Преобразование бюджета в формат OpenBudgetFormat 1.0 с помощью OpenRefine

Спасибо, что проявляете интерес к нашему проекту LenObl2014.ru – Открытый бюджет Ленинградской области 2014.

В рамках работы над проектом Открытого бюджета Ленинградской области 2014 г. нами был разработан формат OpenBudgetFormat 1.0 (OBF) представления бюджета, который соответствует международным стандартам и является более подходящим для загрузки на платформу Open Spending, в отличие от стандартного формата, который предлагается государственными финансовыми органами России. Более подробную информацию об этом формате вы можете найти на сайте: http://lenobl2014.ru/format. Далее мы расскажем, как вы можете самостоятельно преобразовать полученный от органов власти или с их официального сайта документ бюджета в формат OBF.

1. Требования к исходному файлу.

Пример исходного файла бюджета вы можете скачать тут. В данном руководстве рассмотрено преобразование стандартного файла программного бюджета в формате xls из приложений закона о бюджете. Итоговый файл можно скачать тут.

2.Пошаговый алгоритм преобразования файла бюджета

Для преобразования данных необходимо установить OpenRefine. Скачать его и прочитать все необходимые инструкции можно по ссылке: http://openrefine.org.

После этого необходимо запустить OpenRefine, открыть браузер и перейти по ссылке http://localhost:3333/. Открыть вкладку Create project и загрузить файл с необходимыми данными – бюджетом региона (рис. 1). Предпочитаемый формат – xls.

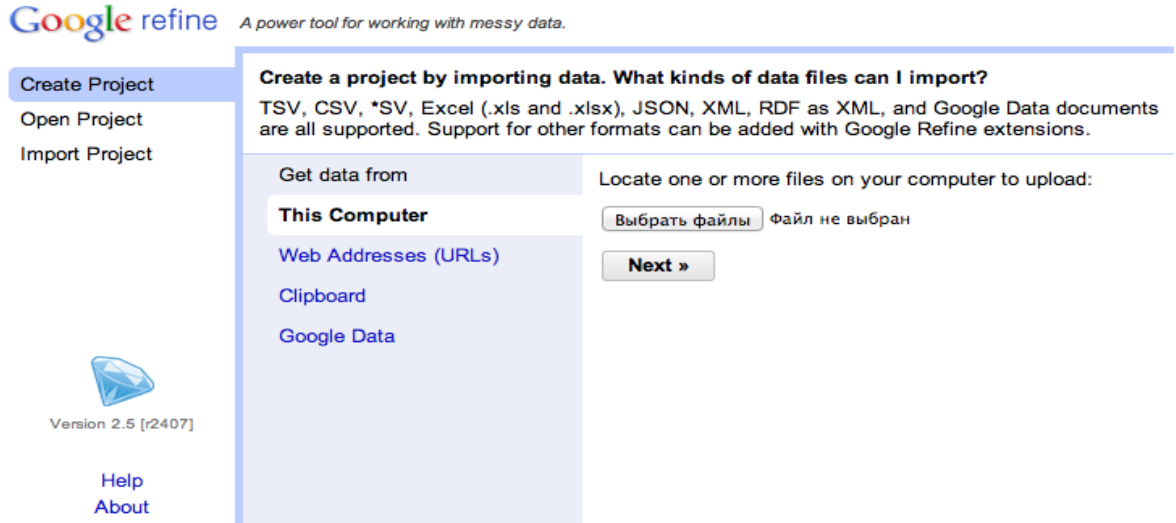


Рисунок 1. Окно создания проекта

После загрузки массива данных вам будет предложены некоторые настройки импорта файла (рис. 2).

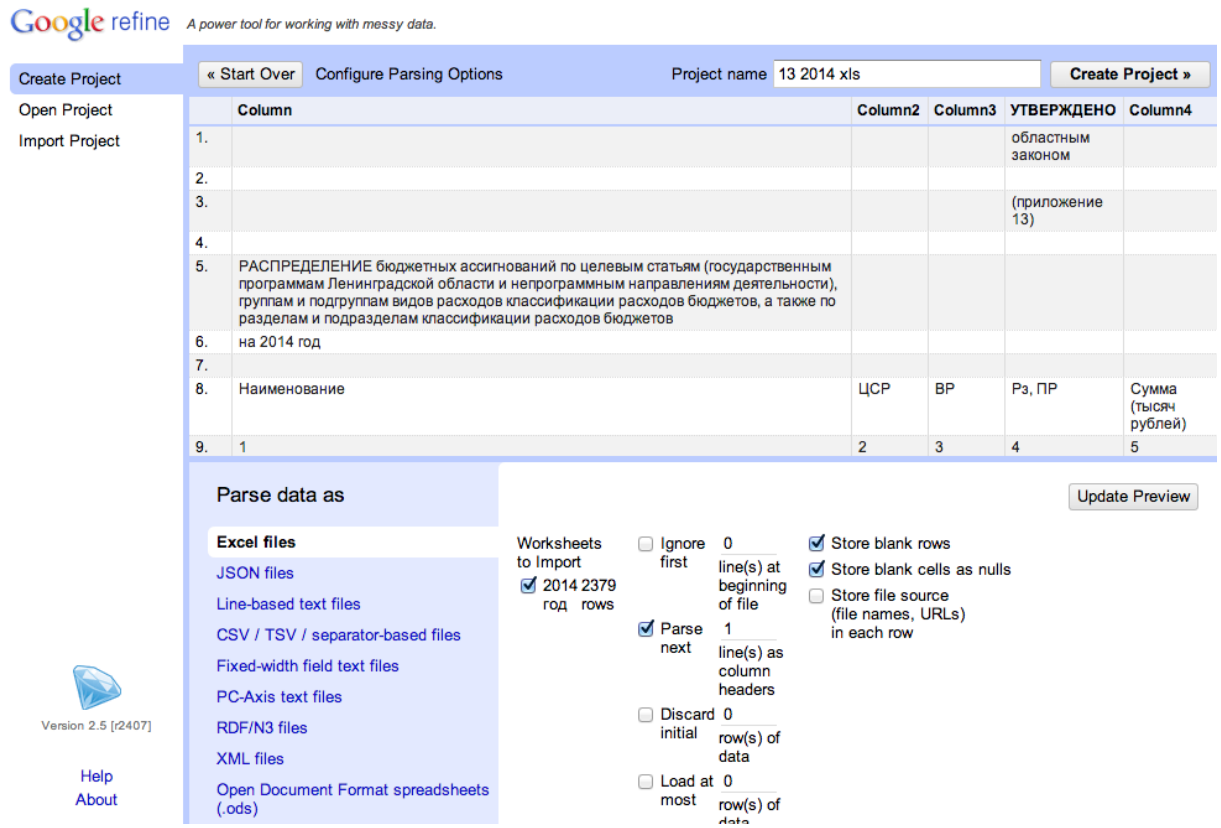


Рисунок 2. Настройка импорта файла

В данном окне вы можете указать следующее:

- ввести название проекта (Project name);
- в левой части экрана (синее меню) выбрать тип загружаемого файла (в примере загружается файл формата Excel);
- указать, с какой строки необходимо обрабатывать файл с помощью поля «ignore first … lines» (в примере данные необходимо обрабатывать с строки 8, с которой начинается таблица расходов);
- отметить строку, в которой содержатся названия столбцов с помощью поля «parse next … line(s) as column headers» (в примере на рис. 2 должна быть выделена первая строка).

После обработки настроек вы попадете в главное окно программы, в котором происходит обработка данных (рис. 3). Далее описан пошаговый алгоритм преобразования файла с объяснением каждой вводимой команды.

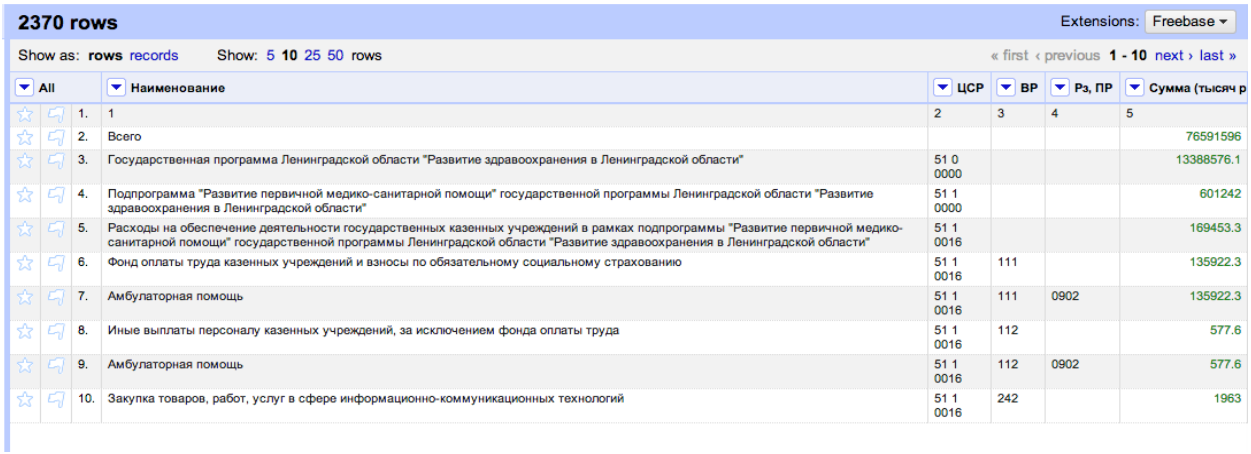


Рисунок 3. Главное окно программы OpenRefine

1. Удаление лишних строк. Первые две строки не содержат полезной информации, поэтому их можно удалить. Для удаления ненужных строк необходимо нажать на «звездочку» или «флажок» – рис. 4.1., сделать фасет по всем строкам, отмеченным «звездочкой» – рис. 4.2. (выбрать All – Facet – Facet by star), выбрать в фасете «Starred rows» все строки, соответствующие этому условию (true) – рис. 4.3. и удалить их, выбрав меню All – Edit rows – Remove all matching rows.

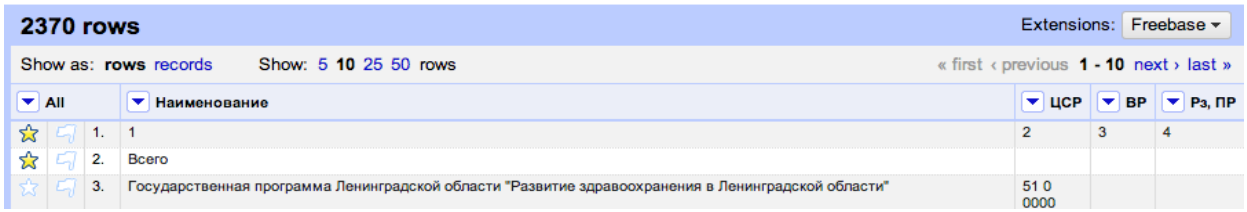


Рисунок 4.1.

Рисунок 4.2.

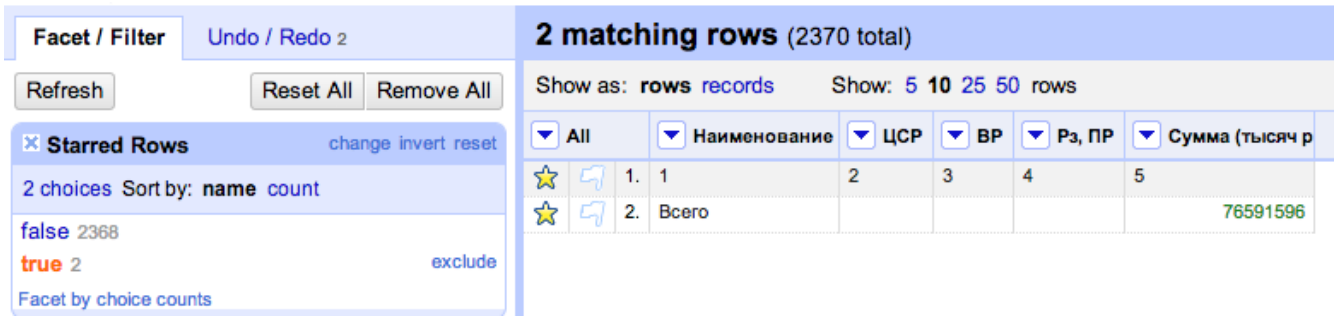


Рисунок 4.3.

Рисунок 4.4.

2. Добавление отдельных столбцов с названием и кодом программ. Кодом программы являются первые две цифры из столбца «ЦСР» с пятью нулями. Поэтому для добавления графы с кодом программы необходимо скопировать первые две цифры столбца «ЦСР» и добавить к ним выражение: « 0 0000». Для этого необходимо в меню столбца «ЦСР» (кнопка с треугольником слева от названия столбца) выбрать пункт редактирование столбца (Edit column) и в раскрывшемся подменю выбрать пункт создания нового столбца , основанного на данном столбце (Add column, based on this column…). В результате этого откроется окно, предназначенное для ввода команд (рис. 5). В нем необходимо указать название нового столбца (New column name) и ввести команду в поле Expression. В нижней части окна во вкладке Preview можно увидеть результат выполнения команды, просмотреть историю ввода команд (вкладка History), просмотреть избранные команды (вкладка Starred) или прочитать справку (вкладка Help).

Для добавления столбца с кодом программы необходимо ввести следующую команду: **substring (value, 0,2) + " 0 0000"**. Данная команда позволяет взять первые две цифры (0,2) из выбранного столбца (value) и добавить к ним выражение « 0 0000». Команды выполняются для каждой строки отдельно, то есть для первой строки нового столбца будет взято значение первой строки столбца value, для второй строки – второй и т.д. **Пример:** если в исходном столбце (value) значение первой строки «51 1 0016», то значением первой строки нового столбца будет выражение «51 0 0000», если значение второй строки исходного столбца «53 2 0110», то значением второй строки нового столбца будет выражение «53 0 0000».

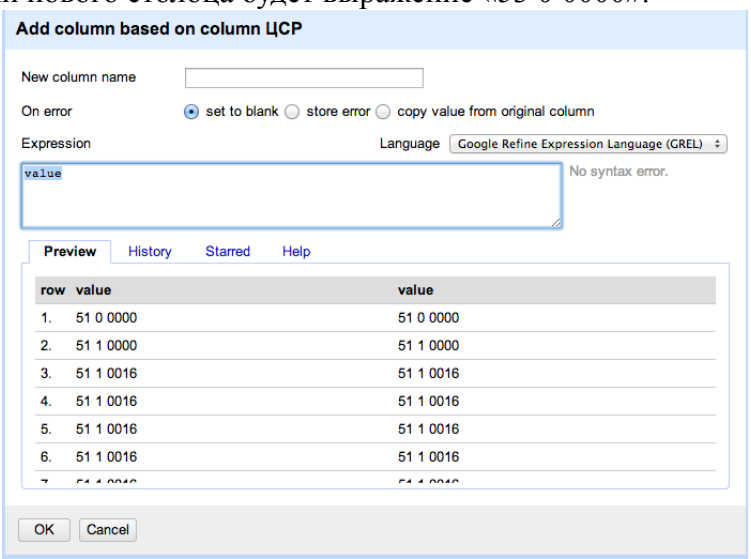


Рисунок 5.

Для добавления наименований программ есть два способа: добавление их вручную, если значений не очень много, или добавление с помощью ввода соответствующей команды. Рассмотрим первый способ. OpenRefine позволяет группировать ячейки в столбце по их содержимому. Этот процесс называется фасетом (facet). Для этого необходимо в меню столбца выбрать пункт текстовый фасет: Facet – Text facet. Тогда в левой части экрана появится окно фасета со списком всех встречающихся значений ячеек в данном столбце и их количестве (рис. 6).

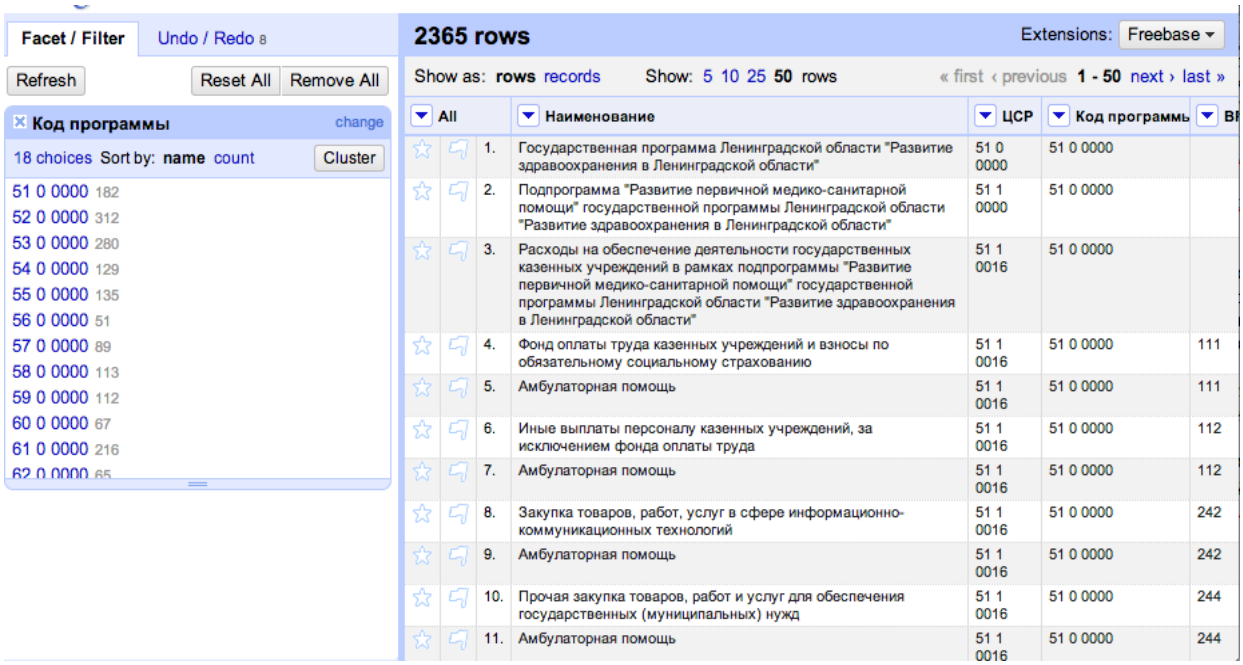


Рисунок 6.

Выбор группы ячеек осуществляется нажатием на их значение, при этом выбранные группы подсвечиваются оранжевым цветом, а в верхней части экрана над столбцами и их названиями отображается количество выделенных строк и их общее количество (например, 182 matching rows (2635 total). Также при наведении на строку с группой ячеек появляются кнопки edit (для редактирования значений ВСЕХ ячеек этой группы) и include/exclude, позволяющие выбирать или отменить выбор данной группы (рис. 7). Например, на рис. 7 выбрана группа ячеек со значением «51 0 0000» и показано меню для группы «52 0 0000». Если мы выберем edit в этом меню и введем в появившемся окне значение «52 1 0000», то оно изменится во всех 312 строках.

Одновременно можно работать с несколькими фасетами, созданными для разных столбцов. При работе с фасетами необходимо учитывать следующее. Если ни одна группа ни в одном фасете не выбрана, то показываются фасеты по всем ячейкам таблицы. Если в одном фасете выбрана одна или несколько групп, то показываются все значения ячеек только в этой группе. Например, есть таблица из ста строк и двух столбцов: столбец «число» с цифрами от 0 до 99 и столбец «количество десятков» с цифрами от 0 до 9. Если создать фасеты по этим столбцам и не выбирать ни одной группы, то в фасете по столбцу «число» будет 100 групп (со значениями от 0 до 99), а в фасете по столбцу «десятки» будет 10 групп (со значениями от 0 до 9). Если выбрать одну из групп в фасете по столбцу «десятки», например «2», то в окне с фасетами по столбцу «число» останется только 10 значений (от 20 до 29).

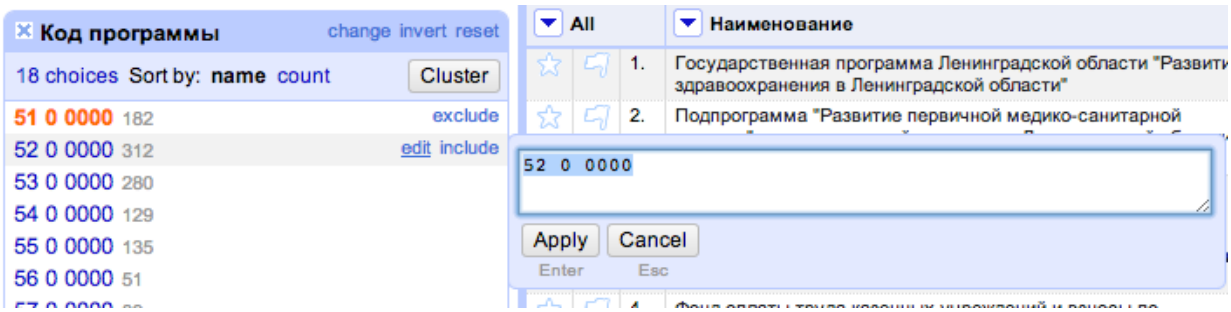


Рисунок. 7

Для добавления наименования программ необходимо сначала добавить новый столбец «Наименование программы». Он может быть создан с пустыми ячейками. Для этого необходимо в меню любого столбца выбрать пункт «Add column based on this column» и в поле Expression ввести кавычки: **“”**. После этого необходимо создать еще один текстовый фасет, но уже по новому столбцу «Наименование программы». Пример результата представлен на рис. 8.

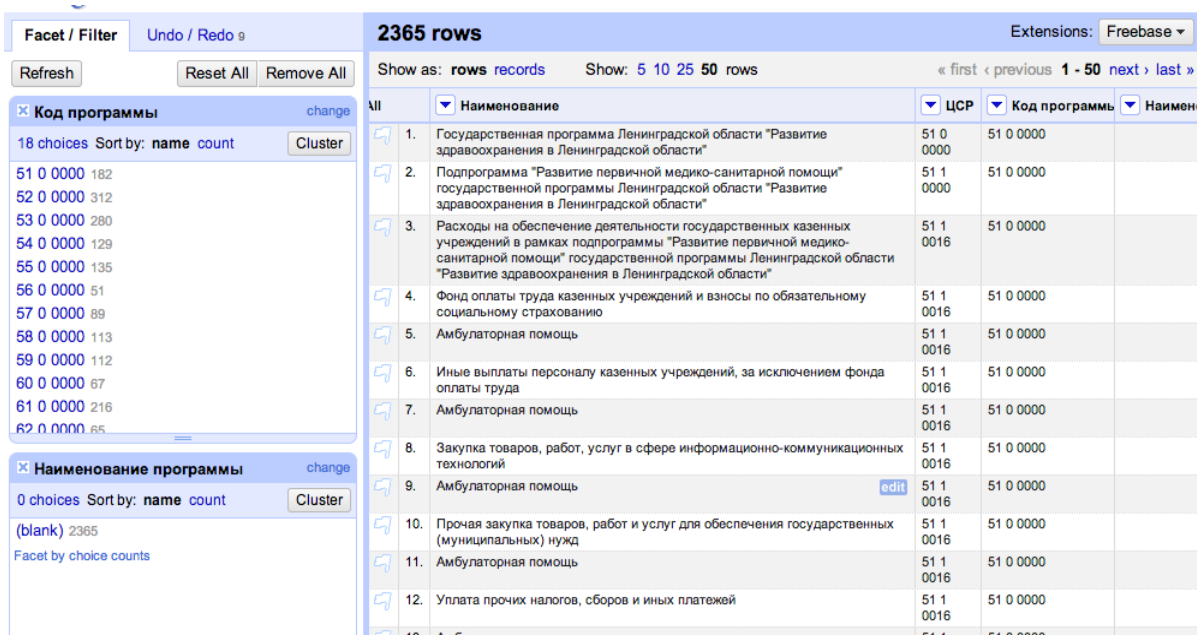


Рисунок 8.

Обычно в файле бюджета наименования программ указаны в столбце «Наименование» в тех ячейках, в которых значение ячейки столбца «ЦСР» совпадает с кодом программы. Поэтому для добавления наименования программ необходимо выполнить следующие команды:

a. Выбрать группу значений в фасете «Код программы» (например, 51 0 000)

b. Найти строку, в которой значение ячейки столбца «ЦСР» совпадает с данной группой значений, и скопировать ее значение

c. Выбрать группу значений в фасете «Наименование программы» (в нем должна быть всего одна группа с пустыми ячейками – blank, так как мы создавали столбец с незаполненными ячейками.

d. Выбрать edit, в появившееся окно «вставить» скопированное значение (рис. 9), и нажать Apply (подтвердить изменение) или Cancel (если необходимо отменить изменения ячеек).

e. После этого необходимо выбрать другую группу в фасете «Код программы» и повторить все предыдущие действия.

f. Чтобы проверить, что внесены наименования программ для всех ячеек, можно создать фасет по столбцу «Наименование программ» и, просмотрев появившиеся группы значений, убедиться, что в них нет группы “blank”.

g. В бюджете Ленинградской области наименования всех программ начинаются с выражения: «Государственная программа Ленинградской области» для упрощения и сокращения наименований это выражение можно заменить на слово «Программа». Для этого необходимо в меню столбца «Наименование программы» выбрать Edit cells – Transform… и в поле Expression ввести команду: **"Программа" + substring (value, 47)**.

h. Строки, в которых указаны наименования программ в столбце «Наименование» можно удалить (они не содержат дополнительной информации, см. страницу на сайте «Формат публикации бюджетов на примере Бюджета Ленинградской области»). Для этого необходимо в меню столбца «ЦСР» выбрать пункт Text filter (фильтры позволяют выбирать ячейки, в которых встречается данное значение.) и в появившемся слева окне ввести значение « 0 0000». В результате выполнения данного действия должно остаться количество строк, равное количеству программ в бюджете (для Ленинградской области это число – 18). Эти строки можно удалить, выбрав в меню All пункт Edit rows – Remove all matching rows.

Рисунок. 9

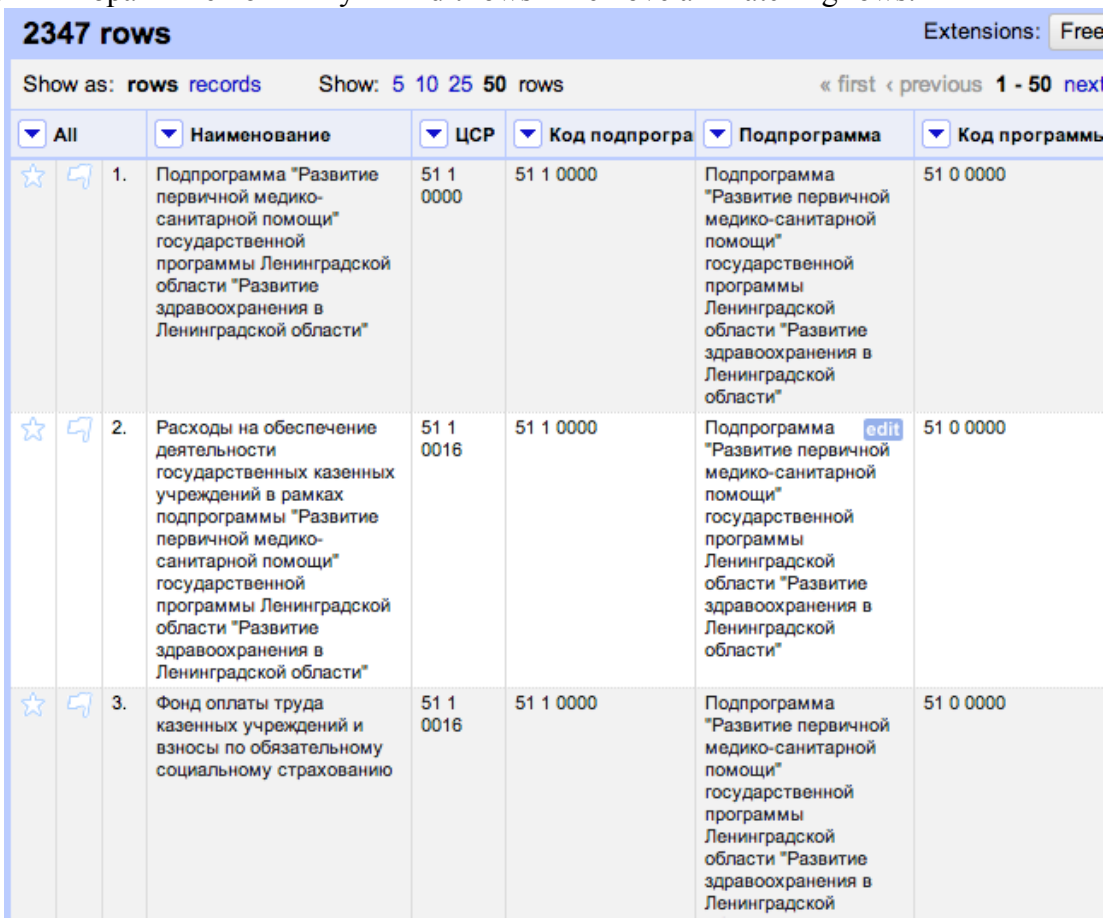
3. Следующим шагом можно добавить наименования и коды подпрограмм. Коды подпрограмм добавляются также как и коды программ, за исключением вводимой команды в поле Expression: **substring (value, 0, 4) + " 0000"**. Данная команда позволяет взять первые четыре символа исходной ячейки (столбец «ЦСР») и добавить к ним выражение “ 0000” (название данного столбца: “Код подпрограммы”).

Наименования подпрограмм можно добавить автоматическим способом, создав новый столбец на основе столбца «Код подпрограммы» и введя команду в поле Expression:

cell.cross("Budget LO", "ЦСР")[0].cells["Наименование"].value.

Данная команда позволяет в проекте “Budget LO” (название проекта указано рядом с логотипом Google Refine) ячейке строки N нового столбца присвоить значение той ячейки из столбца “Наименование”, значение которой равно значению ячейки столбца «Код подпрограммы» в строке N. Например, в результате выполнения данной команды для трех строк, в которых значение “Код подпрограммы” равно “51 1 0000” в столбце “Подпрограмма” присвоено значение ячейки из первой строки столбца “Наименование” (значение ячейки столбца “ЦСР” в первой строке = значениям ячеек столбца “Код подпрограммы” первых трех строк), рис. 10.

Удалить ненужные строки (те, в которых в столбце “Наименование” указаны наименования подпрограмм) можно создав текстовый фильтр по столбцу “ЦСР” со значением “0000” и выбрав в меню All пункт Edit rows – Remove all matching rows.



2347 rows Extensions: Free

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

All		Наименование	ЦСР	Код подпрогра	Подпрограмма	Код программы
1.		Подпрограмма "Развитие первичной медико-санитарной помощи" государственной программы Ленинградской области "Развитие здравоохранения в Ленинградской области"	51 1 0000	51 1 0000	Подпрограмма "Развитие первичной медико-санитарной помощи" государственной программы Ленинградской области "Развитие здравоохранения в Ленинградской области"	51 0 0000
2.		Расходы на обеспечение деятельности государственных казенных учреждений в рамках подпрограммы "Развитие первичной медико-санитарной помощи" государственной программы Ленинградской области "Развитие здравоохранения в Ленинградской области"	51 1 0016	51 1 0000	Подпрограмма edit "Развитие первичной медико-санитарной помощи" государственной программы Ленинградской области "Развитие здравоохранения в Ленинградской области"	51 0 0000
3.		Фонд оплаты труда казенных учреждений и взносы по обязательному социальному страхованию	51 1 0016	51 1 0000	Подпрограмма "Развитие первичной медико-санитарной помощи" государственной программы Ленинградской области "Развитие здравоохранения в Ленинградской	51 0 0000

Рисунок 10.

4. Добавление кода и наименования статей расходов. Для добавления кода статей расходов необходимо создать новый столбец на основе столбца “ЦСР”, в появившемся окне ввести команду: **substring (value, 5)** и название столбца “Код СР”. Результатом выполнения команды будут четырехзначеные значения статей расходов.

Добавление наименования статей расходов происходит в 2 действия. Сначала необходимо добавить наименования для тех строк, в которых значения ячеек в столбце «Вид расходов» (ВР) пустые. Для этого необходимо сделать текстовый фасет по столбцу «Вид расходов» и выбрать группу значений «(blank)». Для всех этих строк необходимо добавить новый столбец, основанный на столбце «Наименование» с теми же значениями (для этого в появившемся окне в поле Expression необходимо ввести команду: **value**, и название столбца «Наименование СР»). Также для этих строк необходимо создать столбец с кодами статей расходов. Для этого надо создать столбец на основе столбца «Код СР», аналогично ввести команду **value** в поле

Expression и название «ЦСР2». После этого надо закрыть окно с фасетом, чтобы были выбраны все строки. И создать новый столбец на основе столбца «Код СР» и в появившемся окне в поле Expression ввести команду: **cell.cross("Budget LO", "ЦСР2")[0].cells["Наименование СР"].value**.

Следующим шагом необходимо удалить все вспомогательные столбцы («Наименование СР» и «ЦСР2»). Для этого нужно в меню столбца выбрать пункт Edit column – Remove column. И удалить ненужные строки, в которых значения ячеек столбца «Вид расходов» (ВР) являются пустыми. Для этого с помощью фасета на столбце «Вид расходов» надо выбрать все строки с пустыми ячейками и удалить их в меню All, выбрав пункт Edit rows – Remove all matching rows.

5. Добавление наименований Вида расходов. Наименования видов расходов можно добавить вручную (их должно быть не очень много) или как в предыдущем пункте с помощью создания двух вспомогательных столбцов. Для этого надо выбрать строки, в которых пустыми являются ячейки столбца «Раздел и подраздел» (Рз, ПР). Добавить столбец («Код ВР») с теми же значениями, что и значения столбца «Вид расходов», и столбец («ВР2») с теми же значениями, что и значения столбца «Наименование». Закрыв фасет и выбрав все строки, выполнение команды добавит наименования видов расходов для всех строк. Вспомогательные столбцы «Код ВР» и «ВР2», а также строки, в которых значения столбца «Рз, ПР» являются пустыми, можно удалить.

6. Следующим шагом является изменение сумм расходов. Их нужно преобразовать из тысяч рублей в рубли. Для этого нужно создать новый столбец на основе столбца с суммами расходов и в появившемся окне в поле Expression ввести команду: **if (contains(value, "."), value.replace(".", "") + "00", value + "000")**. Данная команда выполняет следующее: если в значении ячейки содержится «.», то она удаляется и добавляются два нуля, если не содержится – добавляются три нуля. Например, если сумма расходов составляет «60.1» тыс. руб., то в результате выполнения команды будет получена сумма «60100» рублей, если сумма расходов составляет «50» тыс. руб., то результатам выполнения команды будет значений «50000» рублей.

Перед удалением исходного столбца с суммами расходов необходимо проверить для нескольких строк (для целых значений и для дробных), что суммы преобразованы правильно.

Данный пункт может незначительно отличаться в зависимости от формата, в котором указаны расходы. Возможно, в качестве разделителя в файле будет использоваться не точка, а запятая, в этом случае в команде необходимо изменить **"."** на **","**. Если в нем использованы и точки и запятые, тогда перед выполнением указанной команды вам нужно выполнить команду, заменяющую все точки на запятые: **value.replace(“.”,”,”)**, или наоборот, команду, заменяющую все запятые на точки: **value.replace(“,”,”.”)**. Еще одним вариантом формата представления сумм расходов может быть наличие пробелов, удалить которые можно командой: **value.replace(“ ”,””)**.

7. Добавление даты расходов. В файлах бюджетов может быть указан только год (так как фактическая дата расходов неизвестна). Для этого необходимо создать новый столбец на основе любого другого с помощью пункта Edit column – Add column based on this column и в появившемся окне в поле Expression ввести команду: **“2014 год”**. Все остальные столбцы (контакты и пр.) создаются аналогиным образом.

Заключение

Полученный формат представления данных о расходах удобнее как для визуализации, так и для изучения экспертами. Вас не должно смущать, что в итоговом файле получилось значительно меньше строк, чем в исходном – никакая часть информации не потеряна, а значения сумм по отдельным группам расходов будут автоматически сформированы сервисом OpenSpending.

Автор руководства

Ольга Пархимович, независимый эксперт в области открытых данных
olya.parkhimovich@gmail.com

Редактор

Виталий Власов, руководитель Фонда “Открытый город”
inxaoc@gmail.com

Благодарности

Авторы хотели бы выразить благодарности команде Комитета Финансов Ленинградской области за содействие в реализации данного проекта, а именно:

Андрей Сытник, Начальник Департамента "Открытого бюджета" Комитета финансов Ленинградской области

Александр Зарецкий и Олег Мишкорудный, Департамент "Открытого бюджета" Комитета финансов Ленинградской области