Metadata Extraction

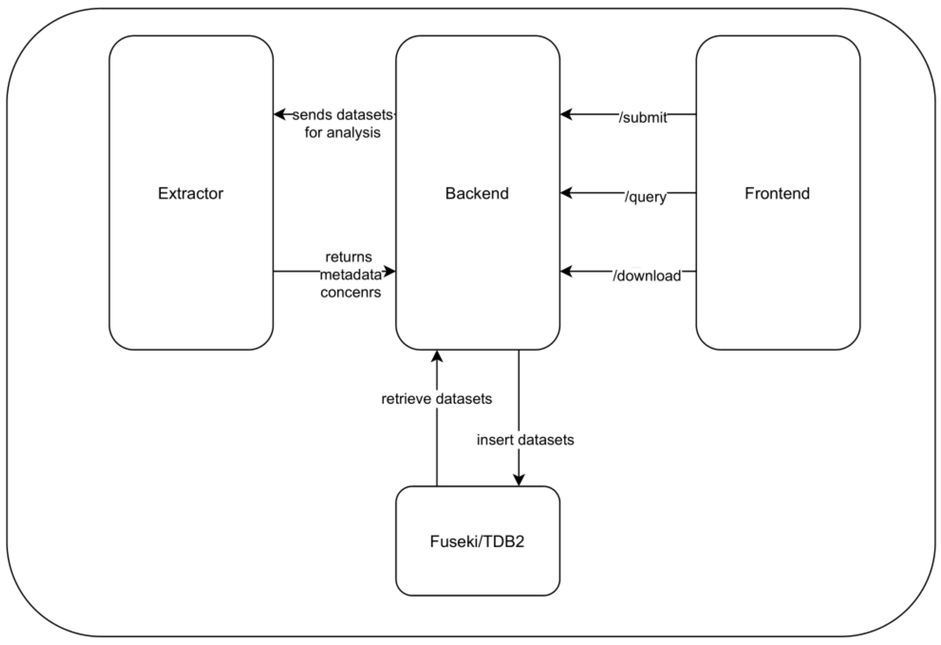
Fraunhofer FIT

Abhishek Nadgeri[1], Omar Ejjeh[2], Lei Wang[3], David Abdelsalam[4], Salman Tariq[5]

Computer Science 5 - Information Systems and Databases

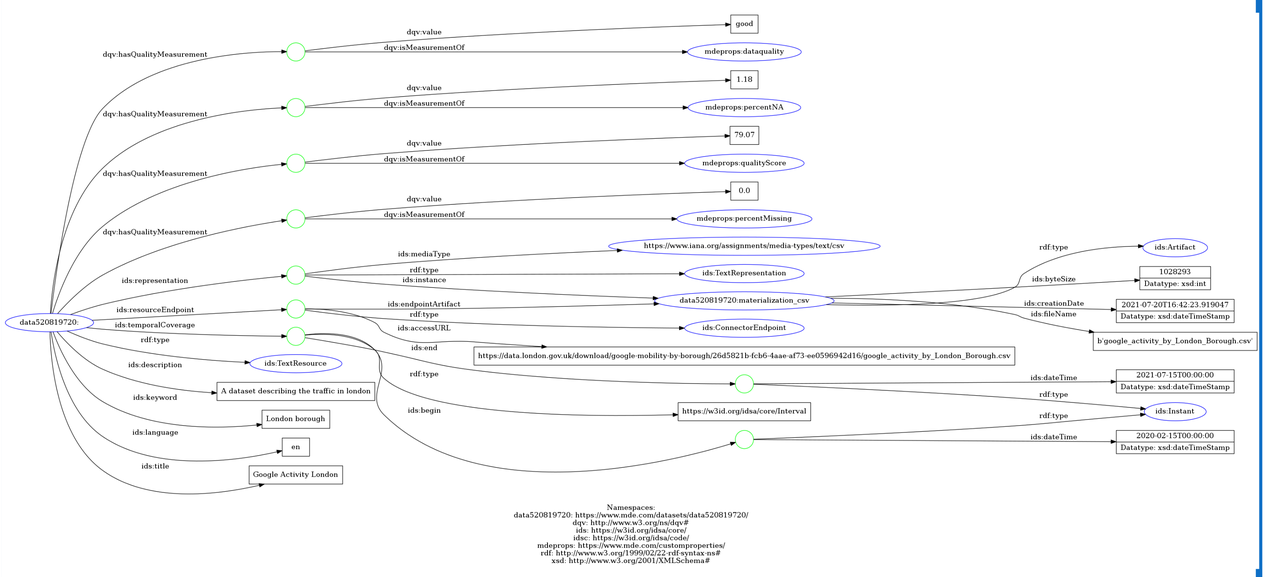
RWTH AACHEN UNIVERSITY

Metadata Extraction Overview[1,2]



- Extract metadata from CSV files
- Metadata is stored in RDF triplestore
- Users can query for relevant datasets based on their metadata
- Users can submit new datasets for extraction and integration into the database
- Users can download relevant metadata files from the frontend

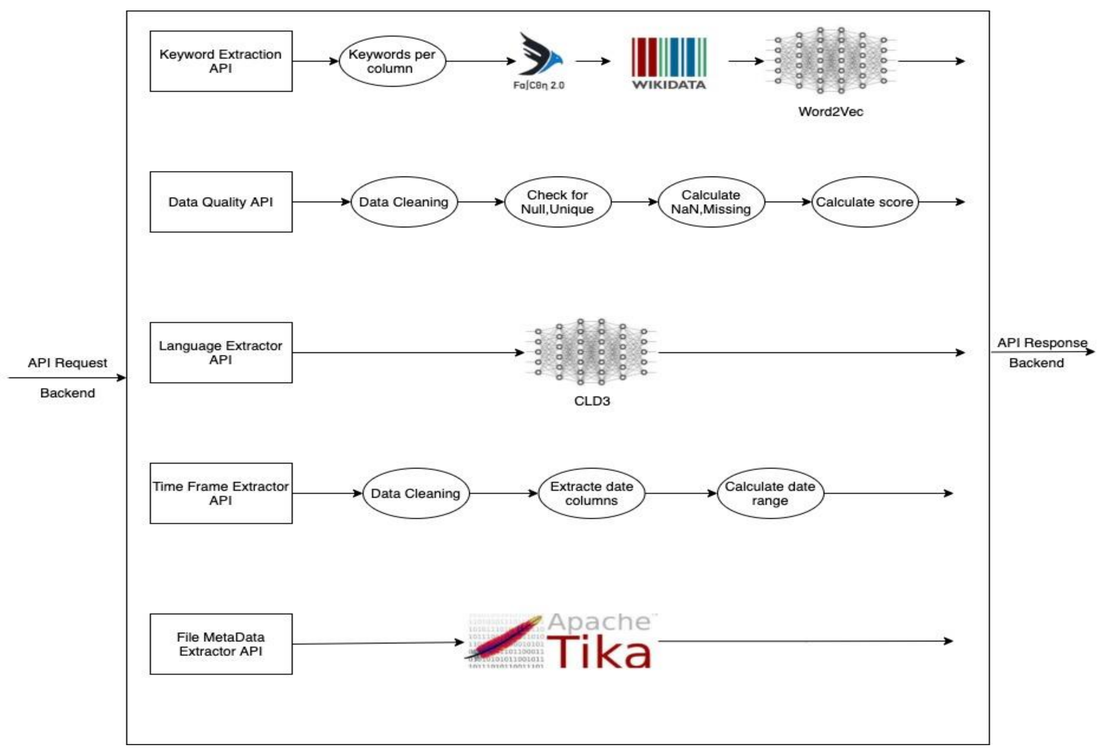
Metadata Schema[2]



- Uses IDS Information Model to describe datasets
- Datasets represented by resources of type ids:TextResource
- Included metadata concerns: ids:description, ids:title, ids:keyword, ids:language, ids:resourceEndpoint, ids:temporalCoverage, ids:Representation,
- New namespace mdeprops for custom dataquality properties
- Dataquality represented using dqv vocabulary

Extractor[1,3,4,5]

Overview of the extractor architecture



Extractor functionality

- **Keyword:**
Use word2vec and dbpedia lookup to find relevant keywords without further user input

- **Quality:**
Calculate measure of data quality based on missing values etc.

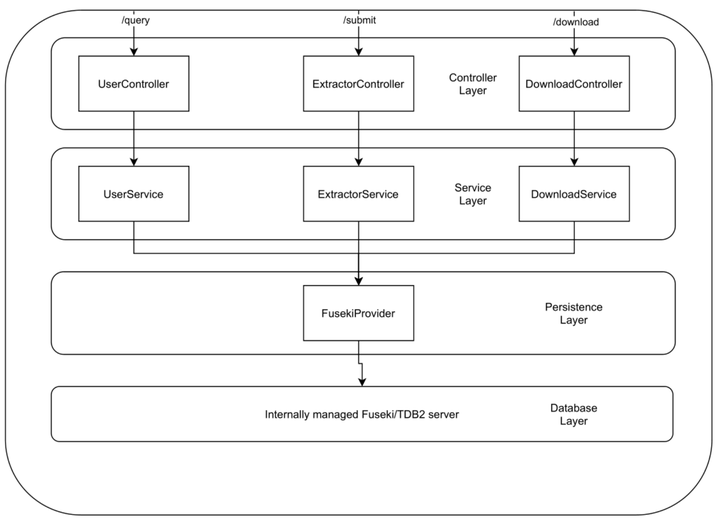
- **Language:**
Use CLD3 to discover the language of the dataset

- **Timeframe:**
For datasets with time column, calculate the timeframe

- **File metadata:**
Use tika to extract filesize, filetype etc.

Backend[2]

Overview of the backend architecture



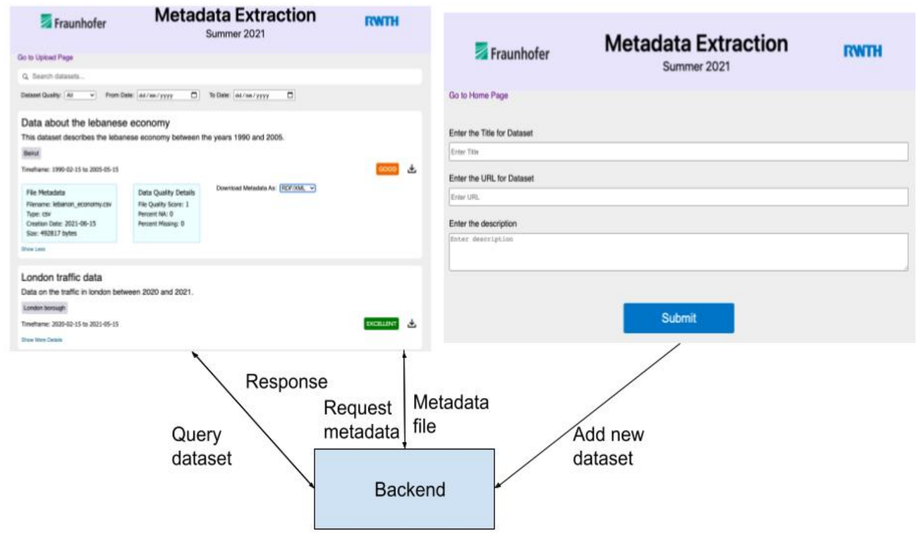
- **Controller:**
One controller per API

- **Service:**
One service per API

- **Persistence:**
FusekiProvider initializes and queries database

- **Database:**
Fuseki/TDB2 as database

Frontend[1]



- **Query page:** allows looking up datasets based on their Metadata and down-loading them

- **Admin page:** allows users to upload new datasets