УДК 004.9
DOI 10.52575/2687-0932-2024-51-3-735-748

Цифровой инструмент автоматизации процессов сбора, хранения и обработки данных об инновационном развитии регионов

Бекетов С.М., Федяевская Д.Э., Схведиани А.Е., Редько С.Г., Бурлуцкая Ж.В.
Санкт-Петербургский политехнический университет Петра Великого
Россия, 195251, Санкт-Петербург, ул. Политехническая, д. 29
E-mail: salbek.beketov@spbpu.com

Аннотация. Данная работа посвящена разработке цифрового инструмента для автоматизации процессов сбора, хранения и обработки данных. В рамках исследования рассматриваются вопросы поддержания актуальности данных для цифровых инструментов исследования социально-экономических систем за счет адаптивных средств синхронизации данных с открытыми источниками. В исследовании поднимаются проблемы обработки больших данных, связанные с особенностями источников данных и соответствующих им проблем с качеством данных и непостоянством структур данных. На основании проанализированной информации разрабатываются функциональные требования к инструменту работы с данными с учетом особенностей источников данных, а также требований к использованию данных со стороны вычислительных моделей социально-экономических систем. Результатом работы является цифровой инструмент автоматизации процессов сбора, хранения и обработки данных, реализованный в качестве функционального модуля цифровой платформы инновационного развития регионов. Разработанное решение может быть адаптировано для аналогичных конфигурируемых систем хранения и обработки данных, в частности в цифровых платформах. Исследование реализовано в рамках проекта по разработке цифровой платформы региональной инновационной системы Российской Федерации как драйвера устойчивого развития.

Ключевые слова: цифровые инструменты, цифровая платформа, инновационное развитие регионов, автоматизированная загрузка, цифровые данные

Финансирование: исследование выполнено при поддержке Министерства науки и высшего образования Российской Федерации (государственное задание № 075-03-2024-004 от 17.01.2024).

Для цитирования: Бекетов С.М., Федяевская Д.Э., Схведиани А.Е., Редько С.Г., Бурлуцкая Ж.В. 2024. Цифровой инструмент автоматизации процессов сбора, хранения и обработки данных об инновационном развитии регионов. Экономика. Информатика. 51(3): 735–748. DOI 10.52575/2687-0932-2024-51-3-735-748

A Digital Tool for Automating the Processes of Collecting, Storing and Processing Data on the Innovative Development of Regions

Salbek M. Beketov, Darya E. Fedyaevskaya, Angi E. Skhvediani, Sergey G. Redko, Zhanna V. Burlutskaya
Peter the Great St. Petersburg Polytechnic University
29 Politechnicheskaya St, Saint-Petersburg 195251, Russia
E-mail: salbek.beketov@spbpu.com

Abstract. This work is devoted to the development of a digital tool for automating the processes of data collection, storage and processing. The research examines the issues of maintaining the relevance of data for digital tools for the study of socio-economic systems through adaptive means of synchronizing data with open sources. The study raises the problems of big data processing related to the peculiarities of data sources and corresponding problems with data quality and the variability of data structures. Based on the analyzed

information, functional requirements to the data management tool are developed, taking into account the characteristics of data sources, as well as data usage requirements for computational models of socio-economic systems. The result of the work is a digital tool for automating the processes of data collection, storage and processing, implemented as a functional module of the digital platform for innovative development of regions. The developed solution can be adapted for similar configurable data storage and processing systems, in particular, in digital platforms. The research was implemented as part of a project to develop a digital platform for the regional innovation system of the Russian Federation as a driver of sustainable development.

Keywords: digital tools, digital platform, innovative development of regions, automated download, digital data

Funding: The research is funded by the Ministry of Science and Higher Education of the Russian Federation (contract No. 075-03-2024-004 dated 17.01.2024).

For citation: Beketov S.M., Fedyaevskaya D.E., Skhvediani A.E., Redko S.G., Burlutskaya Z.V. 2024. A Digital Tool for Automating the Processes of Collecting, Storing and Processing Data on the Innovative Development of Regions. Economics. Information technologies, 51(3): 735–748. DOI 10.52575/2687-0932-2024-51-3-735-748

Введение

Цифровая трансформация затрагивает широкий спектр областей, включая социально-экономические исследования и соответствующие им интеллектуальные системы поддержки принятия решений [Зверева и др., 2019]. Этот процесс является приоритетным в рамках обеспечения глобального устойчивого развития, так как использование цифровых инструментов обеспечивает экономию времени и ресурсов, а также доступ к анализу больших объемов данных [Cenamor et al., 2019; Hasell et al., 2020], предоставляющих новые возможности для исследования систем.

Одним из инструментов цифровизации социально-экономических систем являются цифровые платформы, предоставляющие полный цикл работы с данными, включая: сбор, обработку, визуализацию и дальнейшее использование в вычислительных инструментах [Gorodetsky et al., 2019; Belov et al., 2021]. Однако точность цифровой платформы напрямую зависит от актуальности и качества данных, в свою очередь определяемого особенностями источников данных и системы взаимодействия с ними. Поскольку основными источниками данных об инновационном развитии регионов являются государственные офисы статистики, необходимо разработать эффективные методы сбора, хранения и обработки информации, соответствующие особенностям источников данного типа [Winther et al., 2019; Mathieu et al., 2021]. Стоит отметить, что источники данного типа содержат в основном неструктурированные динамические данные [Mehmood, Anees, 2022], что повышает риски возникновения ошибок при их сборе и обработке.

Автоматизация этих процессов сбора, хранения и обработки данных для цифровых платформ происходит на нескольких уровнях (уровень сбора данных, уровень хранения данных, уровень обработки данных, уровень представления данных) и включает в себя разработку взаимосвязанных модулей, обеспечивающих эффективную работу платформы в целом. В рамках проекта по разработке цифровой платформы региональной инновационной системы Российской Федерации как драйвера устойчивого развития уже разработан прототип цифровой платформы региональной инновационной системы Российской Федерации как драйвера устойчивого развития [Bolsunovskaya et al., 2023], и на данном этапе необходима ее доработка в части расширения функциональности за счет нового модуля, предназначенного для сбора, хранения и обработки данных об инновационном развитии регионов.

Таким образом, целью данной работы является разработка модуля для обеспечения сбора, хранения и обработки данных для цифровой платформы анализа инновационного

развития регионов. В рамках данной цели были решены следующие задачи: определение функциональных требований к цифровому инструменту автоматизации, определение исходных данных, разработка модуля анализа и обработки данных, разработка модуля автоматизированной загрузки данных.

Материалы и методы исследования

Перед началом разработки автоматизированной системы сбора, хранения и обработки данных для цифровой платформы анализа инновационного развития регионов необходимо определение функциональных требований к платформе, определение формата исходных данных, анализ цикла работы с ними и определение особенностей сбора и обработки данных для социально-экономического моделирования.

Исходные данные

Источники данных играют важную роль в научных исследованиях, представляя собой цифровые или физические хранилища, в которых хранятся информационные материалы различных форматов, включая таблицы, файлы и прочее. Для целей исследований источники данных обычно классифицируются на внутренние и внешние.

Внутренние данные находятся внутри системы, доступной для исследователей. В контексте социально-экономических и социотехнических систем внутренние источники данных могут включать в себя бухгалтерскую отчетность, отчеты отделов, экспертные мнения внутренних специалистов, данные, собранные с датчиков и прочее.

Внешние данные – это данные, собранные сторонними субъектами, хранятся во внешних источниках. Особенностью внешних данных является их возможная труднодоступность (в том числе за плату) и неполнота, а также низкое качество, связанное с ограниченной информацией о методах сбора и обработки данных. Внешние источники могут включать государственные и негосударственные публикации, услуги синдикатов и прочие внешние источники информации. Также при использовании внешних данных необходимо учитывать неизвестность методов сбора и обработки, которые могут сказаться на результатах анализа [Талканбаева, 2019].

Источники данных могут быть получены посредством трех основных методов: экспериментального, эмпирического (на основе исторических данных) и экспертного.

Экспериментальный метод основан на непосредственных наблюдениях и измерениях в реальном времени. Этот метод часто применяется в случае небольших систем, где объекты исследования доступны непосредственно для наблюдения. Однако данный подход требует значительных затрат времени и ресурсов, особенно в случае работы с комплексными системами.

Эмпирический метод базируется на использовании исторических данных, которые могут быть записаны и подвергнуты анализу. В контексте цифровизации различных аспектов жизни человека объемы данных, доступных для анализа, существенно увеличились. Эти данные могут быть использованы для создания моделей с использованием двух основных подходов: прямого использования исторических данных и построения теоретико-вероятностных распределений на их основе. Второй подход предпочтителен, так как позволяет учитывать различные статистические свойства данных. Однако он требует значительного объема данных и может потребовать сложной их обработки.

Экспертный метод основан на оценках экспертов. В случаях, когда исторические данные недоступны или неприменимы, а также для новых систем, оценки экспертов могут служить единственным источником формализованных данных. Существуют различные подходы к использованию экспертных оценок, включая метод Дельфи, который предполагает получение согласованных оценок от группы экспертов путем итеративного опроса и обсуждения. Этот метод может быть особенно полезен для моделирования

крупных систем, где доступ к непосредственным данным ограничен или отсутствует [Гинцяк и др., 2023].

При анализе крупных социально-экономических систем большую часть данных можно получить эмпирически из внешних источников: офисов статистики и отчетов органов исполнительной власти. При рассмотрении в качестве объекта исследования региональной инновационной системы следует обратить внимание на индексы научно-технологического и инновационного развития и данные, которые в них используются. Такими рейтингами являются следующие: Индекс научно-технологического развития, Национальный рейтинг научно-технологического развития субъектов Российской Федерации, Рейтинг инновационного развития субъектов Российской Федерации, Рейтинг инновационных регионов России SMART. Данные, используемые в рейтингах, пересекаются, что связано с опорой на открытые источники данных: Росстат, ЕМИСС, Минобрнауки России, Минпросвещения России, «Научная электронная библиотека» (e-library), Минэкопромразвития России. Такие источники данных содержат информацию предварительно обработанную (агрегированную) и представленную в виде цифровых документов (машиночитаемого формата).

При работе с одним внешним источником данных, как правило, этап предварительной обработки может быть пропущен, поскольку данные, полученные из этого источника, часто уже прошли процесс очистки и агрегации. Однако, когда речь идет об анализе сложных социально-экономических систем, важно понимать, что такие системы включают в себя множество взаимосвязанных переменных и факторов, которые могут быть представлены различными источниками данных.

Использование различных источников данных является ключевым для получения более полного и точного представления о системе. Однако эти данные могут различаться по различным качественным характеристикам, таким как формат, точность, полнота и даже надежность. Точность данных также может варьироваться в зависимости от того, как они были собраны и обработаны.

Именно поэтому при анализе сложных социально-экономических систем необходимо проводить комплексное исследование, которое учитывает различные источники данных. Этот процесс требует дополнительного этапа обработки, включающего в себя сравнение, сопоставление, стандартизацию и, возможно, даже объединение данных из разных источников. Такой подход позволяет учесть разнообразие данных и уменьшить искажения или ошибки, которые могут возникнуть из-за различий в данных.

Цикл работы с данными

В рамках разработки системы сбора, обработки и хранения данных необходимо определить цикл работы с данными в рамках эксплуатации цифровой платформы. Цикл работы с данными представлен на рис. 1.

Сбор данных представляет собой начальный этап цикла работы с данными, где информация собирается из различных источников. Сбор данных может включать в себя проведение экспериментов, анкетирование, интервью, наблюдение, а также использование открытых источников данных или доступ к уже существующим базам данным [Баклыская и др., 2023]. Для обеспечения повторяемости и воспроизводимости исследования необходима тщательная документация процесса сбора данных, включающая в себя описание источников, методов и условий сбора.

После сбора данных происходит их анализ и обработка на предмет целостности, точности и соответствия заранее установленным критериям, то есть парсинг и валидация данных. Данный этап включает в себя удаление дубликатов, исправление ошибок, преобразование данных в удобный формат и проверку на наличие недостающих или аномальных значений [Макаров, Намиот, 2023]. Отдельное внимание уделяется проверке качества данных и их соответствию ожидаемым стандартам и требованиям исследования.

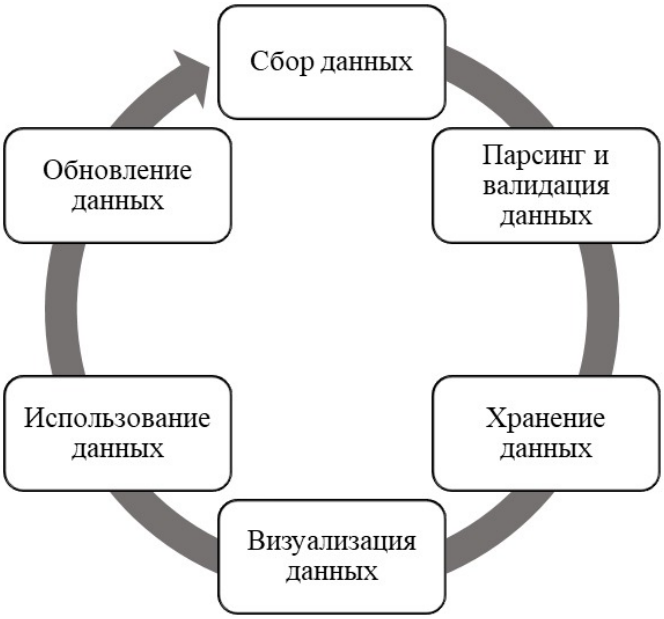

Рис. 1. Цикл работы с данными
Fig. 1. The data processing cycle

Для эффективного управления данными требуется выбор подходящего механизма их хранения, например, реляционная или нереляционная база данных, файловая система или облачное хранилище. При выборе формата и структуры хранения учитываются объем и типы данных, а также требования к доступу и безопасности [Хомоненко и др., 2023]. Важно также обеспечить резервное копирование данных для предотвращения потери информации.

Для более глубокого анализа данных важно визуализировать их, чтобы выявить закономерности, тренды и взаимосвязи. Визуализация может включать построение графиков, диаграмм, тепловых карт и других графических представлений данных. Визуализация позволяет исследователям лучше понимать структуру и характеристики данных, а также делиться результатами исследования с другими заинтересованными сторонами.

После обработки и визуализации данные готовы к использованию для достижения целей исследования. Возможно применение статистических методов, машинного обучения или других аналитических подходов для извлечения знаний из данных [Саханевич, 2020]. Результаты анализа могут быть использованы для формулирования гипотез, выявления закономерностей, подтверждения или опровержения теорий, а также принятия решений в различных областях науки и практики.

Цикл работы с данными является итеративным процессом, поэтому важно регулярно обновлять данные в соответствии с поступающей информацией или изменениями в окружающем мире [Хромова, Петросян, 2023]. Данный этап может включать в себя добавление новых данных, коррекцию ошибок или обновление алгоритмов обработки данных для улучшения качества и точности исследования. Обновленные данные позволяют улучшать результаты исследования и адаптировать их к изменяющимся условиям и требованиям.

Функциональные требования

В соответствии с особенностями выбранных источников данных и цикла работы с данными в рамках исследования социально-экономических систем были определены основные функциональные требования для инструмента сбора, хранения и обработки данных: настраиваемая автоматизированная загрузка данных из открытых источников офисов официальной статистики; парсинг и структурная валидация загружаемых данных (соответствие шаблонам и структурам текущих данных; семантическая валидация данных (соответствие ожидаемым значениям); интеграция новых данных в единую структуру

данных; хранение и идентификация данных; обеспечение доступа к данным (посредством специализированных запросов); визуализация данных; фильтрация данных; запись данных, полученных в результате имитационных экспериментов.

Таким образом, система должна обеспечивать автоматизированную загрузку данных, первичную обработку, хранение данных, доступ к данным из интерфейса платформы, фильтрацию данных и обработку запросов [Ricciato et al., 2019]. Предметная область не требует хранения всех данных о сущностях (как это происходит в информационно-поисковых системах) [Constantinides et al., 2018]. Следовательно, должен быть определен набор данных, необходимых для программной обработки и получения в результате массивов заданных показателей [Jovanovic et al., 2022].

Результаты и их обсуждение

Разработка модуля загрузки, анализа и обработки данных

В системе работы с данными платформы анализа социально-экономических систем необходимо производить двухуровневый процесс обработки. Модуль загрузки данных производит цикл ETL: сбор данных из источников, валидация и обработка, запись в хранилище. Затем модуль анализа собирает данные из базы данных для применения методов математического и статистического анализа данных, записывает в базу данных в отдельные отношения. Модуль обработки данных производит выборки данных для их визуализации пользователю на платформе. Так как двухуровневая система работы с данными имеет особенность неоднократной записи данных в хранилище, необходимо обеспечить разработку специальной структуры базы данных для устранения аномалий добавления, удаления и изменения данных.

ETL-процесс (extract – transform – load) направлен на автоматизацию процесса сбора, обработки и записи данных в хранилище. Сбор производится путем обращения к одному или нескольким источников данных. Обработка данных необходима для выборки данных, очистки и преобразования данных под необходимый формат.

ETL-процессы применяются в контексте формирования электронной демографии в качестве эффективного инструмента для социальных исследований и мониторинга данных о населении. Для успешной реализации данной задачи предлагается использовать технологию ETL с целью извлечения данных из различных источников и последующей консолидации. Предложенная архитектура многоуровневой системы обеспечивает возможность оптимизации процесса загрузки новых данных, что, в свою очередь, позволяет осуществлять анализ в реальном времени. Данная методология включает в себя изменение отношений данных или обновление значений с целью определения влияния различных изменений на демографическую ситуацию [Алгулиев и др., 2019; Шикульский и др., 2022].

ETL-процессы также используются в отрасли розничной торговли, играют ключевую роль в анализе данных о закупках и продажах товаров. Например, при анализе данных о продажах магазина процесс извлечения (Extract) включает сбор информации о продажах из базы данных OLTP, содержащей транзакционные данные о каждой покупке. Затем данные подвергаются преобразованию (Transform), включая очистку, агрегацию, объединение с другими данными и применение бизнес-правил для подготовки к анализу. Наконец, подготовленные данные загружаются (Load) в аналитическую платформу, где они используются для проведения различных аналитических операций, включая прогнозирование спроса, анализ эффективности продаж и оптимизацию ассортимента товаров. Такой подход к ETL-процессам помогает компаниям в розничной торговле принимать обоснованные бизнес-решения на основе актуальных и точных данных о продажах и потребительском поведении [Mehmood, Anees, 2022].

После сбора данных из источников важной составляющей является проведение валидации – проверки данных на соответствие формату, полноте, отсутствию дубликатов.

Она может быть разделена на две составляющие: структурную и семантическую. Методы структурной валидации оценивают соответствие качественных параметров данных. В данном процессе используются методы для валидации: типов данных (данные из одного источника могут отличаться по точности: целые и дробные значения, проверка типов данных позволяет привести значения к единому формату); зависимостей и связей (проверка позволяет оценить наличие ранее установленных связей, таких как заполнение поля только в случае полноты другого); структуры входных файлов (наличию необходимых полей во входных файлах); ссылочной целостности.

Структурная валидация является предварительным шагом перед мэппингом данных (data mapping). На основе шаблона мэппинга и проведенных проверок производится трансформация исходных данных в определенный ранее исходный формат. Семантическая валидация проверяет, соответствуют ли данные ожидаемым значениям. При этом проверяется, например, соответствие ограничениям и наличие выбросов.

Загрузка данных является заключительным этапом записи данных в хранилище. Необходимо обеспечить механизм, позволяющий откатывать транзакцию в хранилище данных при возникновении ошибок для обеспечения целостности хранилища. Описание процесса представлено на рис. 2.

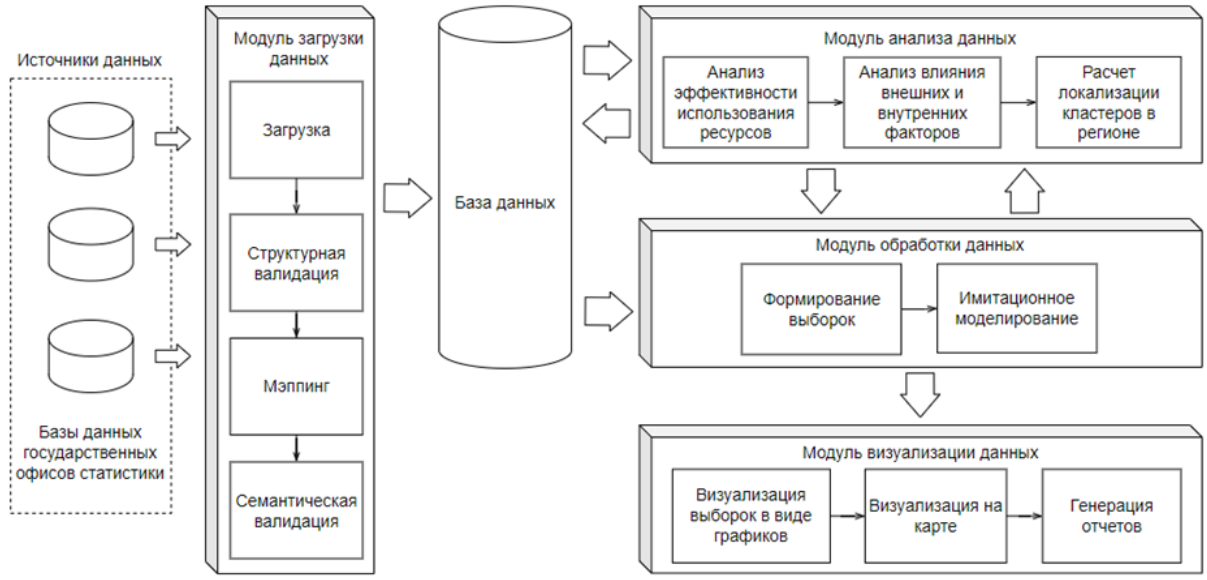

Рис. 2. Описание процесса загрузки, анализа и обработки данных
Fig. 2. Description of the data loading, analysis and processing process

Таким образом, цикл обработки данных начинается с загрузки данных с офисов официальной статистики [Winther et al., 2019], таких как РОССТАТ, ЕМИСС. Основной задачей официальных офисов статистики является сбор, обработка и анализ данных для определения состояния экономики и социальных процессов в стране. Они собирают и агрегируют большие наборы данных [Shastitko, Markova, 2020], которые потенциально могут быть использованы в социально-экономических исследованиях, в том числе моделировании социально-экономических систем. Исследования могут быть проведены над всеми сферами, которые отражены в отчетах офисов статистики.

В модулях анализа и обработки данных платформа обращается к хранилищу данных, которое является ядром систем. Модуль анализа данных включает в себя анализ эффективности и использования ресурсов, анализ влияния внешних и внутренних факторов и расчет локализации кластеров в регионе. Проходя через модуль обработки, данные могут быть уже визуализированы (в виде графиков, на карте, в отчетах). В хранилище должны находиться данные об объекте исследования (регион, город, сфера деятельности и т. д.) для

временного периода исследования. Таким образом, необходим инструмент добавления новых кортежей данных в хранилище, при этом обновление и изменение уже записанных данных не предусматривается. Такая система загрузки укладывается в аддитивную модель хранилища [Gorodetsky et al., 2019].

Разработка модуля автоматизированной загрузки данных

При управлении платформой могут быть рассмотрены несколько вариантов загрузки данных в хранилище:

– загрузка данных вручную (сотрудник заходит на сайты официальных офисов статистики, производит выгрузку необходимых файлов, обрабатывает их, производит загрузку в базу данных, обращаясь к системе управления базой данных);

– автоматизированная загрузка данных (система производит этап загрузки данных с официальных офисов статистики полностью без участия сотрудника);

– частичная автоматизация процесса (система производит какую-то часть процесса, сотрудник привлекается на подпроцессы).

Автоматизированная загрузка данных – достаточно сложная и комплексная задача, требующая разработки архитектуры модуля с учетом всех возможных операций с данными. Однако автоматизированная загрузка данных позволяет снизить потребление ресурсов на сопровождение платформы, а также исключить возможные ошибки, связанные с человеческим фактором. Преимущества такого способа загрузки данных обладают значительной ценностью для проекта в целом. На данном этапе исследования было принято решение спроектировать и реализовать модуль автоматизированной загрузки данных.

Основные связи модуля – это связь с базой данных цифровой платформы и с источником данных – официальными офисами статистики. Дополнительное управление модулем осуществляется пользователем платформы через пользовательский интерфейс, таким образом добавляется еще две связи.

Модуль загрузки данных должен выполнять следующие функциональности:

– взаимодействие с модулем конфигурации, что является частью модуля загрузки данных, который позволяет настроить параметры и конфигурации загрузки данных: источники, формат, режим загрузки и т. д.;

– соединение с государственными офисами статистики и сбор данных;

– парсинг – это процесс обработки и разбора информации, хранимой в источнике данных для её дальнейшей загрузки в хранилище;

– построение модели данных предполагает определение сущностей и связей между ними;

– проведение тестов, что позволяет проверить качество данных: соответствие формату, точность, полнота, отсутствие дубликатов и т. п.;

– проведение транзакции в хранилище данных. Модуль загрузки создает транзакцию, вносит изменения в базу данных и выполняет транзакцию; при возникновении ошибок модуль отменяет все изменения, что гарантирует целостность базы.

Изложенные функциональности модуля загрузки данных охватывают важные аспекты процесса сбора, обработки и сохранения информации. Однако среди них выделяются несколько блоков, которые могут потенциально вызвать проблемы в процессе реализации.

В рамках соединения с государственными офисами статистики и сбора данных причиной проблемы является то, что государственные структуры часто имеют ограниченный доступ к данным из-за политики безопасности, технических ограничений или сложностей в обмене информацией. Вследствие этого наблюдается непредсказуемость форматов данных, изменения в структуре информации, сложности в обработке больших объемов данных, а также соблюдении требований по безопасности и конфиденциальности.

Парсинг данных также является наиболее проблемным, так как разнообразие источников данных приводит к необходимости создания универсальных алгоритмов парсинга, что может быть сложно из-за непредсказуемой структуры данных. Данная проблема может быть охарактеризована неполной или некорректной документацией источников данных, изменениями в форматах или структуре данных без предварительного уведомления, а также обработкой специфических типов данных.

Также проблемой может являться проведение транзакции в хранилище данных, так как гарантирование целостности данных требует точного контроля за транзакциями. Появляется необходимость внедрения надежных механизмов отката транзакций при сбоях, обеспечение достаточной производительности при выполнении транзакций, а также обработка конфликтов при параллельных транзакциях.

Решение этих проблем требует тщательного анализа требований, разработки надежных алгоритмов и реализации механизмов обработки и восстановления при ошибках. Также важно учитывать технические особенности и специфику работы с каждым источником данных, а также предусмотреть механизмы мониторинга и отчетности для своевременного выявления и устранения проблем.

Механизмы мониторинга, а точнее обращение пользователя к модулю загрузки, происходит через интерфейс. Через него модуль сообщает о возникающих ошибках: соединения с сайтом офиса статистики, формата данных, записи в базу. Алгоритм автоматической загрузки данных описан с помощью UML-диаграммы и представлен на рис. 3.

Разработанный модуль загрузки данных обращается к сайтам официальных офисов статистики, производя проверку на наличие новых документов. Затем при обнаружении обновлений модуль производит их выгрузку и парсинг. На основании полученных данных строится модель данных, включающая в себя структуру данных (описание основных сущностей и их атрибутов, таких как таблицы, поля данных, связи между ними и их типы данных), поля данных, связи между таблицами, типы данных и ограничения, правила целостности.

После построения модели данных и загрузки информации из официальных источников, производится тестирование, которое представляет собой важный этап для обеспечения качества данных. На данном этапе осуществляется сравнение выгруженных данных с оригинальными источниками (например, сайтами официальных офисов статистики), чтобы удостовериться в точности и полноте выгрузки. Это включает в себя сопоставление значений показателей, дат и других характеристик данных. Проверяется также соответствие формата и типов данных, загруженных из источников, требованиям и ожидаемым структурам в модели данных. Осуществляется анализ данных на наличие дубликатов, то есть повторяющихся записей, которые могут исказить результаты анализа. Если обнаружены дубликаты, принимаются меры по их удалению или коррекции. Проводится также анализ на предмет наличия ошибок или неточностей в данных, включая проверку на наличие неожиданных значений, выбросов или пропущенных данных.

На основании результатов тестирования система принимает решение о том, следует ли загружать данные в хранилище или нет. Если данные прошли проверку успешно и не содержат серьезных ошибок, их можно передать в базу данных для дальнейшего анализа и использования. В случае выявления проблем, данные могут быть отклонены, и требуется корректировка или повторная загрузка. Таким образом, проведение тестов на этом этапе гарантирует высокое качество данных, что является ключевым аспектом в процессе автоматизации сбора и обработки информации.

Возвращаясь к вопросу об актуальности разработки модуля автоматической загрузки данных, стоит отметить, что загрузка данных вручную является причиной возникновения ряда проблем, например: ошибка ввода, необходимость постоянного мониторинга сайтов государственных офисов статистики в ожидании обновления документов, временные затраты сотрудников. Разработка и внедрение системы загрузки данных позволяет снизить совокупную стоимость владения и эксплуатации цифровой платформы [Лычагин, Позин,

2011]. Более того, разработка подобных модулей уже ведется и для других интеллектуальных систем управления социально-экономическими системами [Энгель, Энгель, 2018; Романчуков и др., 2020], что также подтверждает актуальность и востребованность подобных исследований.

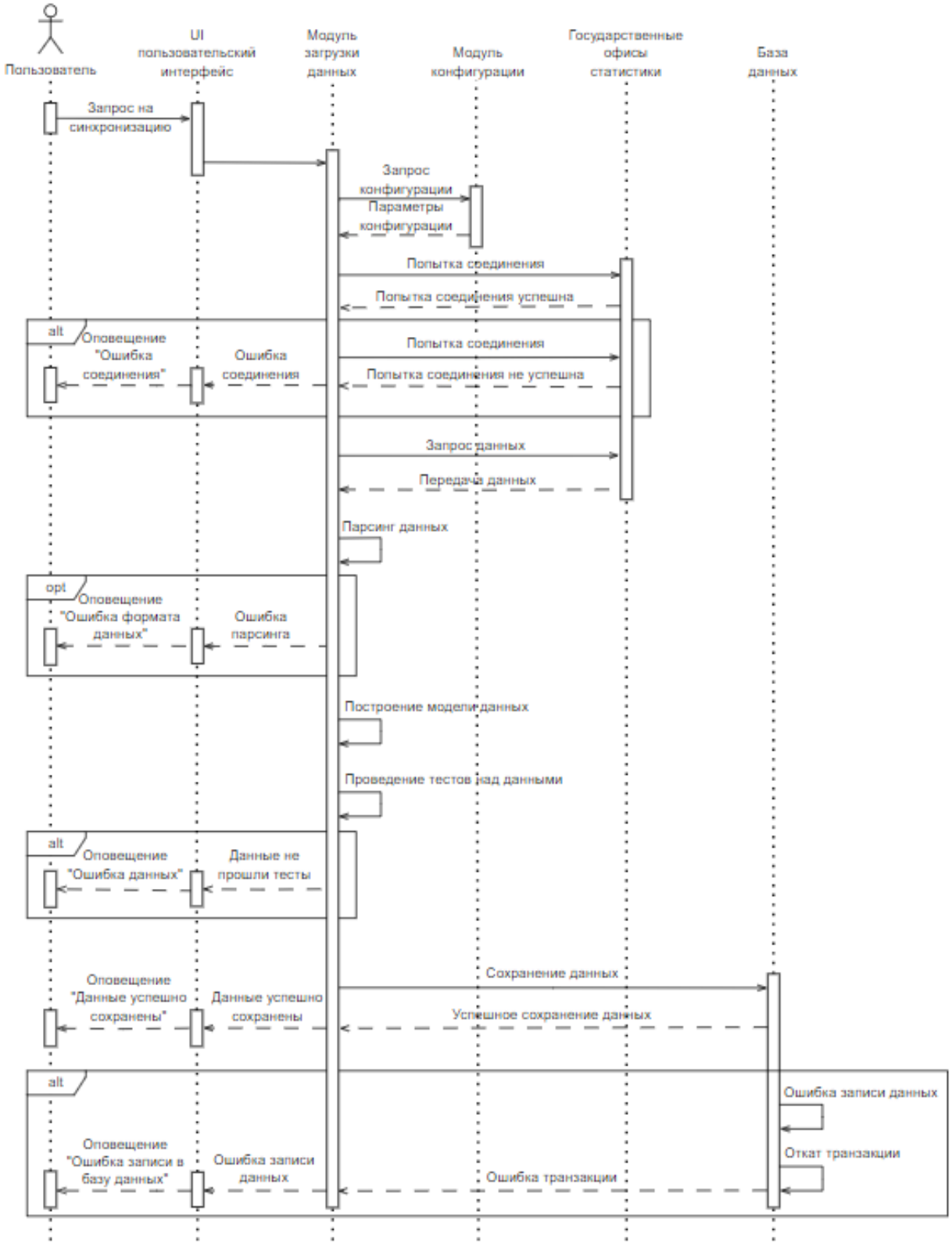

Рис. 3. UML-диаграмма последовательности работы модуля загрузки данных
Fig. 3. UML diagram of the sequence of the data loading module operations

Обсуждение результатов исследования

Преимуществами представленного решения в сравнении с аналогами [Энгель, Энгель, 2018; Романчуков и др., 2020] являются интегрированные инструменты валидации данных, как структурной, так и семантической, что обеспечивает устойчивость системы – изменение структуры файла или неожиданные значения не приведут к ее сбою. Однако стоит отметить и текущие недостатки разработанной системы. На данном этапе разработки модуль не обрабатывает некоторое количество ошибок. Первая ошибка связана с форматом данных: система не добавит данные другого формата, поэтому эти ошибки будут обрабатываться человеком. Еще одним аспектом является обработка ошибок. Сейчас обработка замыкается на пользователе, которому приходят уведомления.

Перспективой доработки модуля является подключение нейросетевых алгоритмов по примеру исследования [Романчуков и др., 2020] для распознавания полей (заголовков) и последующая смена формата.

Заключение

В рамках данной работы рассматриваются преимущества применения цифровых технологий для автоматизации процессов сбора, хранения и обработки данных с учетом особенностей источников данных для исследования социально-экономических систем, а также соответствующих им проблем с качеством данных и непостоянством структур данных. Результатом работы является цифровой инструмент автоматизации процессов сбора, хранения и обработки данных, реализованный в качестве функционального модуля цифровой платформы инновационного развития регионов. Разработанное решение обеспечивает автоматизированное обращение к сайтам государственных офисов статистики, экспорт и обработку данных, проверку и последующую загрузку данных в хранилище.

Решаемая проблема, связанная с созданием автоматизированного модуля сбора, хранения и обработки данных для цифровых инструментов исследования социально-экономических систем, в частности цифровых платформ, представляет собой типовую задачу, а значит используемые методы и подходы, разработанные в ходе этой работы, могут быть успешно адаптированы и применены для других проектов. Так, данное исследование не только способствует развитию конкретной платформы, но и представляет ценный опыт, который может быть широко применен в других проектах и областях деятельности.

Исследование реализовано в рамках проекта по разработке цифровой платформы региональной инновационной системы Российской Федерации как драйвера устойчивого развития.

Список литературы

Алгулиев Р.М.О., Алыгулиев Р.М.О., Юсифов Ф.Ф.О., Алекперова И.Я.Г. 2019. Формирование электронной демографии как эффективного инструмента социальных исследований и мониторинга данных о населении. Вопросы государственного и муниципального управления, (4): 61–86.

Баклыская Л.Е., Чукмарева Е.А., Фишева Н.О. 2023. Атриум на территории университетского кампуса. Урбанистика, (3): 44–59. DOI: 10.7256/2310-8673.2023.3.43888.

Гинцяк А.М., Бурлуцкая Ж.В., Федяевская Д.Э., Поспелов К.Н., Ракова В.В. 2023. Цифровое моделирование социотехнических и социально-экономических систем. DOI.10.18720/SPBPU/2/i23-253.

Зверева А.А., Беляева Ж.С., Сохаг К. 2019. Влияние цифровизации экономики на благосостояние в развитых и развивающихся странах. Экономика региона, 15(4): 1050–1062. DOI: 10.17059/2019-4-7.

Лычагин К.А., Позин Б.А. 2011. Снижение совокупной стоимости владения информационно-аналитической системой за счет создания системы интеграции данных. Открытое образование, (2): 238–242.

Макаров А.В., Намиот Д.Е. 2023. Обзор методов очистки данных для машинного обучения. International Journal of Open Information Technologies, 11(10): 70–78.

Романчуков С.В., Лызин И.А., Марухина О.В. 2020. Информационная система для анализа и моделирования социального и экономического развития региона. Информационные и математические технологии в науке и управлении, 3(19): 96–104.

Саханевич Д.Ю. 2020. Исследование подходов и методов применения искусственного интеллекта и машинного обучения в социально-экономических процессах. Вестник Омского университета. Серия «Экономика», 18(2): 65–79. DOI: 10.24147/1812-3988.2020.18(2).65–79.

Талканбаева Р.А. 2019. Цифровизация должна начинаться с регионов. Вестник Академии государственного управления при Президенте Кыргызской Республики, (26): 38–41.

Хомоненко А.Д. 2023. О надежности и доступности объектных хранилищ данных. Интеллектуальные технологии на транспорте, (S1): 123–128.

Хромова А.Р., Петросян Л.Э. 2023. Анализ уязвимостей в системах безопасности данных. Инженерный вестник Дона, (6 (102)): 67–76.

Шикульский М.И., Медведева О.В., Баркова В.М., Плешакова Л.А. 2022. Применение ETL-процессов для автоматизации анализа данных по розничным продажам. Инженерно-строительный вестник Прикаспия, 4 (42): 108–113. DOI: 10.52684/2312-3702-2022-42-4-108-113.

Энгель Е.А., Энгель, Н.Е. 2018. Модернизация программного модуля «Загрузка данных для интеллектуальной модели». Вестник Хакасского государственного университета им. Н.Ф. Катанова, (23): 25–30.

Belov S., Ilina A., Javadzade J., Kadochnikov I., Korenkov V., Pelevanyuk I., Semenov R., Zrelov P., Tarabrin V. 2021. Analytical platform for socio-economic studies. In CEUR Workshop Proceedings, 9: 619–623. DOI: 10.54546/MLIT.2021.81.99.001.

Bolsunovskaya M.V., Kudryavtseva T.Y., Rudskaya I.A., Gintciak A.M., Zhidkov D.O., Fedyaevskaya D.E., Burlutskaya Z.V. 2023. Digital Platform for Modeling the Development of Regional Innovation Systems of Russian Federation. International Journal of Technology, 14(8): 1779–1789. DOI: 10.14716/ijtech.v14i8.6843.

Cenamor J., Parida V., Wincent J. 2019. How entrepreneurial SMEs compete through digital platforms: The roles of digital platform capability, network capability and ambidexterity. Journal of Business Research, 100: 196–206. DOI: 10.1016/j.jbusres.2019.03.035.

Constantinides P., Henfridsson O., Parker G.G. 2018. Introduction—platforms and infrastructures in the digital age. Information Systems Research, 29(2): 381–400. DOI: 10.1287/isre.2018.0794.

Gorodetsky V.I., Laryukhin V.B., Skobelev P.O. 2019. Conceptual Model of a Digital Platform for Cyber-Physical Management of a Modern Enterprises Part 1. Digital Platform and Digital Ecosystem. Mekhatronika, Avtomatizatsiya, Upravlenie, 20(6): 323–332. DOI: 10.17587/mau.20.323-332.

Hasell J., Mathieu E., Beltekian D., Macdonald B., Giattino C., Ortiz-Ospina E., Roser M., Ritchie H. 2020. A cross-country database of COVID-19 testing. Scientific Data, 7(1): 345. DOI: 10.1038/s41597-020-00688-8.

Jovanovic M., Sjödin D., Parida V. 2022. Co-evolution of platform architecture, platform services, and platform governance: Expanding the platform value of industrial digital platforms. Technovation, 118: 102218. DOI: 10.1016/j.technovation.2020.102218.

Mathieu E., Ritchie H., Ortiz-Ospina E., Roser M., Hasell J., Appel C., Giattino C., Rodes-Guirao. 2021. A global database of COVID-19 vaccinations. Nat Hum Behav, (5): 947–953. DOI: 10.1038/s41562-021-01122-8.

Mehmood E., Anees T. 2022. Distributed real-time ETL architecture for unstructured big data. Knowledge and Information Systems, 64(12): 3419–3445. DOI: 10.1007/s10115-022-01757-7.

Ricciato F., Wirthmann A., Giannakouris K., Skaliotis M. 2019. Trusted smart statistics: Motivations and principles. Statistical Journal of the IAOS, 35(4): 1–15. DOI: 10.3233/SJI-190584.

Shastitko A.E., Markova O.A. 2020. An old friend is better than two new ones? Approaches to market research in the context of digital transformation for the antitrust laws enforcement. Voprosy Ekonomiki, (6): 37–55. DOI: 10.32609/0042-8736-2020-6-37-55.

Winther K.T., Hoffmann M.J., Boes J.R., Mamun O., Bajdich M., Bligaard T. 2019. Catalysis-Hub. org, an open electronic structure database for surface reactions. Scientific data, 6(1): 75. DOI: 10.1038/s41597-019-0081-y.

References

Alguliyev R.M., Aliguliyev R.M., Yusifov F.F., Alekperova I.Y. 2019. Developing Electronic Demography as an Effective Tool for Social Research and Monitoring Population Data. Public Administration Issue, (4): 61–86. (in Russian).

Baklyskaia L.E., Chukmareva E.A., Fisheva N.O. 2023. Atrium na territorii universitetskogo kampusa [Atrium on the university campus]. Urbanistika, (3): 44–59. DOI: 10.7256/2310-8673.2023.3.43888.

Gintsyak A.M., Burlutskaya ZH.V., Fedyaevskaya D.E., Pospelov K.N., Rakova V.V. 2023. Digital modeling of sociotechnical and socio-economic systems. DOI.10.18720/SPBPU/2/i23-253. (in Russian).

Zvereva A.A., Belyaeva Zh.S., Sohag K. 2019. Impact of the Economy Digitalization on Welfare in the Developed and Developing Countries. Ekonomika regiona, 15(4): 1050–1062. DOI: 10.17059/2019-4-7. (in Russian).

Lychagin K.A., Pozin B.A. 2011. Snizhenie sovokupnoj stoimosti vladenija informacionno-analiticheskoj sistemoj za schet sozdanija sistemy integracii dannyh [Reducing the total cost of ownership of an information and analytical system by creating a data integration system]. Otkrytoe obrazovanie, (2): 238–242.

Makarov A.V., Namiot D.E. 2023. Overview of data cleaning methods for machine learning. International Journal of Open Information Technologies, 11(10): 70–78. (in Russian).

Romanchukov S.V., Lyzin I.A., Marukhina O.V. 2020. Information system for analysis and modeling of social and economic development of the region. Information and mathematical technologies in science and management, 3(19): 96–104. (in Russian).

Sakhanevich D.Yu. 2020. Research of approaches and methods of applying artificial intelligence and machine learning in socio-economic processes. Herald of Omsk University. Series "Economics", 18(2): 65–79. DOI: 10.24147/1812-3988.2020.18(2).65-79. (in Russian).

Talkanbaeva R.A. 2019. Digitalization must begin from regions. Vestnik akademii gosudarstvennogo upravlenija pri prezidente kyrgyzskoj respubliki, (26): 38–41. (in Russian).

Khomonenko A.D. 2023. About the reliability and availability of object data stores. Intellectual technologies on transport, (S1): 123–128. (in Russian).

Hromova A.R., Petrosjan L.E. 2023. Analiz ujazvimostej v sistemah bezopasnosti dannyh [Analysis of vulnerabilities in data security systems]. Inzhenernyj vestnik Dona, (6 (102)): 67–76.

Shikulskiy M.I., Medvedeva O.V., Barkova V.M., Pleshakova L.A. 2022. Application of etl processes for automation of retail sales data analysis. Inzhenerno-stroitel'nyj vestnik Prikaspija, 4 (42): 108–113. DOI: 10.52684/2312-3702-2022-42-4-108-113. (in Russian).

Engel E.A., Engel N.E. 2018. Modernizacija programmnogo modulja "Zagruzka dannyh dlja intellektual'noj modeli" [Modernization of the software module "Data loading for an intelligent model"]. Vestnik Hakasskogo gosudarstvennogo universiteta im. NF Katanova, (23): 25–30.

Belov S., Ilina A., Javadzade J., Kadochnikov I., Korenkov V., Pelevanyuk I., Semenov R., Zrelov P., Tarabrin V. 2021. Analytical platform for socio-economic studies. In CEUR Workshop Proceedings, 9: 619–623. DOI: 10.54546/MLIT.2021.81.99.001.

Bolsunovskaya M.V., Kudryavtseva T.Y., Rudskaya I.A., Gintciak A.M., Zhidkov D.O., Fedyaevskaya D.E., Burlutskaya Z.V. 2023. Digital Platform for Modeling the Development of Regional Innovation Systems of Russian Federation. International Journal of Technology, 14(8): 1779–1789. DOI: 10.14716/ijtech.v14i8.6843.

Cenamor J., Parida V., Wincent J. 2019. How entrepreneurial SMEs compete through digital platforms: The roles of digital platform capability, network capability and ambidexterity. Journal of Business Research, 100: 196–206. DOI: 10.1016/j.jbusres.2019.03.035.

Constantinides P., Henfridsson O., Parker G.G. 2018. Introduction—platforms and infrastructures in the digital age. Information Systems Research, 29(2): 381–400. DOI: 10.1287/isre.2018.0794.

Gorodetsky V.I., Laryukhin V.B., Skobelev P.O. 2019. Conceptual Model of a Digital Platform for Cyber-Physical Management of a Modern Enterprises Part 1. Digital Platform and Digital Ecosystem. Mekhatronika, Avtomatizatsiya, Upravlenie, 20(6): 323–332. DOI: 10.17587/mau.20.323-332.

Hasell J., Mathieu E., Beltekian D., Macdonald B., Giattino C., Ortiz-Ospina E., Roser M., Ritchie H. 2020. A cross-country database of COVID-19 testing. Scientific Data, 7(1): 345. DOI: 10.1038/s41597-020-00688-8.

Jovanovic M., Sjödin D., Parida V. 2022. Co-evolution of platform architecture, platform services, and platform governance: Expanding the platform value of industrial digital platforms. Technovation, 118: 102218. DOI: 10.1016/j.technovation.2020.102218.

Mathieu E., Ritchie H., Ortiz-Ospina E., Roser M., Hasell J., Appel C., Giattino C., Rodes-Guirao. 2021. A global database of COVID-19 vaccinations. Nat Hum Behav, (5): 947–953. DOI: 10.1038/s41562-021-01122-8.

Mehmood E., Anees T. 2022. Distributed real-time ETL architecture for unstructured big data. Knowledge and Information Systems, 64(12): 3419–3445. DOI: 10.1007/s10115-022-01757-7.

Ricciato F., Wirthmann A., Giannakouris K., Skaliotis M. 2019. Trusted smart statistics: Motivations and principles. Statistical Journal of the IAOS, 35(4): 1–15. DOI: 10.3233/SJI-190584.

Shastitko A.E., Markova O.A. 2020. An old friend is better than two new ones? Approaches to market research in the context of digital transformation for the antitrust laws enforcement. Voprosy Ekonomiki, (6): 37–55. DOI: 10.32609/0042-8736-2020-6-37-55.

Winther K.T., Hoffmann M.J., Boes J.R., Mamun O., Bajdich M., Bligaard T. 2019. Catalysis-Hub. org, an open electronic structure database for surface reactions. Scientific data, 6(1): 75. DOI: 10.1038/s41597-019-0081-y.

Конфликт интересов: о потенциальном конфликте интересов не сообщалось.
Conflict of interest: no potential conflict of interest related to this article was reported.

Поступила в редакцию 19.08.2024
Поступила после рецензирования 05.09.2024
Принята к публикации 06.09.2024

Received August 19, 2024
Revised September 05, 2024
Accepted September 06, 2024

ИНФОРМАЦИЯ ОБ АВТОРАХ

Бекетов Сальбек Мустафаевич, аналитик лаборатории «Цифровое моделирование индустриальных систем», Санкт-Петербургский политехнический университет Петра Великого, г. Санкт-Петербург, Россия

Федяевская Дарья Эдуардовна, аналитик лаборатории «Цифровое моделирование индустриальных систем», Санкт-Петербургский политехнический университет Петра Великого, г. Санкт-Петербург, Россия

Схведиани Анги Ерастиевич, кандидат экономических наук, доцент Высшей инженерно-экономической школы, Санкт-Петербургский политехнический университет Петра Великого, г. Санкт-Петербург, Россия

Редько Сергей Георгиевич, доктор технических наук, директор Высшей школы проектной деятельности и инноваций в промышленности, Санкт-Петербургский политехнический университет Петра Великого, г. Санкт-Петербург, Россия

Бурлуцкая Жанна Владиславовна, младший научный сотрудник лаборатории «Цифровое моделирование индустриальных систем», Санкт-Петербургский политехнический университет Петра Великого, г. Санкт-Петербург, Россия

INFORMATION ABOUT THE AUTHORS

Salbek M. Beketov, Analyst at the Laboratory of Digital Modeling of Industrial Systems, Peter the Great St. Petersburg Polytechnic University, Saint-Petersburg, Russia

Daria E. Fedyaevskaya, Analyst at the Laboratory of Digital Modeling of Industrial Systems, Peter the Great St. Petersburg Polytechnic University, Saint-Petersburg, Russia

Angi E. Skhvediani, Candidate in Economics Sciences, Associate Professor at the Higher School of Engineering and Economics, Peter the Great St. Petersburg Polytechnic University, Saint-Petersburg, Russia

Sergey G. Redko, Doctor of Technical Sciences, Director of Higher School of Design and Innovation in Industry, Peter the Great St. Petersburg Polytechnic University, Saint-Petersburg, Russia

Zhanna V. Burlutskaya, Junior researcher at the Laboratory of Digital Modeling of Industrial Systems, Peter the Great St. Petersburg Polytechnic University, Saint-Petersburg, Russia