

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

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

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.

Enter **Da**tAasee

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)

Ingest Protocols

- OAI-PMH
- SQL (ie postgres)
- SRU
- ...

Ingest Formats

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

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)

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 \supset Document Model \supset Key-Value Model

Bonus: Why no Knowledge Graph?

Knowledge Graph:

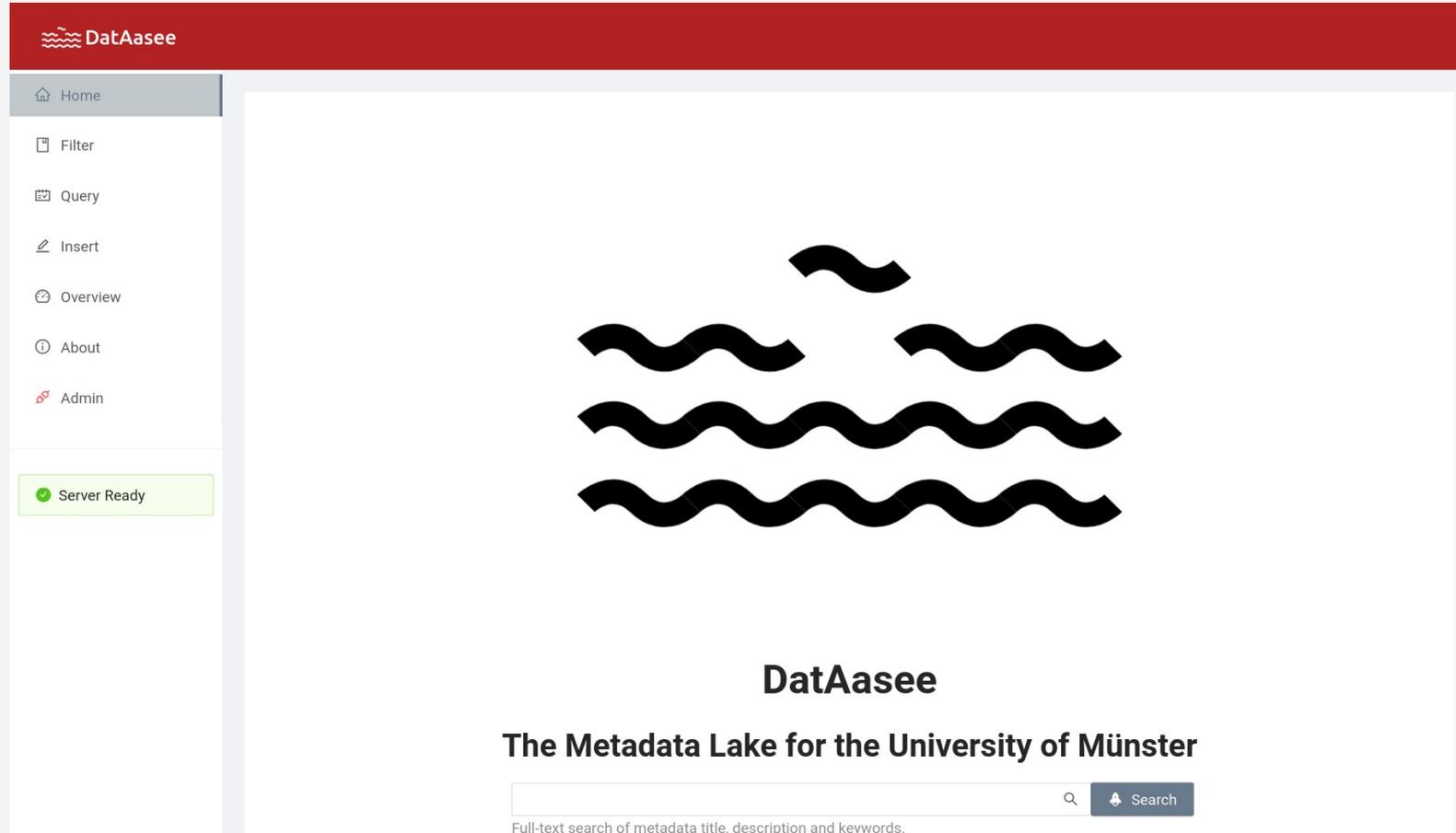
- Subject, Predicate, Object
- Nodes, Edges are URLs

Property Graph:

- Vertex, Edge
- Vertices, Edges are Documents

KG and PG are equivalent (ie annotations \leftrightarrow properties), but ...

Bonus: Prototype Frontend



Bonus: Prototype Frontend (Filter Search)

 DatAasee

Home
Filter
Query
Insert
Overview
About
Admin

✓ Server Ready

Filter Search

Select Filters

Categories

Resource Types

Languages

Licenses

Newest First Oldest First

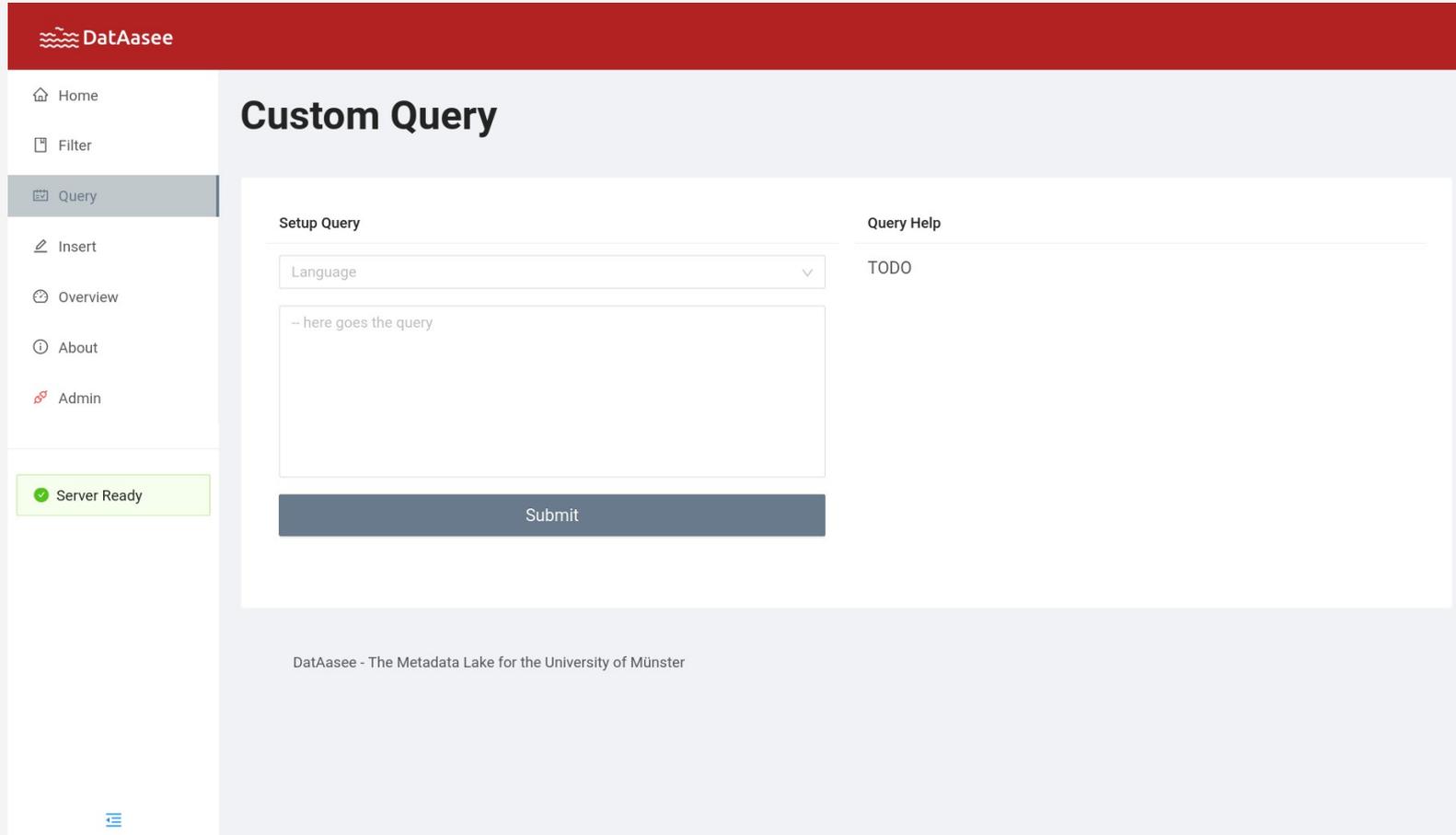
Submit

Filter Help

TODO

DatAasee - The Metadata Lake for the University of Münster

Bonus: Prototype Frontend (Custom Query)



The screenshot displays the DatAasee Custom Query interface. At the top, a red header bar contains the DatAasee logo and name. A left sidebar lists navigation options: Home, Filter, Query (highlighted), Insert, Overview, About, and Admin. Below the sidebar, a green box indicates 'Server Ready'. The main content area is titled 'Custom Query' and is divided into two sections: 'Setup Query' and 'Query Help'. The 'Setup Query' section includes a 'Language' dropdown menu, a large text area for the query (containing the placeholder '-- here goes the query'), and a 'Submit' button. The 'Query Help' section currently displays 'TODO'. At the bottom of the page, the footer text reads 'DatAasee - The Metadata Lake for the University of Münster'.

Bonus: Prototype Frontend (Submit Dataset)

 DatAasee

- Home
- Filter
- Query
- Insert**
- Overview
- About
- Admin

✔ Server Ready

Submit Dataset

Mandatory Metadata

* Title:
Short phrase describing this dataset (max length 255).

* Creator(s):
Role (max 255), name (max 255) and identifier (URI format) of the persons contributing to this dataset (max 255).

Data Steward:

* Publisher:
Institution or company responsible for first publishing this dataset (max 255).

* Published:
First year of publication of this dataset denoted by up to four digits (min -9999, max 9999).

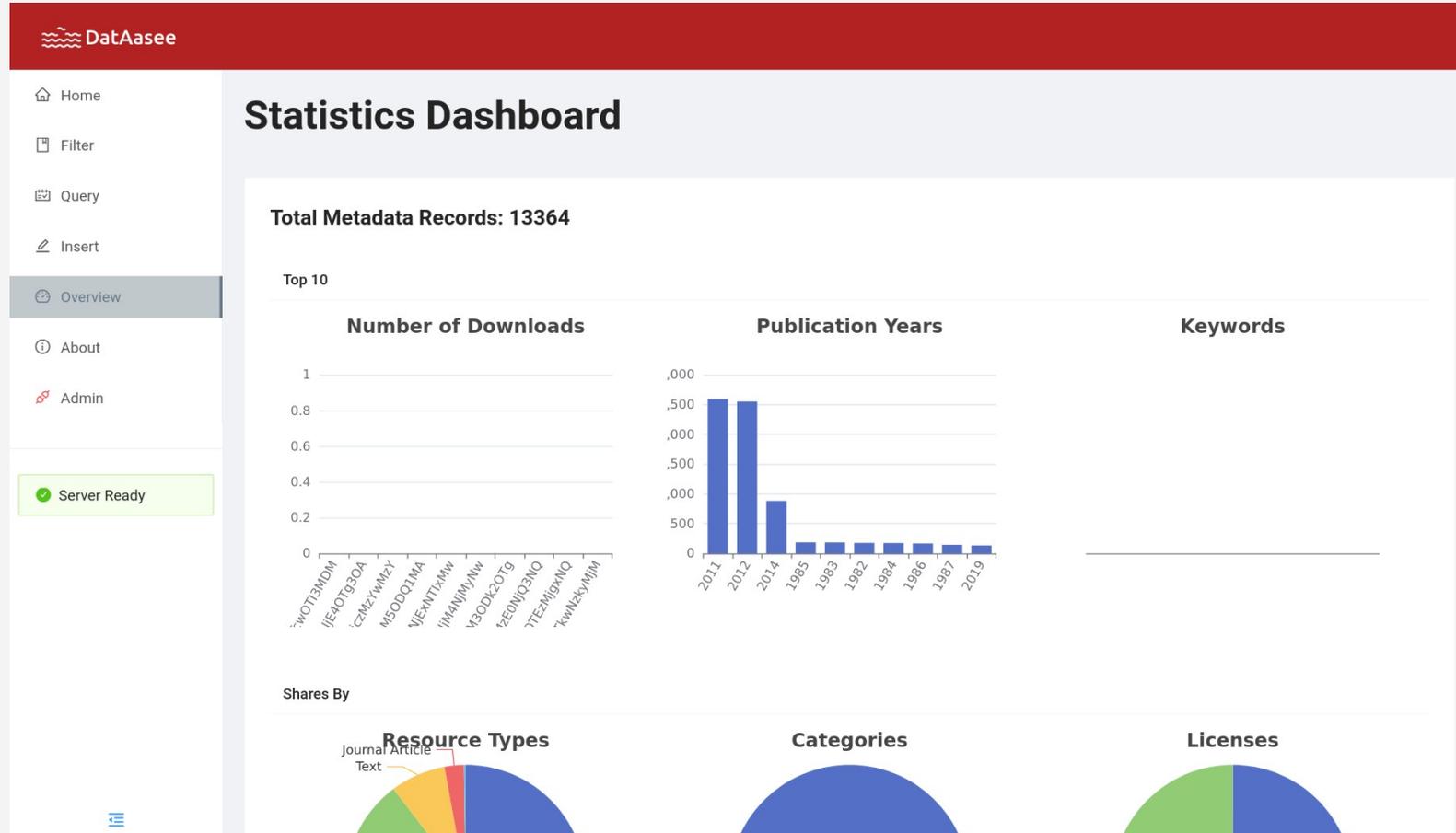
* Resource Type:
Type of resource of this dataset.

Optional Metadata

Insert Help

TODO

Bonus: Prototype Frontend (Statistics Dashboard)



Bonus: Prototype Frontend (API Summary)



DatAasee

Home
Filter
Query
Insert
Overview
About
Admin

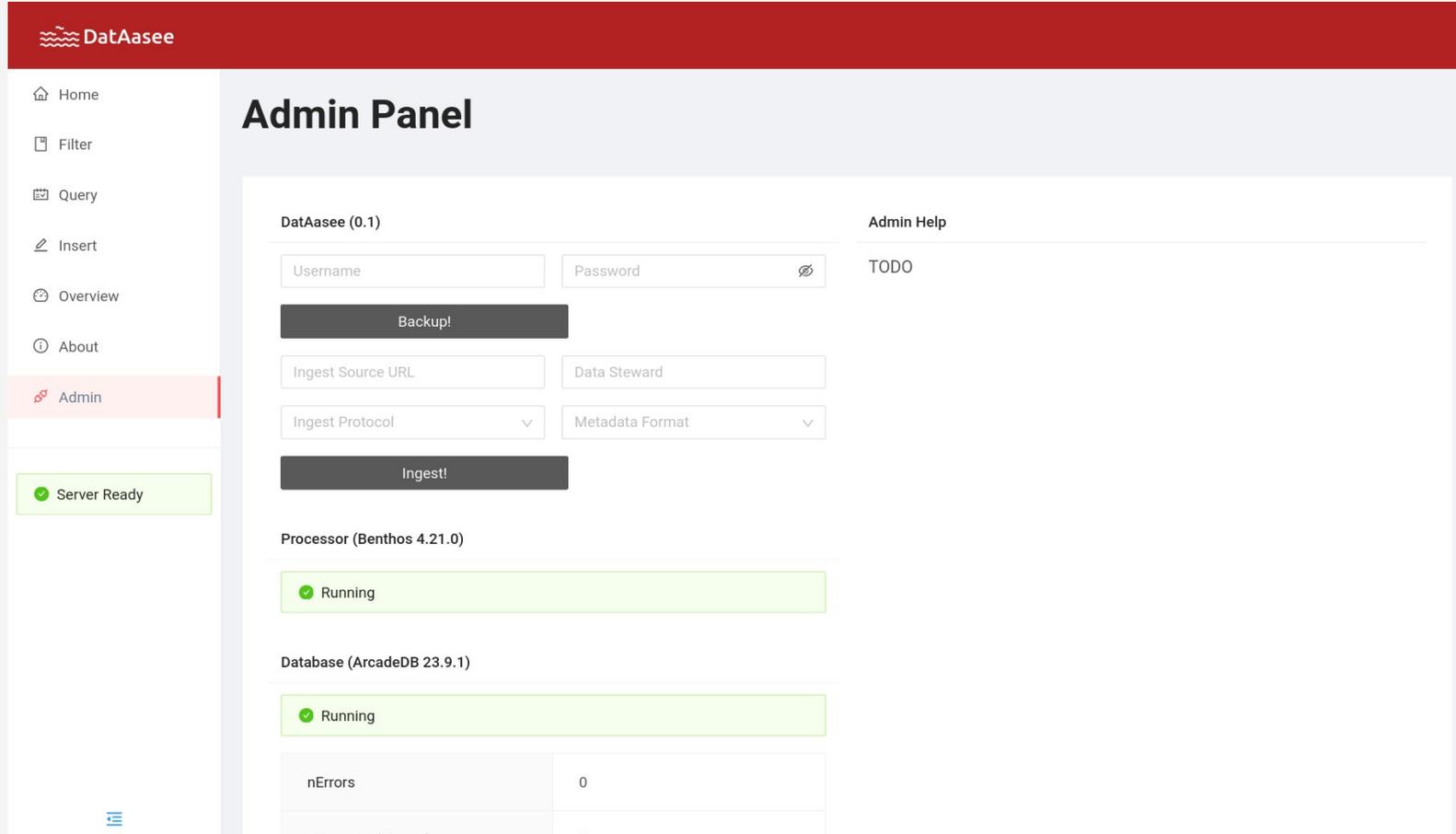
Server Ready

API Summary

Endpoints

Method	Path	Description	Request Schema	Response Schema	Auth
GET	/api	Returns OpenAPI specification if no parameter is given, otherwise returns a request or response schema.	JSON Request Schema	JSON Response Schema	🔒
GET	/attributes	Returns list of enumerated attribute values.	JSON Request Schema	JSON Response Schema	🔒
POST	/backup	Triggers database backup.	JSON Request Schema	JSON Response Schema	🔒
GET	/export	Export a metadata-set to another format.	JSON Request Schema	JSON Response Schema	🔒
POST	/forward	Forward a metadata-set to a third-party service.	JSON Request Schema	JSON Response Schema	🔒
GET	/health	Returns internal status of service components.	JSON Request Schema	JSON Response Schema	🔒
POST	/ingest	Trigger ingest from data source.	JSON Request Schema	JSON Response Schema	🔒

Bonus: Prototype Frontend (Admin Panel)



The screenshot displays the DatAasee Admin Panel. At the top, a red header bar contains the DatAasee logo and name. A left sidebar lists navigation options: Home, Filter, Query, Insert, Overview, About, and Admin (highlighted in red). Below the sidebar, a green box indicates 'Server Ready'. The main content area is titled 'Admin Panel' and is divided into several sections:

- DatAasee (0.1)**: Includes input fields for Username and Password, a Backup! button, Ingest Source URL, Data Steward, Ingest Protocol (dropdown), and Metadata Format (dropdown). An Ingest! button is located below these fields.
- Admin Help**: A section with a 'TODO' item.
- Processor (Benthos 4.21.0)**: A green bar with a checkmark and the text 'Running'.
- Database (ArcadeDB 23.9.1)**: A green bar with a checkmark and the text 'Running'.
- nErrors**: A table with one row showing the value '0'.