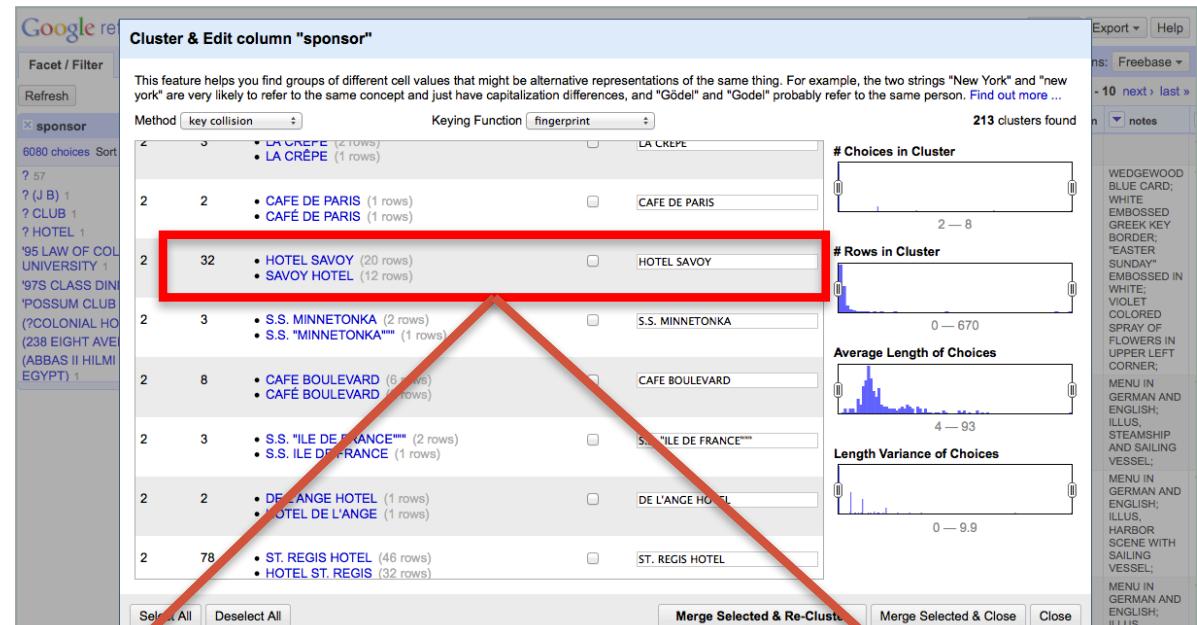
# Kinds of Clustering

- 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

# But beware: Clustering Caveat!

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

*Savoy Hotel  
Strand  
London WC2R 0EU  
United Kingdom*



# Summary: A First Look at OpenRefine

- Creating a New Project
- Basic Normalization
- Different Facets (text, timeline, scatterplot)
- Clustering and Mass Edits
- Operation History: Provenance
- Separate videos:
  - Installing OpenRefine
  - Advanced Operations