**Глава 12. Импорт веб-данных в Power Query**

Это продолжение перевода книги Кен Пульс и Мигель Эскобар. Язык М для Power Query. Главы не являются независимыми, поэтому рекомендую читать последовательно.

[Предыдущая глава](http://baguzin.ru/wp/?p=21109) [Содержание](http://baguzin.ru/wp/?p=20852#content) Следующая глава

Power Query работает очень хорошо, когда данные веб-сайта хранятся в таблицах. Иногда вы сможете импортировать данные из текстового файла. Однако, если данные не отформатированы с помощью тегов таблиц в HTML-коде, все становится довольно сложным, даже если у вас есть некоторые знания HTML.

В качестве примера попытается выяснить, в каких странах лучше всего продавать ваш товар. Вы предполагаете, что это ТОП-10 стран с самым большим населением. Вы нашли нужные данные на веб-сайте ЦРУ. Создайте новую книгу Excel, пройдите по меню *Данные* –> *Из интернета.* При запросе URL-адреса введите: <https://www.cia.gov/library/publications/the-world-factbook/rankorder/2119rank.html>. Нажмите Ok. Вам будет предложено ввести параметры проверки подлинности:



Рис. 12.1. Диалоговое окно веб-аутентификации

В левой части диалогового окна есть несколько параметров, позволяющих выбрать несколько вариантов проверки подлинности. В данный момент остановитесь на варианте *Анонимно*. Вопрос, который вам задают здесь, заключается в том, хотите ли вы использовать эти учетные данные для конкретной страницы или корневого домена. Если у вас нет особых причин, сохраните установку по умолчанию – корневой домен. Нажмите *Подключение*.

Power Query проверяет страницу, ища таблицы. Откроется окно *Навигатор*. В нашем примере есть два варианта *Document* и *Table 0*. Кликните вторую опцию, и в окне появится предварительный просмотр таблицы:



Рис. 12.2. Предварительный просмотр *Table 0* в *Навигаторе*

Таблица выглядит довольно хорошо. Ваша цель – отфильтровать первые десять стран. Кликните *Преобразовать данные*. Переименуйте запрос *Top 10 Population*. *Главная* –> *Сохранить строки* –> *Сохранить верхние строки* –> *10* (рис. 12.3). Завершите импорт: *Главная* –> *Закрыть и загрузить*.



Рис. 12.3. ТОП-10 стран по населению

### Подключение к размещенным в Интернете файлам

Если вы знаете URL-адрес текстового файла, вы можете подключиться непосредственно к нему. Создайте новую книгу Excel, пройдите по меню *Данные* –> *Из интернета.* Введите URL: <https://www.cia.gov/library/publications/the-world-factbook/rankorder/rawdata_2119.txt>. Нажмите Ok. Вы попадаете в окно предварительного просмотра Power Query, минуя шаг проверки подлинности веб-сайта. Причина в том, что во время выполнения предыдущего запроса вы указали, что хотите, чтобы параметры проверки подлинности применялись к корню веб-сайта. Если бы это было первое обращение к сайту, у вас запросили бы параметры проверки подлинности.



Рис. 12.4. Подключение к текстовому файлу в Интернете

Глядя на рис. 12.4, вы можете предположить, что заголовки столбцов не хранятся в текстовом файле (так же, как и даты оценки населения), а были предоставлены HTML веб-страницы. Это не страшно. Нажмите *Преобразовать данные*. Оставьте первые 10 строк, и переименуйте столбцы.

### Подключение к страницам без таблиц

Если веб-страница, к которой вы подключаетесь, не содержит тэгов таблиц в исходном коде HTML, задача усложняется. Откройте в веб-браузере страницу <https://www.cia.gov/library/publications/the-world-factbook/rankorder/2119rank.html>, откройте окно *Инструменты разработчика* (в Chrome можно нажать клавишу F12).



Рис. 12.5. Окно *Инструменты разработчика*

Хитрость заключается в том, чтобы последовательно разворачивая тэги, добраться до таблицы. В правом окне наводя курсор мыши на ту или иную строку кода html, в левом окне будут подсвечиваться различные элементы страницы. Когда подсветка покрывает таблицу, вы разворачиваете элемент, нажав на стрелку, которая указывает вправо. При этом стрелка поворачивается вниз, открывая дополнительные элементы, и вы продолжаете процесс, пока не найдете нужную вам таблицу.

Теперь можно попробовать воспроизвести в запросе Power Query процесс поиска таблицы среди тэгов html. Создайте новый файл Excel. Пройдите по меню *Данные* –> *Из интернета.* Введите URL: <https://www.cia.gov/library/publications/the-world-factbook/rankorder/2119rank.html>. В окне *Навигатор* выберите *Document*:



Рис. 12.6. Окно *Навигатор* для выбора импорта из html

Нажмите *Преобразовать данные*. Откроется редактор Power Query (рис. 12.7). Если кликнуть в пустое пространство справа от слова *Table* (1), в нижней части экрана появится поле с перечислением детских элементов, относящихся к *Table* (2).



Рис. 12.7. Исходная таблица

Раскройте *Table*, кликнув на нее (3 на рис. 12.7). Одна таблица HTML теперь представлена двумя таблицами: HEAD и BODY



Рис. 12.8. Таблица HTML раскрылась в две: HEAD и BODY

Обратите внимание: имена таблиц в столбце *Name* соответствуют тэгам в браузере.



Рис. 12.9. Тэги в браузере

Теперь можно идти параллельно. Сначала смотреть в браузере, какой тэг подсвечивает таблицу, и раскрывать его, а затем переходить в Power Query и раскрывать соответствующую таблицу.



Рис. 12.10. Раскрыт тэг *body* в браузере и таблица BODY Power Query

Проблема в том, что в браузере все теги имеют имена, но в Power Query они не отражаются, поэтому очень легко запутаться. Кроме того, в области *Примененные шаги* маршрут не отслеживается, а с каждым новым раскрытием таблицы формула шага видоизменяется, не давая возможности создать резервную копию или вернуться на шаг назад. Так что, если вы пошли не по той дорожке, единственный выход – начать все сначала.

Если вы всё сделали верно, то доберетесь до нужной таблицы:



Рис. 12.11. Таблицы с численностью населения по странам в кодах html и в Power Query

Дальнейшая обработка таблицы в Power Query будет описана позже.

### Предостережения и разочарования, связанные с веб-опытом

Power Query отлично справляется с файлами и структурированными таблицами, размещенными в Интернете. А вот обработка тэгов, пожалуй, не является сильным местом Power Query (см. [Извлечение данных с web-страниц с помощью кода на языке Python](http://baguzin.ru/wp/?p=14390)). Помните также, что дизайн страниц может меняться, и сложный хорошо разработанный запрос Power Query может из-за этого сломаться.