

Documentation

Table of contents

[Home](#)

[About](#)

[Architecture](#)

[Documentation](#)

[Crawler](#)

[Database](#)

[Server](#)

[WebUI](#)

[Installation](#)

[Usage](#)

[FAQ](#)

About



Metadata-Hub

This application is developed in the course of the [AMOS](#) project.

Motivation

Big-Data environments often comprise large amounts of data that have been integrated from various sources. These can only be turned into valuable information through the use of metadata, which is significantly more lightweight, allowing for faster accessing and handling when compared to the actual data. Therefore, the aim of the Metadata-Hub is to provide a platform-independent retrieval, storage, and query mechanism for metadata on large file systems. This will allow end-user applications to obtain resilient and meaningful data as basis for complex data analyses in order to mine valuable information for the application-specific context.

Goals

Our mission is to create a first **prototype** of the Metadata-Hub, an independent piece of software that intelligently crawls large file systems in order to gain and store metadata about the files it finds. An intelligent algorithm continuously traverses the file system and collects interesting metadata. This metadata is stored in a designated metadata store, thereby generating an easy-to-access and persistent index of the whole file system. Finally, existing end-user applications will be able to query and consume the collected metadata.

Wiki

This wiki is the official documentation of the Metadata-Hub project. All included images are preview images that are linked to the larger original image. Simply click on the image to view it in higher resolution.

Architecture

This chapter provides a short overview about the conceptual architecture of the Metadata-Hub application. It aims to give a high-level understanding of the structure of the application rather than a detailed description of its components. For a more detailed documentation, please refer to the [Documentation](#) chapter.

The Metadata-Hub application consists of three major components that are listed below.

- **Crawler**

This component implements the tree walk algorithm which crawls the target directories/filesystems. The tree walk can be manually configured, e.g. input directories or rough CPU restrictions. It extracts metadata of the traversed files with the [ExifTool](#) and stores this data in the database. Furthermore, it stores state about its execution(s) such that the tree walk can be initialized with already traced data or being paused/continued.

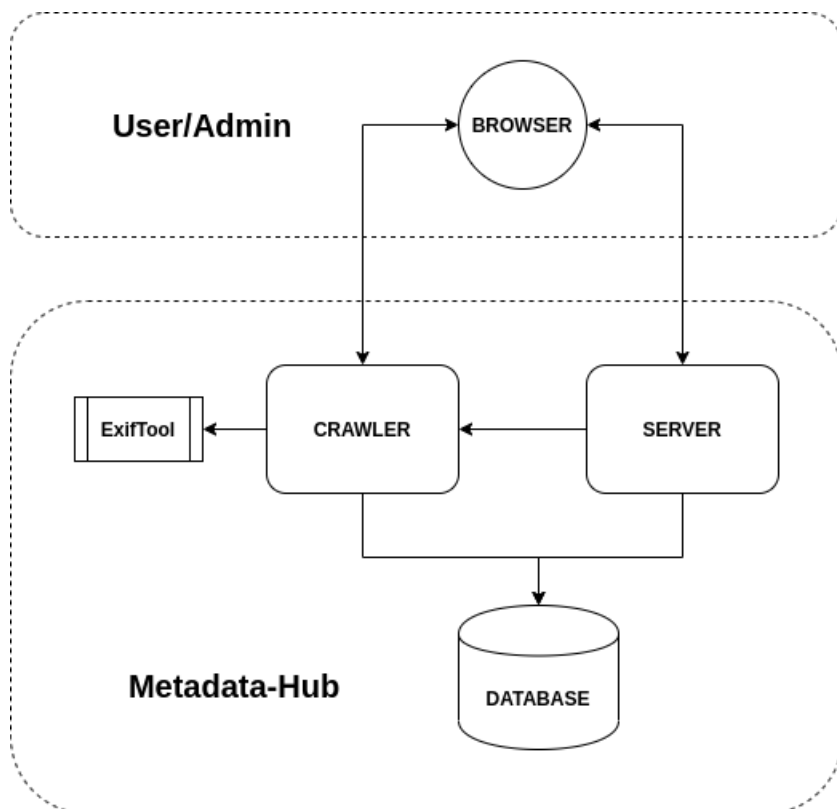
- **Server**

This component provides an interface for queries about the metadata. It uses [GraphQL](#) as the query language. Furthermore, it provides a web interface for the user to create queries without too detailed technical knowledge.

- **Database**

This component is responsible for persisting the metadata. The used database is [PostgreSQL](#). Furthermore, it is used to store state about executions of the tree walk for the controlling mechanism.

The interaction of these components is depicted in the following graphic and explained below.



Both the server and the crawler communicate with the database. Both the server and the crawler are accessible via the user's web browser. The database itself isn't required to be accessed directly by the user.

Documentation

This chapter provides a more detailed overview about the single components. It aims to give a deeper understanding of how they work and what functionality they provide. Though, this is no source code documentation. Therefore, please refer to the [GitHub repository](#).

Crawler

Tree Walk Interface

This section will shortly explain how to access the tree walk interface , as well as the different configuration options provided.

The interface can be accessed in a web browser, by using the address and port specified while starting the docker container. The following gif shows an example of the interface:

- [Interface access example](#)

The different options have the following meaning:

- **Input directories**

This field expects a path to a directory and a corresponding boolean separated by a comma. This value determines, if the tree walk is supposed to be executed recursively (If the value is set to true, all subdirectories will be scanned). You can also enter multiple pairs, by separating them with a semicolon.

An example input could look like this:

```
/home/metadata-hub, True; /home/user, false
```

- **Output directory**

This field is deprecated and is used for debugging purposes. Enter a directory path in the project folder.

- **Trace file**

This field is used to input a directory path to a trace file. This file will be used to store data on the execution. Visited and completed nodes in the tree walks traversal path will be saved, so future tree walk runs can exclude them. This can potentially save time, if you do not want to analyze directories twice.

An example input could look like this:

```
/home/metadata-hub/crawler/trace.log
```

- **ExifTool**

This field is used to choose between the linux and windows executables of the exiftool. Pick the operating system you are currently working on.

- **Clear trace data before start**

This field is used to determine, whether the trace.log should be used or not. If yes is picked, previously visited node of the directory tree will be skipped. If no is picked, the tree walk will traverse the entire directory tree.

- **Power level**

This field is used to determine how many system resources should be reserved for the tree walk execution. 1 provides the system with the minimum amount, 4 with the maximum,

- **Work package size**

This field represents the value of the desired work packages the crawler will scan. The algorithm attempts to evenly split the files of all directories into this size.

- **Update current execution**

This field determines, if the submission that is about to be sent to the crawler will replace the old one. If no is selected the crawler will not accept a new submission and prompt the user to wait.

API

This section will shortly explain the API of the crawler. All these endpoints support `GET` and `POST` requests except `/info` that only supports `GET` requests.

- **/config**

Create a configuration for the tree walk and start it. This endpoint provides the interface described above. After submitting the configuration, a success or error page is shown.

- **/start?config={CONFIG}&update=[True]**

Start the tree walk. This is the proper way to start the tree walk in a automated way.

- `CONFIG` (*required*) This is the configuration of the execution. It can either be a filepath pointing to a valid configuration file or a valid JSON configuration.
- `update` (*optional*) By providing `update=True`, a possible running execution will be stopped and the new one will be started. Without providing `update` or with `update=False`, the request will be ignored if a execution is running/paused.
- **/pause**
Pause a currently running execution of the tree walk. The request will be ignored when the tree walk is currently not running. The execution can be continued or stopped later on
- **/continue**
Continue a paused execution of the tree walk. The request will be ignored if the tree walk is not paused.
- **/stop**
Stop a running or paused execution of the tree walk. The request will be ignored if the tree walk is neither paused nor running.
- **/info**
Retrieve information about the current status of the tree walk. Useful to check the status and progress of a possible running execution.
- **/shutdown**
Shutdown the crawler completely. This will force a possible running execution of the tree walk to end and exit the crawler process.

Database

Database

All file metadata analyzed by the crawler gets inserted in a [PostgreSQL](#) Database. Also information about the crawls started by the crawler Component is stored in the database.

Database Setup

1. Create the Database Role or Import the Database (/metadatabasehub/metadatabasehub-role.sql)
2. Create the Database or Import the Database (/metadatabasehub/metadatabasehub-database.sql)
3. Import the Schemata

Default setup:

- "database-name": "metadatabasehub",
- "database-host": "localhost",
- "database-port": 5432,
- "database-user": "metadatabasehub",
- "database-password": "metadatabasehub",

If another setup wants to be used metadata-hub/configs need to be adapted.

Loading the Schemata:

psql command for Linux

```
psql metadatabasehub -Umetadatabasehub -W -f metadatabasehub-schemata.sql
```

psql command for Windows

```
metadata=> \i c:/dir/dir/metadatabasehub-schemata.sql
```

Database-Schemata

The database uses two tables.

- The "crawls" table saves information about started crawls/tree_walks.
 - The "files" table saves information about all analyzed file metadata, most metadata information lands in the jsonb-File "metadata".
- crawls**

id	dir_path	name	status	crawl_config	analyzed_dirs	starting_time	finished_time	update_time	analyzed_dirs_hash
[PK] bigint	text	text	text	text	jsonb	timestamp with time zone	timestamp with time zone	timestamp with time zone	text

```
id(bigint)                -> primary key (uses autoincrement sequence)
dir_path(text)             -> starting directory of the crawl
name(text)                 -> specified name of the crawler
status(text)               -> status of the crawl (running, finished, suspended, aborted, ...)
crawl_config(text)         -> latest used crawl_config by the crawl, only for user-presentation purposes
analyzed_dirs(jsonb)       -> array of currently analyzed dirs at "update_time"
starting_time(date)        -> start time of the crawler job
finished_time(date)        -> the end of the first crawler-job
update_time(date)          -> time of the latest update of the crawl data
analyzed_dirs_hash(text)   -> sha-256 hash of analyzed_dirs (uses trigger to create hash on inserting and updating)
```

files

id	crawl_id	dir_path	name	type	size	metadata	creation_time	access_time	modification_time	file_hash
[PK] bigint	bigint	text	text	text	bigint	jsonb	timestamp with time zone	timestamp with time zone	timestamp with time zone	text

<code>id(bigint)</code>	-> primary key (uses autoincrement sequence)
<code>crawl_id(bigint)</code>	-> foreign key " <code>crawls.id</code> "
<code>dir_path(text)</code>	-> absolute path of the file
<code>name(text)</code>	-> name of the file
<code>type(text)</code>	-> type of the file
<code>size(bigint)</code>	-> the size in bytes
<code>metadata(jsonb)</code>	-> metadata of the file
<code>creation_time(date)</code>	-> time the file was created
<code>access_time(date)</code>	-> time the file was last accessed
<code>modification_time(date)</code>	-> time the file was last modified
<code>file_hash(text)</code>	-> sha-256 hash of the file

Server

Server

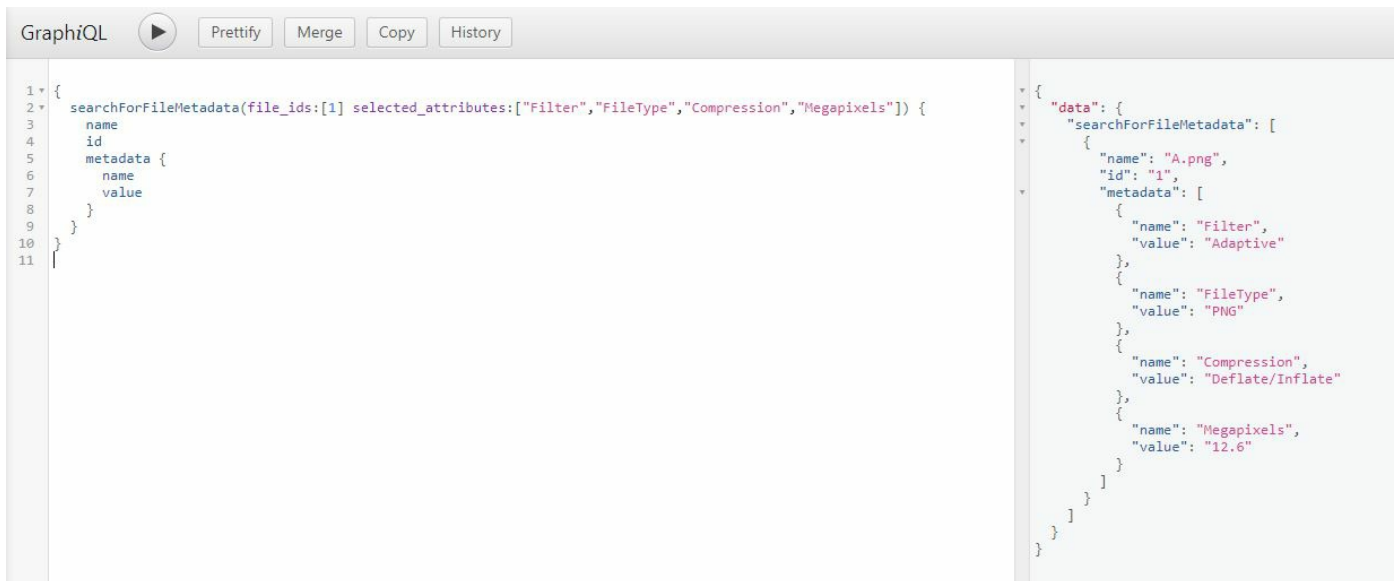
This component uses [GraphQL-Java](#) to query the postgresSQL database for file metadata information. User can directly interact with the API or use the Web User Interface to retrieve metadata information.

Application Programming Interface

The queries are sent to the http-server using the http POST method. Port and URI of the server are set by the configuration file (metdatahub/configs/environments.deployment.json). The default Port is 8080 and the default URI is localhost and the path to the http POST interface is "/graphql".

Using all that information we can now send GraphQL-queries to the server. When the server is running a GraphiQL test console can be found here: <http://localhost:8080/testconsole/> In the console we find information about the queries syntax and what it will retrieve. For the next two queries we will use the GraphiQL console to demonstrate GraphQL.

- A very simple GraphQL query looks like this:



The screenshot shows the GraphiQL web interface. On the left, a query is entered:

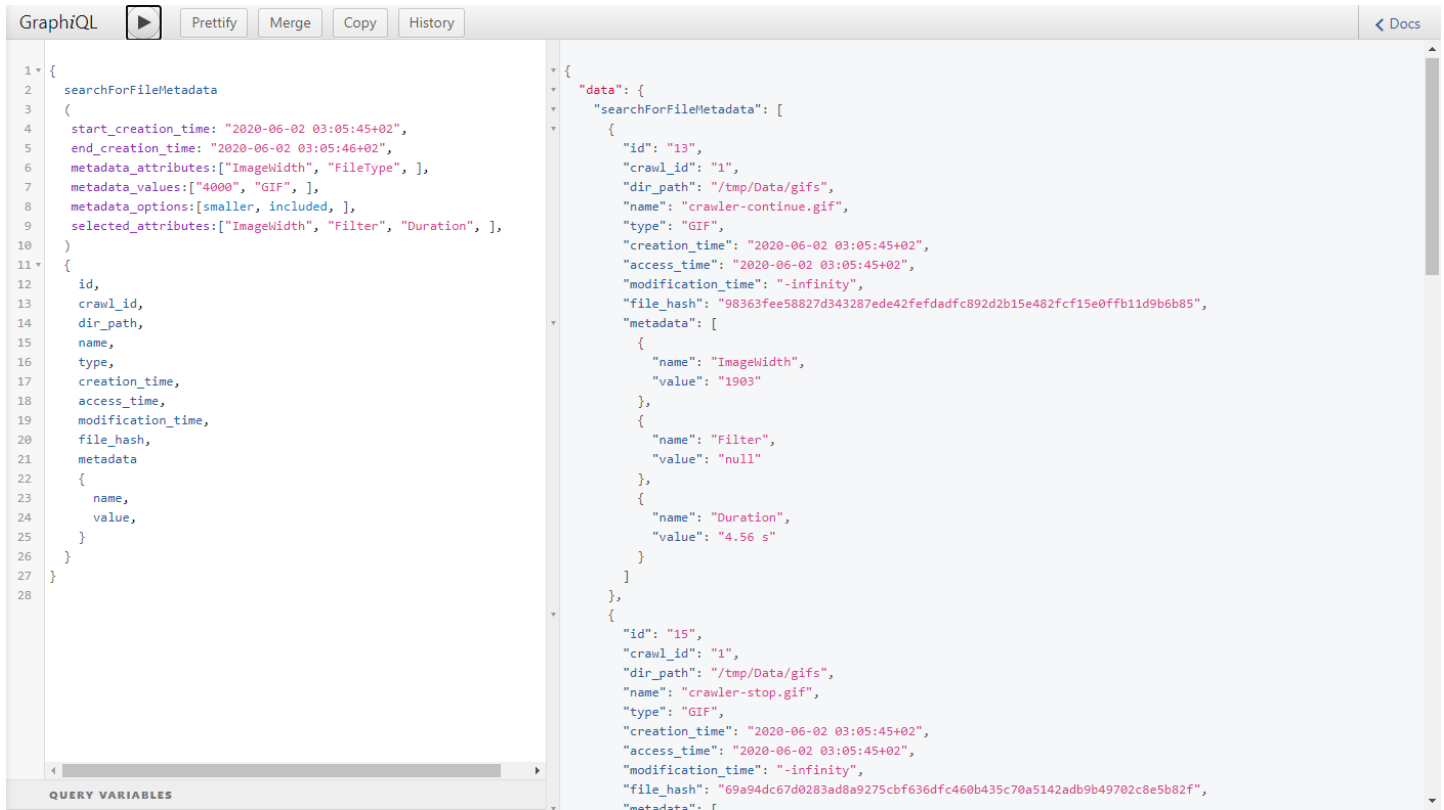
```
1 {
2   searchForFileMetadata(file_ids:[1] selected_attributes:["Filter","FileType","Compression","Megapixels"]) {
3     name
4     id
5     metadata {
6       name
7       value
8     }
9   }
10 }
11
```

 The interface includes buttons for 'Prettify', 'Merge', 'Copy', and 'History'. On the right, the JSON result is displayed:

```
{
  "data": {
    "searchForFileMetadata": [
      {
        "name": "A.png",
        "id": "1",
        "metadata": [
          {
            "name": "Filter",
            "value": "Adaptive"
          },
          {
            "name": "FileType",
            "value": "PNG"
          },
          {
            "name": "Compression",
            "value": "Deflate/Inflate"
          },
          {
            "name": "Megapixels",
            "value": "12.6"
          }
        ]
      }
    ]
  }
}
```

On the left side we can see the query and on the right the result. The Query `"searchForFileMetadata()"` has two Arguments, a list of file ids and a list of metadata attributes we want to have returned. In the curly brackets there are even more fields, like `name`, `id`, `metadata`, they are fields of the object `"File"` and they define how the result will look like. `"metadata"` is a field of `"File"` but its a object on its own so we have to define again in which fields of metadata we are interested in, in our example that is the `name` and value field.

A more complicated GraphQL query looks like this:



Now our new query has 6 arguments.

The **values** we have **set** for **start_creation_time** and **end_creation_time** **only** return files, which were exactly created at **"2020-06-02 03:05:45"**.

The three lists **metadata_attributes**, **metadata_values** and **metadata_options** are related to each other.

In this example we look for all Files that have a metadata attribute of **"ImageWidth"** that has a value which is smaller than **"4000"**. And all Files that have a **"FileType"** attribute which have the pattern **"GIF"** included in their value.

The list of **selected_attributes** specifies that we are only interested in the metadata attributes **"ImageWidth"**, **"Filter"** and **"Duration"**.

Those are the only ones that are returned to us by the query.

The **searchForFileMetadata** query returns an **Object Type** named **"File"** and now we have specified that the query returns all the fields of **"File"**.

`("id", "crawl_id", "dir_path", "name", "type", "creation_time", ... , "metadata")`

GraphQL Schema

Now we will take a closer look at the Graphl Schema which defines the syntax of our query. Only the most important information is displayed here. The exact schema can be found in `/metadata-hub/server/src/main/resources/schema.graphqls`

File

```
type File{
  id: String!
  crawl_id: String!
  dir_path: String!
  name: String!
  type: String!
  metadata: [Metadatum!]!
  creation_time: String
  access_time: String
  modification_time: String
  file_hash: String
}
```

"File" is our object type which gets returned by **searchForFileMeta** and represent our files in the server. It has different fields which are also represented in some way in the database. Most metadata will be in the field **"metadata"**, which is a list of **"Metadatum"** Object Types. The **"Metadatum"** Object Type can be seen directly below.

Metadatum

```
type Metadatum{
  name: String!
  value: String!
}
```

"Metadatum" is the object type that represents one metadatum of a file.

searchForFileMetadata()

```
searchForFileMetadata(file_ids: [Int!], crawl_ids: [Int!], dir_path: String, dir_path_option: MetadataOption,
  file_name: String, file_name_option: MetadataOption, file_type: String, size: Int, size_option: IntOption,
  start_creation_time: String, end_creation_time: String, start_access_time: String, end_access_time: String,
  start_modification_time: String, end_modification_time: String, file_hashes: [String!],
  metadata_attributes: [String!], metadata_values:[String!], metadata_options: [MetadataOption!],
  selected_attributes: [String!], limitFetchingSize: Int) : [File]
```

"searchForFileMetadata()" is our only GraphQL query and it offers a lot of options to specify which file metadata we want to have returned.

searchForFileMetadata:

Searches **for** all **file** metadata dependent **on the** specified search options.
If no options are specified, all **file** metadata **is** returned.

Search Options:

file_id: Only returns **file** metadata belonging **to one of the file** ids **in the list**.

crawl_id: Only returns **file** metadata belonging **to one of the crawl** ids **in the list**.

dir_path: Only returns **file** metadata **where** their directory path matches **the** specified pattern.

Default PatternOption is **"included"**.

dir_path_option: Different PatternOptions can be used, which change how **"dir_path"** gets compared **to** other Strings.

file_name: Only returns **file** metadata **where** their file_name matches **the** specified pattern.

Default PatternOption is **"included"**

file_name_option: Different PatternOptions can be used, which change how **"file_name"** gets compared **to** other Strings.

file_type: Only returns **file** metadata **where the file** type is exactly **"file_type"**

size: Only returns **file** metadata **of** a certain size depending **on the** used IntOption. Default IntOption is **"equal"**.

size_option: Here **the** IntOption **for "size"** can be specified.

!Notice to timestamps!: timestamps follow **the** ISO **8601** standard, e.g. **"2004-10-19 10:23:54+02"**, if no timezone is specified **it is** assumed **to be in the** system's timezone, if no **time** is specified **"00:00:00"** is used.

start_creation_time: Only returns **file** metadata, which got created later **or at the** same point **as "start_creation_time"**.

end_creation_time: Only returns **file** metadata, which got created earlier than **"end_creation_time"**.

start_access_time: Only returns **file** metadata, which got accessed later **or at the** same point **as "start_access_time"**.

end_access_time: Only returns **file** metadata, which got accessed earlier than **"end_access_time"**.

start_modification_time: Only returns **file** metadata, which got modified later **or at the** same point **as "start_modification_time"**.

end_modification_time: Only returns **file** metadata, which got modified earlier than **"end_modification_time"**.

file_hash: Only returns **file** metadata **where the file** has **the** same sha_254 hash **as the** hashes **in the list**.

metadata_attributes: (see **in** metadata_values)

metadata_values: Only returns **file** metadata, **where the file** has **the** attribute specified **in**

metadata_attributes **and its** value matches **the** value **of** metadata_values. Different PatternOptions can be used.

Both lists have **to be the** same **length as** their ordering relates them **to** each other.

Default PatternOption is **"included"**.

metadata_options: Different PatternOptions can be used **for** metadata_values. The index **in** metadata_option relates **to the** index

in meta_data_values e.g. metadata_option[1] will be used **for** metadata_value[1]. If used both lists need **to have the** same **length**.

selected_attributes: For all **file** metadata only **the** specified metadata attributes are returned.

limitFetchingSize: Limits how many files will **get** fetched **by the** search.

Web User Interface

Using the Web UI a User can send queries in GraphQL syntax, but it also provides a form query, where different search options can be used without knowing GraphQL syntax. More information about the Web UI can be found in its own documentation page.

WebUI

Form-Query

Here you can use a special form to ask for the file-data. You can use there advanced filters and limit the json-output to metadata which you are interested in.

The following Inputs are possible:

Form Query Editor

Query-Name [\[?\]](#)

My Testquery

Owner [\[?\]](#)

Robin

Start-DateTime (File created)[\[?\]](#)

2020-05-27

End-DateTime (File created)[\[?\]](#)

2020-05-28

Start-DateTime (File modified)[\[?\]](#)

2020-05-22 07:10:20

End-DateTime (File modified)[\[?\]](#)

2020-07-28 20:35:22

Limit [\[?\]](#)

7

Limit

The max output limit.
Empty means no limit.

Filter: [\[?\]](#)

Pattern included [↕](#)

FileName

dog

Pattern excluded [↕](#)

FileName

rex

Exists (Attribute) [↕](#)

Metadata-Attribute

Value

Which Attributes: [\[?\]](#)

FileName

FileSize

Send

Open Query

Send to GraphQL

Clear All

- **Query-Name:** A name for the query, to find this executed query later easier
- **Owner:** A name of the owner, which is saved along the name, to see who executed the query

The following fields are for limiting the resultset related to different datetimes, the files have

- **Start-DateTime (File created)**
- **End-DateTime (File created)**
- **Start-DateTime (File modified)**
- **End-DateTime (File modified)**
- **Limit:** The result contains at maximum limit entries. No input means all entries.

Filter (List):

- **Filter-Function:** The filter function which take place
 - **Pattern included:** The attribut-value must contain the value
 - **Pattern excluded:** The attribut-value may not contain the value
 - **Equal:** The attribut-value must be exactly the value
 - **Exists (Attribute):** The attribut have to exists.
 - **Greater Than:** The attribut-value must be greater than value
 - **Lower Than:** The attribut-value must be lower than value

[See an Example from the filter](#)

Which Attributes (List):

- **Attribute:** *Here you can insert the Name of an attribute*

[See an Example from the attribute-selection](#)

If you press the send-button the form-data are converted to a *GraphQL-Query* in the background. It is send and you get the result as json. You also can introspect the *GraphQL-Query*, send the query to the *GraphiQL-Console* for better formatting and changing the query with the help of the *GraphiQL-console*. You also have the possibility to clear the whole form.

Hash Query

Here you can enter a *file-hash*. If it is in the database, you get all informations about that file. You get all meta-data. Also you get some informations about the *path* and the *crawl*, which scanned the file.

GraphQL-Query

Here you can enter *GraphQL-Code* and execute it. The result is in json.

GraphiQL

GraphiQL-Console

Here you can use the *GraphiQL-Console*. It's a editor, where you can enter your GraphQL-Code. Its highlighted. More options you have there:

- **Get more informations about our used graphql-Endpoint in the Docu.** (You have to click to the Docu-Button)
- **You can re-execute old queries. They are cached locally on your computer.**
- **Prettify your entered graphql-code.**
- **Execute the graphql-code and get the result as json.**

More informations about GraphQL and GraphiQL you can find in the [server-section](#).

Installation

This chapter will show you how to install the software requirements and the Metadata-Hub application. It is important to mention that this is **no** application that should be used in a production environment because of predefined user/password settings that cannot be changed.

Requirements

The application is published as a Docker image. Thus, it requires your system to have the Docker Engine installed. Therefore, please refer to the official installation instructions of Docker at <https://docs.docker.com/get-docker>. Please make sure to install at least version *19.03* and check the installation before continuing.

Installation

The image is published using the DockerHub registry at [amosproject2/metadatabase](https://hub.docker.com/r/amosproject2/metadatabase). There are two versions of the application you can use:

- `latest` The latest stable version, usually updated once a week
- `dev` The currently developed version, might be unstable

The `latest` version is the recommended one to use, thus

```
$ docker pull amosproject2/metadatabase
```

will pull the image tagged with `latest` by default.

Configuration

The image is build with a default configuration that specifies some mandatory settings that cannot be changed, such as the default database user and port settings inside the container (see more at [Usage](#)).

The storage of the database will kept inside the container by default. Indeed, it can be useful to store this data on the host system to access it later on. Therefore, simply create a Docker volume that is used to store the content of the database.

```
$ docker volume create --name metadatabase-database -d local
```

It is also possible to use multiple volumes to have separate 'databases' for the various container.

Usage

After setting everything up, you are ready to run the image. Therefore, use the following template.

```
docker run \
  -p {HOST_SERVER_PORT}:8080 \
  -p {HOST_CRAWLER_PORT}:9000 \
  -v {DATA}:/filesystem \
  -v metadatabase-database:/var/lib/postgresql/12/main \
  amosproject2/metadatabase
```

The following values have to be specified by the user:

- `HOST_SERVER_PORT`
The port that publishes the *server* with the graphical user interface for data querying on the *host* machine.
- `HOST_CRAWLER_PORT`
The port that publishes the *crawler* for starting/stopping/etc. the crawling mechanism on the *host* machine.
- `DATA`
The directory/filesystem you want to crawl.

The ports *8080* and *9000* must not change thus they are required for internal communication. For example, you can run the image with the following command.

```
docker run \  
-p 9999:8080 \  
-p 9998:9000 \  
-v /home/john/data:/filesystem \  
-v metadatahub-database:/var/lib/postgresql/12/main \  
amosproject2/metadatahub
```

You should be able to access both *localhost:9999* and *localhost:9998* for the corresponding services.

If you encounter any errors, please refer to the [FAQ](#) section.

Usage

Each entry is linked to a GIF that shows the specified functionality.

Crawler

- [How to start the crawler with the web interface](#)
- [How to start the crawler with the REST interface](#)
- [How to get status information](#)
- [How to stop the crawler](#)
- [How to pause the crawler](#)
- [How to continue a paused execution of the crawler](#)

FAQ

I have problems with docker permissions using Linux.

It is most likely your user does not belong to the `docker` group. Please have a look at these [instructions](#).

I have problems with a running container and want to inspect them.

You can start a shell session in your running container:

```
docker exec -it {container-id} /bin/bash
```

This will start a *bash* session inside the container with the ID `container-id`. Furthermore, you can inspect the log files at `/metadatabus/server.log` and `/metadatabus/crawler.log`.