УДК 004.51, 004.62, 004.65

ПРОЕКТИРОВАНИЕ АГЕНТНОЙ СИСТЕМЫ ДЛЯ СБОРА И ОТОБРАЖЕНИЯ ДАННЫХ О НАУЧНЫХ ПУБЛИКАЦИЯХ СТРАН СОЮЗА БРИКС

Соколова Мария Викторовна[1], Кошлань Диана Игоревна[2], Пряхина Дарья Игоревна[3]

[1]Студент;
ГБОУ ВО МО «Университет «Дубна»,
141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: somv.17@uni-dubna.ru.

[2]Инженер-программист;
Объединенный институт ядерных исследований,
Россия, 141980, Московская обл., г. Дубна, ул. Жолио-Кюри, д. 6;
Аспирант;
ГБОУ ВО МО «Университет «Дубна»,
Россия, 141980, Московская обл., г. Дубна, ул. Университетская, д. 19;
e-mail: dkoshlan@jinr.ru.

[3]Научный сотрудник;
Объединенный институт ядерных исследований,
Россия, 141980, Московская обл., г. Дубна, ул. Жолио-Кюри, д. 6;
Старший преподаватель;
ГБОУ ВО МО «Университет «Дубна»,
Россия, 141980, Московская обл., г. Дубна, ул. Университетская, д. 19;
e-mail: pryahinad@jinr.ru.

Работа посвящена проектированию агентной системы для сбора и отображения данных о научных публикациях стран союза БРИКС. Поэтому целью работы является проектирование системы, позволяющей собирать информацию из баз данных (БД), которые содержат сведения о научных публикациях, изданных в ведущих журналах, и предоставлять статистическую информацию о публикационной активности стран БРИКС в удобном для анализа виде. В качестве информационного ресурса, из которого будут извлекаться требуемые данные, выбран сайт natureindex.com. Выбранный ресурс Nature Index содержит ссылки на публикации каждой страны в определенном журнале. В проектируемой системе требуется разработать программу-агент, которая должна собирать с сайта ссылки на публикации и названия журналов этих публикаций. В статье представлены требования к системе, описана структура выбранного информационного ресурса. Проведено проектирование каждой части системы: описана модель БД, представлен алгоритм работы программы-агента и приведена диаграмма классов для ее дальнейшей реализации, обоснован выбор трехуровневой архитектуры веб-приложения и созданы макеты пользовательского интерфейса для отображения информации согласно выдвинутым требованиям и вариантам использования. Результаты проектирования будут использоваться в дальнейшем для разработки описанной системы.

Ключевые слова: проектирование, информационная система, база данных, программа-агент, веб-приложение, публикационная активность, союз БРИКС.

Для цитирования:
Соколова М. В., Кошлань Д. И., Пряхина Д. И. Проектирование агентной системы для сбора и отображения данных о научных публикациях стран союза БРИКС // Системный анализ в науке и образовании: сетевое научное издание. 2021. № 2. С. 18–31. URL : http://sanse.ru/download/433.

DESIGNING AN AGENT SYSTEM FOR COLLECTING AND DYSPLAYING DATA ON SCIENTIFIC PUBLICATIONS OF BRICS UNION COUNTRIES

Sokolova Maria V.[1], Koshlan Diana I.[2], Priakhina Daria I.[3]

[1]Student;
Dubna State University;
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;
e-mail: somv.17@uni-dubna.ru.

[2]Software engineer;
Joint Institute for Nuclear Research,
6 Joliot-Curie Str., Dubna, Moscow region, 141980, Russia;
PhD student;
Dubna State University;
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;
e-mail: dkoshlan@jinr.ru.

[3]Research fellow;
Joint Institute for Nuclear Research,
Russia, 141980, Moscow reg., Dubna, 6 Joliot-Curie st.;
Senior lecturer;
Dubna State University,
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;
e-mail: pryahinad@jinr.ru.

The paper is devoted to the design of an agent system for collecting and displaying data on scientific publications of the BRICS Union countries. Therefore, the aim of the work is to design a system that allows you to collect information from databases that contain information about scientific publications published in leading journals and provide statistical information about the publication activity of the BRICS countries in a convenient form for analysis. The site natureindex.com is selected as an information resource from which the required data will be extracted. The selected Nature Index resource contains links to publications of each country in a particular journal. In the designed system, it is required to develop an agent program that should collect links to publications and journal titles of these publications from the site. The article presents the requirements for the system, describes the structure of the selected information resource. The design of each part of the system is carried out: the database model is described, the algorithm of the agent program is presented and a class diagram for its further implementation is given, the choice of a three-level architecture of the web application is justified and user interface layouts are created to display information according to the advanced requirements and use cases. The design results will be used in the future to develop the described system.

Keywords: design, information system, database, agent program, web application, publication activity, BRICS union.

For citation:
Sokolova M., Koshlan D., Priakhina D. Designing an agent system for collecting and dysplaying data on scientific publications of BRICS union countries. System Analysis in Science and Education, 2021;(2):18–31(In Russ). Available from: http://sanse.ru/download/433.

Введение

Анализ публикационной активности различных стран интересен с точки зрения получения информации о количестве публикаций каждой страны, выделения наиболее популярных тем для исследований, изучения сотрудничества разных стран друг с другом по публикациям, в которых принимали участие конкретные организации стран. В качестве анализируемых государств выбраны страны союза БРИКС [1]: Бразилия, Россия, Индия, КНР, ЮАР.

Для проведения такого анализа необходимо разработать систему для сбора данных о научных публикациях, например, с электронных ресурсов, предоставляющих информацию о различных жур-

налах. Помимо этого, система должна включать приложение с удобным интерфейсом, с помощью которого пользователи смогут просматривать собранную информацию.

К разработке такой системы можно приступать только после ее детального проектирования. Поэтому целью работы является проектирование системы, позволяющей собирать информацию из баз данных (БД), которые содержат сведения о научных публикациях, изданных в ведущих журналах, и предоставлять статистическую информацию о публикационной активности стран БРИКС в удобном для анализа виде.

Требования к системе

К разрабатываемой системе выдвинуты функциональные требования, которые должны быть отражены в макетах пользовательского интерфейса:

- должны быть доступны для просмотра:
  - количество публикаций каждой страны союза БРИКС, согласно выбранному информационному ресурсу;
  - список организаций страны, отсортированный по количеству публикаций;
  - список авторов с наибольшим количеством публикаций для любой организации;
  - список наиболее часто встречающихся ключевых слов для публикаций каждой страны союза БРИКС;
  - список наиболее часто встречающихся ключевых слов для публикаций всех стран союза БРИКС;
  - информация о попарном сотрудничестве стран;
- должна быть реализована возможность осуществлять поиск среди организаций страны по названию организации или по количеству публикаций;
- должна быть реализована возможность запустить процесс обновления статистики;
- во время обновления статистики система должна быть в состоянии в любой момент выдавать требуемую информацию.

К системе выдвинуты требования к удобству использования, которые также должны быть отражены в макетах пользовательского интерфейса:

- должна быть отдельная страница для отображения публикационной активности всех стран союза БРИКС;
- должна быть отдельная страница для отображения списка самых активных организаций, имеющих наибольшее количество публикаций («Топ организаций»), и списка часто используемых ключевых слов для каждой страны («Топ ключевых слов»);
- должна быть отдельная страница для отображения списка авторов, имеющих наибольшее количество публикаций, каждой организации («Топ авторов»);
- должна быть отдельная страница для отображения кнопки запуска обновления статистики;
- поиск по организациям не должен зависеть от регистра.

К требованиям надежности системы относятся следующие:

- после сбоя система должна быть восстановлена без потери данных;
- режим работы системы – 7 дней в неделю без перерывов.

К производительности системы выдвинуты следующие требования:

- первое после запуска системы получение данных об организациях с наибольшим количеством публикаций может занимать не более двух минут, в дальнейшем получение этих данных должно занимать не более 5 секунд;
- получение всех остальных данных в любое время должно занимать не более 5 секунд;
- процесс сбора данных должен длиться не более 24 часов.

Помимо всего перечисленного работоспособность системы не должна зависеть от стороннего программного обеспечения.

Описание информационного ресурса

Журналов, где публикуются статьи ученых из организаций разных стран, немало, но существуют сайты, которые агрегируют ссылки на публикации в различных журналах. С таких сайтов удобно взять список ссылок на публикации для определенной страны и собрать данные об этих публикациях со страниц журналов. Одним из таких агрегаторов является сайт *Nature Index* [2].

Nature Index следит за исследовательскими научными статьями, он обновляется ежемесячно и представляет результаты исследований в разбивке по институтам и странам [2]. Если в статье есть один или несколько соавторов из России, а также один или несколько соавторов из Бразилии, то в рамках данного ресурса считается, что статья относится и к России, и к Бразилии.

Nature Index был выбран, так как он содержит необходимые для сбора данные о публикациях, а также разрешает сканирование себя программами-агентами, согласно файлу *robots.txt* – это текстовый файл, который содержит параметры индексирования сайта для роботов поисковых систем [3]. Такой файл сайта *Nature Index* не содержит параметров для директивы *Disallow* (см. рис. 1), что и разрешает сканирование всех страниц сайта. Это играет важную роль, так как, например, такой информационный ресурс как *Web Of Science* запрещает сканирование своих страниц, согласно файлу *robots.txt* в корне его сайта (см. рис. 2).

Рис. 1. Содержание файла robots.txt сайта Nature Index

Рис. 2. Содержание файла robots.txt сайта Web of science

Выбранный ресурс *Nature Index* содержит ссылки на публикации каждой страны в определенном журнале. В проектируемой системе требуется разработать программу-агент, которая должна собирать с сайта ссылки на публикации и названия журналов этих публикаций.

Информационный ресурс *Nature Index* содержит отдельную страницу для каждой страны, где представлен список тем (см. рис. 3), каждая из которых разворачивается в список журналов (см. рис. 4), где публикуются научные статьи.

Subject	Count	Share
Life Sciences	230	55.79
Physical Sciences	467	154.49
Chemistry	128	57.65
Earth & Environmental Sciences	104	34.91

Рис. 3. Список тем научных исследований на странице информационного ресурса

Subject	Count	Share
Life Sciences	230	55.79
American Journal of Human Genetics	7	0.59
Cancer Cell	1	0.03
Cancer Research	4	1.7
Cell	4	0.44

Рис. 4. Список журналов внутри темы

Каждый журнал из списка разворачивается в список ссылок на публикации, которые ведут на страницу того же информационного ресурса *Nature Index* с краткой информацией о публикации и кнопкой перехода на страницу публикации на сайте конкретного журнала (см. рис. 5). По нажатию на кнопку перехода к статье (*Go to article*) откроется страница, где будут отображены: название статьи, авторы и их аффилиации, ключевые слова и аннотация.

Рис. 5. Кнопка перехода ну страницу журнала с публикацией

Разработка модели базы данных

Для проведения анализа публикационной активности стран необходимо знать, в каком количестве публикаций каждая страна приняла участие; какие организации и авторы страны принимали участие в тех или иных публикациях; с какими организациями связаны авторы публикаций; какие ключевые слова относятся к каждой научной публикации; даты издания научных публикаций.

Для хранения перечисленной информации необходимо спроектировать реляционную БД. При построении модели данных были выделены следующие сущности: Публикация; Ключевое слово; Организация; Автор; Публикация_организация; Публикация_автор; Автор_организация; Публикация_ключевое_слово.

У одной публикации может быть несколько авторов, организаций и ключевых слов. Одна организация, как и один автор, может принимать участие в нескольких публикациях. Одно ключевое слово может относиться к нескольким публикациям. Один автор может относиться к нескольким организациям, а у каждой организации может быть несколько авторов. Исходя из этого спроектирована логическая модель БД для хранения данных (см. рис. 6).

Хранение данных публикации: ее названия, ссылки на сайте журнала, ссылки на страницу на сайте *Nature Index*, где расположена кнопка перехода на страницу журнала с публикацией, названия журнала, аннотации и даты осуществляется в сущности «Публикация». Данные авторов публикаций (их имена) хранятся в сущности «Автор». Данные организаций стран (их названия), принявших участие в публикациях, осуществляется в сущности «Организация». Ключевые слова, относящиеся к публикациям, хранятся в сущности «Ключевое слово». Связи публикаций с организациями, которые принимали участие в их создании, хранятся в сущности «Публикация_организация». Связи публикаций с авторами, которые принимали участие в их создании, хранятся в сущности «Публикация_автор». Связи авторов с организациями, к которым они относятся, хранятся в сущности «Ав-

тор_организация». Связи слов и публикаций, в которых такие слова отмечены как ключевые, хранятся в сущности «Публикация_ключевое_слово».

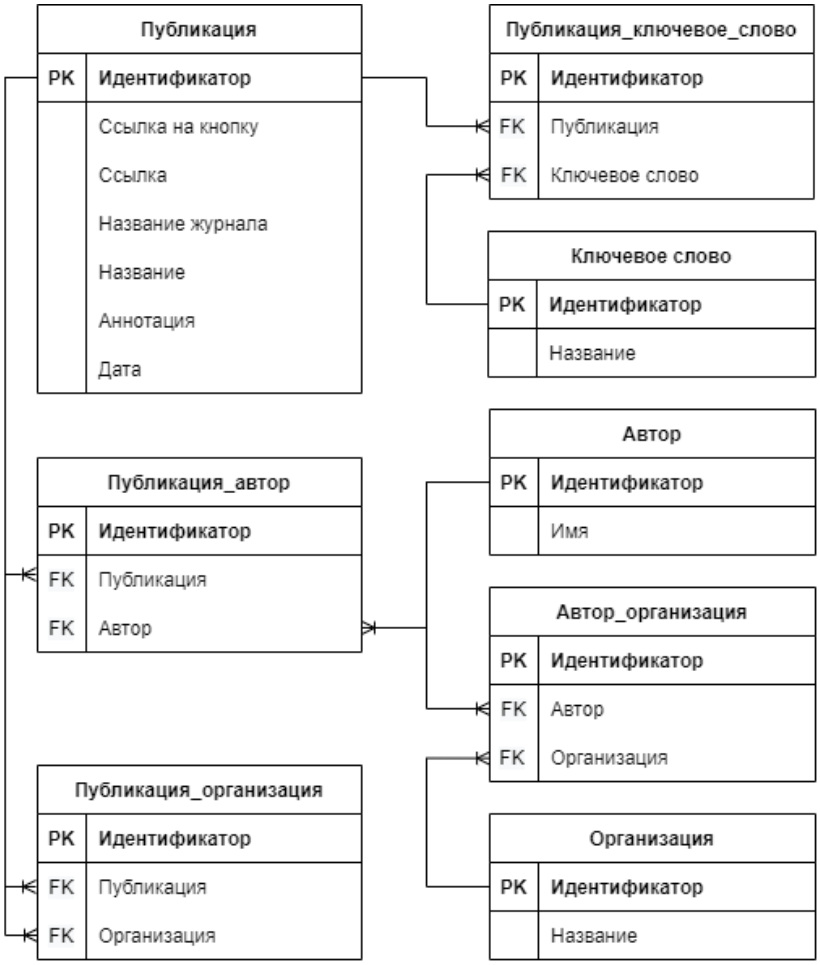

Рис. 6. Логическая модель БД

Проектирование программы-агента

Задачами программы-агента являются: сбор данных публикаций для каждой страны союза БРИКС с информационного ресурса *Nature Index* и доставка собранных данных в систему хранения. Алгоритм работы агента представлен на рисунке 7.

Для сбора данных публикации агент должен перейти на страницу публикации по ссылке для нее. Список ссылок на публикации для каждой страны необходимо собрать с информационного ресурса *Nature Index*.

Для осуществления возможности перехода по ссылкам на публикации, ссылки должны быть собраны в список вместе с соответствующими журналами, для каждого из которых должны быть написаны отдельные методы сбора данных, поскольку информационные ресурсы каждого журнала имеют разную структуру. Для корректного сбора данных каждый журнал должен быть связан со своим методом сбора данных. После установления необходимых связей программой-агентом может производиться непосредственно сбор данных каждой публикации.

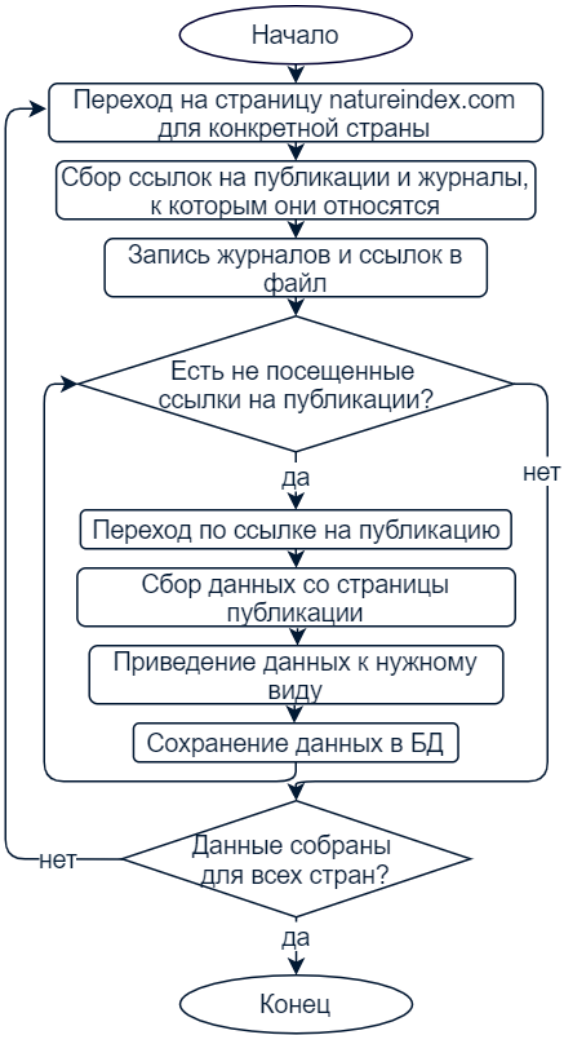

Рис. 7. Алгоритм работы программы-агента

Можно организовать работу программы-агента, создав следующие классы. Процессом работы программного агента может управлять класс *Agent*. Он делегирует обязанность сбора данных классу *Parser*, а обязанности доставки данных в систему хранения – классу *Service*. Таким образом, класс *Agent* передает классу *Parser* информацию о том, какие данные и откуда нужно собрать, а полученные от класса *Parser* данные класс *Agent* передает классу *Service*, чтобы тот позаботился о преобразовании данных к нужному виду и их доставке в БД.

Ссылки на публикации можно записывать в отдельный текстовый *csv*-файл для каждой страны. При этом необходимо учитывать, что сайт каждого научного журнала, публикующего статьи, имеет свою уникальную структуру, из чего вытекает необходимость иметь специальный метод для сбора данных с сайта каждого журнала. Поэтому вместе с каждой ссылкой на публикацию в *csv*-файл необходимо записать и название журнала, где расположена информация о публикации, чтобы затем агент знал, какой метод использовать для сбора данных с конкретной страницы.

Для создания метода сбора данных со страниц конкретного журнала необходимо проанализировать структуру сайта этого журнала. Например, страница одной из публикаций журнала *Nature* имеет вид, представленный на рисунке 8.

Рис. 8. Страница журнала Nature с одной из публикаций

С помощью встроенных инструментов браузера можно идентифицировать каждый элемент страницы. Например, элемент, содержащий название публикации, имеет *css*-класс *c-article-title* (см. рис. 9), который можно использовать для нахождения этого элемента на странице и взятия его текста.

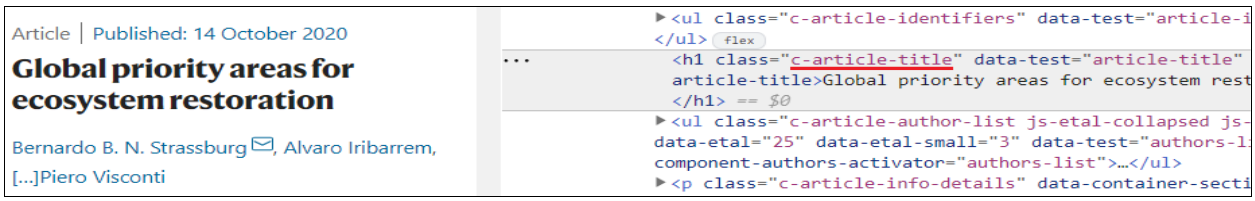

Рис. 9. Элемент страницы с названием публикации и его css-класс

Таким же образом на странице можно найти элементы с аннотацией статьи, ключевыми словами, именами авторов и названиями организаций. Названия организаций, к которым относятся авторы, на сайте журнала *Nature* можно увидеть на карточках авторов (см. рис. 10).

Рис. 10. Карточка автора с его аффилиациями

Таким образом, в каждом методе сбора данных со страниц конкретного журнала должно быть описано нахождение необходимых элементов и способ взятия их текста. Сопоставление журнала и названия метода сбора данных для него также должно быть описано. Все методы сбора данных со страниц журналов должны быть описаны в классе *Parser*.

Таким образом, программа-агент может состоять из четырех классов: *Agent*, *Parser*, *Repo*, *Service*, каждый из которых выполняет свои задачи: *Agent* отвечает за полный процесс сбора и сохранения данных, делегируя задачу сбора данных классу *Parser*, а сохранения данных – классу *Service*, кото-

рый в свою очередь преобразует данные и передает их для непосредственного сохранения в БД классу *Repo*. Диаграмма классов программы-агента представлена на рисунке 11.

Рис. 11. Диаграмма классов программы-агента

Проектирование веб-приложения

Для отображения собранных агентом данных и предоставления информации о публикационной активности стран БРИКС в удобном для анализа виде необходимо разработать веб-приложение.

Для создания приложения была выбрана трехуровневая клиент-серверная архитектура. Такая архитектура подразумевает наличие «тонкого» клиентского приложения для отображения данных, серверного приложения, выполняющего бизнес-логику системы, а также БД. Клиентское приложение не отвечает за бизнес-логику в системе и занимается исключительно отображением данных. Серверное приложение обрабатывает данные, оно ничего не знает о состоянии своих клиентов, но доступно каждому из них. БД занимается хранением данных и может выдавать их по запросу. Клиентское приложение формирует запросы к серверному приложению, чтобы получить данные. В свою очередь

серверное приложение формирует запросы к БД, которая возвращает требуемые данные, которые сервер затем отправляет клиенту для последующего отображения.

Преимущества выбранной архитектуры, в отличие от, например, двухуровневой, состоят в том, что каждый уровень выполняет свои функции и имеет минимальную нагрузку, все уровни могут быть расположены на разных физических серверах, добавлять новые возможности в приложение такой архитектуры проще. Также применение трехуровневой архитектуры снижает нагрузку на сеть, так как передаются только запросы и необходимые данные [4]. Таким образом выбранная архитектура в совокупности с правильно разработанной программой-агентом позволит создать систему, удовлетворяющую предъявленным требованиям надежности, безопасности и работоспособности.

Передача данных между клиентским и серверным приложениями осуществляется по сетевому протоколу *HTTP*. Сетевой протокол — это набор правил, определяющих, как осуществляется соединение и обмен данными между включенными в сеть устройствами. *HTTP* – это клиент-серверный протокол передачи данных прикладного уровня, который позволяет передавать и получать ресурсы, являясь основой любого обмена данными в *web*. *HTTP* подразумевает инициирование запросов клиентом, на которые отвечает сервер [5].

Данные между клиентским и серверным приложениями передаются в формате *JSON*. Преимущества этого формата в том, что он легко читается людьми и не зависит от технологий реализации приложений, которые обмениваются данными, сериализованными с помощью *JSON*. Формат представляет собой набор пар ключ-значение, где в качестве значений могут выступать как одиночные значения, так и списки, словари и т.д [6].

На рисунке 12 приведена диаграмма вариантов использования [7]. Можно выделить два объекта, которые взаимодействуют с проектируемой системой: администратора и пользователя. Пользователь должен иметь возможность просматривать статистику, а администратор – и просматривать ее, и запускать процесс обновления системы программой-агентом. Для запуска процесса обновления администратор должен быть аутентифицирован в системе.

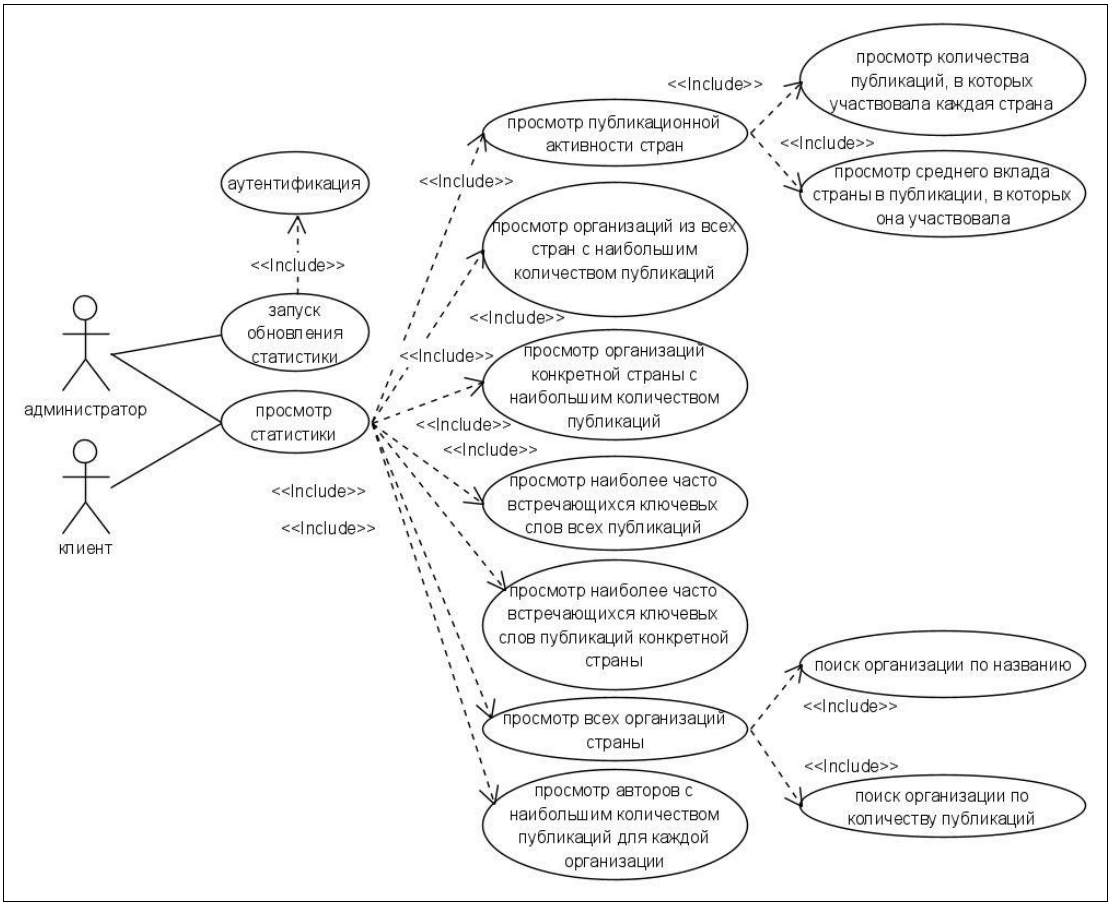

Рис. 12. Диаграмма вариантов использования

Каждый вариант использования требует отображения пользователю определенных данных, которые необходимо взять из системы хранения.

За отображение данных будет отвечать клиентское приложение, которое запрашивает данные у серверного приложения. В свою очередь серверное приложение берет данные из БД и возвращает их клиентскому приложению в определенном виде.

Серверное приложение будет включать в себя программный агент, являясь оберткой для него, предоставляющей возможность запускать работу агента и пользоваться собранными им данными. Также серверное приложение должно содержать функции, к которым клиентское приложение по определенным *URL* может обращаться для получения необходимых данных.

Клиентское приложение должно представлять из себя *Single Page Application* (одностраничное приложение), состоящее из одной страницы с заменяющими друг друга компонентами. Компоненты отображают определенные данные согласно вариантам использования, представленным на рисунке 12. Макеты компонентов для отображения данных были спроектированы с использованием инструмента *Figma* [8].

Так, макет компонента для просмотра публикационной активности стран представлен на рисунке 13. В последствии к компоненту может быть добавлена информация об организациях всех стран с наибольшим количеством публикаций, а также список наиболее встречающихся в публикациях всех стран ключевых слов и вклад страны в публикации, в которых она участвовала, в процентах.

Рис. 13. Макет компонента для просмотра публикационной активности стран

Макет компонента для просмотра организаций страны с наибольшим количеством публикаций, а также с наиболее часто встречающимися ключевыми словами в публикациях страны представлен на рисунке 14.

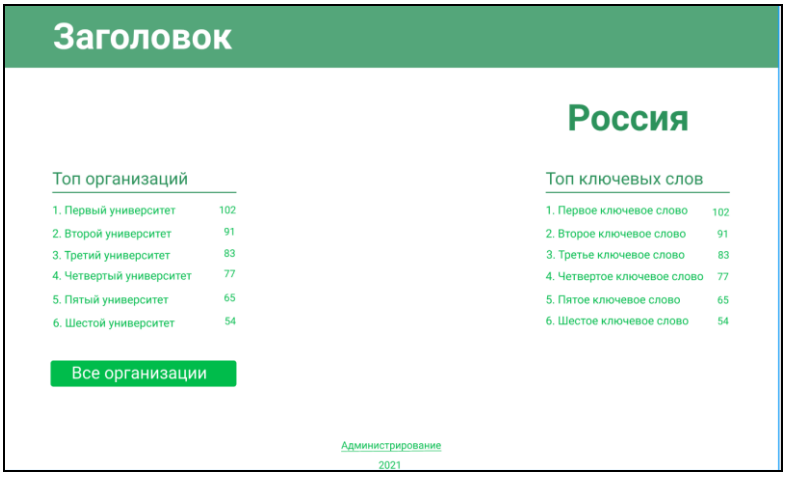

Рис. 14. Макет компонента для просмотра организаций и ключевых слов страны

По нажатию на кнопку «Все организации» должны отображаться все организации выбранной страны. Список организаций должен быть отсортирован по количеству имеющихся публикаций по убыванию. Макет компонента для просмотра всех организаций страны представлен на рисунке 15.

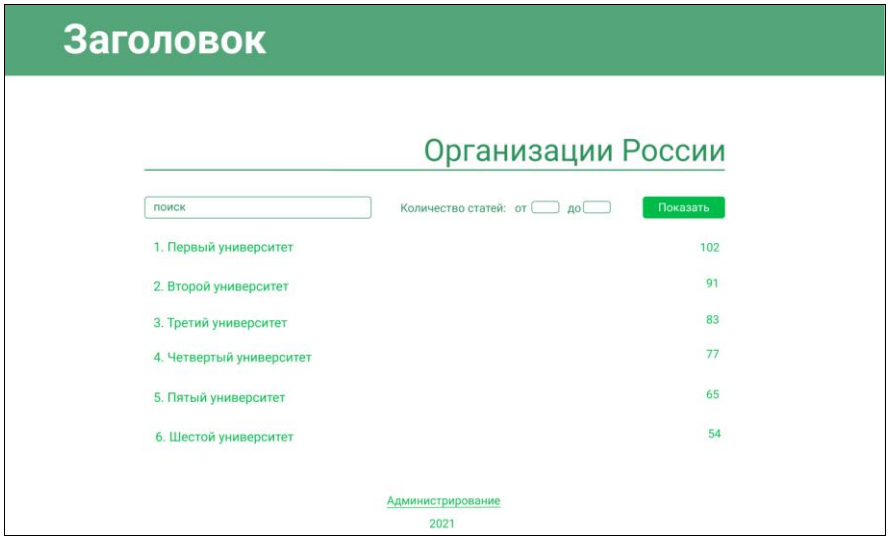

Рис. 15. Макет компонента для просмотра всех организаций страны

При нажатии на одну из организаций должен отображаться список авторов с наибольшим количеством публикаций. Список авторов выбранной организации должен быть отсортирован по количеству публикаций по убыванию. Макет компонента для просмотра авторов с наибольшим количеством публикаций конкретной организации представлен на рисунке 16.

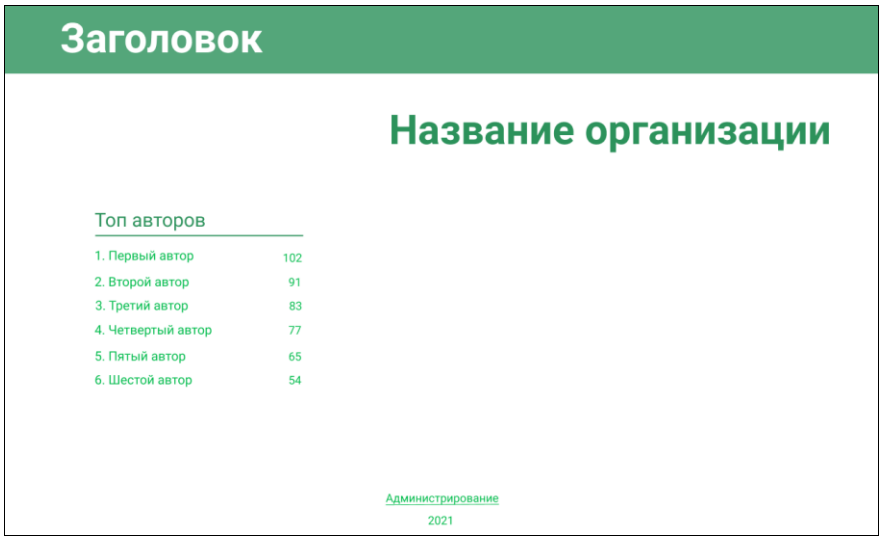

Рис. 16. Макет компонента для просмотра авторов с наибольшим количеством публикаций для конкретной организации

Макет компонента для аутентификации администратора в системе представлен на рисунке 17. После входа в систему администратор должен получить доступ к запуску процесса сбора данных. Макет компонента для запуска сбора данных представлен на рисунке 18. К этому компоненту может быть добавлена возможность выбора страны, для которой необходимо запустить сбор данных.

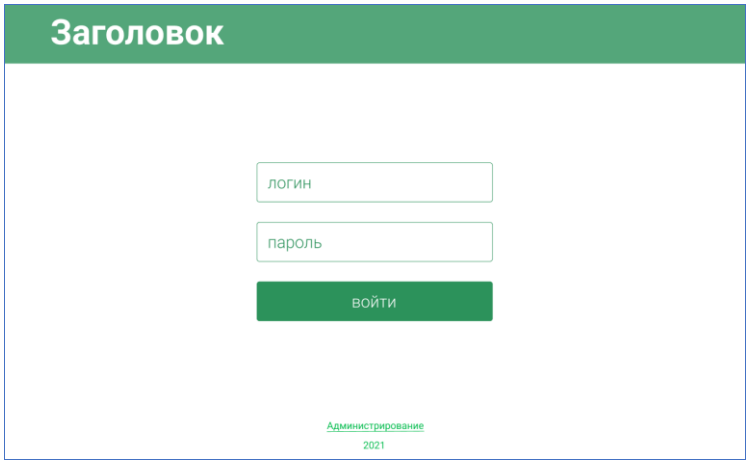

Рис. 17. Макет компонента для аутентификации

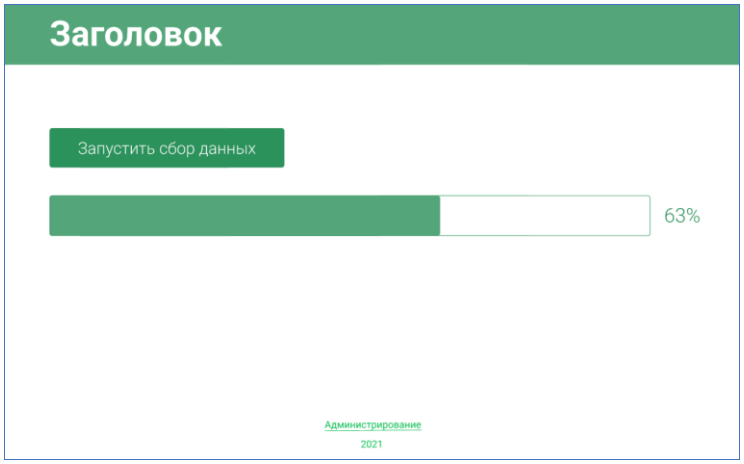

Рис. 18. Макет компонента для запуска сбора данных

Заключение

Для проведения анализа публикационной активности стран союза БРИКС была спроектирована система для сбора данных о научных публикациях с электронного ресурса *Nature Index* и представления их в удобном для пользователя виде.

Разработаны проекты для всех компонентов системы: описана модель БД, представлен алгоритм работы программы-агента и приведена диаграмма классов для ее дальнейшей реализации, обоснован выбор трехуровневой архитектуры веб-приложения и созданы макеты пользовательского интерфейса для отображения информации согласно выдвинутым требованиям и вариантам использования.

Результаты проектирования будут использоваться в дальнейшем для разработки описанной системы.

Список литературы

1. BRICS India 2021. URL: https://brics2021.gov.in/.
2. Nature Index. Springer Nature Limited, 2021. URL : https://www.natureindex.com/.
3. Яндекс Справка. ООО «Яндекс», 2015–2021. URL : https://yandex.ru/support/.
1. Хелпикс.Орг – Интернет помощник. helpiks.org — Хелпикс.Орг, 2014-2021. URL: https://helpiks.org/.
2. MDN Web Docs. Mozilla, 2005-2021. URL: https://developer.mozilla.org/.

3. JSON. URL: http://www.json.org/.
4. Буч Г., Рамбо Д., Якобсон И. Язык UML. Руководство пользователя. — ДМК Пресс, 2007.
5. Figma: the collaborative interface design tool. URL: https://www.figma.com/.