БАЗА ЗНАНИЙ В СИСТЕМЕ РАССЛЕДОВАНИЯ АВАРИЙ И ИНЦИДЕНТОВ НА ПРОИЗВОДСТВЕННЫХ ОБЪЕКТАХ

Маркарян А.О.[1], Поляховская О.А.[2], Якубовская Ю.А.[3]

Ключевые слова: *информационная система, расследование, формализация знаний, база данных, правила обработки данных, обработка естественного языка, регулярные выражения, Lakehouse-архитектура, промышленная безопасность, цифровизация.*

Аннотация

Цель работы: *разработка базы знаний — централизованной системы хранения и автоматизированной обработки данных для расследования причин аварий и инцидентов на производственных объектах, обеспечивающей единый доступ к гетерогенной информации и сокращение времени анализа аварий.*

Методы: *системный анализ процесса расследования и существующих информационных решений в области промышленной безопасности; математический аппарат теории множеств для построения формальной модели данных; методы обработки естественного языка и регулярные выражения для автоматизированного извлечения фактов из разнородных документов; проектирование Lakehouse-архитектуры и технологий для программной реализации системы.*

Результаты: *проведён анализ регламента расследования аварий и существующих информационных решений, выявивший необходимость разработки инструментария для централизованного хранения разнородных материалов расследования с автоматизированным извлечением фактов и расчётом нормативных показателей. Предложена концептуальная модель и осуществлена формализация знаний системы расследования: витрины данных, сущности, характеризующие предметную область, отношения между ними и правила прослеживаемого извлечения фактов из документов разных типов. Определена структура базы данных, предложен алгоритм её автоматизированного наполнения, предусматривающий маршрутизацию документов, применение специализированных парсеров и вычисление показателей по материалам расследования. Разработано программное обеспечение на основе Lakehouse-архитектуры, включая модуль обработки данных и панель администратора. Тестирование на материалах реального расследования показало точность извлечения фактов 91,5%, полноту 88,7%, полное совпадение расчётных показателей с ручным расчётом и сокращение времени формирования знаний по инциденту примерно на 97%.*

DOI: 10.24412/1994-1404-2026-3-51-59

Введение и постановка задачи

Производственные объекты — шахты, заводы и иные промышленные предприятия — играют ключевую роль в экономике страны, однако работа на них сопряжена с риском аварий и инцидентов, способных привести к человеческим жертвам, экологическим катастрофам и значительным экономическим потерям. Несмотря на ужесточение требований промышленной безопасности и модернизацию оборудования, аварийность на опасных производственных объектах сохраняется: по сравнению с 2024 годом число аварий увеличивается, а число внеплановых проверок растёт; большинство аварий по-прежнему связано с неисправностью оборудования и человеческим фактором.

Ключевым инструментом предотвращения повторения инцидентов является качественное и оперативное расследование, направленное на установление цепочки событий и первопричины происшествия. Расследование на промышленном объекте регламентировано

[1] ***Маркарян Анна Оганесовна***, *кандидат технических наук, доцент, доцент кафедры автоматизированных систем управления НИТУ МИСИС, г. Москва, Российская Федерация. ORCID: 0009-0004-2181-2144.*
E-mail: anestrelle@gmail.com

[2] ***Поляховская Ольга Александровна***, *студент кафедры автоматизированных систем управления НИТУ МИСИС, г. Москва, Российская Федерация. ORCID: 0009-0005-4053-6357.*
E-mail: plkhvskl@gmail.com

[3] ***Якубовская Юлия Александровна***, *студент кафедры автоматизированных систем управления НИТУ МИСИС, г. Москва, Российская Федерация. ORCID: 0009-0009-7196-0774.*
E-mail: yakuboyulia@gmail.com

на федеральном уровне и сопровождается формированием комиссии, которая в сжатые сроки собирает и анализирует большой объём разнородной информации: показания датчиков, фото- и видеофиксацию, протоколы осмотра, опросы свидетелей, техническую документацию. На практике эта информация консолидируется вручную в виде текстовых отчётов и таблиц, что ведёт к потере контекста и причинно-следственных связей между событиями, субъективности интерпретации и задержкам в принятии решений. Накопленный опыт расследований сложно систематизировать и использовать для проактивного анализа рисков.

Анализ существующих информационных систем в данной области — зарубежной базы крупных промышленных аварий MHIDAS, отечественного автоматизированного рабочего места «БАЭ» для электроэнергетики, программного комплекса «1С: Охрана труда и промышленная безопасность» и европейской базы eMARS — показал, что ни одно из решений не обеспечивает единого централизованного хранения разнородных данных расследования с автоматическим извлечением фактов и расчётом нормативных показателей применительно к российским производственным объектам, что свидетельствует об актуальности создания такого инструментария.

Проблема расследования аварий рассматривается в научной литературе с разных сторон. Кузнецова Е.А. показывает эволюцию подходов к анализу аварий от линейных моделей до системных, многоуровневых [1], что подтверждается систематическим обзором Ahmadi Rad с соавторами, демонстрирующим превосходство методов системного мышления (STAMP, AcciMap, FRAM, HFACS) над традиционными линейными подходами [2]. Статистический анализ отраслевой аварийности (Стасева с соавторами [3]; Пигилова с соавторами [4]) указывает на устойчивые причины аварий — неисправность оборудования и человеческий фактор. Derdowski и Mathisen [5], а также Ванякин с соавторами [6] обращают внимание на то, что традиционные методики расследования не учитывают психосоциальные и организационные факторы и не успевают за усложнением современного производства. Adumene с соавторами предлагают адаптивную модель на основе таксономии HFACS-MA с байесовскими сетями для количественной оценки человеческого фактора в морских операциях с учётом динамики изменения производительности персонала и неопределённости операционной среды [7], а Климова с соавторами показывают ограниченность классических инструментов (анализ дерева отказов, FMEA, диаграммы Исикавы) при изолированном рассмотрении случаев [8].

Принципиально важной для настоящего исследования является работа Маркарян А.О., предложившей концептуальную модель системы поддержки принятия решений при расследовании авиационных происшествий, включающую шесть специализированных подсистем и завершающий их работу блок хранения и систематизации сведений — базу знаний [9]. Настоящее исследование развивает данное направление, фокусируясь на наполнении и структурной организации именно этого компонента применительно к расследованию аварий на производственных объектах. Архитектурные решения для подобных задач (клиент-серверная модель, разграничение пользователей, модульная организация) представлены в работе Шамина и Зеленко [10]. Вопрос выбора подхода к организации данных (Data Lake против Data Warehouse) рассмотрен Гунтупалли [11]; типовые атрибуты и классификаторы баз данных о техногенных чрезвычайных ситуациях в горнодобывающей отрасли выявлены Кудрявцевым с соавторами [12]. Sahoo с соавторами [13] и Karimi Dehkordi с соавторами [14] показывают эффективность методов машинного обучения и текстовой аналитики для извлечения цепочек событий и факторов риска из неструктурированных описаний инцидентов. Опыт агрегирования разнородных источников (телеметрия, дистанционное зондирование) представлен Ерошенко с соавторами [15]; Захаров с соавторами обосновывают применение протоколов IIoT и подхода ELT, при котором сырые данные сначала загружаются в хранилище, а затем трансформируются под конкретные аналитические задачи [16], что соответствует гибкой природе проектируемой базы знаний. Библиометрический обзор применения машинного обучения для анализа защитных барьеров [17], а также работы Rodriguez и Fernandez [18], Wang с соавторами [19], Thompson и Lee [20], Kumar и Singh [21] подтверждают актуальность внедрения интеллектуальных систем поддержки процесса расследования и важность интероперабельности между источниками данных. Наконец, Потекаева с соавторами подчёркивают необходимость интеграции новой системы в общую информационную архитектуру предприятия, а не её изолированное существование [22]. Проведенный обзор подтвердил актуальность данного исследования, а также заложил основу для выбора технологий создания инструментария, позволяющего автоматизировать процесс извлечения значимых для расследования фактов из множества неструктурированной информации и организовать их хранение.

На основании этого поставлена цель исследования — разработать базу знаний системы расследования аварий и инцидентов на производственных объектах, включающую базу данных с фактами о происшествиях и правила автоматизированного извлечения фактов из гетерогенных источников. В результате должна быть создана централизованная система хранения и обработки данных, обеспечивающая единый доступ к гетерогенной информации и позволяющая сократить время анализа аварий и инцидентов.

Для достижения цели поставлены следующие задачи:

1) построение концептуальной модели;
2) формализация знаний системы расследования;
3) разработка архитектурно-технического решения и программного обеспечения;
4) анализ результатов тестирования на данных реального расследования.

Концептуальная модель

База знаний представляет собой систему, объединяющую базу данных и логические правила автоматизированной обработки исходных материалов, преобразующих разрозненный документно-ориентированный процесс расследования в централизованную цифровую среду. В связи с этим структура разрабатываемой системы включает три компонента: централизованное хранилище (базу) данных; модуль обработки данных, осуществляющий автоматизированное извлечение фактов из разнородных документов и расчёт нормативных показателей; панель администратора для просмотра и ручного редактирования информации экспертами (рис. 1).

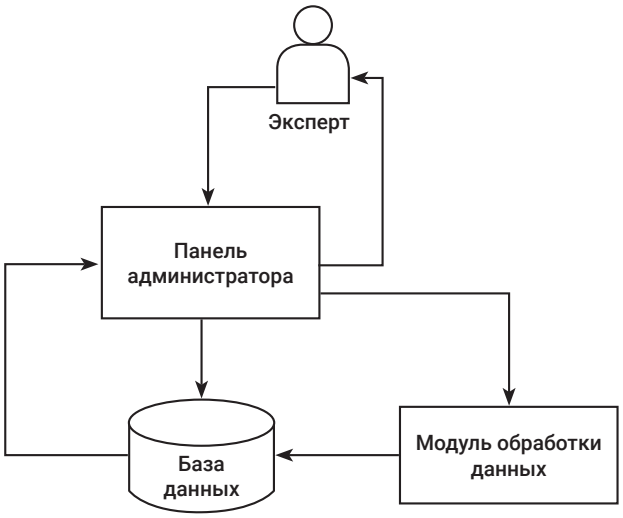


Рис. 1. Концептуальная модель системы

В свою очередь, модуль обработки данных декомпозирован на четыре подсистемы: загрузки и маршрутизации данных; извлечения фактов; вычисления расчётных показателей; хранения и доступа к данным (рис. 2). Такая реализация освобождает экспертов от рутинных операций по консолидации и расчётам, в которых возможна ошибка, и создаёт единую платформу, в которой накапливаются данные обо всех расследованных и расследуемых инцидентах. Ниже описаны формальные модели каждой из подсистем.

Формализация знаний системы расследования

Пусть $U = \{u_1, u_2, \dots, u_n\}$ — множество всех инцидентов, зарегистрированных в системе. Для каждого инцидента u_i определено множество его исходных материалов $P(u_i) = \{p_1, \dots, p_{mi}\}$, которое разбивается на три непересекающихся подмножества:

$$P = P_{struct} \cup P_{unstruct} \cup P_{media} \quad (1)$$

где P_{struct} — формализованные материалы с чёткой структурой (акты осмотра, протоколы замеров, журналы технического обслуживания, паспорта оборудования); $P_{unstruct}$ — неструктурированные текстовые материалы (транскрипции аудиозаписей, опросы свидетелей, объяснительные записки); P_{media} — медиафайлы (графика, аудио, видео, чертежи), не подлежащие обработке.

Множество сущностей предметной области *E* объединяет 12 типов объектов, описывающих инцидент и его контекст: инцидент E_{inc}, предприятие E_{comp}, причастные сотрудники E_{empl}, оборудование E_{eq}, сертификаты соответствия E_{cert}, журналы технического обслуживания E_{maint}, локации (выработки и помещения) E_{loc}, датчики E_{sens}, геологические объекты E_{geo}, эксперты E_{exp}, сейсмические события E_{seis} и нормативные документы E_{reg}.

Дополнительно определены множества: извлечённых фактов F_{inc}, структурированное хранилище *T*, хранилище исходных файлов *O* и множество пользователей системы *C* с ролями {admin, expert, viewer}.

Каждый извлечённый факт $\varphi \in F_{inc}$ представляется кортежем

$$\varphi = \langle e, \text{attr}, \text{value}, \text{source} \rangle \quad (2)$$

где $e \in E$ — сущность предметной области; attr — атрибут сущности; value — извлечённое значение; source = *p* — ссылка на исходный документ, обеспечивающая полную прослеживаемость факта к материалу расследования. Множество фактов по инциденту u_i определяется объединением фактов по всем его материалам.

Работа подсистемы загрузки и маршрутизации основана на действии оператора Г, определяющего тип загружаемого документа по расширению файла и наличию повторяющейся структуры:

$$\Gamma(p) = \begin{cases} structured, \text{если } p \in P_{struct} \\ unstructured, \text{если } p \in P_{unstruct} \\ media, \text{если } p \in P_{media} \end{cases} \quad (3)$$

В подсистеме извлечения фактов функционируют три оператора, в зависимости от типа документа:

1) формализованные документы обрабатываются оператором Ψ_{regex} на основе множества регулярных шаблонов П = $\{\pi_1, \dots, \pi_k\}$, где π_k = ⟨pattern, group_map, attr_target⟩ — регулярное выражение с правилами сопоставления найденных фрагментов с атрибутами сущности:

$$\Psi_{regex}(p) = \\ = \bigcup_{\pi \in \Pi} \{\langle attr, value \rangle \mid \exists match = \\ = \pi.pattern(content(p)), value \\ = extract(match, \pi.group_map)\} \quad (4)$$

Например, шаблон вида «(?:Тип|Модель):\ s*([A-Z0-9/.\-]+)» из текста «Модель: FS300/1.0» возвращает значение MODEL = «FS300/1.0»; аналогичные шаблоны определены для заводского номера, года



Рис. 2. Модуль обработки данных

выпуска, класса искробезопасности, технического состояния и места расположения оборудования.

2) неструктурированные тексты (транскрипции, опросы свидетелей) обрабатываются оператором

$$\Psi_{sem} = \Theta_{NER} \circ \Theta_{segm} \circ \Theta_{token} \quad (5)$$

Он представляет собой композицию токенизации, сегментации на предложения и распознавания именованных сущностей типов PER (персоны), LOC (локации) и ORG (организации), с последующим применением семантических шаблонов, нацеленных на конкретные виды информации: имя опрашиваемого, дату и время опроса, ключевые события по словарю K_{events} = {взрыв, пожар, искра, задымление, обрушение, …}. Так, в предложении «горный мастер Петров находился в лаве 512» распознаются сущности «Петров» → PER и «лава 512» → LOC.

3) медиафайлы обрабатываются оператором Ψ_{media}, фиксирующим только метаданные — имя, описание и ссылку на файл в хранилище.

Подсистема вычисления расчётных показателей построена по принципу действия оператора *M*, который на основе множества фактов F_{inc} вычисляет три группы показателей:

$$M : F_{inc} \rightarrow \{M_{inc}, M_{fire}, M_{vent}\} \quad (6)$$

Общие показатели инцидента M_{inc} включают суммарные потери (сумму погибших и пострадавших) и протяжённость зоны поражения. Вентиляционные показатели M_{vent} — относительный дефицит скорости воздуха $\delta_v = max(0, (v_{norm} - v_{fact}) / / v_{norm} \times 100\%)$, коэффициент утечек воздуха $k_{leak} = (Q_{out} - Q_{in}) / Q_{in}$ и признак соответствия вентиляции нормам. Пожарные показатели M_{fire} вычисляются по R-коэффициентам диагностики эндогенных пожаров:

$$R_1 = \Delta O_2/\Delta CO_2,\ R_2 = \Delta CO/\Delta O_2,\ R_3 = \Delta CO/\Delta CO_2 \quad (7)$$

где ΔO_2, ΔCO_2, ΔCO — отклонения концентраций соответствующих газов в пробе воздуха от атмосферных значений; при $R_1 \geq R_{1_{crit}} = 2{,}5$ фиксируется признак окислительного процесса. Дополнительно вычисляются эффективность дегазации и признак соответствия норме содержания негорючих веществ при оценке пылевзрывозащиты.

Подсистема хранения и доступа к данным предоставляет доступ пользователей $c \in C$ к данным структурированного хранилища *T* посредством трех операторов: фильтрации Ω_{filter} (запрос к таблице по заданным условиям), Ω_{source} (переход от факта φ к ссылке на исходный документ-источник) и CRUD (создание, чтение, обновление и удаление записей).

Полный цикл обработки информации об инциденте u_i представляет собой композицию операторов загрузки, маршрутизации, извлечения фактов, вычисления показателей и записи в хранилище:

$$Y_{total}(u_i) = \Delta_{load}(M(\bigcup_{p \in \Lambda_{load}} \Psi_{\Gamma(p)}(p))) \quad (8)$$

а база данных определяется как объединение структурированных представлений всех зарегистрированных инцидентов:

$$K = \bigcup_{i=1}^{n} S(u_i), \text{ где } S(u_i) = \langle E, F_{inc}, M(u_i) \rangle \quad (9)$$

Архитектурно-техническое решение и программное обеспечение

Архитектурно-техническое решение строится вокруг трёх компонентов: централизованного хранилища данных, реализованного по принципу Lakehouse; модуля обработки входных данных; десктопного приложения панели администратора. Такое решение позволяет членам комиссии по расследованию, не имеющим глубоких навыков программирования, получать удобный графический инструмент для просмотра, добавления, изменения и удаления информации об инцидентах, в то время как расчётные показатели формируются автоматически модулем обработки данных.

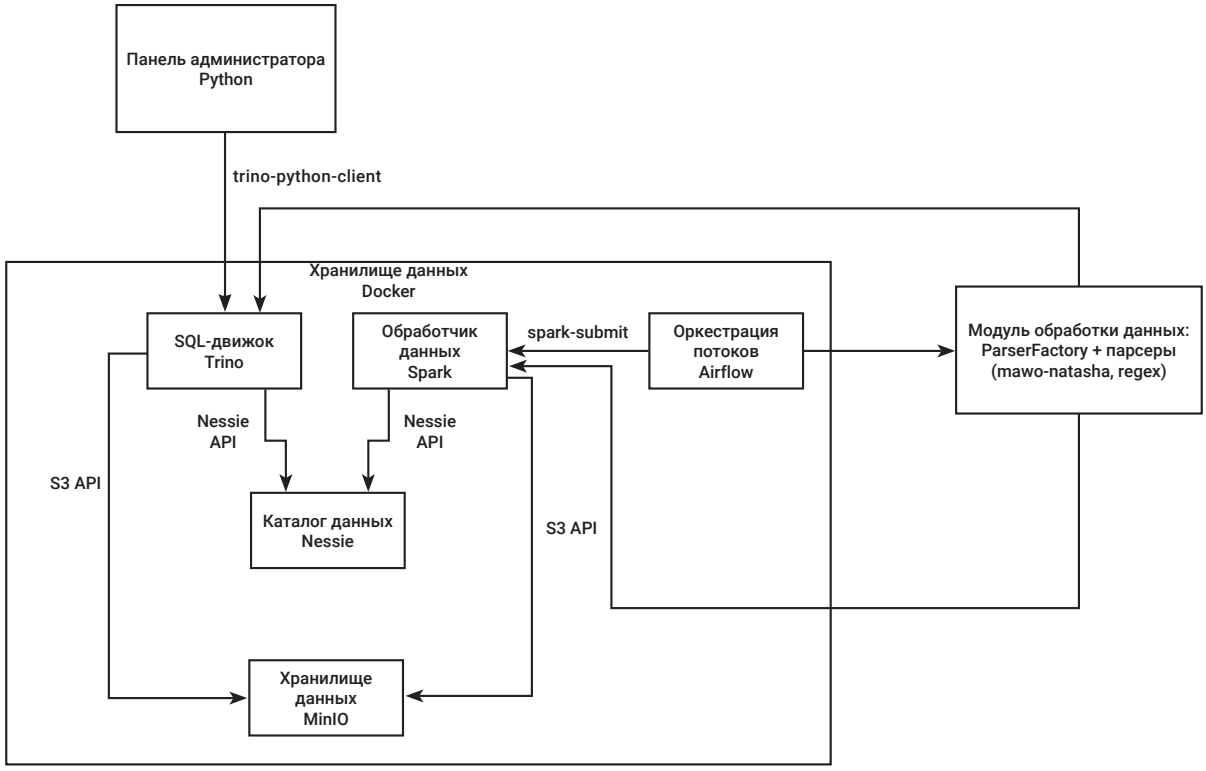


Рис. 3. Архитектурно-техническое решение

Хранилище данных, архитектура которого представлена на рис. 3, построено на трёх открытых технологиях. Объектное хранилище MinIO — S3-совместимое решение для размещения исходных файлов в их первоначальном виде и преобразованных данных в формате Apache Parquet. Распределённый SQL-движок Trino позволяет выполнять запросы к файлам Parquet непосредственно в MinIO без их предварительного перемещения, обеспечивая единый интерфейс доступа к хранилищу через JDBC/ODBC. Выбор архитектуры Lakehouse обусловлен необходимостью одновременной работы с разнородными типами данных — от структурированных таблиц до неструктурированных изображений и документов — при сохранении ACID-транзакций и эффективного выполнения аналитических запросов, что недостижимо при использовании классического Data Lake или Data Warehouse по отдельности. Структурированная часть хранилища организована в виде набора взаимосвязанных таблиц, отражающих ключевые сущности предметной области: инциденты, предприятия, сотрудники, оборудование, локации, документы, извлечённые факты, гипотезы и расчётные показатели. Связи между таблицами реализуют привязку фактов к конкретному инциденту и источнику данных, а также обеспечивают переход от сущностей к связанным документам и результатам расчётов.

Модуль обработки данных реализован как конвейер, оркестрируемый Apache Airflow. Обработка одного инцидента осуществляется по следующему алгоритму.

На вход подаётся множество исходных документов $P(u_i)$ по инциденту u_i. Запускается цикл перебора каждого документа $p \in P(u_i)$. Для текущего документа выполняется загрузка в объектное хранилище O, после чего срабатывает оператор маршрутизации Γ. Документ классифицируется как структурированный, неструктурированный текстовый или медийный.

Если документ признан структурированным, применяется оператор Ψ_{regex}, извлекающий факты на основе регулярных шаблонов. Если документ является неструктурированным текстом (транскрипции, опросы), применяется семантический конвейер Ψ_{sem}, включающий токенизацию, сегментацию и NER-распознавание сущностей типов PER, LOC, ORG, с последующим применением семантических шаблонов. Медийные файлы не обрабатываются автоматически; фиксируются только метаданные и ссылка для ручного анализа.

Все факты, извлечённые из текущего документа, добавляются в множество F_{inc}. Каждый факт представляется кортежем $\varphi = \langle e, \text{attr}, \text{value}, \text{source} \rangle$, где source — ссылка на исходный документ, обеспечивающая прослеживаемость. Цикл повторяется до обработки всех документов.

После завершения цикла на основе F_{inc} оператором M вычисляются три группы показателей: M_{inc}, M_{vent}, M_{fire}. Дополнительно рассчитываются показатели дегазации и пылевзрывозащиты.

Результаты объединяются в структурированное представление $S(u_i)$ и записываются в хранилище T. Доступ пользователей C к данным реализован с помощью операторов фильтрации и CRUD с разграничением по ролям {admin, expert, viewer}. На этом алгоритм завершает свою работу.

Парсинг реализован по принципу фабрики парсеров (Parser Factory): для каждого типа документа создан специализированный парсер, наследующий общий базовый класс и переопределяющий метод разбора содержимого. Документ, не подходящий ни под одно правило маршрутизации, направляется в универсальный парсер, извлекающий ключевые слова для последующего анализа. Извлечённые записи вставляются в соответствующие таблицы хранилища пакетно. Расчётные показатели вычисляются отдельной задачей на движке Apache Spark, запускаемой по завершении загрузки данных.

Панель администратора реализована как десктопное приложение на языке Python с использованием библиотеки PyQt6, что обусловлено требованиями к автономности работы экспертов в защищённом контуре предприятия и возможностью локальной обработки больших объёмов данных без дополнительной нагрузки на серверную инфраструктуру.

Приложение подключается к хранилищу через Trino и предоставляет функции просмотра списка таблиц и их содержимого, выполнения произвольных SQL-запросов с сохранением истории, ручного редактирования значений ячеек, добавления и удаления записей с фиксацией факта удаления в истории изменений для обеспечения прозрачности процесса.

Анализ результатов

Тестирование системы проведено на материалах расследования аварии на угольной шахте в 2023 году. Корпус включал 47 документов разных типов: акты осмотра, протоколы опросов, паспорта оборудования, журналы обслуживания, транскрипции аудиозаписей, результаты газового и воздушного анализа, сертификаты, сейсмические данные и мультимедийные материалы. Были проверены модуль загрузки и маршрутизации, парсеры извлечения фактов, модуль расчётных показателей, панель администратора и инфраструктурная часть решения.

Из 47 документов в хранилище загружено 45; два файла не обработаны из-за некорректной кодировки. Оператор маршрутизации корректно классифицировал все загруженные документы: формализованные материалы обрабатывались регулярными шаблонами, неструктурированные тексты — NLP-конвейером, медиафайлы — оператором регистрации метаданных. Среднее время обработки одного документа составило около 0,72 с.

На выборке из 10 документов проведена экспертная валидация извлечения фактов. Получено 86 верно извлечённых фактов, 8 ложных срабатываний и 11 пропусков, что соответствует точности 91,5%, полноте 88,7% и F1-мере 90,1%. Ошибки в основном связаны с нестандартными сокращениями и отдельными случаями неверной интерпретации числовых значений.

Сравнение результатов работы модуля расчётных показателей с ручным расчётом эксперта показало полное совпадение по общим показателям инцидента. Для пожарных и вентиляционных показателей полученные значения также согласовались с выводами подкомиссий, что подтверждает корректность реализованных расчётных правил.

Функциональное тестирование панели администратора по сценариям просмотра данных, выполнения SQL-запросов, редактирования и удаления записей показало полное соответствие ожидаемому результату. Все проверяемые сценарии были отработаны успешно.

Ключевым эффектом внедрения системы является сокращение трудозатрат на формирование базы знаний по инциденту. При ручном подходе консолидация материалов, извлечение фактов, формирование хронологии и вычисление показателей занимают 8—12 часов. В предлагаемой системе автоматизируются все этапы, кроме первичного поиска и подготовки материалов, что сокращает время формирования базы знаний до 5—10 минут, то есть примерно на 97%.

К ограничениям решения относятся формирование словаря семантических шаблонов на материалах одного инцидента, фиксация для медиафайлов только метаданных без анализа содержимого и необходимость ручного формирования гипотез о причинах аварии. Перспективы развития связаны с расширением корпуса документов, интеграцией методов компьютерного зрения и автоматизацией ранжирования гипотез на основе накопленных данных.

Выводы

В результате проведённого исследования разработана база знаний для расследования аварий и инцидентов на производственных объектах, обеспечивающая централизованное хранение и автоматизированную обработку данных из гетерогенных источников.

Разработанная математическая модель формального представления данных по инцидентам и реализация алгоритма их обработки позволяет унифицировать процесс извлечения знаний из неструктурированной информации, поступающей в процессе расследования аварии. Предложенная комбинация Lakehouse-архитектуры с модулем семантического анализа вносит вклад в методологию построения отраслевых экспертных систем.

На практике подтвердилась эффективность полученного решения, сократив время первичной обработки данных. Это позволяет ускорить рутинные процессы и заложить основу к системному подходу к расследованию аварий на производственных объектах.

Направления дальнейшего развития включают: расширение словаря семантических шаблонов на основе расширенного корпуса документов, интеграцию методов компьютерного зрения для извлечения фактов из медиаматериалов, автоматизацию ранжирования гипотез о причинах аварии на основе расчётных показателей и масштабирование базы знаний на множество инцидентов для последующего сравнительного анализа системных причин аварийности.

Литература

1. Кузнецова Е.А. Столетняя история развития методов расследования несчастных случаев на производстве // Экономика, предпринимательство и право. 2021. Т. 11. № 1. С. 193—212.
2. Ahmadi Rad M., Lefsrud L., Hendry M.T. Application of systems thinking accident analysis methods: A review for railways // Safety Science. 2023. Vol. 160. P. 106066.
3. Стасева Е.В., Кузнецов С.Ю., Шаяхметов Р.Р., Идиятуллина З.Ф., Муллагалиев И.Р. Анализ характерных причин аварий и травматизма на объектах газоснабжения // Безопасность техногенных и природных систем. 2021. № 2. С. 2—7.
4. Пигилова Р.Н., Малышева Т.В., Аверьянова Ю.А., Филиппова Ф.М. Статистический анализ аварийности на производстве в результате некорректного функционирования системы энергообеспечения предприятия // Известия Самарского научного центра РАН. 2023. № 2 (112). С. 27—33.
5. Derdowski L.A., Mathisen G.E. Psychosocial factors and safety in high-risk industries: A systematic literature review // Safety Science. 2023. Vol. 157. P. 105948.
6. Ванякин О.В., Сухоруков В.В., Строилов И.А. Актуализация подхода к установлению причин возникновения аварий и инцидентов на опасных производственных объектах в условиях современной добычи и переработки (обогащения) полезных ископаемых // Уголь. 2026. № 2. С. 65—68.
7. Adumene S., Afenyo M., Salehi V., William P. An adaptive model for human factors assessment in maritime operations // International Journal of Industrial Ergonomics. 2022. Vol. 89. P. 103293.
8. Климова И.В., Сазанова Н.В., Махнева А.Н. Опыт проведения технического расследования аварии на опасном производственном объекте нефтепродуктообеспечения // Современные проблемы гражданской защиты. 2023. № 1(46). С. 93—102.
9. Маркарян А.О. Система поддержки принятия решений при расследовании авиационных происшествий // Труды МАИ. 2026. № 141.
10. Шамин Е.А., Зеленко Л.С. Разработка автоматизированного рабочего места следователя // Перспективные информационные технологии (ПИТ 2022): сб. тр. науч.-техн. конф. 2022. С. 95—97.
11. Guntupalli B. Data Lake vs. Data Warehouse: Choosing the Right Architecture // International Journal of Artificial Intelligence, Data Science, and Machine Learning. 2023. Vol. 4. № 4. P. 54—64.
12. Кудрявцев С., Емелин П., Козлов В. Анализ баз данных о техногенных чрезвычайных ситуациях в горнодобывающей отрасли // Механика және технологиялар. 2026. № 3 (89). С. 371—381.
13. Sahoo S., Mukane P., Maiti J., Tewari V.K. A framework for process risk assessment incorporating prior hazard information in text mining models using chunking // Process Safety and Environmental Protection. 2024. Vol. 189. P. 486—499.
14. Karimi Dehkordi M., Sattari F., Lefsrud L. Creating an incident investigation framework for a complex socio-technical system: Application of multi-label text classification and Bayesian network structure learning // Reliability Engineering and System Safety. 2025. Vol. 260.
15. Ерошенко С.А., Матренин П.В., Хальясмаа А.И., Клименко Д.Е., Сидорова А.В. Технология обработки данных для прогнозирования притока воды в водохранилище при использовании дистанционного зондирования Земли и сети метеорологических и гидрологических постов // Проблемы региональной энергетики. 2022. № 4 (56). С. 100—109.
16. Захаров В.Н., Клебанов Д.А., Макеев М.А., Радченко Д.Н. Анализ методов подготовки и преобразования информации, поступающей в хранилища данных для эффективного управления горнотехнической системой // Горный информационно-аналитический бюллетень. 2021. № 6. С. 120—135.
17. Machine learning-based literature review on the concept of safety barriers against hazardous events // Journal of Loss Prevention in the Process Industries. 2024. Vol. 92. P. 105470.
18. Rodriguez M., Fernandez A. Deep learning for accident report classification in high-risk industries // Safety Science. 2026. Vol. 165. P. 106210.
19. Wang H., Li J., Liu Y. Knowledge graph and machine learning for accident pattern recognition // Reliability Engineering and System Safety. 2024. Vol. 248. P. 110124.
20. Thompson M., Lee S. Interoperability in corporate safety management systems: A data integration framework // Journal of Safety Research. 2024. Vol. 90. P. 245—260.
21. Kumar A., Singh R. Implementation of investigation support systems in Indian manufacturing industries: Challenges and success factors // International Journal of Industrial Ergonomics. 2024. Vol. 99. P. 103512.
22. Потекаева Ю.В., Охотников И.В., Сибирко И.В. Формирование информационной архитектуры промышленного предприятия // Экономика: вчера, сегодня, завтра. 2021. Т. 11. № 8-1. С. 191—200.

SECTION:
SYSTEM ANALYSIS, MANAGEMENT, ARTIFICIAL INTELLIGENCE, AND INFORMATION PROTECTION

KNOWLEDGE BASE IN THE SYSTEM OF INVESTIGATING ACCIDENTS AND INCIDENTS AT INDUSTRIAL FACILITIES

Markarian A.O.[4], Polyakhovskaya O.A.[5], Yakubovskaya Yu.A.[6]

Keywords: *information system, investigation, knowledge formalization, database, data processing rules, natural language processing, regular expressions, Lakehouse architecture, industrial safety, digitalization.*

Abstract

Purpose of the study: *development of a knowledge base — a centralized system for storage and automated processing of data for investigating the causes of accidents and incidents at industrial facilities, providing unified access to heterogeneous information and reducing the time of accident analysis.*

Methods: *system analysis of the investigation process and existing information solutions in the field of industrial safety; mathematical apparatus of set theory for constructing a formal data model; natural language processing methods and regular expressions for automated fact extraction from heterogeneous documents; Lakehouse architecture design and technologies for software implementation of the system.*

Results: *an analysis of the accident investigation regulations and existing information solutions was conducted, revealing the need to develop a toolkit for centralized storage of heterogeneous investigation materials with automated fact extraction and calculation of regulatory indicators. A conceptual model was proposed and knowledge formalization of the investigation system was carried out: data marts, entities characterizing the subject area, relationships between them, and rules for traceable fact extraction from documents of various types. The database structure was defined, and an algorithm for its automated population was proposed, providing for document routing, application of specialized parsers, and calculation of indicators based on investigation materials. Software based on the Lakehouse architecture was developed, including a data processing module and an administrator panel. Testing on real investigation materials showed fact extraction precision of 91.5%, recall of 88.7%, complete coincidence of calculated indicators with manual calculations, and a reduction in the time of knowledge formation for an incident by approximately 97%.*

References

1. Kuznetsova E.A. Stoletniaia istoriia razvitiia metodov rassledovaniia neschastnykh sluchaev na proizvodstve // Ekonomika, predprinimatel'stvo i pravo. 2021. T. 11. No. 1. S. 193—212.
2. Ahmadi Rad M., Lefsrud L., Hendry M.T. Application of systems thinking accident analysis methods: A review for railways // Safety Science. 2023. Vol. 160. P. 106066.
3. Staseva E.V., Kuznetsov S.Iu., Shaiakhmetov R.R., Idiiatullina Z.F., Mullagaliev I.R. Analiz kharakternykh prichin avarii i travmatizma na ob"ektakh gazosnabzheniia // Bezopasnost' tekhnogennykh i prirodnykh sistem. 2021. No. 2. S. 2—7.
4. Pigilova R.N., Malysheva T.V., Aver'ianova Iu.A., Filippova F.M. Statisticheskii analiz avariinosti na proizvodstve v rezul'tate nekorrektnogo funktsionirovaniia sistemy energoobespecheniia predpriiatiia // Izvestiia Samarskogo nauchnogo tsentra RAN. 2023. No. 2 (112). S. 27—33.
5. Derdowski L.A., Mathisen G.E. Psychosocial factors and safety in high-risk industries: A systematic literature review // Safety Science. 2023. Vol. 157. P. 105948.

[4] ***Markarian Anna Oganesovna***, *Ph.D. (Technology), Associate Professor of the Department of Automated Control Systems, National University of Science and Technology MISIS, Moscow, Russian Federation. ORCID: 0009-0004-2181-2144.*
E-mail: anestrelle@gmail.com

[5] ***Olga A. Polyakhovskaya***, *Student, Automated Control Systems Department, National University of Science and Technology MISIS, Moscow, Russian Federation. ORCID: 0009-0005-4053-6357.*
E-mail: plkhvskl@gmail.com

[6] ***Yulia A. Yakubovskaya, Student***, *Automated Control Systems Department, National University of Science and Technology MISIS, Moscow, Russian Federation. ORCID: 0009-0009-7196-0774.*
E-mail: yakuboyulia@gmail.com

6. Vaniakin O.V., Sukhorukov V.V., Stroilov I.A. Aktualizatsiia podkhoda k ustanovleniiu prichin vozniknoveniia avarii i intsidentov na opasnykh proizvodstvennykh ob"ektakh v usloviiakh sovremennoi dobychi i pererabotki (obogashcheniia) poleznykh iskopaemykh // Ugol'. 2026. No. 2. S. 65—68.
7. Adumene S., Afenyo M., Salehi V., William P. An adaptive model for human factors assessment in maritime operations // International Journal of Industrial Ergonomics. 2022. Vol. 89. P. 103293.
8. Klimova I.V., Sazanova N.V., Makhneva A.N. Opyt provedeniia tekhnicheskogo rassledovaniia avarii na opasnom proizvodstvennom ob"ekte nefteproduktoobespecheniia // Sovremennye problemy grazhdanskoi zashchity. 2023. No. 1(46). S. 93—102.
9. Markarian A.O. Sistema podderzhki priniatiia reshenii pri rassledovanii aviatsionnykh proisshestvii // Trudy MAI. 2026. No. 141.
10. Shamin E.A., Zelenko L.S. Razrabotka avtomatizirovannogo rabochego mesta sledovatelia // Perspektivnye informatsionnye tekhnologii (PIT 2022): sb. tr. nauch.-tekhn. konf. 2022. S. 95—97.
11. Guntupalli B. Data Lake vs. Data Warehouse: Choosing the Right Architecture // International Journal of Artificial Intelligence, Data Science, and Machine Learning. 2023. Vol. 4. No. 4. P. 54—64.
12. Kudriavtsev S., Emelin P., Kozlov V. Analiz baz dannykh o tekhnogennykh chrezvychainykh situatsiiakh v gornodobyvaiushchei otrasli // Mekhanika zhäne tekhnologiialar. 2026. No. 3 (89). S. 371—381.
13. Sahoo S., Mukane P., Maiti J., Tewari V.K. A framework for process risk assessment incorporating prior hazard information in text mining models using chunking // Process Safety and Environmental Protection. 2024. Vol. 189. P. 486—499.
14. Karimi Dehkordi M., Sattari F., Lefsrud L. Creating an incident investigation framework for a complex socio-technical system: Application of multi-label text classification and Bayesian network structure learning // Reliability Engineering and System Safety. 2025. Vol. 260.
15. Eroshenko S.A., Matrenin P.V., Khal'iasmaa A.I., Klimenko D.E., Sidorova A.V. Tekhnologiia obrabotki dannykh dlia prognozirovaniia pritoka vody v vodokhranilishche pri ispol'zovanii distantsionnogo zondirovaniia Zemli i seti meteorologicheskikh i gidrologicheskikh postov // Problemy regional'noi energetiki. 2022. No. 4 (56). S. 100–109.
16. Zakharov V.N., Klebanov D.A., Makeev M.A., Radchenko D.N. Analiz metodov podgotovki i preobrazovaniia informatsii, postupaiushchei v khranilishcha dannykh dlia effektivnogo upravleniia gornotekhnicheskoi sistemoi // Gornyi informatsionno-analiticheskii biulleten'. 2021. No. 6. S. 120—135.
17. Machine learning-based literature review on the concept of safety barriers against hazardous events // Journal of Loss Prevention in the Process Industries. 2024. Vol. 92. P. 105470.
18. Rodriguez M., Fernandez A. Deep learning for accident report classification in high-risk industries // Safety Science. 2026. Vol. 165. P. 106210.
19. Wang H., Li J., Liu Y. Knowledge graph and machine learning for accident pattern recognition // Reliability Engineering and System Safety. 2024. Vol. 248. P. 110124.
20. Thompson M., Lee S. Interoperability in corporate safety management systems: A data integration framework // Journal of Safety Research. 2024. Vol. 90. P. 245—260.
21. Kumar A., Singh R. Implementation of investigation support systems in Indian manufacturing industries: Challenges and success factors // International Journal of Industrial Ergonomics. 2024. Vol. 99. P. 103512.
22. Potekaeva Iu.V., Okhotnikov I.V., Sibirko I.V. Formirovanie informatsionnoi arkhitektury promyshlennogo predpriiatiia // Ekonomika: vchera, segodnia, zavtra. 2021. T. 11. No. 8-1. S. 191—200.