

Transforming Data from the Web with OpenRefine

Wikimedia Wikimeet India 2021

Jinoy Tom Jacob
User: Gnoeee

Ranjith Siji
User:Ranjithsiji

Wikimedians of Kerala User Group ([Q57414284](#))

What is Openrefine

- A tool based on Java
- Free and Open Source Software
- Download from openrefine.org
- Run in your favourite browser
- Powerful tool for working with messy data

Why OpenRefine ?

- A Powerful tool for exploring, cleaning, and transforming data from one format into another.
- Used for scraping data, clean it and connect it with knowledge bases, like Wikidata, the free and open knowledge base.
- It always keeps **your data private**; It's **open sourced**.
- Previously known by Freebase Gridworks (2010); Google Refine (2012) and currently a community-supported project.

Why OpenRefine ?

- Can import different data files; like CSV, Excel, JSON, Google data
- Filter the rows to display using facets (filtering criteria)
- Export is supported in TSV, CSV, Excel, Google sheets and custom template for outputting data, like as MediaWiki table.

Data Transformation

- Data **transformation can be simple** or complex based on the required changes to the data.
- Transformation expressions written in General Refine Expression Language (GREL), Jython
- No formulas are stored in the cells. Only used to transform the data.

Wikidata is Short

Statement =

Item -> Property -> Value

Eg: Aamir Khan -> Occupation-> film actor

(Q9557) -> (P106) -> (Q10800557)

Reconciliation

- It's the process of linking free-text tabular cells to identifiers in knowledge bases.
- It have built-in reconciliation capabilities to reconcile tabular data.
- Use multiple columns and match them against values of properties in Wikidata

Live Demo - Open Refine

- First you can go on <https://openrefine.org> to download the tool.
- Download the zip, unzip it and open it. Open folder and click on the blue diamond icon; The tool opens in your browser

OpenRefine tool



Create Project

Open Project

Import Project

Language Settings

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:

Choose Files

No file chosen

Next »

**Home Page of the tool that gets
opened in local browser**

Facet

5 rows

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

▼ All	▼ #	▼ School Name	▼ Block Name	▼ Local Body Type	▼ Assembly Name	▼ Finance Type
Transform		22070 - G. H. S. Kannattupadam	Kodakara	Varandarappilly(P)	Pudukkad	Government
Facet			Ollukkara	Madakkathara(P)	Ollur	Government
Edit rows						
Edit columns			Mala	Kuzhur(P)	Kodungallur	Government
View						

Facet by star

Facet by flag

Facet by blank (null or empty string)

Blank values per column

Blank records per column

Non-blank values per column

Non-blank records per column

The Facets in OpenRefine that helps to filter the rows by stared/flag/null rows

Editing a Column

3 rows Schema Issues **1** Preview

Show as: **rows** records Show: 5 10 25 50 rows « first < > last »

▼ All	▼ School Name	▼ Block Name	▼ Local Body Type	▼ Assembly Name	▼ Finance Type	▼ Level Name	▼ School Phone	▼ School WIK
☆	1. 22070 - G. H. S. Kannattupadam	Kodakara	Varandarappilly(P)	Pudukkad	Government	1 -10	04802760379	
☆	2. 22081 - G. H. S. S. Kattilapoovam	Ollukkara	Madakkathara(P)	Ollur	Government	1 -10	04872695264	
☆	3. 23033 - G H S KUZHUR	Mala	Kuzhur(P)	Kodungallur	Government	1 -10	04802779496	

Split column School Name into several columns

How to Split Column

☒ by separator
Separator ☐ regular expression
Split into columns at most (leave blank for no limit)

☐ by field lengths

List of integers separated by commas, e.g., 5, 7, 15

After Splitting

☒ Guess cell type
☒ Remove this column

We can sperate a cell into different ones by a seperator

Reconciliation

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

▼ All	▼ School Name 1	▼ School Name 2	▼ Block Name	▼ L
★	1.	22070	Kodakara	Varan
★	2.	22081	Ollukkara	Mada
★	3.	23033	Mala	Kuzhu

- Facet
- Text filter
- Edit cells
- Edit column
- Transpose
- Sort...
- View
- Reconcile
 - Start reconciling...
 - Facets
 - Actions
 - Copy reconciliation data...
 - Use values as identifiers
 - Add entity identifiers column

**Select the column to
reconcile the data in cells
with knowledge base.
Reconcile > Start reconciling**

Reconciliation

Include data that can be matched with other data

We can Reconcile the data we have with matched with data with Wikidata

Select one of the method for reconcile the data

Reconcile column "School Name 2"

» Access [Service API](#)

Reconcile each cell to an entity of one of these types:

☒ high school
Q9826

Also use relevant details from other columns:

Column	Include?	As Property
School Name 1	<input checked="" type="checkbox"/>	Kerala state school code
Start Grade	<input type="checkbox"/>	
End Grade	<input type="checkbox"/>	
Hi	<input type="checkbox"/>	
Block Name	<input type="checkbox"/>	
Local Body Type	<input type="checkbox"/>	
Assembly Name	<input type="checkbox"/>	
Finance Type	<input type="checkbox"/>	
Level Name	<input type="checkbox"/>	
School Phone	<input type="checkbox"/>	
School WIKI	<input type="checkbox"/>	
Location	<input type="checkbox"/>	

☐ Reconcile against type:

☐ Reconcile against no particular type

☒ Auto-match candidates with high confidence

Maximum number of candidates to return

Add Standard Service...

Start ReconcilingCancel

Data Retrieving

School Name 2	Hi	Block Name	Loc
Facet	എച്ച്. എസ്. കന്നട്‌പാടം	Kodakara	Varanda
Text filter	എച്ച്. എസ്. എസ്. കടവിലപ്പാടം	Ollukkara	Madakkara
Edit cells			
Edit column	Split into several columns...		
Transpose	Join columns...		
Sort...	Add column based on this column...		
View	Add column by fetching URLs...		
Reconcile	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		

Retrieve data from reconciled data.

Edit column > Add columns from reconciled values

Retrieving labels, description or Qids

Add columns from reconciled column School Name 2

Add Property

SPARQL: Lml

Suggested Properties

Qid

Preview

School Name 2	Lml remove configure
Ghs Kannattupadam	ഗവ എച്ച് എസ് കന്നാറ്റുപാടം
Ghss Kattilapooam	ജി.എച്ച്. എസ് എസ് കട്ടിലപ്പാടം
Ghs Kuzhur	ജി. എച്ച്. എസ്. കുഴൂർ

Fetching data from Web

Add column by fetching URLs based on column School Name 1

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 ▼

'https://sametham.kite.kerala.gov.in/'+value

No syntax error.

Preview

[History](#)

[Starred](#)

[Help](#)

row	value	'https://sametham.kite.kerala. ...
1.	22070	https://sametham.kite.kerala.gov.in/22070
2.	22081	https://sametham.kite.kerala.gov.in/22081
3.	23033	https://sametham.kite.kerala.gov.in/23033

OK

Cancel

We can fetch websites for scrapping data from the web

We can use expressions for fetching data from the web

Note: Texts are included in single inverted comma

Scrapping the data

▼ School Name 1	▼ HTML
22070	<pre><!doctype html> <!--[if lt IE 7]> <html class="no-js lt-ie9 lt-ie8 lt-ie7" lang=""> <![endif]--> <!--[if IE 7]> <html class="no-js lt-ie9 lt-ie8" lang=""> <![endif]--> <!--[if IE 8]> <html class="no-js lt-ie9" lang=""> <![endif]--> <!--[if gt IE 8]><!--> <html class="no-js" lang="en"> <style> @page { size: auto; margin: 3mm; } @media print { a { display:none; } aside.left-panel { display:none; } } @media screen and projection { a { display:inline; } } @media print { title, #header, #footer .header, .footer { visibility: hidden } } .panhead { background-color: #2980b9 !important; color: #ffffff; } </style></pre>

We can scrap the data from the HTML codes easily with some small small cleanups

We can use several methods in OR for scrapping the data that we need.

Here I am using a simple method to scrape data without using any expressions or codes. We can use split column method to scarp data here

Note: If you doesn't have any programming skills to you can use this tool for cleanup :)

Wikidata Scheme

3 rows Schema * Issues 1 Preview

Extensions: Wikidata ▼

The Wikidata schema below specifies how your tabular data will be transformed into Wikidata edits. You can drag and drop the column names below in most input boxes: for each row, edits will be generated with the values in these columns.

School Name 1 Start Grade End Grade School Name 2 Hi Block Name Local Body Type Assembly Name Finance Type Level Name School Phone

Click on Edit Wikidata Schema for adding Wikidata scheme which looks similar to Wikidata layout

- Edit Wikidata schema
- Manage Wikidata account
- Import schema
- Export schema
- Upload edits to Wikidata
- Export to QuickStatements

A Wikidata Scheme with labels in Hindi language is added along with some statements

School Name 2

remove

Terms

Label ▼ hi Hi remove

☐ override if present

+ add term

Statements

start grade Start Gra... remove

+ add qualifier

► 0 references

+ add reference

+ add value

end grade End Gra... remove

+ add qualifier

► 0 references

+ add reference

+ add value

+ add statement

+ add item

Editing Wikidata


3 rows

Schema

Issues **6**

Preview

Extensions: Wikidata ▾

**Statements without references.**
Most statements should have references. You can add them easily in the schema.

6

Always check for issues before uploading data into Wikidata

Check the Preview view for viewing how the edits will after the data gets uploaded to Wikidata

3 rows

Schema

Issues **6**

Preview

This tab shows the first edits (out of 3) that will be made once you upload the changes to Wikidata. You can use facets to inspect the edits on particular items.

[Ghs Kannattupadam \(Q64091191\)](#)

Label (do not override)	गि एच् एस् कन्नट्टुपाडम् (Hindi)
end grade (P7095)	<div>tenth grade (Q3269996)</div> <div>0 references</div>
start grade (P7086)	<div>first grade (Q8563383)</div> <div>0 references</div>

You can directly upload the data from Openrefine to Wikidata or can use export directly from here to Quickstatements



Thank you

Jinoy Tom Jacob
User: Gnoeee

Ranjith Siji
User:Ranjithsiji

Wikimedians of Kerala User Group (Q57414284)
<https://w.wiki/t9>