

Использование



Оглавление

УСЛОВНЫЕ ОБОЗНАЧЕНИЯ.....	2
2. ОСНОВЫ РАБОТЫ.....	3
2.1. ЗАГРУЗКА ФАЙЛОВ И ПРОЕКТЫ	3
2.1.1. Перед началом	3
2.1.2. Упражнение 1. Создание проекта	3
2.2. СВОДКА (FACETING).....	4
2.2.1. Перед началом	4
2.2.2. УПРАЖНЕНИЕ 2. Сводка и массовое редактирование	4
2.2.4. УПРАЖНЕНИЕ 4. Сводка и пробелы II.....	6
2.2.5. УПРАЖНЕНИЕ 5. Сводка и дубликаты	7
2.3. ПРИМЕНЕНИЕ ФИЛЬТРОВ	7
2.3.1. УПРАЖНЕНИЕ 6. Основные фильтры	7
2.3.2. УПРАЖНЕНИЕ 7. Расширенный фильтр I	8
2.3.3. УПРАЖНЕНИЕ 8. Расширенный фильтр II	9
2.4. КЛАСТЕРИЗАЦИЯ (РАЗДЕЛЕНИЕ НА ГРУППЫ)	9
2.4.1. УПРАЖНЕНИЕ 9. Базовая кластеризация.....	9
2.5. ЭКСПОРТ.....	12

Автор концепции и содержания упражнений: Néstor Beltrán
Учебные данные: Наталья Иванова
Перевод с английского: Максим Шашков
Редакция: Николай Груданов

Руководство подготовлено для курсов повышения квалификации BIODATA при участии Глобальной информационной системы о биоразнообразии GBIF

УСЛОВНЫЕ ОБОЗНАЧЕНИЯ

Формулы (для копирования)

Синий текст

Пример: ...затем вставьте выражение
^[a-z]

genus

^[a-z]

☒ case sensitive☒ regular expression

Команды в Refine

Красный текст

Пример: ...и далее выполняете Text facet

Facet

Text filter

Edit cells

Edit column

Transpose

Sort...

View

Reconcile

Text facet

Numeric facet

Timeline facet

Scatterplot facet

Custom text facet...

Custom numeric facet...

Customized facets

Название столбцов

Зеленый текст

Пример: ...найдите столбец Cat. Numb

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

All			Cat. Numb.	University	Collector
☆	↶	7.	UWP:157339	University of Guatemala	Betancur J
☆	↶	8.	UWP:157339	University of Guatemala	Betancur H
☆	↶	224.	UWP:122471	University of Guatemala	Vargas P
☆	↶	225.	UWP:122471	University of Guatemala	Vargas I

Гиперссылки

www.gbif.org

Меню столбца



2. ОСНОВЫ РАБОТЫ

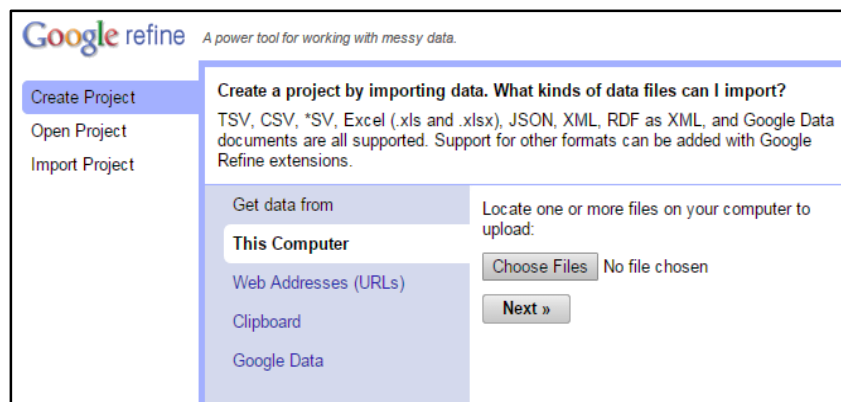
2.1. ЗАГРУЗКА ФАЙЛОВ И ПРОЕКТЫ

2.1.1. Перед началом

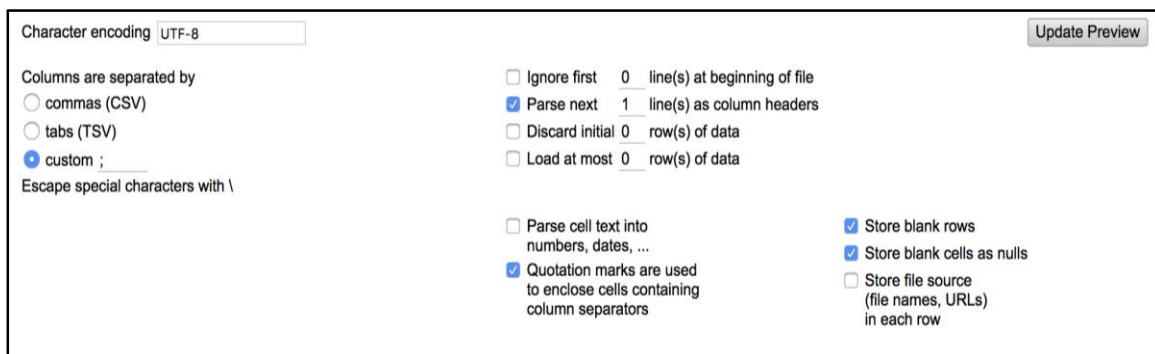
Загрузка данных может осуществляться из различных источников (*типов файлов*): TSV, CSV, SV, Excel (.xls и .xlsx), JSON, XML, RDF и данных XML в виде документов Google. Загрузка данных состоит из двух этапов: первый – это загрузка файла, а второй – создание проекта.

2.1.2. Упражнение 1. Создание проекта

1. Загрузите основной файл с данными Data_Cleaning_OpenRefine_DATA_EXAMPLE_TJ.csv.
2. Запустите *OpenRefine* (GoogleRefine), нажмите на **Create Project (Создать проект)**, и далее выполните последовательность команд **Get data from (Получить данные из...)** > **This Computer (Этот компьютер)**, затем нажмите на **Choose Files (Выбрать файл)**. Выберите файл.
3. Нажмите на **Next (далее)**.



4. Появится меню параметров разбивки строк (Parsing). Выберите параметры, как указано на картинке:




5. Сверху справа в поле название проекта (Project Name), переименуйте ваш файл как [ВашеИмя]UseCase1OpenRefine, нажмите **Create Project (Создать проект)** и Вы будете готовы к работе! *(Настоятельно рекомендуется для названий файлов использовать только латиницу. Здесь и далее курсивом комментарии переводчика)*

2.2. СВОДКА (FACETING)

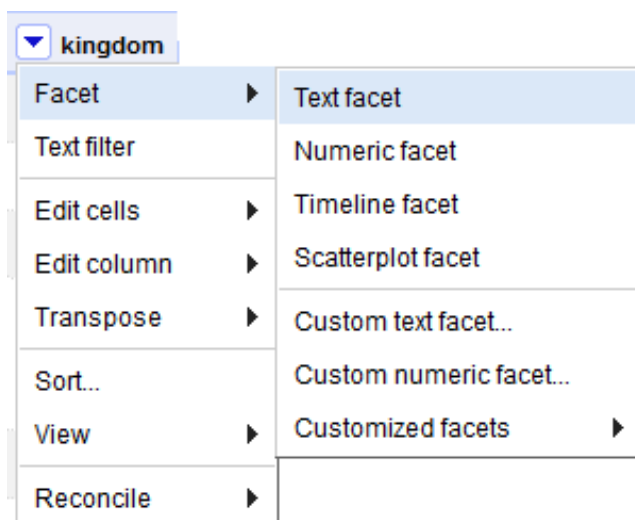
2.2.1. Перед началом

Сводка (Faceting) – это функция, позволяющая получить общую обзорную картину данных и, применив фильтр, увидеть только тот набор строк, которые необходимо просмотреть или отредактировать. Эта функция облегчает использование и анализ данных и может быть применена к ячейкам, содержащим любой текст, числа и даты.

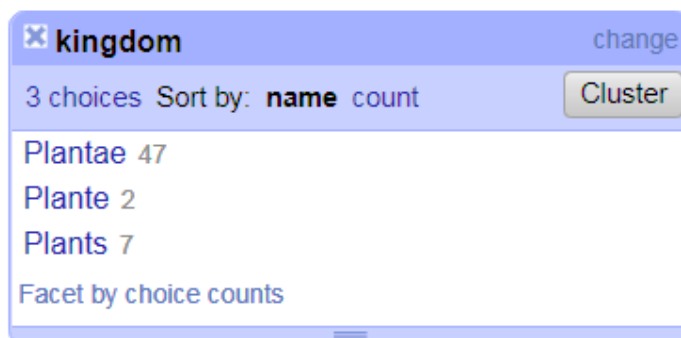
2.2.2. УПРАЖНЕНИЕ 2. Сводка и массовое редактирование

Найдите столбец **kingdom**, и нажмите на меню столбца  и далее выполните последовательность команд

Facet (Сводка) > Text facet (Текстовая сводка) как изображено на картинке:

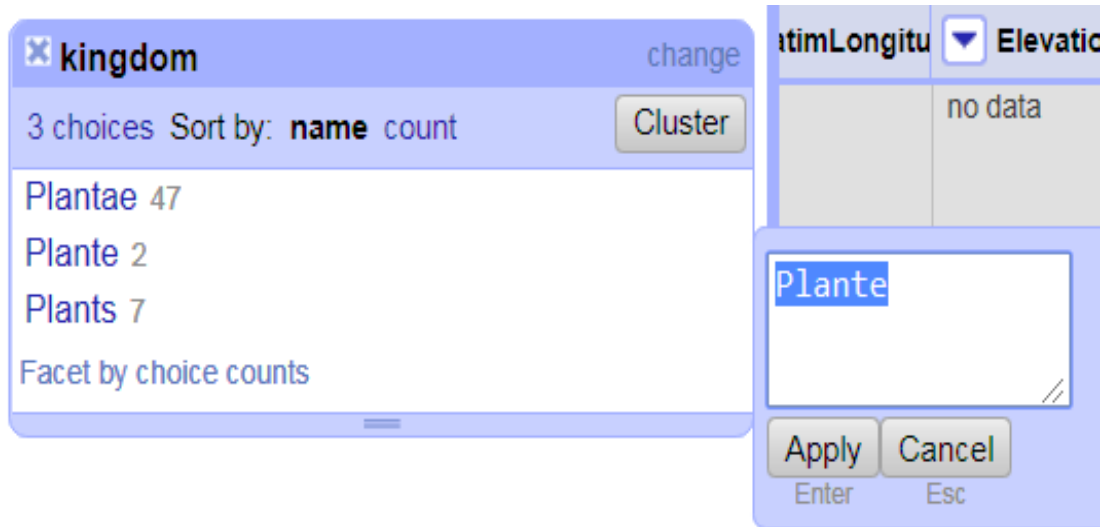


В левой части экрана появится окно с названием столбца, это собственно сводка:




Нажмите на **count** (число) для сортировки по числу записей, затем нажмите на **name** (имя) чтобы отсортировать в алфавитном порядке.

Исправьте орфографические ошибки. Наведите курсор на текст в окне и нажмите на **edit** (редактировать), затем исправьте ошибку в текстовом поле, и, чтобы сохранить исправления, нажмите на **apply** (применить).



Все значения будут исправлены автоматически.

2.2.3. УПРАЖНЕНИЕ 3. СВОДКА И ПРОБЕЛЫ I

1. Найдите столбец **Country col.**, нажмите на меню столбца  и выполните **Text Facet** (Текстовая сводка).




На первый взгляд, названия страны кажется написанным правильно, но сводка выявила три разных значения из-за лишних пробелов в конце названия.

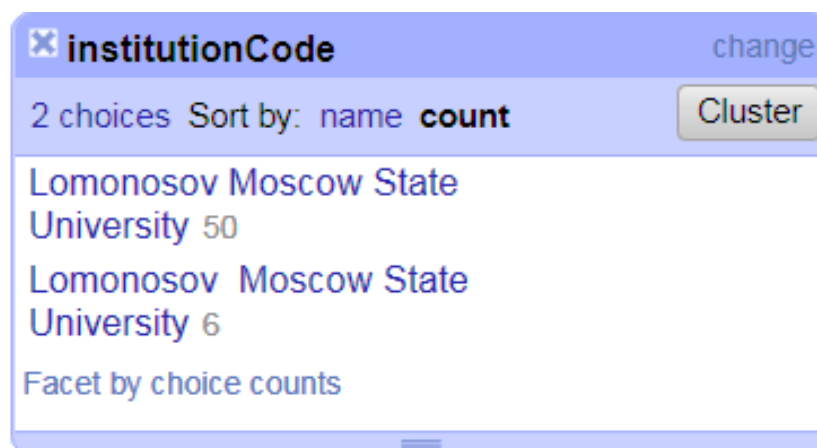
- Исправьте ошибку через меню столбца **Country col.**, выполнив команду **Edit Cells** (Редактировать ячейки) > **Common transforms** (Общие преобразования) > **Trim leading and trailing whitespace** (Удалить ведущие и конечные пробелы). Вы увидите уведомление:

Text transform on 9 cells in column Country col.: value.trim()
Undo

- Теперь проверьте окно сводки - в нём останется только одно значение.

2.2.4. УПРАЖНЕНИЕ 4. Сводка и пробелы II

- Найдите столбец **institutionCode** и нажмите на меню столбца  затем выполните **Text facet** (Текстовая сводка). Затем нажмите на **count** (число). Окно сводки покажет следующее:



Как видно на картинке, *Lomonosov Moscow State University* в списке с 50 экземплярами, но *Lomonosov Moscow State University* также присутствует на втором месте с 6 экземплярами

- Исправьте ошибку в столбце **institutionCode**, через меню столбца выполнив последовательность команд **Edit Cells** (Редактировать ячейки) > **Common transforms** (Общие преобразования) > **Collapse consecutive whitespaces** (Удалить последовательные пробелы).
- Как только лишние пробелы будут удалены, *Lomonosov Moscow State University* будет представлена в списке только один раз с 56 записями.

2.2.5. УПРАЖНЕНИЕ 5. Сводка и дубликаты

1. Для столбца **CatalogNumber**, выполните команду **Facet (сводка) > Customized facets > Duplicates facet (Сводка по дубликатам)**. Сводка покажет 4 дубликата.

2. Нажмите на **true (верно)**, и Вы увидите значения в главном окне:

The screenshot shows the 'Facet / Filter' interface. On the left, the 'catalogNumber' facet is expanded, showing 'false' with 52 counts and 'true' with 4 counts. The 'true' option is selected. On the right, a table displays 4 matching rows. The table has columns for 'All', 'catalogNumber', 'institutionCode', and 'recordedBy'.

All	catalogNumber	institutionCode	recordedBy
19.	MW0895886	Lomonosov Moscow State University	Pimenov
20.	MW0895886	Lomonosov Moscow State University	Kluikov
48.	MW0805651	Lomonosov Moscow State University	Lipschitz
49.	MW0805651	Lomonosov Moscow State University	Pavlov

После проверки по этикеткам образцов исправьте значения правильными каталожными номерами, нажав **edit (редактировать)** непосредственно в ячейке:

MW0895886 Pimenov

MW0858942 Kluikov

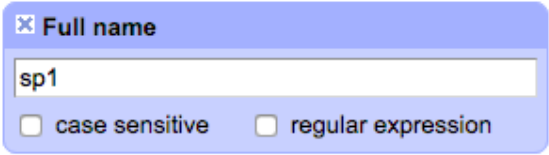

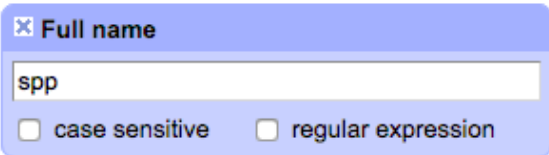

MW0805651 Lipschitz

MW0807201 Pavlov

2.3. ПРИМЕНЕНИЕ ФИЛЬТРОВ

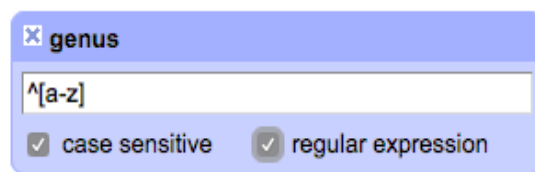
2.3.1. УПРАЖНЕНИЕ 6. Основные фильтры

1. Перейдите снова в меню столбца **Full name** и выполните **Text facet (текстовая сводка)** чтобы показать значения, в меню столбца **Full name** нажмите на **Text filter (текстовый фильтр)**, примените следующие фильтры исправьте значения как показано ниже:

Фильтр	Как исправить	Правильное значение
	Редактировать (Edit) прямо в ячейке	Phleum
	Редактировать (Edit) прямо в ячейке, поставив флажок С учетом регистра (Case sensitive)	Juniperus
	1. В меню  столбца Full name , выполните Edit cells (Редактировать ячейки) > Transform... (Преобразование) 2. В текстовом поле вставьте формулу <code>value.replace(" spp", "")</code> 3. нажмите OK	Anemone Poa Calamagrostis

2.3.2. УПРАЖНЕНИЕ 7. Расширенный фильтр I

1. Найдите столбец **genus** и вызовите **Text filter (Текстовый фильтр)**.
2. Поставьте флажки на пунктах **regular expression (регулярное выражение)** и **case sensitive (с учетом регистра)**, затем вставьте в поле фильтра выражение `^[a-z]`



Это регулярное выражение выберет строки, в которых первая буква в нижнем регистре (строчная).

3. Название рода должно начинаться с заглавной буквы, внесите соответствующие исправления.

Примечание: Если Вы хотите узнать больше о регулярных выражениях, перейдите по [ссылке](#).

2.3.3. УПРАЖНЕНИЕ 8. Расширенный фильтр II

1. Найдите столбец **Full name** и вызовите **Text filter** (Текстовый фильтр).
2. Установите флажки на пунктах **regular expression** (регулярное выражение) и **case sensitive** (с учётом регистра), затем вставьте в поле фильтра выражение `^[A-Z].*\s[A-Z]`



Это регулярное выражение выберет строки, которые начинаются с заглавной буквы, за которой следуют любые символы, затем пробел, затем снова заглавная буква.

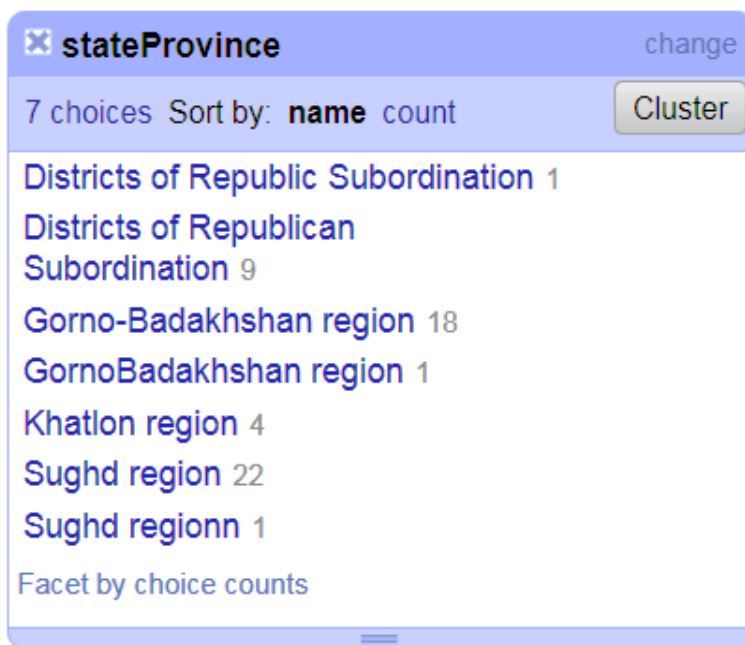
3. Видовой эпитет должен начинаться со строчной буквы, внесите соответствующие исправления.

Примечание: Если Вы хотите узнать больше о регулярных выражениях, перейдите по [ссылке](#).

2.4. КЛАСТЕРИЗАЦИЯ (РАЗДЕЛЕНИЕ НА ГРУППЫ)

2.4.1. УПРАЖНЕНИЕ 9. Базовая кластеризация

1. В меню столбца **stateProvince** выполните команду **Text facet** (текстовая сводка).



Имейте в виду, что правильные названия округов следующие:
Sughd region, Khatlon region, Gorno-Badakhshan region и Districts of Republican Subordination

2. Сверху справа окна сводки нажмите на **Cluster (Кластеризовать)**, появится новое окно:

Cluster & Edit column "stateProvince"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method **key collision** Keying Function **fingerprint** 1 cluster found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	18	<ul style="list-style-type: none">Gorno-Badakhshan region (17 rows)GomoBadakhshan region (1 rows)	<input type="checkbox"/>	<input type="text" value="Gorno-Badakhshan region"/>

Select All Unselect All

Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

3. Теперь Вы можете увидеть информацию о кластерах:

- **Cluster size (Размер кластера)**: число вариантов, которые алгоритм кластеризации определил одинаковыми.
- **Row count (Число строк)**: число записей (строк) для каждого значения в кластере.
- **Values in cluster (Значения в кластере)**: фактические значения, которые алгоритм определяет как одинаковые. Здесь же отображается число записей для каждого значения, а также есть возможность просматривать содержимое кластера в отдельных вкладках браузера (*ссылка **Browse this cluster** снизу справа под значениями*).
- **Merge? (Объединить?)**: поставьте флажок, чтобы объединить значения в одно.
- **New cell value (Новое значение ячейки)**: значение, которое будет задано для каждой записи в кластере. По умолчанию это значение большинства записей. Вы также можете выбрать любое значение, чтобы сделать его **Новым значением ячейки**.

Примечание: Если Вы хотите узнать больше о кластеризации, перейдите по [ссылке](#).

4. Нажмите на **Select All (Выделить всё)** и затем на **Merge Selected & close (Объединить выбранное и закрыть)**, Вы увидите уведомление:

Mass edit 19 cells in column stateProvince Undo

5. Для того, чтобы исправить оставшиеся названия округов, снова выполните команду **Cluster (Кластеризовать)** в окне сводки для столбца **stateProvince**.
6. В окне кластеризации и редактирования в поле **Keying Function (Ключевая функция)**, выберите значение **ngram-fingerprint**, затем установите значение 1 в поле Ngram Size. Нажмите клавишу ввода.
7. Нажмите на **Select All (Выделить всё)** и затем на **Merge Selected & close**, Вы увидите уведомление:

Mass edit 33 cells in column stateProvince Undo

8. Теперь все названия исправлены и окно сводки должно выглядеть так, как указано на картинке ниже:



2.5. ЭКСПОРТ

Исправленные данные можно экспортировать несколькими способами, но в большинстве случаев будет полезен следующий:

В правом верхнем углу нажмите на Export (Экспортировать) и выберите Custom tabular exporter... (Экспорт таблицы с пользовательскими настройками).

Вы увидите окно экспорта:

The screenshot shows the 'Custom Tabular Exporter' dialog box. It has four tabs: 'Content', 'Download', 'Upload', and 'Option Code'. The 'Content' tab is active. Inside, there are two main sections. The left section, 'Select and Order Columns to Export', contains a list of columns with checkboxes: 'catalogNumber' (checked), 'institutionCode', 'Collector', 'identifiedBy', 'individualCount', 'DA', 'MO', 'YE', and 'Country col.'. Below this list are 'Select All' and 'De-select All' buttons. The right section, 'Options for catalogNumber', contains settings for 'For reconciled cells, output' with radio buttons for 'Matched entity's name', 'Matched entity's ID', and 'Cell's content'. There are also checkboxes for 'Link to matched entity's page' and 'Output nothing for unmatched cells'. Below these are date and time format options: 'ISO 8601, e.g., 2011-08-24T18:36:10+08:00', 'Short locale format', 'Long locale format', 'Medium locale format', 'Full locale format', and 'Custom'. There are also checkboxes for 'Use local time zone' and 'Omit time'. At the bottom of the dialog, there are three checkboxes: 'Output column headers' (checked), 'Output empty rows (ie all cells null)' (checked), and 'Ignore facets and filters and export all rows' (unchecked). A 'Cancel' button is at the bottom left.

1. На вкладке **content (содержание)** Вы можете выбрать столбцы, которые хотите экспортировать, если Вы выберите **Ignore facets and filters and export all rows (игнорировать сводки и фильтры и экспортировать все строки)**, все сводки и фильтры будут проигнорированы, это будет полезно, если Вы забыли отменить их перед экспортом.
2. Перейдите на вкладку **Download** и выберите разделитель, который Вы предпочитаете. Не изменяйте другие параметры без необходимости.

Вы также можете экспортировать проект целиком, чтобы открыть его в OpenRefine на другом компьютере через последовательность команд **Export (Экспортировать) > Export project (Экспортировать проект)**. В данном случае Вы получите файл не для работы с ним в редакторе электронных таблиц или текстовом, а файл формата GZIP, который будет доступен только через OpenRefine.