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 currenltly 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)

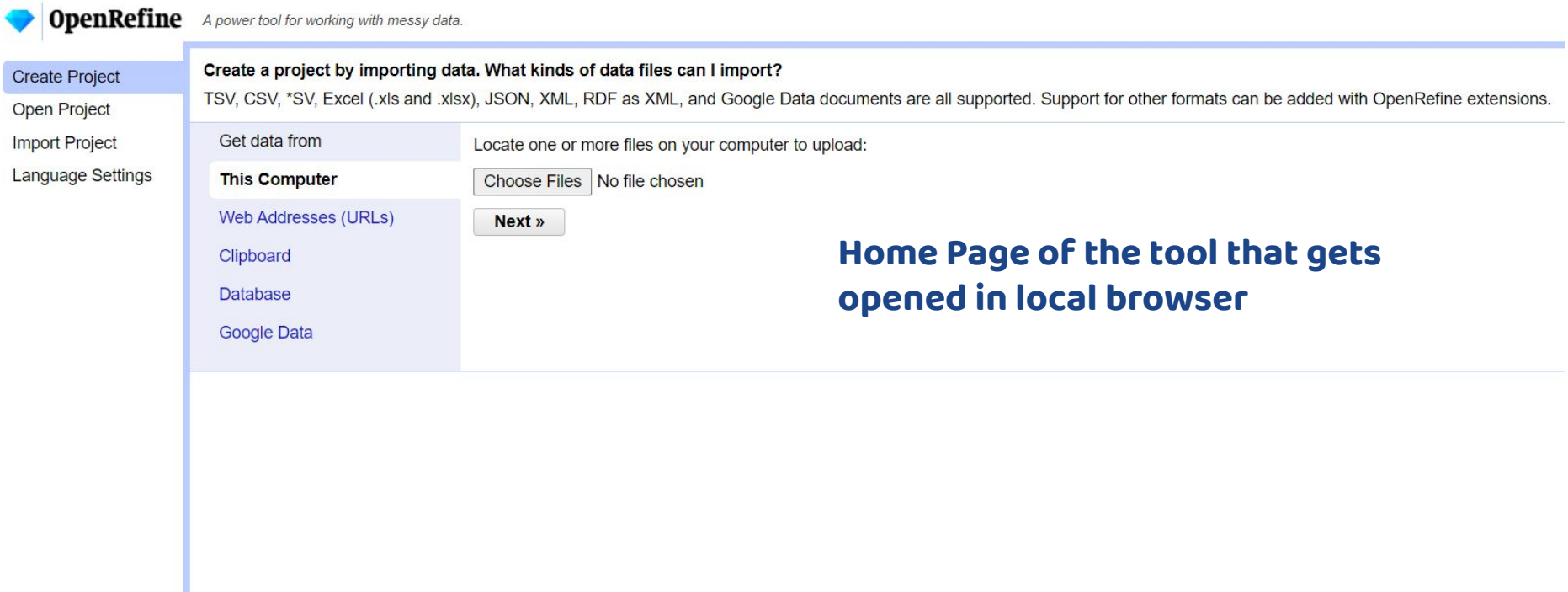
Reconciliation

- 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

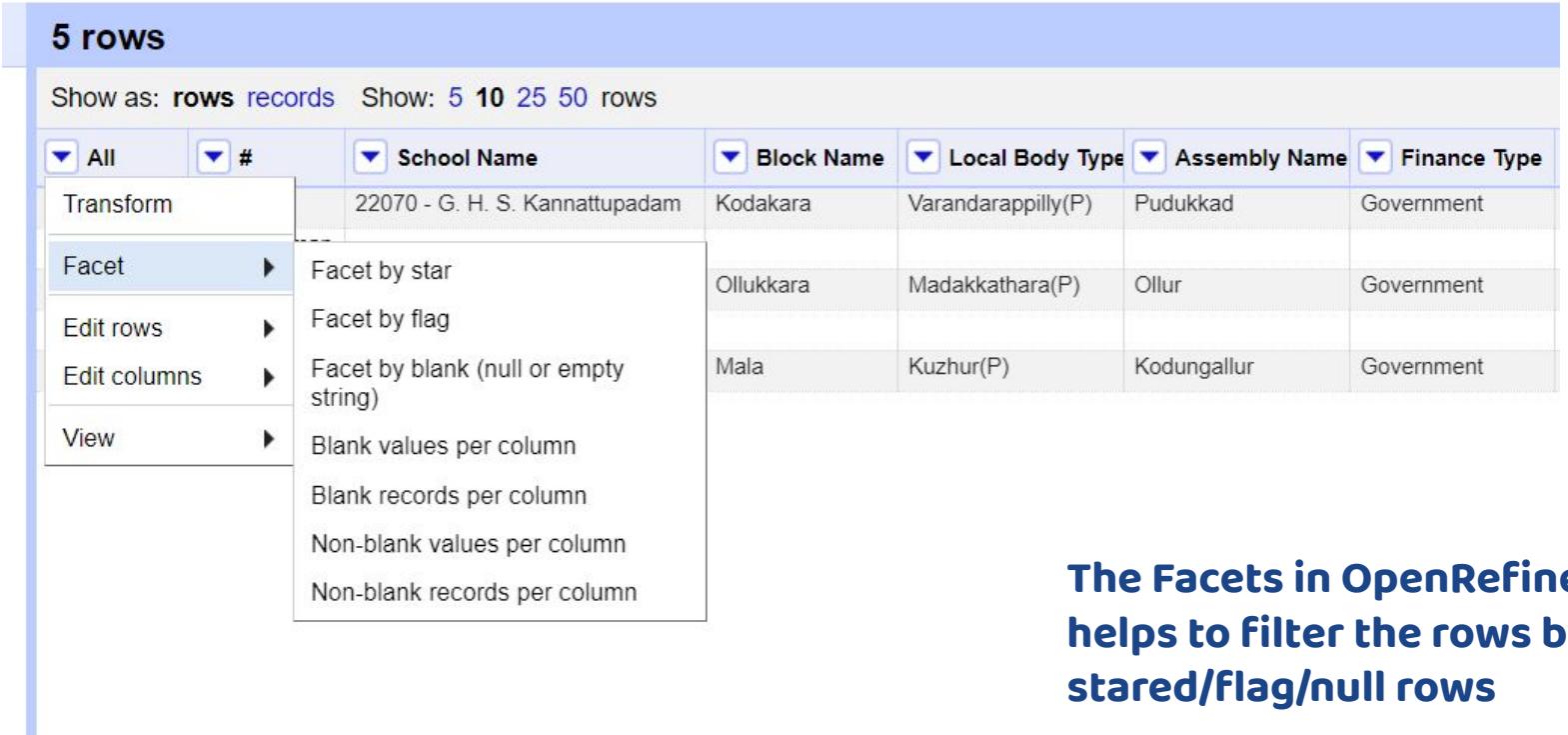
- First you can go on https://openrefine.org to download the tool.
- Downlaod the zip, unzip it and open it. Open folder and click on the blue diamond icon); The tool opens in you browser

OpenRefine tool



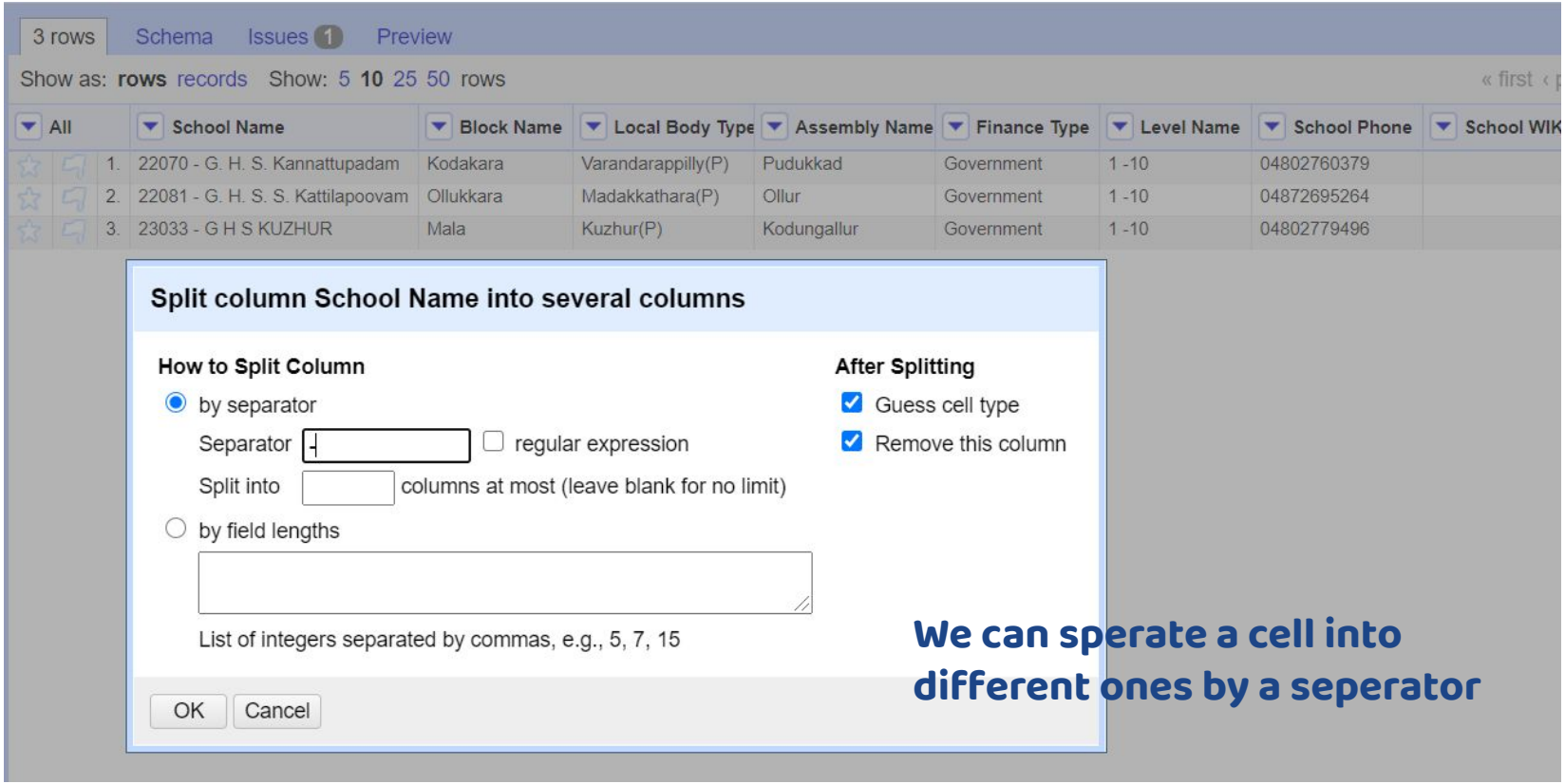
Home Page of the tool that gets opened in local browser

Facet



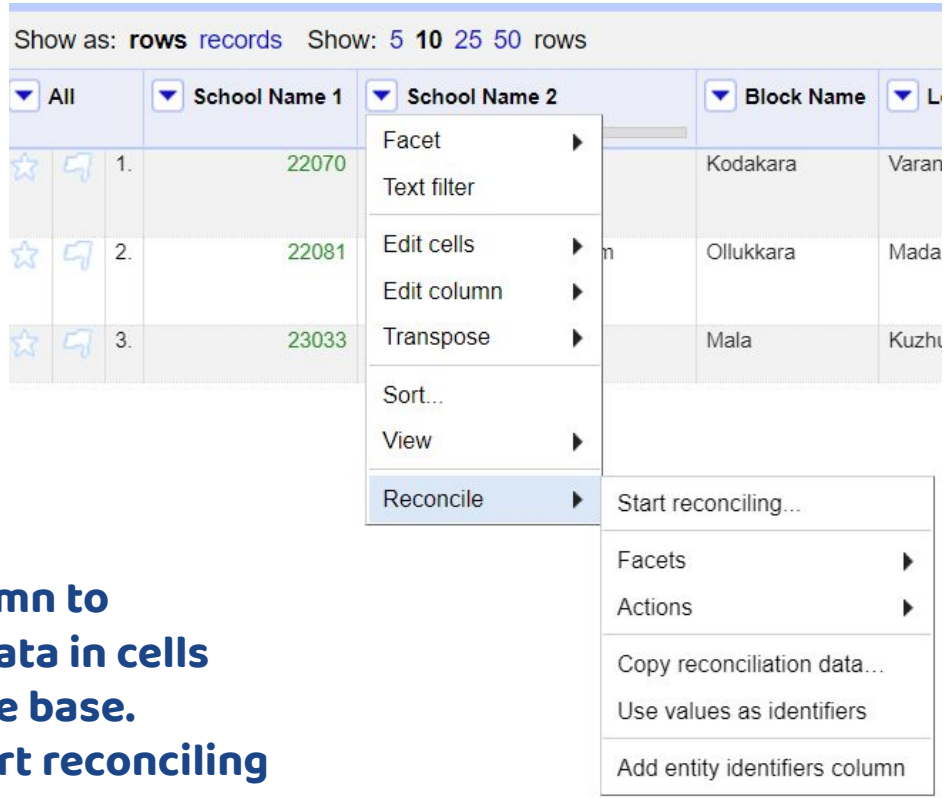
The Facets in OpenRefine that helps to filter the rows by stared/flag/null rows

Editing a Column



We can sperate a cell into different ones by a seperator

Reconciliation



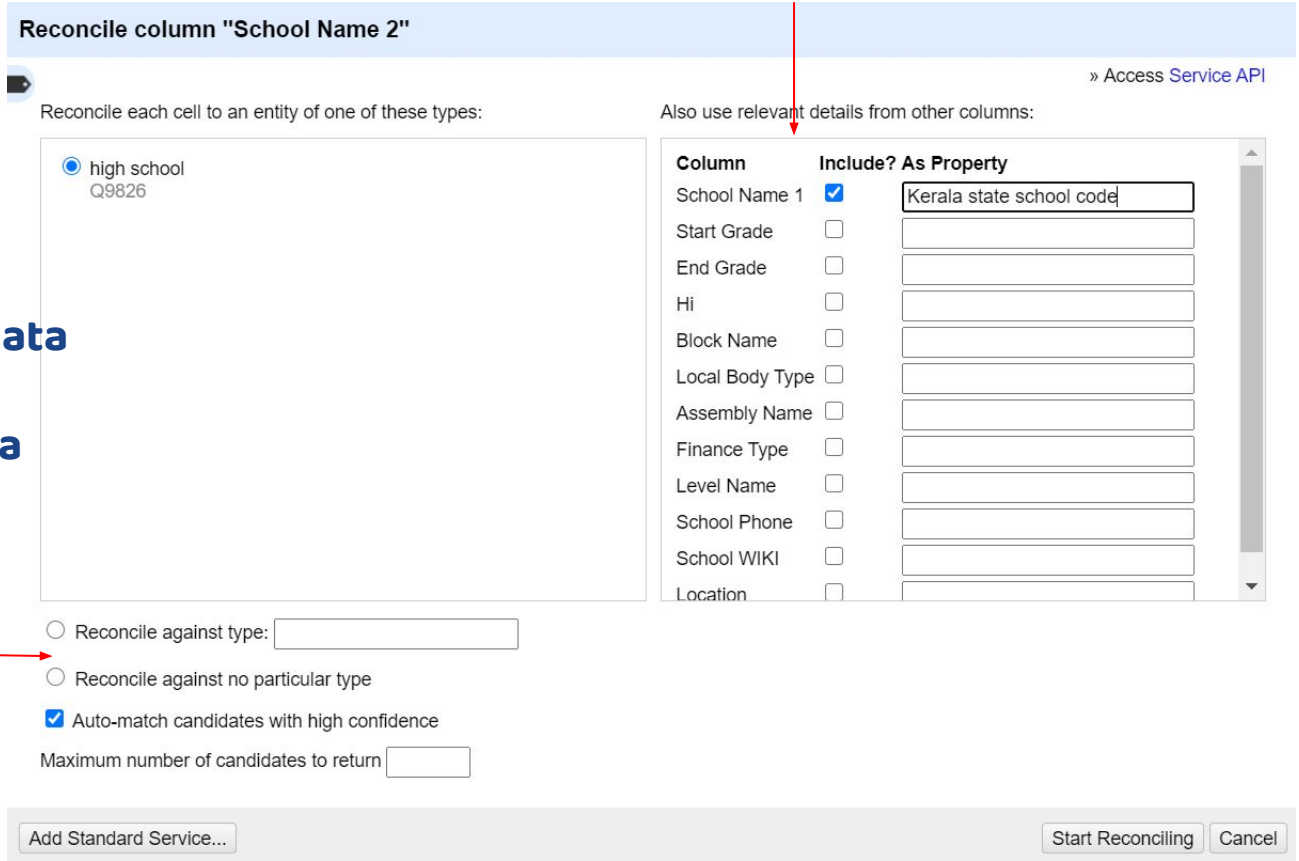
Select the column to reconcile the data in cells with knowledge base.
Reconcile > Start reconciling

Reconciliation

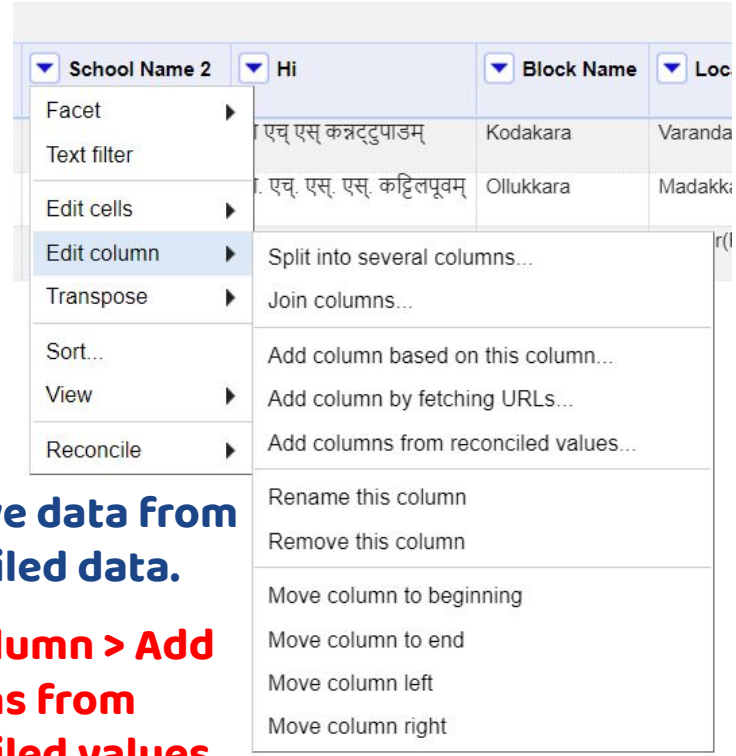
Include data that can be matched with other data

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

Select one of the method for reconcile the data



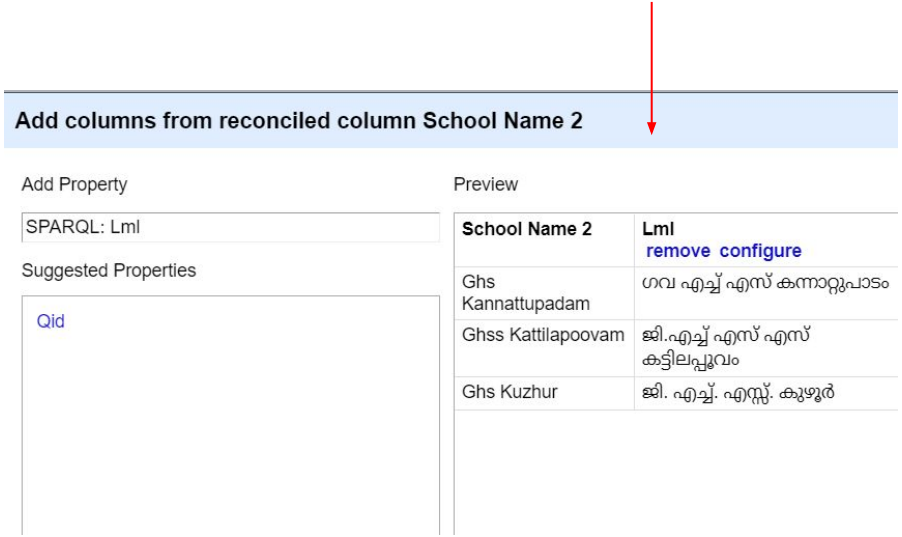
Data Retrieving



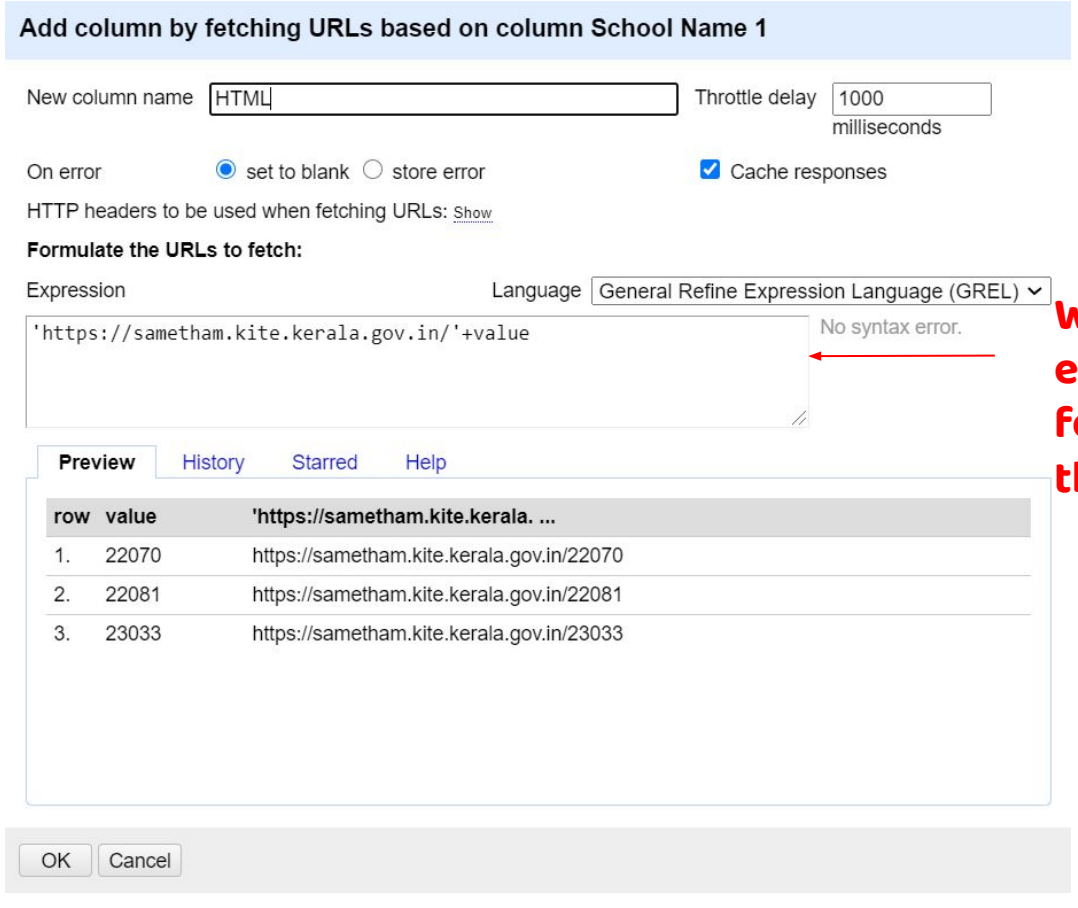
Retrieve data from reconciled data.

Edit column > Add columns from reconciled values

Retrieving labels, description or Qids



Fetching data from Web



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	<Idoctype 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>

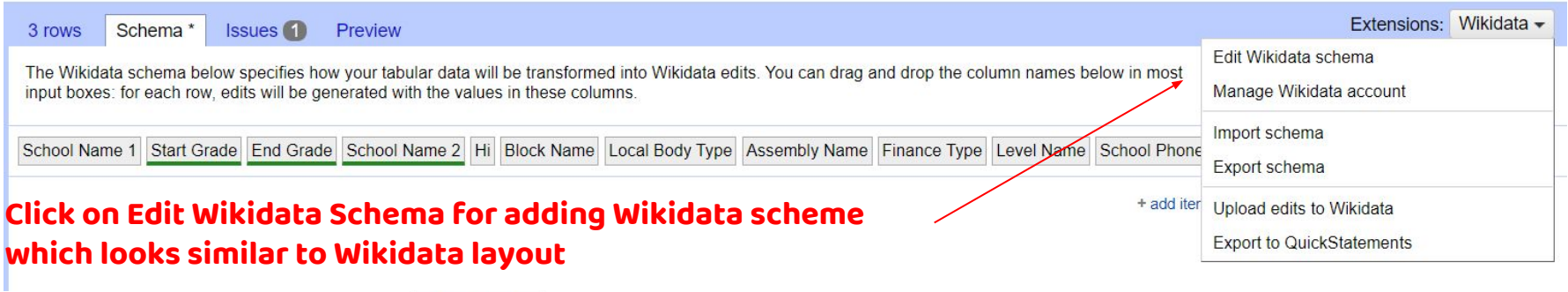
We can scrap the data from the HTML codes easily with some small small cleanups

We can use several menthods 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

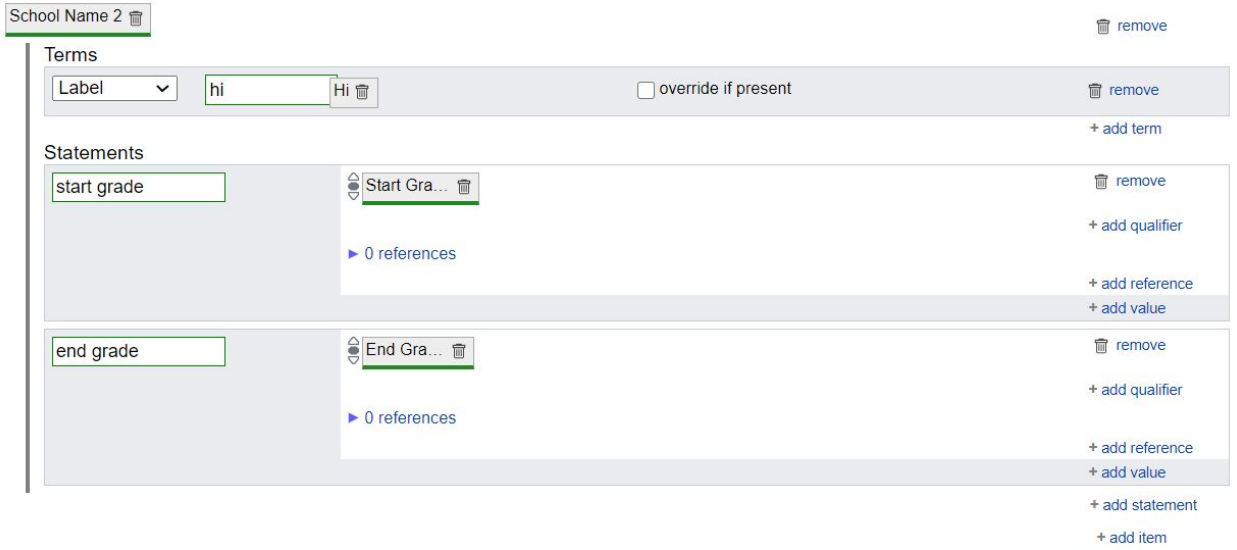
Note: If you doen't have any programming skills to you can use this tool for cleanup :)

Wikidata Scheme

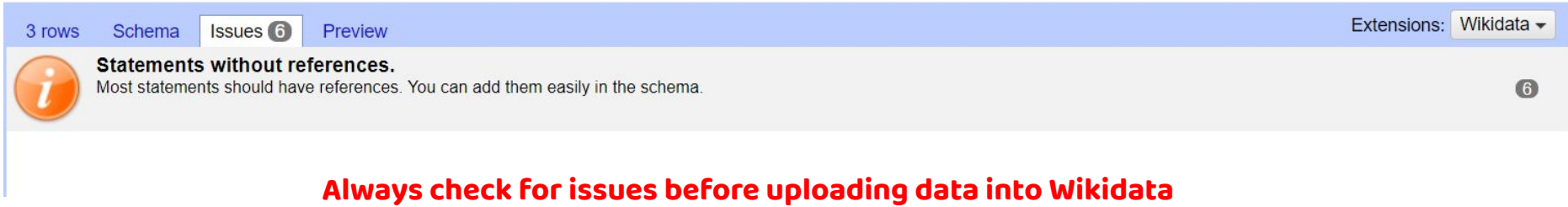


Click on Edit Wikidata Schema for adding Wikidata scheme which looks similar to Wikidata layout

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



Editing Wikidata



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

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