# СИСТЕМНЫЙ АНАЛИЗ, УПРАВЛЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ

УДК 342.843

Вестник УрФО № 1(43) / 2022, с. 28-34



Мирвода С.Г., Поршнев С.В., Рябко Н.Ю.

DOI: 10.14529/secur220104

# АВТОМАТИЗАЦИЯ ПРОЦЕДУРЫ ДОСТУПА К ЭЛЕКТОРАЛЬНЫМ ДАННЫМ, РАЗМЕЩЕННЫХ НА САЙТЕ ЦЕНТРАЛЬНОЙ ИЗБИРАТЕЛЬНОЙ КОМИССИИ РОССИЙСКОЙ ФЕДЕРАЦИИ

В статье описан программный инструмент, осуществляющий автоматическую выгрузку электоральных данных о выборах в 2018 г. Президента Российской Федерации (РФ), размещенных на официальном сайте Центральной избирательной комиссии (ЦИК), включая: сводные данные по РФ; сводные данные, представленные избирательными комиссиями субъектов РФ; сводные данные, представленные территориальными избирательными комиссиями (ТИК), созданными в субъектах РФ; данные, представленные участковыми избирательными комиссиями (УИК), относящимися к соответствующим ТИК, и ее размещение в виде таблицы (размером 21 столбец на 100757 строк) в текстовом файле формата RFC 4180.

Наличие столь представительной выборки данных позволит оценить адекватность статистических методов, используемых в электоральной криминалистике для выявления фальсификации результатов выборов.

**Ключевые слова:** выборы, электоральные данные, избирательная комиссия, веб-страница, веб-скрепинг, анализ данных, информационный поиск.

Mirvoda S.G., Porshnev S.V., Ryabko N.Yu.

# AUTOMATION OF THE PROCEDURE FOR ACCESSING ELECTORAL DATA POSTED ON THE WEBSITE OF THE CENTRAL ELECTION COMMISSION RUSSIAN FEDERATION

The article describes software that automatically loads electoral data on the 2018 presidential elections of the Russian Federation posted on the official website of the central committee of the elections. The data includes summary data for the Russian Federation, summary data provided by the election commissions of the subjects of the Russian Federation, summary data provided by territorial election commissions; the data provided by precinct election commissions. The data is presented in the form of a table (21 columns by 100757 rows) in a text file according to the RFC 4180 format.

Such a representative sample of the election data will allow us to assess the adequacy of statistical methods used in electoral criminology to detect falsification of election results.

Keywords: election, election data, election commission, webpage, web-scrapping, data retrieval, data mining.

#### Введение

Сегодня проблемы, связанные с оценкой достоверности результатов тех или иных выборов в органы власти и местного самоуправления в РФ и выявления возможных фальсификаций электоральных данных, являются предметом острых научных и политических дискуссий, в которых принимают участие политики, политологи, юристы, специалисты по прикладной статистике (электоральные статистики), а также политически активные избиратели) (см., например, [1,2]). При этом, в большинстве случаев, мнения участников дискуссий по обсуждаемой проблеме, в зависимости занимаемой ими политической позиции, оказываются диаметрально противоположными: от полного отрицания каких-либо фальсификаций электоральных данных (например, члены ЦИК и избирательных комиссий нижних уровней), до полного отрицания возможности проведения в РФ честных выборов (например, кандидаты, не получившие поддержки избирателей, а также члены либерально настроенного сообщества электоральных статистиков (см., например, [3]). В, как очевидно, этой ситуации требуется независимая оценка адекватности используемых ими методов и результатов.

При этом основная проблема состоит в отсутствии электоральных данных, полученных по результатам выборов, отсутствие фальсификаций в которых доказана с помощью методов, отличных от методов электоральной статистики (например, на основе данных, представленных независимыми наблюдателями, данных экзит-поллов и т.д.), анализ которых мог бы определить набор «эталонных» статистических характеристик. Однако, выводы об отсутствия фальсификаций электоральных данных, сделанные на основе использования косвенных методов, также могут быть поставлены под сомнение. В этой ситуации представляется целесообразным обратиться к электоральным данным, полученным в ходе выборов, в которых преимущество одного из кандидатов над другими кандидатами было столь велико, что его победа очевидна. Таковыми были выборы Президента РФ в 2018 г., электоральные данные которых размещены на сайте ЦИК РФ [4], на котором, однако, не предусмотрена возможность автоматической выгрузки всех или части электоральных данных в виде, пригодном для последующего анализа. В этой связи извлечение электоральных данных оказывается возможным только в ручном режиме, что весьма трудоемко, так как по данным ЦИК число избирательных участков на этих выборах превысило 97 тысяч.

В статье описана разработанная авторами технология, обеспечивающая автоматическую выгрузку электоральных данных о результатах выборов Президента РФ в 2018 г.

## 2. Структура электоральных данных о результатах выборов Президента РФ в 2018 г.

Напомним, что система избирательных комиссий

по выборам Президента РФ в 2018 имела следующую иерархическую структуру: ЦИК, избирательные комиссии (ИК) субъектов РФ, территориальные ИК, участковые ИК. Всего в 2018 г. было сформировано: 85 ИК субъектов РФ 85 (в том числе, ИК «Территория за пределами РФ», к которой были отнесены 394 УИК, сформированные за пределами РФ, а также ИК г. Байконур (Республика Казахстан), на территории которого были сформированы 7 участковых ИК), 3177 территориальных ИК, 97314 участковых ИК.

Итоговые электоральные данные, находящиеся в свободном доступе, размещены на сайте ЦИК РФ [5]. Эти данные имеют иерархическую структуру: вершина дерева — данные ЦИК (сводные электоральные данные по всей РФ), ветви первого уровня — данные иК субъектов РФ (сводные электоральные данные по соответствующему субъекту РФ), ветви второго уровня — данные территориальных ИК (сводные данные по территориальным ИК соответствующих субъектов РФ, ветви третьего уровня — данные участковых ИК. Информационные модели электоральных данных, размещаемых на вершине, а также первом и втором уровнях описанной выше иерархической структуры, одинаковы. Они имеют следующую структуру полей:

- 1. Название ИК.
- 2. Число избирательных комиссий на следующем уровне иерархии.
- 3. Число избирательных комиссий, приступивших к передаче информации.
- 4. Дата и время подписания протокола соответствующей ИК.
- Число избирателей, включенных в список избирателей.
- 6. Число избирательных бюллетеней, полученных участковой избирательной комиссией.
- 7. Число избирательных бюллетеней, выданных избирателям, проголосовавшим досрочно.
- 8. Число избирательных бюллетеней, выданных в помещении для голосования в день голосования.
- 9. Число избирательных бюллетеней, выданных вне помещения для голосования в день голосования
  - 10. Число погашенных избирательных бюллетеней.
- 11. Число избирательных бюллетеней в переносных ящиках для голосования.
- 12. Число бюллетеней в стационарных ящиках для голосования.
- 13. Число недействительных избирательных бюллетеней.
- 14. Число действительных избирательных бюллетеней
  - 15. Число утраченных избирательных бюллетеней
- 16. Число избирательных бюллетеней, не учтенных при получении.
- 17. Число избирательных бюллетеней, поданных за Бабурина Сергея Николаевича.

- 18. Число избирательных бюллетеней, поданных за Грудинина Павла Николаевича.
- 19. Число избирательных бюллетеней, поданных за Жириновского Владимира Вольфовича.
- 20. Число избирательных бюллетеней, поданных за Путина Владимира Владимировича.
- 21. Число избирательных бюллетеней, поданных за Собчак Ксению Анатольевну.
- 22. Число избирательных бюллетеней, поданных за Сурайкина Максима Александровича.
- 23. Число избирательных бюллетеней, поданных за Титова Бориса Юрьевича.

- 24. Число избирательных бюллетеней, поданных за Явлинского Григория Алексеевича.
- В информационной модели структуры электоральных данных на третьем уровне иерархии отсутствует поле № 2. Для отображения электоральных данных используются древовидно организованные html-страницы. Пример фрагмента отображения электоральных данных на html-странице, предоставленных ЦИК ИК Республики Адыгея, представлен на рисунке 1.

Отметим, что сайт ЦИК [4] не предоставляет заинтересованным пользователям возможностей для

### Выборы Президента Российской Федерации

Назад к списку

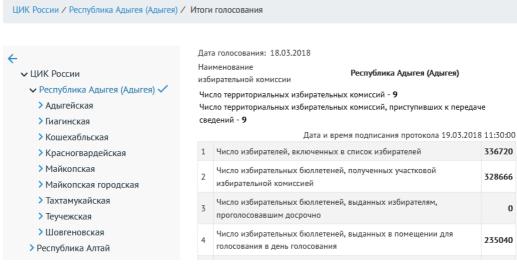


Рис. 1. Фрагмент html-страницы

выгрузки всех или части данных в пригодном для дальнейшего анализа виде. В этих условиях извлечь требующиеся электоральные данные оказывается возможным только в ручном режиме, открывая соответствующие html-страницы сайта ЦИК, выделяя электоральные данные на открытой странице и далее экспортируя их через буфер обмена в тот или иной файл. Для экспорта всех электоральных данных, размещенных на сайте ЦИК, потребуется открыть более 97 тысяч html-страниц и далее выполнить описанную выше последовательность действий, что потребует значительных временных затрат. Кроме того, в ходе экспорта электоральных данных, размещенных на сайте ЦИК, вручную, нельзя исключить внесения в них случайных ошибок, а потому потребуется последующая проверка экспортированных данных, автоматизация которой является самостоятельной нетривиальной задачей. В этой связи авторами был разработан программный инструмент, обеспечивший автоматизированную выгрузку электоральных данных о результатах выборов Президента РФ в 2018 г., представленных ИК различных уровней, с сайта ЦИК [4].

# 3. Технология экспорта электоральных данных о результатах выборов Президента РФ в 2018 г. с сайта ЦИК

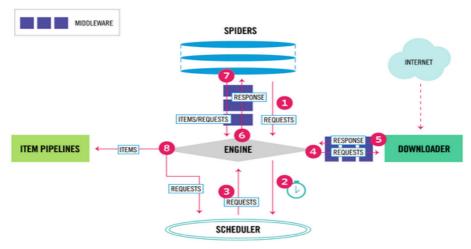
В связи с тем, что доступ к электоральным данным, размещенным на сайте ЦИК [4], как в ручном, так и автоматическом режимах, можно получить либо в процессе обхода всего сайта по заранее заданному алгоритму, либо открывая ссылки, найденные внутри html-страниц, было принято решение использовать технологию веб-скрапинга (англ. – scraping) или сайтскрапинг [5].

Отметим, что разрабатываемый программный инструмент, основанный на выбранной технологии, кроме собственно подключения к удалённому серверу, получения и анализа текста html-страницы должен обеспечивать:

 управление нагрузкой и пропускной способностью с целью обеспечения частоты автоматически выполняемых запросов максимально близкой к частоте запросов, отправляемых обычным пользователем, чтобы на удалённом сервере не включились механизмы ограничения пропускной способности, известные как троттлинг или дросселирование [6], а также не произошла блокировка запросов с данного IP адреса;

- при получении ошибок сервера автоматическую повторную отправку и обработку запросов;
- возможность работы длительный срок без утечек памяти при работе со строками;
- поддержку различных механизмов выборки данных, извлекаемых из запрошенных страниц.

Для реализации разрабатываемого программного инструмента для автоматического излечения электоральных данных, размещенных на сайте ЦИК [4], был выбран один из наиболее на популярном на сегодняшний день инструментарий Scrapy (набор утилит и базовых классов), используемый для извлечения данных из веб-страниц, написанный на языке Python [7]. Инструментарий Scrapy, имеющий асинхронную конвейерную архитектуру, представленную на рисунке 2, позволяет обрабатывать необходимое количество запросов, не перегружая при этом удалённый сервер и не предъявляя существенных требований для клиентского места.



Puc. 2. Архитектура Scrapy [7]: Middleware(response/request/spiders) – пользовательский код осуществляющий обработку данных, engine/item pipelines/scheduler – механизмы распределения нагрузки, downloader – механизм загрузки данных по протоколу http

Из рисунка 2 видно, что архитектура инструментария Scrapy удовлетворяет требованиям к искомому инструменту и обеспечивает базовый набор функциональности.

Единицей функциональности в инструментарии Scrapy является объект Spider – некоторая часть кода (класс), унаследованного от базового класса, в котором пользователь должен реализовать 2 механизма:

- 1. механизм поиска и обхода ссылок;
- 2. механизм извлечения данных.

Запросы из Spider поступают в механизм обработки и с помощью планировщика управляемым способом скачивают данные из внешних источников и передают обратно на обработку.

Обработка загруженных данных, представленных в виде объекта в оперативной памяти, выполняется при помощи механизма селекторов [8]. Доступны два вида селекторов:

- 1. CSS-селекторы, реализуемые на языке CSS селекторов [9];
- 2. XPath-селекторы, реализуемые на языке XML Path [10]

Рассмотрим алгоритм работы сайта:

- 1. По нажатию на элемент дерева, происходит запрос на сервер по ссылке http://www.vybory.izbirkom.ru/region/izbirkom?action=show&root=1&tvd=1001000 84849067&vrn=100100084849062&prver=0&region=1& sub\_region= 1&type=226, содержащей идентификаторы выбранного в дереве элемента.
- 2. После получения страницы браузером, в правой части экрана отображается таблица с данными, представленными при помощи html-атрибута (рис. 1).
- 3. При разворачивании узла дерева на сервер уходит Ајах-запрос для получения дочерних элементов. Адреса всех запросов для получения дочерних элементов задаются в соответствии со следующим шаблоном: {root}/region/izbirkom?action=tvdTree&tvd children=true&vrn={vrn}&tvd={tvd}, где
  - 1. root http://www.vybory.izbirkom.ru
- 2. vrn и tvd идентификаторы из ссылки на текущую страницу

Таким образом, алгоритм получения данных, записанный на псевдоязыке, реализуется выполнением следующей последовательности действий:

1. Начинаем со первой страницы.

- 2. Извлекаем из страницы коды vrn и tvd.
- 3. На основе этих кодов формируем URL для получения данных дерева.
- 4. Из полученных JSON данных (см. рисунок 3) извлекаем ссылки на страницы нижних уровней.
- 5. Создаём объект scrapy.Request для загрузки данных по полученному URL.

6. При получении данных ищем в них при помощи XPath-запроса //table[@class='table-bordered table-striped table-sm']/tr таблицу с данными и извлекаем их в словарь проиндексированный идентификаторами строк таблицы.

Далее на языке Python была разработана программная реализация вышеуказанного псевдоалго-

```
"id": 100100084849066,
"text": "ЦИК России",
"href": "region/izbirkom?action=show&root=0&tvd=100100084849066&vrn=100100084849062&prver=0
"prver": "0",
"isUik": false,
"selected": true.
"load_on_demand": false,
"children": [{
       "id": 100100084849067,
        "text": "Республика Адыгея (Адыгея)",
        "href": "region/izbirkom?action=show&root=1&tvd=100100084849067&vrn=100100084849062
        "prver": "0",
        "isUik": false
        "selected": false,
        "load on demand": true,
        "children": []
    }, {
        "id": 100100084849068,
        "text": "Республика Алтай",
        "href": "region/izbirkom?action=show&root=1&tvd=100100084849068&vrn=100100084849062
        "prver": "0",
        "isUik": false,
        "selected": false,
        "load_on_demand": true,
        "children": []
        "id": 100100084849069,
        "text": "Республика Башкортостан",
        "href": "region/izbirkom?action=show&root=1&tvd=100100084849069&vrn=100100084849062
```

Рис. 3. Фрагмент представление дочерних элементов в виде JSON

### Таблица 1

### Описание данных

Наименование колонки	Тип данных	Описание
l1	Строка	наименование ИК субъекта РФ
I2	Строка	Наименование территориальной ИК субъекта РФ
13	Строка	Наименование участковой ИК 3 уровня, сформированной территориальной ИК субъекта РФ
c1	Целое число	Число избирателей, включенных в список избирателей
c2	Целое число	Число избирательных бюллетеней, полученных участковой избирательной комиссией
c3	Целое число	Число избирательных бюллетеней, выданных избирателям, проголосовавшим досрочно
c4	Целое число	Число избирательных бюллетеней, выданных в помещении для голосования в день голосования
c5	Целое число	Число избирательных бюллетеней, выданных вне помещения для голосования в день голосования
с6	Целое число	Число погашенных избирательных бюллетеней
с7	Целое число	Число избирательных бюллетеней в переносных ящиках для голосования
c8	Целое число	Число бюллетеней в стационарных ящиках для голосования
с9	Целое число	Число недействительных избирательных бюллетеней

Наименование колонки	Тип данных	Описание
c10	Целое число	Число действительных избирательных бюллетеней
c11	Целое число	Число утраченных избирательных бюллетеней
c12	Целое число	Число избирательных бюллетеней, не учтенных при получении
c13	Целое число	Бабурин Сергей Николаевич
c14	Целое число	Грудинин Павел Николаевич
c15	Целое число	Жириновский Владимир Вольфович
c16	Целое число	Путин Владимир Владимирович
c17	Целое число	Собчак Ксения Анатольевна
c18	Целое число	Сурайкин Максим Александрович
c19	Целое число	Титов Борис Юрьевич
c20	Целое число	Явлинский Григорий Алексеевич



Рис. 4. Фрагмент скриншота Excel с открытым CSV-файлом

ритма, запускаемая с помощью утилиты runspider. Haпример, так: runspider cik/spiders/data\_spider.py.

## Результатов автоматической выгрузки с сайте ЦИК результатов выборов Президента РФ в 2018 г.

Для автоматической выгрузки электоральных данных, представленных ИК после выборов Президента РФ в 2018 г., с сайта ЦИК [4] был использован обычный офисный компьютер (AMD Ryzen 5 4 ГГц, 32 Гб ОЗУ, Интернет со скоростью 50Мб/с. На выгрузку данных потребовалось 8 часов, в течение которых было выполнено 201156 http-запросов и скачано около 10 ГБ сырых данных. Далее из первичных («сырых») данных извлечено 11Мб полезных данных, которые были сохранены в CSV-файл (в формате RFC 4180) со следующей структурой (таблица 1).

Фрагмент скриншот программы Excel с открытым файлом, содержащим результаты выгрузки электоральных данных, представленных ИК после выборов Президента РФ в 2018 г., подтверждающий работоспособность разработанного программного инструмента приведен на рисунке 4. Данный файл содержит 100577 строк.

Для верификации результатов выгрузки электоральных данных, представленных ИК после выборов Президента РФ в 2018 г., выгружаемые данные также в текстовом файле, объемом 392 МБ. Результаты выборочного сравнения соответствующих данных в CSV и текстовом файлах показали их идентичность.

#### Заключение

Разработан программный инструмент и реализована автоматическая выгрузка электоральных данных о выборах в 2018 г. Президента Российской Федерации (РФ), размещенных на официальном сайте ЦИК, включая: сводные данные по РФ; сводные данные, представленные избирательными комиссиями субъектов РФ; сводные данные, представленные территориальными избирательными комиссиями (ТИК), созданными в субъектах РФ; данные, представленные участковыми избирательными комиссиями (УИК), относящимися к соответствующим ТИК, и ее размещение в виде таблицы (размером 21 столбец на 100757 строк) в файле формата CSV.

Далее электоральные данные о выборах Президента РФ в 2018 г. Будут использованы для оценки адекватности статистических методов, используемых в электоральной криминалистике для выявления фальсификации результатов выборов.

### Литература

- 1. Доклад Российского общественного института избирательного права (РОИИП) «Математические инструменты делигитимации выборов»//И.Б. Борисов, И.В Задорин, А.В. Игнатов, В.Н. Марачевский, В.И. Федоров/ –М.: РОИИП, 2020. 76 с. URL: http://www.roiip.ru/images/data/gallery/0\_299\_Matematicheskie\_instrumenti\_delegitimatsii\_viborov.pdf (дата обращения 19.02.2022).
- 2. Шень А. Как доклад РОИИП делегитимирует выборы. URL: https://trv-science.ru/2020/09/kak-doklad-roiip-delegitimiruet-vybory/ (дата обращения 19.02.2022).

- 3. URL:http://electoralpolitics.org/ru/articles/vozmozhnosti-matematicheskikh-metodov-po-vyjavleniju-elektoralnykh-falsifikatsii/
  - 4. (дата обращения 19.02.2022).
  - 5. URL: http://www.vybory.izbirkom.ru (дата обращения 19.02.2022).
  - 6. Райн М. Скрапинг веб-сайтов с помощью Python. М.: ДМК Пресс, 2016. 280 с.
  - 7. URL: https://www.nginx.com/blog/rate-limiting-nginx/
  - 8. URL: https://scrapy.org/ (дата обращения 19.02.2022).
  - 9. URL: https://docs.scrapy.org/en/latest/topics/selectors.html (дата обращения 19.02.2022).
  - 10. URL: https://developer.mozilla.org/ru/docs/Web/CSS/CSS\_Selectors (дата обращения 19.02.2022).
  - 11. URL: https://www.w3.org/TR/xpath/all/ (дата обращения 19.02.2022).

#### References

- 1. Doklad Rossiyskogo obshchestvennogo instituta izbirateľnogo prava (ROIIP) «Matematicheskiye instrumenty deligitimatsii vyborov»//I.B. Borisov, I.V Zadorin, A.V. Ignatov, V.N. Marachevskiy, V.I. Fedorov/—M.: ROIIP, 2020. 76 s. URL: http://www.roiip.ru/images/data/gallery/0\_299\_Matematicheskie\_instrumenti\_delegitimatsii\_viborov.pdf (data obrashcheniya 19.02.2022).
- 2. Shen' A. Kak doklad ROIIP delegitimiruyet vybory. URL: https://trv-science.ru/2020/09/kak-doklad-roiip-delegitimiruet-vybory/ (data obrashcheniya 19.02.2022).
- 3. URL: http://electoralpolitics.org/ru/articles/vozmozhnosti-matematicheskikh-metodov-po-vyiavleniiu-elektoralnykh-falsifikatsii/ (data obrashcheniya 19.02.2022).
  - 4. URL: http://www.vybory.izbirkom.ru (data obrashcheniya 19.02.2022).
  - 5. Rayn M. Skraping veb-saytov s pomoshch'yu Python. –M.: DMK Press, 2016. 280 s.
  - 6. URL: https://www.nginx.com/blog/rate-limiting-nginx/
  - 7. URL: https://scrapy.org/ (data obrashcheniya 19.02.2022).
  - 8. URL: https://docs.scrapy.org/en/latest/topics/selectors.html (data obrashcheniya 19.02.2022).
  - 9. URL: https://developer.mozilla.org/ru/docs/Web/CSS/CSS\_Selectors (data obrashcheniya 19.02.2022).
  - 10. URL: https://www.w3.org/TR/xpath/all/ (data obrashcheniya 19.02.2022)

**МИРВОДА Сергей Геннадьевич,** заместитель генерального директора по технологиям ООО «Октоника». 620014, г. Екатеринбург ул. 8 марта 2. E-mail: sergey@mirvoda.com

**MIRVODA Sergey Gennadievch,** CTO, Octonica LLL, 8th March St., 2. 620014, Yekaterinburg, Russia. E-mail: sergey@mirvoda.com

**ПОРШНЕВ Сергей Владимирович,** доктор технических наук, профессор, директор Учебно-научного центра «Информационная безопасность» федерального государственного автономного образовательного учреждения высшего образования «Уральский федеральный университет им. первого Президента России Б.Н. Ельцина». 620002, г. Екатеринбург, ул. Мира, 32. E-mail: s.v.porshnev@urfu.ru

**PORSHNEV Sergey Vladimirovich,** Doctor of Technical Sciences, Professor, Director of the Educational and Scientific Center «Information Security» of the Federal State Autonomous Educational Institution of Higher Education «Ural Federal University named after the first President of Russia B.N. Yeltsin». 620002, Yekaterinburg, st. Mira, 32. E-mail: s.v.porshnev@urfu.ru

**РЯБКО Николай Юрьевич,** аспирант федерального государственного автономного образовательного учреждения высшего образования «Уральский федеральный университет им. первого Президента России Б.Н. Ельцина». 620002, г. Екатеринбург, ул. Мира, 32. E-mail: N. Yu.Ryabko@urfu.ru

**RYABKO Nikolay Yurievich,** post-graduate student of the Federal State Autonomous Educational Institution of Higher Education "Ural Federal University named after the first President of Russia B.N. Yeltsin". 620002, Yekaterinburg, st. Mira, 32. E-mail: N.Yu.Ryabko@urfu.ru