Open Refine

Open Refine è un'applicazione Web per la pulizia, la trasformazione e l'arricchimento di dataset. Può essere scaricato al seguente indirizzo: http://openrefine.org. Nella sezione download, si può scaricare l'ultima versione stabile (OpenRefine 2.7-rc2 Release Candidate 2), disponibile per Windows, Linux e Mac OS. Dopo aver scaricato ed installato il programma, lo si può eseguire. Una volta eseguito, l'applicazione può essere utilizzata dal browser al seguente indirizzo: http://127.0.0.1:3333.
Open Refine permette di eseguire diverse operazioni. In questa guida vedremo solo le seguenti:

1. upload di un dataset in formato CSV (sono supportati anche altri formati)
2. manipolazione delle colonne
    a. aggiunta di una nuova colonna sulla base di quelle esistenti
    b. estrazione di elementi da una colonna
3. arricchimento del dataset con dati provenienti dall'esterno

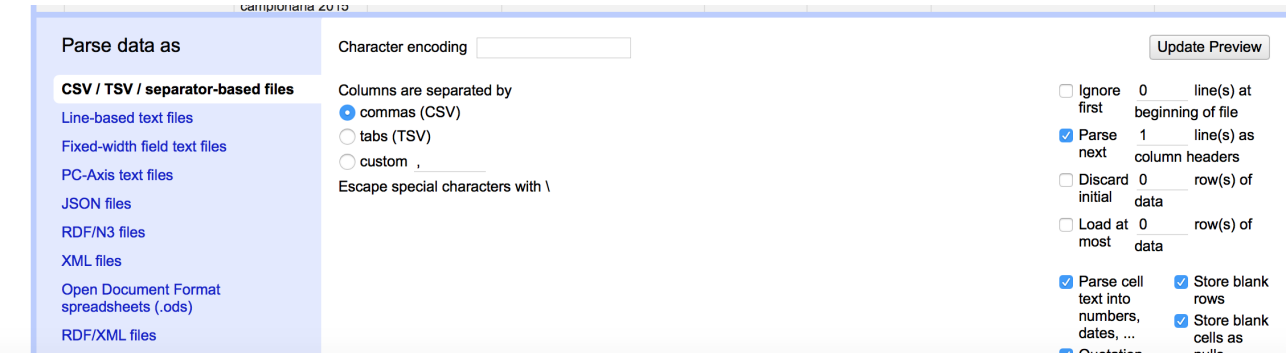
Per la manipolazione dei dati, Open Refine usa il linguaggio GREL (General Refine Expression Language)[1].

Upload di un dataset

A titolo di esempio prendiamo il dataset contenente la produzione editoriale della Regione Toscana nel 2015 e nel 2014, disponibili rispettivamente ai seguenti indirizzi:

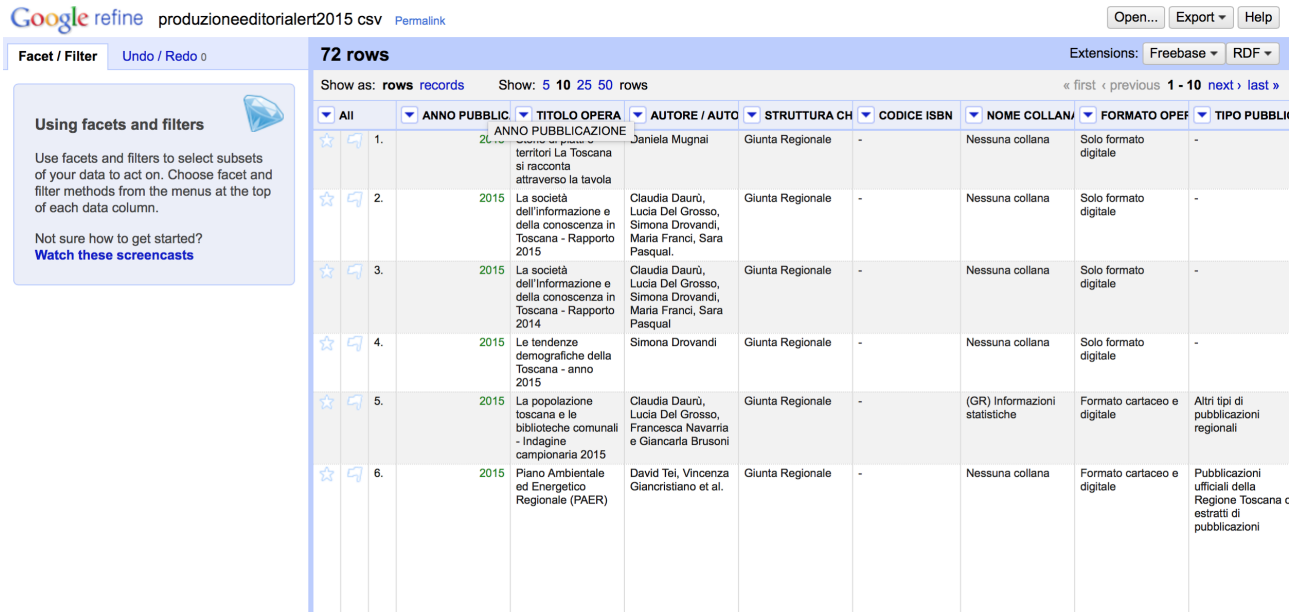
- http://dati.toscana.it/dataset/rt-proded-2015
- http://dati.toscana.it/dataset/rt-proded2014

Scarichiamo entrambi i dataset.Concentriamoci ora solo sul dataset del 2015. Accediamo ad Open Refine e selezioniamo la voce Create Project dal menu sulla sinistra. Poi selezioniamo dal computer il file che vogliamo importare e premiamo il pulsante next. A questo punto compare un'anteprima del progetto. Nella parte bassa, appare un altro menu (vedi figura sottostante), da cui è possibile selezionare alcune opzioni, come ad esempio il carattere di separazione dei campi del CSV (indicato dall'intestazione *Columns are separated by*) ed altre opzioni sulla destra (ad esempio *Parse next 1 line(s) as next column headers* per settare i nomi delle colonne). Un'altra opzione, situata in alto a sinistra, da la possibilità di settare la codifica. Nel nostro caso, il dataset nell'anteprima presenta alcuni problemi di encoding, per cui possiamo selezionare la codifica utf8 dall'opzione *Character encoding*. A questo punto i problemi di encoding del nostro dataset dovrebbero essere risolti.



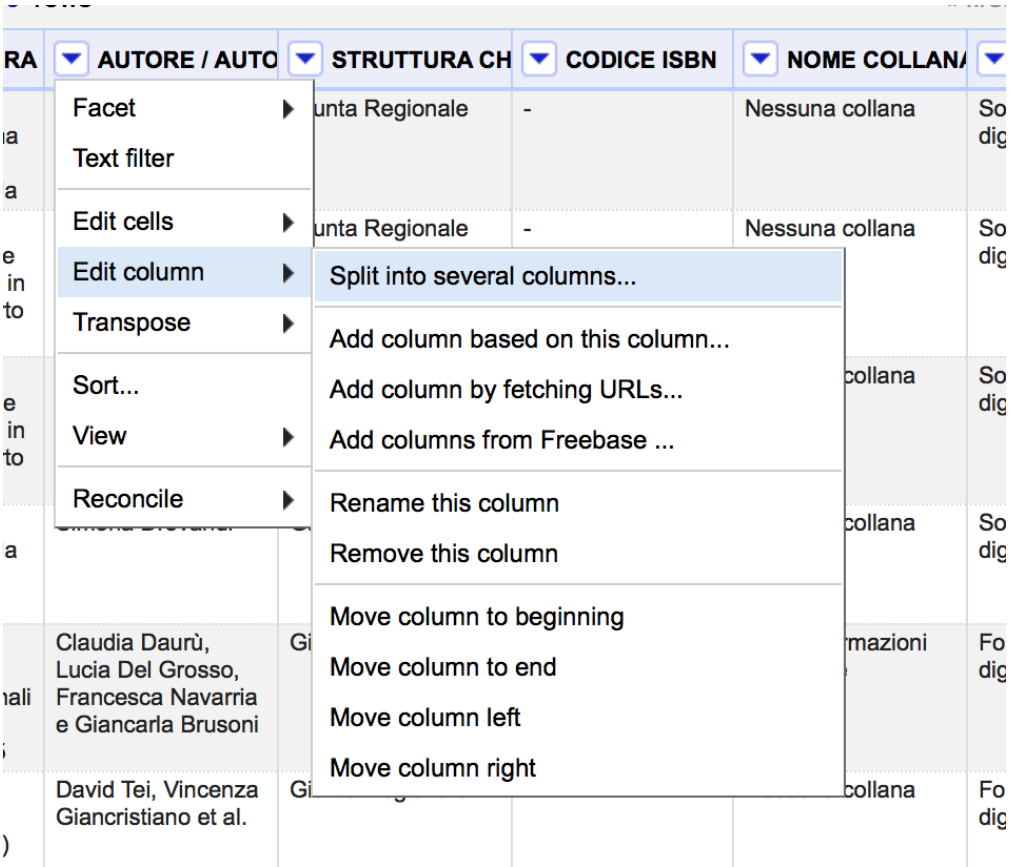
A questo punto, nella parte in alto a destra della pagina possiamo cambiare il nome del progetto (alla voce *Project Name*) e poi premere il tasto *Create Project*. La schermata dell'applicazione dovrebbe apparire come nella seguente figura:

[1] https://github.com/OpenRefine/OpenRefine/wiki/General-Refine-Expression-Language

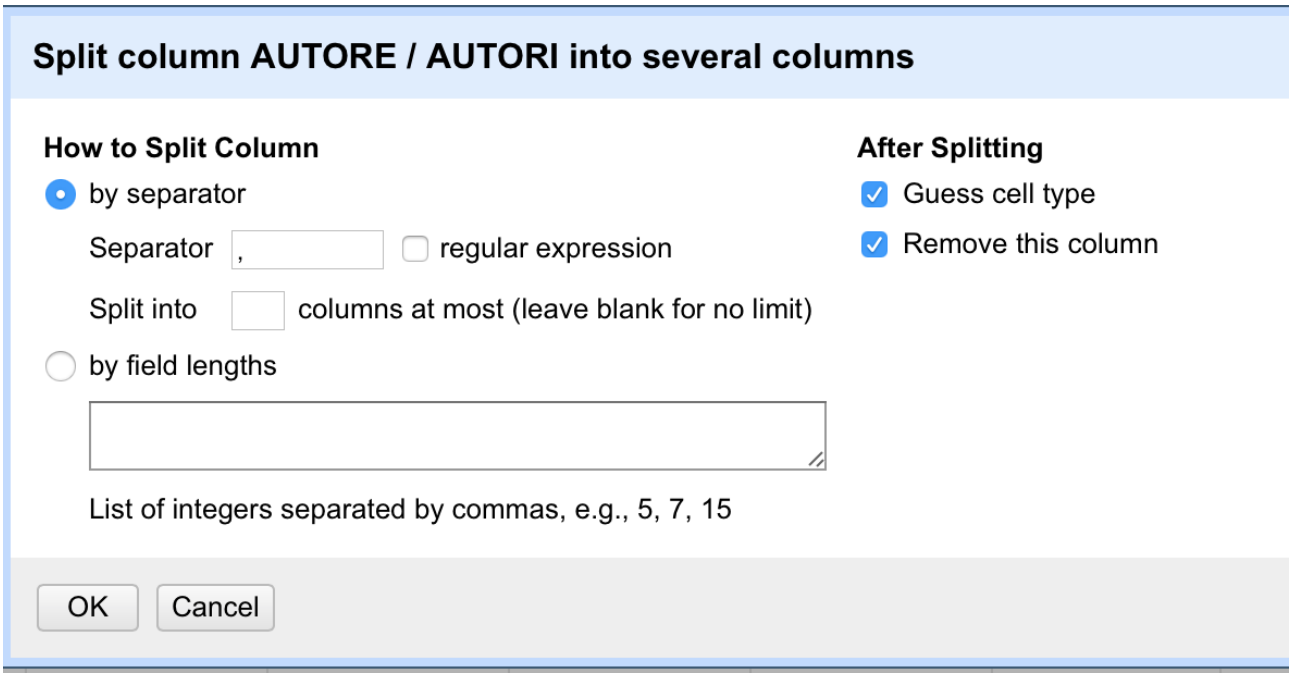


Manipolazione colonne

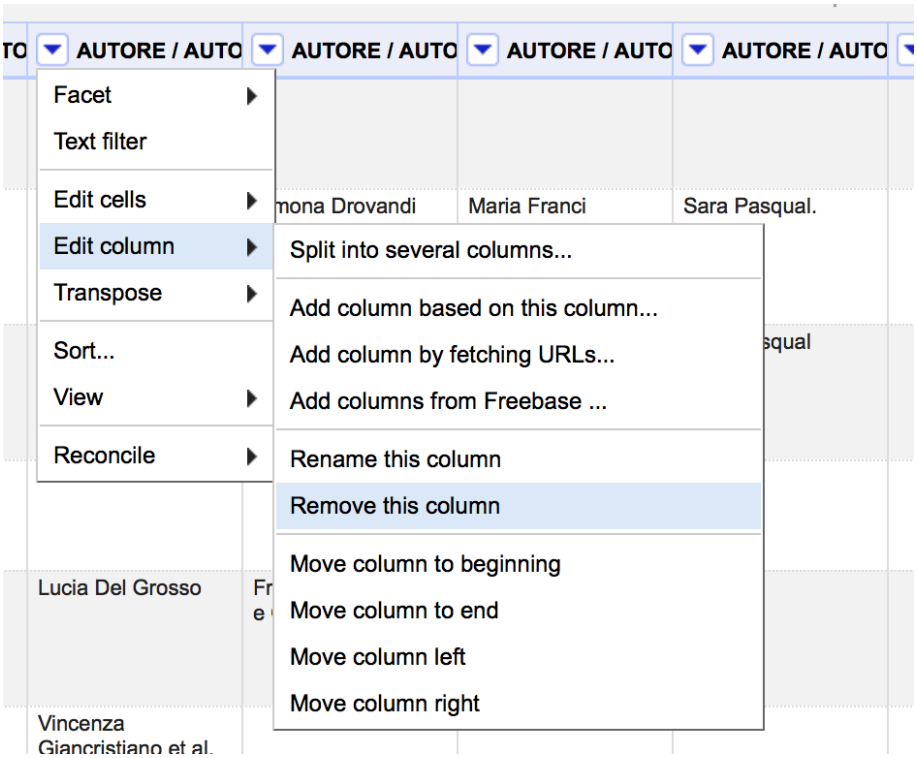
Come prima cosa, vogliamo separare gli autori nella colonna Autore e creare una colonna per ogni autore. Per fare questo, in corrispondenza del titolo della colonna autore, cliccare sulla freccia e dal menù a tendina che si apre (vedi figura sottostante), selezionare edit column e poi split into several columns.



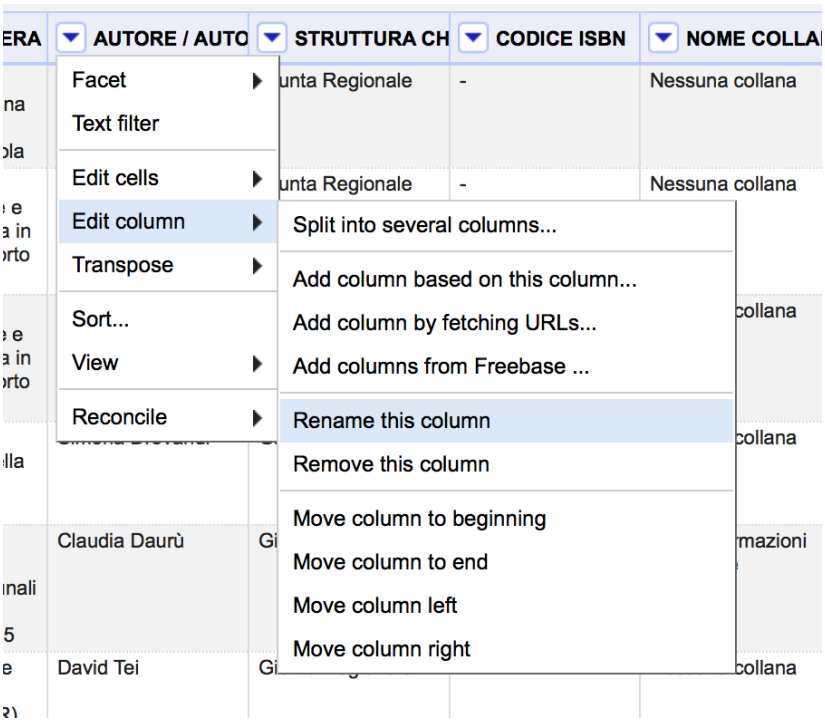
Nella finestra che si apre (vedi figura sotto) è possibile selezionare varie opzioni, tra cui la modalità con cui effettuare la divisione della colonna. Nel nostro caso, va bene l'opzione di default, cioè l'uso della virgola come separatore dei campi. Premiamo il tasto ok e vediamo che la nostra tabella è cambiata. La colonna originale è stata rimossa e sono state create tante colonne, ognuna avente come nome il nome vecchio della colonna seguito da un numero progressivo.



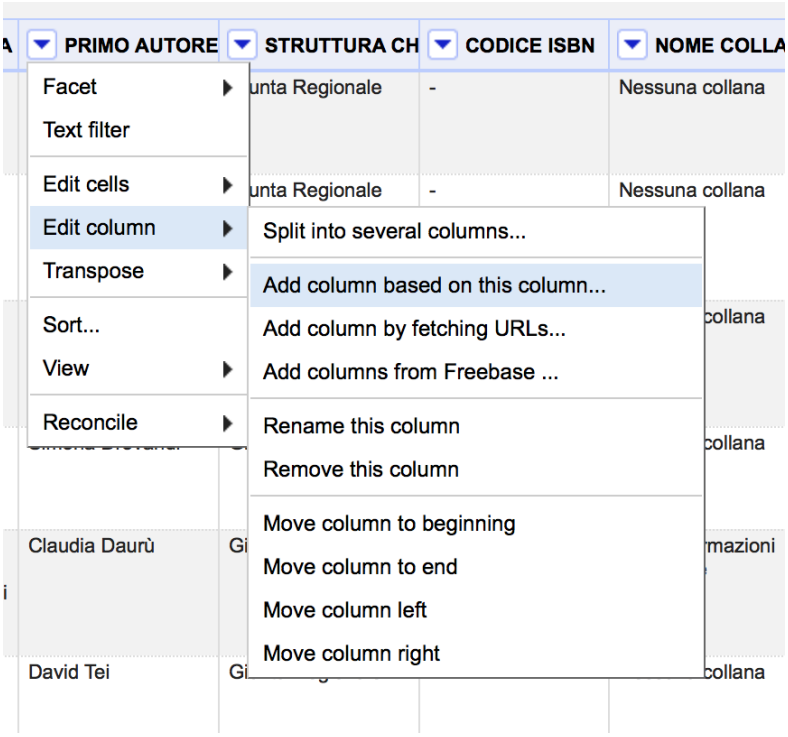
Ora rimuoviamo tutte le colonne aggiunte, tranne la prima. In questo modo avremo solo il primo autore di ogni opera editoriale. Per fare questo, per ogni colonna da rimuovere, selezioniamo, sempre dalla freccia posta vicino al nome della colonna, l'opzione edit column e poi remove this column (vedi figura sotto).



A questo punto abbiamo solo una colonna con il nome del primo autore. Rinominiamo la colonna in "Primo Autore". Per fare questo dalla freccia posta vicino al nome della colonna selezioniamo edit column e poi Rename this column:



Ora, supponiamo di voler creare due nuove colonne, contenenti una il nome e l'altra il cognome del primo autore. Per fare questo dalla freccia selezioniamo Edit Column e poi Add column based on this column:



Nella finestra che si apre occorre indicare il nome della colonna che si vuole creare e poi l'espressione in linguaggio GREL per manipolare la colonna. Apriamo una parentesi sul linguaggio GREL. In GREL è possibile utilizzare variabili, controlli e funzioni. Le variabili rappresentano le righe della tabella. Esistono diversi tipi di variabili:

- value - rappresenta il valore della cella corrispondente alla colonna selezionata

- cell - rappresenta cella corrispondente alla colonna selezionata. Su questa variabile si possono fare due operazioni:
    - cell.value, che corrisponde a value
    - cell.recon che permette di ottenere i risultati di un processo di riconciliazione con dati esterni
- cells - rappresenta l'intera riga. Per accedere al valore di una colonna occorre usare la seguente sintassi: cells["nome colonna"].value.

Per quanto riguarda i controlli, ne esistono di diverso tipo, tra cui l'if e il foreach. Per maggiori dettagli si rimanda alla documentazione[2]. Infine, vi sono le funzioni, che possono essere di diverso tipo. Tra le più importanti vi sono quelle di manipolazione delle stringhe. Per maggiori dettagli si rimanda alla documentazione[3].

Ritorniamo al nostro dataset. Abbiamo detto che vogliamo creare due nuove colonne, una contenente il nome e l'altro il cognome dell'autore. Per fare questo, potremmo usare la split column che abbiamo utilizzato in precedenza, ma perderemmo la colonna originale. Per mantenere la colonna di partenza, possiamo usare la add column. Nel campo riservato all'espressione GREL, possiamo creare la prima colonna contenente il nome usando la funzione substring, che estrae una sottostringa dalla stringa passata. Applichiamo al valore della cella corrente (rappresentata dalla variabile value) la funzione substring, che riceve gli estremi della sottostringa (indice di partenza, indice di arrivo). L'indice di partenza è 0 mentre l'indice di arrivo è dato dalla posizione del carattere spazio, che può essere identificata attraverso la funzione indexOf, applicata sempre alla variabile value:

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

In questo modo otteniamo il nome dell'autore. Per ottenere il cognome, invece occorre creare una nuova colonna, quindi selezionare nuovamente add column based on this column e poi manipolare la stringa tramite GREL nel seguente modo: occorre estrarre dalla variabile value una sottostringa che parte dalla posizione del carattere spazio + 1:

```
value.substring(value.indexOf(" ")
```

A questo punto il risultato dovrebbe apparire come nella figura seguente:

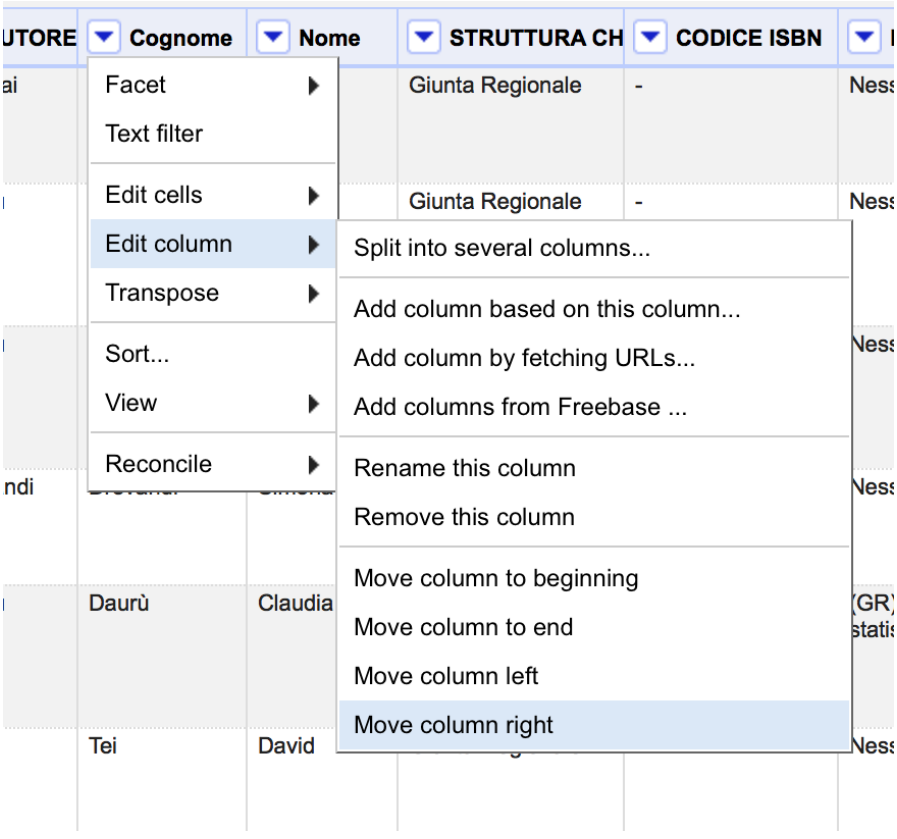
[2] https://github.com/OpenRefine/OpenRefine/wiki/GREL-Controls
[3] https://github.com/OpenRefine/OpenRefine/wiki/GREL-String-Functions

Show as: **rows** records Show: 5 **10** 25 50 rows « first ‹ previous **1 - 10**

	ANNO PUBBLIC	TITOLO OPERA	PRIMO AUTORE	Cognome	Nome	STRUTTURA CH	CODICE ISBN	NOME COLLAN	
1.	edit 2015	Storie di piatti e territori La Toscana si racconta attraverso la tavola	Daniela Mugnai	Mugnai	Daniela	Giunta Regionale	-	Nessuna collana	Solo digit
2.	2015	La società dell'informazione e della conoscenza in Toscana - Rapporto 2015	Claudia Daurù	Daurù	Claudia	Giunta Regionale	-	Nessuna collana	Solo digit
3.	2015	La società dell'Informazione e della conoscenza in Toscana - Rapporto 2014	Claudia Daurù	Daurù	Claudia	Giunta Regionale	-	Nessuna collana	Solo digit
4.	2015	Le tendenze demografiche della Toscana - anno 2015	Simona Drovandi	Drovandi	Simona	Giunta Regionale	-	Nessuna collana	Solo digit
5.	2015	La popolazione toscana e le biblioteche comunali - Indagine campionaria 2015	Claudia Daurù	Daurù	Claudia	Giunta Regionale	-	(GR) Informazioni statistiche	Forn digit
6.	2015	Piano Ambientale ed Energetico Regionale (PAER)	David Tei	Tei	David	Giunta Regionale	-	Nessuna collana	Forn digit

Edit this cell

Per spostare una colonna verso destra o verso sinistra si può selezionare sempre dalla freccia il menù edit column e poi move column left o right (vedi figura sotto).

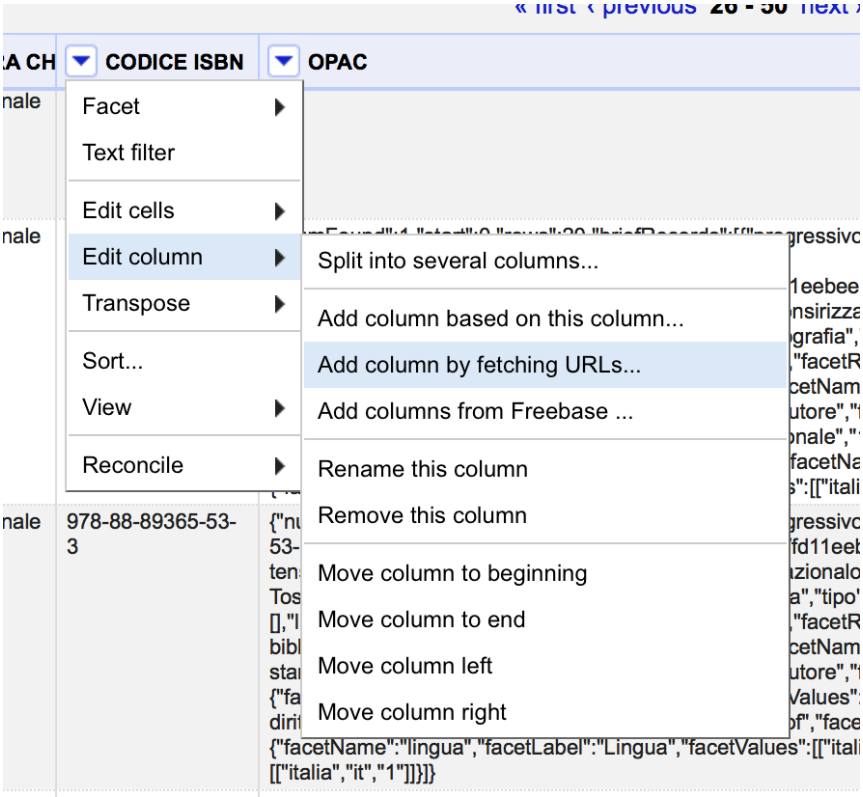


Arricchimento Dataset

Supponiamo ora di voler aggiungere, per i libri provvisti di codice ISBN, anche il codice identificativo. Per fare questo possiamo interrogare il sito opac.sbn.it, che mette a disposizione una API che riceve in ingresso il codice ISBN e restituisce una serie di informazioni. L'APi può essere invocata nel seguente modo:

http://opac.sbn.it/opacmobilegw/search.json?isbn=numeroisbn

Il risultato è un json che contiene anche il codice identificativo del libro. Nel nostro caso, selezioniamo la colonna corrispondente all'ISBN, poi edit column e poi add column by fetching URLs:



Nella casella corrispondente al linguaggio GREL dobbiamo fare una distinzione tra le righe che contengono l'ISBN e quelle che non lo contengono. Possiamo identificare le righe che non contengono l'ISBN come quelle che hanno lunghezza 1. Per cui possiamo usare un controllo di tipo IF per verificare se la riga contiene o meno il codice ISBN. Il controllo if funziona come segue:

if(condizione, espressione vera, espressione falsa)

Se si verifica la condizione, è eseguita l'espressione vera, altrimenti quella falsa. Nel nostro caso:

if(value.length() == 1, null, *carica_codice_identificativo*)

Per caricare il codice identificativo, occorre effettuare la chiamata all'API. Per fare questo, basta inserire l'URL dell'API tra virgolette e aggiungere eventuali variabili tramite l'operatore +. Nel nostro caso occorre specificare l'ISBN, che di volta in volta sarà uguale al valore della riga corrente (value):

"http://opac.sbn.it/opacmobilegw/search.json?isbn=" + value

Quindi il codice completo da inserire nella casella GREL è questo:

```
if(value.length() == 1, null,
"http://opac.sbn.it/opacmobilegw/search.json?isbn=" + value)
```

Diamo un nome alla nuova colonna (ad esempio JSON) e premiamo ok. Aspettiamo il risultato. Quando il sistema ha completato la procedura, compare una nuova colonna chiamata JSON, che contiene un JSON con tutte le informazioni estratte dall'API. Ora per estrarre il campo codice identificativo, dobbiamo fare un parsing del JSON. Selezioniamo la colonna JSON e dal menu edit column selezioniamo add column based on this column. Inseriamo il nome della nuova colonna (Codice Identificativo) e nello spazio relativo al linguaggio GREL inseriamo il codice per manipolare il JSON. Come prima cosa, dobbiamo fare un controllo: se il campo è null, non dobbiamo fare niente, altrimenti dobbiamo fare il parsing del JSON. Usiamo la funzione isNull che verifica se una variabile è null o meno:

if(isNull(value), null, *parsing_del_json*)

Occupiamoci ora del parsing del JSON. Applichiamo alla variabile value la funzione parseJson() che trasforma il JSON in un array, per cui poi possiamo accedere direttamente ai campi dell'array. Il codice identificativo si trova nella prima posizione dell'array briefRecords. Per cui possiamo accedervi nel seguente modo:

value.parseJson()["briefRecords"][0]["codiceIdentificativo"]

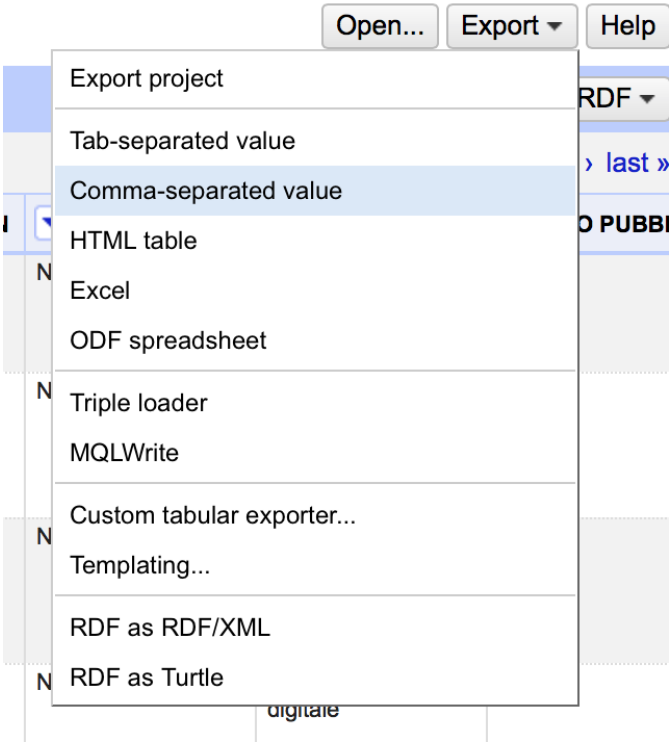
Il codice completo da inserire nello spazio riservato al linguaggio GREL è il seguente:

```
if(isNull(value),                                                    null,
value.parseJson()["briefRecords"][0]["codiceIdentificativo"])
```

Ora premiamo il tasto ok e vedremo comparire una nuova colonna con il codice identificativo in corrispondenza delle righe con codice ISBN.

Gestione di più dataset (Open Refine + PhpMyadmin)

Supponiamo di voler integrare i due dataset (2014 e 2015) in modo da ottenerne uno solo. La prima cosa da fare è controllare che i due dataset abbiano la stessa struttura. Per fare questo possiamo aprire i due file con un programma che legge i CSV e controllare a mano i nomi dei campi. La discriminante tra i due dataset è solo nell'anno. Se i due dataset non dovessero avere una colonna anno, si potrebbe aggiungere usando Open Refine (come descritto in precedenza), aggiungendo una colonna con un valore fisso corrispondente all'anno. Nel nostro caso, il campo anno è presente. Il dataset corrispondente all'anno 2014 presenta delle colonne anomale in fondo, per cui occorre caricarlo in Open Refine ed eliminare le colonne in eccesso.
Successivamente esportiamo il dataset usando il pulsante export posto in alto a destra della schermata. Selezioniamo il formato CSV.



Importiamo poi uno dei due dataset in phpmyadmin (prima occorre creare un database e poi caricare il dataset nel database). Poi selezioniamo la tabella creata e selezioniamo l'opzione importa. Stiamo attenti ad importare il nuovo dataset dentro la tabella e non nel database. Quindi selezioniamo l'altro dataset e premiamo il tasto esegui. A questo punto i due dataset dovrebbero essere uniti.