



# Data Journalism

## Data Cleansing - Open Refine

# 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.

$$X \text{ POUNDS} \cdot 0.454 = Y \text{ KG}$$

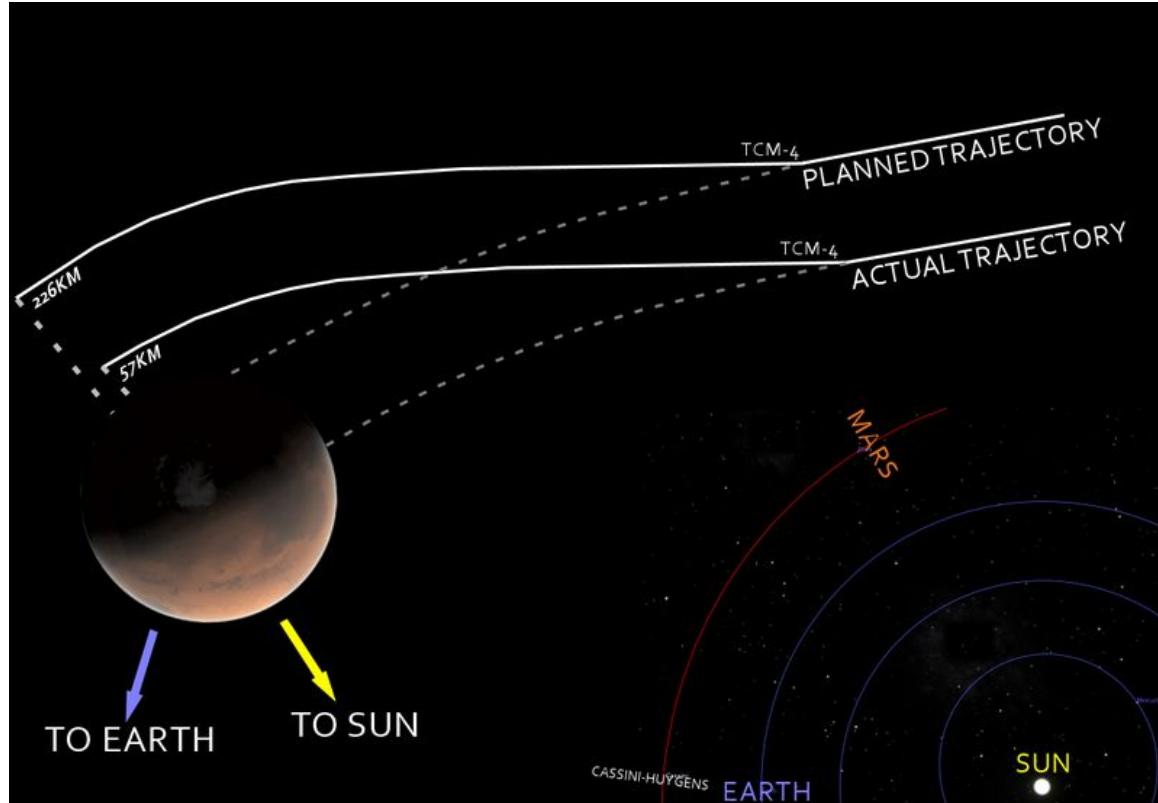
EXAMPLE:

$$100 \text{ lbs} = ? \text{ KG}$$

$$100 \cdot 0.454 = 45.4 \text{ KG}$$



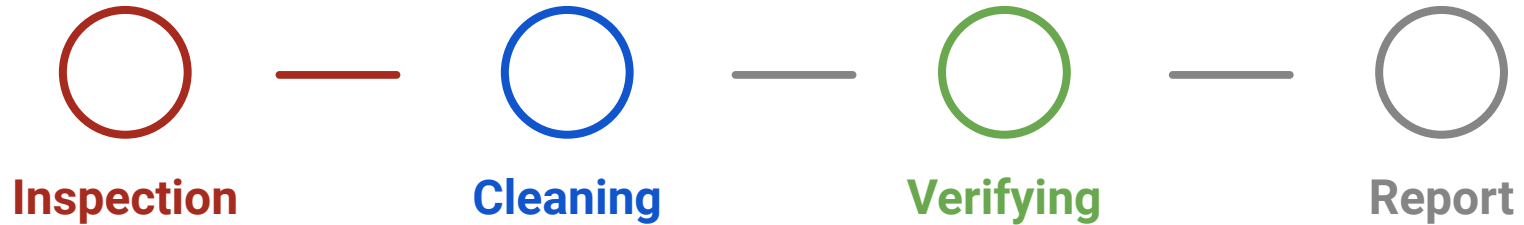
# Mars Climate Orbiter 1999



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

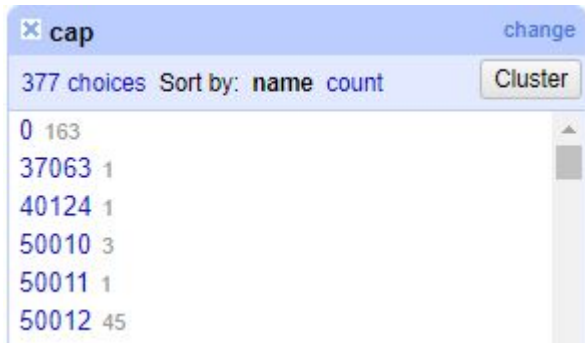
# The method



# Inspection

For each column calculate a Summary Statistics

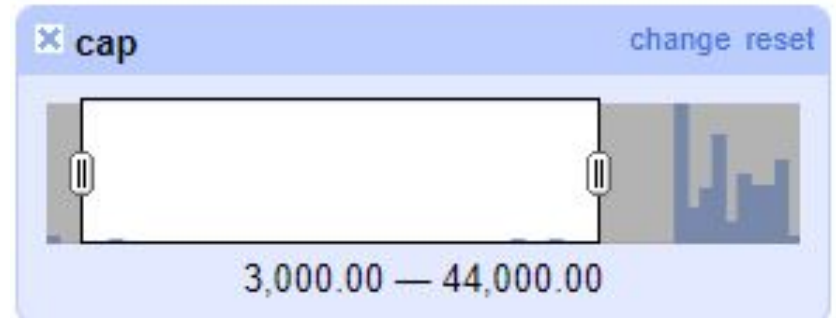
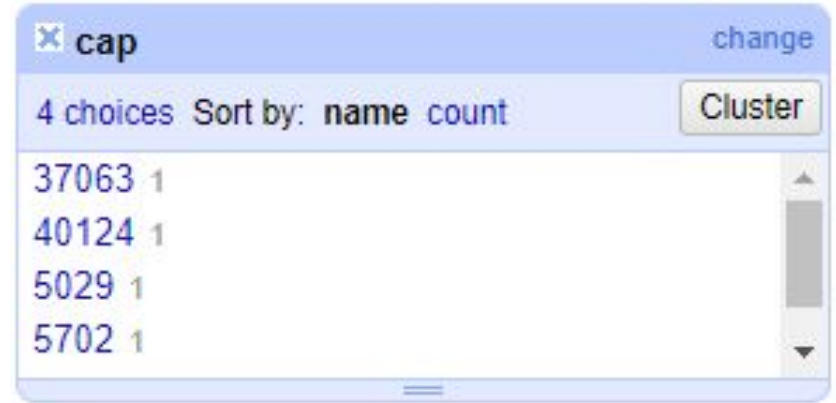
- Is the data column recorded as a string or number?
- How many values are missing?
- How many unique values in a column?



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.





# 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)

Different Formats (10.5432 Vs 10,5432)

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

Lateral spaces (“B&B” vs “ B&B” or “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 Salvosi, 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](https://openrefine.org)

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

OpenRefine 3.8.0, released on April 29, 2024

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

An aerial photograph of the Tuscan landscape. The image shows rolling hills with a mix of green and golden-brown fields. A winding road, lined with tall cypress trees, curves through the valley. In the middle ground, a large, two-story stone house with a tiled roof is visible. The overall scene is peaceful and scenic, typical of the Tuscan countryside.

# 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://www.datiopen.it/en/opendata/Regione\\_Toscana\\_Strutture\\_ricettive](http://www.datiopen.it/en/opendata/Regione_Toscana_Strutture_ricettive)

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|9.9674972000000004|045001AAT0013|Agriturismi|Selva, 0 - Loc. Selva
13843|FIORENTINI GIANLUCA|44.2166706|9.9674972000000004|045001AAT0014|Agriturismi|sanacco, 1
13832|VILLA MIMOSA|44.174129100000002|9.912286299999999|045001AFR0003|Affittacamere|Via MAE
13783|DEMY|44.215124000000003|9.9673911000000004|045001ALB0002|Alberghi - Hotel|Via Salucci,
13717|PASQUINO|44.2166706|9.9674972000000004|045001ALB0003|Alberghi - Hotel|PIAZZA MAZZINI 2
13848|CASA BARANI|44.205532599999998|9.9698068000000006|045001ALL0003|Alloggi Privati|Via SF
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|9.9674972000000004|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|9.9395416999999995|045001CAV0003|Case per Vacanze|Saig
14053|GIUNASCO|44.315319600000002|9.9956531000000002|045002AAT0002|Agriturismi|Giunasco - Lo
```



# Load Accomodation data set

1 Create Project

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

Create Project (circled in red)

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 (circled in red) Nessun file selezionato

Next »

2 Load data

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

# Different way to access data

Get data from

**This Computer**

Web Addresses (URLs)

Clipboard

Database

Google Data

Locate one or more files on your computer to upload:

Scegli file `strutturericettiveXall.csv`

**Next »**

Get data from

This Computer

**Web Addresses (URLs)**

Clipboard

Database

Google Data

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

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

Add Another URL

**Next »**



# Different way to access data

Get data from

This Computer

Web Addresses (URLs)

Clipboard

**Database**

Google Data

New connection

SAVED CONNECTIONS

New Connection Editor

Name: 127.0.0.1

Type: MariaDB

Host: localhost

Port: 3306

User: root

Password: Enter Database Password

Database: Enter Database

Test

Save

Connect

**Set the name****When ready**

Create Project

« Start Over

Configure Parsing Options

Project name 

Tags

**Create Project »**

Open Project

Import Project

Language Settings

	id	nome	lat	lon	codeserc	tipologia	indirizzo	cap	comune	provincia	stelle	er
1.	14017	FORNI ROSAIA	44.2270033	10.0292237999999954	045001AAT0001	Agriturismi	PIAZZA PUCCINI 1 - Loc. Olivola	54011	Aulla	MS	0	lu
2.	13995	POW WOW DI GRULLI ARISTIDE	44.232061100000003	10.0497774999999999	045001AAT0006	Agriturismi	San Domenico la Cavana, 0 - Loc. Bigliolo	54011	Aulla	MS	0	in
3.	13820	VALLE FIORITA	44.225616500000001	10.0181369999999999	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

**UTF-8**

Update Preview

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 0 row(s) of data

initial

 Load at 0 row(s) of data

most

 Use " to enclose cells containing column

character separators

 Parse cell text into

numbers, dates, ...

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

Version 3.3 [58b839b]

Preferences

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

Facet & Filter Results

Number of selected data

Click on this arrow to facet or filter column data

Export the transformation results



OpenRefine Strutture Ricettive Regione Toscana 02-07-2019 [Permalink](#)

Open... **Export** Help

Facet / Filter

Undo / Redo 16 / 16

**16655 rows**

Extensions: Wikidata

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

All	id	nome	lat	lon	Codice esercizio	tipologia		
☆	1.	14017	Facet			Agriturismi	Pi	
☆	2.	13995	Text filter			Agriturismi	Se	
☆	3.	13820	Edit cells			Agriturismi	Vi	
☆	4.	13924	Edit column			Agriturismi	Be	
☆	5.	13843	Transpose			Agriturismi	Se	
☆	6.	13832	Sort...			Agriturismi	Se	
☆	6.	13832	View			Affittacamere	Vi	
☆	7.	13783	Reconcile			Affittacamere	Fe	
☆	7.	13783	Demy	44.21512	9.96739	045001ALB0002	Alberghi - Hotel	Vi
☆	8.	13717	Pasquino	44.21667	9.9675	045001ALB0003	Alberghi - Hotel	Pi

### 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](#)

Undo all your operations

Data Exploration  
Use

## **Facet & Filter**

to select subsets of your data to act on

The image shows a software interface with a table and a context menu. The table has columns for 'tipologia', 'indirizzo', 'cap', 'comune', and 'provincia'. The context menu is open over the 'tipologia' column, showing options like 'Facet', 'Text filter', 'Edit cells', 'Edit column', 'Transpose', 'Sort...', 'View', 'Reconcile', and 'Extract named entities...'. The 'Text facet' option is highlighted with a red box.

	tipologia	indirizzo	cap	comune	provincia
1					
5					
2					
3					
4			54010	Aulla	MS
3			54011	Aulla	MS

**Text Facet on  
“tipologia”  
column**

# Text Facet

In italian: sfaccettature

technically is an histogram





# CAP della Regione Toscana

I Codici di  
Avviamento Postale  
della Regione  
Toscana sono  
compresi tra **50010** e  
**59100**.

## CAP della Toscana per Provincia

Provincia	CAP
Arezzo	52010 - 52100
Firenze	50010 - 50145
Grosseto	58010 - 58100
Livorno	57014 - 57128
Lucca	55010 - 55100
Massa-Carrara	54010 - 54100
Pisa	56010 - 56128
Pistoia	51010 - 51100
Prato	59011 - 59100
Siena	53011 - 53100



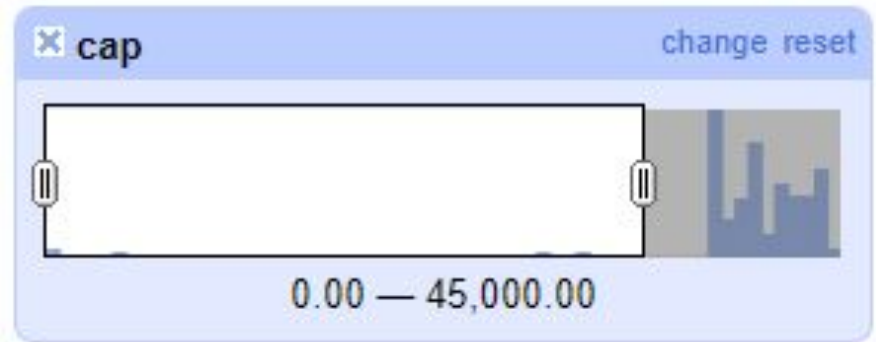
# Combining Facet

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

Click on "0" to see the 163 accomodations without zip code



A text facet for the field 'cap'. The title bar shows 'x cap' and 'change'. Below the title bar, it says '5 choices Sort by: name count' and has a 'Cluster' button. The main area displays a list of choices with their counts:

0	163
37063	1
40124	1
5029	1
5702	1

Facet by choice counts

# 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 PUCCI Olivola
☆	2.	13995	44.2320611	10.0497775	045001AAT0006	Agriturismi	San Domenico 0 - Loc. Bigliolo
☆	3.	13989		029412	045001AAT0011	Agriturismi	Via COLLINA 7
☆	4.	13820					A DI BEL a
☆	5.	13924					0 - Loc.
☆	6.	13843					o, 1 - Lo
☆	7.	13832					ESTRO Albiano M
☆	8.	13783					ucci, 9
☆	9.	13717	44.2166706	9.9			A MAZZI
☆	10.	13848	44.2055326	9.9			RINI 7/A
☆	11.	14000	44.22548	10.			LLINA 8
☆	12.	13803	44.2320611	10.0			ONTE BA
☆	13.	17369	44.2166706	9.9674972	045001ALL0010	Alloggi Privati	Loc. Bigliolo Libertà, 14/F - I - Fraz. Albiano
☆	14.	21952	44.1741291	9.9122863	045001ALL0012	Alloggi Privati	Via Amola, 18 -

**To title  
case**

# Edit cells ▶ Transforms

	lon	codeserc	tipologia	indirizzo
33	Facet		Agriturismi	PIAZZA PUCCINI 1 - Loc. C
11	Text filter		Agriturismi	San Domenico la Cavana, Bigliolo
48	<b>Edit cells</b>		<b>Transform...</b>	loc. Olivc
65	Edit column		Common transforms	NE - Loc
06	Transpose		Fill down	va
06	Sort...		Blank down	Quercia
91	View		Split multi-valued cells...	RRARI 7
24	Reconcile		Join multi-valued cells...	
	Extract named entities...		Cluster and edit...	
06	9.9073	045001ALB0003	Hotel	FRATELLI MARCONI 22 - Fraz
26	9.96981	045001ALL0003	Alloggi Privati	Via SPRINI 7/A - Fraz. Spri

# Edit cells ▶ Transforms

Custom text transform on column nome

Expression Language **General Refine Expression Language (GREL)** ▼

value

new History Starred Help

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

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

OK Cancel

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.name.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`**

Preview

History

Starred

Help

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  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	▼ pro
Facet ▶	54011	Aulla	MS
Text filter ▶	54011	Aulla	MS
Edit cells ▶	54011	Aulla	MS
Edit column ▶			
Transpose ▶			
Sort...			
View ▶			
Reconcile ▶			
Extract named entities...			
PIAZZA MAZZINI 22 - Fraz. pip			
Via SPRINI 7/A - Fraz. Sprini			
Via COLLINA 8 - Loc. Olivola			
Via MONTE BARDELLI - Loc. E			
Libertà 14/F - Loc. ALLIA - Fraz. Albiano Magra	54010	Aulla	MS

- Facet ▶
- Text filter ▶
- Edit cells ▶
- Edit column ▶
- Transpose ▶
- Sort...
- View ▶
- Reconcile ▶
- Extract named entities...
- PIAZZA MAZZINI 22 - Fraz. pip
- Via SPRINI 7/A - Fraz. Sprini
- Via COLLINA 8 - Loc. Olivola
- Via MONTE BARDELLI - Loc. E
- Libertà 14/F - Loc. ALLIA - Fraz. Albiano Magra

- 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

## 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()`

Preview

History

Starred

Help

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

OK

Cancel

# 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

Keying Function

128 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
6	7929	<ul style="list-style-type: none"><li>Via (5736 rows)</li><li>VIA (2154 rows)</li><li>via (32 rows)</li><li>Via (5 rows)</li><li>Via (1 rows)</li><li>vIA (1 rows)</li></ul>	X	Via
5	82	<ul style="list-style-type: none"><li>Località (72 rows)</li><li>LOCALITA (3 rows)</li><li>LOCALITA' (3 rows)</li><li>Localita' (3 rows)</li><li>località (1 rows)</li></ul>	X	Località
	115	<ul style="list-style-type: none"><li>Lungomare (65 rows)</li><li>LUNGOMARE (23 rows)</li><li>Lungomare, (22 rows)</li><li>lungomare (4 rows)</li><li>LUNGOMARE, (1 rows)</li></ul>	X	Lungomare
5	79	<ul style="list-style-type: none"><li>C.S. (66 rows)</li><li>c.s. (7 rows)</li><li>C.S., (3 rows)</li></ul>	X	C.S.

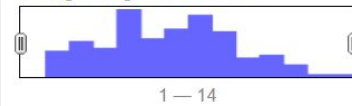
### # Choices in Cluster



### # Rows in Cluster



### Average Length of Choices



### Length Variance of Choices



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

Select All Unselect All

Export Clusters

Merge Selected & Re-Cluster

Merge Selected & Close

Close

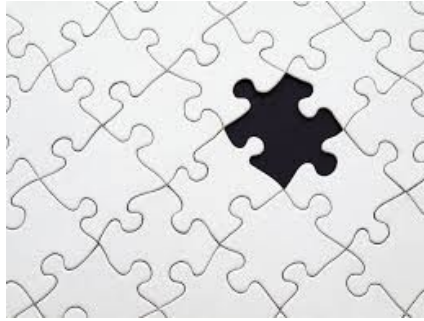


Data enrichment/augmentation



# Why

## Fill missing data



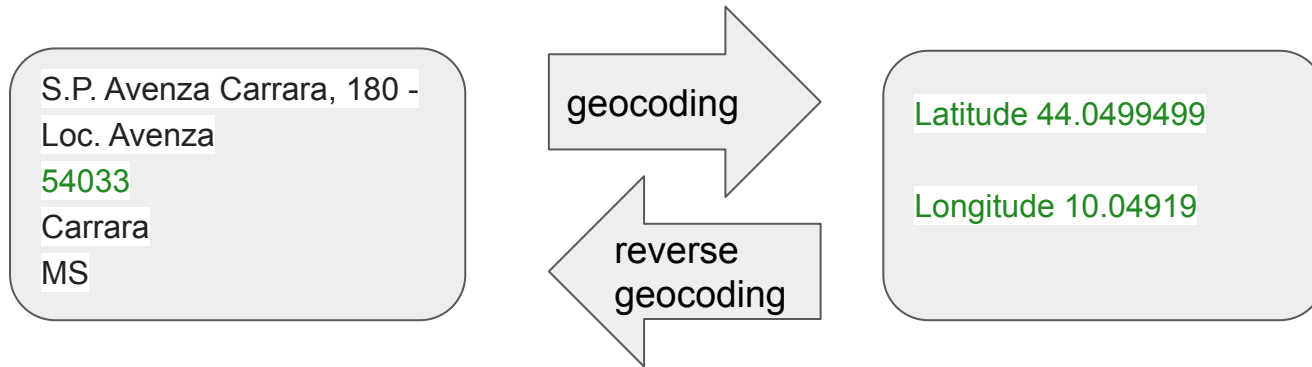
## Validate data

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



# Openstreetmap Json Result

<https://nominatim.openstreetmap.org/search?q=Via Moruzzi 1, Pisa&format=json>

```
[
  {
    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

3 rows

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

All	IndirizzoStadio	Serie	Campionato	Giro
1.	gio a	b	2017	d
2.	E, Italy	b	2017	d
3.				

Context menu for the 'IndirizzoStadio' column:

- Facet
- Text filter
- Edit cells
- Edit column**
  - Split into several columns...
  - Add column based on this column...
  - Add column by fetching URLs...**
  - Add columns from Freebase ...
- Transpose
- Sort...
- View
- Reconcile
- Extract named entities...
- Remove this column
- Move column to beginning
- Move column to end
- Move column left
- Move column right

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.eyJ1IjoieYXF1YWJsdWUiLCJhIjoieY2tZXFnbmR0MnJjODJ2bnc2Znp0bGc3MCI9.YmF\\_-yyqzeCeoPdOSob7g](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.eyJ1IjoieYXF1YWJsdWUiLCJhIjoieY2tZXFnbmR0MnJjODJ2bnc2Znp0bGc3MCI9.YmF_-yyqzeCeoPdOSob7g)

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:  Throttle delay:  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:  No syntax error.

**Preview** History Starred Help

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

OK Cancel

Attesa risposta

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

Escape: [GREL string function](#)

# Extract Latitude from Json

The screenshot shows a data table with three rows of addresses. A context menu is open over the 'Json' column, and the option 'Add column based on this column...' is highlighted with a red box. The table data is as follows:

	All	IndirizzoStadio	Json	Serie
1.	☆	Via Sandriana, 80046 San Giorgio a Cremano NA, Italy	Facet Text filter Edit cells Edit column Transpose Sort... View Reconcile Extract named entities...	b
2.	☆	Via Melissano, 73055 Racale LE, Italy	Split into several columns... Add column based on this column... Add column by fetching URLs... Add columns from Freebase ...	
3.	☆	Via di Settebagni, 340, 00139 Roma RM, Italy	Rename this column Remove this column Move column to beginning Move column to end Move column left Move column right	

The 'Json' column contains the following JSON data for the third row:

```
metropolitan area of Roma", "short_name": "Lazio", "administrative_area_level_2": "Lazio", "administrative_area_level_1": "Lazio", "short_name": "IT", "types": [ "locality", "postal_code" ] }, "formatted_address": "Via di Settebagni, 340, 00139 Roma RM, Italy", "geometry": { "location": { "lat": 41.9662928029149, "lng": 12.5471990802915 }, "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

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

Expression  Language General Refine Expression Language (GREL) ▾

No syntax error.

**Preview** History Starred Help

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,"lat":43.817535,"lon":10.7272541,"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

OK Cancel

Grel Function



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



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



Reconciling

# Reconciling

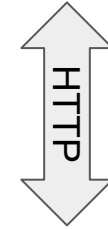
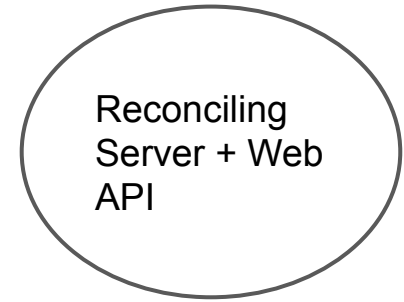
Reconciliation is the process of matching your dataset with that of an **external authoritative source**

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



Web Browser



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

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

RDF ▾

> last

nt@gr

gopiet

ends.it

erossit

uelune

@gma

ino.it

lia.lu.it

sonbe

ora@g

miamibeb.com

www.miamibeb.com



Create a Report

Facet / Filter

Undo / Redo 122 / 122

Extract...

Apply...

Filter:

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 expression value.trim()
- Text transform on cells in column indirizzo 3 using expression 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

Select All

Unselect All

```
[
  {
    "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 column nome"
  },
  {
    "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

[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](#)

