

# **Laboratorio Progettazione Web**

## **Open Refine**

Angelica Lo Duca

# Cos'è

- Open Refine è un tool per la pulizia, la trasformazione e l'estensione dei dati.
- E' disponibile a questo indirizzo
  - <http://openrefine.org>
  - Scaricare la versione Google Refine (stable)
- Per la manipolazione dei dati si usa il linguaggio GREL
  - <https://github.com/OpenRefine/OpenRefine/wiki/General->

# Per iniziare

- Scaricare/Installare Open Refine
- Lanciare il programma
- Dal browser (Chrome, Safari ecc) andare su
  - <http://127.0.0.1:3333>

# Per iniziare

- Scaricare dal didawiki il file Libri.csv
- Caricarlo in Open Refine
  - Selezionare Create Project
  - Selezionare il file Libri.csv
  - Premere Next
  - In basso selezionare
  - In alto selezionare
    - create project

Character encoding

Columns are separated by

commas (CSV)

tabs (TSV)

custom \t \_\_\_\_\_

Escape special characters with \

6 rows

Extensions: Freebase ▾

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

« first < previous **1 - 6** next »

<input type="checkbox"/> All	<input type="checkbox"/> Libro	<input type="checkbox"/> Autore/i	<input type="checkbox"/> Lingua originale	<input type="checkbox"/> Prima pubblicaz	<input type="checkbox"/> Vendite (appross	
<input type="checkbox"/>	<input type="checkbox"/>	1. A Tale of Two Cities (Racconto di due città)	Charles Dickens	inglese	1859	200 milioni[1]
<input type="checkbox"/>	<input type="checkbox"/>	2. And Then There Were None (Dieci piccoli indiani)	Agatha Christie	inglese	1939	100 milioni[4]
<input type="checkbox"/>	<input type="checkbox"/>	3. Le Petit Prince (Il piccolo principe)	Antoine de Saint-Exupéry	francese	1943	140 milioni[3]
<input type="checkbox"/>	<input type="checkbox"/>	4. The Hobbit (Lo Hobbit)	J. R. R. Tolkien	inglese	1937	100 milioni[8]
<input type="checkbox"/>	<input type="checkbox"/>	5. The Lord of the Rings (Il Signore degli Anelli)	J. R. R. Tolkien	inglese	1954, 1955	150 milioni[2]
<input type="checkbox"/>	<input type="checkbox"/>	6. 紅樓夢 (Il sogno della camera rossa)	Cao Xueqin	cinese	1754[5], 1791[6]	100 milioni[7]

# Manipolazione Stringhe

- Nel campo autore separare il nome e il cognome e creare una colonna per il nome e una per il cognome
  - Cliccare sulla freccia corrispondente al titolo della colonna Autore

The screenshot shows a data table with the following columns: Autore/i, Lingua originale, Prima pubblicaz, and Vendite. The 'Autore/i' column header has a dropdown menu open, showing options like Facet, Text filter, Edit cells, Edit column, Transpose, Sort..., View, and Reconcile. The 'Edit column' option is selected, and a sub-menu is visible with options like 'Split into several columns...', 'Add column based on this column...', 'Add column by fetching URLs...', 'Add columns from Freebase ...', 'Rename this column', 'Remove this column', 'Move column to beginning', 'Move column to end', 'Move column left', and 'Move column right'.

	Autore/i	Lingua originale	Prima pubblicaz	Vendite
		inglese	1859	200 milioni[1
iani)		inglese	1939	100 milioni[4
	ry	francese	1943	140 milioni[3
		inolese	1937	100 milioni[8
				50 milioni[2
a rossa)				00 milioni[7

# Manipolazione Stringhe

- Uso `substring(miastringa, posizione_iniziale, posizione_finale)` per estrarre una sottostringa contenente il nome e il cognome
  - Esempio Charles Dickens
    - il nome si trova a partire dalla posizione 0 fino allo spazio escluso
    - il cognome si trova dalla posizione `spazio+1` fino alla fine
    - per identificare la posizione del carattere spazio uso la funzione `indexOf(miastringa, carattere_da_cercare)`

## Add column based on column Autore/i

New column name

On error

set to blank  store error  copy value from original column

Expression

Language

Google Refine Expression Language (GREL) 

```
substring(value,0, indexOf(value, " "))
```

No syntax error.

**Preview**

History

Starred

Help

row	value	substring(value,0, indexOf(value, " "))
1.	Charles Dickens	Charles
2.	Agatha Christie	Agatha
3.	Antoine de Saint-Exupéry	Antoine
4.	J. R. R. Tolkien	J.
5.	J. R. R. Tolkien	J.
6.	Cao Xueqin	Cao

OK

Cancel

## Add column based on column Autore/i

New column name

On error

- set to blank  store error  copy value from original column

Expression

Language

Google Refine Expression Language (GREL) 

```
substring(value,indexOf(value, " ")+1)
```

No syntax error.

**Preview**

History

Starred

Help

row	value	substring(value,indexOf(value, " ")+1)
1.	Charles Dickens	Dickens
2.	Agatha Christie	Christie
3.	Antoine de Saint-Exupéry	de Saint-Exupéry
4.	J. R. R. Tolkien	R. R. Tolkien
5.	J. R. R. Tolkien	R. R. Tolkien
6.	Cao Xueqin	Xueqin

OK

Cancel

# Esercizio

- Separare il titolo in inglese dal titolo in italiano
  - Come soluzione si può usare anche l'opzione split in several columns.
    - La differenza con l'approccio che abbiamo usato consiste nel fatto che attraverso l'uso della split in several columns la colonna originale è cancellata

# Soluzione

- Esempio stringa:
  - A tales of two cities (Racconto di due città)
- Titolo in Inglese
  - `substring(value,0, indexOf(value, "("))`
- Titolo in Italiano
  - `substring(value,indexOf(value, "(")+1, indexOf(value, ")"))`

# Arricchimento Dati

- Aggiungere data di nascita e morte degli autori
  - usare add Column by fetching url
    - usare DBpedia per recuperare i dati
      - <http://dbpedia.org> è la versione Linked Data di Wikipedia

Facet	▶	"link": [], "vars": ["subject", "birth", "death"], "results
Text filter		ered": true, "bindings": [ { "subject": { "type": "uri", "va
Edit cells	▶	edia.org/resource/Charles_Dickens" }, "birth": { "typ
Edit column	▶	datatype": "http://www.w3.org/2001/XMLSchema#date
Transpose	▶	07" }, "death": { "type": "typed-literal", "datatype":
Sort...		http://www.w3.org/2001/XMLSchema#date" "value": "1870-06
View	▶	Split into several columns...
Reconcile	▶	Add column based on this column...
		Add column by fetching URLs...
		Add columns from Freebase ...
		Rename this column
		Remove this column
		Move column to beginning
		Move column to end
		Move column left
		Move column right
J. R. R. Tolkien		"http://ww
		"death": {
		"http://ww
		{ "head": {
		false, "orc
		"http://dbp
		literal", "d
		"1892-01-
		"http://ww

# Query a DBpedia

- "http://dbpedia.org/sparql?query=" + escape("select distinct ?subject, ?birth, ?death where { ?subject foaf: name \" + value + "\"@en; dbp:birthDate ?birth; dbp: deathDate ?death} LIMIT 100", "url") + "&format=JSON"

```
{ "head": { "link": [], "vars": ["subject", "birth", "death"] }, "results": { "distinct": false, "ordered": true, "bindings": [ { "subject": { "type": "uri", "value": "http://dbpedia.org/resource/Charles_Dickens" }, "birth": { "type": "typed-literal", "datatype": "http://www.w3.org/2001/XMLSchema#date", "value": "1812-02-07" }, "death": { "type": "typed-literal", "datatype": "http://www.w3.org/2001/XMLSchema#date", "value": "1870-06-09" } ] ] }
```

- `["results"]["bindings"][0]["birth"]["value"]`

## Add column based on column Dbpedia

New column name

On error

set to blank  store error  copy value from original column

Expression

Language

Google Refine Expression Language (GREL) 

```
value.parseJson()["results"]["bindings"][0]["birth"]["value"]
```

No syntax error.

**Preview**

History

Starred

Help

row	value	value.parseJson()["results"]["bindings"][0]["birth"]["value"]
1.	<pre>{ "head": { "link": [], "vars": ["subject", "birth", "death"] }, "results": { "distinct": false, "ordered": true, "bindings": [ { "subject": { "type": "uri", "value": "http://dbpedia.org/resource/Charles_Dicken" }, "birth": { "type": "typed-literal", "datatype": "http://www.w3.org/2001/XMLSchema#date" "value": "1812-02-07" }, "death": { "type": "typed-literal", "datatype":</pre>	1812-02-07

OK

Cancel