© ***Владимирова А.В.***
ИВ РАН

СТАТИСТИКА ВНЕШНЕЙ ТОРГОВЛИ СТРАН ЮГО-ВОСТОЧНОЙ АЗИИ: ПРОБЛЕМА ЦЕЛОСТНОСТИ ДАННЫХ, ПРЕДОСТАВЛЯЕМЫХ ЧЕРЕЗ API

Введение

В области историко-экономических исследований анализ временных рядов объемов экспорта и импорта занимает особое место. Задачи отслеживания трендов и выявления закономерностей протекания исторических процессов во международной торговле стали особенно актуальны в эпоху глобализации и очень часто такие работы оказываются полезными ученым разных дисциплин, в том числе и востоковедам, не редко занимающимся вопросами на пересечении наук.

Логично, что серьезные усилия и средства были направлены именно на создание открытых баз данных по торговле между странами. Нам теперь становится доступно больше источников данных, больше показателей и, главное, временных периодов. Так, российские историки запустили проект, в ходе которого стало возможным посмотреть в интернете данные о внешней торговле Российской империи в 19 в. и по европейской, и по азиатской границам[1].

С одной стороны, эти усилия оказывают положительное влияние на науку, в частности, многообразие данных, которые легко скачиваются из интернета, позволяет быстро развивать методы анализа. С другой стороны, в эпоху революции данных и текущего кризиса воспроизводимости в науке[2] критерии оценки качества источника меняются принципиально. Все чаще на слуху оказывается принцип GIGO – «мусор на входе, мусор на выходе», напоминающий аналитикам, что даже самая сложная математическая модель не даст надежный результат, если исходные данные ненадлежащего качества.

Понимание важности дискуссии о качестве данных привело к росту работ, раскрывающих многообразие существующих

и только еще возникающих проблем, но, главное, предлагающих различные методы их решения. В этой статье мы обращаемся к одной из них, связанной напрямую с ростом объемов доступной информации и необходимостью эффективно ею делиться. На примере работы ресурса, предоставляющего доступ к крайне популярной базе «Комтрейд ООН» [3], мы продемонстрируем возможные проблемы доступа к данным с использованием программного интерфейса приложений (API).

Востоковеды и программные интерфейсы приложений

Термин «большие данные» уже прочно вошел не только в лексикон аналитиков, но и в лексикон средств массовой информации. Объемы данных растут очень быстро и все чаще становится сложно скачать их в виде одного файла. Конечно, кажется, что для востоковедов, занимающихся историческими процессами вопрос передачи данных по сети более актуален, чем для тех, кто изучает современность и работает со статистикой за один год. Между тем, это не так.

При создании онлайн ресурсов в подавляющем большинстве случаев тип информации в базе не столь важен и обычно программисты выбирают те средства, которые популярны и им более привычны. Сейчас в области распространения и обмена информацией одной из широко распространившихся практик стало использование программных интерфейсов приложений (application programming interface), чаще в России проходящих под англоязычной аббревиатурой «API». Предоставление доступа к данным из API стало нормой для провайдеров и соответствующие страницы или даже разделы с документацией появляются на сайтах государственных и международных организаций. Пользователи онлайн ресурсов могут не осознавать, что данные, которые они видят на веб-странице, в какой-то момент были переданы через API и отсутствие этого знания не мешает им успешно с ними работать. Между тем, проблемы с API неотвратимо влияют на качество данных, которые они получают.

Главной задачей этой статьи стало выявление трудностей, которые могут возникать при использовании API, даже в ситуации, когда мы получаем доступ к данным авторитетных орга-

низаций, таких как Организация Объединенных Наций (ООН). Мы также считаем крайне важным поделиться с коллегами нашей практикой решения проблем, с которыми приходится сталкиваться при оценке качества массива статистических данных по внешней торговле стран Юго-Восточной Азии.

Обоснование выбора базы «Комтрейд ООН» как кейса для анализа работы с API

База «Комтрейд ООН» по внешней торговле является одной из самых популярных в историко-экономических исследованиях по целому ряду причин. Во-первых, ООН считается надежным и независимым источником данных, к тому же, эта авторитетная международная организация получает статистику внешней торговли напрямую от государственных статистических агентств, таких как Федеральная служба государственной статистики (Росстат). Во-вторых, набор данных содержит множество важнейших показателей, доступных для 283 акторов (стран, регионов, территорий, мира в целом), начиная с 1988 г., при этом в базе четко указывается, какая страна (reporter) предоставила конкретную цифру для отражения своих торговых отношений со своей страной-партнёром (partner).

Очевидно, что это большая по объему база данных, содержащая миллионы записей, по которой можно проводить поиск, используя достаточно сложные или специфические запросы. Например, можно посмотреть, что Бруней в 1993 г. импортировал из Таиланда 109 кг. незамороженных осьминогов на сумму в 767 американских долларов.

Портал данных «Комтрейд ООН» постоянно совершенствуется, и мы можем видеть, что его создатели отслеживают новейшие тренды в области анализа данных. Достаточно зайти на главную страницу сайта и просто взглянуть на предлагаемую визуализацию. Учитывая, насколько детализированы данные в этой базе и очевидную заинтересованность в развитии средств их анализа, вполне ожидаемо, что для этого источника тоже доступно скачивание из API. На сайте также размещены описания с примерами запросов данных на популярных среди аналитиков и ученых языках программирования Python и R. Более того, предоставляется не только бесплатный доступ к

данным с четко обозначенными ограничениями по количеству запросов, но и платная подписка, позволяющая серьезно нагружать сервера.

Эти возможности крайне важны для тех, кто проводит историко-экономические исследования или работает со временными рядами в рамках иных научных дисциплин, поскольку сложно скачать базу данных, содержащую более миллиона записей, используя интерфейс вебсайта «Комтрейд ООН». Если нужны показатели для пары стран, то очень просто посмотреть или загрузить данные на соответствующей странице на сайте ООН[4]. Но получать данные уже для 20 пар стран за несколько лет придется множеством запросов, указывая максимально по 5 лет, 5 стран-источников данных и 5 стран-партнеров. Тратить время и силы на то, чтобы скачать и объединить эти данные было бы неэффективным решением, к тому же, как будет продемонстрировано далее, еще и потенциально ведущим к проблемам с надежностью информации.

Конечно, можно загрузить базу «Комтрейд ООН» с других ресурсов, однако в этом случае возникает целый ряд серьезных вопросов, связанных с ответственностью исследователя. Главный из них - может ли автор работы гарантировать, что на другом ресурсе предоставлены именно данные ООН и в полном объеме. В данном случае также может быть спорным указание в качестве источника данных «Комтрейд ООН».

Так, мы проверили одного из самых авторитетных провайдеров данных «Knoema» и обнаружили различия между базой, которую они предоставляли как «Комтрейд ООН»[5] и базой на оригинальном сайте. В ответ на наш запрос, названная компания провела внутреннюю проверку и пришла к выводам, что у них действительно возникли проблемы со скачиваем, связанные с принципами работы с данными в ООН и приняла решение изменить практику работы с информационным порталом этой организации. Следует отметить, что серьезность проблем с расхождениями будет, конечно, зависеть от объекта и предмета конкретного исследования. Так, если в фокусе работы страны Юго-Восточной Азии, то потеря данных даже для одной из 10 стран АСЕАН может существенно повлиять на результаты.

Методология мониторинга целостности данных, предоставляемых через «UN Comtrade API»

Поставив вопрос о надежности данных, получаемых аналитиками по официальному «UN Comtrade API», мы решили провести мониторинг, выступив как пользователь, не обладающий специальной подпиской или особыми правами доступа. Мы написали скрипт в свободной программной среде вычислений R с использованием кода запросов, опубликованного на сайте ООН[6], который автоматически скачивал данные в течение месяца с 01 по 31 мая 2018 г. Этот период был выбран, т.к. в основе нашей методологии мониторинга лежал принцип комплиментарной недели. Иначе говоря, нам нужно было обеспечить загрузку данных в будние, выходные и праздничные дни в разное время. Мы разделили сутки на четыре периода по 6 часов, т.к. на момент проведения нашего исследования пользователь без подписки был ограничен по количеству запросов (1000 в час) и соответственно в скрипт было добавлено время ожидания между обращениями к серверам ООН.

Мы разработали различные подходы к формированию запросов, но в этой статье представим лишь результаты для запроса агрегированных данных по торговле товарами для каждой страны по одному году с 1988 по 2017 гг. с шагом в 5 лет. Цикл начинается с 1988 г., затем последовательно запрашиваются данные для 1993, 1998, 2003, 2008 и 2013 гг., а после чего происходит возвращение к 1989 г. и перебор продолжается по тому же принципу. Каждое обращение содержит запрос данных для 5 стран и всех их партнеров из списка, указанного в рекомендациях на сайте ООН. Таким образом, на каждый год приходится 59 запросов, после чего данные объединяются и в файл с результатами мониторинга записывается сколько ненулевых наблюдений удалось получить. При этом фиксируются и другие показатели, такие как дата и время запроса, коды стран в запросе, возникающие ошибки и их тип, сколько раз пришлось делать запрос, прежде чем сервер выслал данные и т.д. Скачивание данных производится на одном и том же компьютере, без обновления программного или аппаратного обеспечения, по одному и тому же каналу доступа к сети Интернет.

Результаты мониторинга целостности данных, предоставляемых через «UN Comtrade API»

В результате проведенного мониторинга скачивания базы «Комтрейд ООН» с использованием «UN Comtrade API» мы выявили несколько важных для исследователей особенностей такого доступа к статистике внешней торговли. В этой статье обратимся к трем из них: пропуски в данных, замена и удаление записей.

Пропуски в данных. Мы выяснили, что «UN Comtrade API» не является надежным методом передачи базы «Комтрейд ООН» и в ответ на один и тот же запрос можно получить разное количество записей. Иначе говоря, скачав данные один раз и не зная точно сколько их должно быть, пользователь не может быть уверен, что полученная информация целостная. Пропуски в данных могут достигать 42%. При этом, какие-то года более стабильно скачиваются, чем другие (Таб.1).

На сайте ООН представлена документация по запросу информации о доступности данных через API[7], но рекомендации и примеры по написанию кода в R[8] не содержат проверки на целостность и предупреждений о возможных пропусках. Между тем, в свете полученных нами результатов, она должна проводиться в обязательном порядке и быть интегрирована в скрипт закачки. Однако, мы выяснили, что сделать это может быть проблематично, особенно для ученых, только начинающих работать с «Комтрейд ООН». Наши рекомендации по решению этой проблемы будут представлены в следующем разделе.

Замена данных. Мы зафиксировали изменения в «Комтрейд ООН» без публикации на сайте объявления о смене версии базы данных или списка скорректированных показателей. Например, экспорт Дании в ее страну-партнер Сингапур стал составлять не $392 817 046, а $393 475 612. На момент написания статьи эта цифра опять незначительно отличалась от той, что была в базе 31 мая 2018 г. ($393 514 220). С точки зрения статистического анализа, маловероятно, что такие изменения повлияют на результат, но с точки зрения обработки данных, это может стать проблемой.

Таблица 1. *Мониторинг пропусков в данных при скачивании с использованием официального программного интерфейса приложений «UN Comtrade API»*

год	максимум скаченных наблюдений	максимум пропусков в данных	год	максимум скаченных наблюдений	максимум пропусков в данных	год	максимум скаченных наблюдений	максимум пропусков в данных
1988	3496	25%	**1998**	36264	13%	**2008**	53806	7%
1989	7568	15%	**1999**	38191	0%	**2009**	53794	15%
1990	8880	0%	**2000**	45437	0%	**2010**	55201	22%
1991	11055	19%	**2001**	46296	4%	**2011**	54819	3%
1992	15899	1%	**2002**	46984	2%	**2012**	55023	5%
1993	19615	7%	**2003**	48393	9%	**2013**	54893	16%
1994	26142	0%	**2004**	50066	3%	**2014**	53948	15%
1995	29946	6%	**2005**	50643	7%	**2015**	52550	7%
1996	32306	4%	**2006**	51964	5%	**2016**	50962	6%
1997	34809	2%	**2007**	53357	7%	**2017**	34912	42%

Примечание: Доступ осуществлялся с 1.05.2018 г. по 31.05.2018 г. с использованием свободной программной среды вычислений R. Код написан на основе документации «UN Comtrade API», опубликованной на сайте ООН (https://comtrade.un.org/data/Doc/api/ex/r).

Если не знать о подобной практике изменения базы и не вносить советующую проверку в скрипт, то можно получить сходный результат: по итогам месяца мы обнаружили в наших данных дубликаты потоков экспорта и импорта для отдельных пар стран, включая и страны Юго-Восточной Азии.

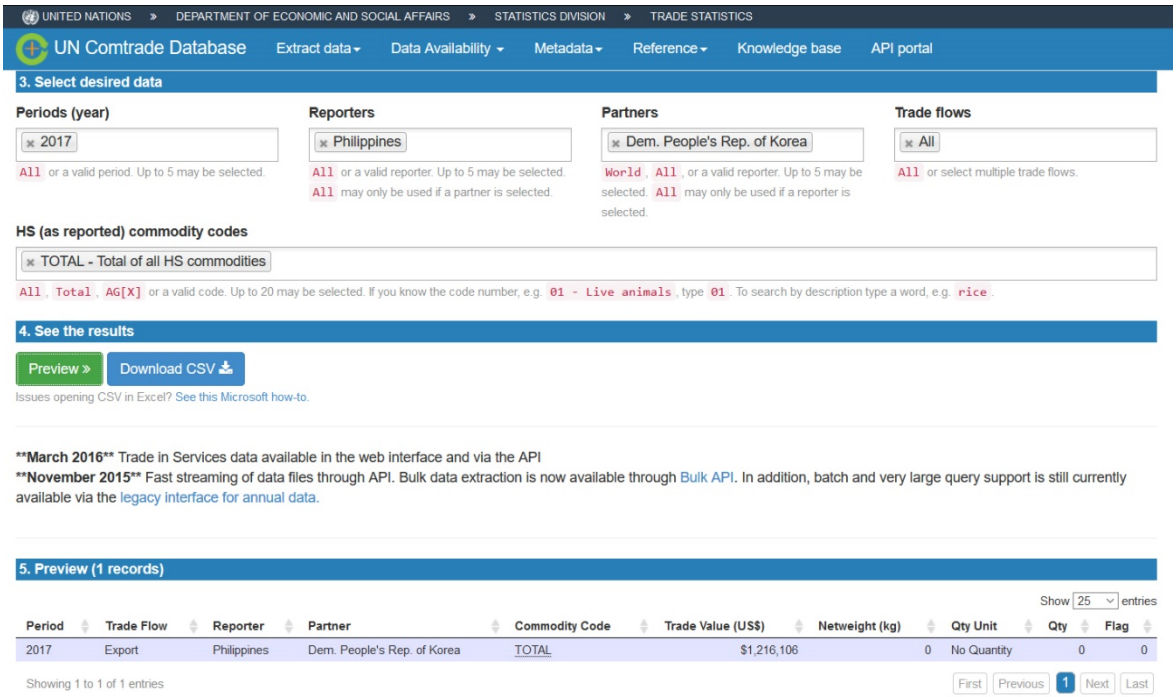

Рис.1. *Подтверждение удаления ранее существовавшего показателя импорта для диады Филиппины - КНДР из базы данных «Комтрейд ООН»*

Удаление данных. Вероятно, с точки зрения регионоведа и историка, одной из самых интересных практик работы сотрудников Статистического отдела ООН с базой «Комтрейд ООН» является удаление данных. За май 2018 г. мы обнаружили исчезновение 3 записей за 2014 г., 5 - за 2015 г., 10 - за 2016 г. и 8 - за 2017 г. Среди них есть весьма любопытные случаи, так, в мае 2018 г. из базы был удален показатель импорта Филиппин из Корейской Народно-Демократической Республики (КНДР) за 2017 г., но оставлен показатель экспорта (Рис.1). Причины удаления этой конкретной записи на сайте не разъясняются, поэтому мы можем только строить свои предположения. Тем не менее, ученым при написании статей желательно не просто указывать «Комтрейд ООН» как источник информа-

ции, но и приводить точную дату доступа к ресурсу, как того, например, требует стандарт для оформления цитирования ГОСТ Р 7.0.5-2008. Иначе может оказаться, что данные не просто немного отличаются, но и вообще не существуют.

Между тем, историкам и ученым, специализирующимся на отдельных странах, было бы, конечно, полезно ознакомиться с такими удаленными показателями. Мы считаем, что выявление в ходе нашего мониторинга этой характеристики важно не только с точки зрения правильной отсылки к источнику, но и с точки зрения организации работы с базой. Видимо, тем исследователям, которые постоянно анализируют статистику внешней торговли имеет смысл периодически, раз в две недели или раз в месяц, скачивать и архивировать текущую версию этой базы.

Возможные решения проблемы контроля целостности информации, скачиваемой через «UN Comtrade API»

Как уже ранее было сказано, в рамках нашего мониторинга мы выявили проблему с целостностью информации в базе «Комтрейд ООН», предоставляемой через API. Очевидным решением было бы включение в скрипт показателя о количестве доступных записей, который ООН не скрывает, однако, у нас возникли с этим определенные сложности. Мы продемонстрируем проблемы с реализацией контроля целостности на примере данных по внешней торговле товарами Лаоса по всем потокам в классификации HS за 2011 г.

Начнем с того, что узнаем, сколько записей по внешней торговле можно скачать для Лаоса. Для него их будет относительно немного, поэтому мы просто скачаем их, используя интерфейс сайта «Комтрейд ООН»[9]. Выбрав Лаос в качестве рапортующей страны мы получаем файл со 131 записью, а выбрав Лаос в качестве страны-партнера – со 183. Тот же массив закачивается и при использовании API. Таким образом, мы получаем контрольный показатель, который теперь необходимо найти в информации о доступности данных на сайте ООН. Сразу уточним, что запрос о наличии записей в текущей версии «Комтрейд ООН» можно сделать только по стране-источнику

данных, поэтому мы возьмём 131 за искомый контрольный показатель.

Узнать о доступности данных можно несколькими способами, включая специальный запрос к API. Мы начнем с меню «Доступность Данных» (Data Availability) и представленного на сайте специального раздела, в котором можно посмотреть на аналитическую панель, созданную на основе «Tableau»[10]. В принципе, подобный дашборд является уже привычным и проверенным средством интерактивной визуализации информации, но видимо, команда сайта «Комтрейд ООН» пока еще разрабатывает и апробирует этот инструмент. На момент написания нашей статьи его виджеты показывают, когда для каждой страны в последний раз обновлялись данные и для каждого периода указывается количество стран и когда они были включены в базу. Мы попытались узнать про доступность данных для Лаоса за 2011 г. (Рис.2), выбрав эту страну в предлагаемых панелях, но смогли выяснить только то, что данные за этот период были внесены в базу или обновлены в 2017 г.

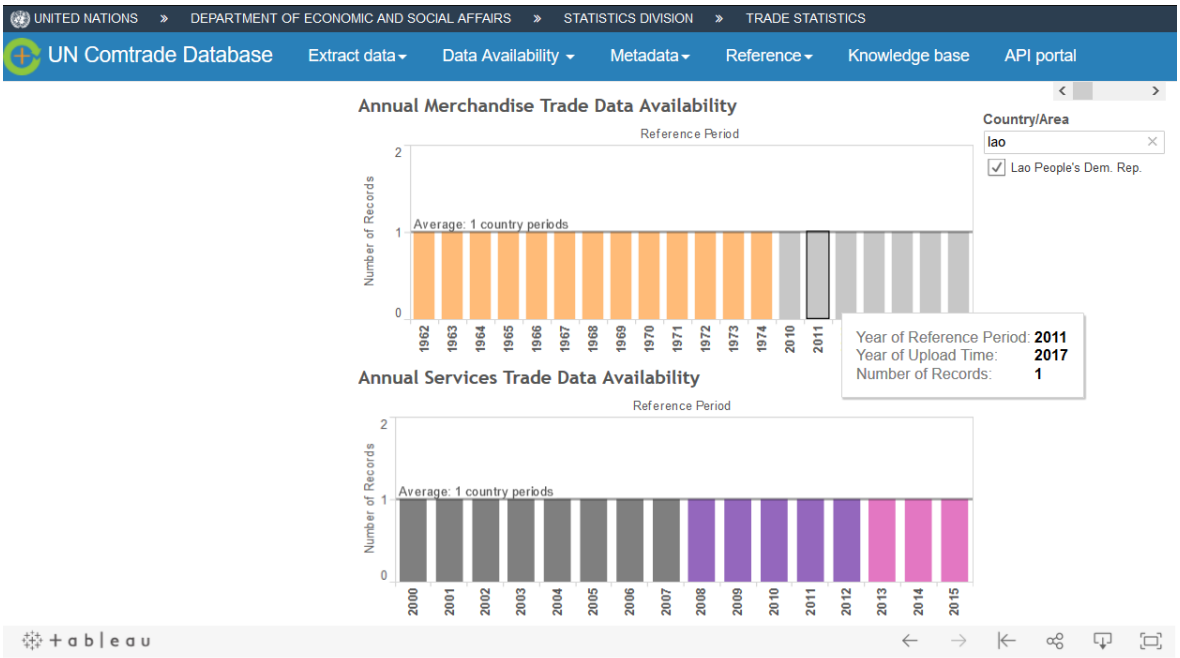

Рис.2. *Дашборд, представляющий информацию о доступности данных по внешней торговле Лаоса в 2011 г. на ресурсе «Комтрейд ООН»*

Очевидно, что таким путем нужный нам показатель мы не найдем, но поскольку этот дашборд не единственный источник информации о доступности данных на сайте «Комтрейд ООН», то можно использовать альтернативы. Самая простая – скачать соответствующий файл «DataAvailabilityComtrade.csv» на главной странице этого веб-ресурса[11]. Конечно, немного странно, что ссылка на файл размещена там, а не в меню «Доступность Данных», поскольку такое ее расположение может вызывать определённые проблемы с поиском. Тем не менее, файл есть и содержит информацию по количеству записей для Лаоса с указанием года, даты последнего обновления и количества записей для разных классификаторов (HS, H0, H1, H2, H3, BEC, S2, S3, S4, S1, EB02). Однако, согласно этому файлу в 2011 г. для классификаторов HS или H3 должно быть доступно 16452 записи. Аналогичная цифра возвращается и при запросе через API (см. например: https://comtrade.un.org/api/refs/da/view?type=C&freq=A&ps=2011&px=HS&r=418).

Тех, кто работал с предыдущей версией сайта «Комтрейд ООН», эта цифра не удивит и не введет в заблуждение, поскольку они сразу увидят, что в файле просто отсутствует дополнительный код классификации, позволяющий уточнять товарную группу и ее агрегированные показатели. Если бы при запросе информации о доступности данных можно было, как и прежде, указать параметр «cc=TOTAL», то мы бы получили информацию о доступности 131 записи для Лаоса, как рапортующей страны. Т.е. эта цифра соответствует нашему выбранному контрольному показателю и именно ее мы искали. Соответственно, у нас появляется возможность реализовать проверку целостности, т.к. теперь мы знаем, где можно узнать правильное количество записей для всех стран, рапортующих о показателях. К счастью, в настоящее время эту информацию можно найти на пока еще доступной прежней версии сайта «Комтрейд ООН»[12] и хотелось бы верить, что эта крайне необходимая функция уточнения запроса появится в новой версии.

Мы не будем приводить в этой статье информацию о том, сколько для каждого года должно быть записей, т.к. эта цифра может меняться. Следует помнить, что в течение любого месяца возможно и удаление, и внесение новых записей в базу дан-

ных «Комтрейд ООН». Исходя из вышесказанного, мы настоятельно советуем включать проверку целостности информации в скрипт закачки данных, особенно, если требуется получить данные за последние годы.

Заключение

В качестве заключения мы хотели бы представить несколько практических рекомендаций по работе с данными «Комтрейд ООН», часть из которых будет актуальной и для баз из других источников.

Во-первых, крайне желательно скачивать данные с сайта организации, выступающей как первоисточник, в нашем случае с сайта ООН. При использовании сторонних онлайн ресурсов даже с хорошей репутацией нет полной гарантии, что там представлены актуальные и полные данные, особенно, если у источника данных есть проблемы предоставлением доступа к ним через API. Надо понимать, что в случае возникновения вопросов о пропусках или изменениях ответственность ложится на исследователя, а не на стороннего провайдера данных.

Надо обязательно проверять, сколько записей должно быть в полной версии базы. Этот показатель можно использовать и как контрольный, если исследователь все же решит скачать версию со стороннего ресурса. Для «Комтрейд ООН» требование о проверке целостности информации является обязательным, причем если скачиваются отдельные показатели, а не все доступные данные, то за сведениями о доступности данных следует обратиться к прежней версии сайта.

Забирать данные «Комтрейд ООН» из интернета лучше одним днем и давать сноску с указанием конкретной даты. Учитывая, что при изучении исторических процессов может потребоваться скачать несколько временных рядов, пользователю придется выбрать доступ через API, но и по времени, и с точки зрения дальнейшей работы, эти усилия окупятся. Скрипт позволит архивировать в автоматическом режиме разные версии базы, изменения в которой сами по себе могут представлять значительный интерес для историко-экономических исследований.

[1] Валетов Т.Я. Проект «Статистика внешней торговли Российской империи»: характеристика источника и цифрового ресурса // Историческая информатика. 2017. № 1. С. 5–14.
[2] Baker M. 1,500 Scientists Lift the Lid on Reproducibility [Электронный ресурс]. URL: http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970 (дата обращения: 05.02.2017).
[3] UN Comtrade [Электронный ресурс]. URL: https://comtrade.un.org/ (дата обращения: 19.12.2017).
[4] Download Trade Data [Электронный ресурс]. URL: https://comtrade.un.org/data (дата обращения: 12.06.2018).
[5] UN Comtrade: Merchandise Trade by Commodity, HS - As Reported [Электронный ресурс]. URL: https://knoema.com//COMTRADE2015R1/un-comtrade-merchandise-trade-by-commodity-hs-as-reported (дата обращения: 03.04.2018).
[6] Using the UN Comtrade data API with R [Электронный ресурс]. URL: https://comtrade.un.org/data/Doc/api/ex/r (дата обращения: 23.04.2018).
[7] The UN Comtrade Data Extraction API [Электронный ресурс]. URL: https://comtrade.un.org/data/doc/api (дата обращения: 25.11.2018).
[8] Using the UN Comtrade data API with R [Электронный ресурс]. URL: https://comtrade.un.org/data/Doc/api/ex/r (дата обращения: 23.04.2018).
[9] Download Trade Data [Электронный ресурс]. URL: https://comtrade.un.org/data (дата обращения: 12.06.2018).
[10] Data Availability [Электронный ресурс]. URL: https://comtrade.un.org/data/da (дата обращения: 30.11.2018).
[11] DataAvailabilityComtrade.csv [Электронный ресурс]. URL: https://comtrade.un.org/api/refs/da/view?fmt=csv (дата обращения: 30.11.2018).
[12] Data Availability by Year [Электронный ресурс]. URL: https://comtrade.un.org/db/mr/daYearsResults.aspx (дата обращения: 30.11.2018).