Structured & Unit Independent Search for DBRepo

Martin Weise, Sotirios Tsepelakis, Nikola Lukic,
Max Spannring, Gökay Güçlü, Geoffrey Karnbach

Motivation for a Database Repository

Databases as important resources for research & industry

1. Database paradigm is well-understood
2. Cost-efficient storage systems for data in use
3. Repositories as established systems to make research data FAIR

Devise a system that combines technological infrastructure with repository work-processes to provide machine-understandable data in databases.

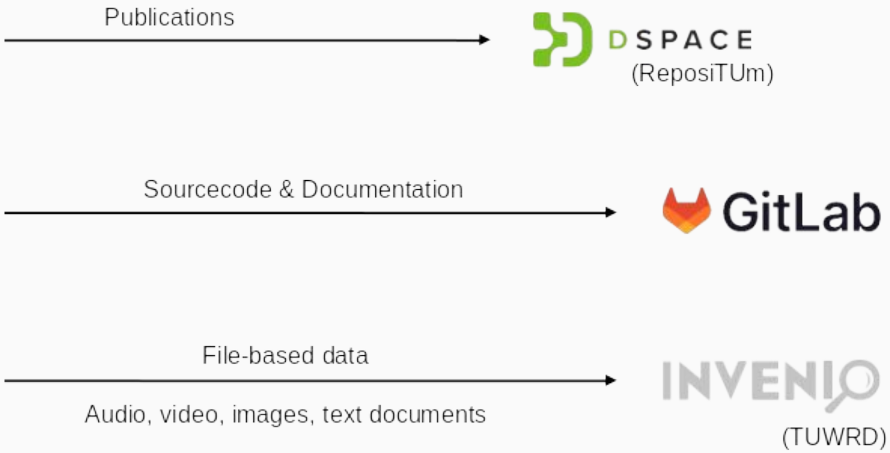
CAVEAT: RDA WGDC recommendations https://doi.org/10.1162/99608f92.be565013 & technical knowledge gaps

Motivation for a Database Repository

TUWRD can handle collection of files

How about relational data in databases?

- Releasing a **data dump** every x amount of time?
- Adding **continuous data** streams, e.g. IoT?
- How to update / correct data in those databases?
- Allow **reproduction** of any subset?

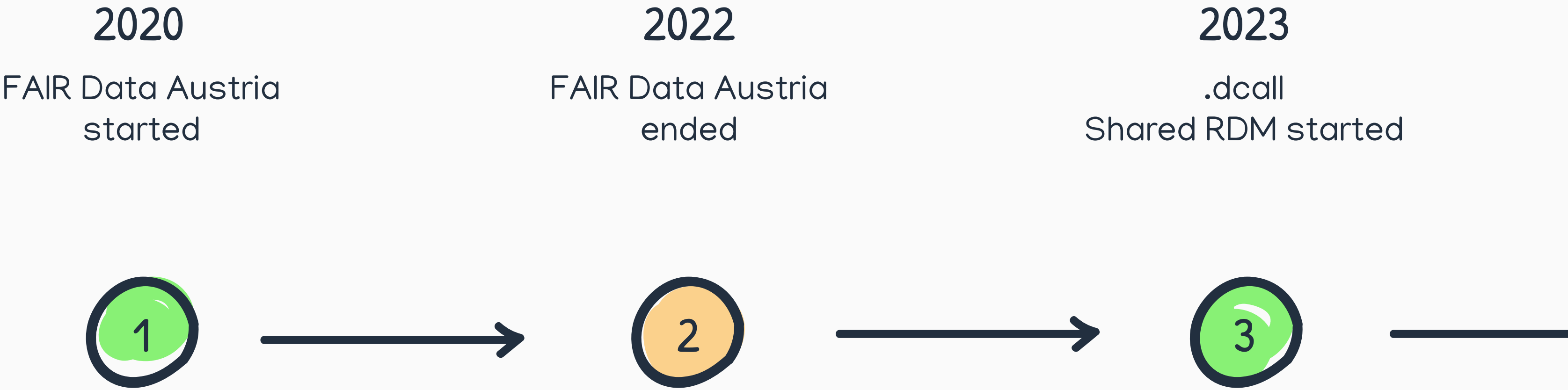


Publications → DSPACE (ReposiTUm)

Sourcecode & Documentation → GitLab

File-based data → INVENIO (TUWRD)

Audio, video, images, text documents



2020

FAIR Data Austria
started

1

2022

FAIR Data Austria
ended

2

2023

.dcall
Shared RDM started

3

PREVIOUS ISSUES

Wildcard search not accurate enough

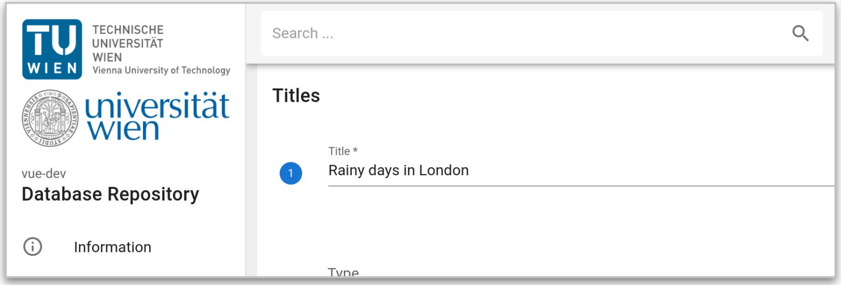
ElasticSearch license change 2021

Authentication directly to the database in the UI (unsafe!)

Does not scale

Not enough data for meaningful queries

Old



Title *
Rainy days in London

… among many others

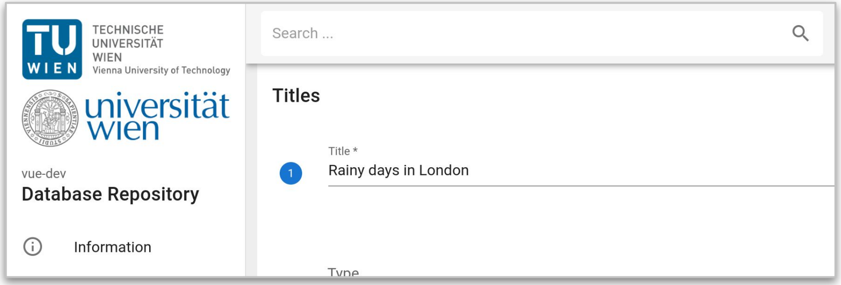
WP1 - Extension of Indexed Metadata

GOALS

Extend the indexed metadata in the search service to cover semantic concepts and units of measurement of columns for tables

Allow structured search through facets that assist users in filtering results based on semantic concept and/or unit of measurement

Old



TECHNISCHE UNIVERSITÄT WIEN Vienna University of Technology

universität wien

vue-dev

Database Repository

Information

Search ...

Titles

Title *

1 Rainy days in London

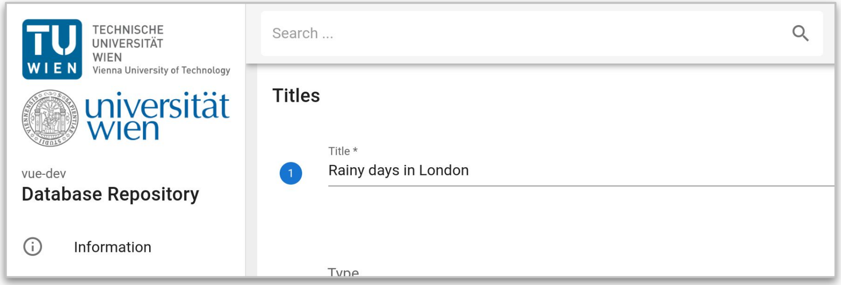
Type

WP1 - Extension of Indexed Metadata

DELIVERIES

1. Faceted browsing based on semantic concepts
2. Faceted browsing based on units of measurement
3. Increase supported ontologies, e.g. I-ADOPT

Old



Title *
Rainy days in London

Refactor ElasticSearch components

/database

/table

/column

/user

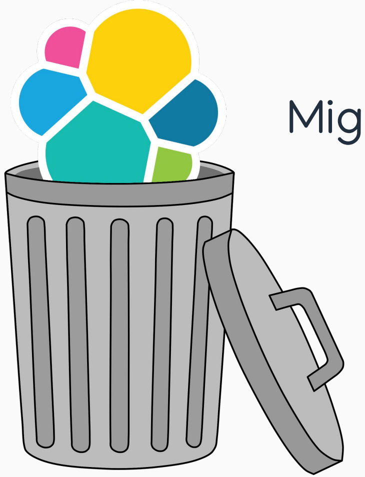
/view

/identifier

/concept

/user

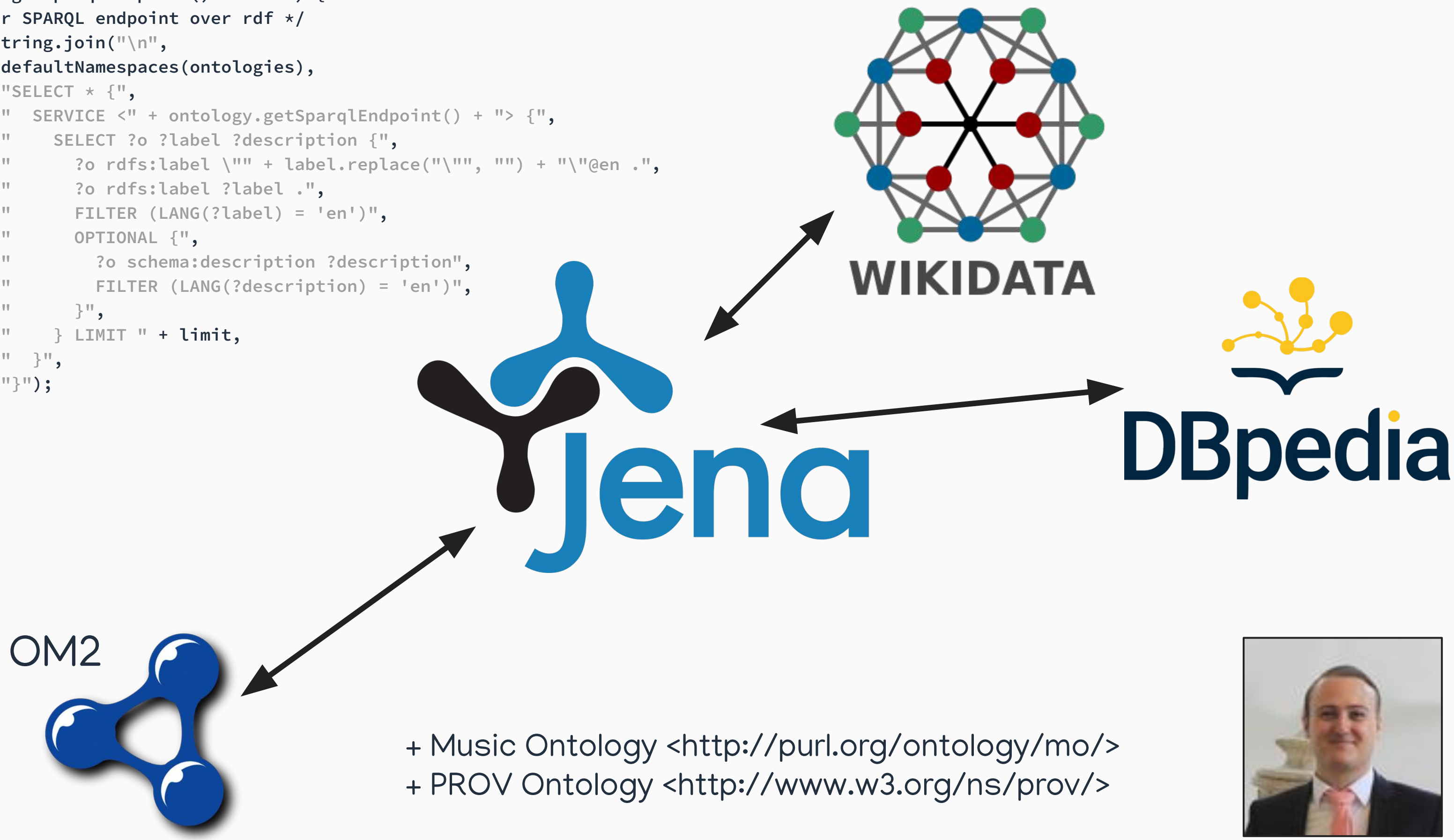
OpenSearch



Migrate queries

Martin Weise

```
default String ontologyToFindByLabelQuery(List<Ontology> ontologies, Ontology ontology, String label, Integer limit) {
        if (ontology.getSparqlEndpoint() != null) {
            /* prefer SPARQL endpoint over rdf */
            return String.join("\n",
                    defaultNamespaces(ontologies),
                    "SELECT * {",
                    "  SERVICE <" + ontology.getSparqlEndpoint() + "> {",
                    "    SELECT ?o ?label ?description {",
                    "      ?o rdfs:label \"" + label.replace("\"", "") + "\"@en .",
                    "      ?o rdfs:label ?label .",
                    "      FILTER (LANG(?label) = 'en')",
                    "      OPTIONAL {",
                    "        ?o schema:description ?description",
                    "        FILTER (LANG(?description) = 'en')",
                    "      }",
                    "    } LIMIT " + limit,
                    "  }",
                    "}");
        }
    }
    ...
}
```

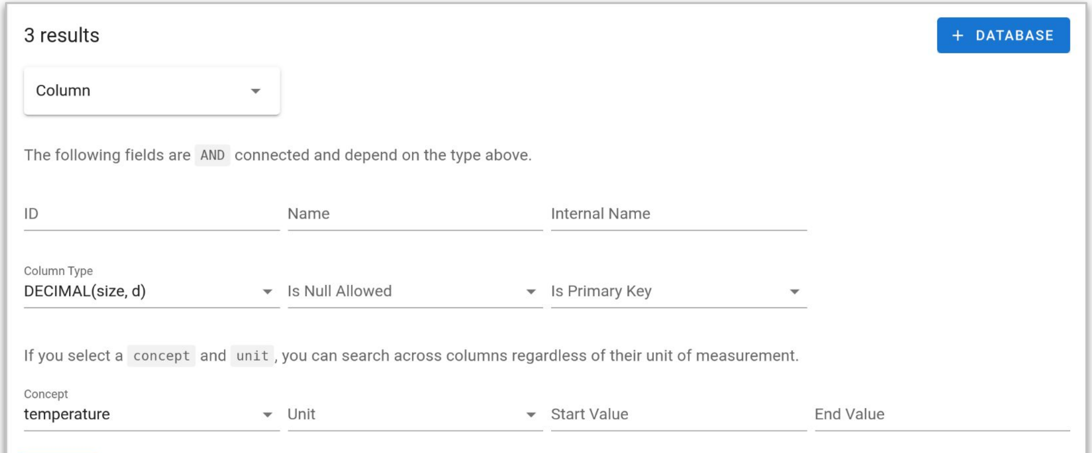


+ Music Ontology <http://purl.org/ontology/mo/>
+ PROV Ontology <http://www.w3.org/ns/prov/>

Martin Weise

NEW UI

NEW SERVICE



Column

The following fields are `AND` connected and depend on the type above.

Column Type: DECIMAL(size, d)

If you select a `concept` and `unit`, you can search across columns regardless of their unit of measurement.

Concept: temperature

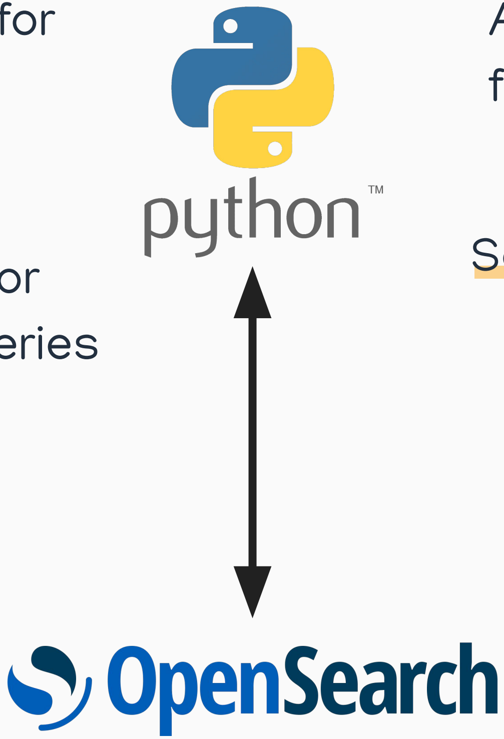
python™

Sotirios Tsepelakis

Geoffrey Karnbach

OpenSearch client for structured search

Authentication hidden from UI



Enough data for meaningful queries

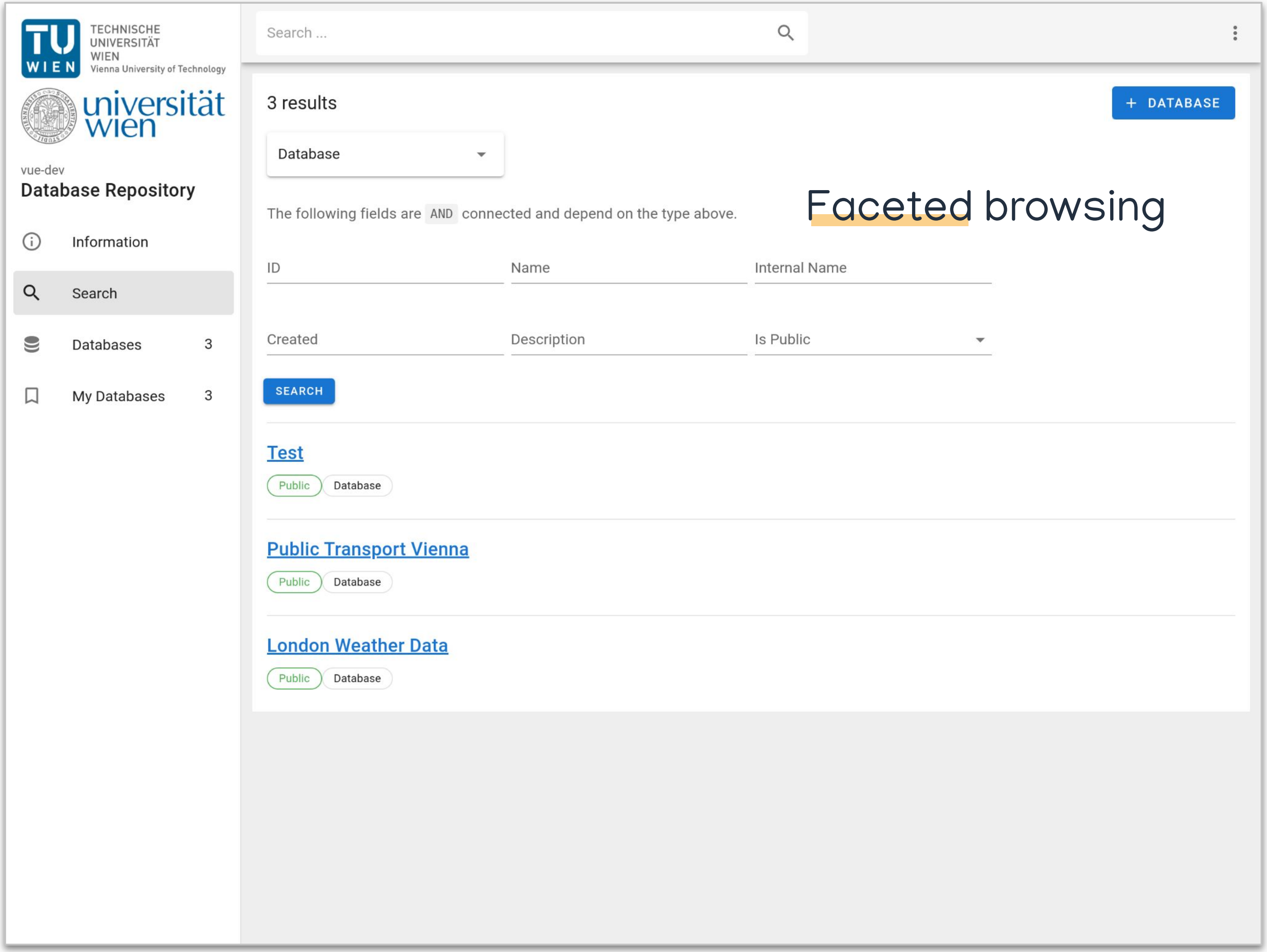
Scales well

Scales well

Indexed metadata

Geoffrey Karnbach

Faceted browsing



vue-dev

Database Repository

3 results

Database

The following fields are `AND` connected and depend on the type above.

Test

Public Database

Public Transport Vienna

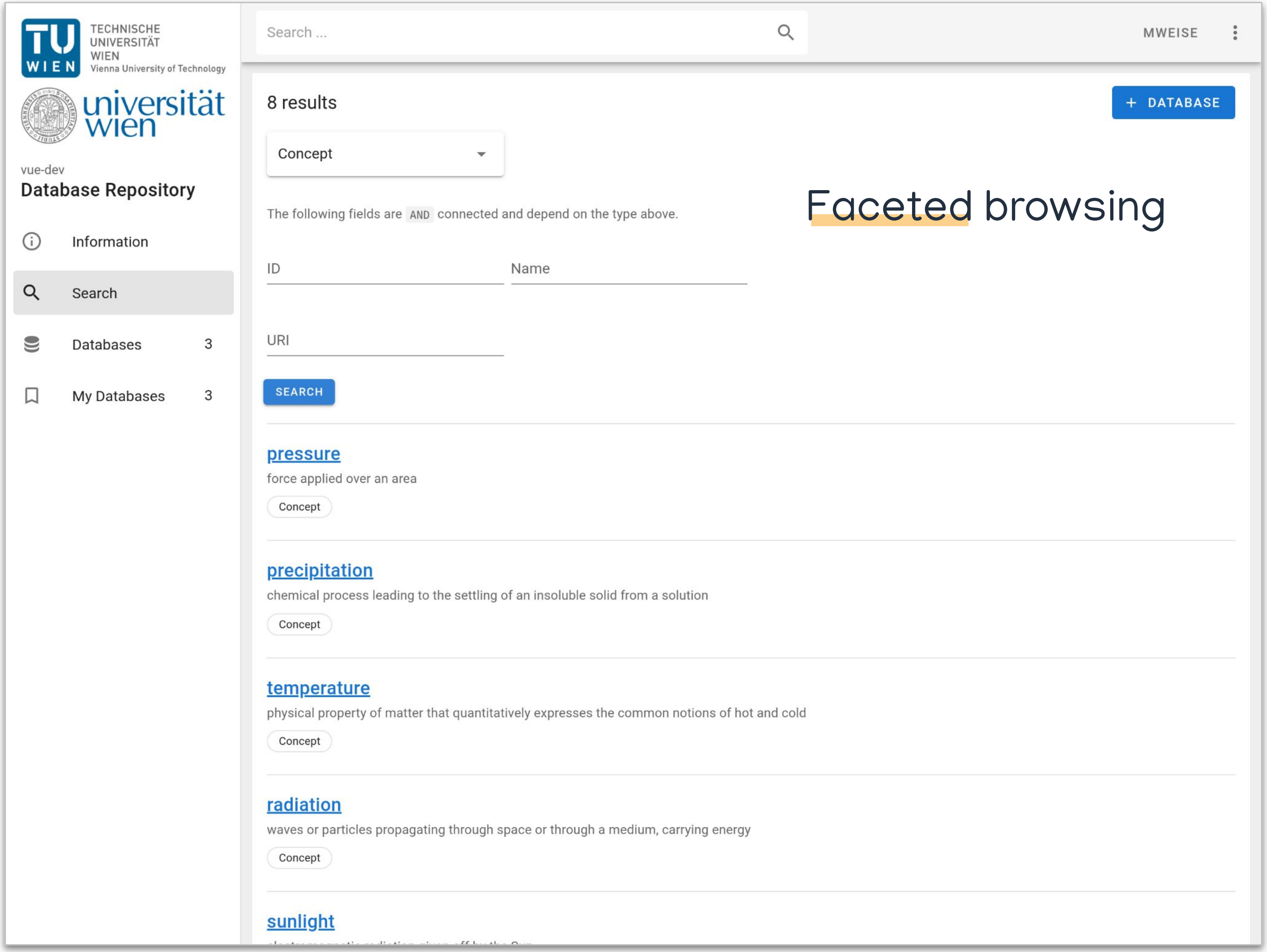
Public Database

London Weather Data

Public Database

Sotirios Tsepelakis

Faceted browsing



TECHNISCHE UNIVERSITÄT WIEN Vienna University of Technology

universität wien

vue-dev

Database Repository

Information

Search

Databases 3

My Databases 3

Search ...

MWEISE

8 results

+ DATABASE

Concept

The following fields are `AND` connected and depend on the type above.

ID

Name

URI

SEARCH

[pressure]

force applied over an area

Concept

[precipitation]

chemical process leading to the settling of an insoluble solid from a solution

Concept

[temperature]

physical property of matter that quantitatively expresses the common notions of hot and cold

Concept

[radiation]

waves or particles propagating through space or through a medium, carrying energy

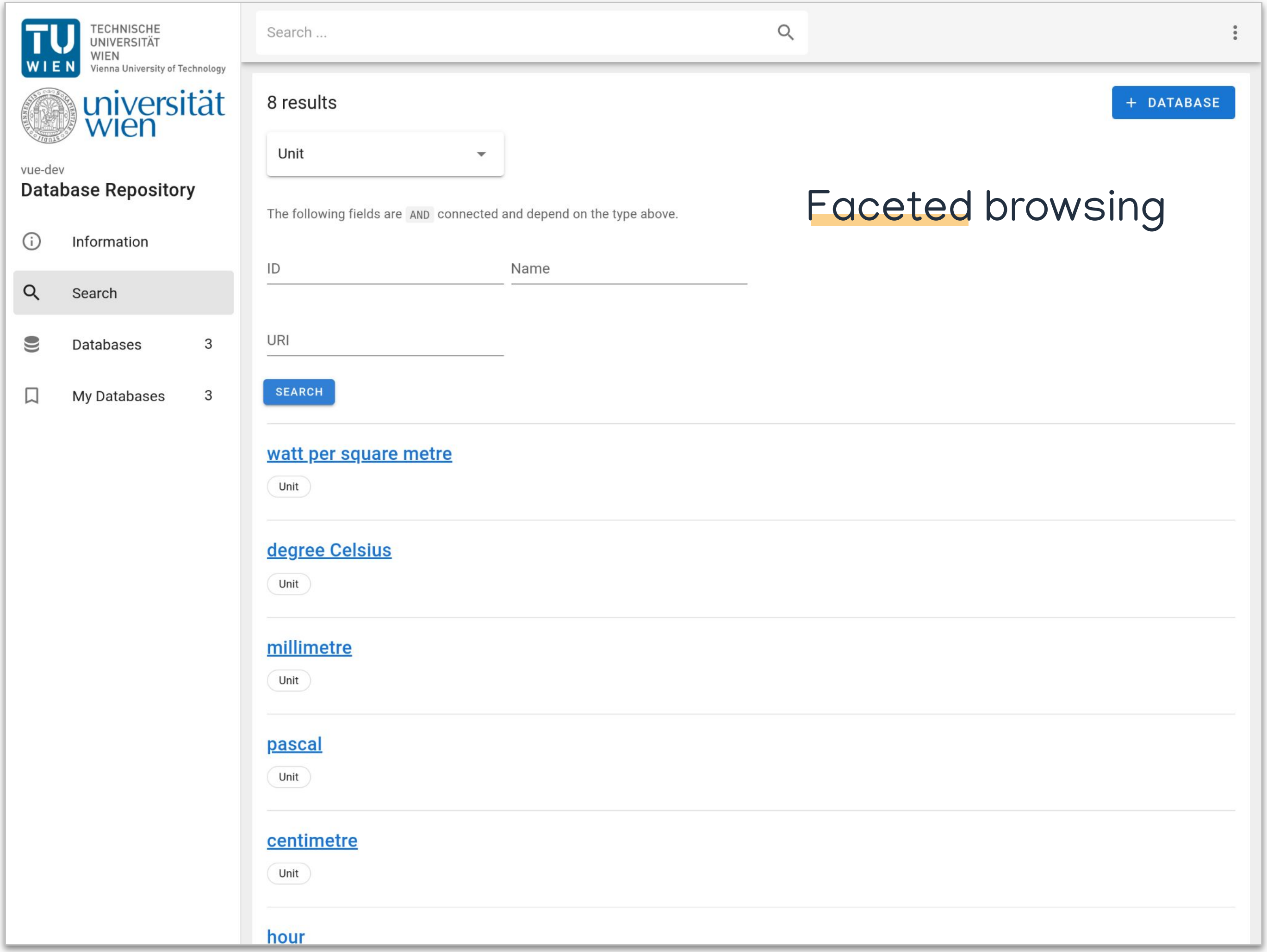
Concept

[sunlight]

electromagnetic radiation given off by the Sun

Sotirios Tsepelakis

Faceted browsing



8 results

Unit

The following fields are `AND` connected and depend on the type above.

ID

Name

URI

SEARCH

watt per square metre

Unit

degree Celsius

Unit

millimetre

Unit

pascal

Unit

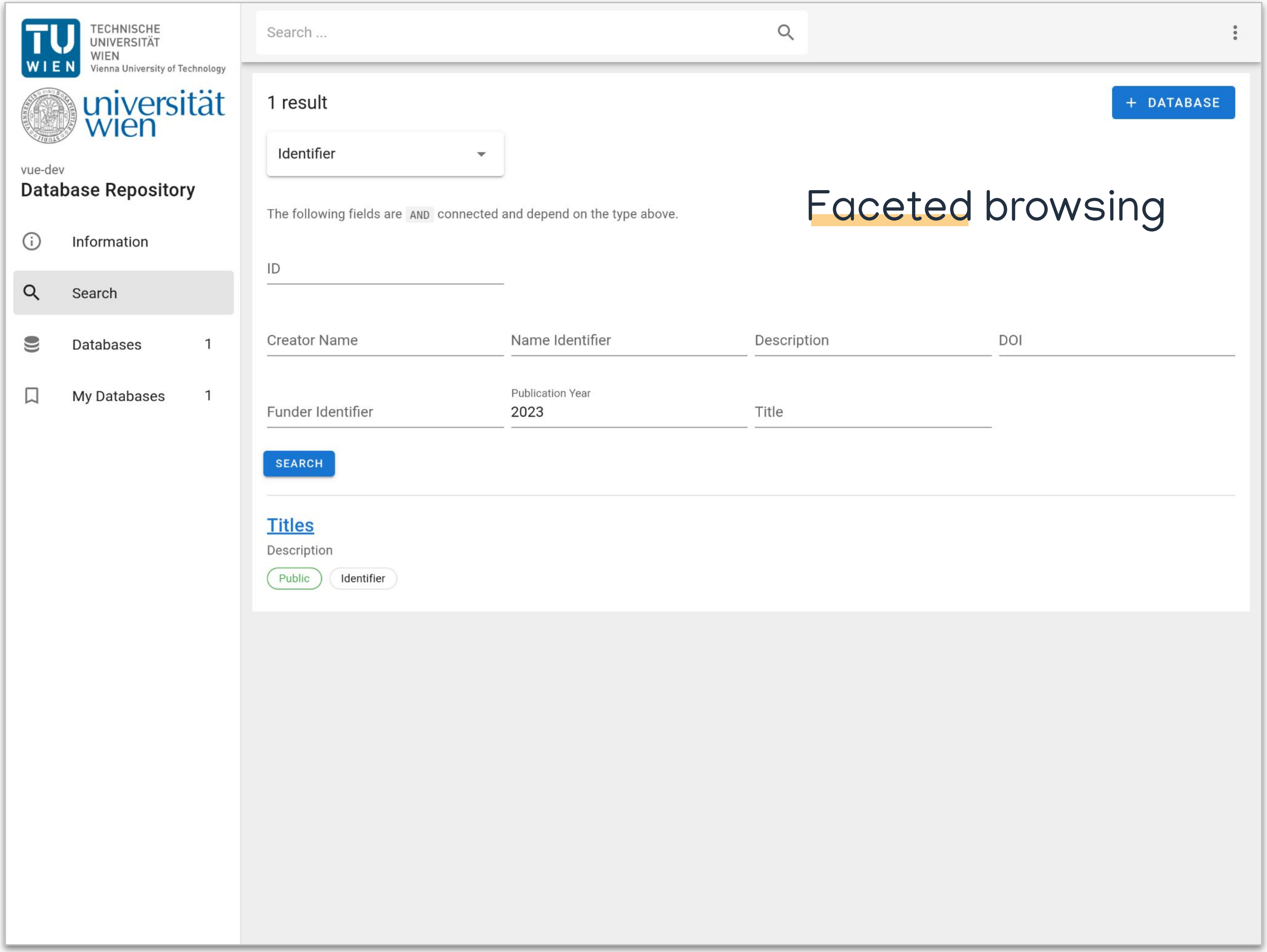
centimetre

Unit

hour

Sotirios Tsepelakis

Faceted browsing



1 result

Identifier

The following fields are `AND` connected and depend on the type above.

ID

Creator Name

Name Identifier

Description

DOI

Funder Identifier

Publication Year: 2023

Title

SEARCH

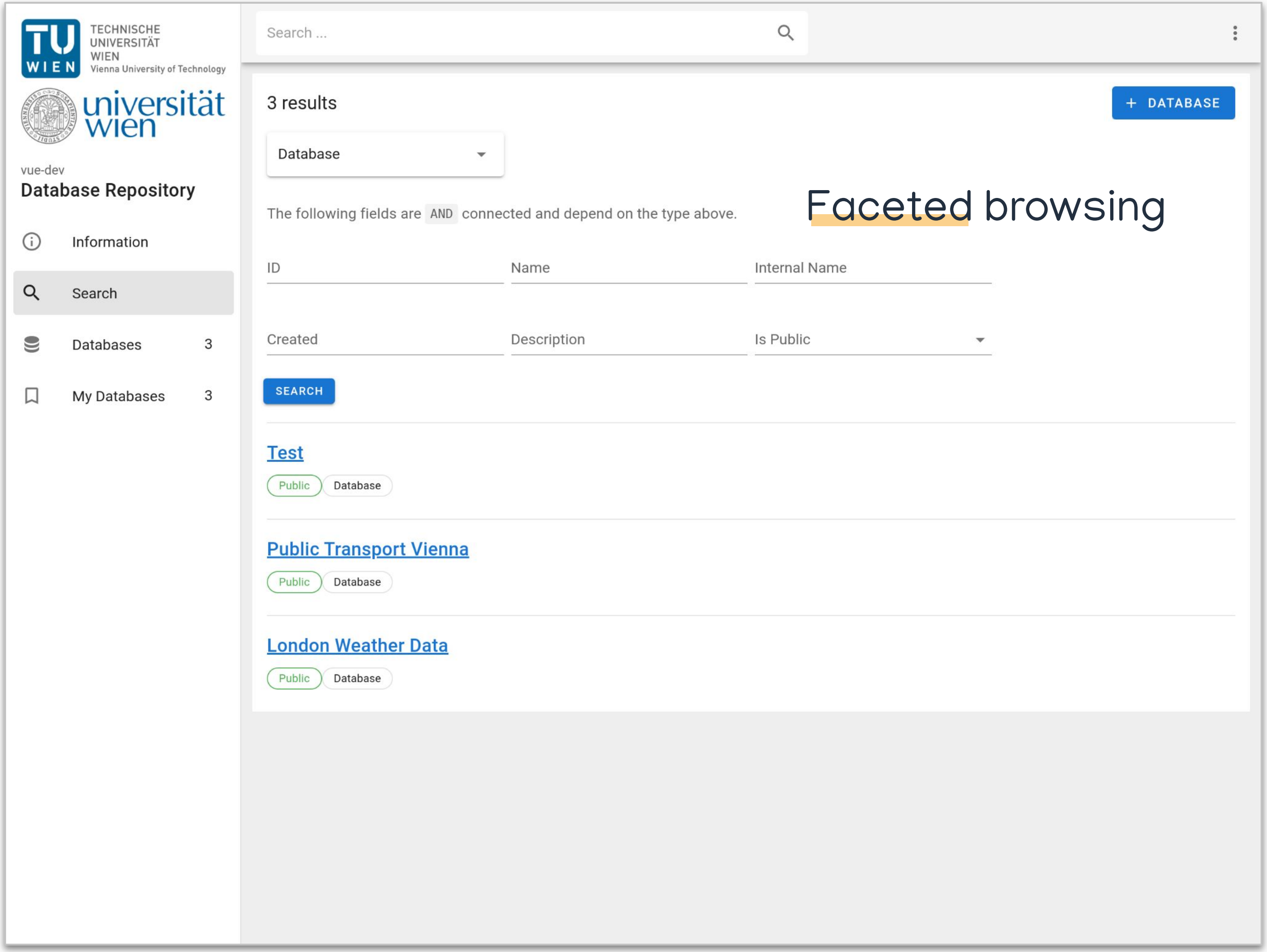
Titles

Description

Public | Identifier

Sotirios Tsepelakis

Faceted browsing



3 results

Database

The following fields are `AND` connected and depend on the type above.

ID

Name

Internal Name

Created

Description

Is Public

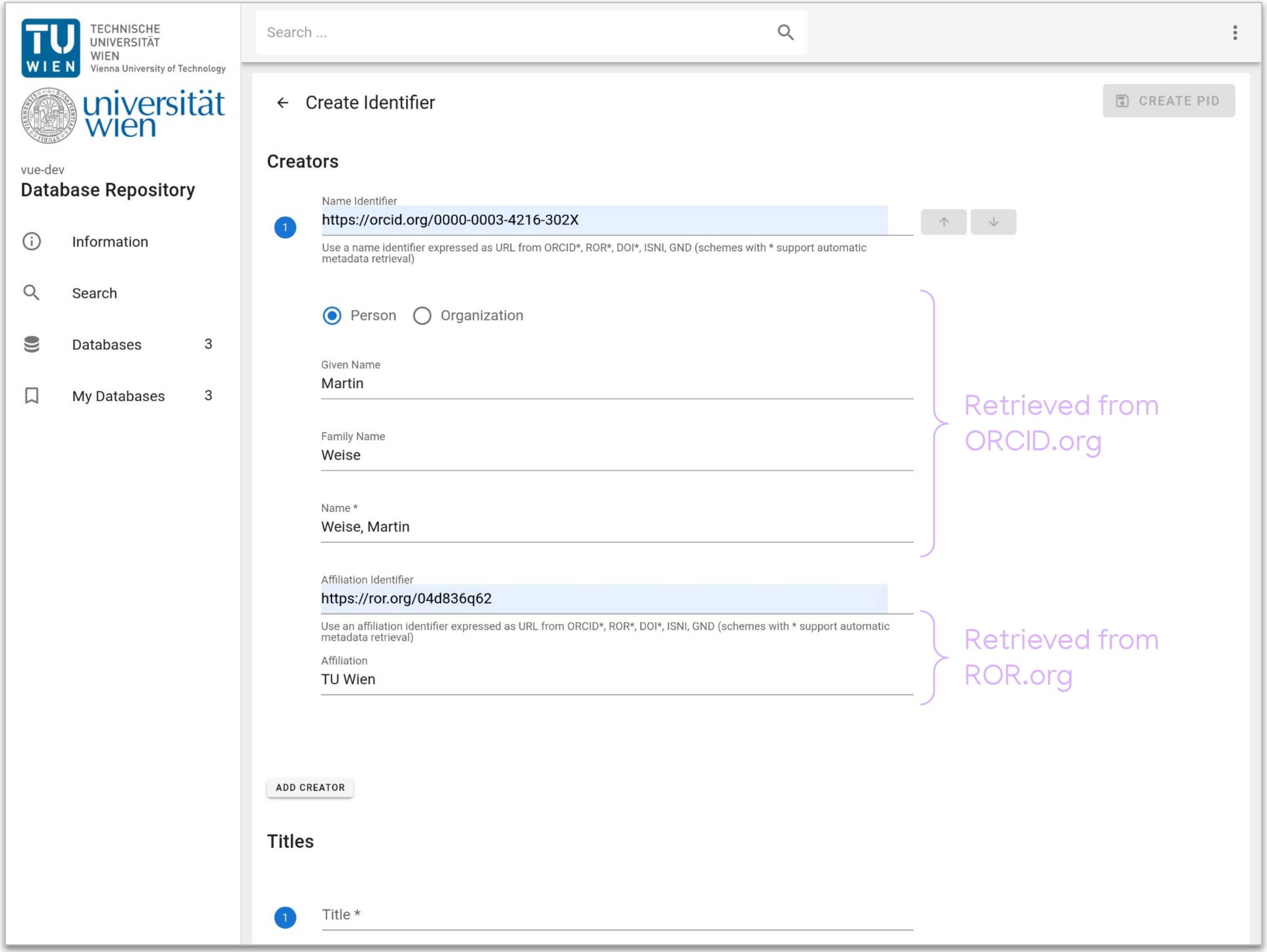
Test
Public Database

Public Transport Vienna
Public Database

London Weather Data
Public Database

Sotirios Tsepelakis

Resolve external identifiers to increase metadata quality



Create Identifier

Creators

1

Name Identifier: https://orcid.org/0000-0003-4216-302X

Use a name identifier expressed as URL from ORCID*, ROR*, DOI*, ISNI, GND (schemes with * support automatic metadata retrieval)

● Person ○ Organization

Given Name: Martin

Family Name: Weise

Name *: Weise, Martin

Affiliation Identifier: https://ror.org/04d836q62

Use an affiliation identifier expressed as URL from ORCID*, ROR*, DOI*, ISNI, GND (schemes with * support automatic metadata retrieval)

Affiliation: TU Wien

Retrieved from ORCID.org

Retrieved from ROR.org

Titles

1

Title *

Martin Weise

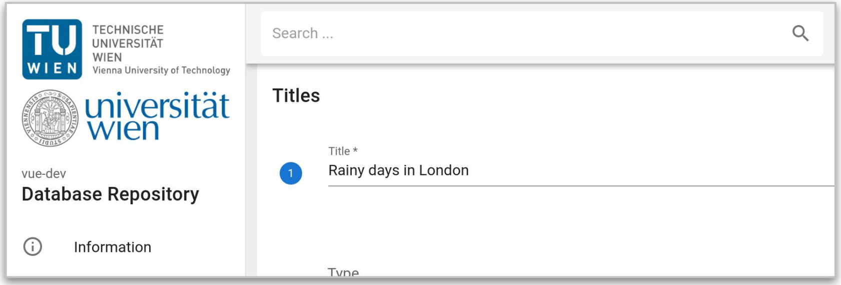
WP2 - Conversion between Units of Measurement

GOALS

Extend the metadata stored for each column that contains measurements to also allow the collection of metadata to enhance the conversation between units of measurement

Extend the search further to allow unit-independent search within the Ontology of units of measurements

Old



vue-dev
Database Repository

Titles

Title *
1 Rainy days in London

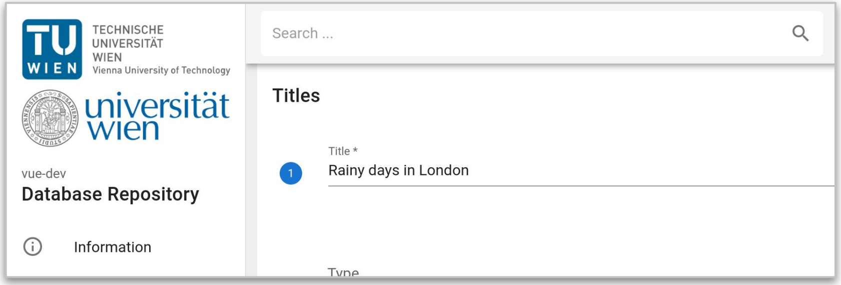
Type

WP2 - Conversion between Units of Measurement

DELIVERIES

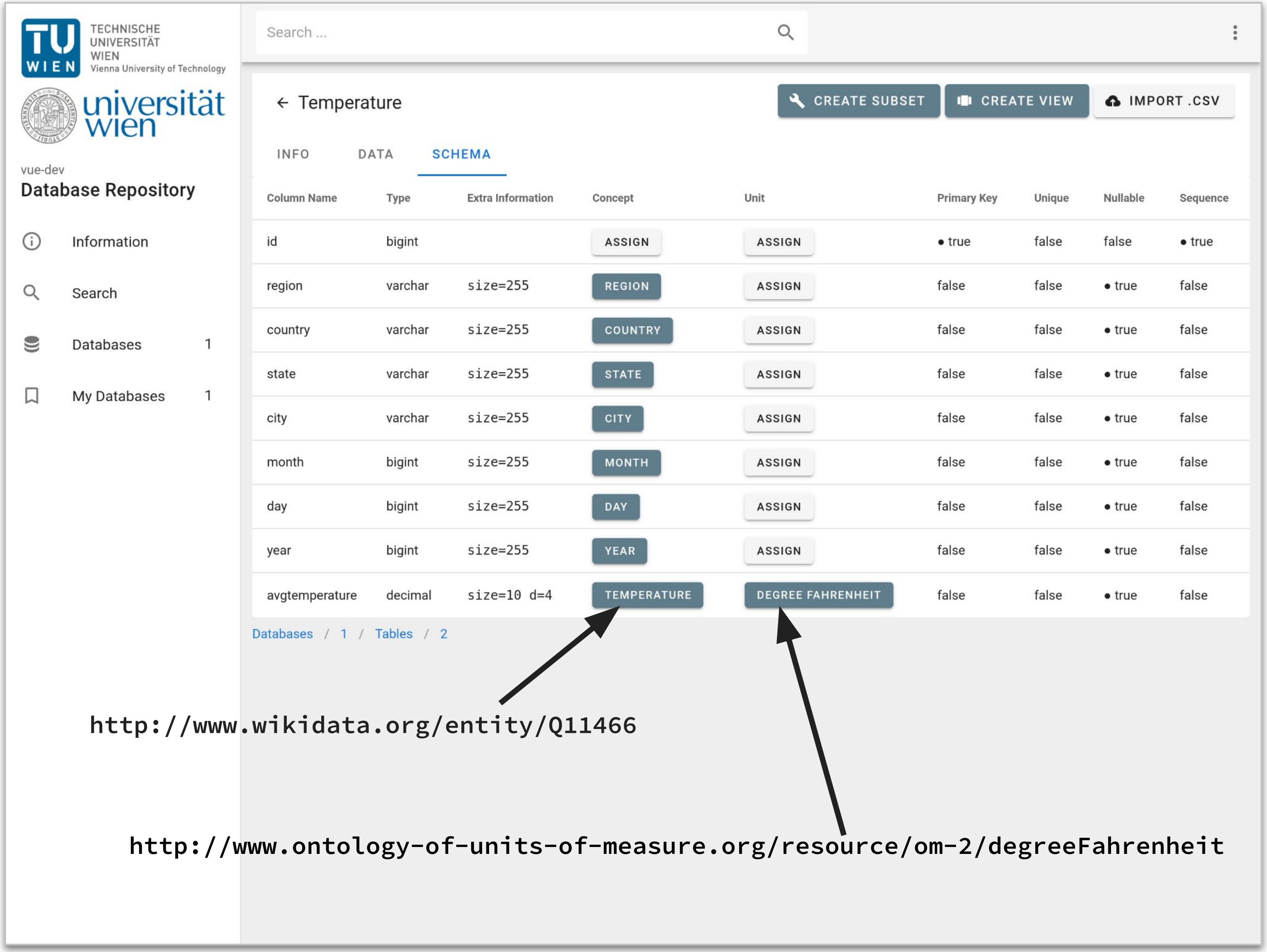
Conversation between Units of Measurement possible

Old



Title *
Rainy days in London

Collect semantic metadata



← Temperature

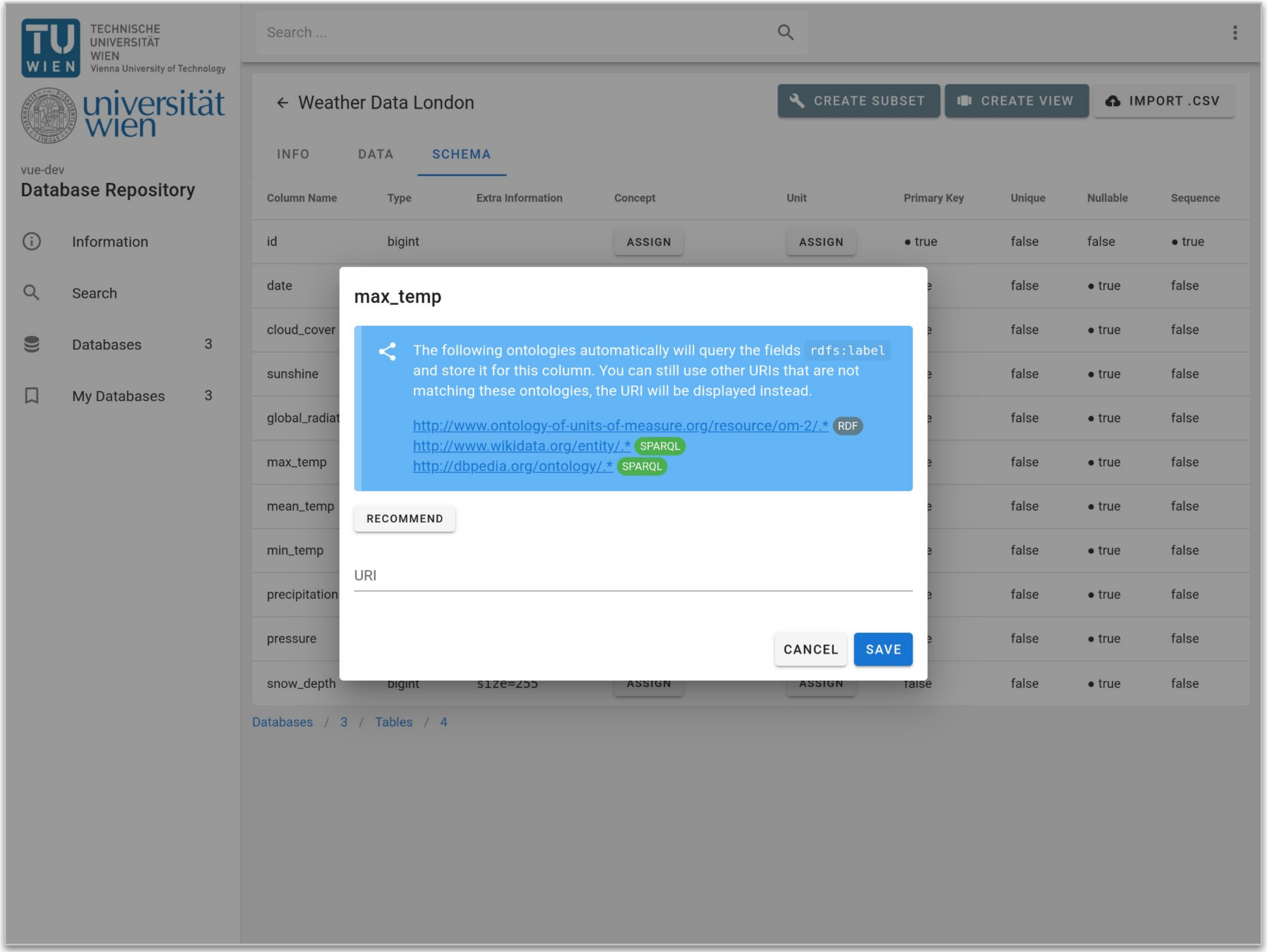
Column Name	Type	Extra Information	Concept	Unit	Primary Key	Unique	Nullable	Sequence
id	bigint		ASSIGN	ASSIGN	● true	false	false	● true
region	varchar	size=255	REGION	ASSIGN	false	false	● true	false
country	varchar	size=255	COUNTRY	ASSIGN	false	false	● true	false
state	varchar	size=255	STATE	ASSIGN	false	false	● true	false
city	varchar	size=255	CITY	ASSIGN	false	false	● true	false
month	bigint	size=255	MONTH	ASSIGN	false	false	● true	false
day	bigint	size=255	DAY	ASSIGN	false	false	● true	false
year	bigint	size=255	YEAR	ASSIGN	false	false	● true	false
avgtemperature	decimal	size=10 d=4	TEMPERATURE	DEGREE FAHRENHEIT	false	false	● true	false

Databases / 1 / Tables / 2

http://www.wikidata.org/entity/Q11466

http://www.ontology-of-units-of-measure.org/resource/om-2/degreeFahrenheit

Martin Weise



TECHNISCHE UNIVERSITÄT WIEN Vienna University of Technology

universität wien

vue-dev

Database Repository

- Information
- Search
- Databases 3
- My Databases 3

Search ...

← Weather Data London

CREATE SUBSET | CREATE VIEW | IMPORT .CSV

INFO | DATA | SCHEMA

Column Name	Type	Extra Information	Concept	Unit	Primary Key	Unique	Nullable	Sequence
id	bigint		ASSIGN	ASSIGN	● true	false	false	● true
date					e	false	● true	false
cloud_cover					e	false	● true	false
sunshine					e	false	● true	false
global_radiat					e	false	● true	false
max_temp					e	false	● true	false
mean_temp					e	false	● true	false
min_temp					e	false	● true	false
precipitation					e	false	● true	false
pressure					e	false	● true	false
snow_depth	bigint	size=255	ASSIGN	ASSIGN	false	false	● true	false

Databases / 3 / Tables / 4

max_temp

The following ontologies automatically will query the fields `rdfs:label` and store it for this column. You can still use other URIs that are not matching these ontologies, the URI will be displayed instead.

http://www.ontology-of-units-of-measure.org/resource/om-2/.* RDF
http://www.wikidata.org/entity/.* SPARQL
http://dbpedia.org/ontology/.* SPARQL

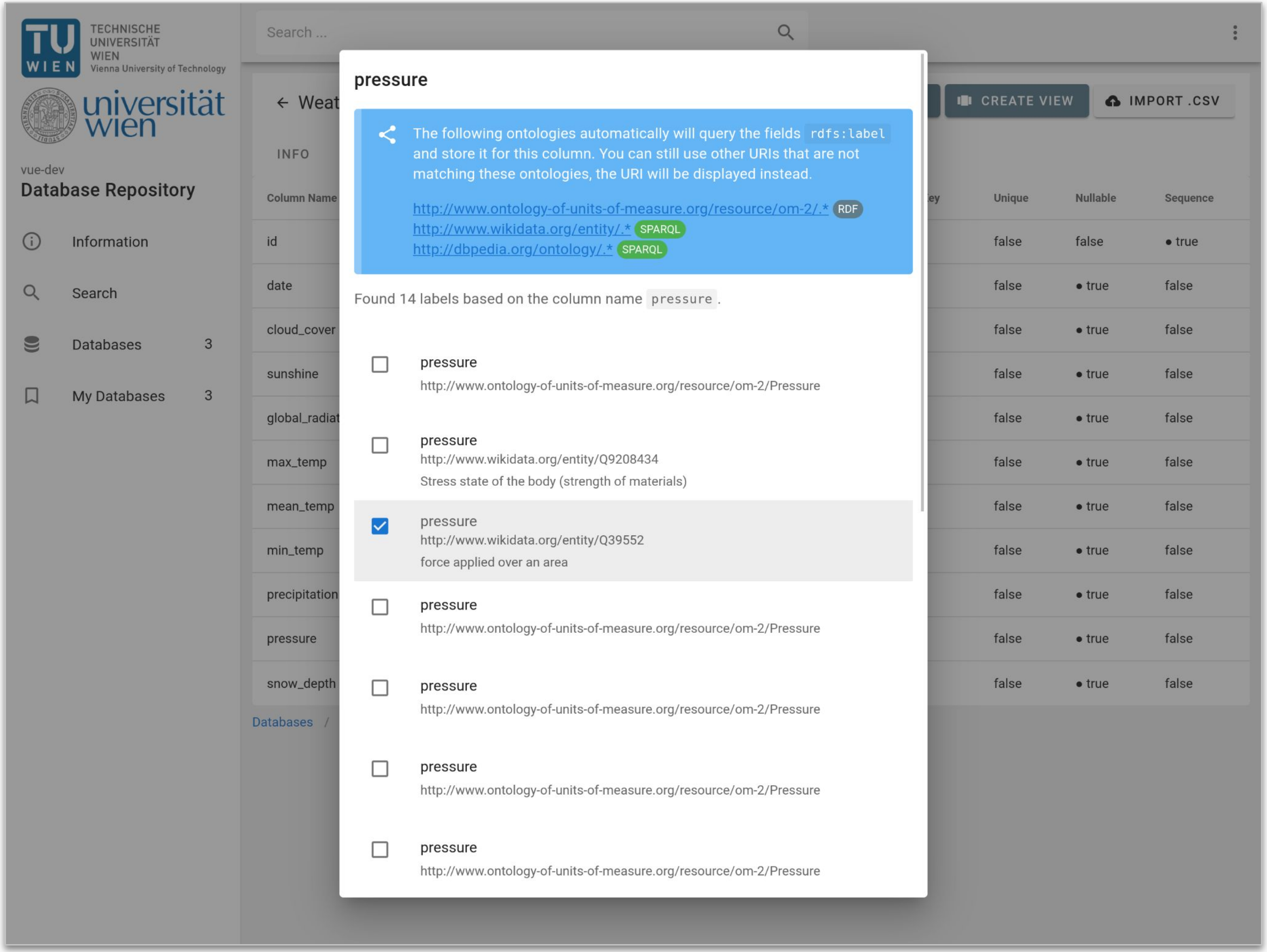
RECOMMEND

URI

CANCEL | SAVE

Recommend semantic metadata based on column name

Martin Weise



Database Repository

Column Name	ey	Unique	Nullable	Sequence
id		false	false	● true
date		false	● true	false
cloud_cover		false	● true	false
sunshine		false	● true	false
global_radiat		false	● true	false
max_temp		false	● true	false
mean_temp		false	● true	false
min_temp		false	● true	false
precipitation		false	● true	false
pressure		false	● true	false
snow_depth		false	● true	false

pressure

The following ontologies automatically will query the fields `rdfs:label` and store it for this column. You can still use other URIs that are not matching these ontologies, the URI will be displayed instead.

http://www.ontology-of-units-of-measure.org/resource/om-2/.* RDF
http://www.wikidata.org/entity/.* SPARQL
http://dbpedia.org/ontology/.* SPARQL

Found 14 labels based on the column name `pressure`.

☐ pressure
http://www.ontology-of-units-of-measure.org/resource/om-2/Pressure

☐ pressure
http://www.wikidata.org/entity/Q9208434
Stress state of the body (strength of materials)

☑ pressure
http://www.wikidata.org/entity/Q39552
force applied over an area

☐ pressure
http://www.ontology-of-units-of-measure.org/resource/om-2/Pressure

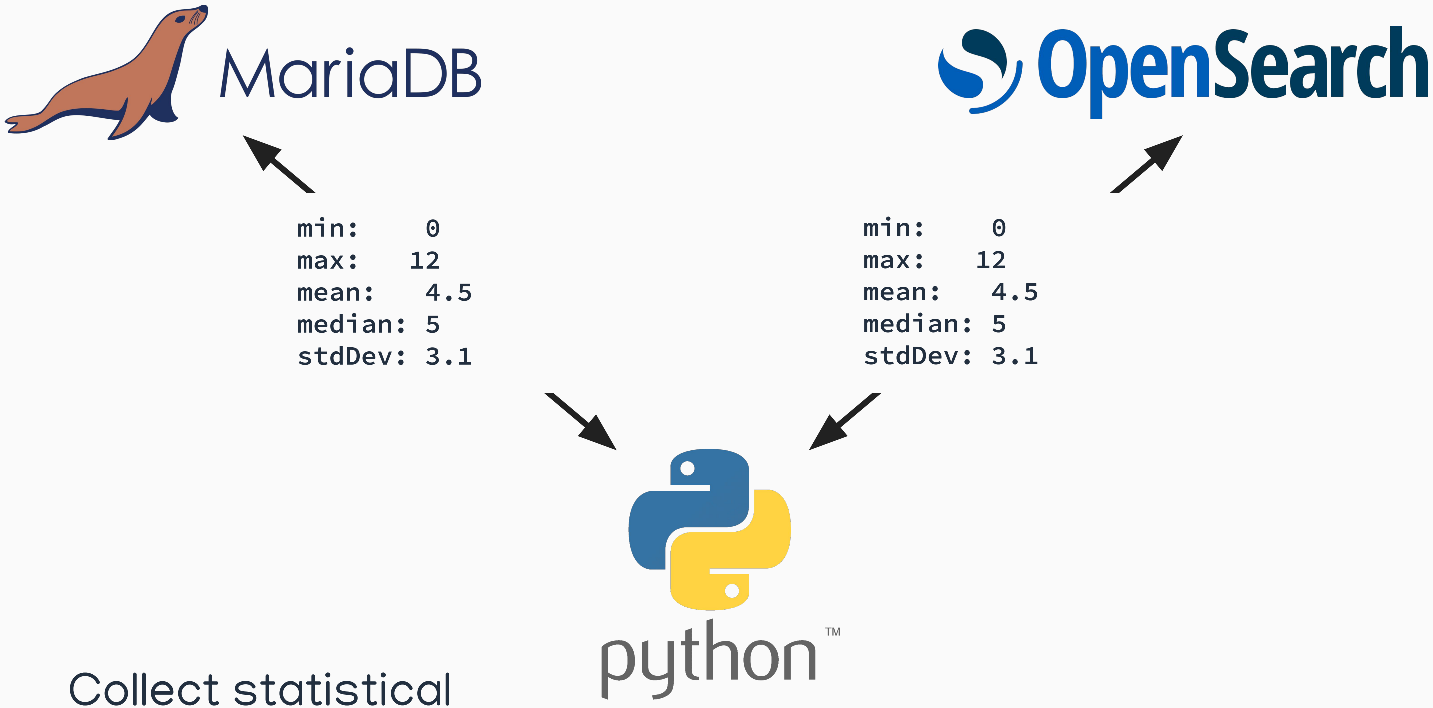
☐ pressure
http://www.ontology-of-units-of-measure.org/resource/om-2/Pressure

☐ pressure
http://www.ontology-of-units-of-measure.org/resource/om-2/Pressure

☐ pressure
http://www.ontology-of-units-of-measure.org/resource/om-2/Pressure

Use external metadata to increase quality

Martin Weise



```
min:     0
max:    12
mean:    4.5
median: 5
stdDev: 3.1
```

```
min:     0
max:    12
mean:    4.5
median: 5
stdDev: 3.1
```

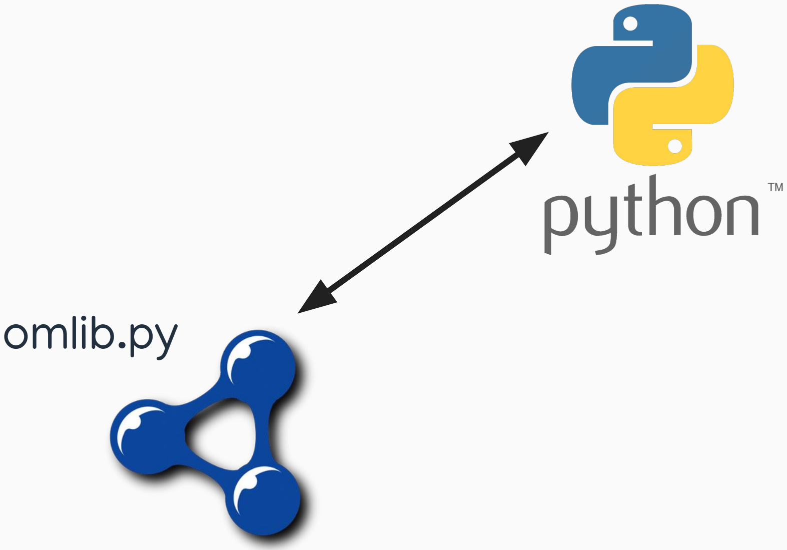
Collect statistical properties for each column

Nikola Lukic

```
for unit_uri in unit_uris:
    gte = t1
    lte = t2
    if unit_uri != field_value_pairs["unit.uri"]:
        target_unit = unit_uri_to_unit(unit_uri)
        if not Unit.can_convert(base_unit, target_unit):
            continue
        gte = om(t1, base_unit).convert(target_unit)
        lte = om(t2, base_unit).convert(target_unit)
        searches.append({"index": "column"})
        searches.append({
        "query": {
        "bool": {
            "must": [
                {
                    "match": {
                        "concept.uri": {
                            "query": field_value_pairs["concept.uri"]
                        }
                    }
                },
                {
                    "range": {
                        "val_min": {
                            "gte": gte
                        }
                    }
                },
                {
                    "range": {
                        "val_max": {
                            "lte": lte
                        }
                    }
                },
                {
                    "match": {
                        "unit.uri": {
                            "query": unit_uri
                        }
                    }
                }
            ]
        }
    }
    })
```

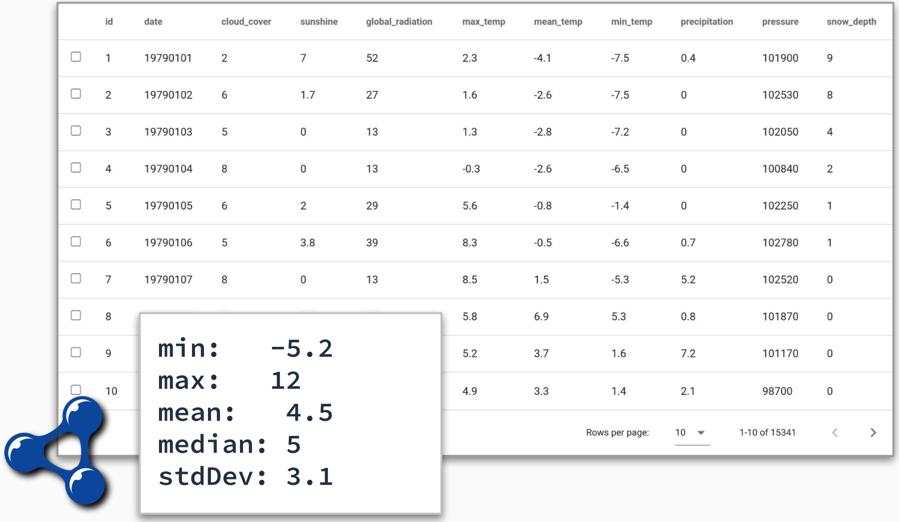
Convert statistical properties not in the target unit

Search unit-independent between [t1, t2]



Max Spannring

EXAMPLE

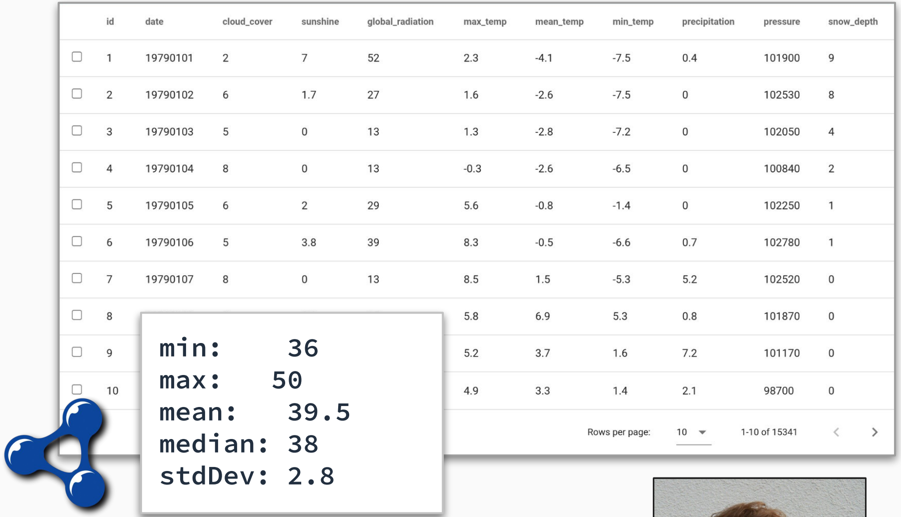


☐	id	date	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth
☐	1	19790101	2	7	52	2.3	-4.1	-7.5	0.4	101900	9
☐	2	19790102	6	1.7	27	1.6	-2.6	-7.5	0	102530	8
☐	3	19790103	5	0	13	1.3	-2.8	-7.2	0	102050	4
☐	4	19790104	8	0	13	-0.3	-2.6	-6.5	0	100840	2
☐	5	19790105	6	2	29	5.6	-0.8	-1.4	0	102250	1
☐	6	19790106	5	3.8	39	8.3	-0.5	-6.6	0.7	102780	1
☐	7	19790107	8	0	13	8.5	1.5	-5.3	5.2	102520	0
☐	8					5.8	6.9	5.3	0.8	101870	0
☐	9					5.2	3.7	1.6	7.2	101170	0
☐	10					4.9	3.3	1.4	2.1	98700	0

Rows per page: 10 ▾ 1-10 of 15341 ‹ ›

```
min:    -5.2
max:    12
mean:    4.5
median:  5
stdDev:  3.1
```

Temperature / °C



☐	id	date	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth
☐	1	19790101	2	7	52	2.3	-4.1	-7.5	0.4	101900	9
☐	2	19790102	6	1.7	27	1.6	-2.6	-7.5	0	102530	8
☐	3	19790103	5	0	13	1.3	-2.8	-7.2	0	102050	4
☐	4	19790104	8	0	13	-0.3	-2.6	-6.5	0	100840	2
☐	5	19790105	6	2	29	5.6	-0.8	-1.4	0	102250	1
☐	6	19790106	5	3.8	39	8.3	-0.5	-6.6	0.7	102780	1
☐	7	19790107	8	0	13	8.5	1.5	-5.3	5.2	102520	0
☐	8					5.8	6.9	5.3	0.8	101870	0
☐	9					5.2	3.7	1.6	7.2	101170	0
☐	10					4.9	3.3	1.4	2.1	98700	0

Rows per page: 10 ▾ 1-10 of 15341 ‹ ›

```
min:     36
max:    50
mean:    39.5
median:  38
stdDev:  2.8
```

Temperature / °F

"Give me tables with concept Temperature and °C between [0, 10]"

Max Spannring

omlib.py

EXAMPLE

“convert °F from (°C × 9/5) + 32”

“Do you contain values 0–10?”

```
min:    -5.2
max:    12
mean:    4.5
median: 5
stdDev: 3.1
```

Temperature / °C

“Do you contain values 32–50?”

```
min:     31
max:    50
mean:    39.5
median: 38
stdDev: 2.8
```

Temperature / °F

“Give me tables with concept Temperature and °C between [0, 10]”

Max Spannring

Helm charts for Kubernetes deployments

Generic open-source cloud deployment (for any cloud)

`oci://dbrepo.azurecr.io/helm/dbrepo-core`

TU Wien flavored cloud deployment (for any cloud)

`oci://dbrepo.azurecr.io/helm/dbrepo-tuwien`

+ SSO proxy
+ Prometheus monitoring
+ Grafana dashboard

Martin Weise

Future Work

VISION

Collaboration of TUWRD and DBRepo across VREs:

1. Get database snapshots from TUWRD (or other file-based repository)
2. Add semantic context for machine-understandability and explore tabular data in VRE
3. Seamlessly store finished research artifacts in TUWRD (e.g. plots) and data that produced these plots as queries/views in DBRepo