

**ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ
INFORMATION SYSTEM AND TECHNOLOGIES**

УДК 004.62,004.42

DOI: 10.18413/2518-1092-2025-10-3-0-1

Резниченко О.С.

**МЕТОДЫ И ПРОГРАММНЫЕ ИНСТРУМЕНТЫ
ДЛЯ ИЗВЛЕЧЕНИЯ ПУБЛИКАЦИОННЫХ ДАННЫХ
НАУЧНОЙ ОРГАНИЗАЦИИ, РАЗМЕЩЕННЫХ
НА ПОРТАЛЕ ELIBRARY.RU**Белгородский государственный национальный исследовательский университет,
ул. Победы, 85, г. Белгород, 308015, Россия

e-mail: oreznichenko@bsuedu.ru

Аннотация

Публикационные отчеты занимают важную долю среди всей отчетности научной организации. Они нужны не только для текущего мониторинга, но и для планирования будущей деятельности. Отчетность включает в себя в том числе информацию о количестве и списочном составе публикаций в различных разрезах и типах, например, РИНЦ, публикации из RSCI, журналов из списка ВАК, журналов из «Белого списка» и других. Сейчас основной источник данных для отчетов – это научная электронная библиотека eLIBRARY.RU. Списки публикаций из профиля организации в eLIBRARY.RU можно получить путем ручной выгрузки публикационных данных, используя API, используя возможности аналитической системы SCIENCE INDEX или парсинге веб-содержимого. В статье описывается разработка алгоритмов и инструментов выгрузки и парсинга содержимого страниц, которые позволяют получать сведения о названиях, списках авторов и источниках публикаций в табличном виде. Автоэкспорт веб-страниц и их анализ основан на использовании эмуляции работы пользователя в браузере и использовании серии опорных тегов для парсинга. Созданные скрипты можно скомпилировать в исполняемый файл и сократить время сохранения и преобразования веб-страниц eLIBRARY.RU на 5 минут для каждых 100 статей поисковой выдачи при ручной обработке.

Ключевые слова: наукометрия; публикационная активность; публикационная отчетность; веб-скрейпинг; парсинг; eLIBRARY.RU; Microsoft Excel; SCIENCE INDEX; Selenium WebDriver; Python; BeautifulSoup; pandas.DataFrame

Для цитирования: Резниченко О.С. Методы и программные инструменты для извлечения публикационных данных научной организации, размещенных на портале eLIBRARY.RU // Научный результат. Информационные технологии. – Т.10, №3, 2025. – С. 3-19. DOI: 10.18413/2518-1092-2025-10-3-0-1

Reznichenko O.S.

**METHODS AND SOFTWARE TOOLS
FOR UPLOADING PUBLICATION DATA OF A SCIENTIFIC
ORGANIZATION POSTED ON THE ELIBRARY.RU**Belgorod State National Research University,
85 Pobedy St., Belgorod, 308015, Russia

e-mail: oreznichenko@bsuedu.ru

Abstract

Publication reports are an important part of the overall reporting process in a scientific organization. They are necessary not only for monitoring the organization's status but also for planning future activities. Publication reports include information on the count and list of publications in various formats, such as articles from the Russian Science Citation Index, articles in journals from the HAC

list, articles in journals from the "White List," and many others. Currently, the main source of input data for publication reports is the scientific electronic library eLIBRARY.RU. Lists of publications from the organization's profile on eLIBRARY.RU can be manually uploaded, through the API, using tools from the SCIENCE INDEX analytical system, or via web scraping techniques. This article describes the development of algorithms and software tools for extracting and scraping page content, which allow for the retrieval of data about titles, author lists, and journal information in spreadsheet form. The automatic saving of web pages and their scraping is based on emulation browser user actions and the use of specific HTML tags. The created Python scripts can be compiled into standalone executable files, reducing the time required to upload and transform the content of eLIBRARY.RU web pages into spreadsheets by 5 minutes for every 100 articles in search results during manual processing.

Keywords: scientometrics; publication activity; publication reports; web scraping; eLIBRARY.RU; Microsoft Excel; SCIENCE INDEX; Selenium WebDriver; Python; BeautifulSoup; pandas.DataFrame

For citation: Reznichenko O.S. Methods and software tools for uploading publication data of a scientific organization posted on the eLIBRARY.RU // Research result. Information technologies. – T.10, №3, 2025. – P. 3-19. DOI: 10.18413/2518-1092-2025-10-3-0-1

ВВЕДЕНИЕ

Важную роль среди отчетов научной организации о своей деятельности является публикационная отчетность. Показатели публикационной активности необходимы для принятия управленческих решений, так как на основе этих показателей Минобрнауки определяет грантовое финансирование, а также объемы финансирования в рамках программы «Приоритет-2030». В Белгородском государственном национальном исследовательском университете (далее – БелГУ, университет) с различной периодичностью приходится формировать несколько видов публикационных отчетов. Среди внешней публикационной отчетности можно упомянуть следующие её виды:

- отчеты о количестве публикаций в наукометрических реферативных базах Scopus и Web of Science Core Collection в рамках выполнения реализации программы стратегического академического лидерства «Приоритет-2030» [1];
- отчеты о деятельности научно-образовательного центра мирового уровня «Инновационные решения в АПК» [2];
- отчеты о составе публикаций, посвященных вопросам обеспечения условий для обучения и социализации инвалидов и лиц с ограниченными возможностями здоровья в рецензируемых научных журналах;
- списки монографий, глав монографий и словарей;
- список публикаций, посвященных вопросам реализации бережливого производства.

В качестве внутренней периодической отчетности выступают следующие публикационные отчеты:

- состав и количество статей университета в журнал из «Белого списка» [3];
- количество публикаций структурных подразделений университета в реферативных базах Scopus и Web of Science Core Collection [4];
- количество и состав публикаций отдельных сотрудников университета в реферативных базах Scopus, Web of Science Core Collection.

Кроме этого, периодический мониторинг количества публикаций в различных разрезах позволяет строить прогнозные модели на будущие периоды и формировать на их основе как планы публикационной активности конкретных научных подразделений, так и план публикационной активности в целом по университету.

Большую часть требуемых публикационных отчетов можно получить, анализируя списки публикаций, размещенных в научной электронной библиотеке eLIBRARY.RU (далее – elibrary.ru)

[5]. Профиль научной организации на сайте elibrary.ru позволяет отображать списки аффилированных публикаций согласно следующим актуальным для университета фильтрам:

- публикации организации, включенные в РИНЦ [6];
- публикации, имеющие DOI;
- статьи в журналах, входящих в Russian Science Citation Index [7];
- статьи в журналах, включенных в «Белый список»;
- статьи в российских журналах, включенных в перечень ВАК [8].

Сформированный после применения соответствующих фильтров список публикаций на портале elibrary.ru представляет собой последовательность веб-страниц на каждой из которых отображается до 100 публикаций при условии осуществления авторизованного входа на портал (рис. 1).

Всего найдена 261 публикация с общим количеством цитирований: 74. Показано на данной странице: с 1 по 100.		
№	Публикация	Цит.
1.	ГЕНЕТИЧЕСКИЕ ДЕТЕРМИНАНТЫ УРОВНЯ АНГИОТЕНЗИНПРЕВРАЩАЮЩЕГО ФЕРМЕНТА (ДАННЫЕ ПОЛНОГЕНОМНЫХ ИССЛЕДОВАНИЙ) ☐ Камышникова Л.А., Ефремова О.А., Фентисов В.В., Болховитина О.А., Чурнов М.И. ↓ Артериальная гипертензия. 2024. Т. 30. № 6. С. 537-552.	0
2.	СРАВНИТЕЛЬНАЯ МЕТАГЕНОМИКА ДЛЯ МОНИТОРИНГА СКРЫТОЙ ДИНАМИКИ АЛЬГО-БАКТЕРИАЛЬНОГО СООБЩЕСТВА СТОЧНЫХ ВОД ПОД ДЕЙСТВИЕМ ЛЕКАРСТВЕННЫХ ВЕЩЕСТВ ☐ Зайцев П.А., Шурыгин Б.М., Родин В.А., Панова Т.В., Зверева М.Э., Скрипникова Е.В., Соловченко А.Е. ↓ Российские нанотехнологии. 2024. Т. 19. № 4. С. 505-519.	0
3.	РАДИОНУКЛИДЫ (^{40}K, ^{232}Th, ^{226}Ra, ^{137}Cs, ^{90}Sr) В АГРОЭКОСИСТЕМАХ ЦЕНТРАЛЬНОГО ЧЕРНОЗЕМЬЯ ☐ Лукин С.В. ↓ Доклады Российской академии наук. Науки о Земле. 2024. Т. 519. № 2. С. 156-162.	0
4.	РАЗЛИЧИЕ СПЕКТРОВ ПАРАМЕТРИЧЕСКОГО РЕНТГЕНОВСКОГО ИЗЛУЧЕНИЯ ИЗ ПОРОШКОВЫХ МИШЕНЕЙ И ТЕКСТУРИРОВАННЫХ ФОЛЫГ ☐ Кишин И.А., Киданова Е.Ю., Кубанкин А.С., Нажмудинов Р.М., Алексеев В.И., Елисеев А.Н. Письма в журнал Физика элементарных частиц и атомного ядра. 2024. Т. 21. № 4 (255). С. 782-786.	0
5.	ИЗМЕРЕНИЯ ПОЛЯРИЗАЦИИ ПУЧКОВ ДЕЙТРОНОВ И ПРОТОНОВ НА НУКЛОТРОНЕ ☐ Волков И.С., Ладыгин В.П., Схоменко Я.Т., Гурчин Ю.В., Исупов А.Ю., Янек М., Карачук Ю.Т., Хренов А.Н., Курилкин П.К., Ливанов А.Н., Пиядин С.М., Резников С.Г., Терёхин А.А., Тишевский А.В., Аверьянов А.В., Черных Е.В., Енаше Д., Кривенков Д.О., Внуков И.Е. Письма в журнал Физика элементарных частиц и атомного ядра. 2024. Т. 21. № 4 (255). С. 732.	0
6.	ЭКОЛОГИЧЕСКАЯ ОЦЕНКА СОДЕРЖАНИЯ ТЯЖЕЛЫХ МЕТАЛЛОВ И МЫШЬЯКА В ПОЧВАХ И СЕЛЬСКОХОЗЯЙСТВЕННЫХ РАСТЕНИЯХ ЦЕНТРАЛЬНОГО ЧЕРНОЗЕМЬЯ ☐ Лукин С.В. ↓ Почвоведение. 2024. № 10. С. 1414-1423.	0
	Версии: ECOLOGICAL ASSESSMENT OF CONCENTRATIONS OF HEAVY METALS AND ARSENIC IN SOILS AND CROPS OF THE CENTRAL CHERNOZEMIC REGION Lukin S.V. Eurasian Soil Science. 2024. Т. 57. № 10. С. 1709-1717.	0

Рис. 1. Часть внешнего вид профиля научной организации в режиме отображения списка публикаций

Fig. 1. User interface of a scientific organization profile in the publication list display mode

ОБЗОР МЕТОДОВ ИЗВЛЕЧЕНИЯ ПУБЛИКАЦИОННЫХ ДАННЫХ ИЗ ELIBRARY.RU

Среди возможных способов получения отфильтрованного списка публикаций в удобном для анализа виде (например, в виде электронной таблицы) авторы предполагают следующие варианты:

- 1) сохранение веб-страниц с результатами поиска публикаций, открытие их в Microsoft Office Excel (далее – Excel) с последующей обработкой содержимого этих веб-страниц;
- 2) использование возможностей интерфейса прикладного программирования API [9];
- 3) использование возможностей информационно аналитической системы SCIENCE INDEX (далее – система Science Index) [10];
- 4) синтаксический анализ списка публикаций на портале онлайн;
- 5) синтаксический анализ сохраненных списков публикаций с портала в виде выгруженных веб-страниц.

Рассмотрим достоинства и недостатки каждого из вариантов решения задачи.

Среди достоинств первого метода можно указать отсутствие необходимости использовать какие-либо программные инструменты, кроме интернет-браузера и процессора электронных таблиц. Главным недостатком этого метода является объем временных затрат, который составляет по оценкам авторов не менее 300 секунд на каждые 100 статей при условии уверенного владения исполнителем основами работы в процессорах электронных таблиц. Оценочное время получено авторами на персональном компьютере на базе процессора Intel Intel Core i7-4770 с 16 Гб оперативной памяти с операционной системой Microsoft Windows 11 Pro (сборка 26100.3194), посредством браузера Opera One (версия: 117.0.5408.39) и процессора электронных таблиц Microsoft Office Excel 2016. Следует также учесть, что для публикаций с большим количеством соавторов не все они попадут в соответствующий столбец, так как представлены в урезанном виде на исходной веб-странице.

Использование встроенного в elibrary.ru API позволяет выгружать данные о публикациях научной организации в формат xml. Вопросы получения информации о публикационной активности научной организации посредством встроенного API рассмотрены в работе [11]. Данный способ доступен только научным организациям, имеющим, кроме доступа к системе Science Index, дополнительную лицензию, дающую право на увеличение спектра и количества сведений, которые можно выгружать посредством API [9].

В БелГУ имеется лицензия на доступ к системе Science Index, предполагающая в том числе наличие неограниченного экспорта списка публикаций в формате XML. Вопросы разработки и использования инструментов преобразования данных, выгруженных в формате xml, в электронную таблицу формата Excel рассмотрены в работе [12]. Содержимое выходного файла, полученного в результате применения этого инструмента легко можно анализировать и подвергать дальнейшей трансформации для получения сведений о публикациях в различных разрезах. В тоже время, ориентация на использования данного метода сопряжено с несколькими особенностями, ограничивающими его применение:

- обязательное наличие платной лицензии на доступ к системе Science Index;
- необходимость каждый раз адаптировать входные данные, полученные на этапе их подготовки, для корректной работы этого инструментария;
- некорректной работе функции экспорта в xml на портале elibrary.ru при использовании некоторых типов фильтров (например, для фильтра статей в журналах, включенных в «Белый список»), что приводит к невозможности получить актуальное количество публикаций;
- нестабильной работе функции фильтрации публикаций научной организации по некоторым фильтрам. Периодически при фильтрации может возникать ошибка типа «Microsoft OLE DB Provider for ODBC Drivers error '80040e31'».

Онлайн синтаксический разбор веб-страницы с результатами фильтрации списка публикаций научной организации также связан с несколькими трудностями:

- необходимо устанавливать и настраивать специализированные библиотеки, эмулирующие работу интернет-браузеров, так как elibrary.ru не формирует веб-страницы с результатами поиска на основе сгенерированной адресной ссылки, в структуру которой входят параметры поискового запроса, что делает невозможным пакетно обрабатывать эти поисковые результаты программными инструментами. Кроме того, портал elibrary.ru имеет встроенную защиту от парсинга, которую можно обойти только используя эмуляцию действий пользователя в интернет-браузере, например, посредством Selenium WebDriver. Пример разработки подобного инструментария для парсинга содержимого профилей авторов рассматривается в исследовании [13];

- имеющийся инструмент парсинга содержимого веб-страниц портала elibrary.ru, описанный в источнике [14] также предназначен только для экспорта списка публикаций конкретного автора по его уникальному идентификатору, и, аналогично, предполагает использование Selenium WebDriver.

Видно, что описанные выше способы реализации метода онлайн парсинга предполагают наличие у исполнителя навыков владения информационными технологиями выше среднего, так как сопровождаются настройкой и использованием дополнительных программных инструментов. Кроме того, скрипты, описанные в источниках выше, предполагают значительную их доработку для извлечения списка публикаций именно из профиля научной организации, включающую разработку сценария эмуляции действий пользователя для Selenium WebDriver. При этом процесс эмуляции, как и в случае ручной навигации по параметрам фильтров и их результатов может сопровождаться ошибками доступа к базе данных elibrary.ru.

В качестве альтернативы использования эмуляторов браузеров может выступать только ручная постраничная загрузка результатов поиска в отдельные html-файлы. Подобная работа может быть выполнена не менее чем за 30 секунд на каждые 100 статей поисковой выдачи.

Подытоживая все вышесказанное, автор предлагает остановиться на реализации метода, предполагающего автоматическое сохранение веб-страниц со списком публикаций и синтаксический анализ этих списков, так как остальные методы либо занимают много времени, либо привязаны к наличию лицензии на использование API.

ФОРМАТ И СТРУКТУРА ВХОДНЫХ И ВЫХОДНЫХ ДАННЫХ

В качестве исходных данных для разрабатываемого программного обеспечения выступают наборы html-файлов, представляющие собой сохраненные вручную через браузер веб-страницы с результатами фильтрации списка публикаций научной организации (рис. 2). Из рисунка очевидно, что, анализируя текст html-файла можно получить табличные данные, имеющие структуру, представленную в таблице 1.

```

2081 <td align="left" valign="top">
2082 <span style="overflow-wrap: break-word;">
2083
2084 <a href="https://elibrary.ru/item.asp?id=80303583"><b><span style="line-height:1.0;">КОНСТИТУЦИОННЫЕ КРИТЕРИИ
ЭФФЕКТИВНОСТИ ПУБЛИЧНОЙ ВЛАСТИ В РОССИИ</span></b></a><br><font color="#00008f"><i>Мархгейм М.В., Бойченко Е.М.
</i></font><br>
2085 <font color="#00008f">
2086
2087
2088 <a href="https://elibrary.ru/contents.asp?id=80303514">Право и государство: теория и практика</a>.
2089 2024. <a href="https://elibrary.ru/contents.asp?id=80303514&selid=80303583">№11 (239)</a>. С.
2090 248-251.
2091 </font></span><font color="#00008f">
2092 </font></td><td align="center" valign="middle" class="select-tr-right">
2093 0</td></tr>
2094
2095
2096
2097 <tr valign="middle" id="arw80303581">
2098 <td align="center" valign="top" class="select-tr-left"><font color="#00008f"><b>29.</b></font><br><input type=
"checkbox" name="tid80303581" id="tid80303581" onclick="if (this.checked) {RowSelect(&quot;arw80303581&quot;;,1)}
else {RowSelect(&quot;arw80303581&quot;;,0)}"><div id="pdf_80303581" name="pdf_80303581" style=""><a href=
"javascript:load_article(80303581)"></a>
2099 </div></td>
2100 <td align="left" valign="top">
2101 <span style="overflow-wrap: break-word;">
2102
2103 <a href="https://elibrary.ru/item.asp?id=80303581"><b><span style="line-height:1.0;">К ВОПРОСУ О КОНСТИТУЦИОННОЙ
ОСНОВЕ РОДИТЕЛЬСКИХ ПРАВ (ОБЯЗАННОСТЕЙ) И ФОРМАЛИЗАЦИИ КАТЕГОРИИ "РОДИТЕЛЬ"</span></b></a><br><font color=
"#00008f"><i>Комнатная Ю.А., Котарева О.В., Чистыхина М.В.</i></font><br>
2105 <font color="#00008f">
2106
2107
2108 <a href="https://elibrary.ru/contents.asp?id=80303514">Право и государство: теория и практика</a>.
2109 2024. <a href="https://elibrary.ru/contents.asp?id=80303514&selid=80303581">№11 (239)</a>. С.
2110 240-243.
2111 </font></span><font color="#00008f">
2112 </font></td><td align="center" valign="middle" class="select-tr-right">
2113 0</td></tr>

```

Рис. 2. Пример содержимого входного html-файла для анализа
Fig. 2. Example of the contents of the input HTML file for analysis

Таблица 1

Структура табличного файла с выходными данными

Table 1

Structure of the output spreadsheet file

Название полей	Описание назначения полей и их типы
IDs	идентификатор публикации на портале elibrary.ru, целое
Авторы	список авторов, разделенный запятыми, строка
Название	название публикации, строка
Источник	название источника, год издания, номер тома и выпуска, диапазон страниц расположения статьи в источнике
URL	интернет-ссылка на публикацию на портале elibrary.ru, строка

Для задач разработки интерес представляет в первую очередь html-код выгруженных страниц, а именно текстовое содержимое внутри цепочки определенных тегов.

Таким образом разрабатываемое программное обеспечение должно уметь автоматически сохранять информацию о публикациях научной организации из портала elibrary.ru в наборы веб-страниц формата *.html (рис. 3), а также преобразовывать её в один файл электронных таблиц формата Excel.

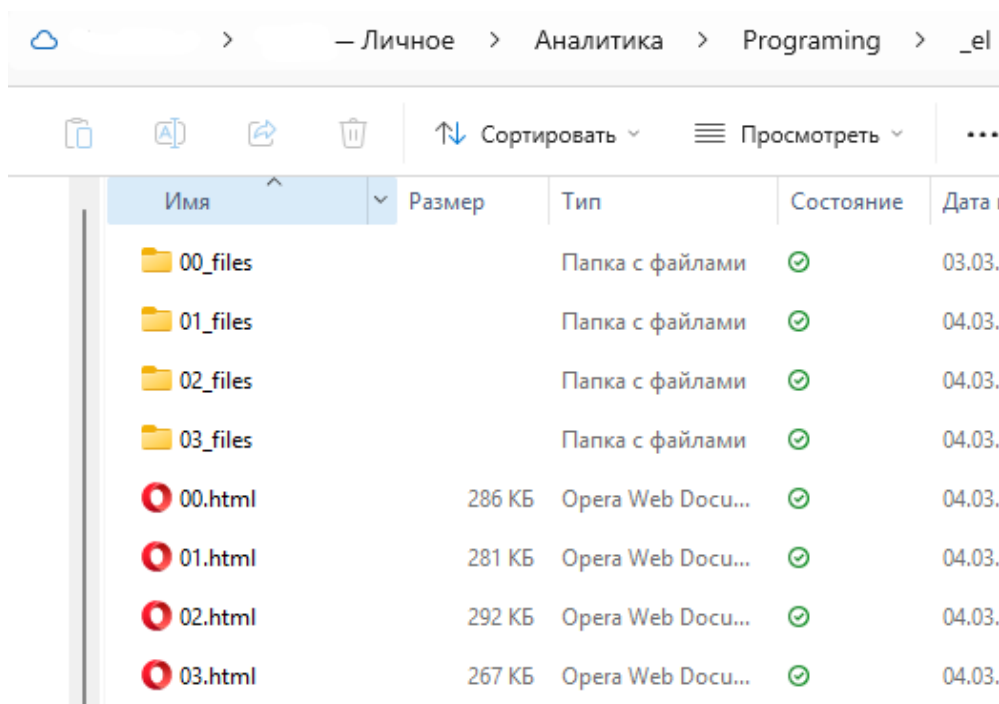


Рис. 3. Содержимое папки “_el” с исходными файлами для анализа
Fig. 3. Contents of the “_el” folder with source files for analysis

ОПИСАНИЕ ИНСТРУМЕНТОВ И СРЕДСТВ РАЗРАБОТКИ ПРИЛОЖЕНИЙ

В данном разделе перечисляется программное обеспечение, которое потребуется для получения исходных данных, а также набор средств и сред разработки, которые необходимы для написания исходного кода и генерации исполняемого файла из полученного кода. Среди необходимого для реализации исследования программного обеспечения можно указать следующее:

- интернет-браузер с возможностью сохранения веб-страниц в формате html и поддерживаемый библиотекой Selenium WebDriver, например, Google Chrome, Microsoft Edge, Mozilla Firefox или Apple Safari [15];
- интерпретатор Python версии не ниже 3.12.6 [16];
- JetBrains PyCharm (Community Edition) версии не ниже 2024.3.4 в качестве интегрированной среды разработки [17];
- программная библиотека pandas для генерации и обработки табличных данных [18];
- программная библиотека BeautifulSoup для синтаксического разбора html- и xml-документов [19];
- программная библиотека Selenium WebDriver как инструмент для автоматизации действий веб-браузера;
- библиотека Pyinstaller для упаковки Python-приложений в standalone-исполняемые файлы для ОС Windows [20].

После компиляции исполняемого файла программы процесс преобразования данных для потенциального пользователя можно ограничить только необходимостью использовать браузер и скомпилированные исполняемые программы.

ОПИСАНИЕ ЭТАПОВ РАЗРАБОТКИ ПРОГРАММНОГО ИНСТРУМЕНТАРИЯ

Общий принцип действия разрабатываемых скриптов для экспорта данных из elibrary.ru и формирования табличного файла со сведениями о публикациях научной организации на основе синтаксического анализа содержимого нескольких сохраненных страниц с портала elibrary.ru

основан на постолбцовом формировании таблицы с выгруженными публикационными данными. Алгоритм автоматической фильтрации и сохранения результатов применения фильтра состоит из нескольких этапов (рис. 4).

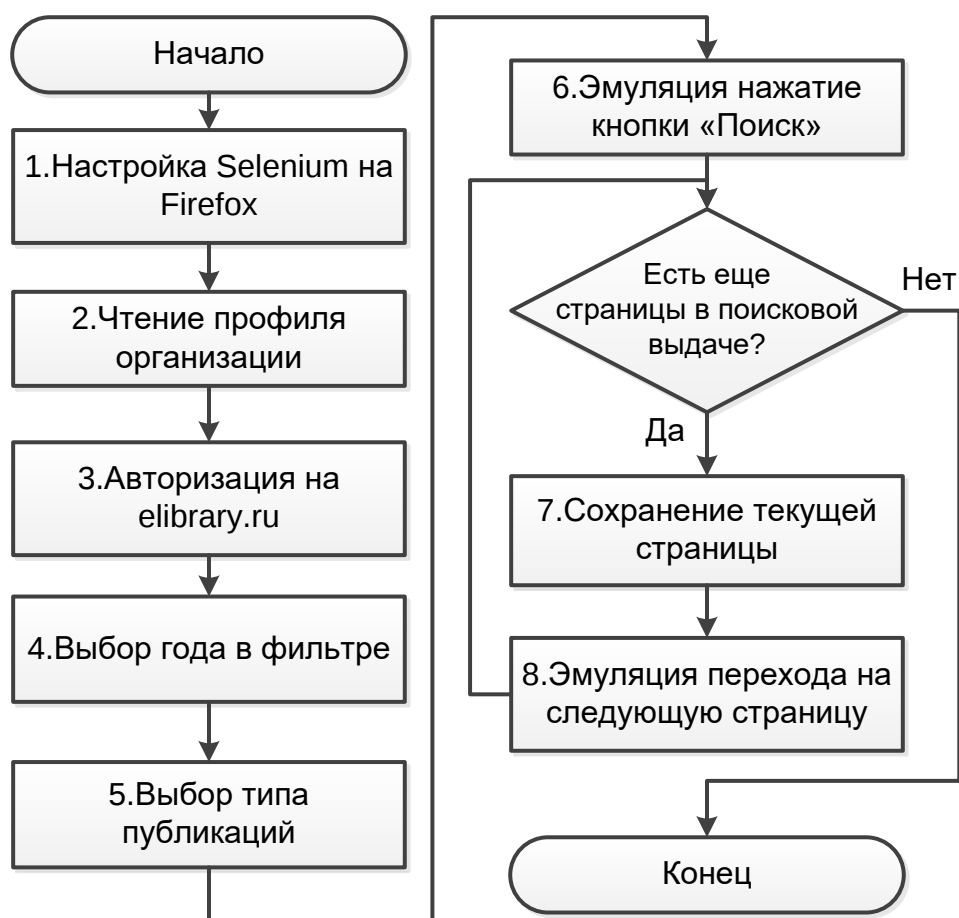


Рис. 4. Блок-схема алгоритма автоматического экспорта страниц из портала eLIBRARY.RU

Fig. 4. Flowchart of the algorithm for automatic export of web pages from the eLIBRARY.RU

На базе алгоритма был разработан скрипт, применение которого приведет к созданию набора html-файлов в папке “_el” с выходными публикационными данными. В основе скрипта лежит проект, представленный в работе [14]. Для составления собственной версии скрипта используются следующие шаги, включающие программные методы и опорные теги html-кода страниц:

Шаг 1 – настройка Selenium WebDriver на эмуляцию в Mozilla Firefox;

Шаг 2 – открытие профиля научной организации в режиме просмотра списка статей посредством метода `driver.get`.

Шаг 3 – авторизация на портале `elibrary.ru` (теги `<input name='login'>`, `<input name='password'>` и `<div onclick='check_all()'>`).

Шаг 4 – раскрытие списка годов и выбор конкретного года (свойства `'id="hdr_years"'` и `'id="year_2024"'`);

Шаг 5 – выбор типа публикаций из всплывающего списка (свойство `'value="3"'`, где ‘3’ значение, соответствующее варианту “статьи, в журналах, входящих в Russian Science Citation Index” в всплывающем списке);

Шаг 6 – эмуляция нажатия кнопки «Поиск» (признак `'//td[6]/div'`);

Шаг 7 – сохранение текущей страницы в файл с помощью метода `driver.page_source`;

Шаг 8 – эмуляция нажатия ссылки «Следующая страница», на основе поиска ссылки по тексту.

Одна из оригинальных составляющих скрипта реализует эмуляцию процесса авторизации на портале `elibrary.ru` (шаг 3 алгоритма на рисунке 1). Открывать профиль научной организации в

режиме просмотра списка публикаций можно и без авторизации, однако при автоматизации процесса сохранения веб-страниц настройки фильтров могут сбрасываться в процессе перехода к следующей странице с результатами поиска. Скрипт авторизации приведен в листинге 1.

Листинг 1. Скрипт авторизации на портале elibrary.ru

Source Code 1. Authorization script for eLIBRARY.RU

```
driver.find_element(By.XPATH, "///input[@name='login']").send_keys("login")
driver.find_element(By.XPATH, "///input[@name='password']").send_keys("password")
time.sleep(sleep_seconds)
driver.find_element(By.XPATH, "///div[@onclick='check_all()']").click()
```

Полностью весь скрипт автоэкспорта публикационных веб-данных в файлы приведен на хостинге ИТ-проектов GitHub [22].

Каждому из столбцов выходной таблицы, с приведенной выше структурой (табл. 1), будут соответствовать набор признаков во входном html-файле, на основе которого будет определяться принадлежность текста внутри тегов к тому или иному столбцу. Признаки представляют собой цепочки опорных тегов и их свойств. Конкретные особенности выбора состава тегов и их свойств будет рассмотрен далее. Общий алгоритм работы программы синтаксического анализа входных файлов, основанный на использовании методов класса BeautifulSoup, представлен на рисунке 5



Рис. 5. Блок-схема общего алгоритма программы для синтаксического офлайн-анализа сохраненных публикационных данных из портала eLIBRARY.RU

Fig. 5. Flowchart of the main algorithm of the software for offline web scraping of saved publication data from the eLIBRARY.RU

При извлечении ссылок на статьи в реализации шага 4 общего алгоритма и заголовков статей (шаг 5) необходимо учитывать, что на странице с результатами поиска могут присутствовать несколько версий одной и той же статьи, например, выходные данные статьи в переводной версии журнала (см. рис. 1). При реализации шага 4 из содержимого веб-страницы сначала извлекаются наборы текста тегов `<href>` по схеме: для всех `<href>` из списка всех тегов `<a>`, которые входят в теги `<td>`. Сокращенно здесь и далее подобные цепочки тегов будут обозначаться как `td -> a -> href`. Затем из полученного списка отфильтровываются только непустые элементы, внутри которых встречается текст `'elibrary.ru/item.asp?id='` (листинг 2).

Листинг 2. Исходный код скрипта получения списка ссылок на публикации из файлов, выгруженных на портале eLIBRARY.RU

Source Code 2. Source code of the script for obtaining a list of URLs to publications from pages uploaded from the eLIBRARY.RU

```
links_eng = [a.get('href') for a in (td.find('a') for td in bs.find_all('td')) if a]
urls_100 = [i for i in links_eng if (i is not None) and ('elibrary.ru/item.asp?id='
in i)]
```

В процессе реализации шага 5 из веб-страницы извлекаются тексты тегов `<span>` с атрибутом `style="line-height:1.0;"` в виде списка с именем `tituls_100` (листинг 3).

Листинг 3. Исходный код скрипта для получения списка заголовков статей

Source Code 3. Source code of the script for getting the list of article titles

```
tituls_100 = bs.find_all('span', attrs={ "style" : "line-height:1.0;"})
tituls_100 = [i.text for i in tituls_100]
```

Перед выполнением шага 6 формируется список `urls_eng` из всех ссылок на дублирующие версии статей согласно цепочке тегов `span class = "menuug" -> a -> href`

Шаг 6 реализован в виде двух вложенных циклов, в которых из списков `tituls_100` и `urls_100` удаляются элементы с номерами, предшествующими номерам элементов из списка `urls_eng`, то есть в списках со ссылками и заголовками статей остаются только вторые версии ссылок и заголовков (рис. 6).

Особенность реализации скрипта для шага 7 общего алгоритма (см. рис. 5) заключается в том, что на исходной веб-странице есть много элементов, подходящих под соответствующую опорную цепочку тегов `td -> font color=#00008f`, поэтому необходимо дополнительно урезать получившийся после применения фильтра по опорным тегам список до элемента, равного значению “Всего найдено”. Получившийся урезанный список включает чередующиеся элементы, включающие порядковый номер публикации в списке с результатами поиска и список соавторов. Чтобы оставить только списки соавторов применяется срез по четным элементам списка `authors_vyh`:

```
authors_vyh = authors_vyh[1::2]
```

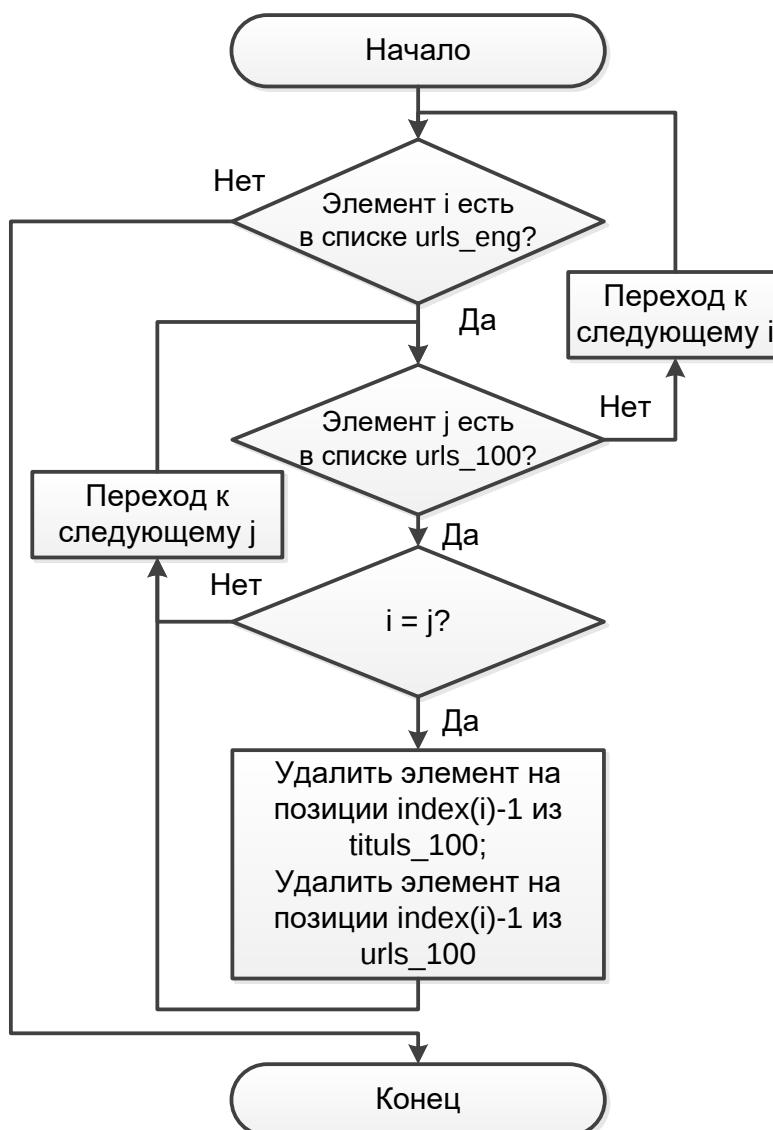


Рис. 6. Блок-схема алгоритма получения окончательного варианта списков ссылок и заголовков статей без дублирования

Fig. 6. Flowchart of the algorithm for obtaining the final version of lists of URLs and article titles without duplication

В выходных данных публикаций в файле с поисковыми результатами не у каждой публикации имеется список соавторов, например, если публикация представляет собой сборник статей по материалам конференции (рис. 7).

Всего найдено 4518 публикаций с общим количеством цитирований: 446. Показано на данной странице: с 101 по 200.		
№	Публикация	Цит.
101.	ФИЛОЛОГИЯ И НОВОЕ ЗНАНИЕ ☐ Сборник материалов IV Российской научной конференции с международным участием / Белгород, 2024.	0

Рис. 7. Пример выходных данных публикации без списка её соавторов

Fig. 7. Example of publication output data without a list of its co-authors

В связи с этим необходимо в качестве элементов списка, в которых текст не заключен в тег `<i>` вставлять пустую строку. Для реализации корректировки итогового списка используется сравнение двух списков – набора списка соавторов в тексте тегов `<i>` (переменная `authors_err`) и полного списка с авторами и возможными выходными данными (переменная `authors_vyh`). Блок-схема алгоритма получения списка соавторов публикаций приведена на рисунке 8.

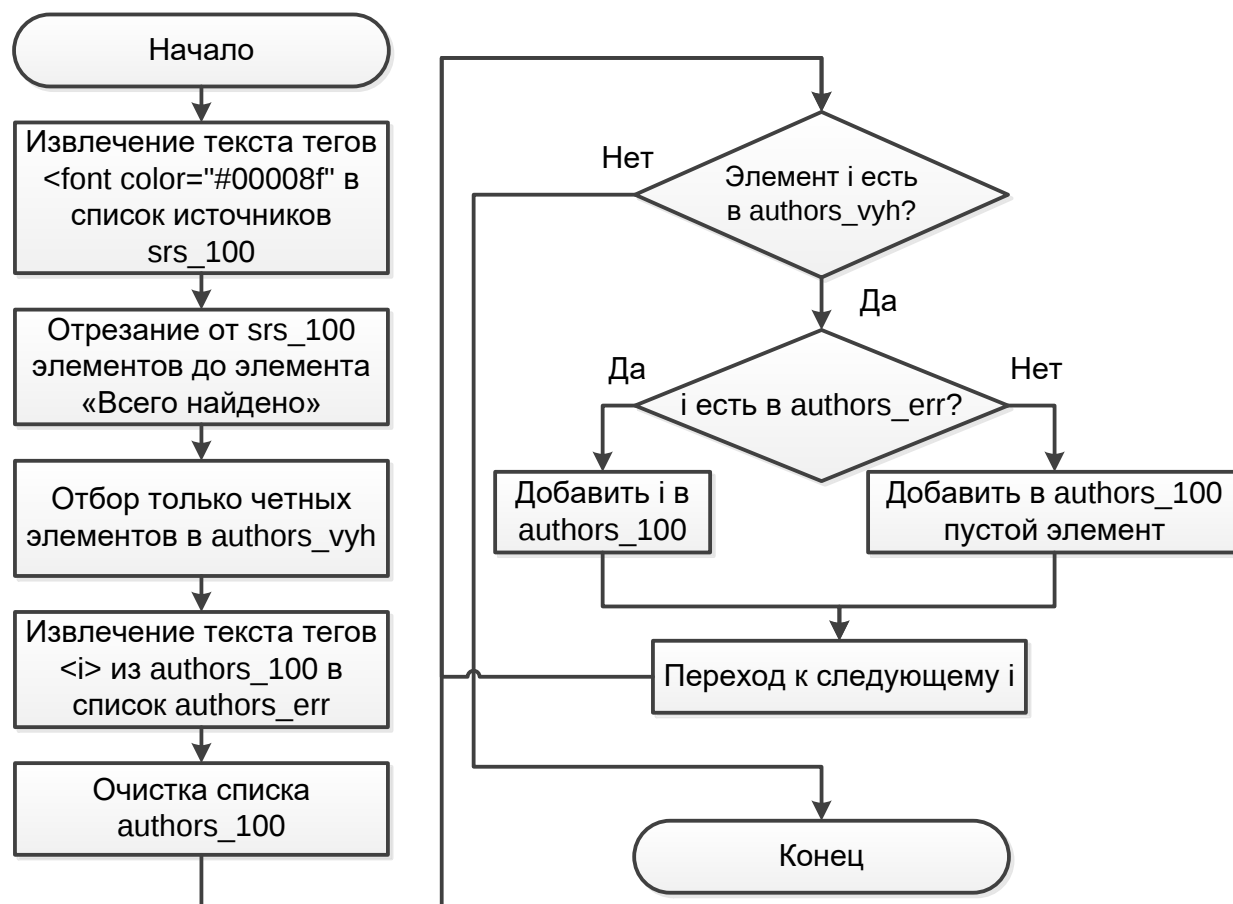


Рис. 8. Блок-схема алгоритма получения списка соавторов публикаций
Fig. 8. Flowchart of the algorithm for obtaining a list of co-authors of publications

Не менее трудоемким является алгоритм получения информации об источниках публикаций (шаг 8 общего алгоритма). Это обусловлено тем, что, если использовать опорный тег `<font color=\"#00008f\">`, извлекаются не только сведения об источниках, но и списки соавторов. После урезания полученного списка до элемента «Всего найдено» необходимо удалить из него элементы, присутствующие в списке с соавторами (переменная `authors_100`). Кроме это необходимо выполнить следующие операции:

- 1) урезать список, удалив из начала 2 лишних незначащих элемента;
- 2) убрать элементы равные символу конца строки `'\n'`;
- 3) убрать номера публикаций на нечетных местах;
- 4) удалить в элементах списка символы конца строки `('\'n\')` и возврат каретки `('\'r\')`.

Весь алгоритм получения информации об источниках статей приведен на рисунке 9.

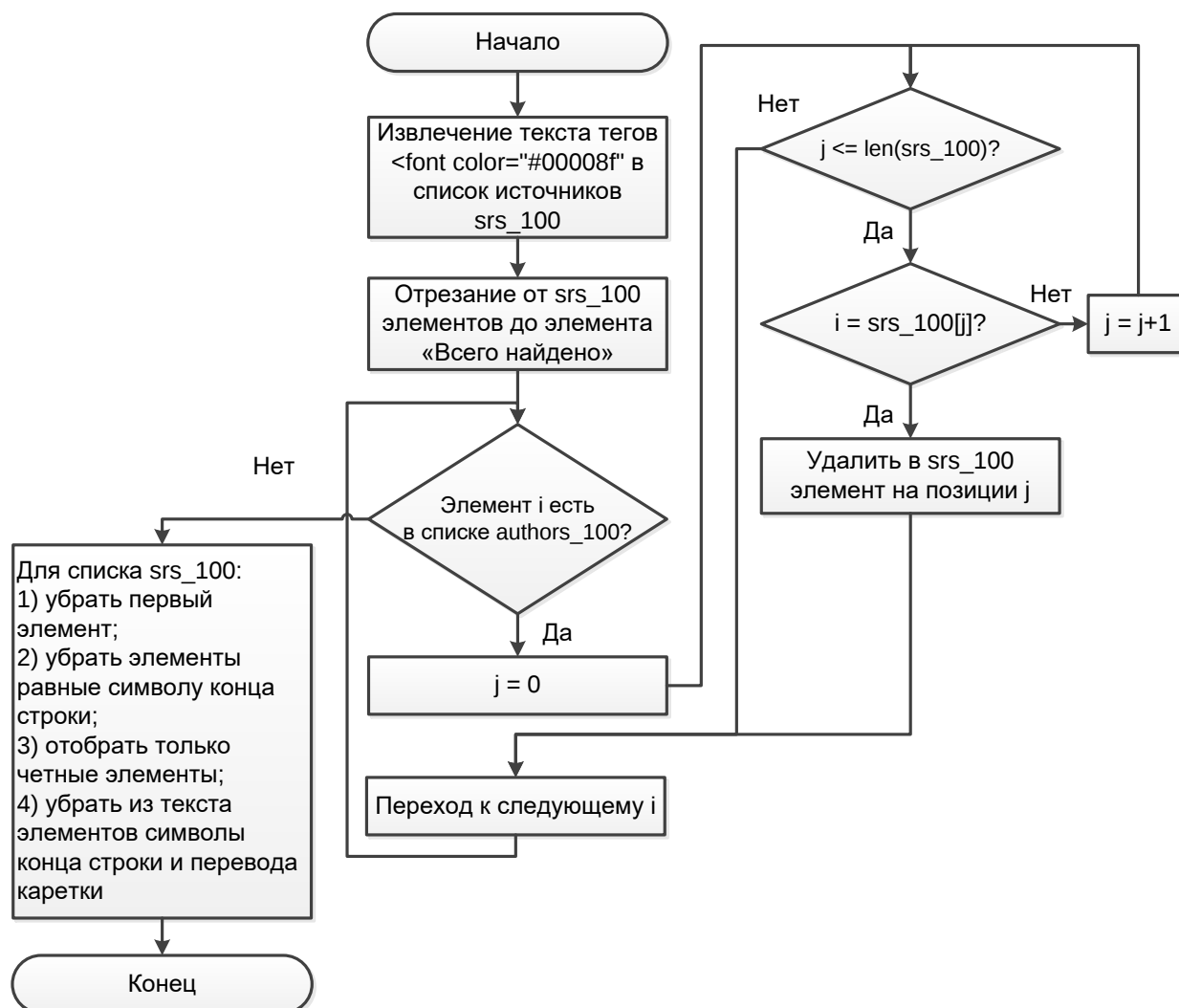


Рис. 9. Блок-схема алгоритма получения информации об источниках статей
Fig. 9. Flowchart of the algorithm for obtaining data about article sources

Последнее поле в структуре таблицы 1 – это идентификатор публикации на портале elibrary.ru. Содержимое этого поля получается в результате извлечения значений атрибута arw из цепочки тегов tr -> id (если тег <id> встречается в <tr>) (листинг 4).

Листинг 4. Исходный код скрипта получения уникальных идентификаторов публикации на портале eLIBRARY.RU

Source code 4. Source code of the script for obtaining unique articles identifiers on the eLIBRARY.RU

```

tr_100 = bs.find_all('tr')
id_100 = [i.get('id') for i in tr_100 if i.has_attr('id')]
id_100 = [i for i in id_100 if 'arw' in i]
id_100 = [i[3:] for i in id_100]

```

В результате реализации скриптов получают 5 списков, которые распределяются по результирующему фрейму данных tempdf, а тот, в свою очередь добавляется к результирующему фрейму данных elxls (листинг 5).

Листинг 5. Исходный код скрипта для сборки результирующего фрейма данных elxls
Source code 5. Source code of the script for assembling the resulting DataFrame 'elxls'

```
tempdf = pd.DataFrame()
tempdf.insert(0, 'IDs', id_100)
tempdf.insert(1, 'Авторы', authors_100)
tempdf.insert(2, 'Название', tituls_100)
tempdf.insert(3, 'Источник', srs_100)
tempdf.insert(4, 'URL', urls_100)
elxls = pd.concat([elxls, tempdf])
```

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И ИХ ОБСУЖДЕНИЕ

Разработанный программный инструментарий позволяет получать только самые базовые сведения о публикациях: название статьи, список соавторов и сведения об источнике, включая номер тома и выпуска (для журналов), а также диапазоны страниц размещения статьи в источнике. В скрипте не реализовано извлечение количества цитирований статей, но при необходимости это может быть выполнено добавлением скрипта, аналогичного представленному в листинге 2 на базе опорного тега `<td align="center" valign="middle" class="select-tr-right">`. Кроме этого, в случае, когда в публикации много соавторов, то их список в соответствующей строке электронной таблицы будет урезанным.

При изменении состава опорных тегов или их атрибутов, исходный код разработанных скриптов может претерпеть незначительные изменения, но общие принципы реализации останутся нетронутыми. Однако программный инструментарий может стать непригодным для использования без значительной модификации в случае изменении платформы, на основе которой работает портал elibrary.ru.

Код разработанных скриптов можно сократить за счет еще большего уровня использования функционального программирования в форме генератора списков [21]. Использование Selenium WebDriver позволило адаптировать программу под работу в режиме онлайн, что устранило необходимость вручную длительно сохранять веб-страницы с результатами поиска статей.

При помощи разработанных инструментов нет возможности получить такие важные характеристики публикаций как DOI, ISSN журналов и списки аффилиаций авторов. Если у научной организации есть лицензия на использования системы Science Index, то для получения таких характеристик необходимо использовать алгоритм и инструменты, разработанные в исследовании [12].

Разработанный программный инструментарий при незначительной доработке можно также использовать для целей преобразования содержимого веб-страниц в электронные таблицы для решения следующих задач при работе на портале elibrary.ru:

- обработка результатов расширенного поиска;
- синтаксического анализа подборок публикаций;
- анализ списка публикаций авторов;
- анализ информации об авторах научной организации.

ЗАКЛЮЧЕНИЕ

В статье были проанализированы методы извлечения сведений из веб-страниц с выходными данными публикаций научной организации на портале elibrary.ru в формат электронной таблицы, были кратко рассмотрены особенности их использования, а также их достоинства и недостатки. В качестве реализации одного из методов был разработан автоэкспортер контента веб-страниц из elibrary.ru в html-файлы, а также синтаксический анализатор, позволяющий решить задачу извлечения и сохранения содержимого выгруженных файлов в формате Excel. Исходный код анализаторов можно легко скомпилировать в исполняемый файл Windows посредством инструментов библиотеки Pyinstaller (команда `"python -m PyInstaller --onefile [имя файла].py"` или

запускать посредством самого интерпретатора Python на операционных системах семейства Linux при условии наличия в папке со скриптом каталога “_el” с набором html-файлов. Использование этого программного инструментария позволяет сократить минимум на 5 минут время преобразования содержимого веб-страниц в электронную таблицу на каждые 100 публикаций результатов фильтрации и получить данные в форме, пригодной для дальнейшей обработки и анализа.

Список литературы

1. Бабынина Л.С., Грунина И.С. Программа "Приоритет-2030" как развитие концептуальных под-ходов в оценке деятельности вузов // KANT. 2024. №1 (50). С. 4-11. EDN: AAMGCU. DOI: 10.24923/2222-243X.2024-50.1 URL: https://www.elibrary.ru/download/elibrary_65310545_87336556.pdf (дата обращения: 03.04.2025)
2. Формирование инновационной образовательной системы в НИУ «БелГУ» как основы подготовки кадров мирового уровня на базе НОЦ «Инновационные решения в АПК» / О.Н. Полухин, А.В. Маматов, И.В. Спичак и др. // Достижения науки и техники АПК. 2020. Т. 34. № 9. С. 9–13. DOI: 10.24411/0235-2451-2020-10902. URL: https://www.elibrary.ru/download/elibrary_44095025_41119820.pdf (дата обращения: 03.04.2025)
3. Kochetkov D.M. Russian Journal Whitelist: Questions to be answered // Science Editor and Publisher. 2022. № 7(2). С. 185-190. URL: <https://doi.org/10.24069/SEP-22-48> (дата обращения: 03.04.2025)
4. Moed H.F., Markusova V., Akoev M. Trends in Russian research output indexed in Scopus and Web of Science // Scientometrics. 2018. № 116. С. 1153-1180. URL: <https://doi.org/10.1007/s11192-018-2769-8> (дата обращения: 03.04.2025)
5. О проекте eLIBRARY.RU. Научная электронная библиотека eLIBRARY.RU, 2025. URL: https://elibrary.ru/elibrary_about.asp (дата обращения: 03.04.2025)
6. Российский индекс научного цитирования. Научная электронная библиотека eLIBRARY.RU, 2025. URL: https://elibrary.ru/projects/citation/cit_index.asp (дата обращения: 03.04.2025)
7. О проекте Russian Science Citation Index. Научная электронная библиотека eLIBRARY.RU, 2025. URL: https://www.elibrary.ru/rsci_about.asp? (дата обращения: 03.04.2025)
8. Перечень рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук (по состоянию на 10.06.2024 года), 2024. URL: <https://vak.minobrnauki.gov.ru/uploader/loader?type=19&name=91107547002&f=23267> (дата обращения: 03.04.2025)
9. Интерфейс программирования приложений API. Научная электронная библиотека, 2025. URL: https://www.elibrary.ru/projects/api/api_info.asp (дата обращения: 03.04.2025)
10. Информационно-аналитическая система SCIENCE INDEX. Научная электронная библиотека, 2025. URL: https://www.elibrary.ru/projects/science_index/science_index_org_info.asp? (дата обращения: 03.04.2025)
11. Гинтофт А.С., Новгородов П.А., Коренев А.Н. Интеграционный модуль импорта наукометрических данных из научной электронной библиотеки. Свидетельство о государственной регистрации программы для ЭВМ № 2021663072 от 04.08.2021. URL: <https://www.elibrary.ru/item.asp?id=46484977> (дата обращения: 03.04.2025)
12. Резниченко О.С. Алгоритмы и инструменты обработки данных о публикациях научных организаций, размещенных на портале eLIBRARY.RU // Экономика. Информатика. 2025. № 52(1). С. 181-193. DOI: 10.52575/2687-0932-2025-52-1-181-193.
13. Парсинг электронных ресурсов. Библиотека selenium или fake useragent? / А.Ю. Попов, М.В. Ремез, Е.В. Жилина, М.И. Ожиганова // Информатизация в цифровой экономике. 2022. Т. 3. № 4. С. 197-210. DOI: 10.18334/ide.3.4.115219
14. Запекин С., Шитов В., Заваруев И. Парсер научных журналов eLibrary // GitHub. 2021. URL: <https://github.com/Lfdd/Parser> (дата обращения: 03.04.2025)
15. Supported Browsers. Selenium, 2025. URL: <https://www.selenium.dev/documentation/webdriver/browsers/> (дата обращения: 03.04.2025)
16. van Rossum Guido at al. Python, 2025. URL: <https://www.python.org> (дата обращения: 03.04.2025)
17. JetBrains. PyCharm: The Python IDE for data and web professionals, 2025. URL: <https://www.jetbrains.com/pycharm/> (дата обращения: 03.04.2025)

18. Van den Bossche J. at al. pandas.DataFrame, 2024. URL: <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html> (дата обращения: 03.04.2025)
19. Richardson L. Beautiful Soup Documentation, 2015. URL: <https://beautiful-soup.readthedocs.io/en/latest/> (дата обращения: 03.04.2025)
20. Cortesi D. PyInstaller Manual, 2025. URL: <https://pyinstaller.org/en/stable/index.html> (дата обращения: 03.04.2025)
21. van Rossum G. at al. 5.1.3. List Comprehensions. Python, 2025. URL: <https://docs.python.org/3/tutorial/datastructures.html#list-comprehensions> (дата обращения: 03.04.2025)
22. Reznichenko O. Appendix to article "Methods and Software Tools for Uploading Publication Data of a Scientific Organization Posted on eLIBRARY.RU" // GitHub. 2025. URL: https://github.com/leo-phoenix/elibrary_html_to_xls (дата обращения: 03.04.2025)

References

1. Babynina L.S., Grunina I.S. Priority-2030 Program as A Development of Conceptual Approaches in The Assessment of Higher Education Institutions' Performance. KANT. 2024. No. 1(50). P. 4-11. EDN: AAMGCY. DOI: 10.24923/2222-243X.2024-50.1 URL: https://www.elibrary.ru/download/elibrary_65310545_87336556.pdf (accessed: 03 April 2025)
2. Polukhin O.N., Mamatov A.V., Spichak I.V., et al. Formation of an innovative educational system at the Belgorod State National Research University as the basis for training world-class personnel at the Scientific and Educational Center "Innovative Solutions in the Agro-Industrial Complex" in the Belgorod region. Dostizheniya nauki i tekhniki APK. 2020. No. 34(9). P. 9-13. DOI: 10.24411/0235-2451-2020-10902. URL: https://www.elibrary.ru/download/elibrary_44095025_15844399.pdf (accessed: 03 April 2025)
3. Kochetkov D.M. Russian Journal Whitelist: Questions to be answered. Science Editor and Publisher. 2022. No. 7(2). P. 185-190. URL: <https://doi.org/10.24069/SEP-22-48> (accessed: 03 April 2025)
4. Moed H.F., Markusova V., Akoev M. Trends in Russian research output indexed in Scopus and Web of Science // Scientometrics. 2018. No. 116. P. 1153-1180. URL: <https://doi.org/10.1007/s11192-018-2769-8> (accessed: 03 April 2025)
5. About eLIBRARY.RU project. Scientific Electronic Library. 2025. URL: <https://elibrary.ru/defaultx.asp> (accessed: 03 April 2025)
6. Russian Science Citation Index. Scientific Electronic Library. 2025. URL: https://elibrary.ru/projects/citation/cit_index.asp (accessed: 03 April 2025)
7. About RSCI project. Scientific Electronic Library 2025. URL: https://www.elibrary.ru/rsci_about.asp (accessed: 03 April 2025)
8. List of peer-reviewed journals in which the principal scientific findings of dissertations for the degrees of Candidate of Sciences and Doctor of Sciences must be published (as of 10.06.2024). 2024. URL: <https://vak.minobrnauki.gov.ru/uploader/loader?type=19&name=91107547002&f=23267> (accessed: 03 April 2025)
9. Application Programming Interface API. Scientific Electronic Library. 2025. URL: https://www.elibrary.ru/projects/api/api_info.asp (accessed: 03 April 2025)
10. Science Index. Interface for a company representative. Scientific Electronic Library. 2025. URL: https://www.elibrary.ru/projects/science_index/science_index_org_info.asp (accessed: 03 April 2025)
11. Gintoft A.S., Novgorodov P.A., Korenev A.N. Integration Module for Importing Scientific Metric Data from a Scientific Electronic Library. Certificate of state registration of a computer program. 2021. No. RU 2021663072 at 04.08.2021. URL: <https://www.elibrary.ru/item.asp?id=46484977> (accessed: 03 April 2025)
12. Reznichenko O.S. Algorithms and Tools for Processing Data on Scientific Organizations' Articles Uploaded to eLIBRARY.RU. Economics. Information Technologies. 2025. No. 52(1). P. 181-193). DOI: 10.52575/2687-0932-2025-52-1-181-193
13. Popov A.Y., Remez M.V., Zhilina E.V., Ozhiganova M.I. Parsing of electronic resources. Selenium library or fake useragent? Informatization in the Digital Economy. 2022. No. 3(4), P. 197-210. URL: <https://doi.org/10.18334/ide.3.4.115219> (accessed: 03 April 2025)
14. Zapekin S., Shitov V., Zavaruev I. eLibrary Scientific Journal Parser // GitHub. 2021. URL: <https://github.com/Lfdd/Parser> (accessed: 03 April 2025)
15. Supported Browsers. 2025. URL: <https://www.selenium.dev/documentation/webdriver/browsers/> (accessed: 03 April 2025)
16. van Rossum G. at al. Python. 2025. URL: <https://www.python.org> (accessed: 03 April 2025)

17. JetBrains. PyCharm: The Python IDE for Professional Developers. 2025. URL: <https://www.jetbrains.com/pycharm/> (accessed: 03 April 2025)
18. Van den Bossche J. et al. pandas.DataFrame. 2024. URL: <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html> (accessed: 03 April 2025)
19. Richardson L. Beautiful Soup Documentation. 2015. URL: <https://beautiful-soup.readthedocs.io/en/latest/> (accessed: 03 April 2025)
20. Cortesi D. PyInstaller Manual. 2025. URL: <https://pyinstaller.org/en/stable/index.html> (accessed: 03 April 2025)
21. van Rossum G. et al. 5.1.3. List Comprehensions. 2025. URL: <https://docs.python.org/3/tutorial/datastructures.html#list-comprehensions> (accessed: 03 April 2025)
22. Reznichenko O. Appendix to article "Methods and Software Tools for Uploading Publication Data of a Scientific Organization Posted on eLIBRARY.RU" // GitHub. 2025. URL: https://github.com/leo-phoenix/elibrary_html_to_xls (accessed: 03 April 2025)

Резниченко Олег Сергеевич, старший преподаватель кафедры прикладной информатики и информационных технологий, Белгородский государственный национальный исследовательский университет, г. Белгород, Россия

Reznichenko Oleg Sergeevich, Senior lecturer of the Department of Applied Informatics and Information Technology, Belgorod State National Research University, Belgorod, Russia