Web Applications
for the Future Internet

Data Journalism

Data Cleansing - Open Refine

InfoUma 2020-21 Andrea Marchetti

What

Data cleansing definition

“*Process of* ***detecting*** *and* ***correcting*** *corrupt or inaccurate records from data.*”

source Wikipedia

Why

*Better **data** beats fancier **algorithms…***

***Plain and Simple**! If you have a clean dataset, even simple algorithms can learn impressive insights from it!*

*We can make beautiful analyzes but if our data is dirty we expose ourselves to **destructive criticism***

The origins of errors

- user entry errors
- multiple users involved in data input
- corruption in transmission or storage
- join of different data sources
- use of different control data dictionaries
- no use of control data dictionaries
- ...

The Goal is **Data Quality**

Data Quality Criteria

1. Validity
2. Accuracy
3. Completeness
4. Consistency
5. Uniformity

Validity: compliance with defined constraints

Data-Type: values in a particular column must be of a particular datatype, e.g., boolean, numeric, date, etc. For example a latitude should be a float not a string

Range: typically, numbers or dates should fall within a certain range. For example month number should be [1-12] latitude of Tuscany should be [42-45]

Mandatory*:* certain columns cannot be empty. For example the coordinates of accomodation

Unique: a field, or a combination of fields, must be unique across a dataset. For example a civic address

Validity: compliance with defined constraints

***Set-Membership*:** values of a column come from a set of discrete values, e.g. enum values. `For example, a person’s gender may be male or female.`

***Foreign-key*:** as in relational databases, a foreign key column can’t have a value that does not exist in the referenced primary key.

Regular expression patterns: text fields that have to be in a certain pattern. `For example, a date may be required to have the pattern 23-12-2019.`

Cross-field validation: certain conditions that span across multiple fields must hold. `For example, a patient’s date of discharge from the hospital cannot be earlier than the date of admission.`

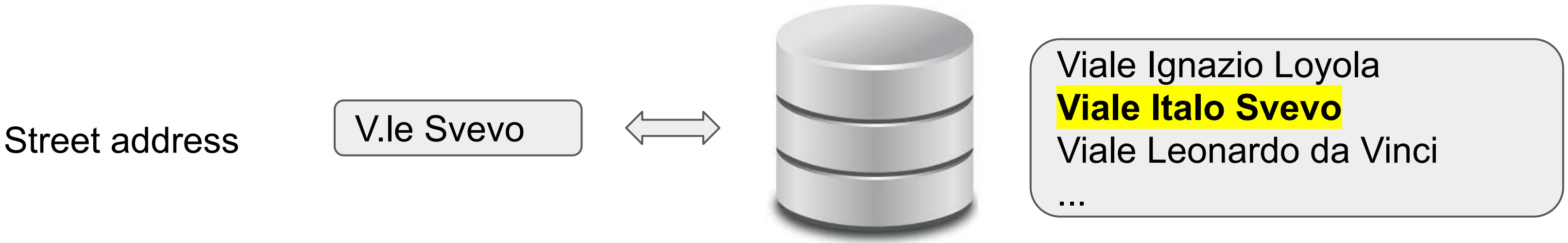
Accuracy

Definition: The **degree of conformity** of a **measure** to a **standard or a true value**

It requires accessing an external source of data that contains the true value.

Such "**gold standard**" data is often unavailable.

`Examples of gold standard: official street name data base`



Street address

V.le Svevo

Viale Ignazio Loyola

Viale Italo Svevo

Viale Leonardo da Vinci

...

Completeness

Definition: **The degree to which all required measures are known.**

Missing data is going to happen for various reasons

You can check why on the data source miss some data and try to fix or you can **exploit external services**

For example for missing geographical coordinates exploit **geocoding services**

Consistency

Definition: The degree to which a set of measures are consistent

Inconsistency occurs when two data items in the data set contradict each other

e.g., a customer is recorded in two different sources as having two different current addresses. A valid age, say 10, mightn’t match with the marital status, say divorced

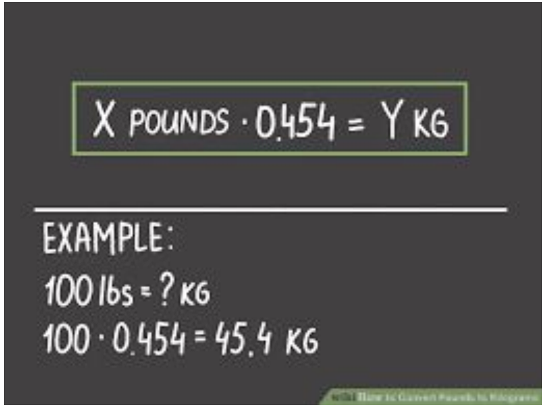
Fixing inconsistency is not always possible: it requires a variety of strategies

e.g., deciding which data were recorded more recently, which data source is likely to be most reliable, or simply trying to find the truth (e.g., calling up the customer).

Uniformity

Definition: The degree to which the data is specified using the same **unit of measure**

E.g. In datasets extracted from different sources, weight may be recorded either in pounds or kilos and must be converted to a single measure using an arithmetic transformation.



X POUNDS · 0.454 = Y KG

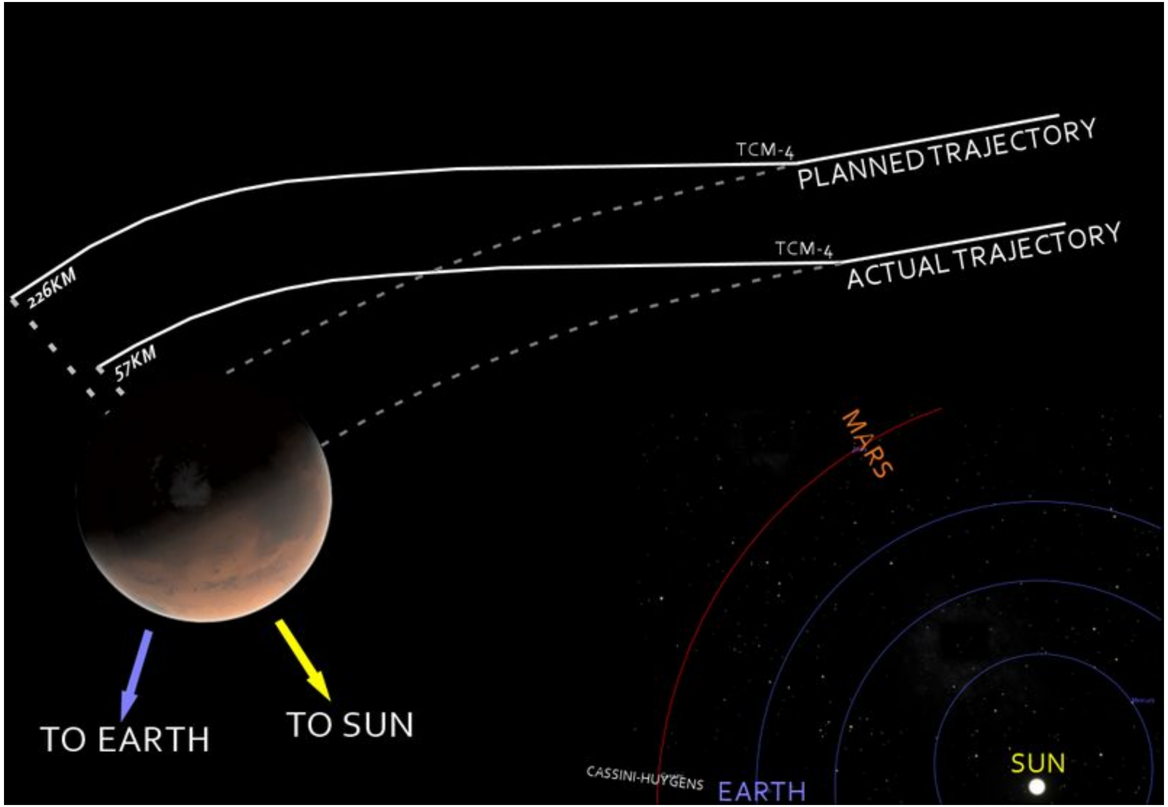
EXAMPLE:

100 lbs = ? KG

100 · 0.454 = 45.4 KG

wikiHow to Convert Pounds to Kilograms

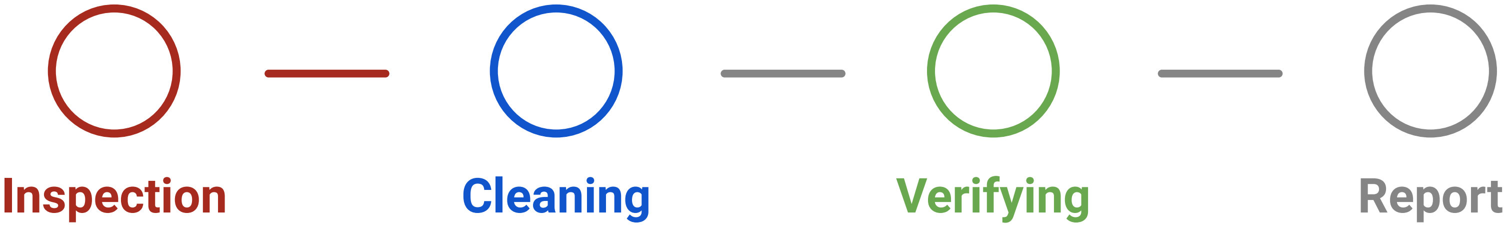
Mars Climate Orbiter 1999



The primary cause of this discrepancy was that one piece of ground software supplied by Lockheed Martin produced results in a United States customary unit, contrary to its Software Interface Specification (SIS), while a second system, supplied by NASA, expected those results to be in SI units, in accordance with the SIS

wikipedia

The method



Inspection

Cleaning

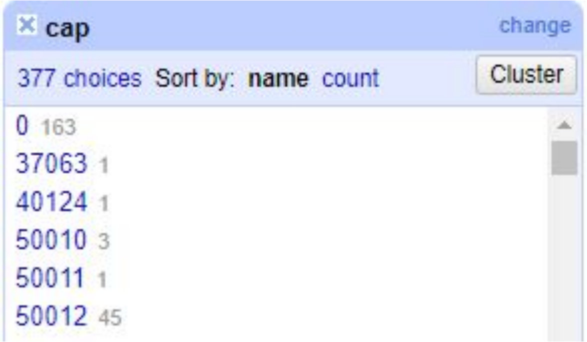
Verifying

Report

Inspection

For each column calculate a Summary Statistics

- Is the data column recorded as a string or number?
- How many values are missing?
- How many unique values in a column?

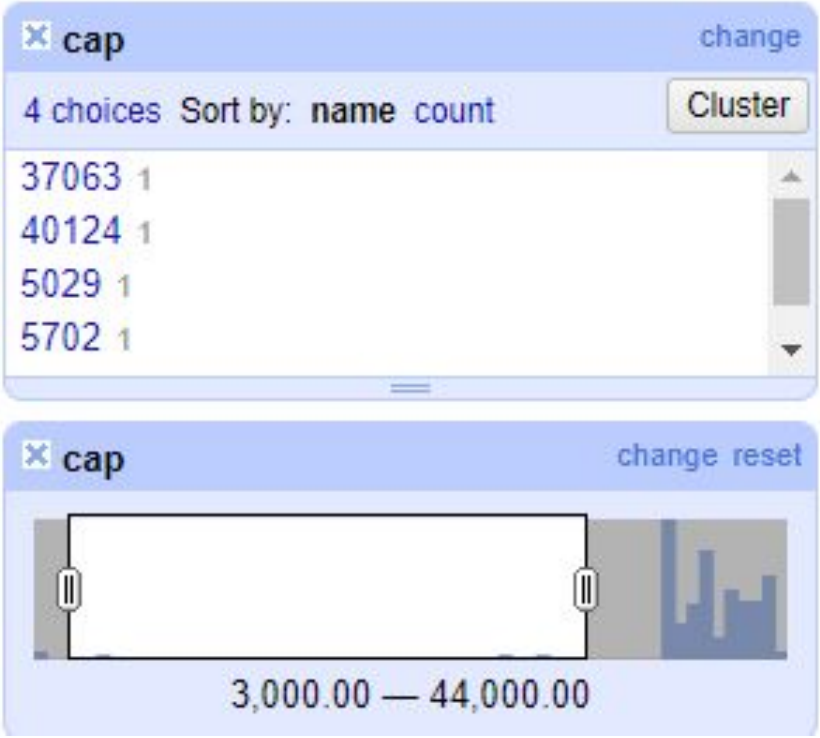


377 choices Sort by: name count

0	163
37063	1
40124	1
50010	3
50011	1
50012	45

Inspection of Data Distribution

Visualizing data distribution with Histograms, and statistical methods such as mean, standard deviation, range, or quartiles, one can find Outliers and thus potential data entry errors that it worths to investigate.



37063	1
40124	1
5029	1
5702	1

3,000.00 — 44,000.00

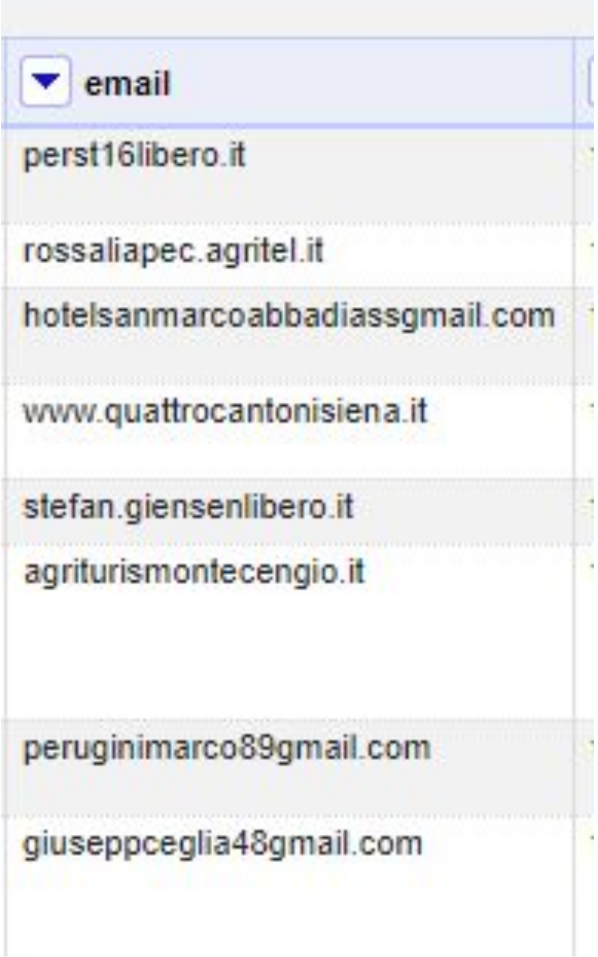
Cleansing

Irrelevant data: remove

Duplicates: remove

Type conversion: fix

Syntax errors: fix



email
perst16libero.it
rossaliapec.agritel.it
hotelsanmarcoabbadiassgmail.com
www.quattrocantonisiena.it
stefan.giensenlibero.it
agriturismontecengio.it
peruginimarco89gmail.com
giuseppceglia48gmail.com

Cleansing and enrichment

Cleansing

- Fixing errors
- Remove duplicate records (rows) or irrelevant data (cols)
- Split multi data columns (address, datetime)

Enrichment

- Filling missing values
- Fixing not normalized values

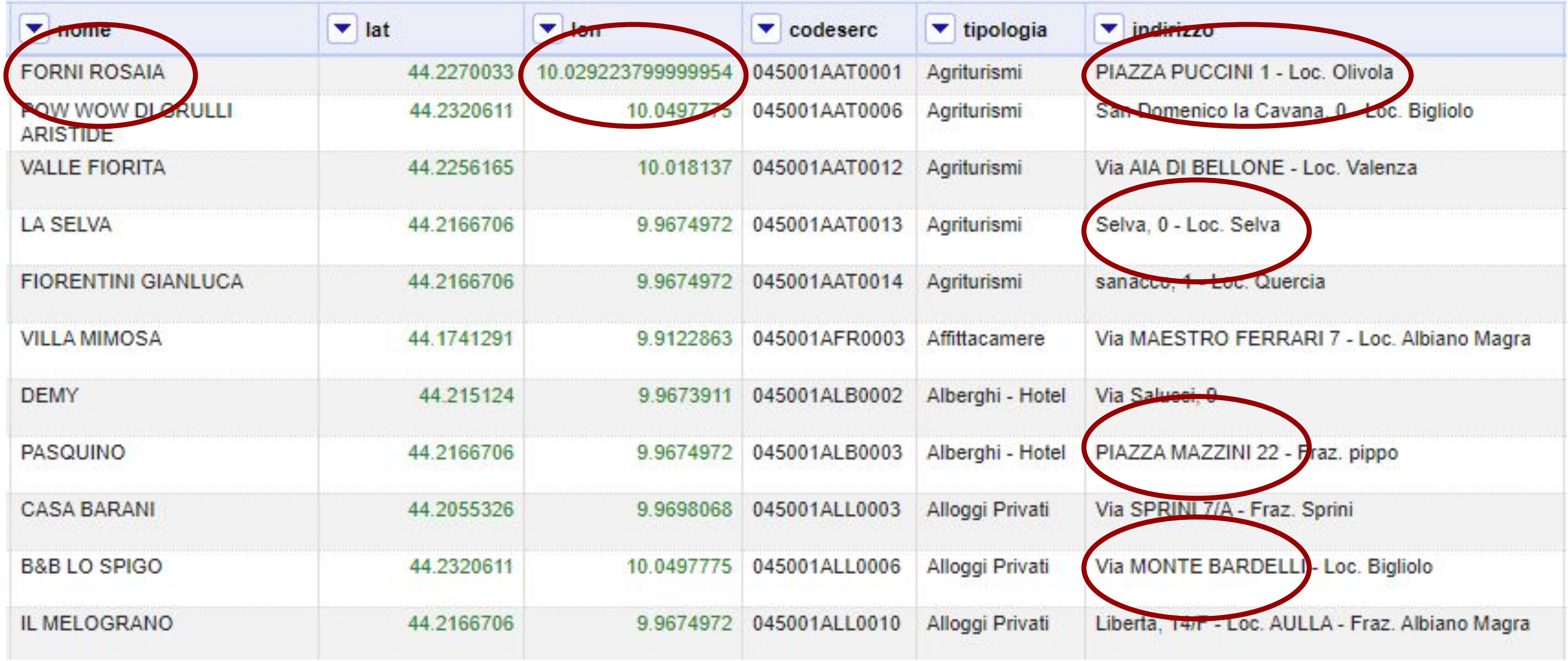
Common errors

String vs numbers (“10,5432” vs 10.5432)

Different Formats (01/09/2016 vs 01-09-2016)

Data inconsistencies (Piazza, P.zza, P.za)

Lateral spaces (“B&B” vs “ B&B”)



nome	lat	lon	codeserc	tipologia	indirizzo
FORNI ROSAIA	44.2270033	10.029223799999954	045001AAT0001	Agriturismi	PIAZZA PUCCINI 1 - Loc. Olivola
POW WOW DI GRULLI ARISTIDE	44.2320611	10.0497775	045001AAT0006	Agriturismi	San Domenico la Cavana, 0 - Loc. Bigliolo
VALLE FIORITA	44.2256165	10.018137	045001AAT0012	Agriturismi	Via AIA DI BELLONE - Loc. Valenza
LA SELVA	44.2166706	9.9674972	045001AAT0013	Agriturismi	Selva, 0 - Loc. Selva
FIORENTINI GIANLUCA	44.2166706	9.9674972	045001AAT0014	Agriturismi	sanacco, 1 - Loc. Quercia
VILLA MIMOSA	44.1741291	9.9122863	045001AFR0003	Affittacamere	Via MAESTRO FERRARI 7 - Loc. Albiano Magra
DEMY	44.215124	9.9673911	045001ALB0002	Alberghi - Hotel	Via Salucci, 0
PASQUINO	44.2166706	9.9674972	045001ALB0003	Alberghi - Hotel	PIAZZA MAZZINI 22 - Fraz. pippo
CASA BARANI	44.2055326	9.9698068	045001ALL0003	Alloggi Privati	Via SPRINI 7/A - Fraz. Sprini
B&B LO SPIGO	44.2320611	10.0497775	045001ALL0006	Alloggi Privati	Via MONTE BARDELLI - Loc. Bigliolo
IL MELOGRANO	44.2166706	9.9674972	045001ALL0010	Alloggi Privati	Liberta, 14/F - Loc. AULLA - Fraz. Albiano Magra

Verifying

- Verify always what have you done.
  - For example, after filling out the **missing data**, they might violate any of the rules and constraints.
- The data cleansing is an iterative process
- It might involve some manual correction if not possible otherwise.

Report

In the end it is necessary to make a **report** of all the changes made, describing the reasons and the methods of data cleansing.

It would be desirable that all changes were automatic and therefore repeatable

Bibliography

Data Cleansing - Wikipedia

The Ultimate Guide To Data Cleaning

A plethora of data cleaning **tools**

- **Text Editor**: Notepad++,

- **Spreadsheet**: Google SpreadSheet, MS Excel
- **Free tools**: Open refine
- **Not free tools**: Trifacta, Paxata, Alteryx
- **Code yourself**: Python with Pandas Library

Open Refine

openrefine.org

A free, open source, multiplatform, **desktop application**

OpenRefine 3.4.1, released on September 20, 2020

User manual https://docs.openrefine.org/

Besides it’s possible:

- extend functionalities with **extension** (Ex Named Entity recognition)
- drive some operations by python (or other languages) scripts

Accomodations in Tuscany

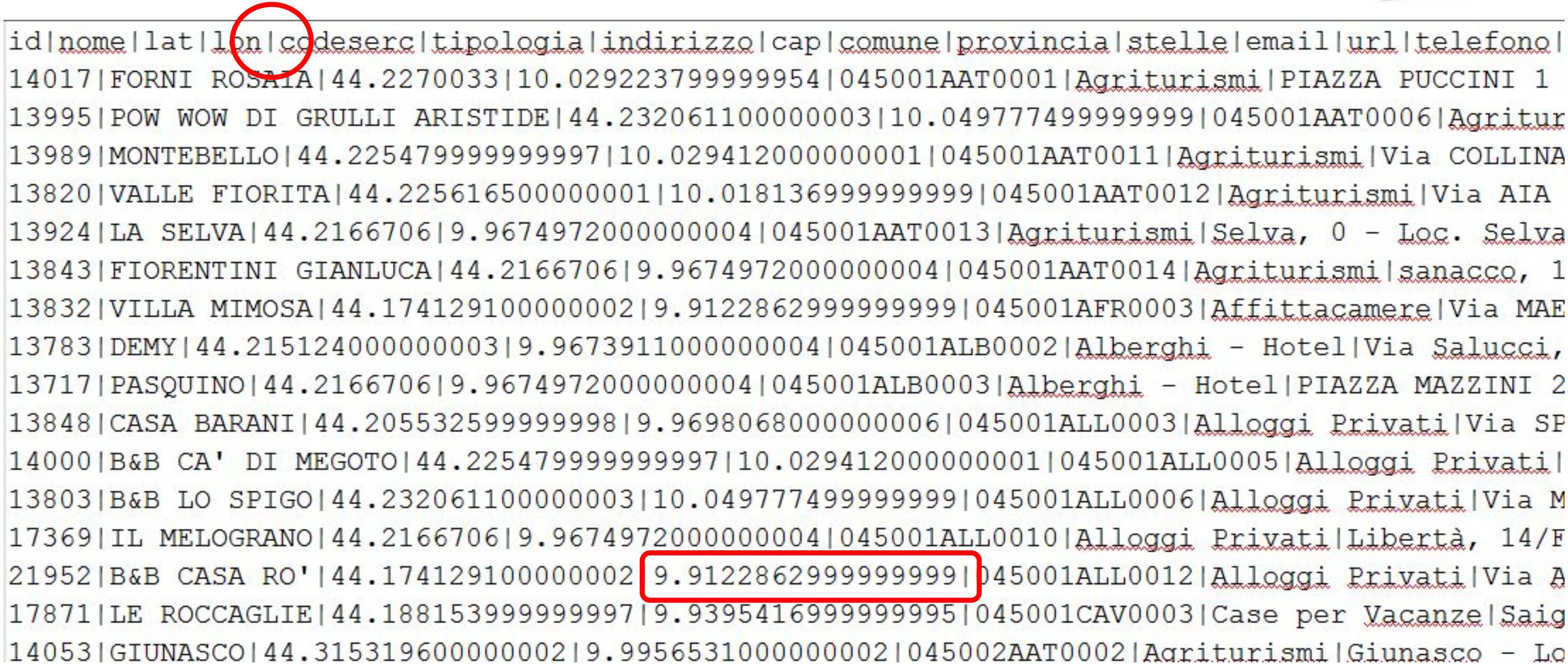
Accomodations in Tuscany

*“L'archivio contiene i nomi e i dati anagrafici (indirizzo, telefono, e-mail, sito web) di tutte le **strutture ricettive della Toscana**, codificate secondo i codici ISTAT e distinte per tipologia (alberghi, agriturismi, ..) e stabilimenti balneari.”*

http://servizi.toscana.it/RT/mappe/strutturericettiveXall.csv

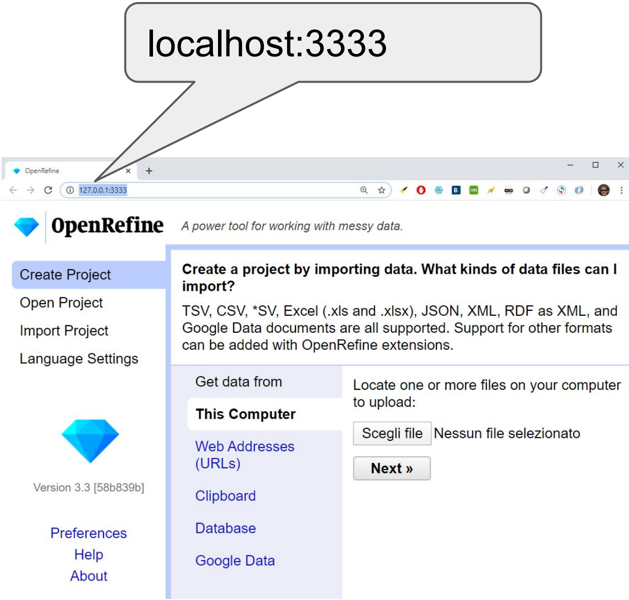
Creator	Area di Coordinamento "Turismo, Commercio e Terziario"
Creation date	28 - 11 - 2013
Last update	02 - 07 - 2019

First view with a text editor



```
id|nome|lat|lon|codeserc|tipologia|indirizzo|cap|comune|provincia|stelle|email|url|telefono|
14017|FORNI ROSAIA|44.2270033|10.029223799999954|045001AAT0001|Agriturismi|PIAZZA PUCCINI 1
13995|POW WOW DI GRULLI ARISTIDE|44.232061100000003|10.049777499999999|045001AAT0006|Agritur
13989|MONTEBELLO|44.225479999999997|10.029412000000001|045001AAT0011|Agriturismi|Via COLLINA
13820|VALLE FIORITA|44.225616500000001|10.018136999999999|045001AAT0012|Agriturismi|Via AIA
13924|LA SELVA|44.2166706|9.967497200000004|045001AAT0013|Agriturismi|Selva, 0 - Loc. Selva
13843|FIORENTINI GIANLUCA|44.2166706|9.967497200000004|045001AAT0014|Agriturismi|sanacco, 1
13832|VILLA MIMOSA|44.174129100000002|9.912286299999999|045001AFR0003|Affittacamere|Via MAE
13783|DEMY|44.215124000000003|9.967391000000004|045001ALB0002|Alberghi - Hotel|Via Salucci,
13717|PASQUINO|44.2166706|9.967497200000004|045001ALB0003|Alberghi - Hotel|PIAZZA MAZZINI 2
13848|CASA BARANI|44.205532599999998|9.969806800000006|045001ALL0003|Alloggi Privati|Via SP
14000|B&B CA' DI MEGOTO|44.225479999999997|10.029412000000001|045001ALL0005|Alloggi Privati|
13803|B&B LO SPIGO|44.232061100000003|10.049777499999999|045001ALL0006|Alloggi Privati|Via M
17369|IL MELOGRANO|44.2166706|9.967497200000004|045001ALL0010|Alloggi Privati|Libertà, 14/F
21952|B&B CASA RO'|44.174129100000002|9.912286299999999|045001ALL0012|Alloggi Privati|Via A
17871|LE ROCCAGLIE|44.188153999999997|9.939541699999995|045001CAV0003|Case per Vacanze|Saig
14053|GIUNASCO|44.315319600000002|9.995653100000002|045002AAT0002|Agriturismi|Giunasco - Lo
```

OpenRefine.exe



localhost:3333

OpenRefine *A power tool for working with messy data.*

- Create Project
- Open Project
- Import Project
- Language Settings

Version 3.3 [58b839b]

Preferences
Help
About

Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with OpenRefine extensions.

Get data from

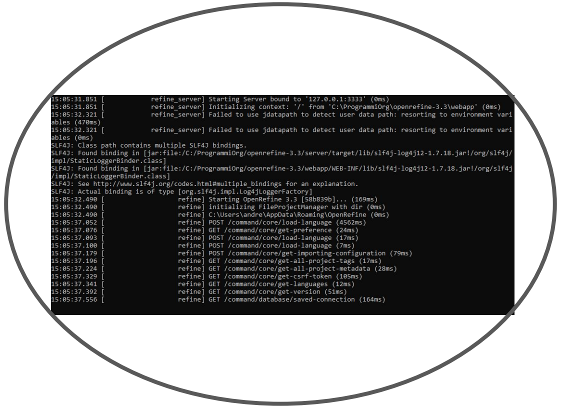
- **This Computer**
- Web Addresses (URLs)
- Clipboard
- Database
- Google Data

Locate one or more files on your computer to upload:

Scegli file Nessun file selezionato

Next »

Web Browser

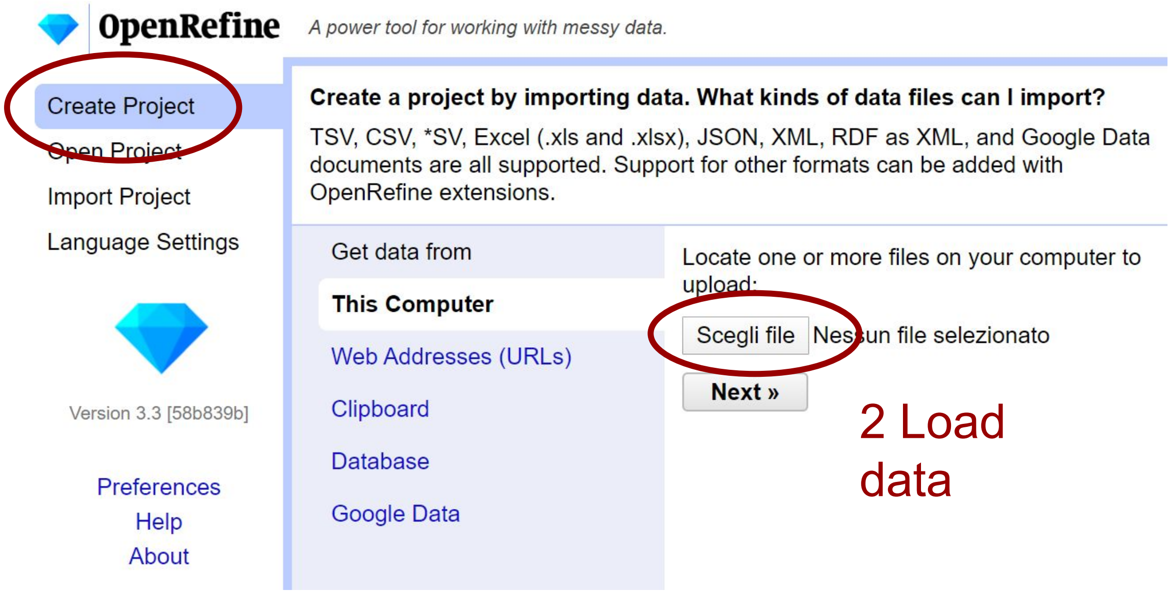


```
15:05:31.851 [            refine_server] Starting Server bound to '127.0.0.1:3333' (0ms)
15:05:31.851 [            refine_server] Initializing context: '/' from 'C:\ProgrammiOrg\openrefine-3.3\webapp' (0ms)
15:05:32.321 [            refine_server] Failed to use jdatapath to detect user data path: resorting to environment vari
ables (470ms)
15:05:32.321 [            refine_server] Failed to use jdatapath to detect user data path: resorting to environment vari
ables (0ms)
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/C:/ProgrammiOrg/openrefine-3.3/server/target/lib/slf4j-log4j12-1.7.18.jar!/org/slf4j/
impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/C:/ProgrammiOrg/openrefine-3.3/webapp/WEB-INF/lib/slf4j-log4j12-1.7.18.jar!/org/slf4j
/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
15:05:32.490 [                   refine] Starting OpenRefine 3.3 [58b839b]... (169ms)
15:05:32.490 [                   refine] initializing FileProjectManager with dir (0ms)
15:05:32.490 [                   refine] C:\Users\andre\AppData\Roaming\OpenRefine (0ms)
15:05:37.052 [                   refine] POST /command/core/load-language (4562ms)
15:05:37.076 [                   refine] GET /command/core/get-preference (24ms)
15:05:37.093 [                   refine] POST /command/core/load-language (17ms)
15:05:37.100 [                   refine] POST /command/core/load-language (7ms)
15:05:37.179 [                   refine] POST /command/core/get-importing-configuration (79ms)
15:05:37.196 [                   refine] GET /command/core/get-all-project-tags (17ms)
15:05:37.224 [                   refine] GET /command/core/get-all-project-metadata (28ms)
15:05:37.329 [                   refine] GET /command/core/get-csrf-token (105ms)
15:05:37.341 [                   refine] GET /command/core/get-languages (12ms)
15:05:37.392 [                   refine] GET /command/core/get-version (51ms)
15:05:37.556 [                   refine] GET /command/database/saved-connection (164ms)
```

Http server on
localhost:3333

Load Accomodation data set

1 Create Project



Create a project by importing data. What kinds of data files can I import?

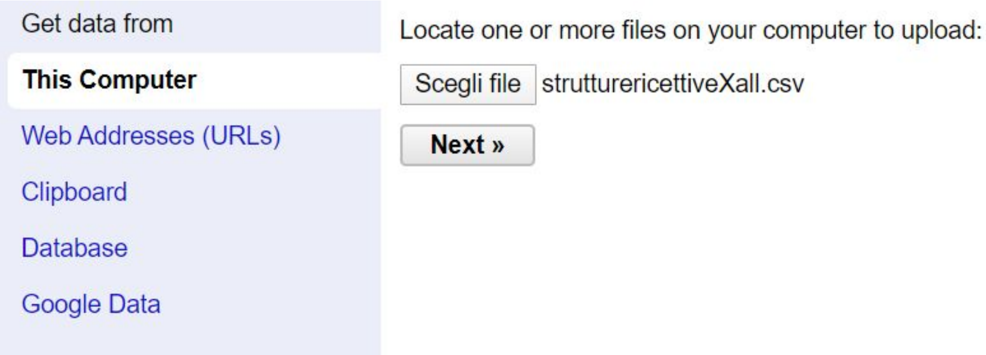
TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with OpenRefine extensions.

Locate one or more files on your computer to upload:

2 Load data

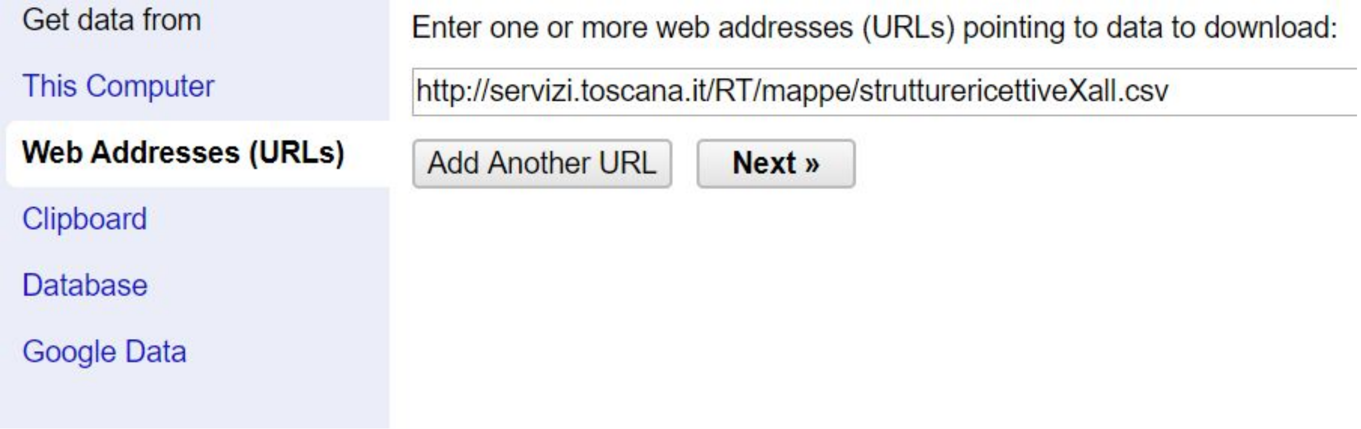
Supported format: CSV, MsExcel, JSON, XML, ...

Different way to access data



Locate one or more files on your computer to upload:

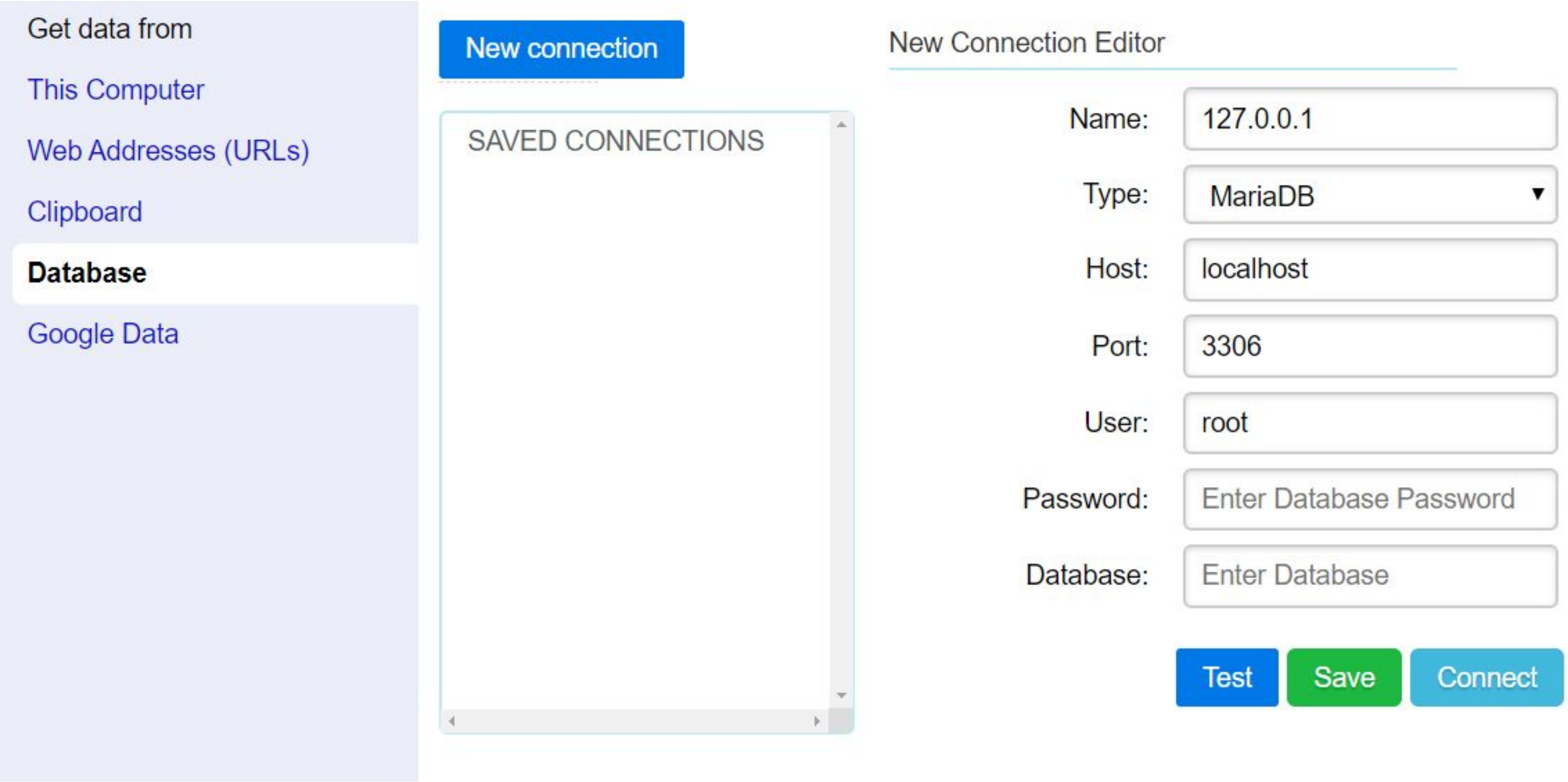
strutturericettiveXall.csv



Enter one or more web addresses (URLs) pointing to data to download:

http://servizi.toscana.it/RT/mappe/strutturericettiveXall.csv

Different way to access data



New Connection Editor

Name: 127.0.0.1

Type: MariaDB ▼

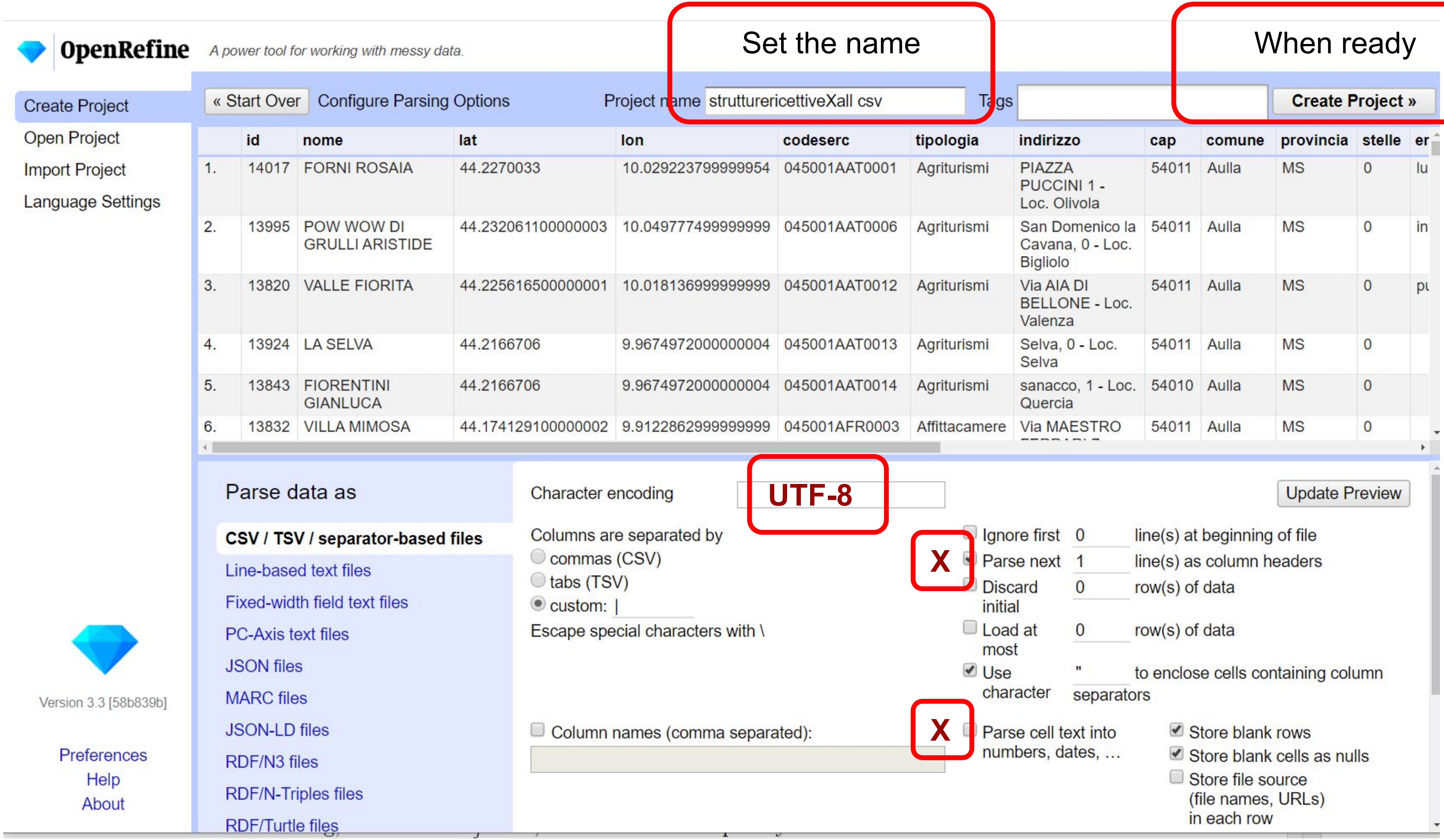
Host: localhost

Port: 3306

User: root

Password: Enter Database Password

Database: Enter Database



Set the name

When ready

UTF-8

X

X

OpenRefine *A power tool for working with messy data.*

Project name: struttthericettiveXall csv

Tags:

	id	nome	lat	lon	codeserc	tipologia	indirizzo	cap	comune	provincia	stelle	er
1.	14017	FORNI ROSAIA	44.2270033	10.029223799999954	045001AAT0001	Agriturismi	PIAZZA PUCCINI 1 - Loc. Olivola	54011	Aulla	MS	0	lu
2.	13995	POW WOW DI GRULLI ARISTIDE	44.232061100000003	10.049777499999999	045001AAT0006	Agriturismi	San Domenico la Cavana, 0 - Loc. Bigliolo	54011	Aulla	MS	0	in
3.	13820	VALLE FIORITA	44.225616500000001	10.018136999999999	045001AAT0012	Agriturismi	Via AIA DI BELLONE - Loc. Valenza	54011	Aulla	MS	0	pu
4.	13924	LA SELVA	44.2166706	9.9674972000000004	045001AAT0013	Agriturismi	Selva, 0 - Loc. Selva	54011	Aulla	MS	0	
5.	13843	FIORENTINI GIANLUCA	44.2166706	9.9674972000000004	045001AAT0014	Agriturismi	sanacco, 1 - Loc. Quercia	54010	Aulla	MS	0	
6.	13832	VILLA MIMOSA	44.174129100000002	9.9122862999999999	045001AFR0003	Affittacamere	Via MAESTRO	54011	Aulla	MS	0	

Parse data as

- **CSV / TSV / separator-based files**
- Line-based text files
- Fixed-width field text files
- PC-Axis text files
- JSON files
- MARC files
- JSON-LD files
- RDF/N3 files
- RDF/N-Triples files
- RDF/Turtle files

Character encoding

Columns are separated by
○ commas (CSV)
○ tabs (TSV)
● custom: |

Escape special characters with \

☐ Column names (comma separated):

☐ 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
☑ Use character " to enclose cells containing column separators
☐ Parse cell text into numbers, dates, ...

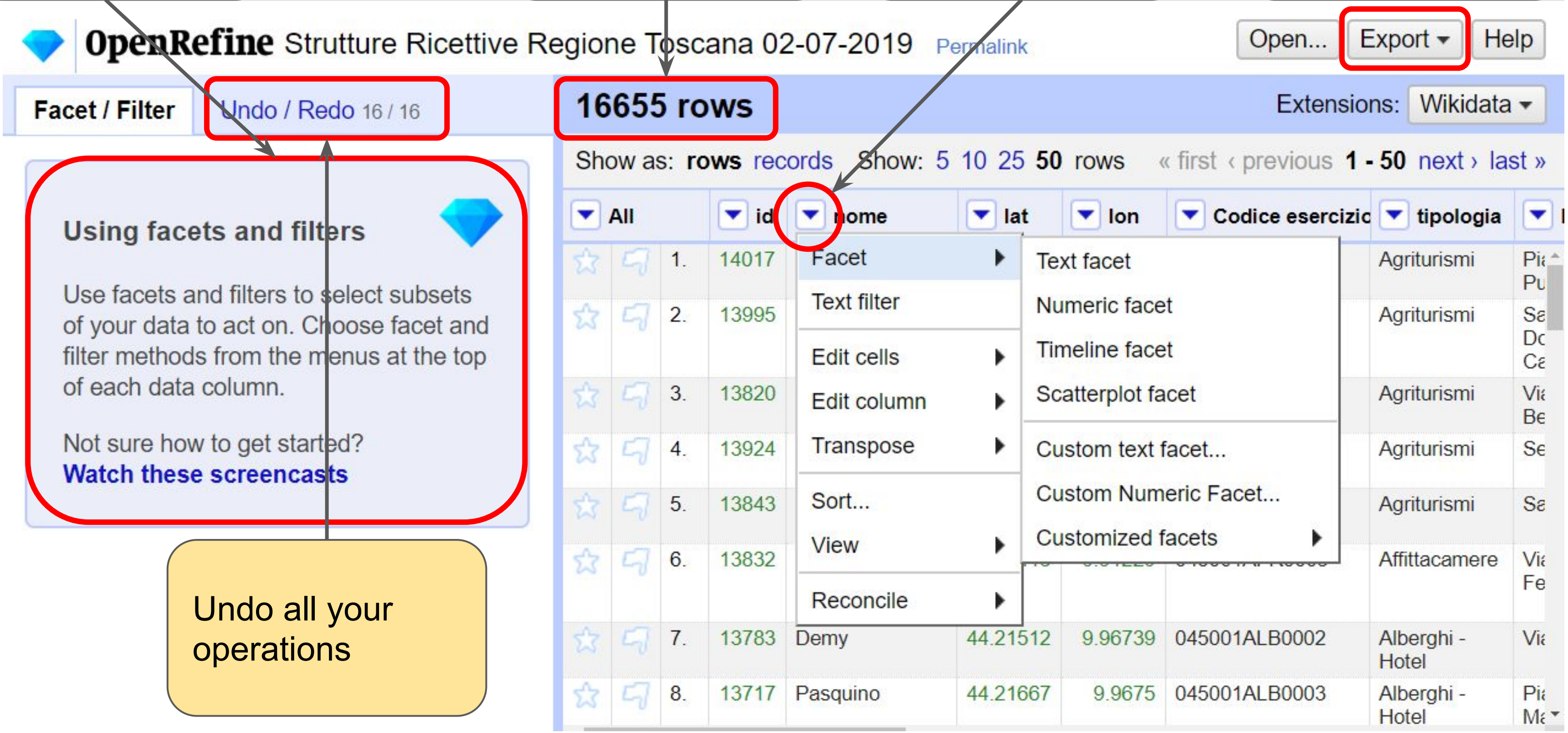
☑ Store blank rows
☑ Store blank cells as nulls
☐ Store file source (file names, URLs) in each row

Facet & Filter Results

Number of selected data

Click on this arrow to facet or filter column data

Export the transformation results



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	nome	lat	lon	Codice esercizio	tipologia	I
1.	14017					Agriturismi	Pia Pu
2.	13995					Agriturismi	Sa Do Ca
3.	13820					Agriturismi	Via Be
4.	13924					Agriturismi	Se
5.	13843					Agriturismi	Sa
6.	13832					Affittacamere	Via Fe
7.	13783	Demy	44.21512	9.96739	045001ALB0002	Alberghi - Hotel	Via
8.	13717	Pasquino	44.21667	9.9675	045001ALB0003	Alberghi - Hotel	Pia Ma

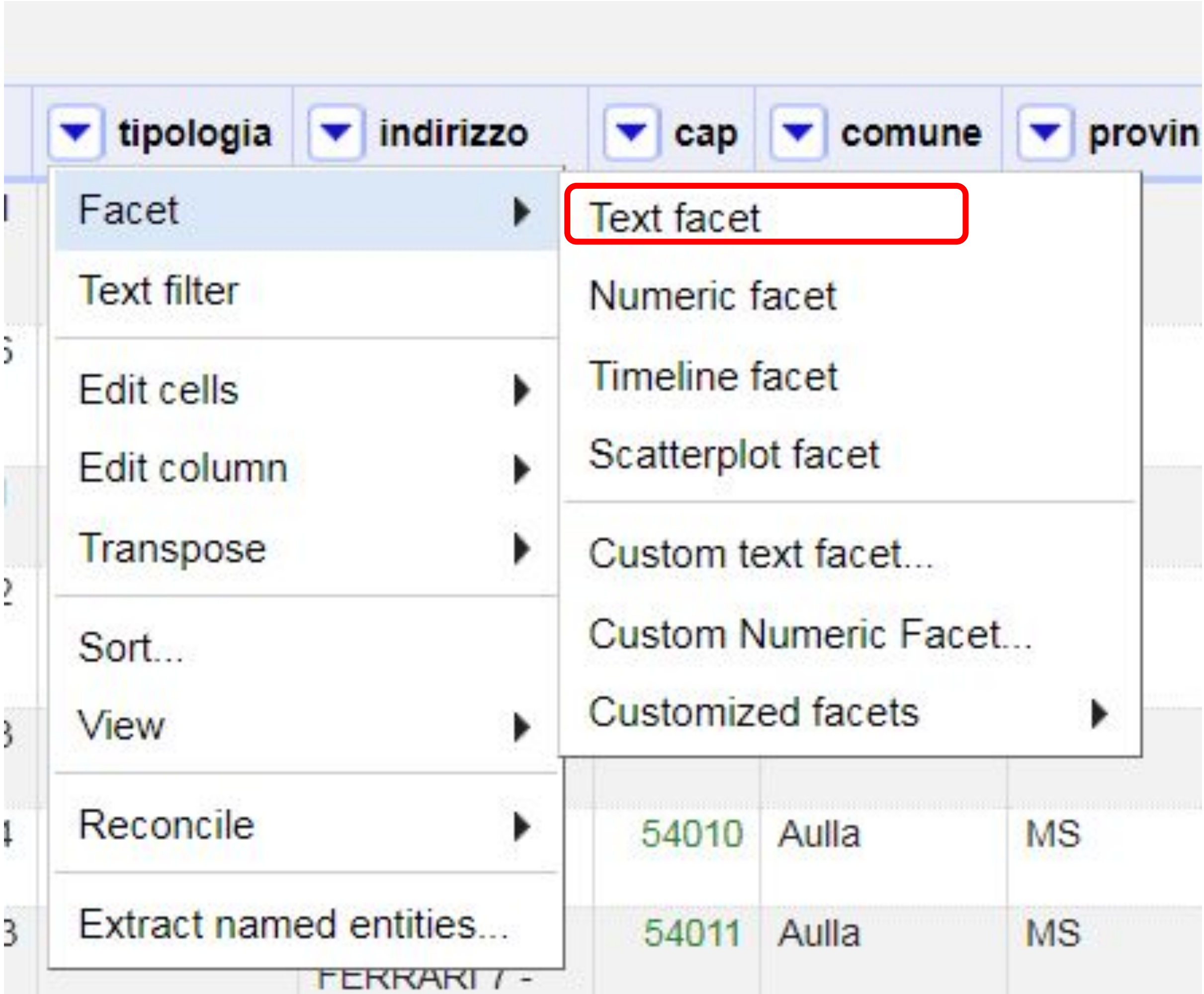
Undo all your operations

Data Exploration

Use

Facet & Filter

to select subsets of your data to act on

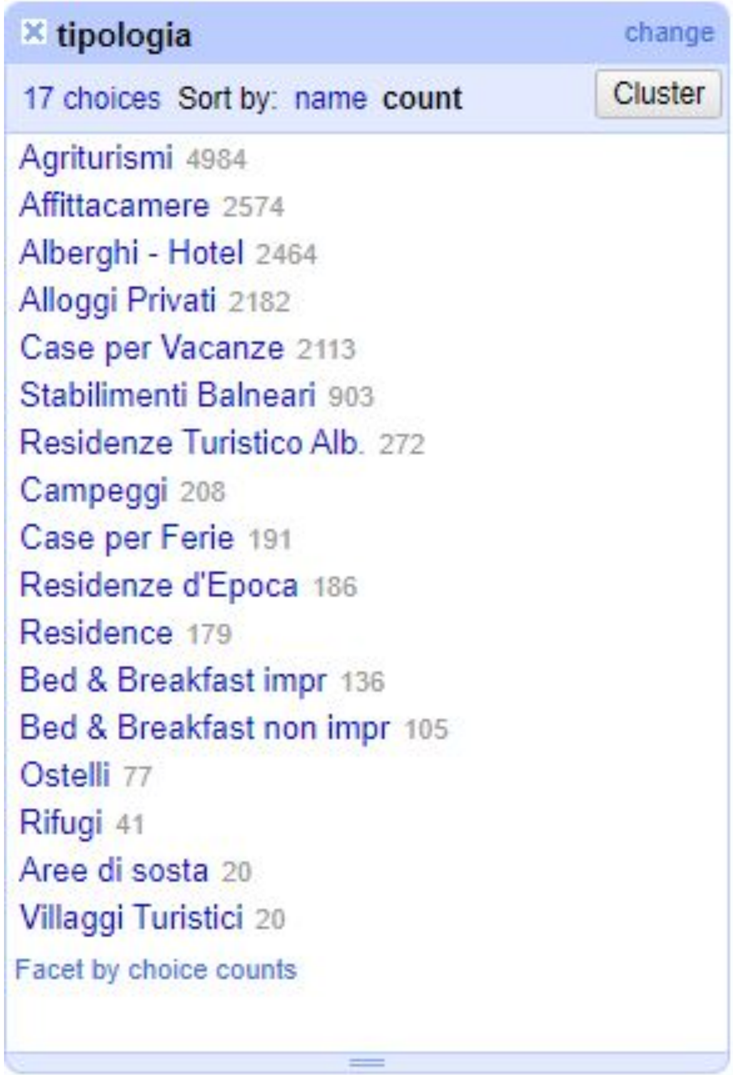


Text Facet on “tipologia” column

Text Facet

In italian: sfaccettature

technically is an histogram



tipologia

17 choices Sort by: name count

Agriturismi	4984
Affittacamere	2574
Alberghi - Hotel	2464
Alloggi Privati	2182
Case per Vacanze	2113
Stabilimenti Balneari	903
Residenze Turistico Alb.	272
Campeggi	208
Case per Ferie	191
Residenze d'Epoca	186
Residence	179
Bed & Breakfast impr	136
Bed & Breakfast non impr	105
Ostelli	77
Rifugi	41
Aree di sosta	20
Villaggi Turistici	20

Numeric Facet

check the limits

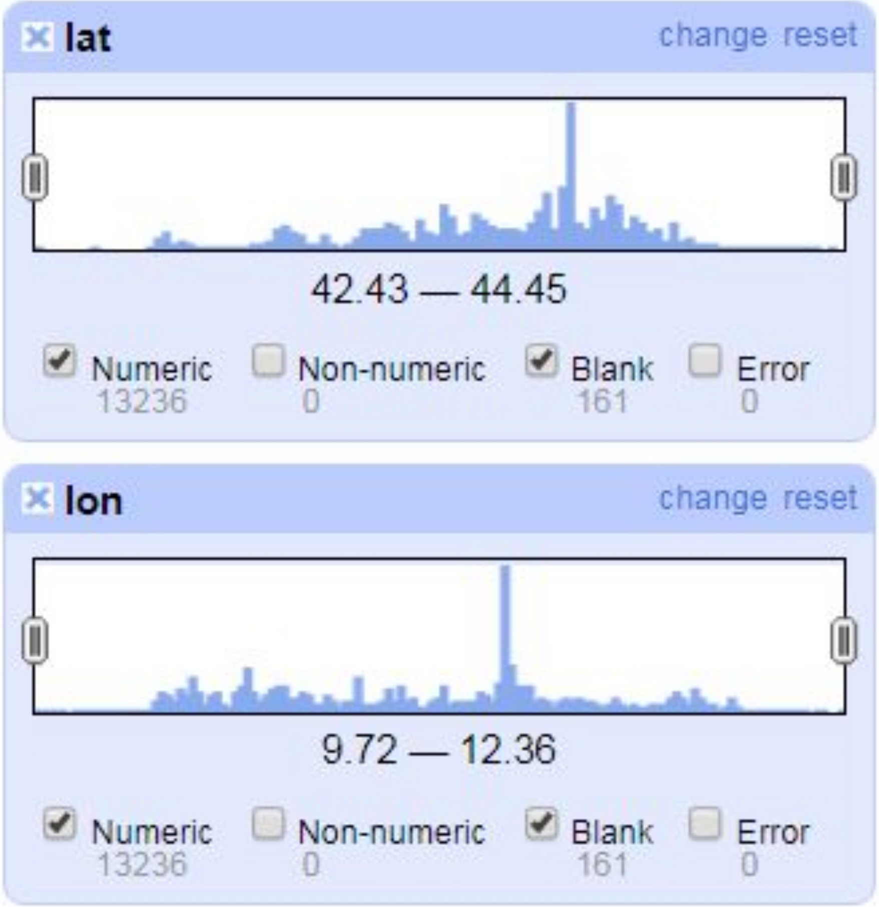
44.62

9.68

Lunigiana
Garfagnana
Montagna Pistoiese
Val di Bisenzio
Mugello
Media Valle
Versilia
Valdinievole
Piana Lucchese
Piana Fiorentina-Pratese-Pistoiese
Valdisieve
Casentino
Valdarno Inferiore
Valtiberina
Empolese
Valdarno Superiore
Valdera
Chianti
Val di Cecina
Valdelsa
Valdichiana
Crete Senesi
Val di Merse
Val di Cornia
Colline Metallifere
Val d'Orcia
Arcipelago Toscano
Amiata
Maremma

12.36

42.23



☒ **lat** change reset

42.43 — 44.45

☑ Numeric 13236 □ Non-numeric 0 ☑ Blank 161 □ Error 0

☒ **lon** change reset

9.72 — 12.36

☑ Numeric 13236 □ Non-numeric 0 ☑ Blank 161 □ Error 0

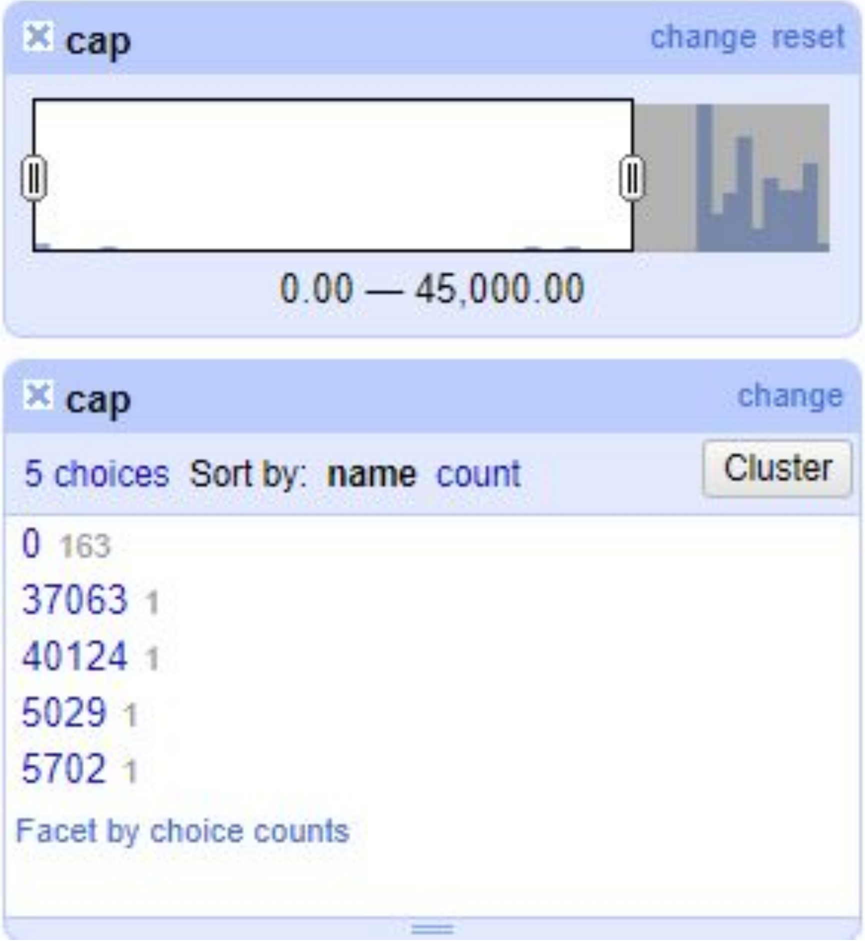
Combining Facet

Select Numeric and then Text facet

Interact on numeric facet to isolate the wrong zip codes

On the Text facet discover the wrong zip codes

Clik on “0” to see the 163 accomodations without zip code



0.00 — 45,000.00

5 choices

0	163
37063	1
40124	1
5029	1
5702	1

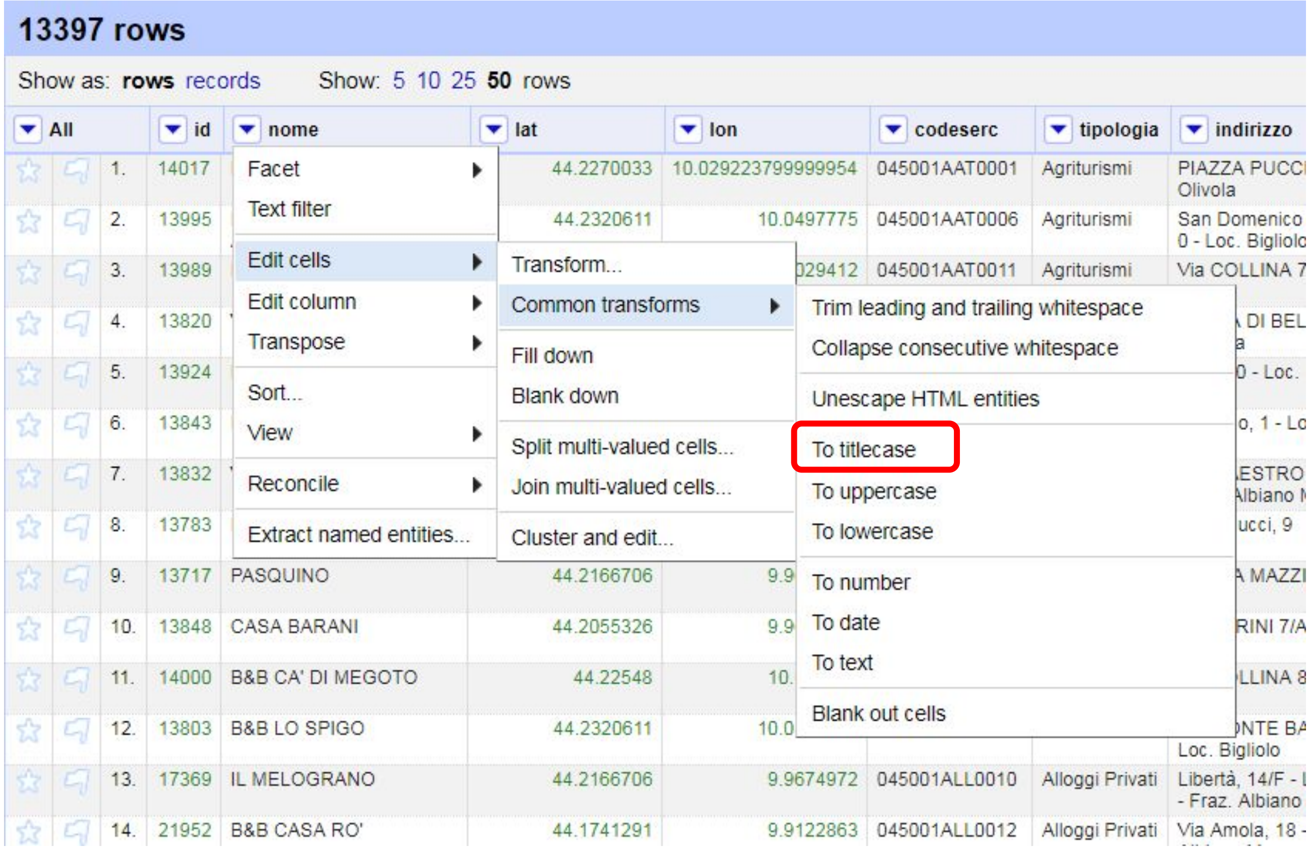
Data Transformation

Data Transformation Overview

- edit cell contents within a particular column
- edit Columns contents such as split or join columns
- add new columns based on existing data, with fetching new information, or through reconciliation
- convert your rows of data into multi-row records.

Edit Cells

Edit cells ◗ Common transforms



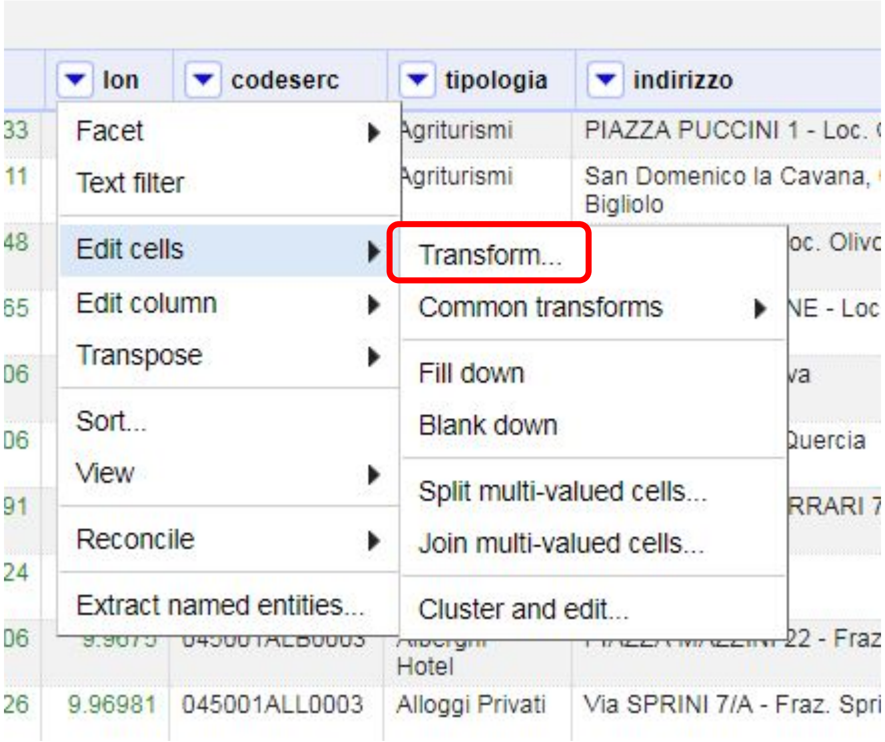
13397 rows

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

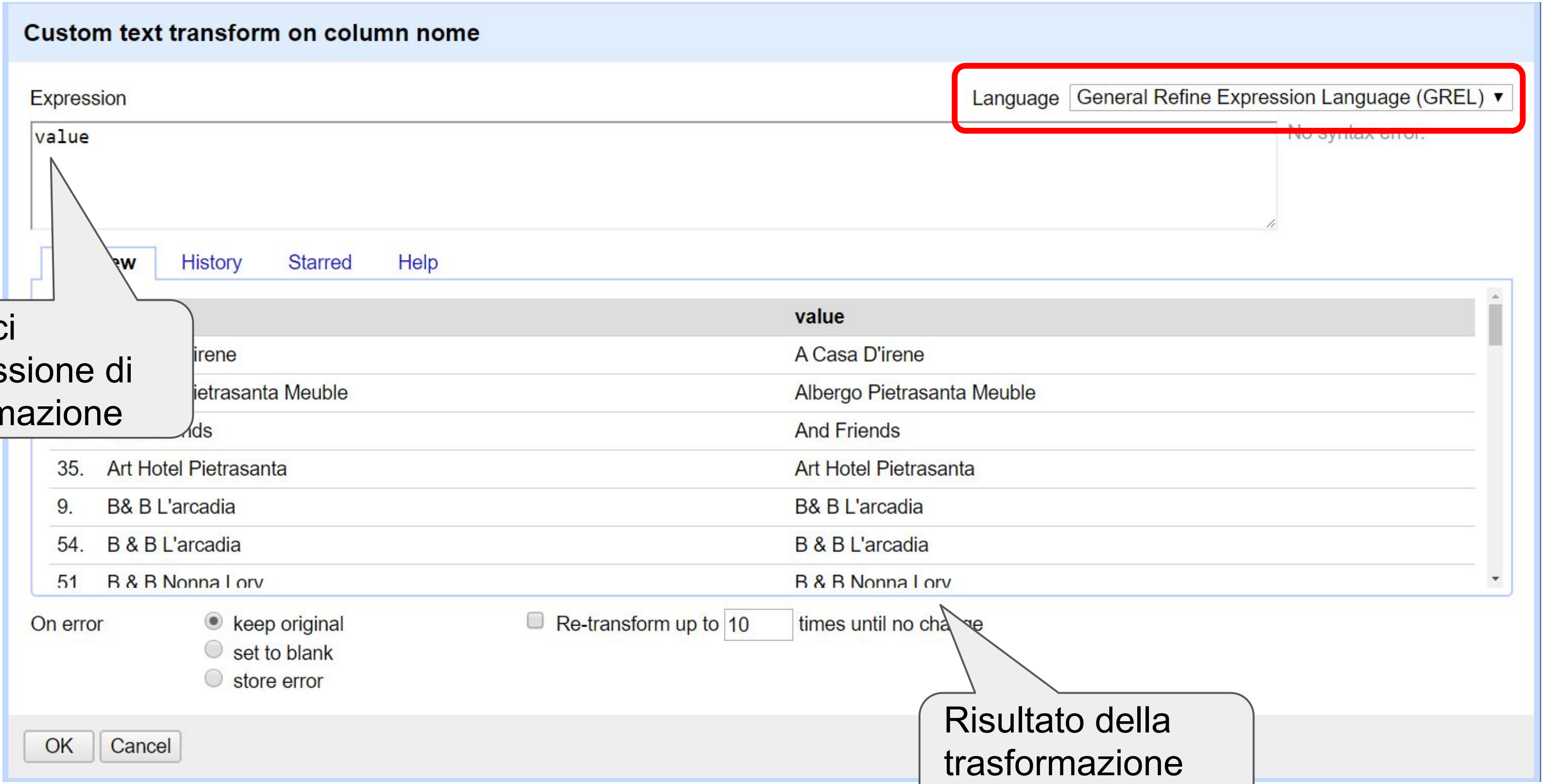
All			id	nome	lat	lon	codeserc	tipologia	indirizzo
☆		1.	14017		44.2270033	10.029223799999954	045001AAT0001	Agriturismi	PIAZZA PUCC Olivola
☆		2.	13995		44.2320611	10.0497775	045001AAT0006	Agriturismi	San Domenico 0 - Loc. Bigliolo
☆		3.	13989			029412	045001AAT0011	Agriturismi	Via COLLINA 7
☆		4.	13820						DI BEL a
☆		5.	13924						0 - Loc.
☆		6.	13843						o, 1 - Lo
☆		7.	13832						ESTRO lbiano I
☆		8.	13783						ucci, 9
☆		9.	13717	PASQUINO	44.2166706	9.9			A MAZZI
☆		10.	13848	CASA BARANI	44.2055326	9.9			RINI 7/A
☆		11.	14000	B&B CA' DI MEGOTO	44.22548	10.			LLINA 8
☆		12.	13803	B&B LO SPIGO	44.2320611	10.0			NTE BA Loc. Bigliolo
☆		13.	17369	IL MELOGRANO	44.2166706	9.9674972	045001ALL0010	Alloggi Privati	Libertà, 14/F - I - Fraz. Albiano
☆		14.	21952	B&B CASA RO'	44.1741291	9.9122863	045001ALL0012	Alloggi Privati	Via Amola, 18 -

To title case

Edit cells ◗ Transforms



Edit cells ➧ Transforms



Custom text transform on column nome

Expression

```
value
```

Language: General Refine Expression Language (GREL) ▾

No syntax error.

		value
	irene	A Casa D'irene
	ietrasanta Meuble	Albergo Pietrasanta Meuble
	ıds	And Friends
35.	Art Hotel Pietrasanta	Art Hotel Pietrasanta
9.	B& B L'arcadia	B& B L'arcadia
54.	B & B L'arcadia	B & B L'arcadia
51	B & B Nonna Lory	B & B Nonna Lory

On error: ● keep original ○ set to blank ○ store error

□ Re-transform up to 10 times until no cha ge

Inserisci l’espressione di trasformazione

Risultato della trasformazione su tutte le celle

General Refine Expression Language - GREL

System Variables

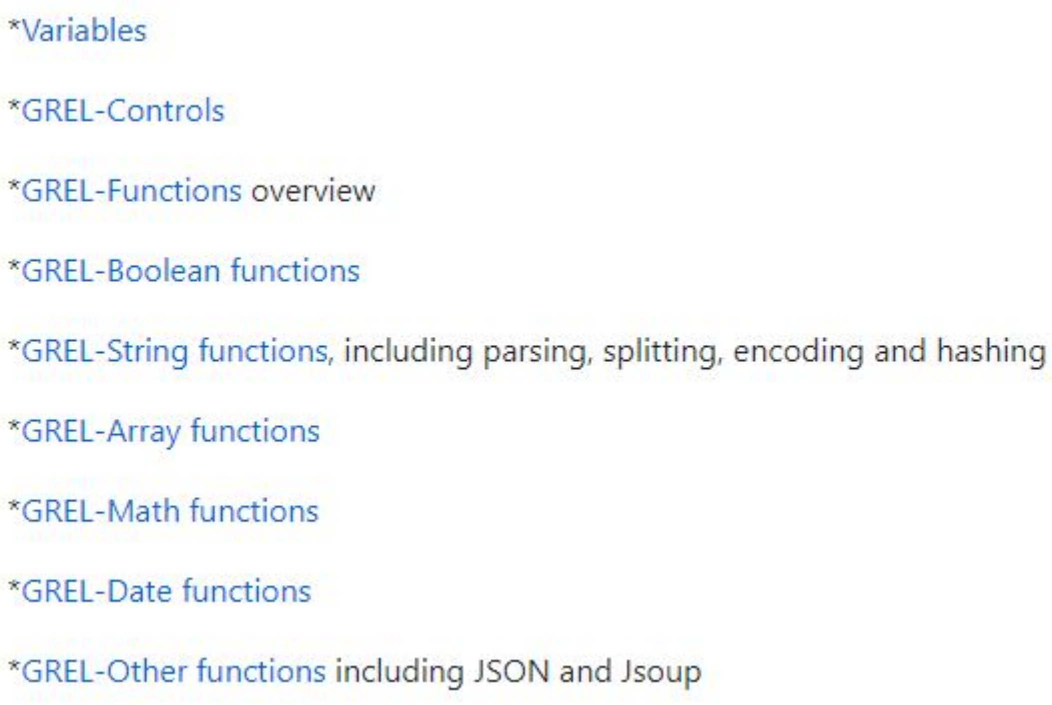
value = value of current cell

cells = cells of the current row
(**cells**.nome.**value**)

Functions

split("division character")

round() = round up



*Variables

*GREL-Controls

*GREL-Functions overview

*GREL-Boolean functions

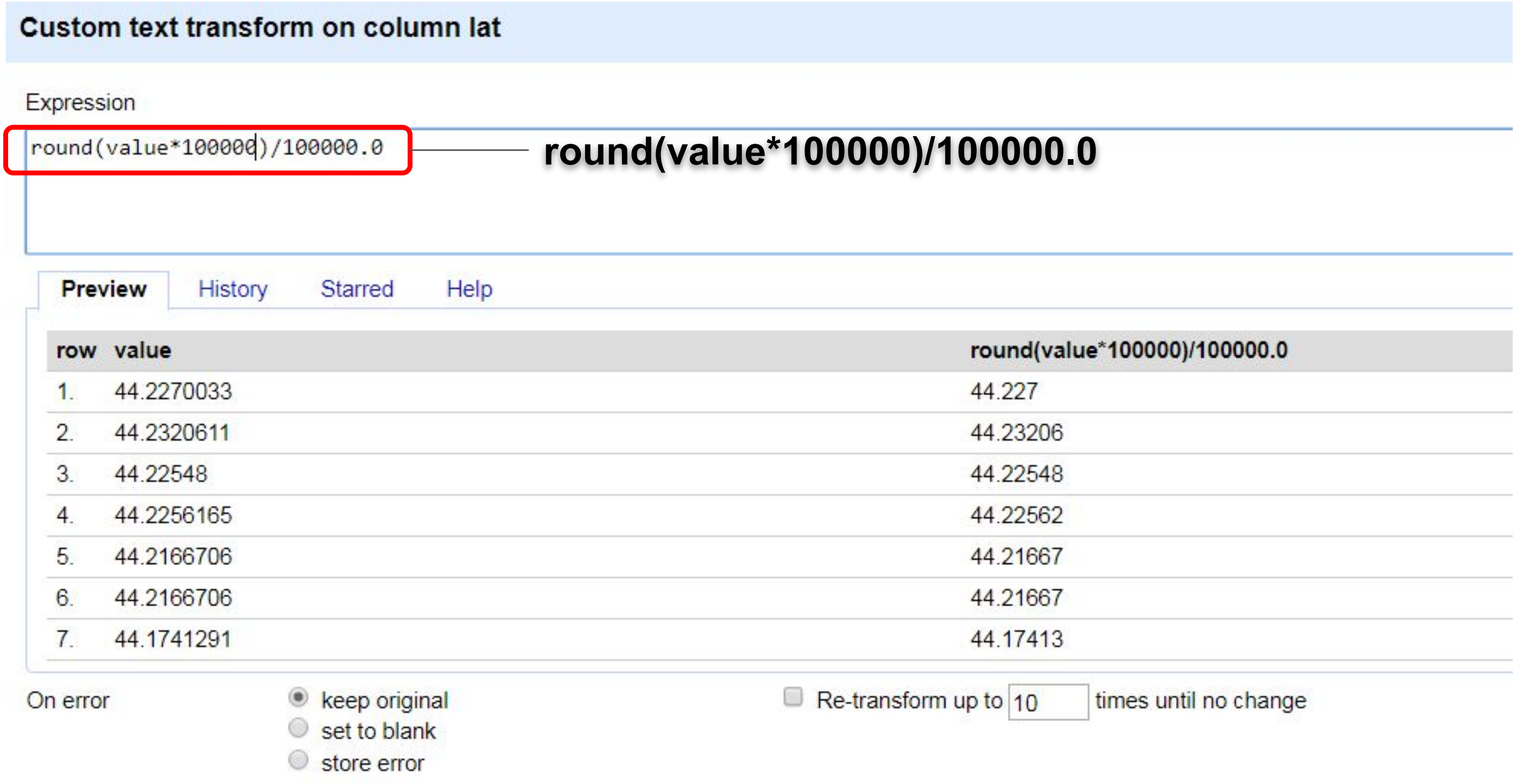
*GREL-String functions, including parsing, splitting, encoding and hashing

*GREL-Array functions

*GREL-Math functions

*GREL-Date functions

*GREL-Other functions including JSON and Jsoup



Custom text transform on column lat

Expression

```
round(value*100000)/100000.0
```

round(value*100000)/100000.0

row	value	round(value*100000)/100000.0
1.	44.2270033	44.227
2.	44.2320611	44.23206
3.	44.22548	44.22548
4.	44.2256165	44.22562
5.	44.2166706	44.21667
6.	44.2166706	44.21667
7.	44.1741291	44.17413

On error

● keep original
○ set to blank
○ store error

□ Re-transform up to 10 times until no change

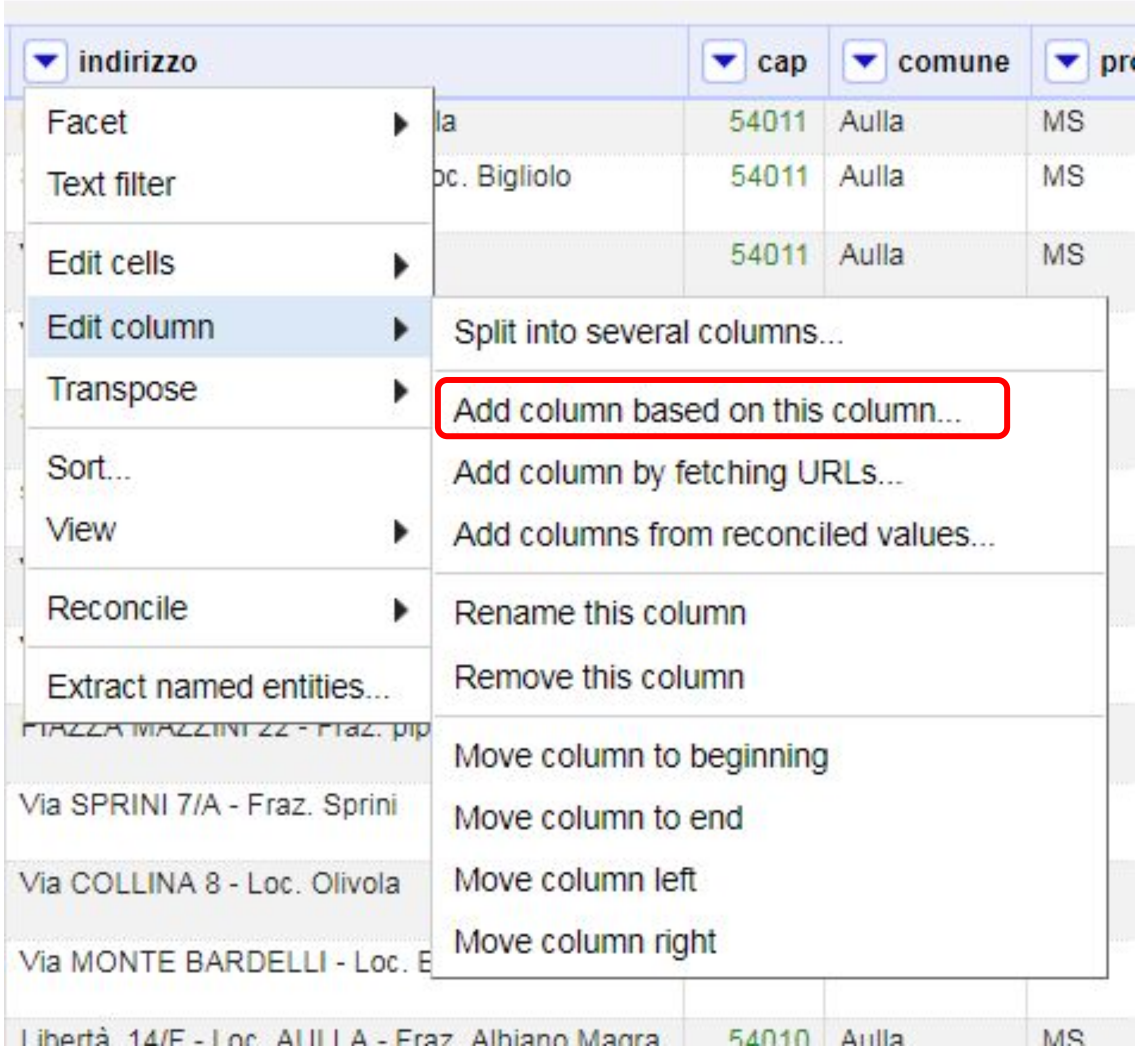
Edit columns

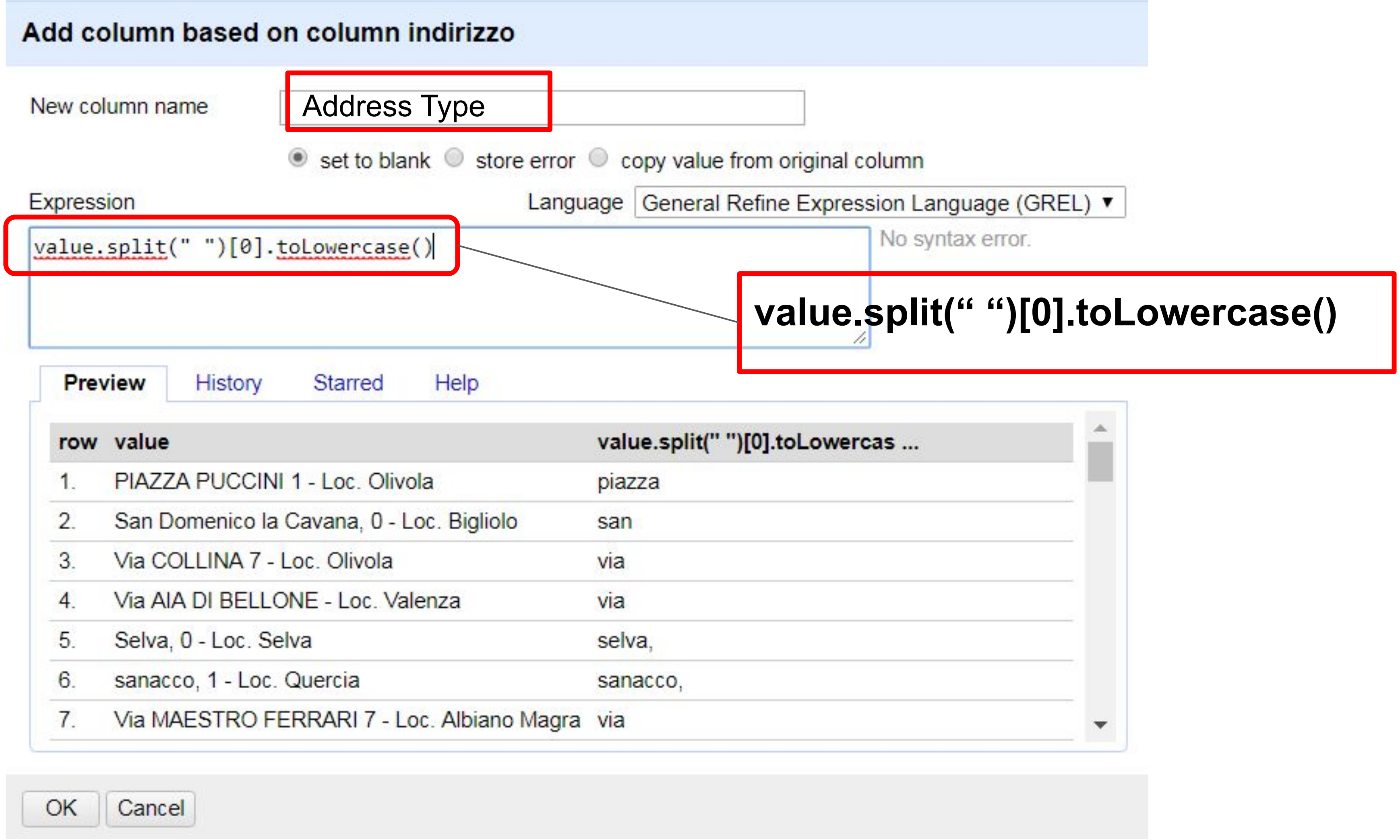
Suppose we want to investigate the type of road on which the accommodation is located

For example, on a state or provincial road, on a road or avenue, etc.

I work on the first word of the address

S.P. Avenza Carrara, 180 - Loc. Avenza


indirizzo
cap
comune
pr
Facet
Text filter
Edit cells
Edit column
Transpose
Sort...
View
Reconcile
Extract named entities...
Split into several columns...
Add column based on this column...
Add column by fetching URLs...
Add columns from reconciled values...
Rename this column
Remove this column
Move column to beginning
Move column to end
Move column left
Move column right
la
oc. Bigliolo
54011
Aulla
MS
54011
Aulla
MS
54011
Aulla
MS
PIAZZA MAZZINI 22 - Fraz. pip
Via SPRINI 7/A - Fraz. Sprini
Via COLLINA 8 - Loc. Olivola
Via MONTE BARDELLI - Loc. B
Libertà, 14/F - Loc. AULLA - Fraz. Albiano Magra
54010
Aulla
MS



Add column based on column indirizzo

New column name: Address Type

● set to blank ○ store error ○ copy value from original column

Expression

Language: General Refine Expression Language (GREL) ▾

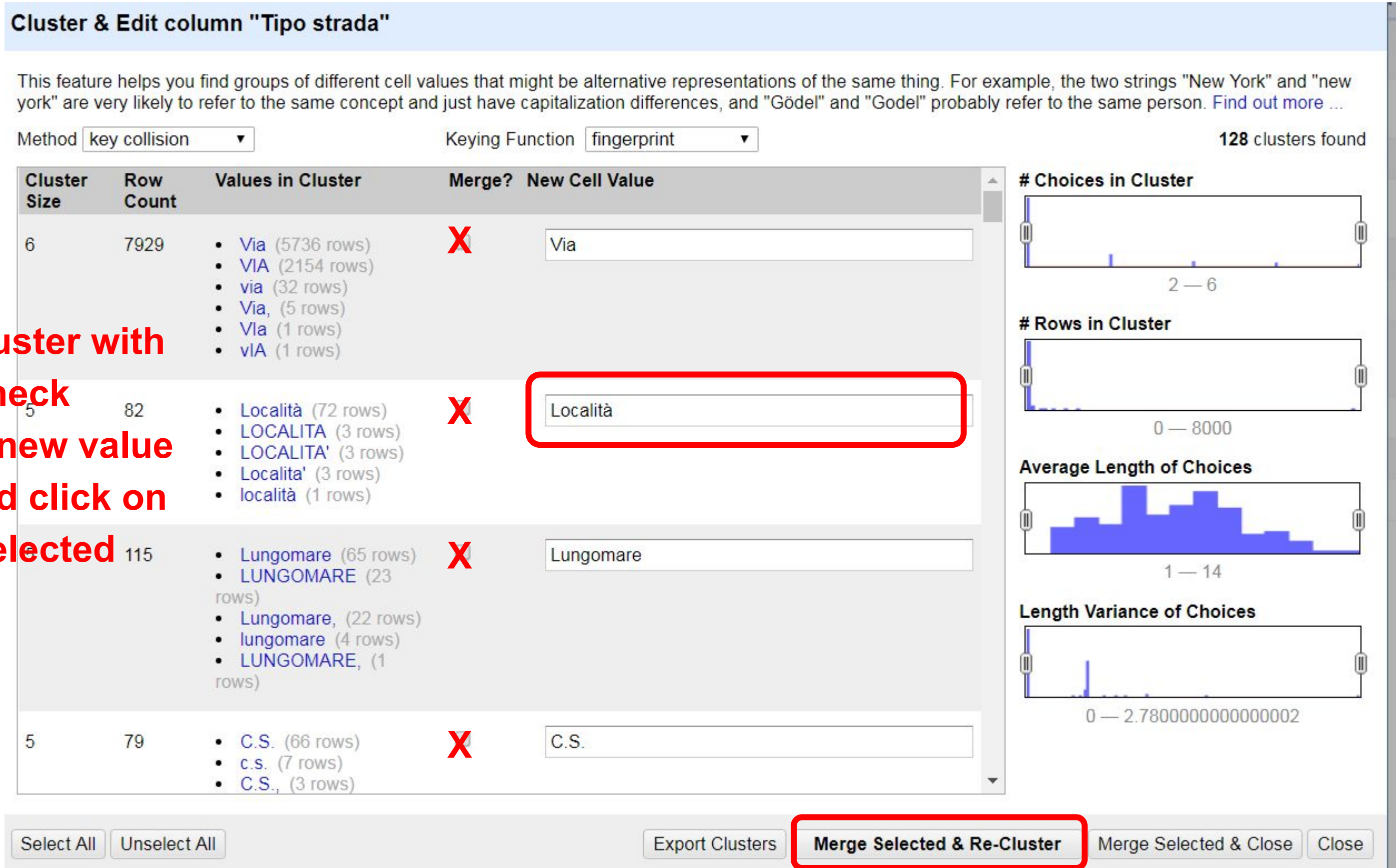
```
value.split(" ")[0].toLowercase()
```

No syntax error.

value.split(“ “)[0].toLowercase()

row	value	value.split(" ")[0].toLowercas ...
1.	PIAZZA PUCCINI 1 - Loc. Olivola	piazza
2.	San Domenico la Cavana, 0 - Loc. Bigliolo	san
3.	Via COLLINA 7 - Loc. Olivola	via
4.	Via AIA DI BELLONE - Loc. Valenza	via
5.	Selva, 0 - Loc. Selva	selva,
6.	sanacco, 1 - Loc. Quercia	sanacco,
7.	Via MAESTRO FERRARI 7 - Loc. Albiano Magra	via

Facet text and cluster on new column



Cluster & Edit column "Tipo strada"

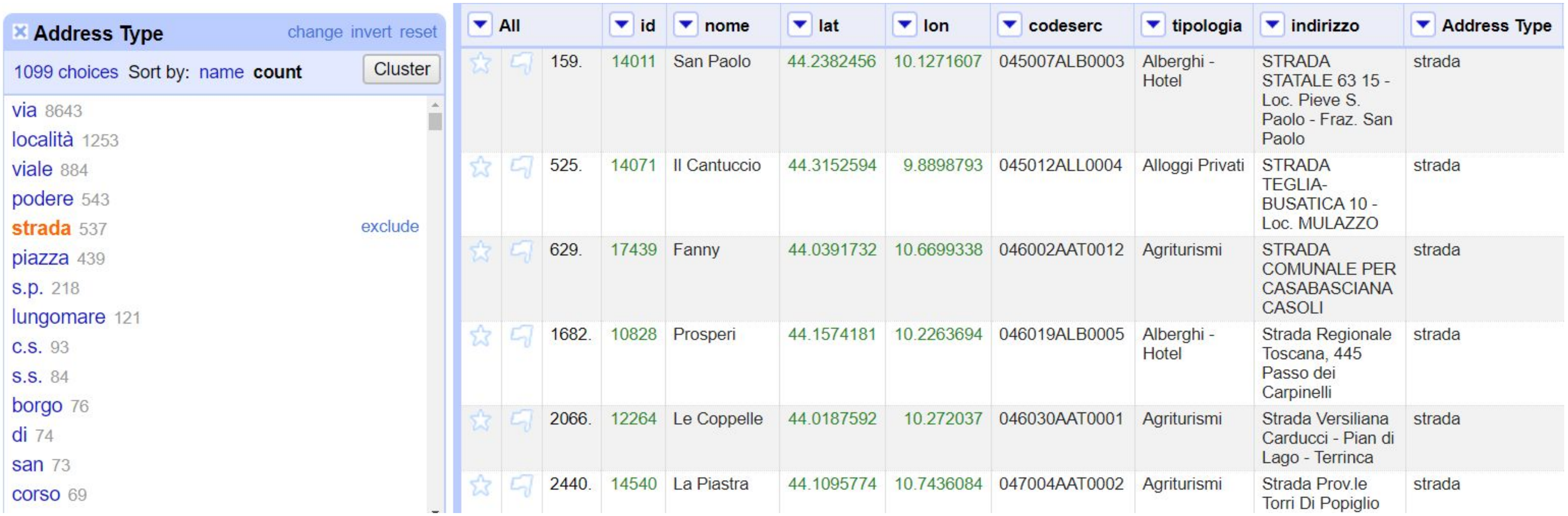
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer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

Method: key collision ▾ Keying Function: fingerprint ▾ **128** clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
6	7929	Via (5736 rows) VIA (2154 rows) via (32 rows) Via, (5 rows) VIa (1 rows) vIA (1 rows)	X	Via
5	82	Località (72 rows) LOCALITA (3 rows) LOCALITA' (3 rows) Localita' (3 rows) località (1 rows)	X	Località
	115	Lungomare (65 rows) LUNGOMARE (23 rows) Lungomare, (22 rows) lungomare (4 rows) LUNGOMARE, (1 rows)	X	Lungomare
5	79	C.S. (66 rows) c.s. (7 rows) C.S., (3 rows)	X	C.S.

1. select cluster with merge check
2. type the new value
3. at the end click on Merge Selected

Check the results with facet text on new column



Address Type change invert reset

1099 choices Sort by: name **count** Cluster

via	8643
località	1253
viale	884
podere	543
strada	537 exclude
piazza	439
s.p.	218
lungomare	121
c.s.	93
s.s.	84
borgo	76
di	74
san	73
corso	69

All	id	nome	lat	lon	codeserc	tipologia	indirizzo	Address Type
☆ 159.	14011	San Paolo	44.2382456	10.1271607	045007ALB0003	Alberghi - Hotel	STRADA STATALE 63 15 - Loc. Pieve S. Paolo - Fraz. San Paolo	strada
☆ 525.	14071	Il Cantuccio	44.3152594	9.8898793	045012ALL0004	Alloggi Privati	STRADA TEGLIA-BUSATICA 10 - Loc. MULAZZO	strada
☆ 629.	17439	Fanny	44.0391732	10.6699338	046002AAT0012	Agriturismi	STRADA COMUNALE PER CASABASCIANA CASOLI	strada
☆ 1682.	10828	Prosperi	44.1574181	10.2263694	046019ALB0005	Alberghi - Hotel	Strada Regionale Toscana, 445 Passo dei Carpinelli	strada
☆ 2066.	12264	Le Coppelle	44.0187592	10.272037	046030AAT0001	Agriturismi	Strada Versiliana Carducci - Pian di Lago - Terrinca	strada
☆ 2440.	14540	La Piastra	44.1095774	10.7436084	047004AAT0002	Agriturismi	Strada Prov.le Torri Di Popiglio	strada

Data enrichment/augmentation

Why

Fill missing data

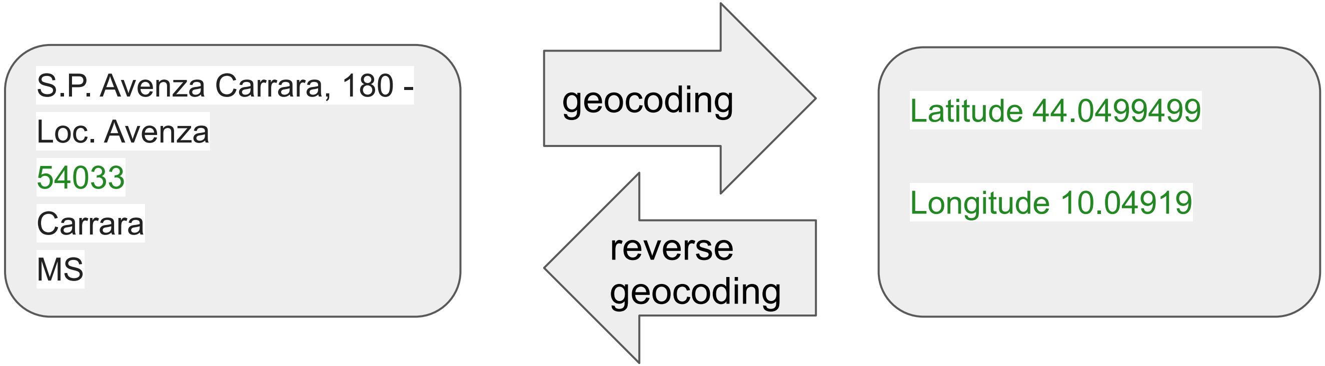
Validate data

In both case you need a **Gold Standard** or **Ground Truth**, something that returns the exact value

Geocoding Vs Reverse Geocoding

Geocoding is the conversion from address to coordinates

Reverse geocoding is the opposite



S.P. Avenza Carrara, 180 -
Loc. Avenza
54033
Carrara
MS

geocoding

reverse geocoding

Latitude 44.0499499

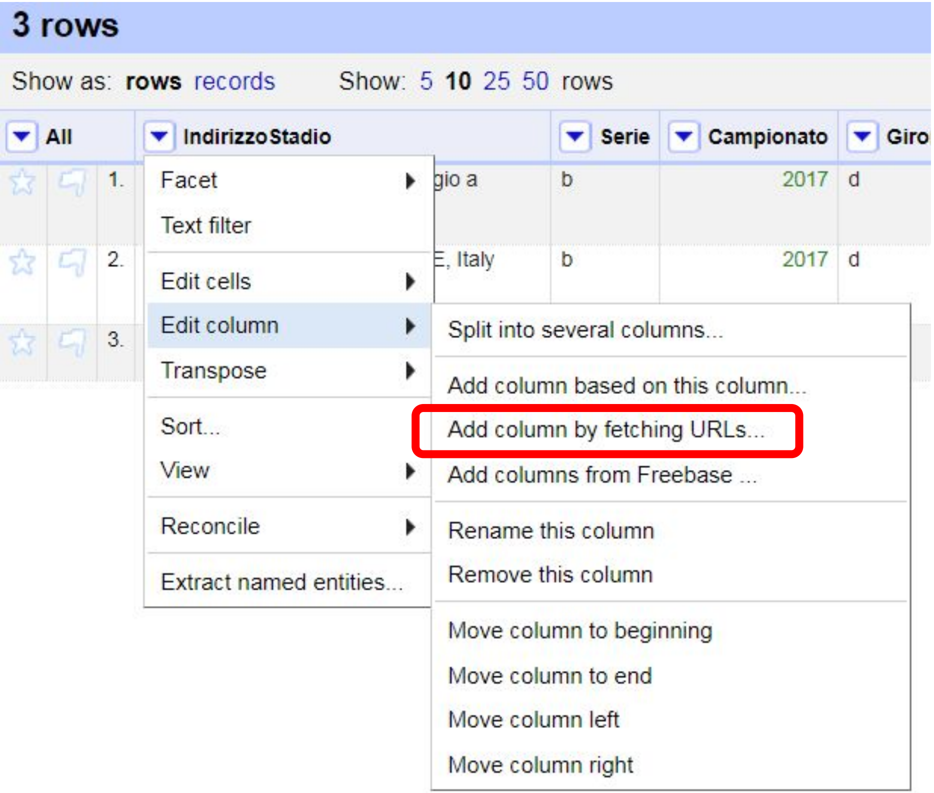
Longitude 10.04919

Openstreetmap Json Result

```
[

          {
                    place_id: "16952760",
                    licence: "Data © OpenStreetMap contributors, ODbL 1.0. https://osm.org/copyright",
                    osm_type: "node",
                    osm_id: "1477804118",
                    boundingbox:
                    [
                              "43.7193809",
                              "43.7194809",
                              "10.4237241",
                              "10.4238241"
                    ],
                    lat: "43.7194309",
                    lon: "10.4237741",
                    display_name: "Area della Ricerca del CNR di Pisa, 1, Via Giuseppe Moruzzi, Don Bosco,
                    Pisa, PI, Tuscany, 56124, Italia",
                    class: "place",
                    type: "house",
                    importance: 0.52025
          }
]
```

Data Enrichment with Open Refine



Edit column ➧ Add a new column by fetching URLs

GeoCoding Service - Web API

Nominatim (*based on Open Street Map DB*) - Documentation

https://nominatim.openstreetmap.org/search?*q*=**Area della Ricerca di Pisa, Via Giuseppe Moruzzi, 1, 56124, Pisa, Tuscany, Italy**

MapBox Documentation

https://api.mapbox.com/geocoding/v5/mapbox.places/**Area della Ricerca di Pisa, Via Giuseppe Moruzzi, 1, 56124, Pisa, Tuscany, Italy**.json?access_token=pk.eyJ1IjoiYXF1YWJsdWUiLCJhIjoiY2ttZXFnbmR0MnljODJ2bnc2Znp0bGc3MCJ9.YmF_-yyqzeCeoePdOSob7g

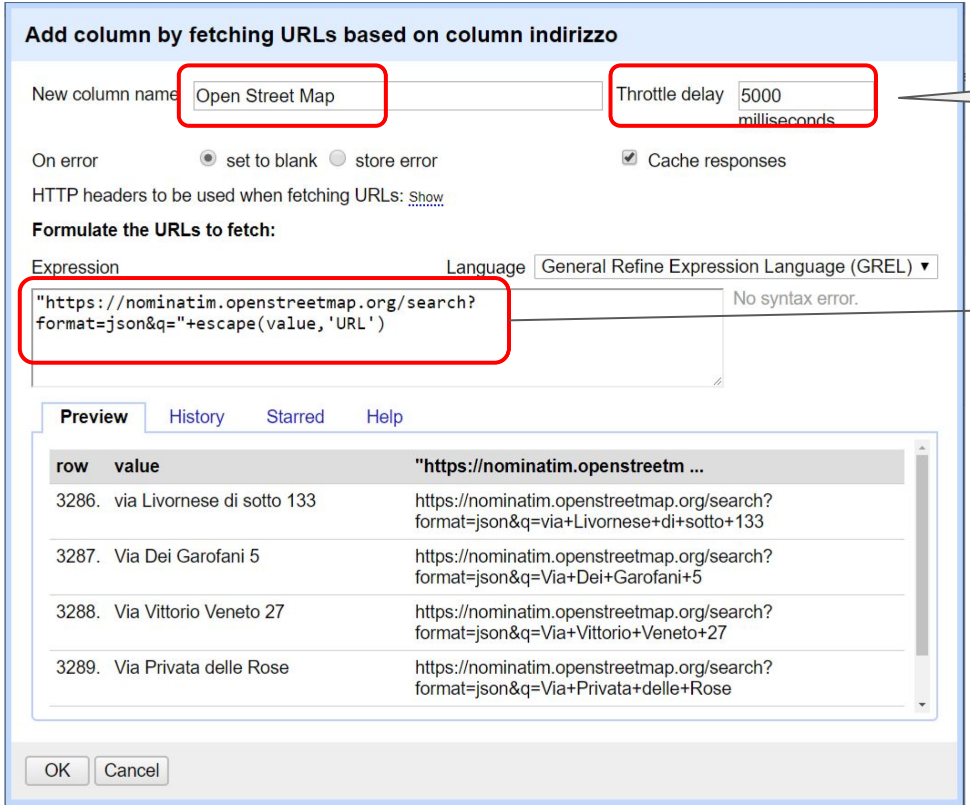
GoogleMap

https://maps.googleapis.com/maps/api/geocode/json?address=**Via Moruzzi 1 Pisa&key=YOUR API KEY**

BingMap

http://dev.virtualearth.net/REST/v1/Locations?countryRegion=IT&locality=Pietrasanta&postalCode=55045&addressLine=Via Provinciale Vallecchia, 85&maxResults=10&key=YOUR API KEY

Call the Geocoding Service



Add column by fetching URLs based on column indirizzo

New column name: Open Street Map

Throttle delay: 5000 milliseconds

On error ● set to blank ○ store error

☑ Cache responses

HTTP headers to be used when fetching URLs: Show

Formulate the URLs to fetch:

Expression

Language: General Refine Expression Language (GREL)

```
"https://nominatim.openstreetmap.org/search?
format=json&q="+escape(value,'URL')
```

No syntax error.

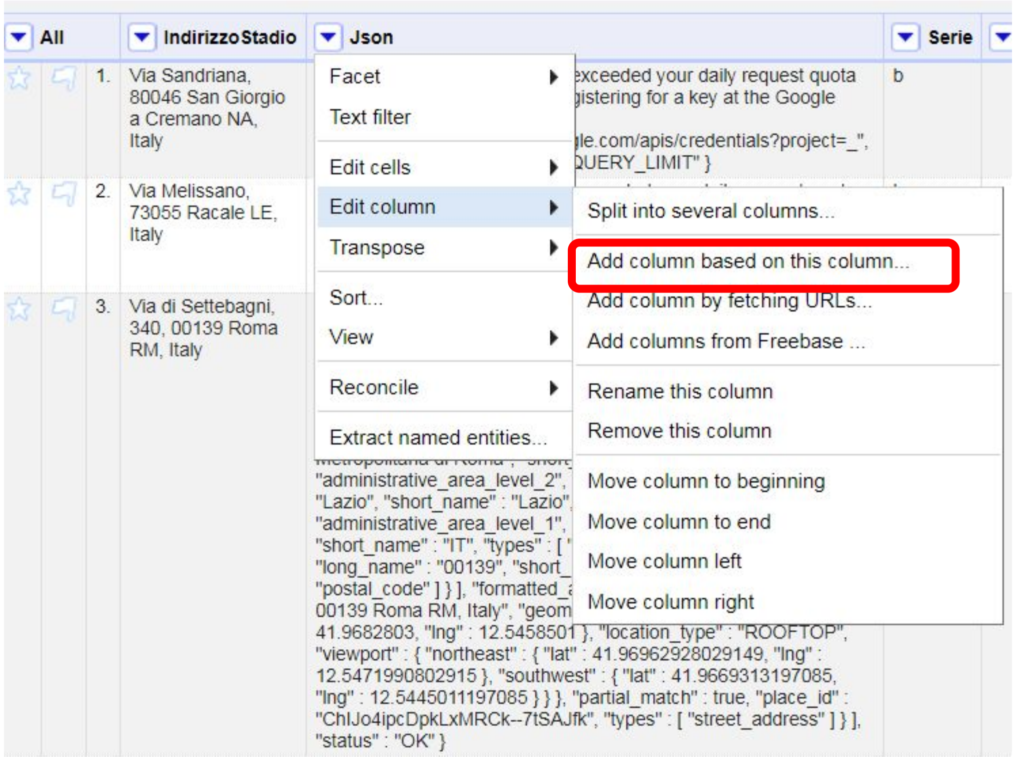
row	value	"https://nominatim.openstreetm ...
3286.	via Livornese di sotto 133	https://nominatim.openstreetmap.org/search?format=json&q=via+Livornese+di+sotto+133
3287.	Via Dei Garofani 5	https://nominatim.openstreetmap.org/search?format=json&q=Via+Dei+Garofani+5
3288.	Via Vittorio Veneto 27	https://nominatim.openstreetmap.org/search?format=json&q=Via+Vittorio+Veneto+27
3289.	Via Privata delle Rose	https://nominatim.openstreetmap.org/search?format=json&q=Via+Privata+delle+Rose

Attesa risposta

"https://nominatim.openstreetmap.org/search?**format=json**&q="+**escape(value,'URL')**

Escape: GREL string function

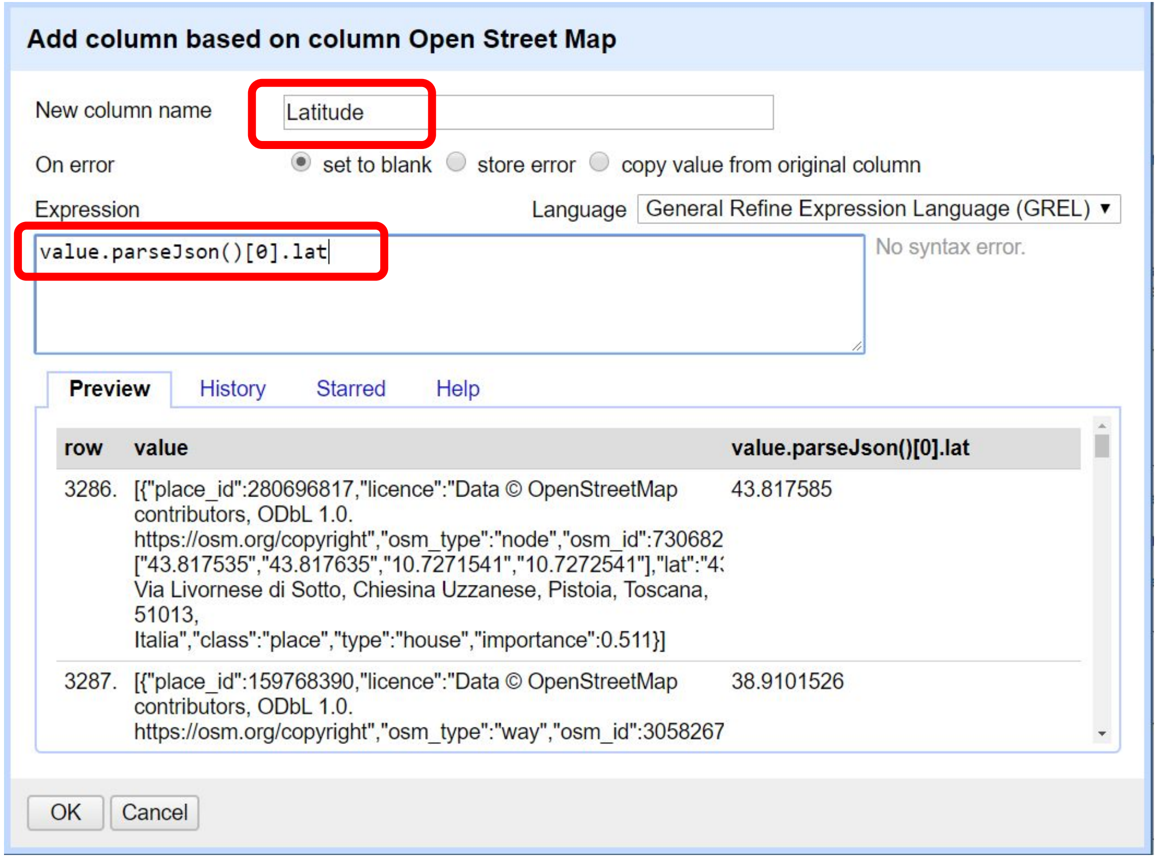
Extract Latitude from Json



All	IndirizzoStadio	Json	Serie
1.	Via Sandriana, 80046 San Giorgio a Cremano NA, Italy	xceeded your daily request quota gistering for a key at the Google le.com/apis/credentials?project=_", QUERY_LIMIT" }	b
2.	Via Melissano, 73055 Racale LE, Italy		
3.	Via di Settebagni, 340, 00139 Roma RM, Italy	"administrative_area_level_2", "Lazio", "short_name" : "Lazio", "administrative_area_level_1", "short_name" : "IT", "types" : [" "long_name" : "00139", "short_ "postal_code"] }], "formatted_ 00139 Roma RM, Italy", "geom 41.9682803, "lng" : 12.5458501 }, "location_type" : "ROOFTOP", "viewport" : { "northeast" : { "lat" : 41.96962928029149, "lng" : 12.5471990802915 }, "southwest" : { "lat" : 41.9669313197085, "lng" : 12.5445011197085 } } }, "partial_match" : true, "place_id" : "ChIJo4ipcDpkLxMRCk--7tSAJfk", "types" : ["street_address"] }], "status" : "OK" }	

I need to create a new column based on Json column

Extract Latitude from Json



Add column based on column Open Street Map

New column name: Latitude

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

Expression

Language: General Refine Expression Language (GREL) ▼

```
value.parseJson()[0].lat
```

No syntax error.

row	value	value.parseJson()[0].lat
3286.	[{"place_id":280696817,"licence":"Data © OpenStreetMap contributors, ODbL 1.0. https://osm.org/copyright","osm_type":"node","osm_id":730682 ["43.817535","43.817635","10.7271541","10.7272541"],"lat":"4 Via Livornese di Sotto, Chiesina Uzzanese, Pistoia, Toscana, 51013, Italia","class":"place","type":"house","importance":0.511}]	43.817585
3287.	[{"place_id":159768390,"licence":"Data © OpenStreetMap contributors, ODbL 1.0. https://osm.org/copyright","osm_type":"way","osm_id":3058267	38.9101526

Grel Function

value.**parseJson()**[0].lat

The Geocoding Service returns always a list of results, we get the first one: [0]

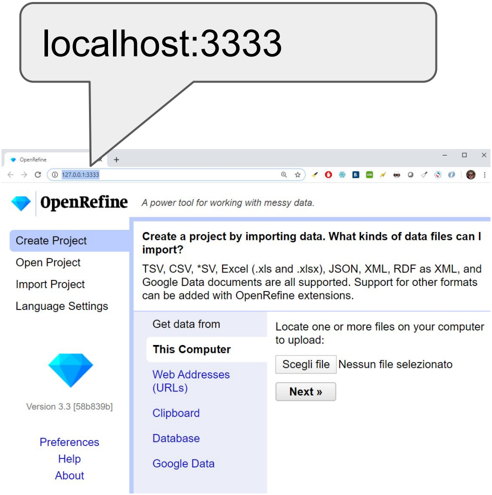
Reconciling

Reconciling

Reconciliation is the process of matching your dataset with that of an **external authoritative source**

To reconcile your OpenRefine project against an external dataset, that dataset must offer a web service that conforms to the Reconciliation Service API standards.

OpenRefine & Reconciling



localhost:3333

OpenRefine

127.0.0.1:3333

OpenRefine *A power tool for working with messy data.*

- Create Project
- Open Project
- Import Project
- Language Settings

Version 3.3 [58b839b]

Preferences
Help
About

Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with OpenRefine extensions.

Get data from

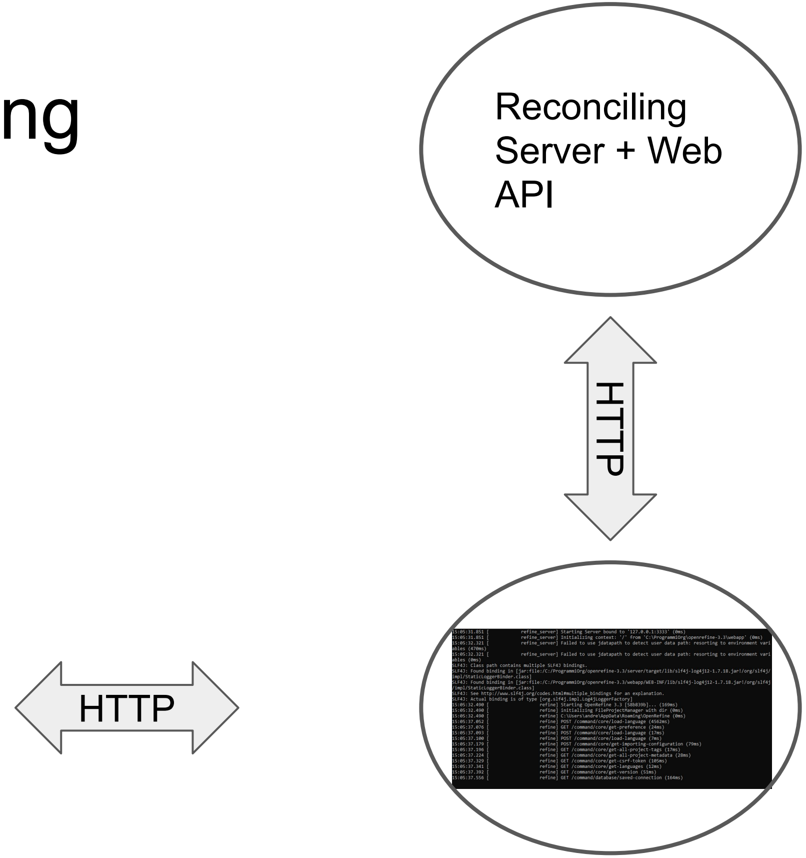
- **This Computer**
- Web Addresses (URLs)
- Clipboard
- Database
- Google Data

Locate one or more files on your computer to upload:

Scegli file Nessun file selezionato

Next »

Web Browser



```
15:05:31.851 [          refine_server] Starting Server bound to '127.0.0.1:3333' (0ms)
15:05:31.851 [          refine_server] Initializing context: '/' from 'C:\ProgrammiOrg\openrefine-3.3\webapp' (0ms)
15:05:32.321 [          refine_server] Failed to use jdatapath to detect user data path: resorting to environment vari
ables (470ms)
15:05:32.321 [          refine_server] Failed to use jdatapath to detect user data path: resorting to environment vari
ables (0ms)
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/C:/ProgrammiOrg/openrefine-3.3/server/target/lib/slf4j-log4j12-1.7.18.jar!/org/slf4j/
impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/C:/ProgrammiOrg/openrefine-3.3/webapp/WEB-INF/lib/slf4j-log4j12-1.7.18.jar!/org/slf4j
/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
15:05:32.490 [                 refine] Starting OpenRefine 3.3 [58b839b]... (169ms)
15:05:32.490 [                 refine] initializing FileProjectManager with dir (0ms)
15:05:32.490 [                 refine] C:\Users\andre\AppData\Roaming\OpenRefine (0ms)
15:05:37.052 [                 refine] POST /command/core/load-language (4562ms)
15:05:37.076 [                 refine] GET /command/core/get-preference (24ms)
15:05:37.093 [                 refine] POST /command/core/load-language (17ms)
15:05:37.100 [                 refine] POST /command/core/load-language (7ms)
15:05:37.179 [                 refine] POST /command/core/get-importing-configuration (79ms)
15:05:37.196 [                 refine] GET /command/core/get-all-project-tags (17ms)
15:05:37.224 [                 refine] GET /command/core/get-all-project-metadata (28ms)
15:05:37.329 [                 refine] GET /command/core/get-csrf-token (105ms)
15:05:37.341 [                 refine] GET /command/core/get-languages (12ms)
15:05:37.392 [                 refine] GET /command/core/get-version (51ms)
15:05:37.556 [                 refine] GET /command/database/saved-connection (164ms)
```

Http server on localhost:3333

Reconciling Targets

- fix spelling or variations in proper names
- clean up manually-entered subject headings against authorities link your data to an existing dataset
- add to an editable platform such as Wikidata
- or see whether entities in your project appear in some specific list, such as the Panama Papers.

Reconciling

Semi-automated

OpenRefine matches your cell values to the reconciliation information as best it can, but human judgment is required to review and approve the results

Iterative

Reconcile multiple times with different settings, and with different subgroups of your data.

External Authoritative Sources

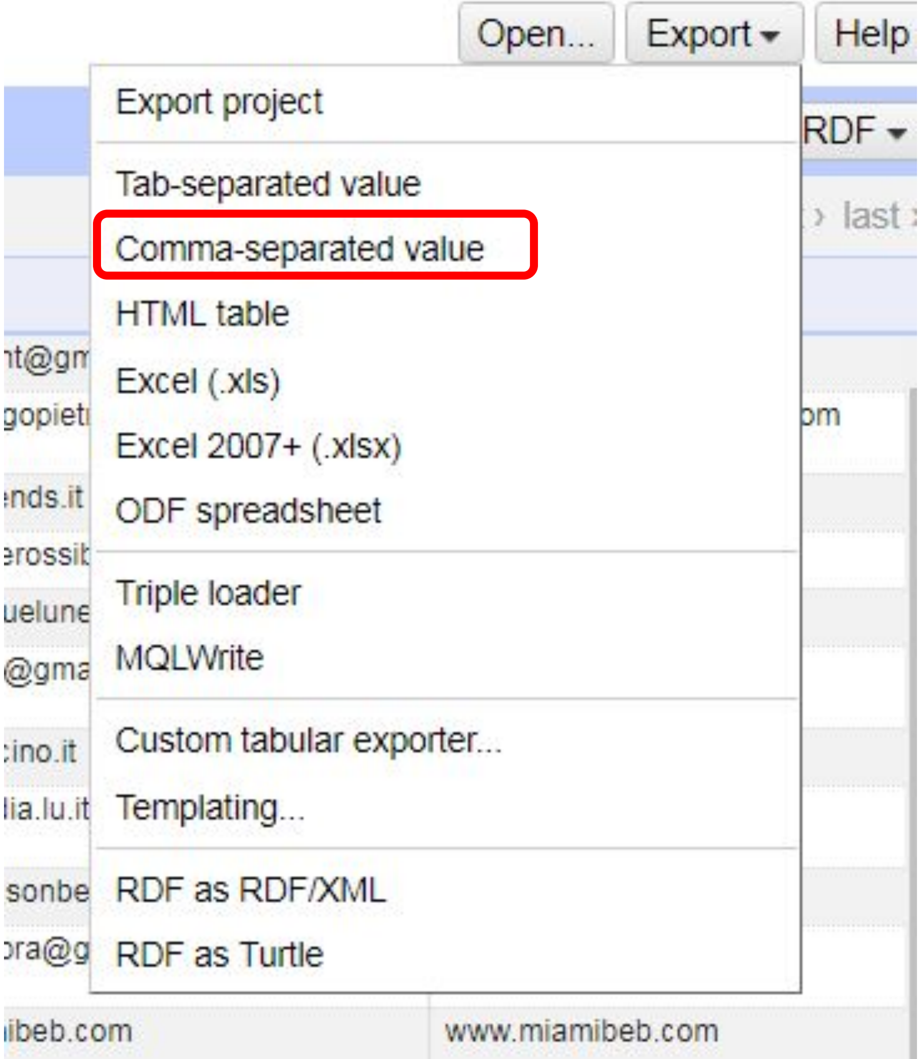
Use an existing reconciliation service

- list of reconcilable authorities
- further list of sources

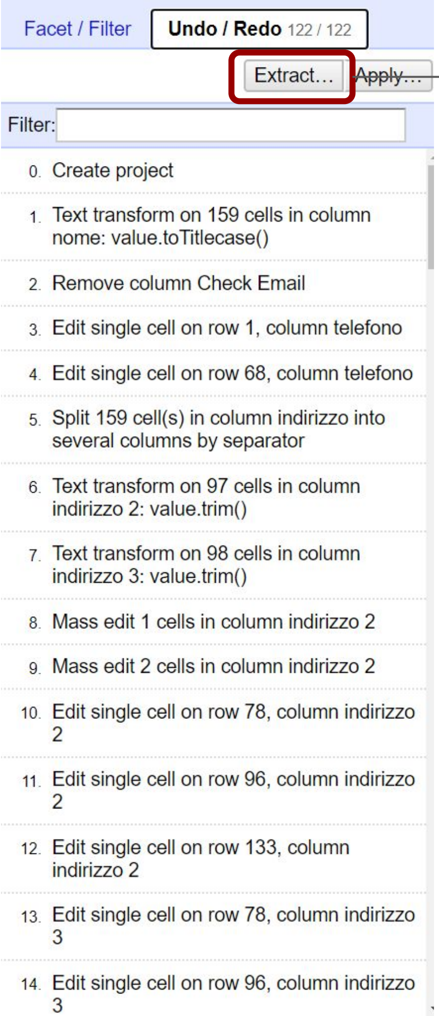
Build your reconciliation service from scratch

Build your reconciliation service from a simple CSV file

Export Data


Open...
Export
Help
RDF
last
Export project
Tab-separated value
Comma-separated value
HTML table
Excel (.xls)
Excel 2007+ (.xlsx)
ODF spreadsheet
Triple loader
MQLWrite
Custom tabular exporter...
Templating...
RDF as RDF/XML
RDF as Turtle
www.miamibeb.com

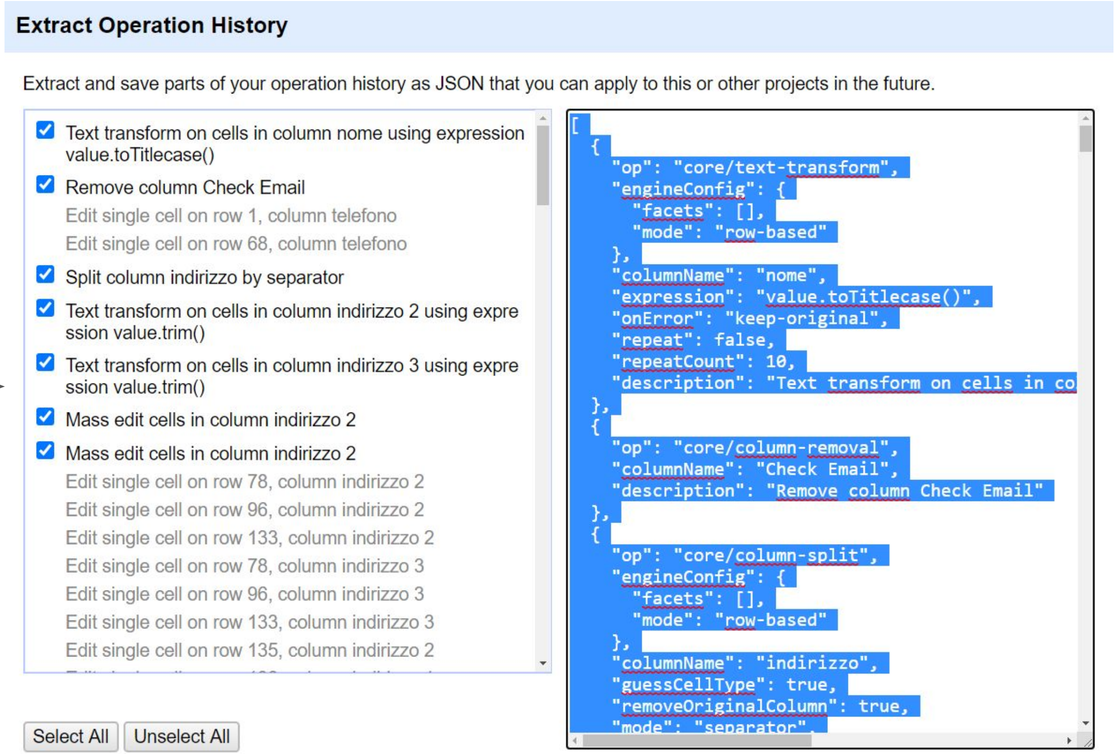
Create a Report



0. Create project
1. Text transform on 159 cells in column nome: value.toTitlecase()
2. Remove column Check Email
3. Edit single cell on row 1, column telefono
4. Edit single cell on row 68, column telefono
5. Split 159 cell(s) in column indirizzo into several columns by separator
6. Text transform on 97 cells in column indirizzo 2: value.trim()
7. Text transform on 98 cells in column indirizzo 3: value.trim()
8. Mass edit 1 cells in column indirizzo 2
9. Mass edit 2 cells in column indirizzo 2
10. Edit single cell on row 78, column indirizzo 2
11. Edit single cell on row 96, column indirizzo 2
12. Edit single cell on row 133, column indirizzo 2
13. Edit single cell on row 78, column indirizzo 3
14. Edit single cell on row 96, column indirizzo 3

The Undo/Redo panel contains the list of all operations made on the dataset

You can **extract** the list of operations in JSON format and save as a Report

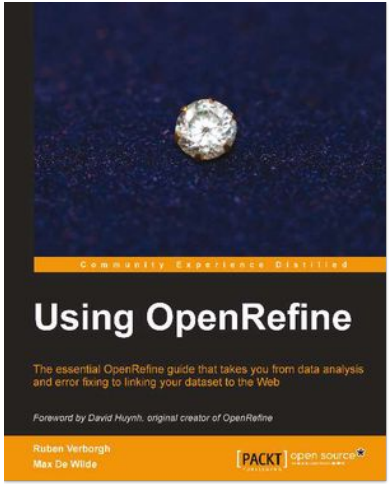

Extract Operation History

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

☑ Text transform on cells in column nome using expression value.toTitlecase()
☑ Remove column Check Email
Edit single cell on row 1, column telefono
Edit single cell on row 68, column telefono
☑ Split column indirizzo by separator
☑ Text transform on cells in column indirizzo 2 using expre ssion value.trim()
☑ Text transform on cells in column indirizzo 3 using expre ssion value.trim()
☑ Mass edit cells in column indirizzo 2
☑ Mass edit cells in column indirizzo 2
Edit single cell on row 78, column indirizzo 2
Edit single cell on row 96, column indirizzo 2
Edit single cell on row 133, column indirizzo 2
Edit single cell on row 78, column indirizzo 3
Edit single cell on row 96, column indirizzo 3
Edit single cell on row 133, column indirizzo 3
Edit single cell on row 135, column indirizzo 2

```
[
  {
    "op": "core/text-transform",
    "engineConfig": {
      "facets": [],
      "mode": "row-based"
    },
    "columnName": "nome",
    "expression": "value.toTitlecase()",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10,
    "description": "Text transform on cells in co
  },
  {
    "op": "core/column-removal",
    "columnName": "Check Email",
    "description": "Remove column Check Email"
  },
  {
    "op": "core/column-split",
    "engineConfig": {
      "facets": [],
      "mode": "row-based"
    },
    "columnName": "indirizzo",
    "guessCellType": true,
    "removeOriginalColumn": true,
    "mode": "separator",
```

Bibliography


Community Experience Distilled
Using OpenRefine
The essential OpenRefine guide that takes you from data analysis and error fixing to linking your dataset to the Web
Foreword by David Huynh, original creator of OpenRefine
Ruben Verborgh
Max De Wilde
[PACKT] open source

Open Refine Home page

Official Documentation

List of Tutorials

Using OpenRefine Ruben Verborgh, Max De Wilde September 2013

General Refine Expression Language

Jython = Python for java platform