

Enhancing Dataset Discovery in Data Catalogs through Knowledge Graphs and Large Language Models

Author: Lucie Vomelová **Supervisor:** prof. Mgr. Martin Nečaský, Ph.D.
Charles University, Faculty of Mathematics and Physics

Motivation

Data catalogs serve as access points for discovering data assets. Unfortunately, many catalogs provide only simple search engines, such as those based on exact string matching. As a result, users often need to run multiple search queries and spend a long time looking through results to find relevant data.

This thesis explores how LLMs and knowledge graphs can be used to improve dataset discovery and searching in data catalogs. We designed, implemented, and evaluated a search system for the Czech National Open Data Catalog (NKOD) and compared it with an existing search engine for the NKOD.

System Introduction

- **Enhanced NKOD Dataset Search**
- Each dataset is represented using its descriptive metadata, such as title, description text, and assigned keywords.
- Processing and enhancement of NKOD data.
- **Search pipeline:** basis of the search engine.
- **Hybrid search:** combining semantic and keyword-based searching.
- **Similar dataset detection** using a knowledge graph.
- **Chatbot** for conversation-based searching.

Architecture

Three main components:

1. **Data processing pipeline:** LLM-enhanced processing of NKOD metadata.
2. **Search platform:** a web application; semantic search, chatbot, and similar dataset detection.
3. **Knowledge base:** NKOD files, document store, vector store, and knowledge graph.

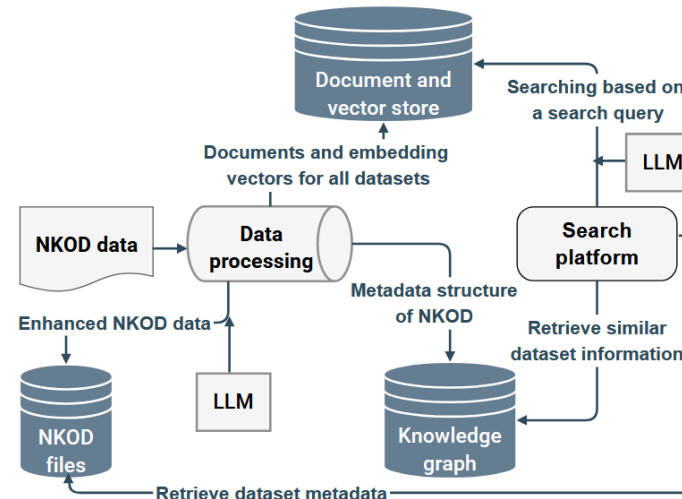


Figure 1: A conceptual system architecture overview

Search pipeline

1. **Query preprocessing:** intent detection and search query expansion.
2. **Retrieval** of relevant datasets based on the expanded search query.
3. **Result postprocessing:** filtering and relevance-based reranking.

Implementation

- Provided as a Python package `EnhancedNkodDatasetSearch`.
- Consists of two subpackages: `data_processing` and `search_platform`.

Evaluation

- Evaluated on the complete NKOD data (over 30,000 datasets).
- Grid search to find candidate configurations.
- User evaluation of candidates to find the best configuration.

Comparison with the existing NKOD application:

- Several advantages of our system identified:
 - Searching is based on the semantic meaning of the search query rather than just keyword matching.
 - Search results ordered by their relevance to the query.
- Evaluation of the new functionality: chatbot and similar dataset detection. Both features were considered useful by all users.

Conclusion

The Enhanced NKOD Dataset Search system enables more effective exploration and discovery of datasets in the NKOD. Its main advantage is semantic search, which considers the meaning of the query rather than relying only on keywords.