DOI 10.37882/2223-2966.2026.02-2.27

МОДЕЛИРОВАНИЕ ИНТЕЛЛЕКТУАЛЬНОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ УПРАВЛЕНИЯ ПЕРСОНАЛЬНЫМИ ДАННЫМИ В ОРГАНИЗАЦИЯХ

MODELING OF AN INTELLIGENT INFORMATION SYSTEM FOR PERSONAL DATA MANAGEMENT IN ORGANIZATIONS

A. Filatov

Summary. The relevance of the study is driven by increasing demands for personal data protection in the digital age. The paper proposes a practical model of an intelligent information system (IIS) designed to automate personal data management processes. Functional requirements for the IIS are presented, including data classification and masking, development of identification rules, and metadata management. Special attention is given to the system architecture, which leverages open-source technologies such as Apache Airflow. The study explores the potential of semantic models and theories, such as Markov processes and fuzzy sets, to optimize the data annotation process.

Keywords: personal data, intelligent information system, data annotation, semantic models, data management.

Филатов Александр Сергеевич
Аспирант, МГУТУ им. К.Г. Разумовского (ПКУ)
shzgudzh@gmail.com

Аннотация. Актуальность исследования обусловлена ростом требований к защите персональных данных в условиях цифровизации. В статье предлагается практическая модель интеллектуальной информационной системы (ИИС), предназначенной для автоматизации процессов управления персональными данными. Представлены функциональные требования к ИИС, включая классификацию и маскирование данных, разработку правил идентификации и управление метаданными. Особое внимание уделено архитектуре системы, которая базируется на использовании технологий с открытым исходным кодом, таких как Apache Airflow. Рассмотрены перспективы применения семантических моделей и теорий, таких как теория Марковских процессов и нечетких множеств, для оптимизации процесса разметки данных.

Ключевые слова: персональные данные, интеллектуальная информационная система, разметка данных, семантические модели, управление данными.

Актуальность проблемы управления персональными данными обусловлена ужесточением требований к защите информации в связи с цифровизацией общества, развитием цифровых сервисов и экосистем. Возрастающие риски утечек, кибератак и неправомерного использования персональных данных ужесточают требования регуляторов. В законе «О персональных данных» (ФЗ-152) установлены несколько ключевых технических требований для компаний, которые обрабатывают персональные данные. *Системы защиты персональных данных*: компании обязаны внедрить системы защиты персональных данных, которые соответствуют угрозам безопасности и защищают данные от несанкционированного доступа, утраты или изменения. *Аудит и мониторинг*: обязателен регулярный контроль за соблюдением требований безопасности, включая внутренний аудит информационных систем.

Для соблюдения требований регуляторов в организациях создаются специализированные департаменты по управлению персональными данными. В виду все большего распространения микросервисной архитектуры [1][2] в разработке цифровых сервисов, задача идентификации персональных данных значительно усложняется. Если при монолитной архитектуре все данные находятся на едином кластере СУБД, то при микросервисном подходе — данные максимально распределены, могут значительно отличаться по структуре хранения и используемым технологиям СУБД. В связи с этим, все больше растет потребность в разработке интеллектуальных информационных систем (ИИС) для автоматизации процессов в области защиты, аудита и мониторинга персональных данных [9].

Цифровой сервис может быть представлен как многофункциональная цифровизированная система (МЦС), где ключевым аспектом является интеграция цифровых инструментов во все слои управления. Важную роль в МЦС играет единое информационное пространство, обеспечивающее оперативность обработки данных и принятие решений. В данных системах крайне важны автоматизация интеграций данных и ответственность за их актуализацию [13]. Основной фокус внимания специалистов по персональным данным направлен на аудит и разметку данных в базах данных сервисов и аналитических хранилищах организаций [14]. При распределенной архитектуре СУБД организациям крайне важно идентифицировать ИС и производные объекты данных, содержащие персональные данные. Такие системы, согласно законодательству, причисляются к *инфор-*

мационным системам персональных данных (ИСПДн), на них накладывается ряд требований и ограничений. Все ИСПДН подлежат классификации по четырём уровням защищённости, определяемым характером обрабатываемых данных и степенью возможных угроз. На организационном уровне оператор ПД обязан назначить ответственное лицо, обеспечить законность обработки данных (наличие согласия субъекта или иной правовой основы), уведомить Роскомнадзор о создании системы и ограничить доступ только уполномоченными сотрудниками. Техническая защита предполагает использование сертифицированных средств защиты информации (ФСТЭК и ФСБ), применение алгоритмов шифрования по ГОСТ, аудит доступа и действий пользователей, а также разграничение прав с помощью ролей и механизмов аутентификации. Дополнительно законом установлены территориальные ограничения, обязывающие хранить персональные данные граждан РФ на серверах внутри страны, а также требования к срокам и объёму обработки информации. Контроль за исполнением возложен на Роскомнадзор, а нарушения влекут административную и, в отдельных случаях, уголовную ответственность.

Автоматизация процессов управления персональными данными возможна в образовательных организациях. Инструментарий предложенный в работе Аютовой И.В. включает разработку моделей и алгоритмов для автоматизированного предпроектного обследования ИСПДн. Эти инструменты позволяют учесть особенности образовательных учреждений, такие как публичность и территориальная разобщённость, и минимизировать трудозатраты, одновременно повышая безопасность данных. Предлагаются модели, основанные на теории марковских процессов и теории нечетких множеств [15]. Голенков В.В. и Гулякина Н.А. разрабатывают методы семантического анализа текстов и их применения в интеллектуальных системах, вне зависимости от стадии опытно промышленной эксплуатации. Основное внимание уделяется созданию моделей, способных структурировать и классифицировать текстовую информацию с учётом её семантических связей. Важным аспектом является использование онтологий и формальных логик, что позволяет системам более точно идентифицировать смысловые отношения между элементами текста. Эти технологии могут применяться для автоматической обработки данных и извлечения знаний, что существенно повышает эффективность интеллектуальных систем. [3] [4]

Формализация задачи

Пусть задано множество таблиц

$$T = \{t_1, t_2, \ldots, t_n\}$$

Где каждая таблица описывается структурой

$$t_i = (FQN_i, D_i, C_i)$$

FQN_i – метаданные: схемы данных, таблицы,

D_i – текстовое описание таблицы,

$C_i = \{c_{i1}, c_{i2}, \ldots, c_{im}\}$ – множество колонок.

FQN_i и $name_{ij}$ представляют собой наименования таблиц и столбцов на латинском языке в двух вариациях: «camel case», например: *smOrdersTotal*, *userClientView*, *orderliness* или «snake case», например: *order_positions*, *user_address*, *rep_lot* для их токенизации, Для текстовых описаний D_i и $name_{ij}$ будет использован вектор нормальной токенизации — разбивка по пробелам и знакам пунктуации [16].

$$tokens = Tok(s)$$

$tokens$ – массив выделенных слов строки s

s – строка текста $(FQN_i, name_{ij}, D_i, name_{ij})$

Каждая колонка c_{ij} задается вектором признаков:

$$c_{ij} = (name_{ij}, type_{ij}, desc_{ij}).$$

$name_{ij}$ – наименование колонки,

$type_{ij}$ – тип данных таблицы : целое, строка, дата, дата и время, булево,

$desc_{ij}$ – текстовое описание колонки

Необходимо провести классификацию на уровне таблицы, определив значение целевой переменной

$$y_i \in \{0,1,2\}$$

где 0 — таблица точно не содержит персональные данные, 1— таблица может содержать персональные данные, 2 — таблица точно содержит персональные данные. В обучающей выборке содержится $N = 30500$ таблиц, для каждой определено множество колонок.

Задача состоит в построении такой функции f, что $\hat{y}_i = f(t_i) y_i$, обеспечивающей уверенности $p_{ik} = P(y_i = k | t_i)$

Проектирование

Процесс разметки персональных данных в широком смысле представляет из себя разметку физических моделей данных (таблиц) ИСПДн системными аналитиками. Методология настоящей разметки формируется и валидируется специалистами по управлению персональными данными. Результатом процесса является

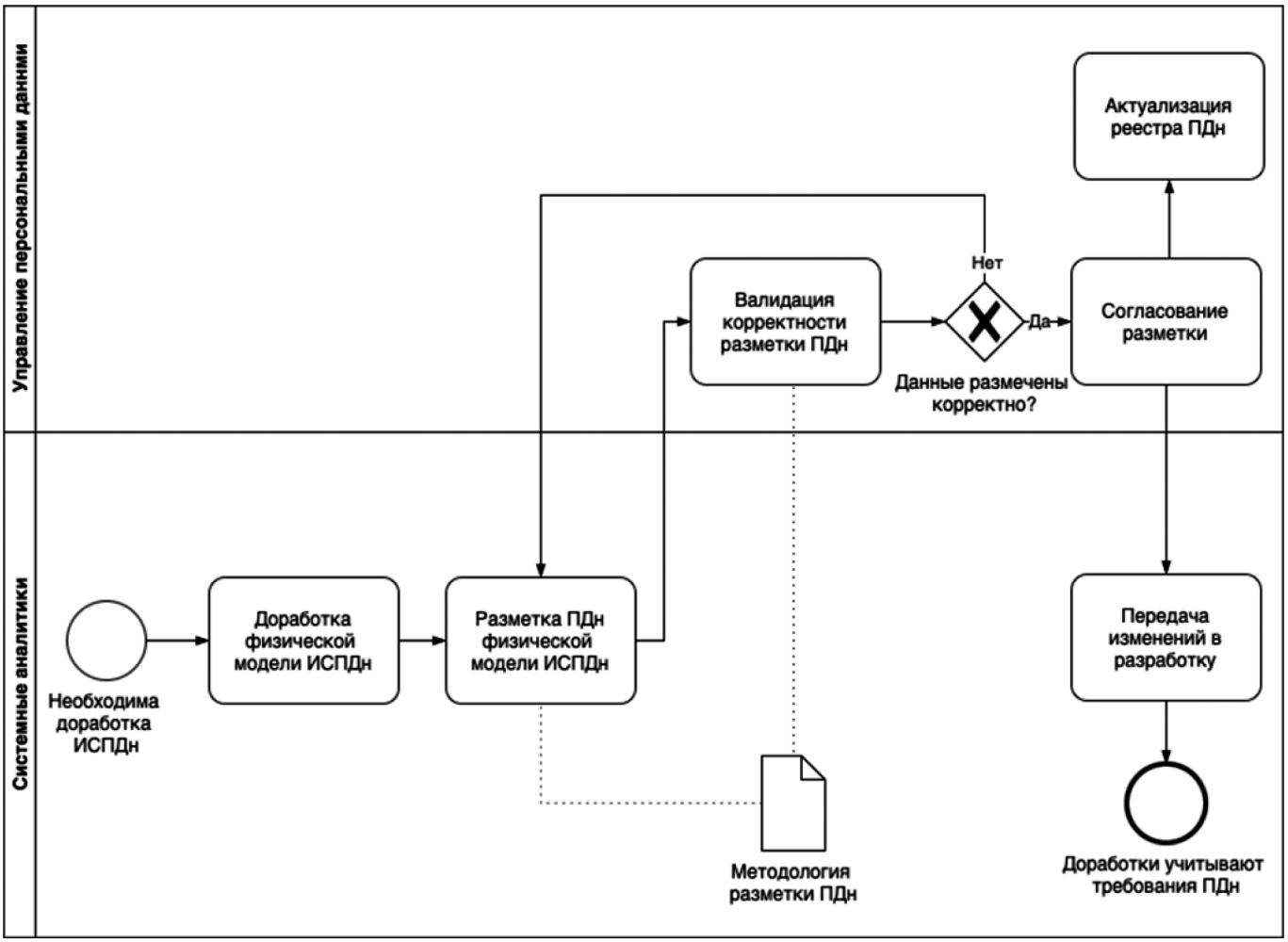

Рис. 1. Верхнеуровневый процесс разметки Персональных данных при разработке СУБД

формирование реестра персональных данных на уровне атрибутов физической модели ИСПДн. Реестр необходим в случаях: запроса пользователя на удаление персональных данных, маскирования персональных данных для просмотра неавторизированными пользователями, ведения учета персональных данных в организации.

Участниками процесса выступают две группы Пользователей:

- Команда Управления персональными данными — осуществляют надзор и валидируют корректность разметки персональных данных
- Системные аналитики — производят разметку на уровне физической модели данных ИСПДн

При принятии решения о классификации ПДн участники ориентируются на набор вводных метаданных и фактических данных. *Наименование таблицы и атрибута* — устанавливаются в соответствие с принятыми конвенциями и содержат указание на содержащиеся данные. *Описание таблицы и атрибута* — бизнес/техническое описание объекта, содержится в документации, либо в метаданных объекта в СУБД ИСПДн. *Фактические данные* — реальные данные внутри объекта, могут содержать номера телефонов, ФИО, денежные величины. *Методология разметки ПДн* — набор правил и методик для классификации ПДн, могут отличаться между организациями в виду специфики.

Назначение и основные функции

Разрабатываемая ИИС предназначена для автоматизации процессов классификации и управления объектами ИСПДн и аналитических хранилищ по уровням содержания персональных данных. ИИС должна удовлетворять следующим требованиям:

- аутентификация и авторизация действий Пользователей;
- визуализация клиентского веб-интерфейса для взаимодействия с серверной частью ИИС, с применением синхронных REST API методов;
- создание, редактирование, просмотр и удаление правил идентификации персональных данных;
- создание, редактирование, просмотр и удаление источников данных (аналитических хранилищ);
- создание, редактирование, просмотр и удаление каталогов базы знаний, включая хранимую историю и обучающие выборки;

- классификация метаданных из ИС-источников согласно правилам идентификации персональных данных;
- маскирование (обезличивание) фактических данных в объектах систем-источников в соответствие с правилами маскирования персональных данных;
- физическое удаление фактических данных в объектах систем-источников в соответствие с регламентами очистки персональных данных;
- экспорт метаданных и классификаций в формате CSV / XML;
- формирование технической отчетности по результатам работы моделей ИИС.

Архитектура

Система должна быть спроектирована как веб-ориентированное программное средство и предоставлять Пользователям возможности запуска интеграционных процедур и актуализации базы знаний. Основные модули проектируемой системы включают в себя:

Пользовательский интерфейс — предназначен для работы Специалистов по персональным данным, необходим для визуального представления результатов работы Интеллектуальной системы, позволяет вносить и актуализировать правила классификации персональных данных. Программная реализация включает в себя технологии JavaScript для визуального отображения объектов, а также REST API для взаимодействия с серверными модулями ИС

Модуль классификации метаданных — основной модуль поддержки принятия решений. Представляет собой семантическую модель, определяющую уровень персональных данных, содержащихся в объекте хранилища данных. На вход модели поступают преобразованные мета-данные из ИСПДн и аналитических хранилищ. В результате работы процедур, на основании внутренней Базы знаний, ранее неклассифицированные данные разбиваются по уровням Персональных данных и маскируются в ИСПДн аналитических хранилищах (источниках).

Модуль интеграций — осуществляет сбор исходных мета-данных из ИСПДН и аналитических хранилищ и преобразовывает их для долгосрочного хранения во внутренней базе знаний. Позволяет запускать процедуры на регламентной основе, либо по запросу Пользователя. Программная реализация модуля использует технологию с открытым исходным кодом Airflow на базе лицензии Apache. Технология позволяет на регламентной основе запускать процедуры обработки больших массивов данных, написанных на языке Python.

База знаний — реляционная СУБД, содержащая: параметры запуска и конфигурации семантической модели, набор Пользовательских правил и процедур, исторические метаданные аналитических хранилищ.

ИИС должна поддерживать ролевую модель доступа к ресурсам. Роль «Администратор» позволяет выполнять все действия в ИИС (просмотр, редактирование, удаление) вне зависимости от ресурса обращения. Роль «Эксперт» включает в себя основные группы Пользователей: Специалистов по персональным данным и Системных аналитиков им доступны действия (просмотр, редактирование, удаление) но с учетом принадлежности к ресурсу (конкретным ИСПДн и хранилищам).

Заключение

Актуальность вопросов управления персональными данными, на ряду с ужесточением требований регуляторов обуславливают необходимость разработки комплексных инструментов и программного обеспечения для автоматизации экспертных процессов. Они должны быть направлены на повышение эффективности и оптимизацию процессов управления ПДн.

Возможным ответом на запрос могут стать интеллектуальные информационные системы, осуществляющие поддержку в процессах принятия экспертных решений. Наиболее перспективной областью является разработка семантических классификационных моделей и соответствующих баз знаний для расширения экспертизы процесса и снижения вовлечения Пользователей в рутинные операции классификации и валидации. Разрабатываемые ИИС должны отвечать современным технологическим и архитектурным принципам, а также требованиям безопасности

В настоящей статье предложены функциональные требования и концептуальная архитектура моделируемой ИИС. В рамках описанного процесса, основной упор ИИС должен быть сделан на модуле классификации метаданных и модуле интеграций. В классификационном модуле могут быть использованы модели, основанные на теории марковских процессов, теории нечетких множеств, использованы онтологии и формальные логики. Интеграционный модуль должен быть гибким в настройке и позволять подключаться к разнообразным источникам ИСПДн, вне зависимости от используемых технологий СУБД, модуль может быть разработан на основе технологии с открытым исходным кодом Apache Airflow. Предложения могут быть использованы организациями при разработке ИИС для решения практических задач в области управления ПДн.

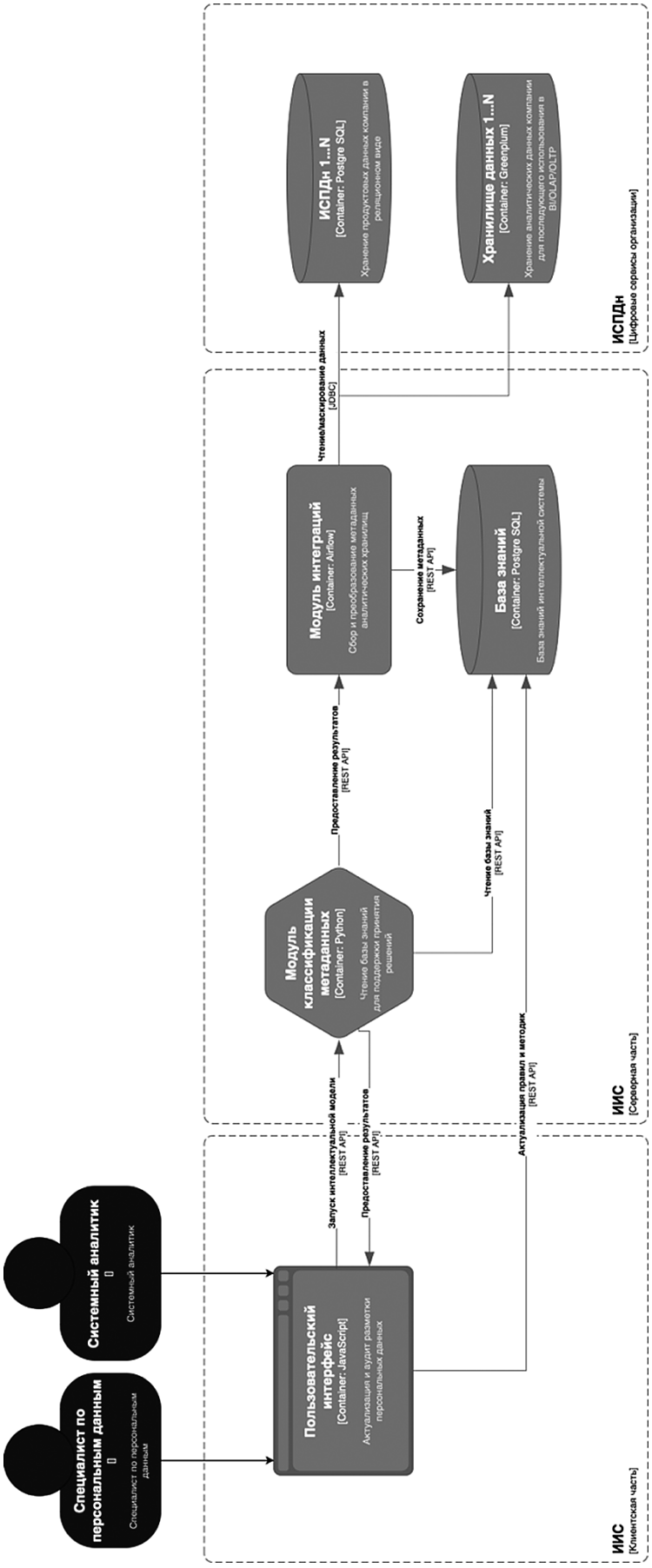

Рис. 2. Архитектурная схема компонентов ИИС

ЛИТЕРАТУРА

1. Маркова В.Д. Цифровизация управления: от АСУ к микросервисам // ЭКО. — 2022. — № 9 (579). — С. 113–129. — DOI: 10.30680/ECO0131-7652-2022-9-113-129.
2. Опарин Г.А., Богданова В.Г., Пашинин А.А. Инструментальные средства автоматизации разработки и применения пакета прикладных микросервисов // Информационные и математические технологии в науке и управлении. — 2024. — № 2 (34). — С. 155–168. — DOI: 10.25729/2413-0133-2024-2-155-168.
3. Голенков В.В., Гулякина Н.А. Методы семантического анализа текстов и их применение в интеллектуальных системах // Материалы международной научно-технической конференции. Минск: БГУИР, 2020. С. 345–350. DOI: 10.35596/2020-01-345-350.
4. Голенков В.В., Гулякина Н.А. Принципы создания интеллектуальных систем на основе семантического анализа // Вестник БГУИР. 2020. № 4. С. 120–125. DOI: 10.35596/2020-04-120-125.
5. Мухаметгалиев А.Ф., Казаков В.А., Гайсин И.Н. Анализ и моделирование информационных процессов в интеллектуальных системах // Научный журнал МОИТ. — 2023. — № 2. — С. 55–65. DOI: 10.17399/moitvivt2023-2-55-65.
6. Борисова А.А., Иванов И.В. Разработка методов и средств для защиты персональных данных в распределенных системах // Информационные технологии. — 2021. — Т. 27, № 4. — С. 345–352. DOI: 10.14357/it20210405.
7. Кирюхина, Е.С. Модели и алгоритмы управления процессом обработки персональных данных в вузе: дис. канд. техн. наук: 05.13.10 / Кирюхина Елена Сергеевна. — Белгород, 2014. — 135 с.
8. Методология разметки персональных данных в образовательных организациях // Научный журнал МОИТ. 2023. № 2. С. 74–82. DOI: 10.17399/moitvivt2023-2-74-82.
9. Остроух А.В. Интеллектуальные системы: монография. Н. Новгород: НКРАС, 2020. 280 с. DOI: 10.32703/2020-11-03.
10. Мариллоннет П., Лоран М., Атес М. Самоуправление персональной информацией: обзор технологий поддержки административных услуг // arXiv preprint arXiv:2109.12968, 2021. DOI: 10.48550/arXiv.2109.12968.
11. Чжан В., Ли М., Ченг Х. и др. Управление персональными данными в соответствии с GDPR: решение на основе блокчейна // arXiv preprint arXiv:1904.03038, 2019. DOI: 10.48550/arXiv.1904.03038.
12. Смирнова Е. А. Защита персональных данных: проблемы и решения // Молодой ученый. 2024. № 5. С. 123–127.
13. Гусев П.Ю., Систематизация и управление доступом к данным в многофункциональной цифровизированной системе // Моделирование, оптимизация и информационные технологии. 2023;11(4). DOI: 10.26102/2310-6018/2023.43.4.025
14. Некрасов А.А., Гаврилов С.О., Беленькая М.Н. Средства создания хранилищ данных // Телекоммуникации и информационные технологии. — 2021. — Т. 8, № 1. — С. 75–80.
15. Аютова И.В. Модели и алгоритмы управления процессом обработки персональных данных в вузе: дис. … канд. техн. наук: 05.13.01 / Аютова Ирина Владимировна. — Сургут: Сургутский государственный университет, 2012. — 155 с.
16. Гречахин В.А. К вопросу о токенизации текста // Международный научно-исследовательский журнал. — 2016. — № 7 (49). — DOI: 10.18454/IRJ.2016.48.070.

Журнал «Современная наука: актуальные проблемы теории и практики»