

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

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

[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 title bar says "OpenRefine" and the status bar says "Bertram". The left sidebar has links for "Create Project" (which is selected), "Open Project", "Import Project", and "Language Settings". Below the sidebar is a diamond icon and the text "Version 2.6-rc.2 [TRUNK]". The main content area has a heading "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. It also mentions support for other formats via extensions. A "Get data from" section shows "This Computer" selected, with a "Choose Files" button and "No file chosen". There are also links for "Web Addresses (URLs)", "Clipboard", and "Google Data". A "Next »" button is visible at the bottom of this section.

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	https://sites.google.com/site/caledoniafarmersmarket/	https://www.facebook.com/Danville.VT.Farmers.Market/		
2.	1011871	Stearns Homestead Farmers' Market	http://StearnsHomestead.com			
3.	1011878	100 Mile Market	http://www.pfcmarkets.com	https://www.facebook.com/100MileMarket/?fref=ts		
4.	1009364	106 S. Main Street Farmers Market	http://thetownofsixmile.wordpress.com/			
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 columns: FMID, MarketName, Website, Facebook, Twitter, and YouTube. The table lists 10 rows of data, each with a star icon and a blue checkmark icon. The data includes:

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

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. A red box highlights this facet, 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/, http://10thStreetCommunityFarmersMarket.com, http://112stMadisonAvenue.com, http://www.12southfarmersmarket.com, http://www.125thStreetFarmersMarket.com, https://www.facebook.com/pages/12th-Brandywine-Urban-Farm-Community-Garden/253769448091860, and https://www.facebook.com/14UFarmersMarket.
 - Facebook:** https://www.facebook.com/Danville.VT.Farmers.Market/, https://www.facebook.com/100MileMarket/?ref=ts, https://www.facebook.com/10thStreetCommunityFarmersMarket, https://www.facebook.com/112stMadisonAvenue, https://www.facebook.com/12SouthFarmersMarket, https://www.facebook.com/125thStreetFarmersMarket, and https://www.facebook.com/14UFarmersMarket.
 - Twitter:** https://twitter.com/FarmMarket125th, @12southfrmsmkt, and https://twitter.com/14UFarmersMkt.
 - Youtube:** https://www.youtube.com/watch?v=... and https://www.youtube.com/watch?v=...
 - Other:** https://www.instagram.com/... and https://www.instagram.com/...
- Toolbar:** Includes buttons for **Open...**, **Export**, and **Help**.
- Header:** Shows the project name **Farmers-Markets** and the URL **127.0.0.1:3333/project?project=1766664496116**.

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">Main Street Farmers Market (9 rows)MAIN STREET FARMERS MARKET (1 rows)Main Street Farmer's Market (1 rows)Main Street Farmers' Market (1 rows)	<input type="checkbox"/>	Main Street Farmers Market
4	5	<ul style="list-style-type: none">Irvington Farmers Market (2 rows)Irvington Farmer's Market (1 rows)Irvington Farmers Market (1 rows)Irvington Farmers' Market (1 rows)	<input type="checkbox"/>	Irvington Farmers Market
3	3	<ul style="list-style-type: none">Wakefield Farmer's Market (1 rows)Wakefield Farmers Market (1 rows)Wakefield Farmers Market (1 rows)	<input type="checkbox"/>	Wakefield Farmer's Market
3	4	<ul style="list-style-type: none">Columbus Farmers' Market (2 rows)Columbus Farmers Market (1 rows)columbus farmers market (1 rows)	<input type="checkbox"/>	Columbus Farmers' Market
3	3	<ul style="list-style-type: none">WATERTOWN FARMERS MARKET (1 rows)Watertown Farmers market (1 rows)Watertown Farmers' Market (1 rows)	<input type="checkbox"/>	WATERTOWN FARMERS M
3	5	<ul style="list-style-type: none">Rochester Downtown Farmers Market (3 rows)Rochester Farmers Market (1 rows)Rochester Farmers Market (1 rows)Rochester Farmers Market (1 rows)	<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, a checkbox for merging, and the resulting new cell value. A red arrow points to the "Select All" button at the bottom left of the cluster table.

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

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

On the right side of the interface, there are four data visualization charts:

- # Choices in Cluster: Histogram showing the distribution of cluster sizes, ranging from 2 to 4.
- # Rows in Cluster: Histogram showing the distribution of row counts per cluster, ranging from 2 to 34.
- Average Length of Choices: Histogram showing the distribution of average lengths of choices, ranging from 13 to 71.
- Length Variance of Choices: Histogram showing the distribution of length variance of choices, ranging from 0 to 2.5.

... resulting in a mass edit ...

The screenshot shows the OpenRefine interface for a project titled "Farmers-Markets". The main view displays 8664 rows of data across several columns: FID, FMID, MarketName, Website, Facebook, Twitter, and YouTube. A yellow header bar indicates "Mass edit 659 cells in column MarketName Undo". A red arrow points from the "MarketName" facet on the left to this bar. Another red arrow points from the "MarketName" facet to the "MarketName" column header in the main view. The facet lists 7846 choices, sorted by name count, with "El Mercado Familiar" at the top.

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

.. 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)

Provenance Information!

Extract and save parts of your operation history as JSON that you can apply to this or other projects in the future.

- e.toNumber()
- Text transform on cells in column updateTime using expression value.toDate()
- Text transform on 8131 cells in column Season1Date using expression value.toDate()
- Text transform on 26 cells in column updateTime: value.toDate()
- Text transform on 8131 cells in column updateTime: value.toString()
- Text transform on 8131 cells in column updateTime: value.toDate()
- Text transform on 286 cells in column street: value.trim()
- Text transform on 8664 cells in column x: value.toString()
- Text transform on 0 cells in column y: value.toNumber()
- Text transform on 8635 cells in column x: value.toNumber()
- Text transform on 8664 cells in column y: value.toString()
- Text transform on 8635 cells in column y: value.toNumber()
- Mass edit cells in column MarketName

Select All Unselect All Close

```
[{"op": "core/mass-edit", "description": "Mass edit cells in column MarketName", "engineConfig": { "mode": "row-based", "facets": [] }, "columnName": "MarketName", "expression": "value", "edits": [ { "fromBlank": false, "fromError": false, "from": [ "Main Street Farmers Market", "MAIN STREET FARMERS MARKET", "Main Street Farmer's Market", "Main Street Farmers' Market" ], "to": "Main Street Farmers Market" }, { "fromBlank": false, "fromError": false, "from": [ "Irvington Farmers Market", "Irvington Farmer's Market", "Irvington Farmers Market ", "Irvington Farmers' Market" ], "to": "Irvington Farmers Market" } ]}
```

Processing History of the "Mass Edit" (update cells in each cluster with canonical name)

Facet / Filter Undo / Redo 13 Extract... Apply... Permalink Open... Export... Help Bertram Extensions: first < previous 1 - 10 next > last Twitter Youtube Other https://v http://ac type=m @12southfrmsmkt @12so https://twitter.com/FarmMarket125th Instagram https://v https://twi https://14UfarmersMkt

More Data Profiling: Timeline Facet

The screenshot shows the OpenRefine interface for a project titled "Farmers-Markets". The main area displays 8664 rows of data in a grid format. The columns represent various categories: Wine, Coffee, Beans, Fruits, Grains, Juices, Mushrooms, PetFood, Tofu, WildHarvested, and updateTime. A sidebar on the left provides instructions on using facets and filters, including a link to "Watch these screencasts". A red arrow points from the "Facet / Filter" button in the top-left corner to a context menu that is open over the "WildHarvested" column header. This menu lists several facet types: 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 "Extensions" button in the top-right corner to a list of available extensions on the right side of the screen. The extensions listed are: Facet (version 8/2016 10:09 PM), Text filter, Edit cells, Edit column, Transpose, Sort..., View, and Reconcile.

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" with a dropdown menu.

Red arrows highlight the facet filter counts (8131, 533) and the text transform message.

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 with a project titled "Farmers-Markets". The main view displays 616 matching rows (8664 total) across various columns: 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 visualization shows a timeline from 2012-01-01 to 2012-05-25, with a specific slice highlighted between 2012-01-01 and 2012-05-22. A red callout box contains the text: "2012 timeslice has data about market offerings!". The OpenRefine interface includes standard navigation buttons like Refresh, Undo / Redo, and a toolbar with Open..., Export, and Help.

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 for a project titled "Farmers-Markets". The main view displays a grid of data with 8664 rows and 17 columns. A yellow status bar at the top indicates "Text transform on 8635 cells in column y: value.toNumber()". A red arrow points from this status bar to the "y" column header. Another red arrow points from a callout box stating "Both x and y columns now have a numeric type!" to the same "y" column header. A third red arrow points from a callout box asking "What and where are the missing cells?" to the status bar. The data grid includes columns for "x" and "y", which are highlighted with a red border. The "y" column contains values like -72.140305, 44.411013, -81.7285969, etc. The "x" column contains values like -77.0320505, 38.9169984, -73.9482477, etc. The rest of the columns contain categorical data represented by Y or N.

x	y	Location	Credit	WIC	WICcash	SFMNP	SNAP	Organic	Bakedgoods	Cheese	Crafts	Flowers	Eggs	Seafood	Health
-72.140305	44.411013		Y	Y	N	Y	N	Y	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	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	N
-77.0320505	38.9169984	Other	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N

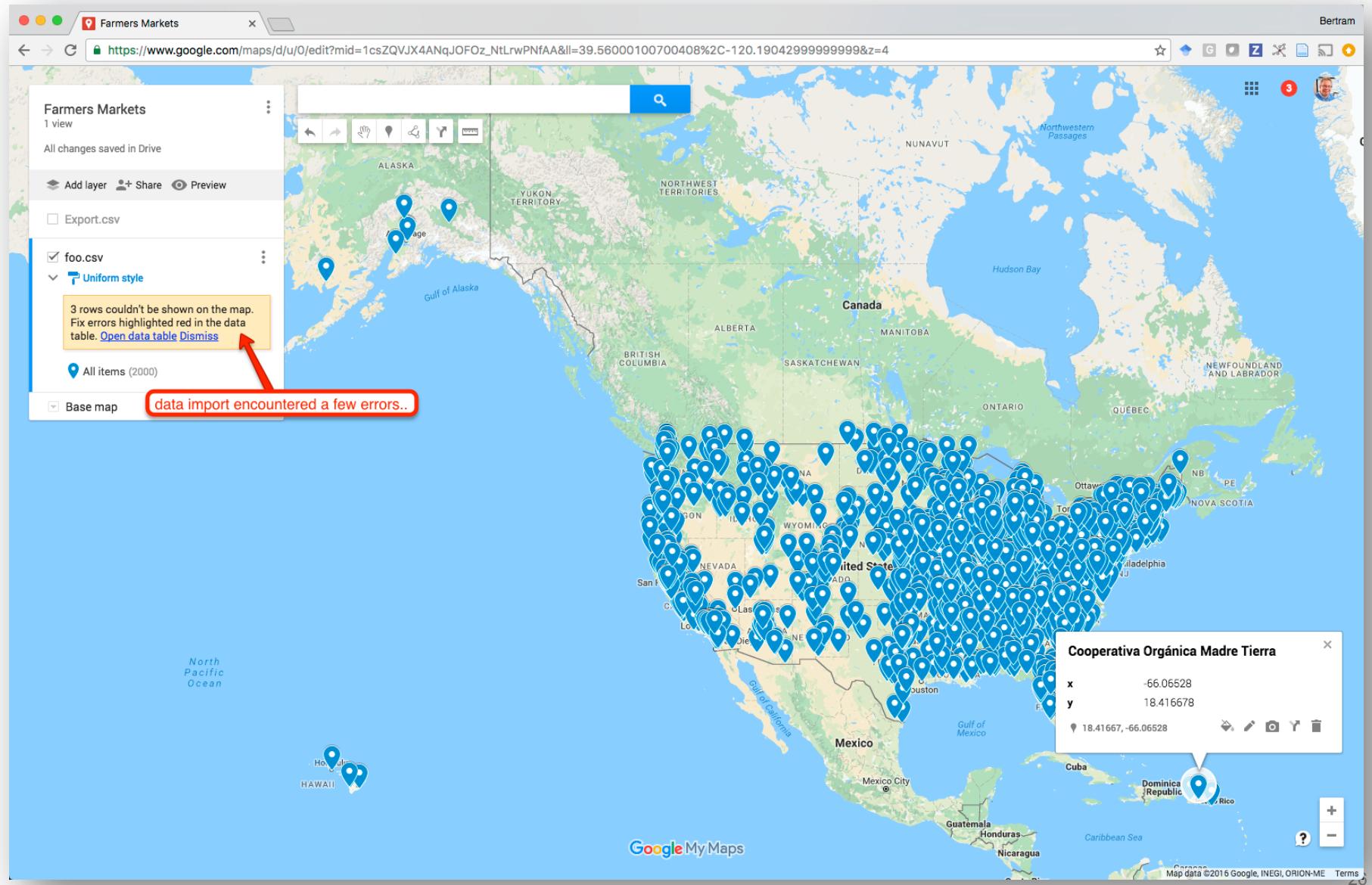
The x,y (longitude,latitude) data lets us use a “gem”: Scatterplot Facet!

The screenshot shows the OpenRefine interface for a project titled "Farmers-Markets". The main view 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	Y
	-75.53446	39.742117	On a farm from: a barn, a	N	N	N	N	Y	N	N

The left sidebar shows a facet for "x (x) vs. y (y)", which includes a map of the United States where data points are plotted. A red arrow points from this sidebar to the "Facet" option in the context menu that is open over the "y" column header. Another red arrow points from the "Scatterplot facet" option in the menu back to the "Facet" button in the sidebar.

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.

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 | 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

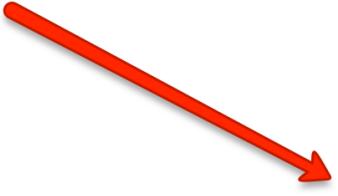
As the first project of NYPL Labs, we're happy to announce that Menus is also the first NYPL project to have a public API. In fact, we use this exact same API to build many of the features of this site.

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.

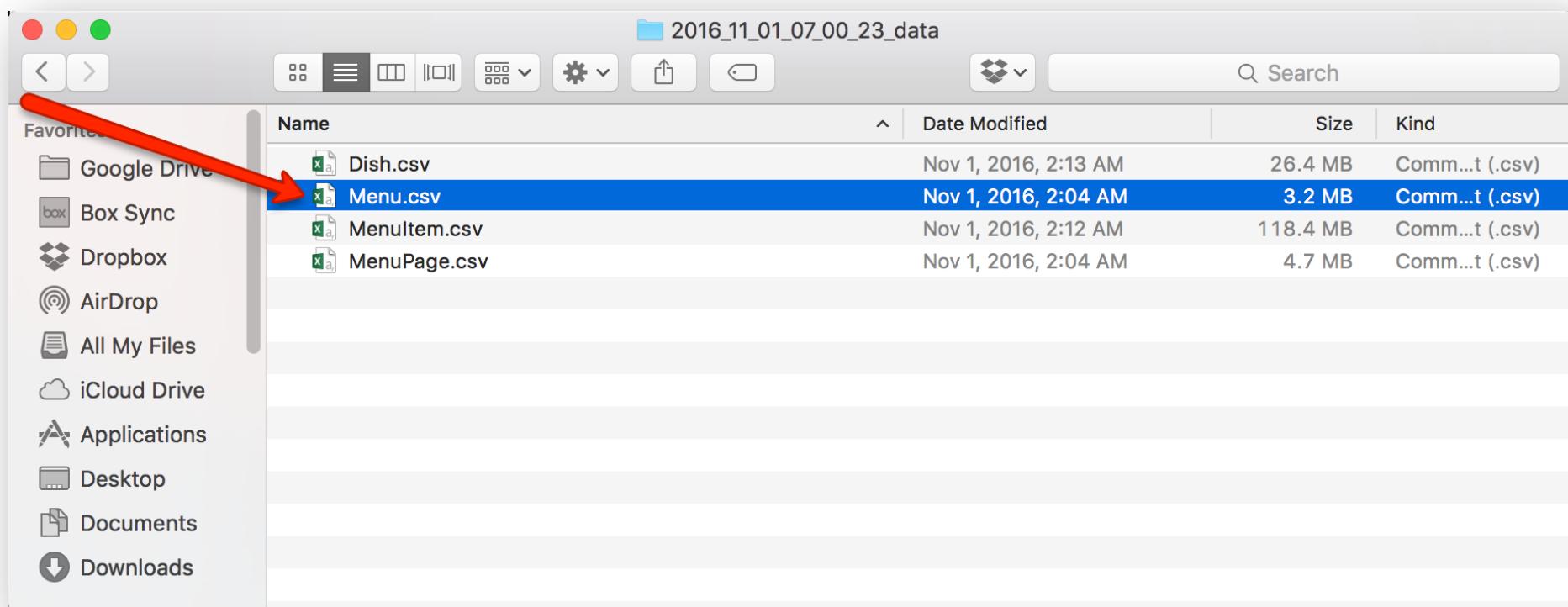
Also, feel free to add issues you've found using the API via our [Github issues](#) page.

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		To number		
							To date		
							To text		
							Blank out cells		
4. 12466			NORDDEUTSCHER LLOYD BREMEN	LUNCH;	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE;	CARD; ILLU; COL; 5.5X8.0;		

A red arrow points to the "Trim leading and trailing whitespace" option in the context menu for the "event" column of row 2.

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;			1
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;		1
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;		1
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;		1



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

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

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

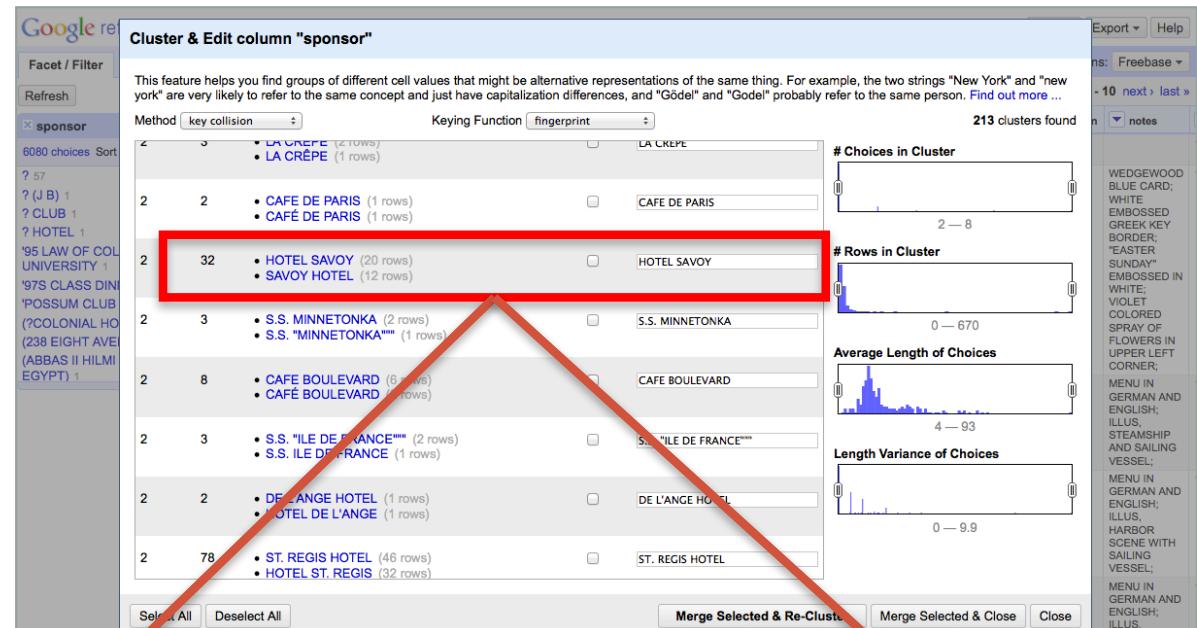
Kinds of Clustering

- 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