**Преобразование бюджета с помощью OpenRefine**

Бюджеты российских регионов публикуются по-разному, причем обычно в формате, который не удобен как для изучения гражданами, так и для машинной обработки. Поэтому для работы с бюджетами необходимо предварительно преобразовывать формат файлов и использовать какие-либо сервисы для их визуализации и изучения данных.

В данном материале рассмотрены три вопроса:

1. Формат публикации бюджетов (на примере одного из регионов)
2. Формат данных, загружаемых в проект OpenSpending
3. Пошаговый алгоритм преобразования файла бюджета

**Формат публикации бюджетов на примере Бюджета Ленинградской области**

Бюджет состоит из закона и нескольких приложений. Наиболее интересными для изучения из них являются программная и ведомственная структуры. В табл. 1 представлен пример данных Программной структуры бюджета ЛО на 2014 год. Ленинградская область предоставляет свои данные в формате XLS, поэтому их в дальнейшем можно будет использовать практически без предварительной обработки. Если интересующие вас данные опубликованы, например, в формате DOC, то вам необходимо предварительно скопировать таблицу с расходами в новый файл, созданный в программе Excel.

Таблица 1. Пример данных Программной структуры бюджета ЛО на 2014 год

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| **РАСПРЕДЕЛЕНИЕ бюджетных ассигнований по целевым статьям (государственным программам Ленинградской области и непрограммным направлениям деятельности), группам и подгруппам видов расходов классификации расходов бюджетов, а также по разделам и подразделам классификации расходов бюджетов** | | | | |
| **на 2014 год** | | | | |
|  |  |  |  |  |
| **Наименование** | **ЦСР** | **ВР** | **Рз, ПР** | **Сумма (тысяч рублей)** |
| 1 | 2 | 3 | 4 | 5 |
| **Всего** |  |  |  | **76 591 596,0** |
| **Государственная программа Ленинградской области "Развитие здравоохранения в Ленинградской области"** | **51 0 0000** |  |  | **13 388 576,1** |
| **Подпрограмма "Развитие первичной медико-санитарной помощи" государственной программы Ленинградской области "Развитие здравоохранения в Ленинградской области"** | **51 1 0000** |  |  | **601 242,0** |
| 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 |  |  | 169 453,3 |
| Фонд оплаты труда казенных учреждений и взносы по обязательному социальному страхованию | 51 1 0016 | 111 |  | 135 922,3 |
| Амбулаторная помощь | 51 1 0016 | 111 | 0902 | 135 922,3 |

Структура данных является следующей. Во 2, 3, 4 столбцах указываются коды классификаторов расходов, необходимые для идентификации каждой строки бюджета. В столбце 5 указана сумма расходов, запланированная по данной строке или группе расходов. Обычно суммы указываются в тысячах рублей, но это не является обязательным. Столбец 1 содержит наименования для всех классификаторов.

Например, в строке 2 (рис. 1) указано только значение 2ого столбца ЦСР (целевая статья расходов) – «51 0 0000». Наименование в данном случае относится к столбцу ЦСР и идентифицирует программу (формат данных для программ: XX 0 0000). Цифра, указанная в столбце «сумма» в данном случае обозначает сумму средств, которые будут потрачены на все расходы в рамках этой программы.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Наименование | ЦСР | ВР | Рз, ПР | Сумма (тысяч рублей) |
| Государственная программа Ленинградской области "Развитие здравоохранения в Ленинградской области" | 51 0 0000 |  |  | 13 388 576,1 |

Рисунок 1. Пример наименования программ

Следующая строка в программной структуре (рис. 2, строка 1) имеет аналогичную структуру, но содержит значение «51 1 0000» и описывает подпрограмму (формат данных подпрограмм: ХХ Х 0000), поэтому цифра, указанная в столбце «сумма» обозначает все расходы в рамках подпрограмм. На рис. 2, строке 2, приведен пример описания расходов в рамках статьи расходов (формат данных статьи расходов: ХХ Х ХХХХ) и указана сумма, которая будет потрачена в рамках данной статьи расходов по всем видам разделов (ВР) и всем разделам и подразделам (Рз, ПР).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Наименование | ЦСР | ВР | Рз, ПР | Сумма (тысяч рублей) |
| Подпрограмма "Развитие первичной медико-санитарной помощи" государственной программы Ленинградской области "Развитие здравоохранения в Ленинградской области" | 51 1 0000 |  |  | 601 242,0 |
| 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 |  |  | 169 453,3 |

Рисунок 2. Пример наименования подпрограмм

На рис. 3 представлен пример записи наименований для видов расходов (строка 1) и разделов и подразделов (строка 2). В строке 1 сумма указывает количество средств, которые будут потрачены на все разделы и подразделы в рамках данного вида расходов, а в строке 2 указана сумма, которая будет потрачена в рамках данного раздела и подраздела.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Наименование | ЦСР | ВР | Рз, ПР | Сумма (тысяч рублей) |
| Фонд оплаты труда казенных учреждений и взносы по обязательному социальному страхованию | 51 1 0016 | 111 |  | 135 922,3 |
| **Амбулаторная помощь** | **51 1 0016** | **111** | **0902** | **135 922,3** |

Рисунок 3. Пример наименования видов расходов и разделов и подразделов

Таким образом, новую информацию предоставляет только последняя рассмотренная строка (в ней заполнены все ячейки), в которой указано количество средств, потраченных на конкретные раздел и подраздел, в рамках данного вида расходов, данной статьи расходов, данной подпрограммы и программы. Остальные строки содержат только наименования соответствующих классификаторов. Значения столбца «сумма» могут быть получены суммированием строк с соответствующими значениями классификаторов. Например, просуммировав все строки бюджета, в которых указано значение ЦСР «51 1 0016», а значение ВР «111», мы получим сумму, равную значению строки 1, рис. 3, а просуммировав все строки, в которых первыми цифрами графы ЦСР являются «51» мы получим сумму, указанную в строке на рис. 1.

После изучения формата представленных данных можно сделать вывод, что данный формат не является оптимальным для изучения и содержит много «лишних» строк, затрудняющих его понимание. Для этого он будет приведен к формату, необходимому для загрузки данных в проект OpenSpending, который даже при ручной обработке является более удобным.

**Формат данных, загружаемых в проект OpenSpending**

В таблице 2 представлен пример преобразованных данных из бюджета Ленинградской области.

Таблица 2. Пример данных в формате проекта OpenSpending

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Год | ИД | Пр | Код Пр | Подпр | Код подпр | СР | Код СР | ВР | Код ВР | Рз и ПР | Код Рз и ПР | Сумма (руб) |
| 2014 | 51100161110902 | Программа "Здравоохранение в ЛО" | 51 0 0000 | Подпрограмма "Развитие медицины" | 51 1 0000 | Расходы на обеспечение деятельности гос.учреждений | 16 | Фонд оплаты труда … | 111 | Амбулаторная помощь | 902 | 135 922 300 |

Изменения заключаются в следующем:

* одним из требований проекта OpenSpending к загружаемым данным является наличие столбца с датой каждой транзакции (строки расходов). Так как в бюджете содержатся только планируемые расходы, поэтому здесь может быть указан только год. Но, например, если вы преобразовываете данные ГосЗакупок (фактические расходы), то вам следует указать более полную дату. Для выполнения требований в файл был добавлена дата – 2014 год (1ый столбец);
* при загрузке данных в OS необходимо выбрать валюту (для России - RUB), поэтому необходимо, чтобы данные в файле были указаны в этой валюте. Для выполнения этого требования была преобразована сумма расходов из тысяч рублей в рубли;
* для сокращения времени, необходимого для изучения данных, а также для группировки этих данных по разным классификаторам в OpenRefine в каждой строке указаны не только классификаторы, но и их наименования.

**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 вкладку Create project и загрузить файл с необходимыми данными – бюджетом какого-то региона (рис. 4). Данные должны быть в формате xls (Excel), csv, json, xml и др.

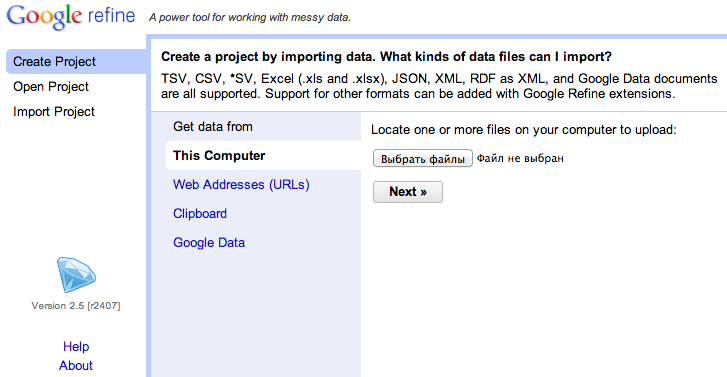


Рисунок 4. Окно создания проекта

После загрузки массива данных вам будет предложены некоторые настройки экспорта файла (рис. 5).

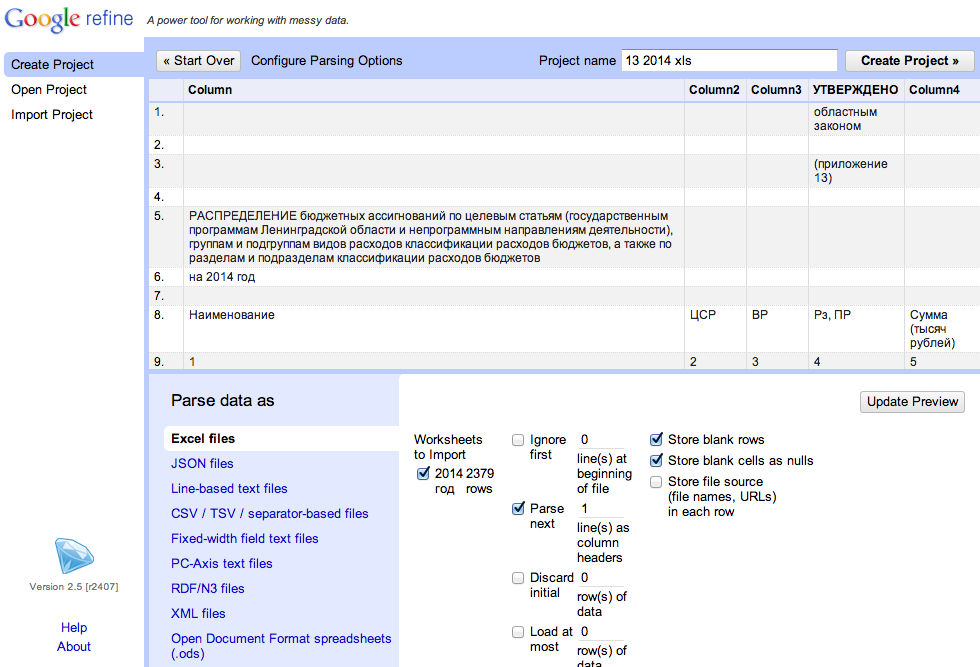


Рисунок 5. Настройка экспорта файла

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. 5 должна быть выделена первая строка).

После обработки настроек вы попадете в главное окно программы, в котором происходит обработка данных (рис. 6). Далее описан пошаговый алгоритм преобразования файла с объяснением каждой вводимой команды.

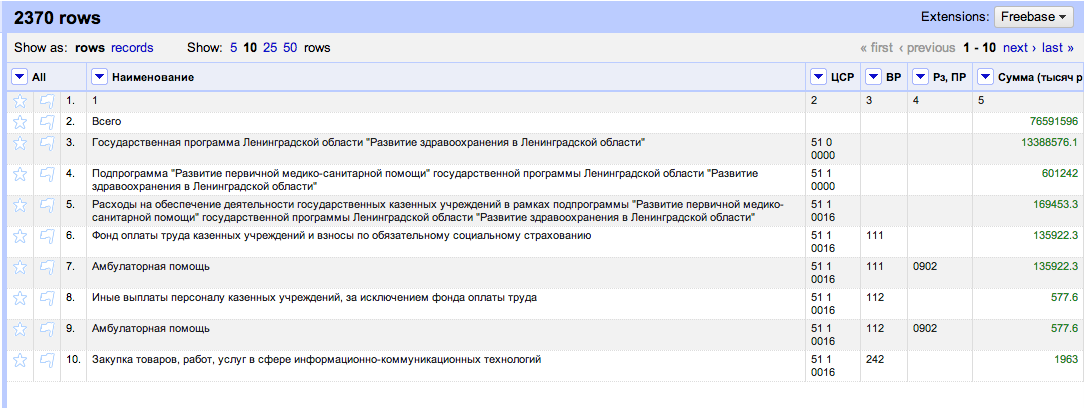


Рисунок 6. Главное окно программы OpenRefine

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. 7.1., сделать фасет по всем строкам, отмеченным «звездочкой» – рис. 7.2. (выбрать All – Facet – Facet by star), выбрать в фасете «Starred rows» все строки, соответствующие этому условию (true) – рис. 7.3. и удалить их, выбрав меню All – Edit rows – Remove all matching rows.

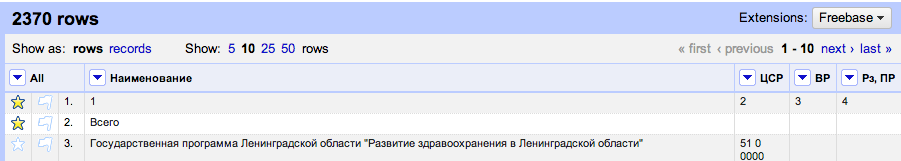


Рисунок 7.1.



Рисунок 7.2.

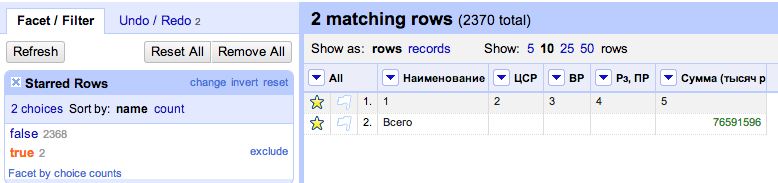


Рисунок 7.3.

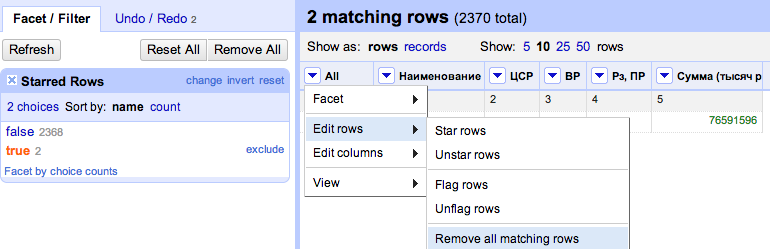


Рисунок 7.4.

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. 8). 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».

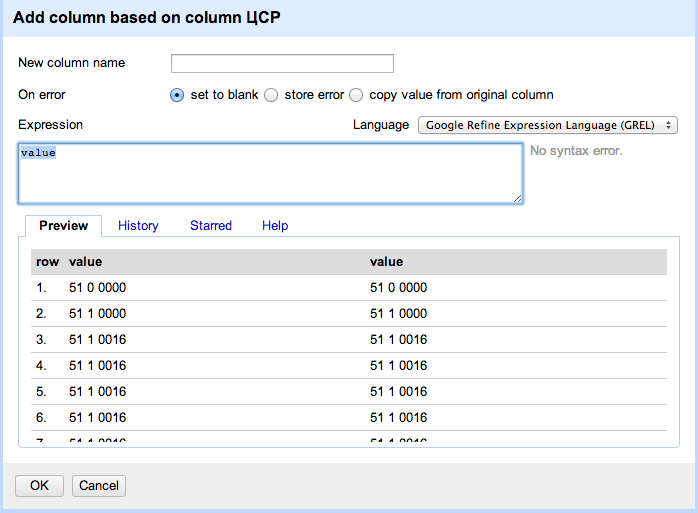


Рисунок 8.

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. 9).

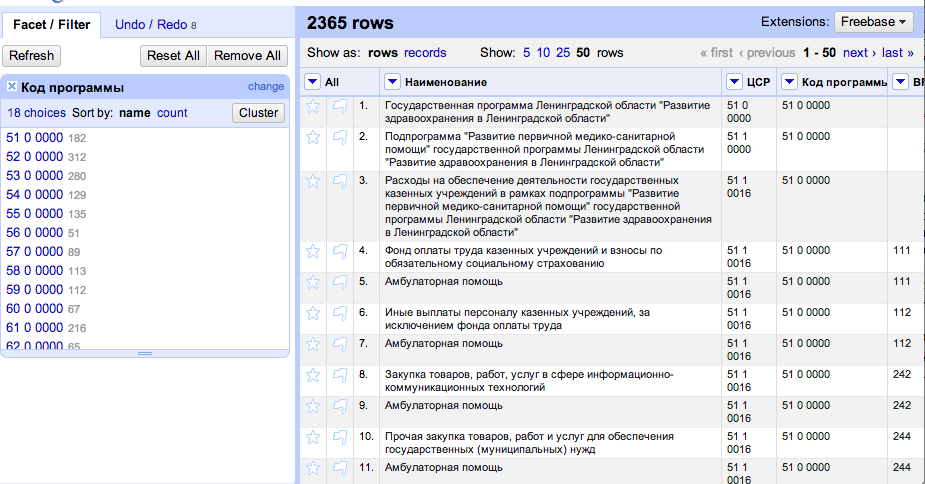


Рисунок 9.

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. 10). Например, на рис. 10 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).

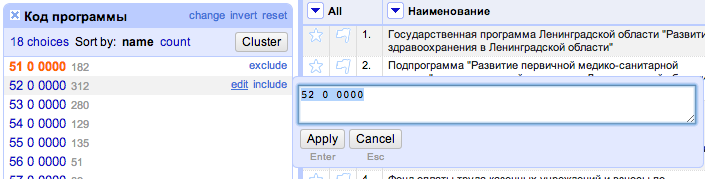


Рисунок. 10

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. 11.

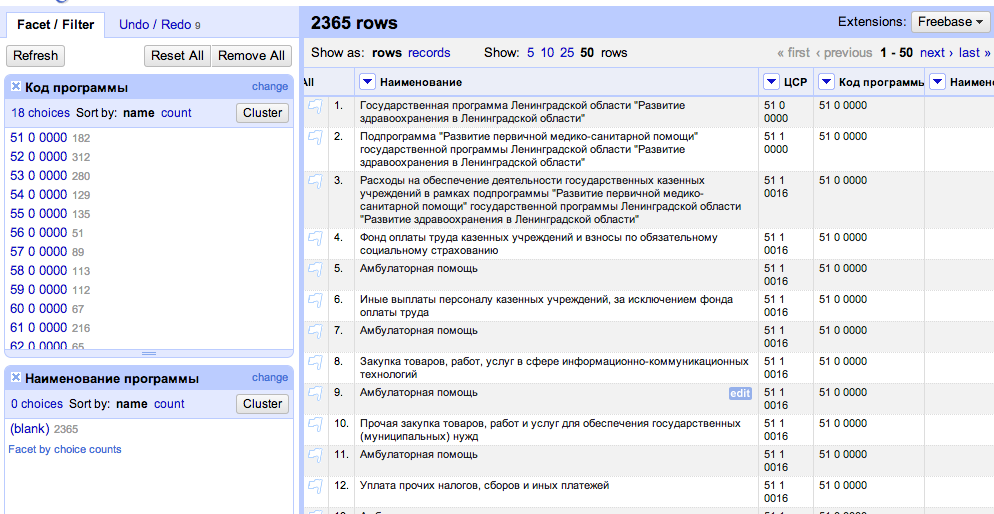


Рисунок 11.

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:

1. Выбрать группу значений в фасете «Код программы» (например, 51 0 000)
2. Найти строку, в которой значение ячейки столбца «ЦСР» совпадает с данной группой значений, и скопировать ее значение
3. 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.
4. Выбрать edit, в появившееся окно «вставить» скопированное значение (рис. 12), и нажать Apply (подтвердить изменение) или Cancel (если необходимо отменить изменения ячеек).
5. После этого необходимо выбрать другую группу в фасете «Код программы» и повторить все предыдущие действия.
6. 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”.
7. 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)**.
8. Строки, в которых указаны наименования программ в столбце «Наименование» можно удалить (они не содержат дополнительной информации, см. Раздел «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.

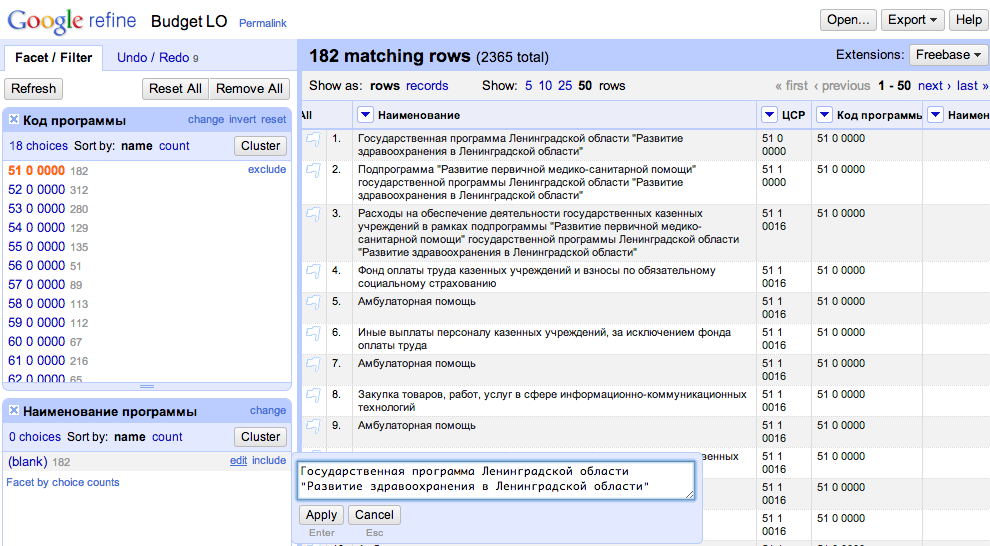


Рисунок. 12

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. 13.

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.

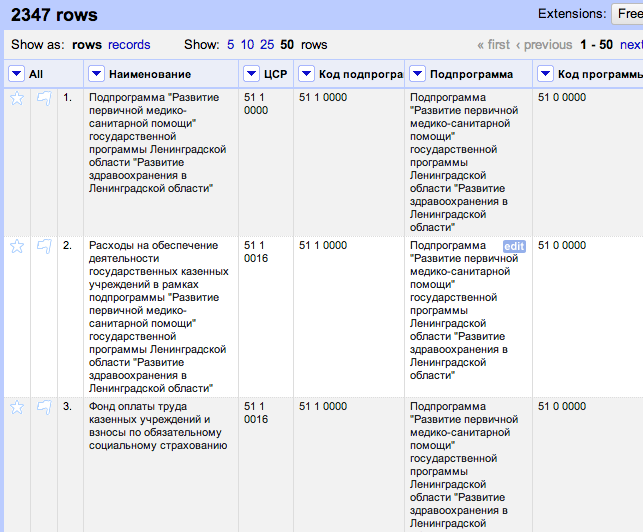


Рисунок 13.

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”**.

**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.