МИИГАиК
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ГЕОДЕЗИИ И КАРТОГРАФИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования
"Московский государственный университет геодезии и картографии" (МИИГАиК)
Кафедра информатики и геоинформационных технологий

Разработка системы автоматической генерации описаний и структуры для открытых наборов данных с использованием нейросетей

Выполнил:
студент группы 2022-ФГиИБ-ПИ-1б
Хоронеко Леонид Александрович
Научный руководитель:
Кондрашин Д. М.

МИИГАиК
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ГЕОДЕЗИИ И КАРТОГРАФИИ

Актуальность

Современные научные репозитории и платформы обмена данными сталкиваются с острой проблемой масштабируемости метаданных. На текущий момент большинство платформ, включая Kaggle и Zenodo, не обеспечивают автоматической генерации описаний и ключевых тегов для загружаемых датасетов — это вынуждает исследователей и администраторов тратить часы на ручное заполнение метаинформации, что замедляет публикацию, снижает качество описаний и затрудняет поиск данных. Отсутствие стандартизированного, машинночитаемого формата вывода (например, JSON) дополнительно усложняет интеграцию с внешними системами. Разработка специализированного инструмента для автоматического анализа структуры данных и генерации многоязычных описаний позволит значительно ускорить процесс публикации, повысить релевантность поиска и укрепить доверие научного сообщества к открытым данным.

МИИГАиК
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ГЕОДЕЗИИ И КАРТОГРАФИИ

Цель и задачи

Разработка программного инструмента для автоматической генерации описаний и ключевых тегов датасетов, обеспечивающего быструю подготовку метаданных и их интеграцию в научные репозитории. Для достижения цели поставлены следующие задачи:

1. Анализ существующих решений
2. Разработка архитектуры системы
3. Создание алгоритмов определения структуры, типа данных и семантики столбцов для генерации осмысленного текстового описания.
4. Генерация ключевых слов и тегов
5. Создание пользовательского интерфейса
6. Тестирование и оптимизация
7. Подготовка документации и сценариев интеграции

МИИГАиК
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ГЕОДЕЗИИ И КАРТОГРАФИИ

Обзор аналогов

В процессе анализа были рассмотрены следующие существующих решения: ydata-profiling, Dataprep.EDA, DataHub AI Documentation и исследовательский фреймворк AutoDDG. Все они работают только на английском языке, требуют навыков программирования или сложной настройки, не генерируют ключевые слова и не позволяют экспортировать результат в формате README.md. Даже локально работающие аналоги недоступны нетехническим пользователям из-за зависимости от Python, а облачные платформы передают данные в облако и не поддерживают русский язык. Разрабатываемое решение устраняет эти пробелы: оно поддерживает русский и английский языки, предлагает простой веб-интерфейс с перетаскиванием файлов, автоматически создаёт теги, экспортирует результат в JSON и README.md, полностью бесплатно и работает без интернета — что делает его первым в своём классе инструментом для русскоязычных исследователей, студентов и администраторов репозиториев.

МИИГАиК
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ГЕОДЕЗИИ И КАРТОГРАФИИ

Используемые технологии

Для реализации проекта выбран лёгкий и автономный технологический стек:

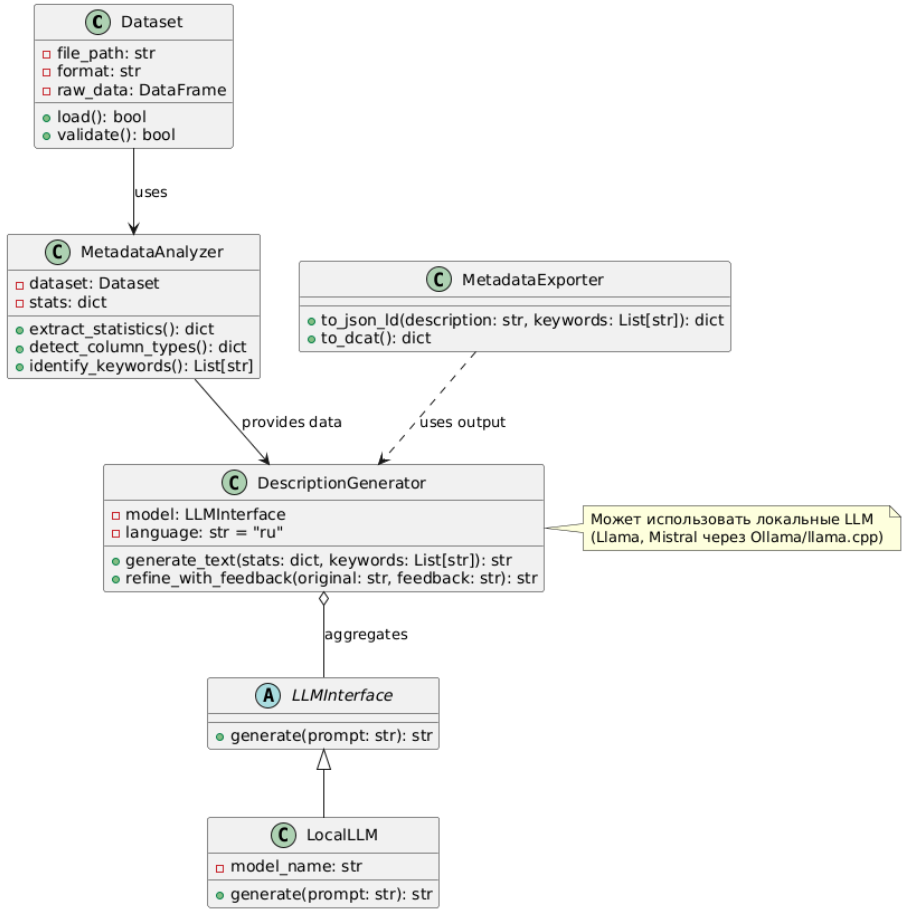
- Python 3.10 — основа системы, обеспечивающая анализ данных и работу с LLM
- Streamlit и Click — веб- и консольный интерфейсы для удобного взаимодействия
- pandas, numpy — для анализа структуры и статистики датасетов
- Ollama / llama.cpp — локальные языковые модели с поддержкой русского языка
- Форматы: CSV, JSON, XLSX, Parquet
- Экспорт: JSON и README.md — для интеграции с репозиториями
- Тестирование: pytest, JMeter

Выбор обусловлен необходимостью создать бесплатное, оффлайн-решение без зависимости от облака, доступное исследователям и администраторам

МИИГАиК
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ГЕОДЕЗИИ И КАРТОГРАФИИ

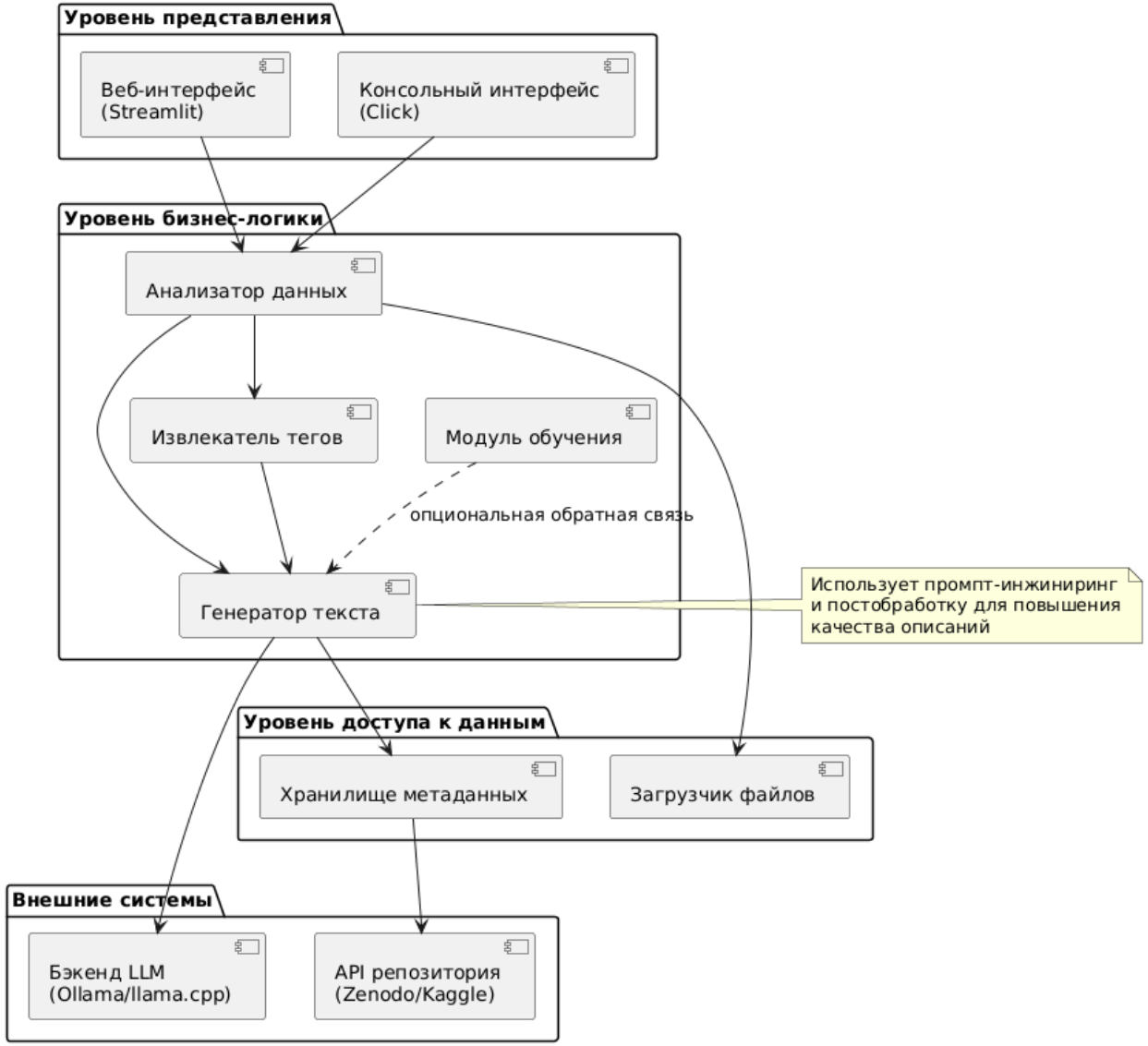
Проектирование и архитектура

Диаграмма классов



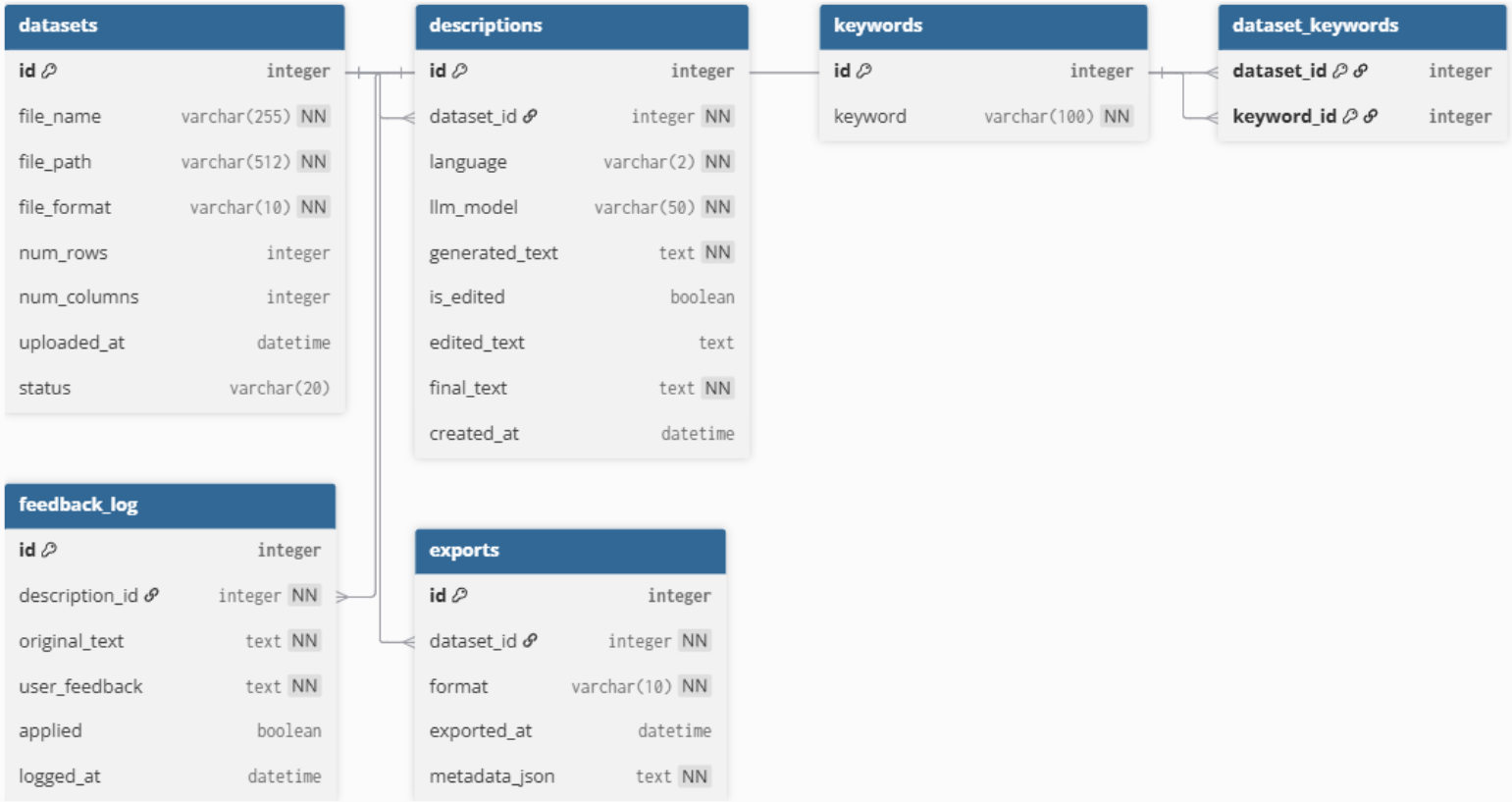
МИИГАиК
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ГЕОДЕЗИИ И КАРТОГРАФИИ

Диаграмма компонентов



МИИГАиК
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ГЕОДЕЗИИ И КАРТОГРАФИИ

ER-диаграмма базы данных

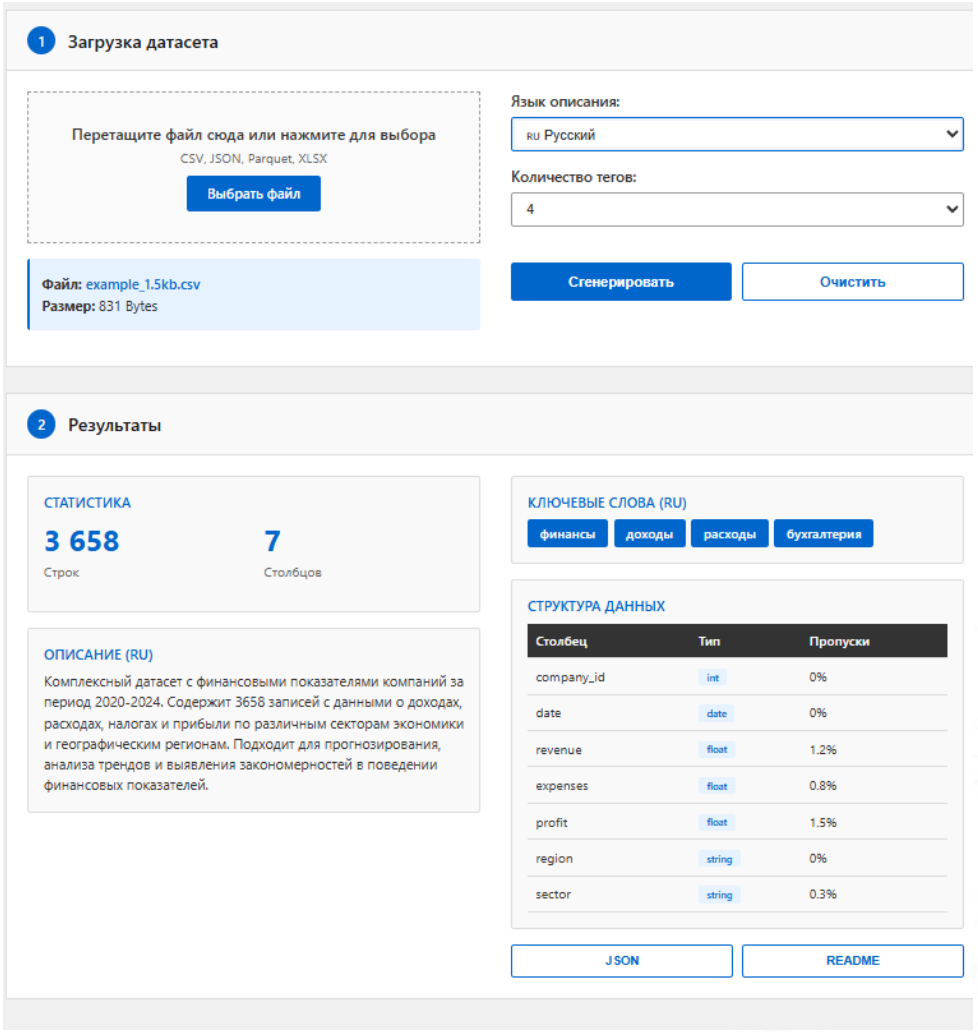


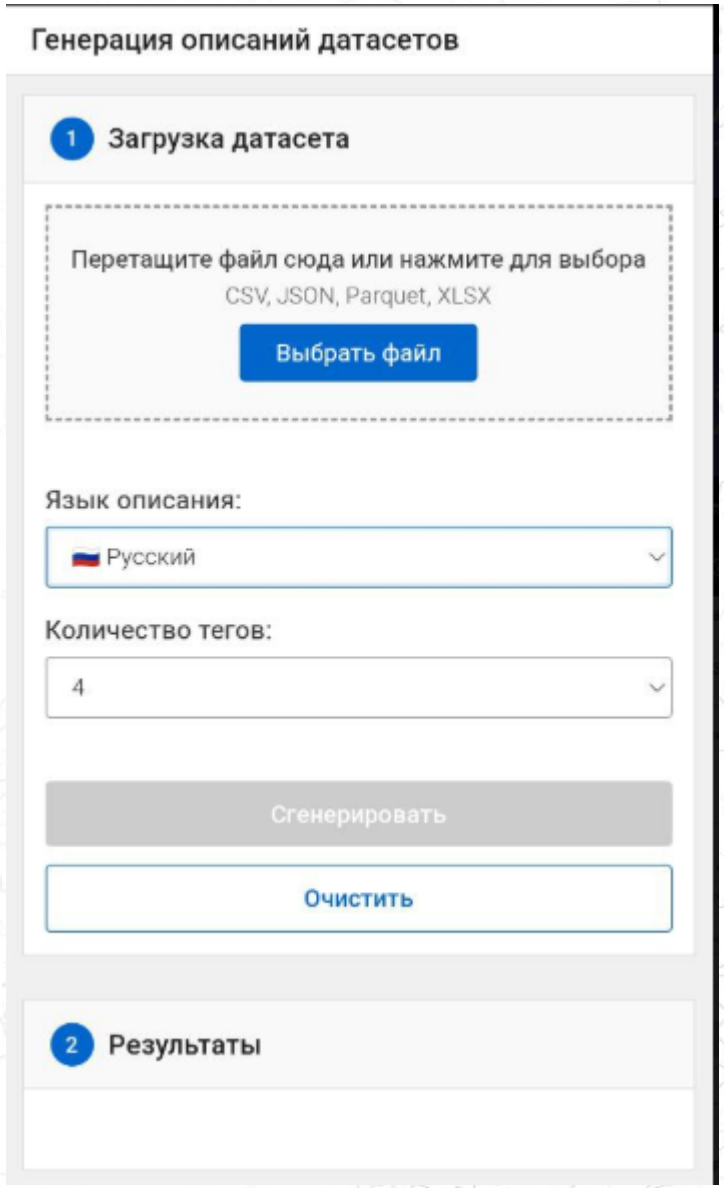
МИИГАиК
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ГЕОДЕЗИИ И КАРТОГРАФИИ

Диаграмма вариантов использования

МИИГАиК
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ГЕОДЕЗИИ И КАРТОГРАФИИ

Демонстрация работы





МИИГАиК
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ГЕОДЕЗИИ И КАРТОГРАФИИ

Результаты и анализ

В ходе тестирования система показала:

- Скорость: среднее время генерации описания — 12 секунд (макс. — 28 сек), что соответствует целевому значению ≤15 сек для 90 % датасетов.
- Потребление памяти: не более 750 МБ (целевой лимит — 800 МБ).
- Точность: экспертная оценка качества описаний — 4.7/5, без вымышленных фактов.
- Релевантность тегов: 89 % ключевых слов признаны экспертами корректными (цель — ≥85 %).
- Юзабилити: 92 % новых пользователей самостоятельно получили результат за ≤3 минуты без инструкций.
- Безопасность: по данным Wireshark, нет исходящих сетевых запросов — все данные обрабатываются локально.
- Экономия времени: подготовка описания сократилась с 60 минут вручную до ≤5 минут (в 12 раз быстрее).
- Покрытие требований: 100 % критически важных и 93 % высокоприоритетных требований выполнены.

Система полностью соответствует целям проекта и готова к использованию..

МИИГАиК
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ГЕОДЕЗИИ И КАРТОГРАФИИ

Перспективы развития

В перспективе развития планируется:

- Реализация полнофункционального терминального (CLI) интерфейса — для пакетной обработки датасетов и интеграции в скрипты и CI/CD-процессы
- Добавление поддержки формата Parquet — расширение совместимости с современными data-science-стеками
- Массовая обработка датасетов — каталогизация целых архивов (например, университетских репозиториев) за один запуск
- Интеграция с GitHub, GitLab, Zenodo, Kaggle — автоматическая генерация README и метаданных при загрузке
- Модуль обучения по правкам пользователя — адаптация описаний на основе обратной связи
- Поддержка дополнительных языков — казахский, белорусский, испанский и др.
- Разработка REST API — для встраивания функционала в корпоративные системы и порталы открытых данных

МИИГАиК
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ГЕОДЕЗИИ И КАРТОГРАФИИ

Заключение

В результате работы был разработан инструмент для автоматической генерации описаний и тегов датасетов, который:

- Позволяет быстро получать понятное описание данных на русском и английском языках
- Поддерживает популярные форматы: CSV, JSON, XLSX
- Генерирует ключевые слова и выдаёт метаданные в JSON и README.md
- Работает локально — без интернета и передачи данных

Прост в использовании — даже без знания программирования

Проект решает реальную проблему: экономит время исследователей и администраторов репозиториев, улучшает качество метаданных и помогает находить нужные данные быстрее.

Готов к использованию в научных, образовательных и аналитических задачах — особенно там, где важны простота, конфиденциальность и поддержка русского языка.

МИИГАиК
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ГЕОДЕЗИИ И КАРТОГРАФИИ

Спасибо за внимание!