Тимофеев Виталий Андреевич, магистрант, федеральное государственное бюджетное образовательное учреждение высшего образования «Московский авиационный институт (национальный исследовательский университет)», Россия, г. Москва

Цвиклич Эрик Томасович, магистрант, федеральное государственное бюджетное образовательное учреждение высшего образования «Московский авиационный институт (национальный исследовательский университет)», Россия, г. Москва

УПРАВЛЕНИЕ МЕТАДАННЫМИ В ГЕТЕРОГЕННЫХ ИНФОРМАЦИОННЫХ ЭКОСИСТЕМАХ: МЕТОДЫ ИНТЕГРАЦИИ И СИСТЕМНЫЕ ВЫЗОВЫ

Статья рассматривает интеграцию метаданных в корпоративные каталоги данных. Анализируются ограничения стандартных коннекторов и ручного управления метаданными при масштабировании. Предлагаются две стратегии: программная интеграция через REST API для гибкого подключения нестандартных систем и применение искусственного интеллекта для автоматического обогащения и классификации метаданных. Выявлены риски ИИ-подхода: галлюцинации моделей и статистические ошибки. Рекомендуется гибридный подход, где автоматизация дополняется экспертной верификацией, обеспечивая актуальность и надежность каталога как основы Data Governance.

The article examines metadata integration in corporate data catalogs. It analyzes limitations of standard connectors and manual metadata management at scale. Two strategies are proposed: programmatic integration via REST API for flexible integration of non-standard systems, and application of artificial intelligence for automatic metadata enrichment and classification. Risks of the AI approach are identified: model hallucinations and statistical errors. A hybrid approach is recommended, where automation is complemented by expert verification, ensuring the accuracy and reliability of the catalog as a foundation for Data Governance.

Ключевые слова: метаданные, корпоративный каталог данных, интеграция данных, REST API, искусственный интеллект, Data Governance, автоматизация.

Keywords: metadata, corporate data catalog, data integration, REST API, artificial intelligence, Data Governance, automation.

Введение

Современные организации функционируют в среде данных исключительной сложности: десятки и сотни разнородных систем — от реляционных баз данных и облачных хранилищ до потоковых платформ и BI-инструментов — генерируют непрерывные потоки информации. Однако масштаб и доступность данных сами по себе не трансформируются в ценность. Напротив, отсутствие четкой структуры, контекста и понимания происхождения данных подрывает доверие к ним, приводя к ошибкам, неэффективности и упущенным возможностям.

Ключом к преодолению этой ситуации является зрелая система управления данными (Data Governance), фундаментом которой служит эффективное управление метаданными. Корпоративный каталог данных, выступая центральным репозиторием этих «данных о данных», призван стать единой точкой доступа, превращая разрозненные технические активы в понятные, доверенные бизнес-активы. Практическая реализация этой цели упирается в критическую проблему интеграции: как обеспечить согласованное, автоматизированное и полное поступление метаданных из всей экосистемы источников в централизованный каталог.

В данной статье рассматриваются современные стратегии и архитектурные подходы к решению этой задачи. Основное внимание уделяется программной интеграции через API как универсальному методу подключения гетерогенных систем, а также оценивается роль искусственного интеллекта в автоматизации наполнения и поддержания актуальности каталога. Цель работы — представить комплексный взгляд на построение масштабируемого и устойчивого контура управления метаданными в распределенной информационной среде.

Метаданные и их роль в управлении данными

Метаданные традиционно определяются как «данные о данных», представляя собой структурированную информацию, которая описывает контекст, характеристики и инструкции по управлению информационными активами. Они включают в себя критически важные сведения о происхождении (lineage), физической структуре (схемах), владении, правах доступа и версиях данных, что позволяет пользователям эффективно находить, понимать и использовать ресурсы в сложной экосистеме организации. В корпоративной среде метаданные принято разделять на три основные категории, каждая из которых решает свою задачу. Технические метаданные описывают инфраструктурный уровень: типы и структуры полей в базах данных, форматы хранения, модели ETL-процессов. Бизнес-метаданные накладывают на эту техническую основу бизнес-контекст, включая определения терминов в глоссариях, описания KPI, правила расчета показателей и ответственных за данные лиц. Операционные метаданные фиксируют фактологию использования: историю выполнения задач, частоту обращений, статистику по качеству и производительности процессов.

Таким образом, метаданные выполняют роль связующей ткани между ИТ-системами и бизнес-пользователями. Они являются не просто описанием, а механизмом, который «оживляет необработанные данные», превращая разрозненные, неясные наборы информации в прозрачные, понятные и, как следствие, доверенные стратегические активы, пригодные для анализа и принятия решений.

Роль каталога данных в использовании метаданных

Корпоративный каталог данных представляет собой централизованный репозиторий метаданных, выступающий в качестве единого и полного инвентаря информационных активов организации. Его основная функция — систематизация и консолидация сведений о данных из множества

разрозненных источников, создавая тем самым целостную картину информационного ландшафта. Практическая ценность такого каталога реализуется через несколько ключевых функций. Прежде всего, это поиск и обнаружение: каталог предоставляет интуитивный интерфейс, позволяющий пользователям быстро находить нужные наборы данных. Однако его роль не сводится лишь к поисковой системе. Каталог критически важен для обеспечения контекста, напрямую связывая технические объекты (таблицы, столбцы) с бизнес-глоссариями и определениями, что позволяет аналитикам понимать смысл данных без обращения к ИТ-отделу. Кроме того, современные платформы поддерживают совместную работу, предлагая функции рейтингов, комментариев и обмена знаниями, что превращает каталог в живую социальную сеть данных.

Например, аналитик может мгновенно найти ключевые показатели продаж, включая их бизнес-определения, историю расчёта и отзывы коллег, вместо того чтобы вручную согласовывать запросы с несколькими отделами. Эта функциональность в полной мере реализует принцип данных по запросу (self-service data), когда бизнес-пользователи самостоятельно получают необходимый контекст: происхождение данных (lineage), ответственных владельцев, информацию о качестве и правилах использования. В результате снижается операционная нагрузка на ИТ-подразделения и ускоряются процессы аналитики. В конечном счёте, каталог становится не просто инструментом учёта, а критическим компонентом для реализации современных архитектурных подходов, таких как Data Mesh и Data Fabric, обеспечивая видимость и управляемость данных как продукта в распределённой экосистеме.

Процесс наполнения каталога данных

Для наполнения каталога данных обычно используются три метода: автоматический сбор через коннекторы, программное управление через API и ручное обогащение.

Встроенные стандартные коннекторы предназначены для автоматического извлечения метаданных из распространённых информационных систем. К ним относятся коннекторы для СУБД (PostgreSQL, Oracle и др.), облачных хранилищ (AWS S3, Databricks), BI-платформ (Power BI, Tableau) и других источников. Эти коннекторы позволяют легко подключать источники и синхронизировать метаданные о их структуре, собирая технические и операционные сведения.

Ручное наполнение — это задача дата-стюардов (Data Stewards) и владельцев данных. Стюарды, как эксперты в предметной области, наполняют каталог бизнес-метаданными: определениями терминов в глоссарии, описанием активов и правилами качества. Обычно дата-стюарды заполняют бизнес метаданные по своим активам: по таблицам и отчетам, которые создают или за наполнение которых несут ответственность. Обычно у Data Stewards есть ограниченные права на управление метаданными: они могут управлять только данными, за которые отвечают и не могут подключать новые источники или удалять старые. Администраторы же имеют права на полное управление каталогом, в том числе на управления источниками данных.

API каталога данных (чаще всего REST API) предоставляет возможности для глубокой интеграции. Через него можно программно передавать информацию о происхождении данных (lineage) и управлять активами на уровне администратора, что требует технических навыков работы с документацией API.

Проблемы интеграции и актуализации метаданных

Интеграция нестандартных источников выходит за рамки простого подключения. Для проприетарных, устаревших (legacy) или специализированных систем, для которых не существует готовых коннекторов, необходима разработка индивидуального решения. Это требует не только единовременной настройки соединения, но и проектирования

устойчивого процесса регулярной синхронизации метаданных. Без такого процесса каталог быстро теряет актуальность, что приводит к расхождениям между реальной структурой источника и её отображением, порождая ошибки и подрывая доверие пользователей.

Проблема рутинного ручного труда становится критическим «узким местом» в масштабируемых средах. Например, если владелец схемы в базе данных по умолчанию является владельцем всех входящих в неё таблиц, дата-стюард вынужден вручную назначать это право для каждого объекта. При смене владельца вся процедура повторяется. Эта рутина не только снижает операционную эффективность, демотивируя специалистов, но и напрямую увеличивает риск человеческих ошибок и несогласованности данных. Кроме того, по мере роста зрелости управления данными в организации требования к метаданным усложняются. Бизнес-подразделения могут запрашивать добавление новых атрибутов для классификации активов — например, пометку о содержании персональных данных или указание критичности таблицы для конкретного процесса. Ручное ведение такого расширяющегося набора метаданных для тысяч объектов быстро становится неподъёмной задачей.

Программная интеграция через API как универсальный коннектор

Первым стратегическим ответом на эти вызовы является использование программных интерфейсов (API) каталога данных, которые, будучи зачастую реализованными по стандарту REST, снимают принципиальные ограничения на выбор инструментов для интеграции. Этот подход трансформирует каталог из статичного репозитория в открытую платформу, доступную для взаимодействия с любой системой, способной отправлять HTTP-запросы. Это открывает возможность создания универсальных адаптеров для подключения проприетарных, legacy-систем или уникальных внутренних разработок, для которых не существует готовых коннекторов. Главная сила API заключается в возможности реализовывать бизнес-логику произвольной сложности,

выходящую далеко за рамки простой синхронизации структуры. Например, можно автоматически наследовать атрибуты (например, назначать владельца всех таблиц схемы на основе одного поля), реализовать сложные правила тегирования на основе анализа имен полей или контента, а также запускать массовые операции при изменении организационной структуры.

Ключевым аспектом является возможность настраивать процесс дообогащения метаданных из сторонних систем, создавая единый контекстный слой. Так, можно автоматически дополнять описание датасета ссылкой на соответствующий отчет в BI-системе (например, Tableau или Power BI), актуализировать информацию о согласовании доступа, синхронизируясь с системой управления привилегиями (например, SailPoint), или подтягивать актуальные контакты ответственных лиц из корпоративного каталога (Active Directory). Для обеспечения актуальности метаданных через API может быть реализован один из двух основных паттернов синхронизации. Наиболее распространен процесс полной синхронизации, который включает:

1. сбор полного снимка метаданных из источника-оригинала;

2. извлечение текущего состояния этих объектов из каталога данных;

3. детальное сравнение двух наборов;

4. выполнение необходимых действий на основе разницы: создание новых объектов, архивация устаревших и обновление измененных.

Для источников, поддерживающих механизм отслеживания изменений, более эффективным является паттерн Change Data Capture (CDC), при котором в каталог поступают только инкрементальные изменения (новые таблицы, модифицированные схемы), что значительно снижает нагрузку на системы и позволяет поддерживать метаданные в состоянии, близком к реальному времени. Более того, такой подход позволяет настраивать частоту синхронизации источника и каталога вплоть до режима реального времени,

что также может быть очень важно для поддержки актуальности и качества данных в каталоге данных.

Таким образом, API-подход позволяет выстроить масштабируемый, устойчивый и автоматизированный контур управления метаданными, минимизируя ручной труд и исключая ошибки согласованности. Однако его реализация — это инженерная задача, требующая выделения разработческих ресурсов, создания инфраструктуры для отказоустойчивого выполнения задач (например, с использованием планировщиков вроде Apache Airflow) и поддержки созданных интеграций на всем протяжении их жизненного цикла, что должно быть учтено в стратегии внедрения.

Использование искусственного интеллекта для интеллектуального управления метаданными

Современные корпоративные каталоги данных переживают трансформацию, обусловленную активным внедрением встроенных инструментов искусственного интеллекта (ИИ). Этот тренд выводит каталоги за рамки пассивных справочников, превращая их в активные, самообучающиеся платформы.

Важной особенностью текущего этапа является то, что ключевые ИИ-возможности, такие как серверы моделей (Model Context Protocol, MCP) или готовые коннекторы к крупным языковым моделям (LLM), всё чаще поставляются «из коробки» в составе платформ ведущих вендоров (Alation, Atlan, Databricks Unity Catalog). Это существенно снижает порог входа для организаций, позволяя им использовать передовые технологии без необходимости самостоятельной разработки сложных ML-инфраструктур. ИИ в каталогах данных способен решать широкий спектр практических задач, существенно автоматизируя работу с метаданными.

К ключевым направлениям относятся: Автоматическое генерирование и обогащение метаданных: создание описаний таблиц, полей и бизнес-

процессов на основе анализа структур данных и существующей документации. Интеллектуальная классификация и тегирование: автоматическое присвоение активам тегов, категорий, выявление данных, подпадающих под действие регуляторных требований (GDPR, PCI DSS).

Мониторинг актуальности и качества: постоянный анализ источников на предмет изменений, устаревания данных или аномалий, влияющих на доверие. Преимущества такого подхода очевидны и значительны. ИИ обеспечивает беспрецедентную скорость и масштабируемость наполнения каталога, обрабатывая тысячи активов параллельно. Это приводит к высокой полноте покрытия и консистентности метаданных, так как правила применяются единообразно ко всем объектам. Автоматизация снижает операционную нагрузку на дата-стюардов и устраняет «узкие места», связанные с нехваткой экспертов. Кроме того, ИИ позволяет проактивно поддерживать актуальность каталога, непрерывно анализируя изменения в источниках данных.

Однако недостатки и ограничения ИИ-подхода остаются серьезными и требуют взвешенного отношения. Главный риск — галлюцинации LLM, когда система с высокой уверенностью генерирует правдоподобные, но фактически неверные описания или классификации. Статистические ошибки (в диапазоне 1.5–4.5% для неоднозначных задач) в масштабах предприятия могут привести к массовому загрязнению каталога. Кроме того, внедрение ИИ для каталога данных также может накладывать значительные расходы на оборудование и IT инфраструктуру компании, особенно если есть ограничения на использование облачных решений.

Таким образом, наиболее эффективной на сегодняшний день признается гибридная стратегия, в которой ИИ выступает в роли мощного ассистента для первичного анализа и предложений, а финальное кураторство, верификация и добавление глубокого бизнес-контекста остаются за экспертами-людьми (data stewards). Этот подход позволяет сочетать масштабируемость и скорость машинной обработки с критическим мышлением и предметными знаниями

человека, обеспечивая высокое качество и надежность метаданных в корпоративном каталоге.

Вывод

Реализация эффективного корпоративного каталога данных в условиях гетерогенной ИТ-инфраструктуры является комплексной задачей, центральным элементом которой становится проблема интеграции метаданных. Как показывает анализ, ключом к её решению служит комбинация стратегических подходов.

Программная интеграция через REST API предоставляет необходимую гибкость и глубину для подключения любых источников, автоматизации сложной бизнес-логики и построения устойчивых конвейеров синхронизации. Это формирует технический фундамент актуальности каталога. В свою очередь, инструменты искусственного интеллекта предлагают мощные возможности для автоматического обогащения, классификации и мониторинга метаданных на масштабе, недоступном при ручном управлении.

Наиболее жизнеспособной на сегодняшний день признается гибридная модель, где API-интеграции обеспечивают надежный сбор и структурную синхронизацию, а ИИ выступает в роли интеллектуального ассистента, чьи предложения проходят верификацию и кураторство со стороны экспертов-людей (data stewards). Именно такой симбиоз технологической автоматизации и человеческой экспертизы позволяет трансформировать корпоративный каталог из пассивного справочника в динамичную платформу, обеспечивающую прозрачность, доверие к данным и ускорение процессов на основе данных в организации.

[1] DAMA International. The DAMA Guide to the Data Management Body of Knowledge (DAMA-DMBOK2R). 2nd ed. Technics Publications, 2024.

[2] Tonnarelli, M., and others. Data Catalog Tools: A Systematic Multivocal Literature Review // ScienceDirect, 2025.

[3] Heterogeneous Data Integration: Challenges and Opportunities // PMC National Center for Biotechnology Information, 2024.

[4] Fusco, G., and others. An Approach for Semantic Integration of Heterogeneous Data Sources // PMC, 2020.

[5] Bernardo, BMV. Data Governance & Quality Management — Innovation and Application // Elsevier, 2024.

[6] The Role of Metadata Management in Data Governance // International Journal of Innovative Research in Management and Production Systems, 2021.

[7] Scheider, S., and others. Exploring Metadata Catalogs in Health Care Data Ecosystems // PMC, 2025.

[8] DataVersity. Fundamentals of Metadata Management // 2025.

[9] Yee, M., and others. The NYU Data Catalog: a Modular, Flexible Infrastructur for Data Sharing // JAMIA, 2023.

[10] Packt Publishing. The Definitive Guide to Data Integration. 2024.

[11] Alation Data Catalog: Product Overview [Электронный ресурс]. URL: https://www.alation.com (дата обращения: 26.11.2025).