

«Предварительная подготовка данных к публикации»

Работа по очистке набора данных при помощи Open Refine

1. Загрузка Open Refine

Зайдите в каталог, где у вас загружен Open Refine

Для того чтобы загрузить Open Refine, необходимо дважды щелкнуть мышью по файлу google-refine (рис. 1). Open Refine загрузится в браузере.

.import-temp	29.10.2013 21:48	Папка с файлами	
licenses	11.12.2011 14:08	Папка с файлами	
server	11.12.2011 14:08	Папка с файлами	
webapp	11.12.2011 14:08	Папка с файлами	
google-refine	11.12.2011 14:08	Приложение	81 КБ
google-refine.l4j	11.12.2011 14:08	Параметры конф...	1 КБ
LICENSE	11.12.2011 14:08	Текстовый докум...	4 КБ
README	11.12.2011 14:08	Текстовый докум...	2 КБ
refine	11.12.2011 14:08	Пакетный файл ...	5 КБ
refine	11.12.2011 14:08	Параметры конф...	1 КБ

Рис. 1 — Загрузка Open Refine

В появившемся окне необходимо закрыть Google Chrome Frame. Для этого нужно нажать на кнопку **Close** в правом верхнем углу (рис. 2).

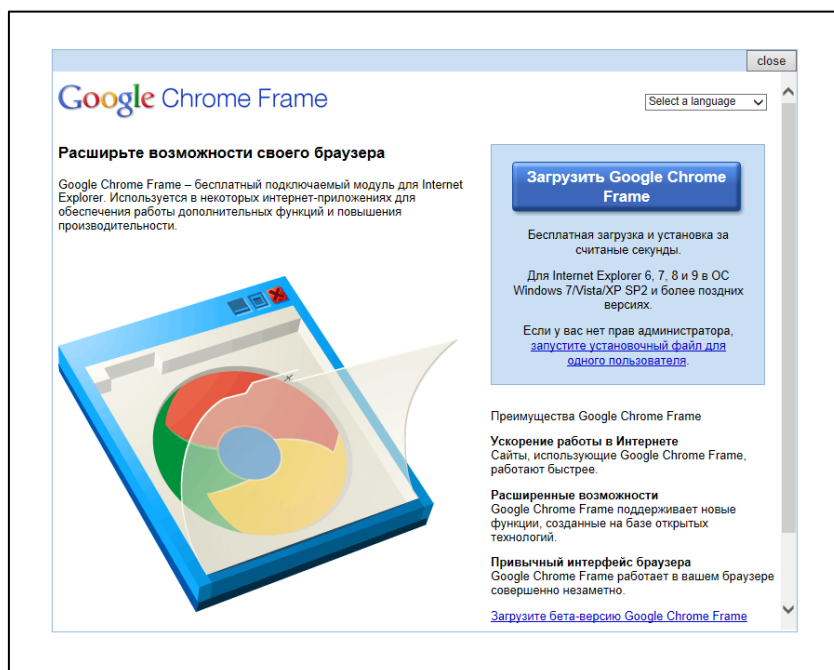


Рис. 2 — Закрыть Google Chrome Frame

2. Создание нового проекта и загрузка в него набора данных

Для того чтобы загрузить набор данных в новый проект удобно воспользоваться опцией **This Computer**, в которой необходимо указать расположение загружаемого набора данных при помощи стандартного диалогового меню, которое отображается после нажатия на кнопку **Обзор...** (рис. 3 и рис. 4).

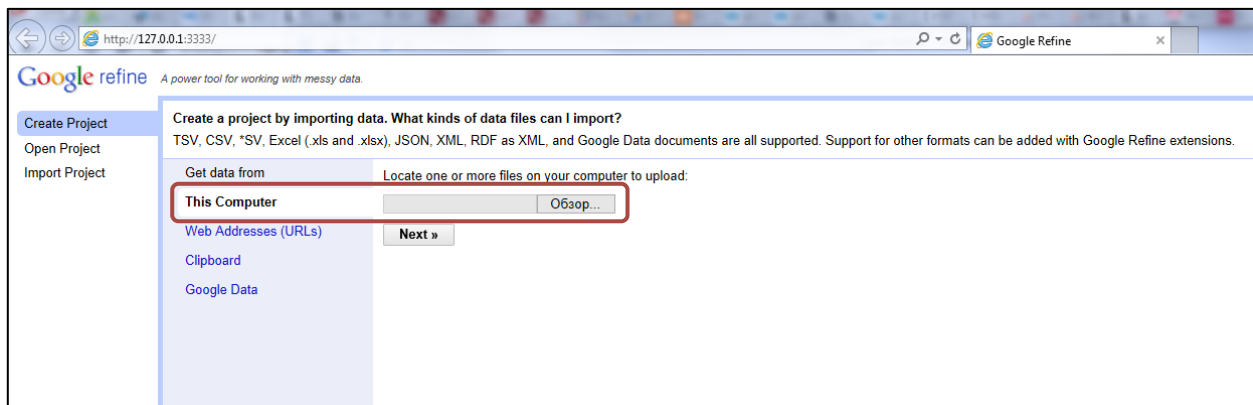


Рис. 3 — Создание нового проекта

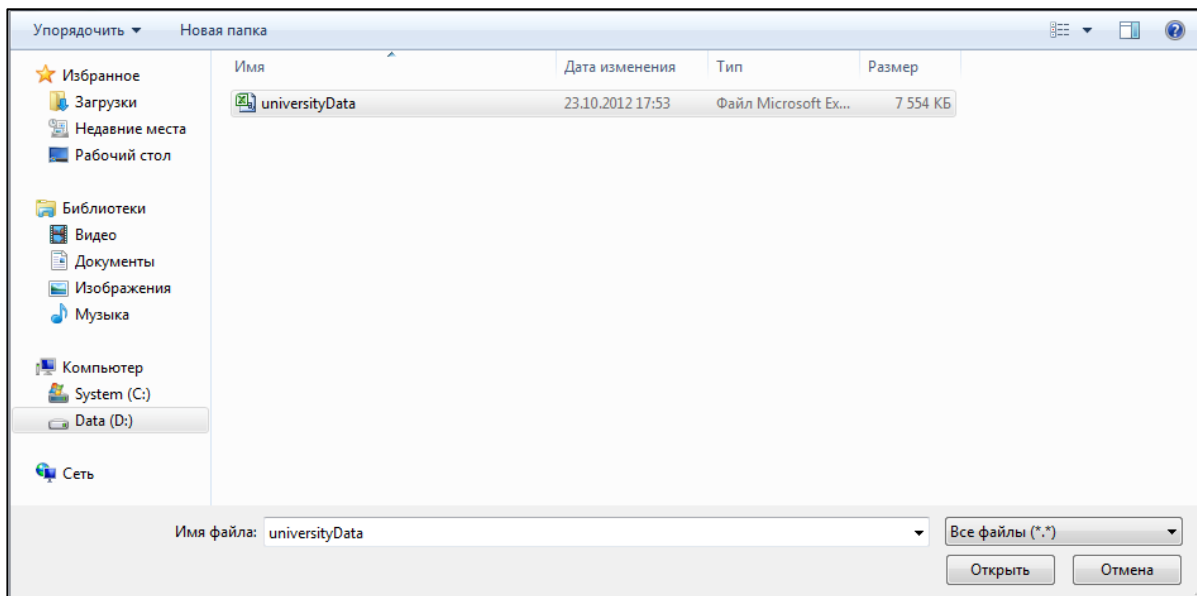


Рис. 4 — Выбор набора данных для загрузки

Выбрать загружаемый набор данных (**universityData.csv**) и затем надо нажать на кнопку **Next** (рис. 5).

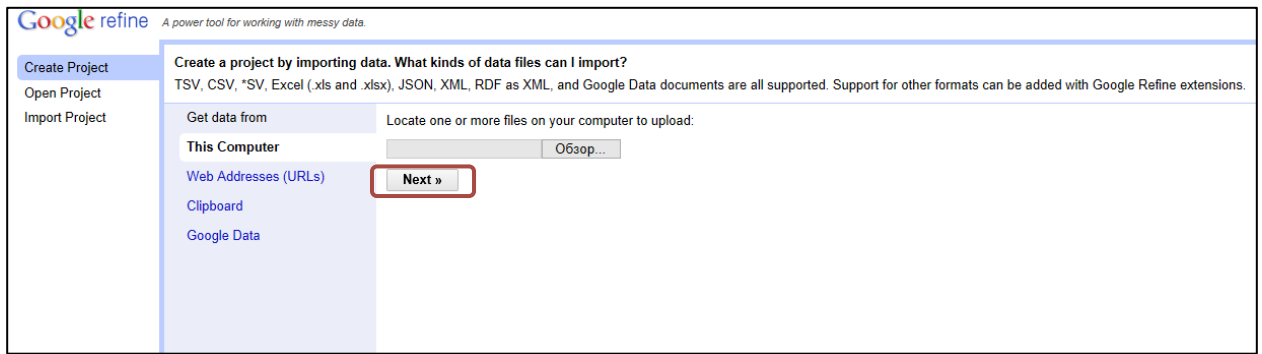


Рис. 5 — Загрузка набора данных

Произойдет загрузка набора данных в новый проект. И затем надо нажать на кнопку **Create Project >>** в верхнем правом углу браузера (рис. 6).

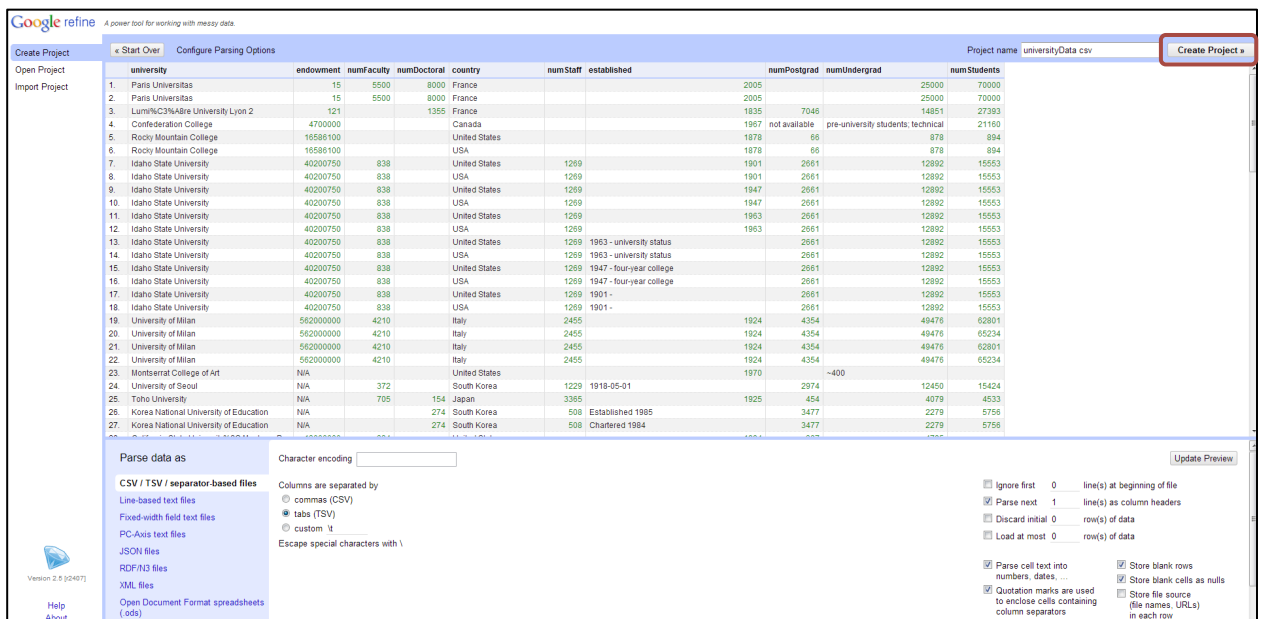



Рис. 6 — Завершение загрузки набора данных

3. Очистка данных в поле country

Данные в колонке **country** (Название страны) содержат различные варианты названия стран. Для того чтобы привести их к единому виду, необходимо щелкнуть мышью по кнопке , которая находится слева от названия колонки **country** и затем выбрать последовательно опции **Edit cells** → **Cluster and edit...** (рис. 7)

«Предварительная подготовка данных к публикации»

Работа по очистке набора данных при помощи Open Refine

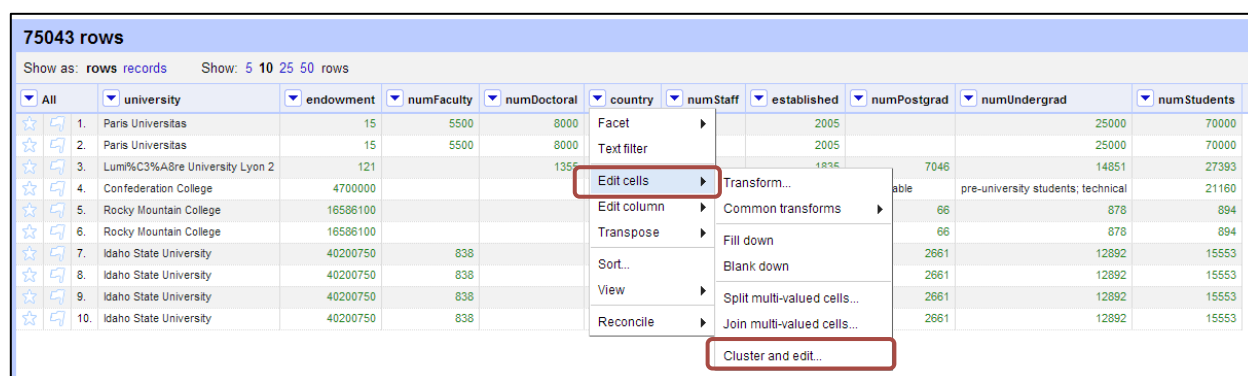


Рис. 7 — Выбор элементов меню для очистки данных в поле **country**

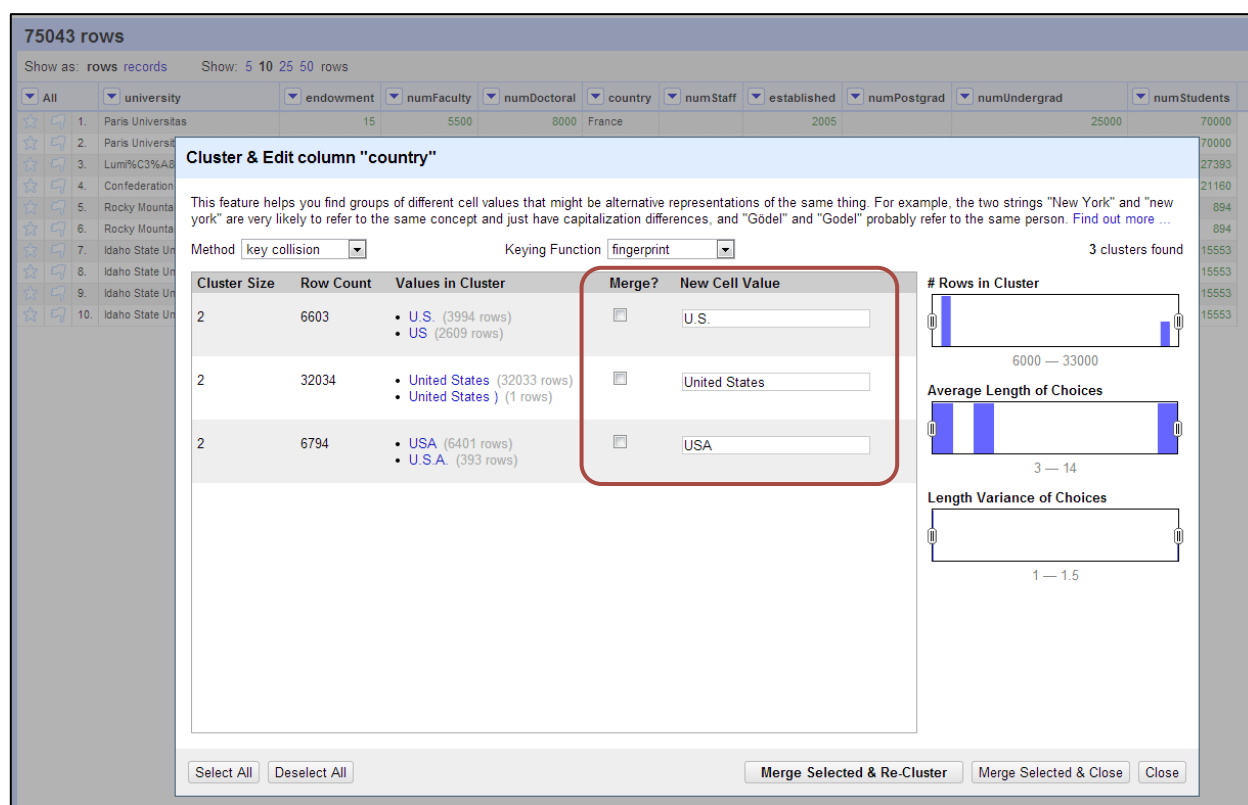
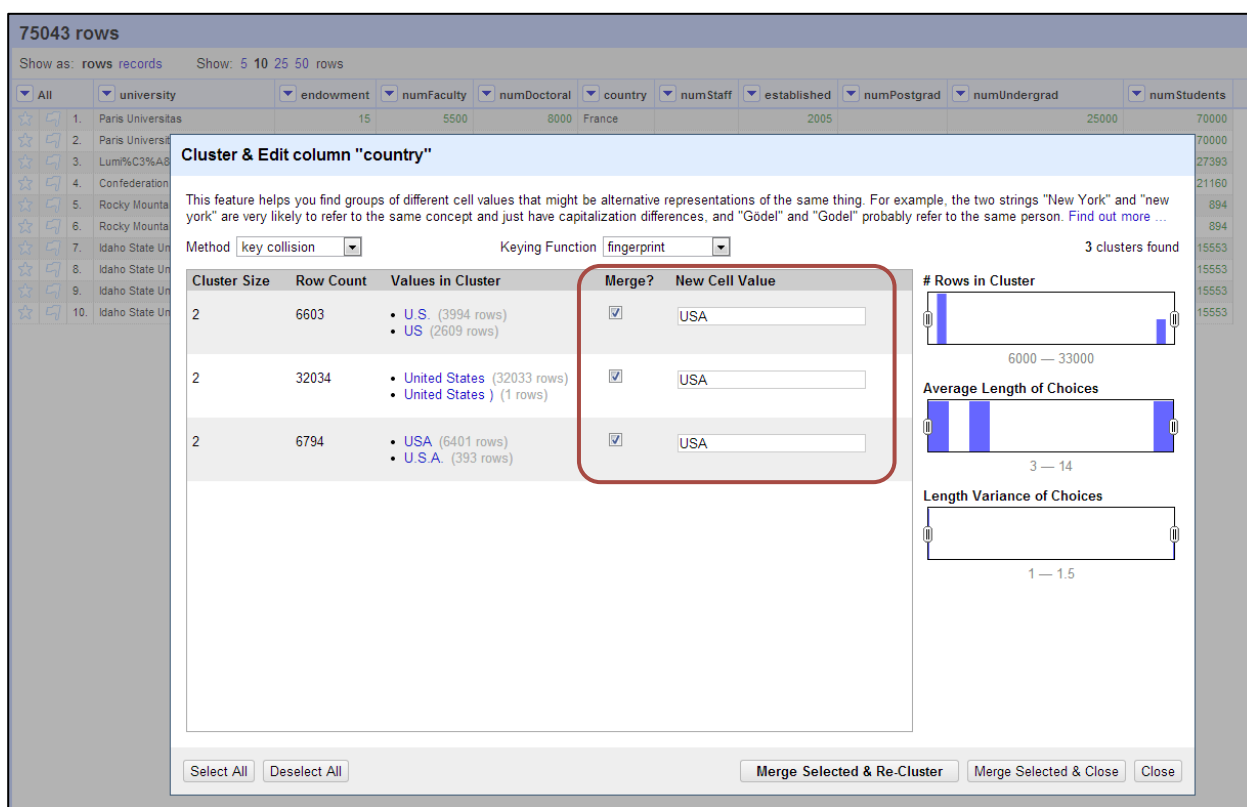


Рис. 8 — Очистка данных в поле **country**

Рис. 9 — Выбор элементов меню для очистки данных в поле **country**

Для наиболее полного поиска наименований, которые могут содержать различные варианты стран, надо выбрать **cologne-phonetic** в выпадающем списке **Key Function** (обратите внимание, что результат зависит от выбранного алгоритма: **fingerprint** и **cologne-phonetic**). И написать новое унифицированное название страны, которое требуется по смыслу (в данном примере: **USA**, **Russia** и **USA**).

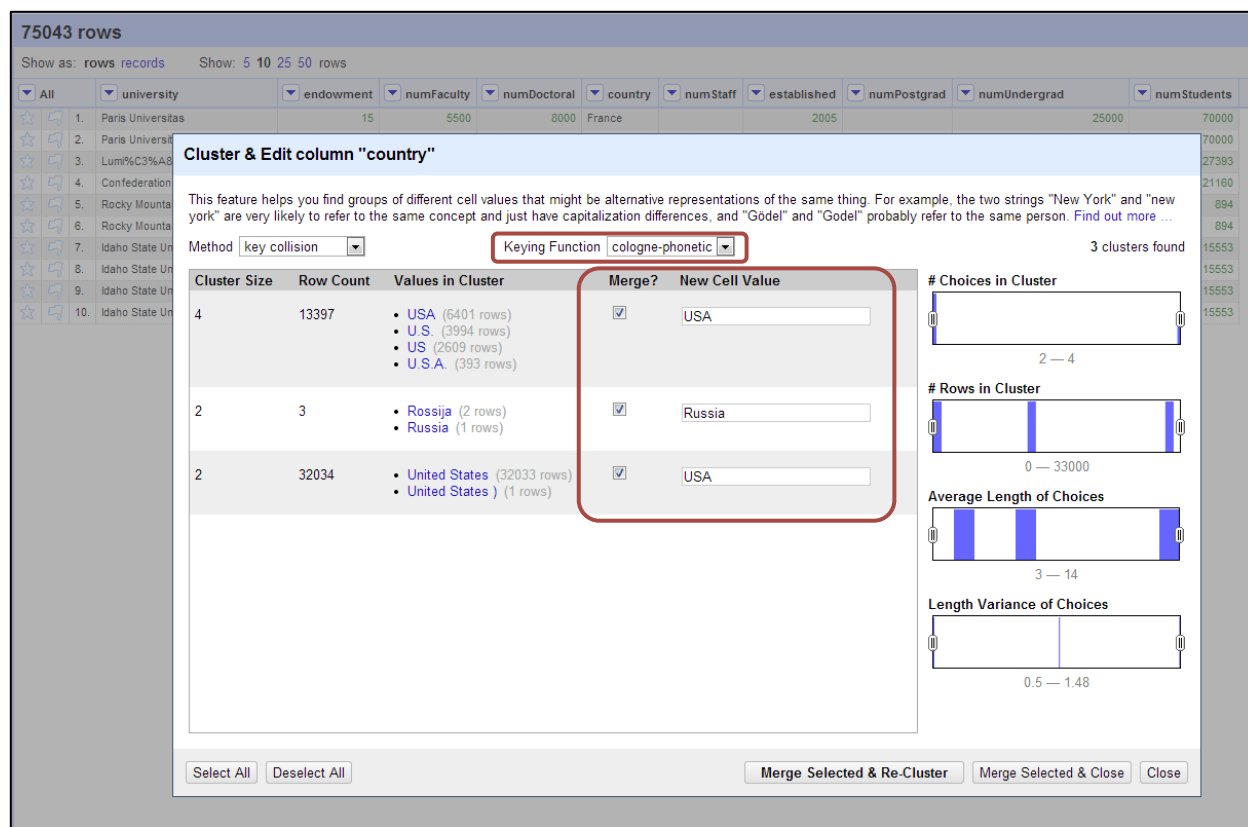
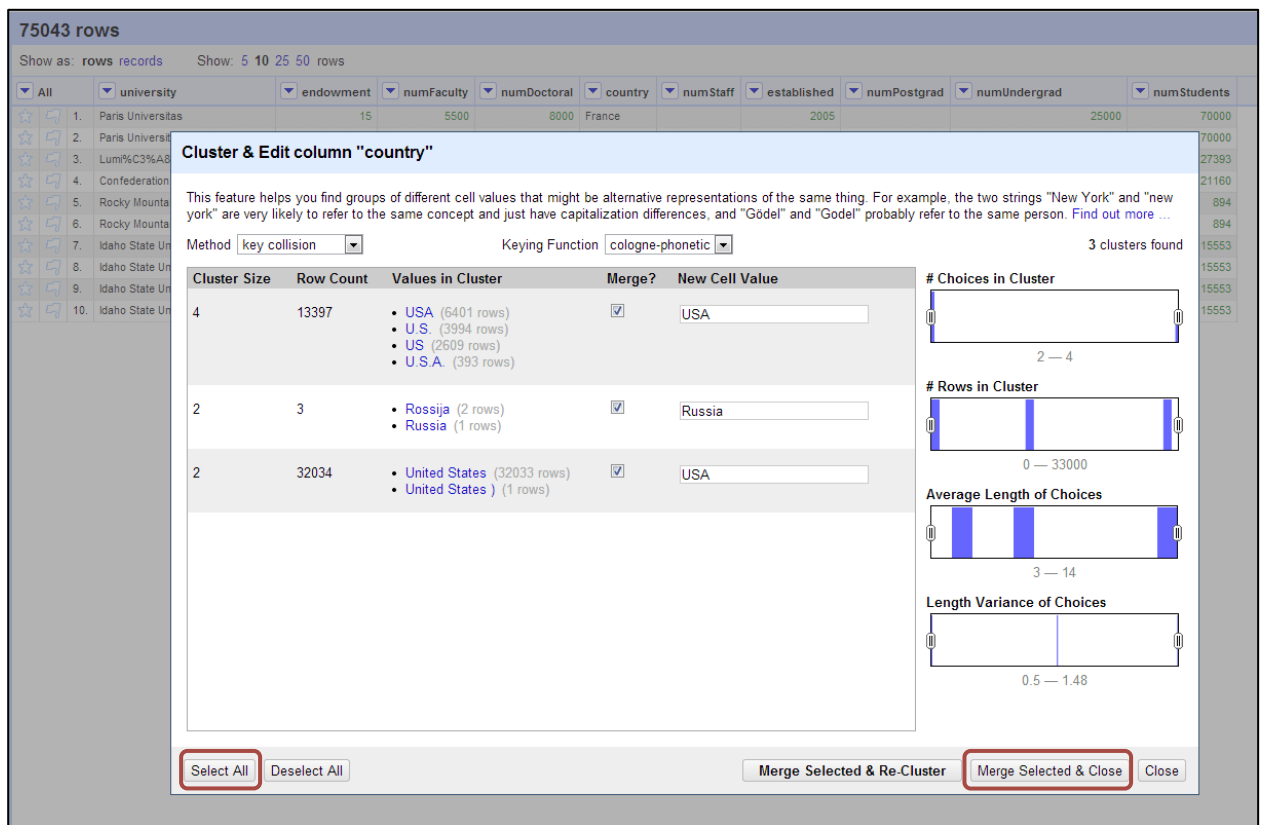


Рис. 10 — Выбор элементов меню для очистки данных в поле **country**


Затем необходимо нажать мышью на кнопку **Select All**, чтобы выделить все поля в колонке **Merge?** И после этого нажать на кнопку **Merge Selected & Close**.

Рис. 11 — Очистка данных в поле **country**

Все различающиеся названия стран стали унифицированы.

4. Очистка данных в поле **numUndergrad** и поле **numStudents**

В этой колонке не все данные имеют числовой формат, многие поля содержат также текст в дополнение к числовым значениям. Эти данные необходимо исправить и привести к единому числовому виду.

Для этого нужно щелкнуть мышью по кнопке , слева от названия колонки **numUndergrad** и затем выбрать последовательно опции **Facet -> NumericFacet** (рис. 12).

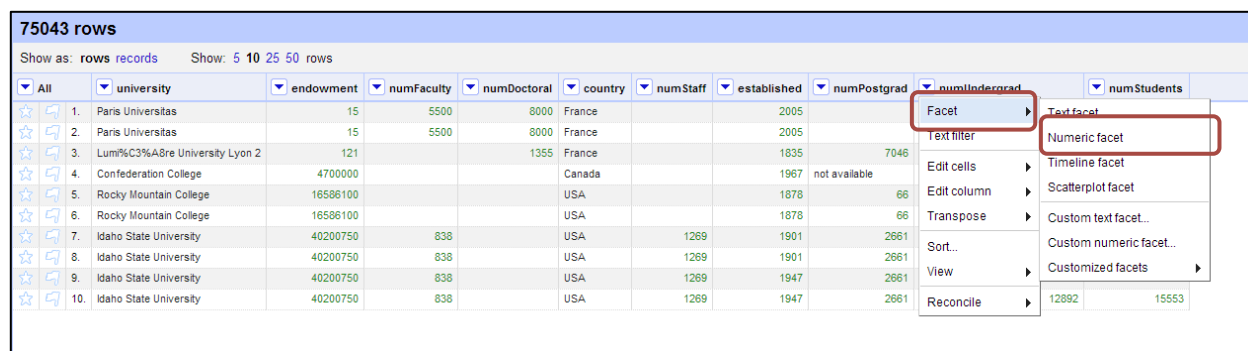


Рис. 12 — Выбор элементов меню для очистки данных в поле **numUndergrad**

На гистограмме слева отобразилось количество **числовых (Numeric)** и **нечисловых (Non-numeric)** записей (рис. 13).

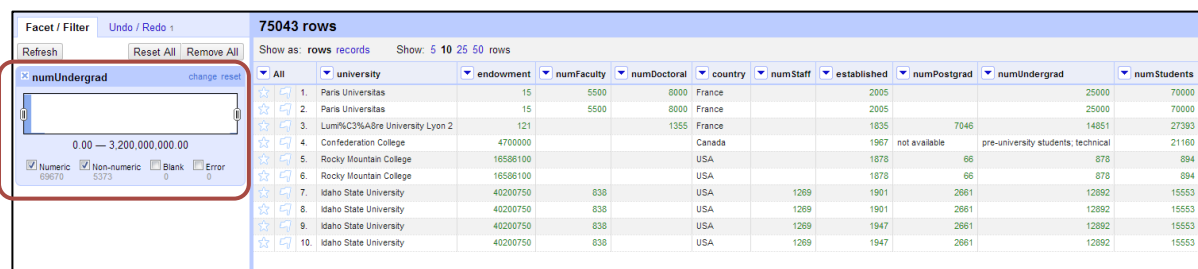


Рис. 13 — Очистка данных в поле **numUndergrad**

Для того чтобы избавиться от знаков «+» и «~» в поле **numUndergrad** необходимо последовательно выбрать опции в выпадающем меню **Edit cells -> Transform**. (рис. 14)

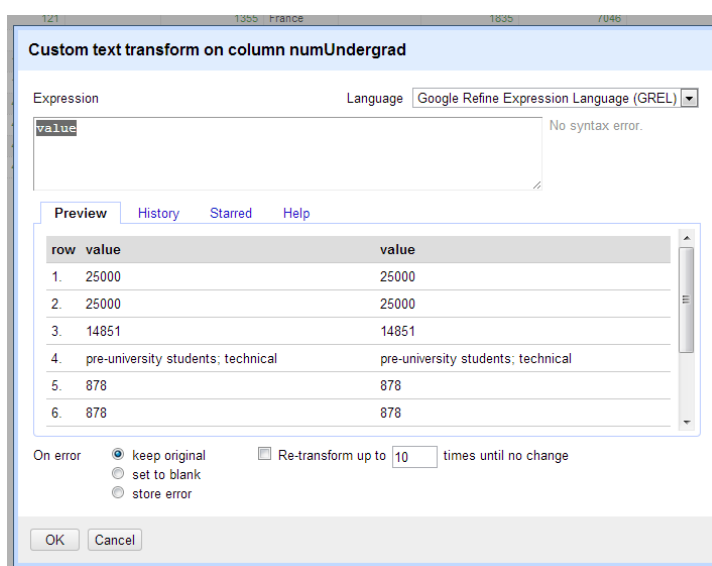


Рис. 14 — Очистка данных в поле **numUndergrad**

Необходимо написать следующее выражение в текстовое поле **Expression**:

value.replace("+", "")

как это продемонстрировано на рис. 15. И затем щелкнуть мышью на кнопке **OK**.

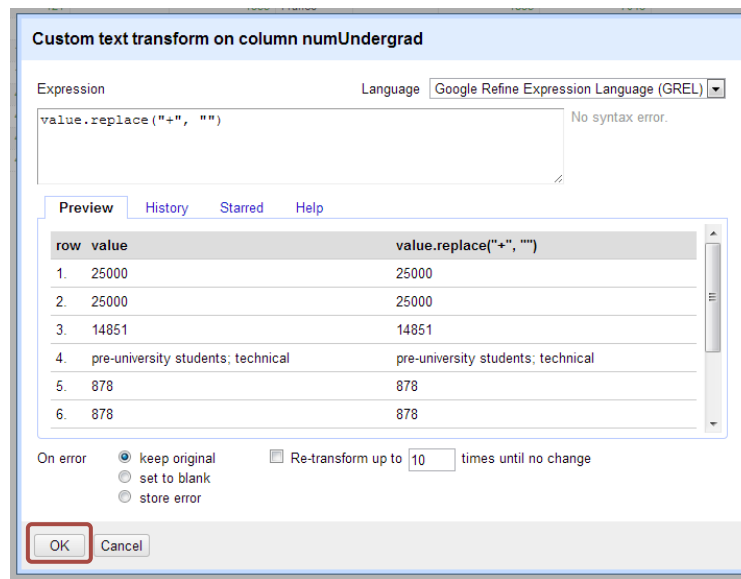


Рис. 15 — Очистка данных в поле **numUndergrad**

Эту же последовательность действий можно выполнить для того, чтобы очистить поле **numUndergrad** от знака ~.

Необходимо написать следующее выражение в текстовое поле **Expression**:

value.replace("~", "").replace(",","")

Необходимо также сконвертировать все нечисловые значения в числовые это можно сделать при помощи последовательного выбора следующих опций меню: **Edit cells -> Common transforms -> To number** (рис. 16).

Работа по очистке набора данных при помощи Open Refine

The screenshot shows the Open Refine interface with a table of university data. The 'numStudents' column contains various values, including '25000', '14851', and '21160'. The 'Edit cells' menu is open, and the 'To number' option is highlighted. The 'Common transforms' menu is also visible, showing options like 'Trim leading and trailing whitespace', 'Collapse consecutive whitespace', 'Unescape HTML entities', 'To titlecase', 'To uppercase', 'To lowercase', 'To number', 'To date', 'To text', and 'Blank out cells'.

Рис. 16 — Очистка данных в поле **numStudents**


Таким образом, можно привести все значения поля **numStudent** к цифровому виду.

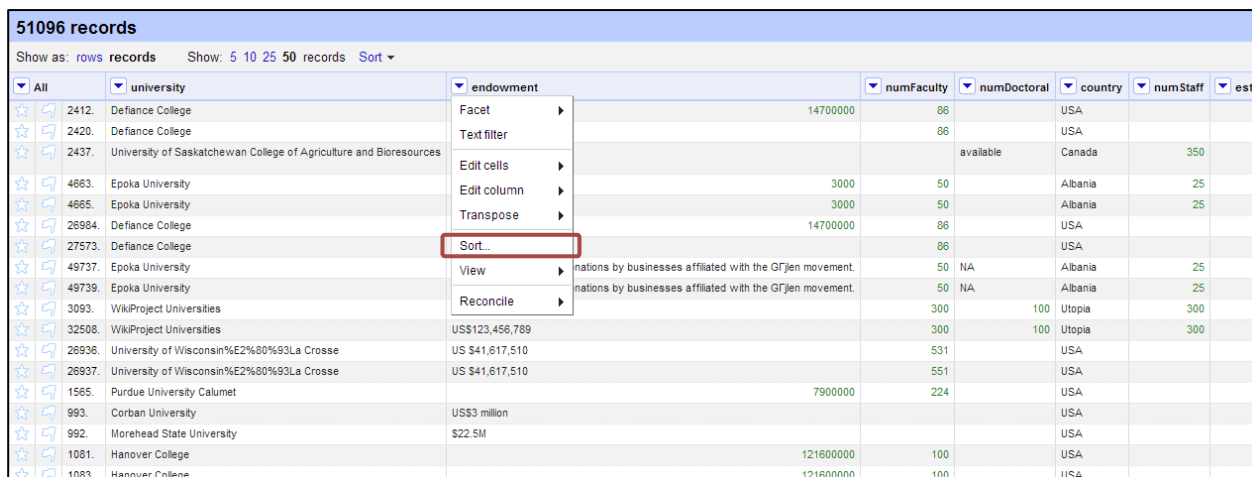
5. Дополнительная функциональность Open Refine

При необходимости можно изменить количество показываемых строк в таблице при помощи строки меню **Show:** (рис. 17).

The screenshot shows the Open Refine interface with a table of university data. The 'Show:' menu is open, and the '51096 rows' option is highlighted. The 'Using facets and filters' sidebar is visible on the left, and the 'Facet / Filter' menu is at the top left.

Рис. 17 — Строка меню **Show:**

В Open Refine также предусмотрена возможность сортировки записей при помощи выбора опции **Sort...** в выпадающем меню, которое вызывается при помощи щелчка мыши на кнопке  (рис. 18).



51096 records

Show as: rows records Show: 5 10 25 50 records Sort ▾

	university	endowment	numFaculty	numDoctoral	country	numStaff	esta
2412.	Defiance College	14700000	86		USA		
2420.	Defiance College	86			USA		
2437.	University of Saskatchewan College of Agriculture and Bioresources			available	Canada	350	
4863.	Epoka University	3000	50		Albania	25	
4865.	Epoka University	3000	50		Albania	25	
26984.	Defiance College	14700000	86		USA		
27573.	Defiance College	86			USA		
49737.	Epoka University	nations by businesses affiliated with the GfJen movement.	50	NA	Albania	25	
49739.	Epoka University	nations by businesses affiliated with the GfJen movement.	50	NA	Albania	25	
3093.	WikiProject Universities		300	100	Utopia	300	
32508.	WikiProject Universities	US\$123,456,789	300	100	Utopia	300	
26936.	University of Wisconsin%E2%80%93La Crosse	US \$41,617,510	531		USA		
26937.	University of Wisconsin%E2%80%93La Crosse	US \$41,617,510	551		USA		
1565.	Purdue University Calumet	7900000	224		USA		
993.	Corban University	US\$3 million			USA		
992.	Morehead State University	\$22.5M			USA		
1081.	Hanover College	121600000	100		USA		
1083.	Hanover College	121600000	100		USA		

Рис. 17 — Сортировка записей