Universität Münster

What is a Metadata Lake and Why are We Building One?

Christian Himpe (University and State Library of Münster)

Maths Meets Information Specialists

MaRDI Workshop @ MPI-MIS (Leipzig), 2023-10-09

living.knowledge

Universität Münster

Wait Aren’t You …

- Worked in MaRDI (TA2)
- Joined ULB Münster 09/2022
- In the department of Information Engineering (RnD)
- Working on centralizing metadata of research data
- Practically, data engineering

Universität Münster

The Issue

- Universities have a lot of research data, artifacts and products.
- University libraries have to organize and catalog research.
- University-wide interdisciplinary research discovery platform needed.
- Research data has to be managed and maintained.
- CRUD (Create, Read, Update, Delete) processes have to be automatized.

An Outside View

“Metadata itself is becoming big data”

- P. Salkar: "**The Rise of the Metadata Lake**"; Towards Data Science, 2021.
- M. Yassin: "**The Role of Metadata and Metadata Lake For a Successful Data Architecture**"; Hyperight, 2022.

Universität Münster

Enter DatAasee

A Metadata Lake for the University of Münster

- Centralized metadata hub
- Research data discovery service
- Extensible automation of (meta)data transport

∗ *BTW: The Aasee is an artificial lake in Münster (next to my office).*

What is a Metadata Lake?

A **metadata lake** is a **data lake** restricted to metadata data!

What is a Data Lake?

A **data lake** is a data architecture for **structured**, **semi-structured** and **unstructured** data!

- structured: data conforming to a (relational) schema
- semi-structured: data with a known (interpretable) format
- unstructured: data with unknown format or in raw form

How Does a Data Lake Differ From a Database?

A **data lake** includes **a database** (DBMS), but requires further components to import, export, transform and store data!

- DBMS: DataBase Management System

How Does a Data Lake Differ From a Data Warehouse?

A **data warehouse** transforms input data to fit its schema, cf. **ETL**!

A **data lake** ingests data as-is and transforms it on-demand, cf. **ELT**!

- ETL: Extract, Transform, Load
- ELT: Extract, Load, Transform

How is Data in a Data Lake Organized?

A **data lake** stores raw data and includes a **metadata catalog** that maintains data locations, their metadata, and transformations!

What Makes a Metadata Lake Special?

The **metadata lake**'s **data lake** and **metadata catalog** **coincide**, this implies incoming data is partially transformed, cf. **EtLT**, to hydrate the catalog aspect of the **metadata lake**.

- EtLT: Extract, transform, Load, Transform

∗ J. Densmore: "**Data Pipelines Pocket Reference**"; O'Reilly, 2021.

How Does a Metadata Lake Differ From a Data Catalog?

A **metadata lake**'s data is metadata while a **data catalog**'s data is databases (and their contents)!

* However their feature sets are quite similar.

DatAasee Features

- Full-text search and filter search
- Query languages: **SQL**, **Gremlin**, **Cypher**, **MQL**, **GraphQL**, (**SPARQL**), (**Redis**)
- Ingest from variable protocol / format combinations
- REST-like HTTP API
- Extensible
- Compatibly containerized (**Docker**, **Podman**, **Kubernetes**)
- Open Source (once released)

Native Schema

- Data URL (Virtual Data Lake)
- Raw Metadata (Data Lake)
- Intra Metadata (Process, Technical, Social, Descriptive [based on **DataCite**])
- Inter Metadata (**DataCite** Relationships, **WEMI**: Work, Expression, Manifestion, Item)

Universität Münster

Ingest Protocols

- **OAI-PMH**
- **SQL** (ie **postgres**)
- **SRU**
- …

Universität Münster

Ingest Formats

- DataCite, CodeMeta
- MARC, MODS, DC
- BibTeX
- Schema.org
- re3data
- DCAT
- …

Universität Münster

DatAasee und MaRDI

- How does MaRDI suggest to report about research products and publications? i.e. BibTeX
- Does MaRDI have list of typical formats of research products? i.e. CSV
- What repositories and services does MaRDI suggest to use? i.e. Zenodo
- What APIs and formats does the MaRDI portal export? i.e. SPARQL endpoint
- Will MaRDI synchronize with other NFDIs wrt metadata? i.e. NFDI4ING

DatAasee Software Stack

- Metadata Catalog: **ArcadeDB** (Multi-Model NoSQL Database)
- EtLT Processor: **Benthos** (Declarative Stream Processor)
- Web Frontend: **Lowdefy** (Declarative Web Framework)

Universität Münster

Summary

- **DatAasee** Metadata Lake
- A central hub for research artifacts and outputs
- For cataloging, measuring, and discovering research

What metadata formats does / will MaRDI use?

Bonus: Why a Multi-Model Database?

Relational Model:

S. Jamil: "**Data Lakes and SQL: A Match Made in Data Heaven**"; KDnuggets, 2023.

Graph Model:

A. Govindarajan: "**Implementing the Metadata Lake...**"; Medium, 2023.

Graph Model ⊃ Document Model ⊃ Key-Value Model

Bonus: Why no Knowledge Graph?

Knowledge Graph:

- Subject, Predicate, Object
- Nodes, Edges are URLs

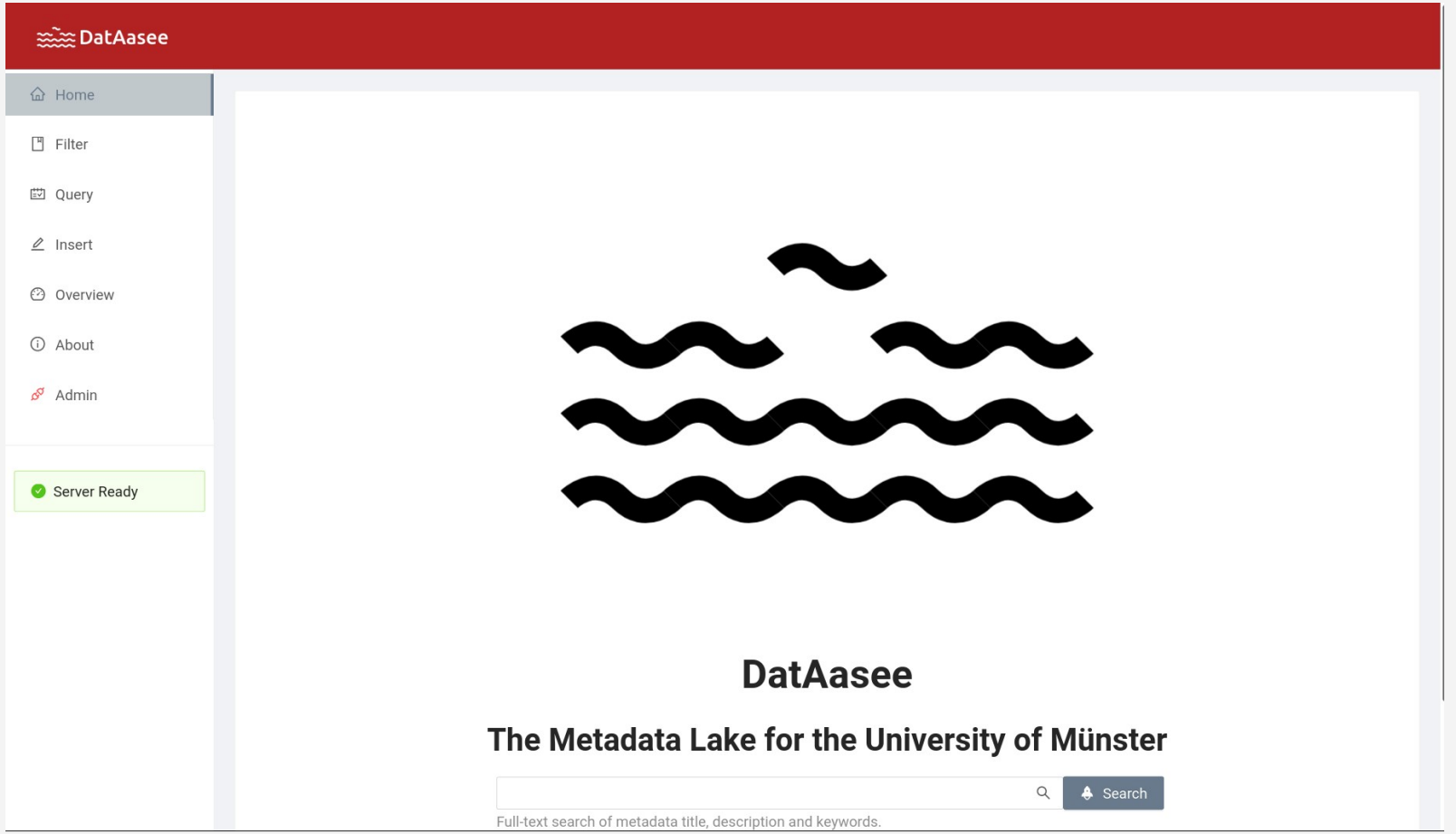
Property Graph:

- Vertex, Edge
- Vertices, Edges are Documents

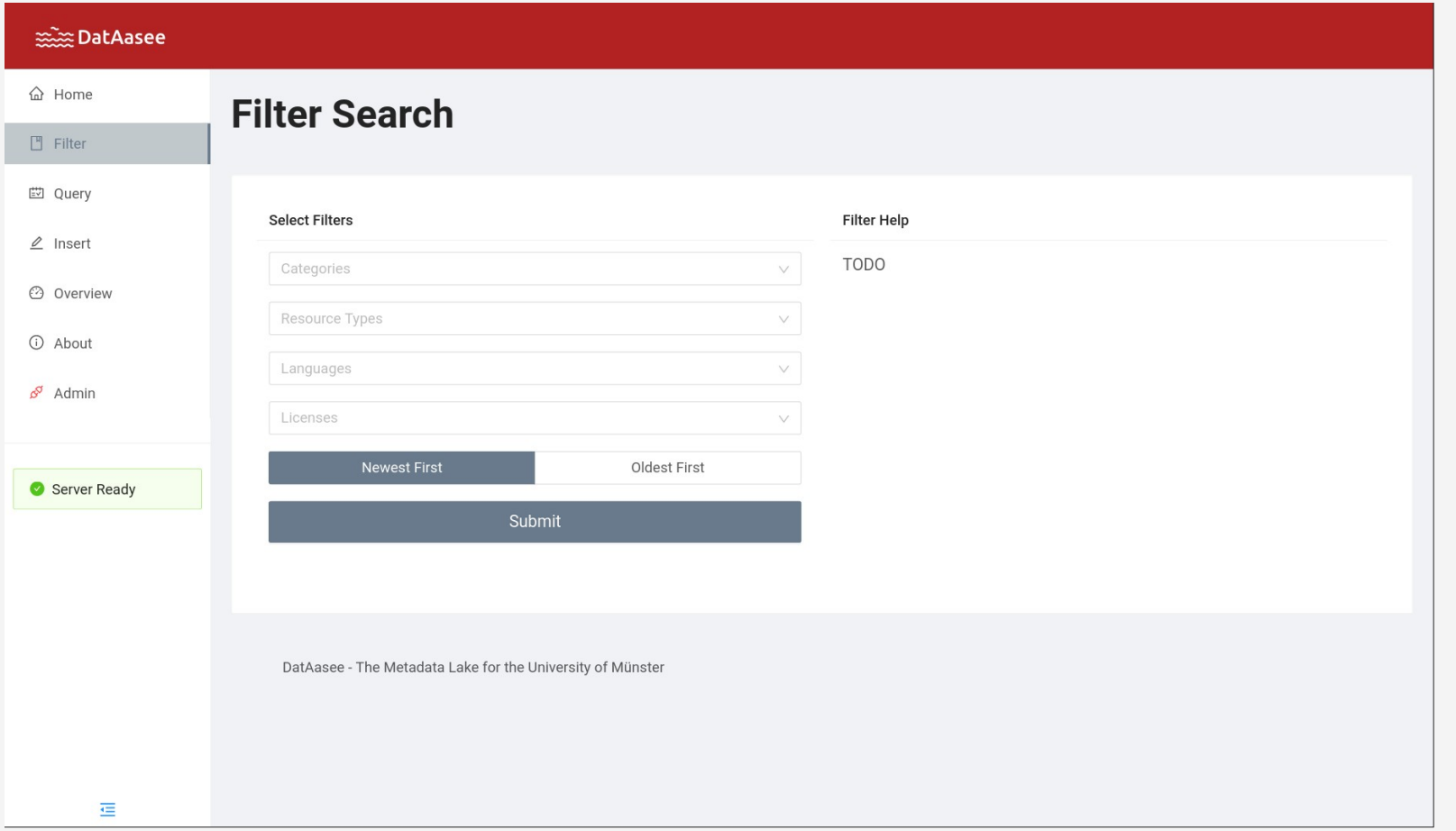
KG and PG are equivalent (ie annotations ↔ properties), but …

Universität Münster

Bonus: Prototype Frontend

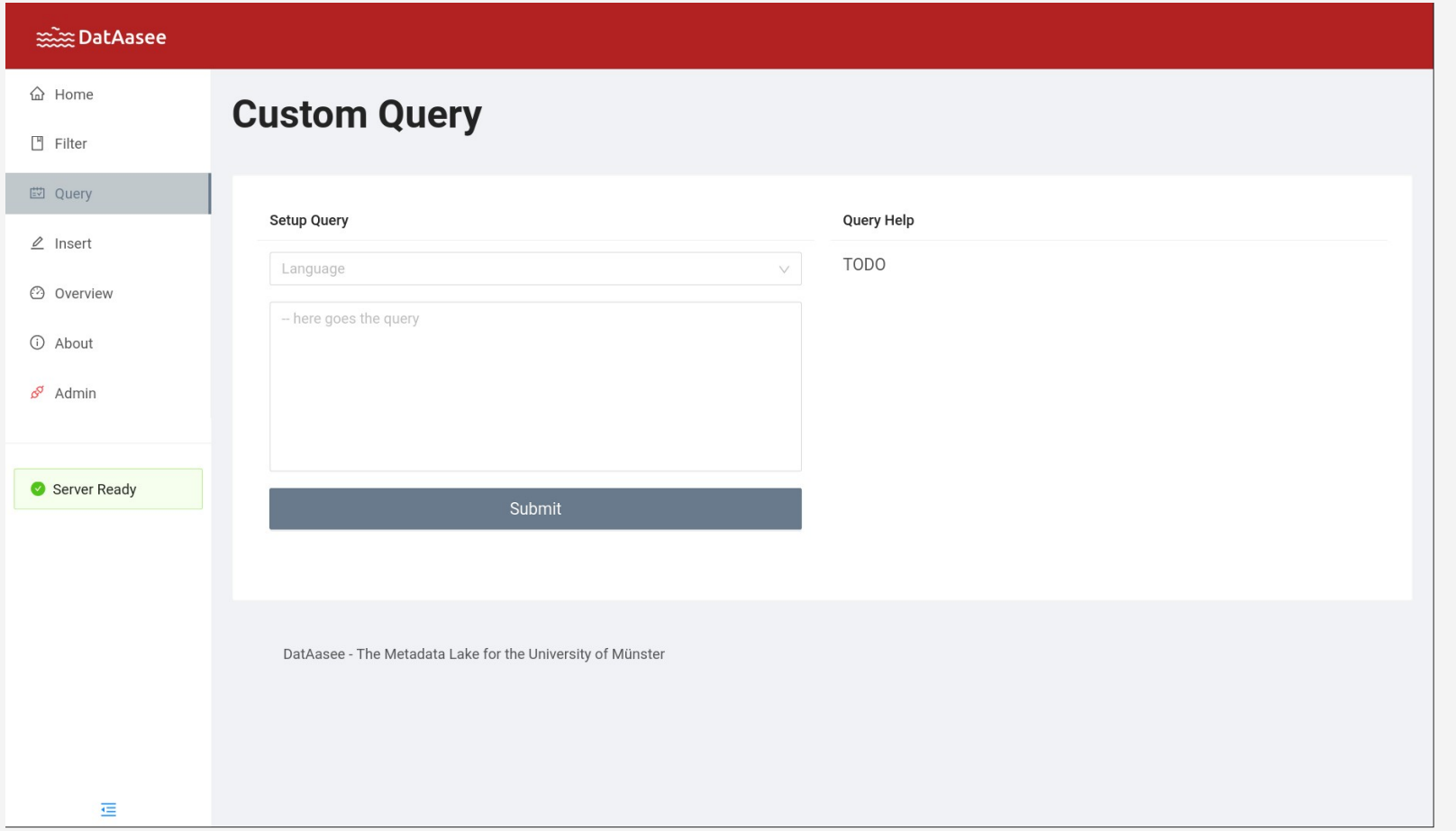


Bonus: Prototype Frontend (Filter Search)



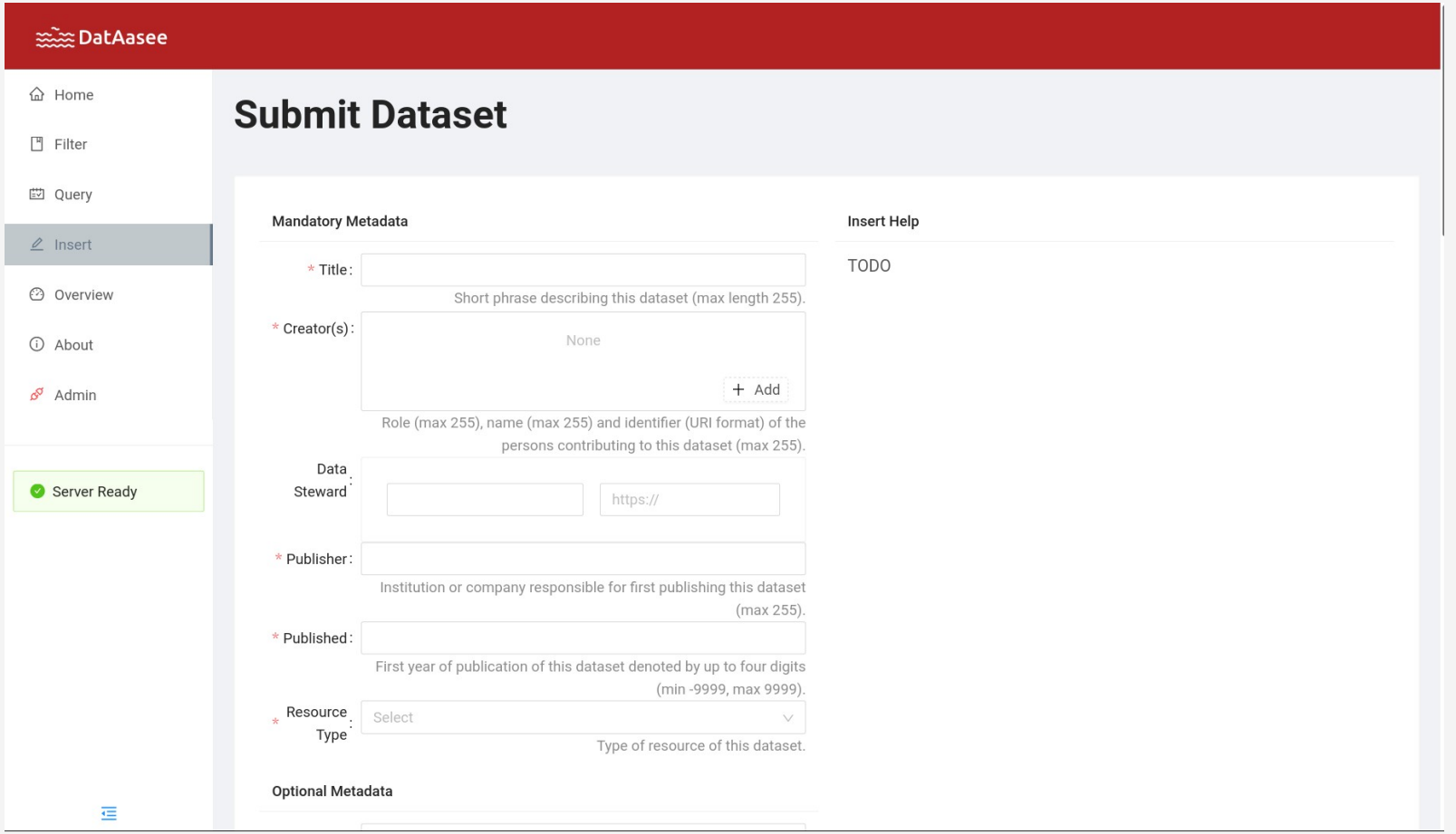
Universität Münster

Bonus: Prototype Frontend (Custom Query)

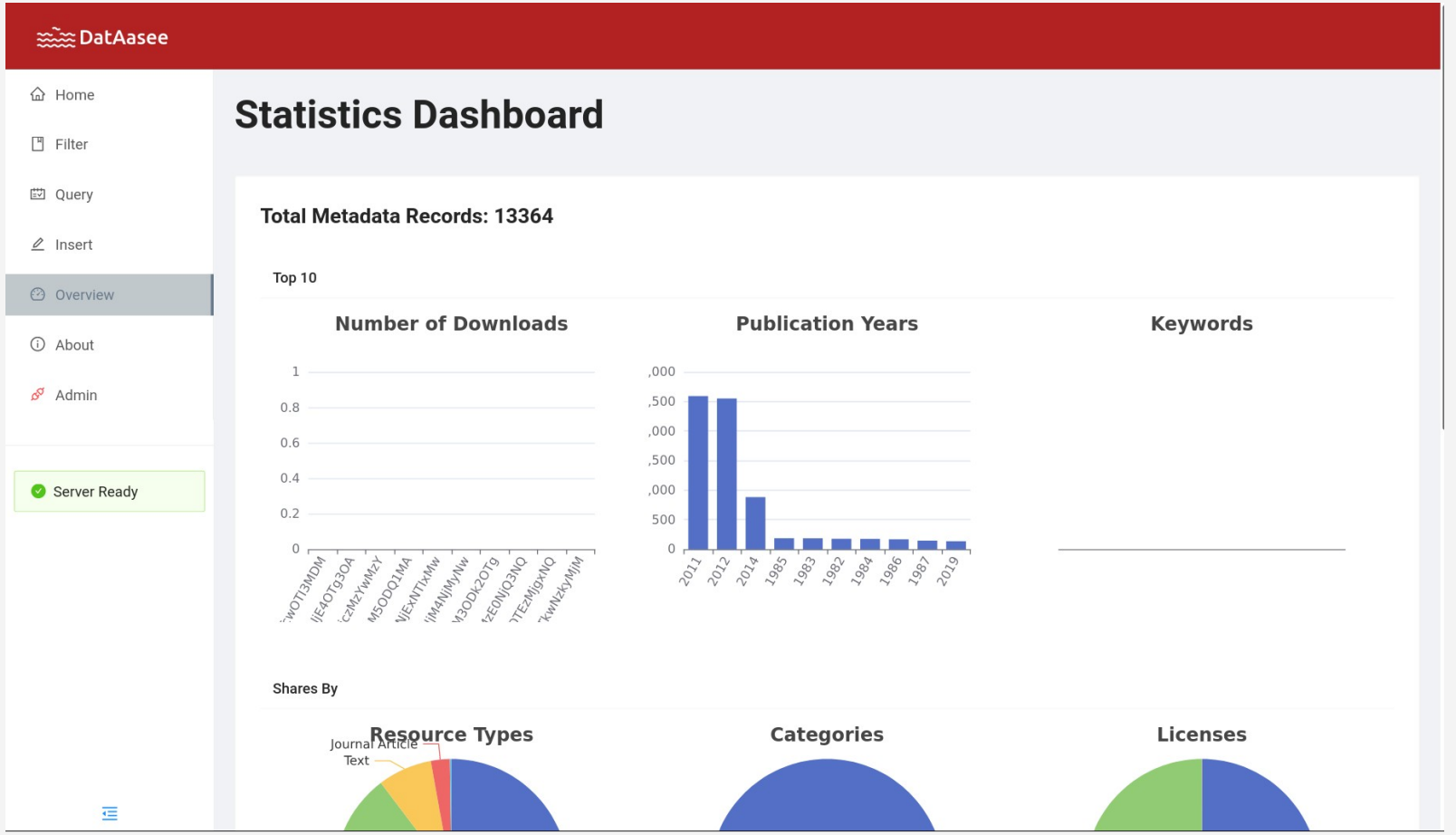


Universität Münster

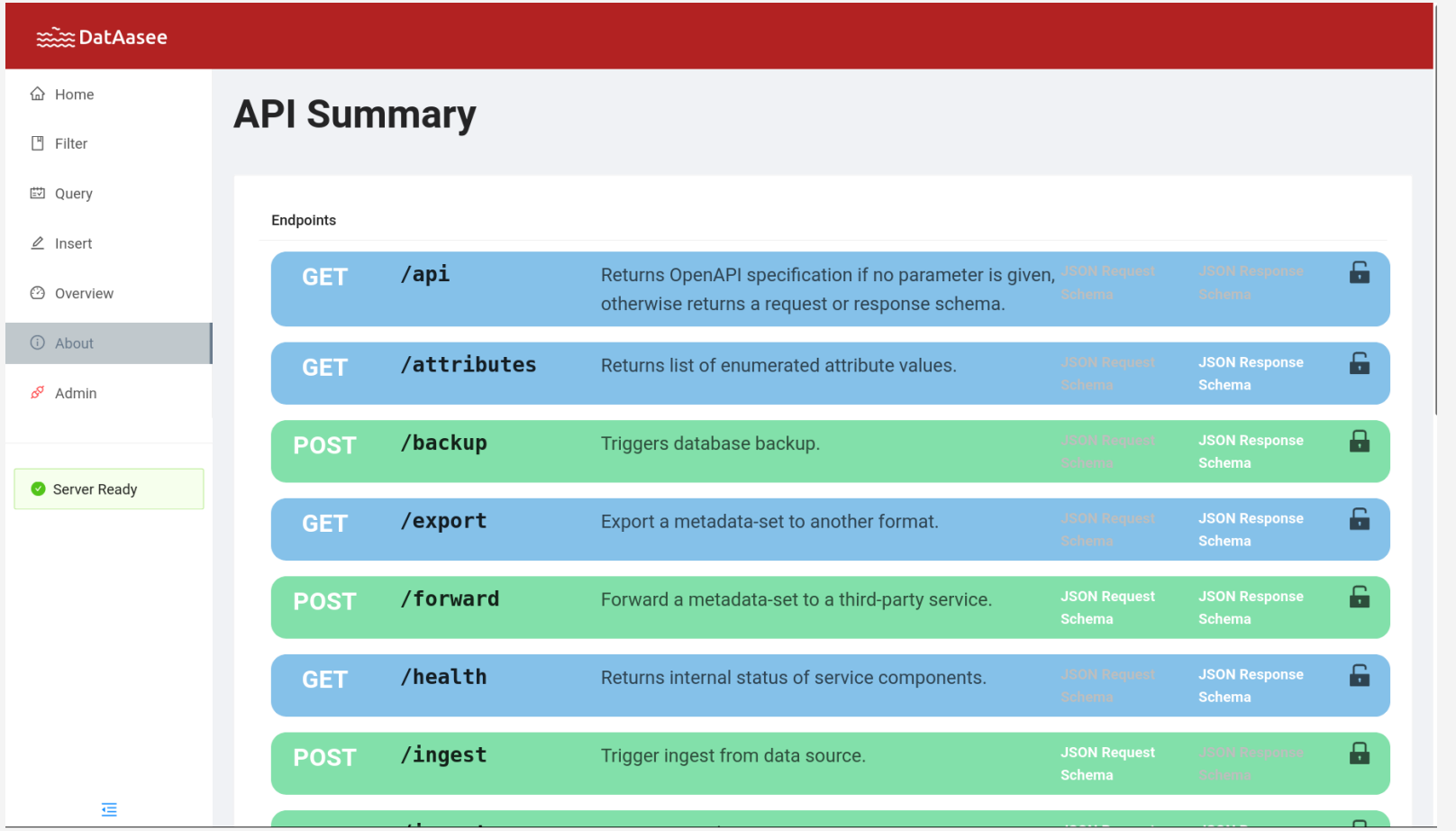
Bonus: Prototype Frontend (Submit Dataset)



Bonus: Prototype Frontend (Statistics Dashboard)



Bonus: Prototype Frontend (API Summary)



Bonus: Prototype Frontend (Admin Panel)