LOD-GEOSS

Implementing FAIR through a distributed data infrastructure

Carsten Hoyer-Klick

Supported by: Federal Ministry for Economic Affairs and Energy on the basis of a decision by the German Bundestag

Research Grant 03EI1005A-G

DLR Deutsches Zentrum für Luft- und Raumfahrt German Aerospace Center

REINER LEMOINE INSTITUT

InfAI Institute for Applied Informatics

Universität Stuttgart IER Institut für Energiewirtschaft und Rationelle Energieanwendung

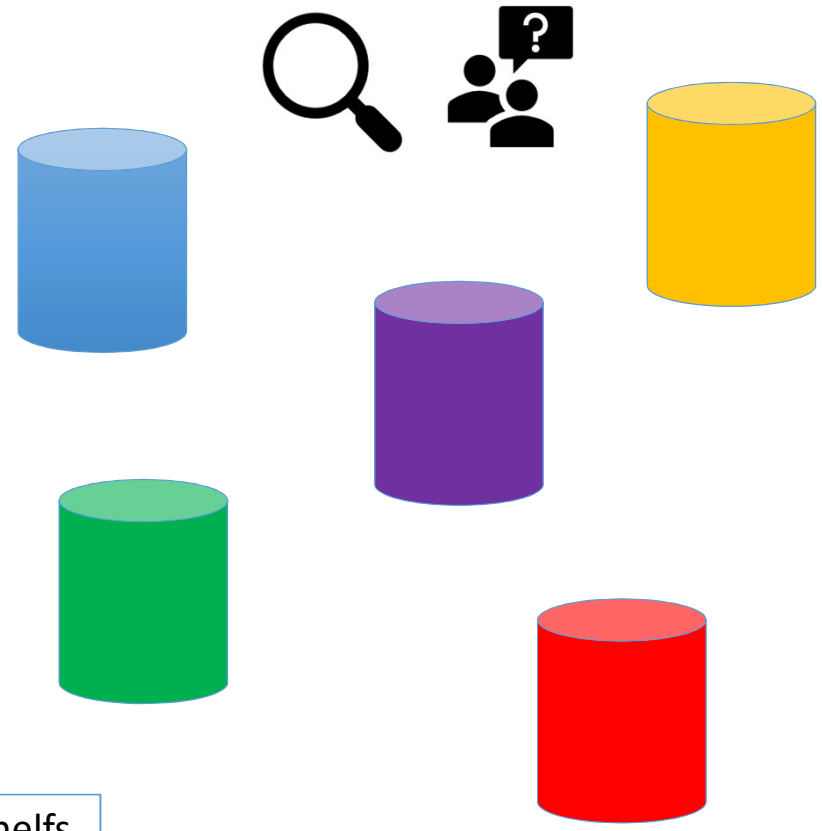
PIK Potsdam Institute for Climate Impact Research

JÜLICH Forschungszentrum

Challenges I: Finding and Accessing Data

- Many data bases exist, each in its own flavor
  - Data access
  - Data format
  - Data licenses (if at all)
  - Sometimes hard to find

- Data collection is a labor intensive task
- Data cleaning, aggregation, etc.
  is repeated by many researchers with different results
- Data quality is often unknown



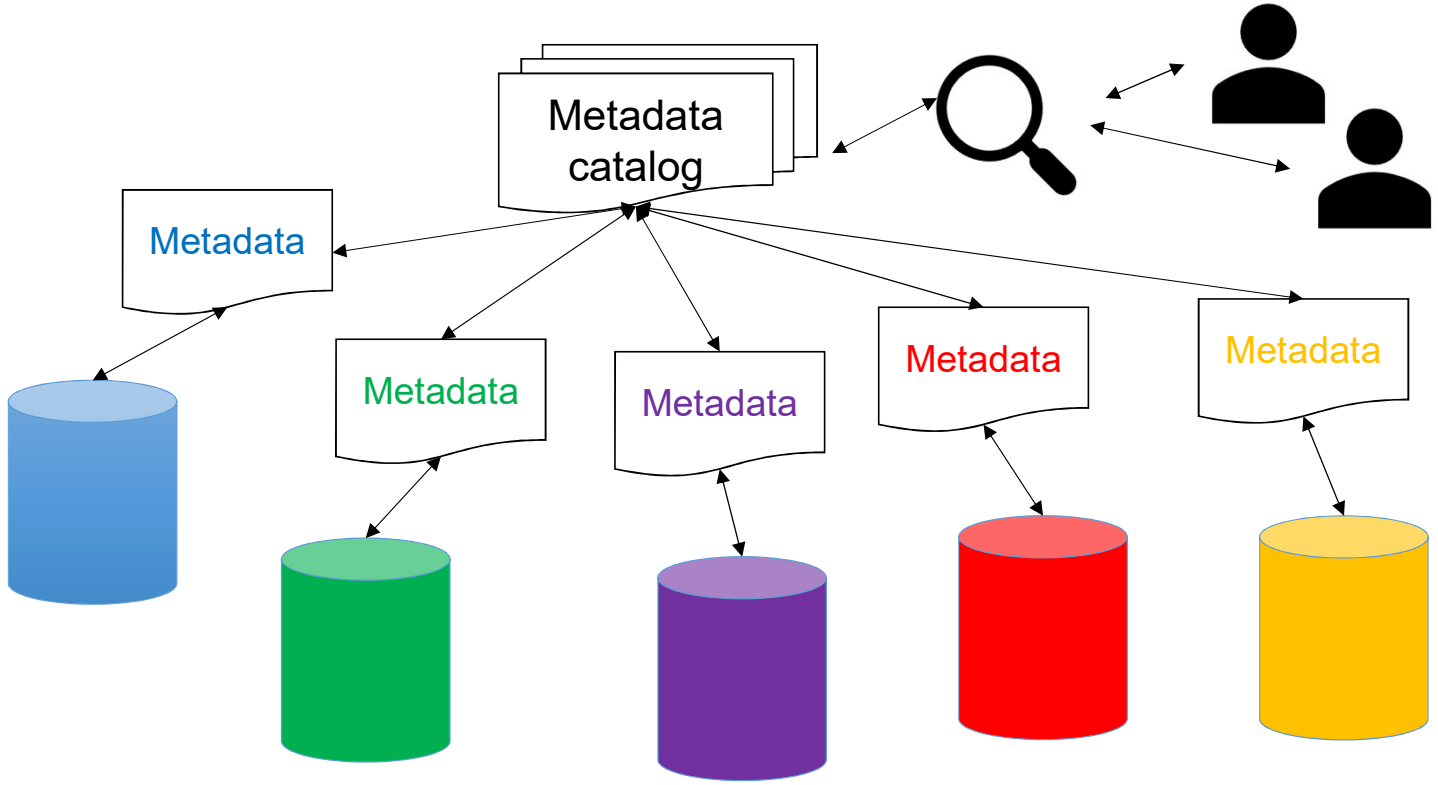
Imagine a library without a catalog and systematic numbers on the shelfs

Research Grant 03EI1005A-G

Solution I: A Metadata Catalog

LOD-GEOSS

- A metadata catalog harvests the (rich) metadata from the available data sources
- The catalog can be used to discover data
- The metadata contains a URI to the actual data
- In case of data bases possibly also an API/Interface description



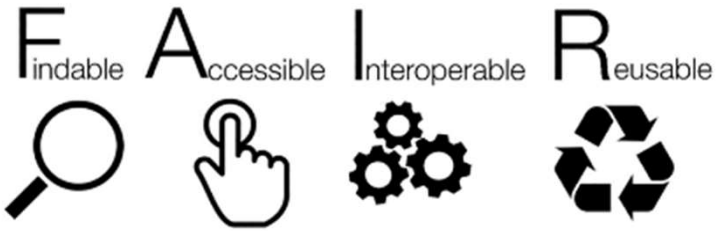
The domain uses the OEP Metadata string: https://openenergy-platform.org/tutorials/jupyter/OEMetadata/

Supported by: Federal Ministry for Economic Affairs and Energy, on the basis of a decision by the German Bundestag
Research Grant 03EI1005A-G

FAIR Data Principles

LOD-GEOSS

- FAIR stands for
  - **Findable**: Data and Metadata should be easy to find for both humans and computers. Machine readable metadata are essential for automatic discovery of data sets and services
  - **Accessible**: Once the user finds the required data, she/he needs to now how they can be accessed
  - **Interoperable**: The data need to be integrated with other data. In addition the need to interoperate with applications or workflows for analysis, storage and processing
  - **Reusable**: The ultimate goal of FAIR is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.



Source: go-fair.org

Supported by: Federal Ministry for Economic Affairs and Energy on the basis of a decision by the German Bundestag
Research Grant 03EI1005A-G

Why Rich Metadata?

LOD-GEOSS

Spinach?

Ravioli?

Without a label, how would you know which one to open?

Metadata contains vital information

		DE_BadenWue	DE_Bayern	DE_Berlin	DE_Brandenburg
processHeat500	t0001	0.000087	0.000076	0.00007	0.00007
processHeat500	t0002	0.000085	0.000076	0.000068	0.000068
processHeat500	t0003	0.000086	0.000076	0.000069	0.000069
processHeat500	t0004	0.000087	0.000077	0.00007	0.00007
processHeat500	t0005	0.000121	0.000116	0.000112	0.000112
processHeat500	t0006	0.000167	0.000168	0.00017	0.00017
processHeat500	t0007	0.000195	0.000199	0.000203	0.000203
processHeat500	t0008	0.000204	0.000208	0.000212	0.000212
processHeat500	t0009	0.000205	0.000209	0.000212	0.000212
processHeat500	t0010	0.000188	0.00019	0.000191	0.000191
processHeat500	t0011	0.000187	0.000188	0.000189	0.000189
processHeat500	t0012	0.000184	0.000187	0.000187	0.000187
processHeat500	t0013	0.000182	0.000184	0.000186	0.000186
processHeat500	t0014	0.000181	0.000183	0.000185	0.000185

What do the collumns stand for?

Supported by: Federal Ministry for Economic Affairs and Energy, on the basis of a decision by the German Bundestag

Research Grant 03EI1005A-G

DLR Deutsches Zentrum für Luft- und Raumfahrt German Aerospace Center · REINER LEMOINE INSTITUT · InfAI Institute for Applied Informatics · Universität Stuttgart IER Institut für Energiewirtschaft und Rationelle Energieanwendung · PIK Potsdam Institute for Climate Impact Research · JÜLICH Forschungszentrum

The OEO Metadata String

LOD-GEOSS

- Developed within the „The Open Energy Familiy“ as part of open science activities within the Domain of Energy Systems Analysis
- Implemented as a JSON-String -> can still be read and edited by humans
- Contains all main elements of metadata
  - **Descriptive**: Descriptions of the content, identification, make the data findable
  - **Structural**: References to other (meta-)data, information about the structure of the data
  - **Administrative**: Management of the data
    - **Use:** How to use the data, protocols, API
    - **Technical:** functions of a systems, digitalization, hardware and software used
    - **Preservation:** how the data is stored
    - **Rights:** Access rights and licenses
    - **Provenance:** Source of the data

Supported by: Federal Ministry for Economic Affairs and Energy on the basis of a decision by the German Bundestag
Research Grant 03EI1005A-G
DLR Deutsches Zentrum für Luft- und Raumfahrt German Aerospace Center
REINER LEMOINE INSTITUT
InfAI Institute for Applied Informatics
Universität Stuttgart IER Institut für Energiewirtschaft und Rationelle Energieanwendung
PIK Potsdam Institute for Climate Impact Research
JÜLICH Forschungszentrum

Metadata example – Wind Data

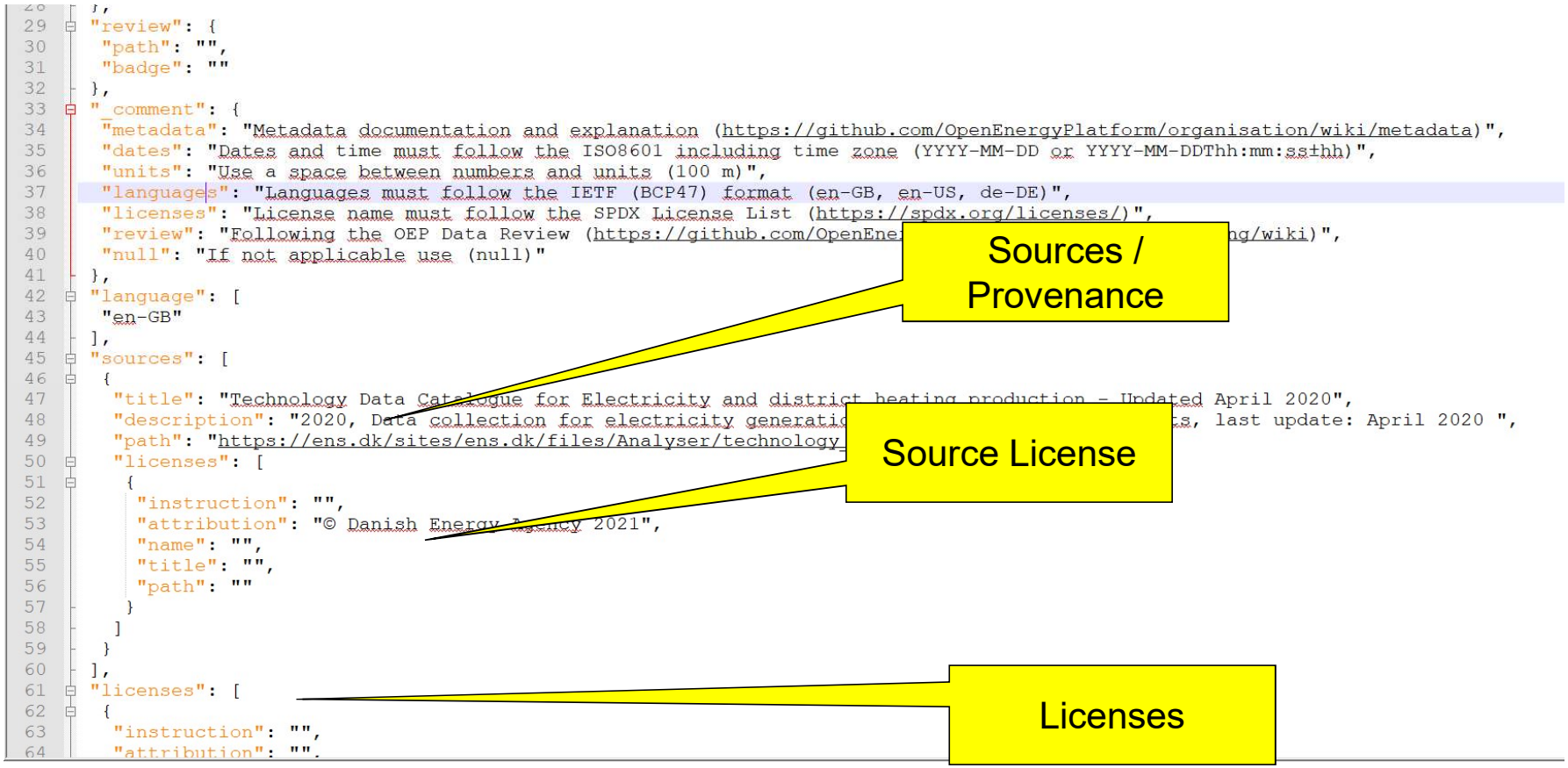

```
{
 "name": "wind_turbine_onshore_lod_geoss",
 "title": "wind_turbine_onshore_LOD_GEOSS",
 "id": "http://openenergyplatform.org/dataedit/view/model_draft/wind_turbine_onshore_lod_geoss",
 "description": "This is the first dataset to test the technology data prototype",
 "keywords": [
  ""
 ],
 "publicationDate": "2021-05-10",
 "context": {
  "homepage": "",
  "documentation": "",
  "sourceCode": "",
  "contact": "",
  "grantNo": "",
  "fundingAgency": "",
  "fundingAgencyLogo": "",
  "publisherLogo": ""
 },
 "spatial": {
  "location": "",
  "extent": "",
  "resolution": ""
 },
 "temporal": {
  "referenceDate": "2021-05-1(
  "timeseries": {}
 },
 "review": {
  "path": "",
  "badge": ""
 },
```


OEP Metadata format and description:

- https://github.com/OpenEnergyPlatform/metadata
- https://github.com/OpenEnergyPlatform/oemetadata/blob/master/metadata/latest/metadata_key_description.md

Research Grant 03EI1005A-G

Metadata example – Wind data

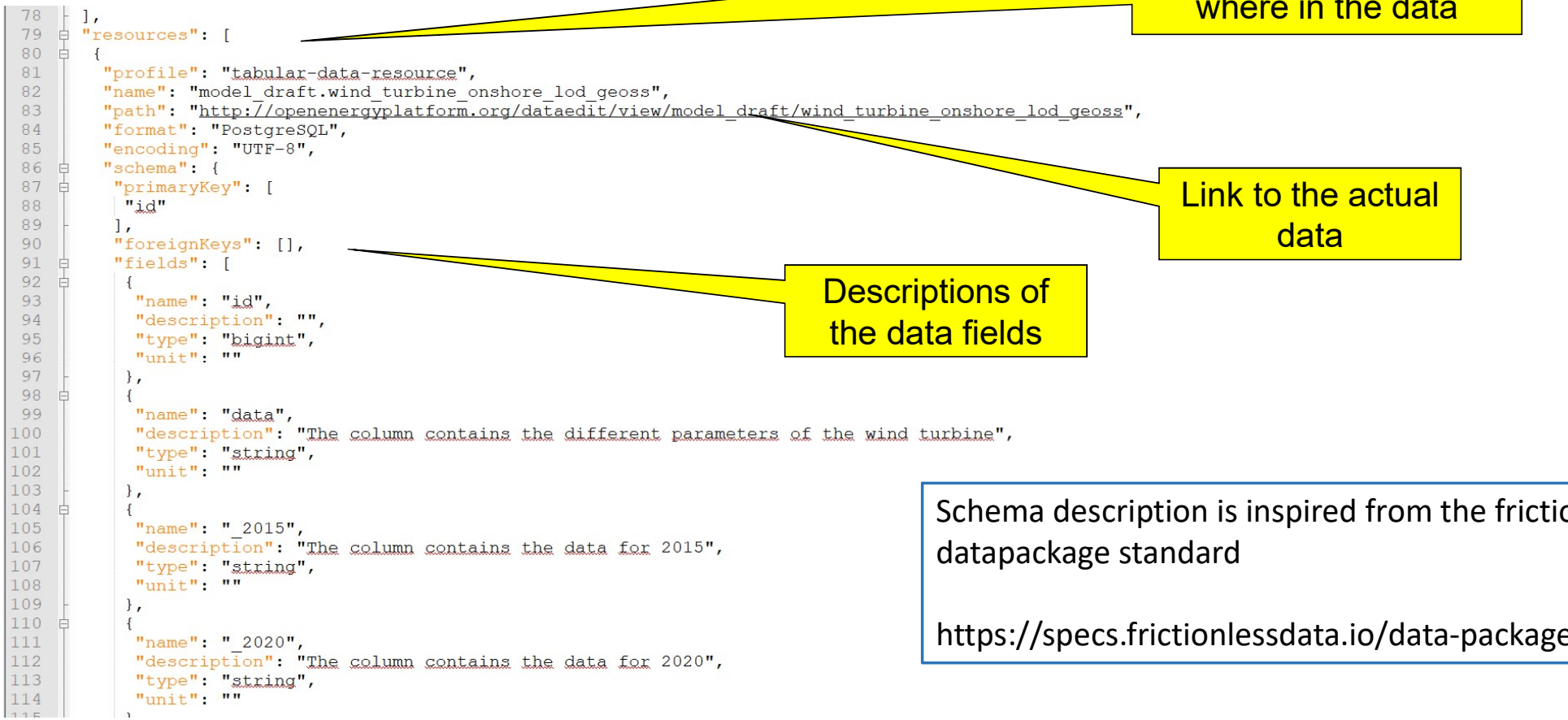


Supported by: Federal Ministry for Economic Affairs and Energy, on the basis of a decision by the German Bundestag

Research Grant 03EI1005A-G

DLR Deutsches Zentrum für Luft- und Raumfahrt German Aerospace Center
REINER LEMOINE INSTITUT
InfAI Institute for Applied Informatics
Universität Stuttgart IER Institut für Energiewirtschaft und Rationelle Energieanwendung
PIK Potsdam Institute for Climate Impact Research
JÜLICH Forschungszentrum

Metadata example – Wind data



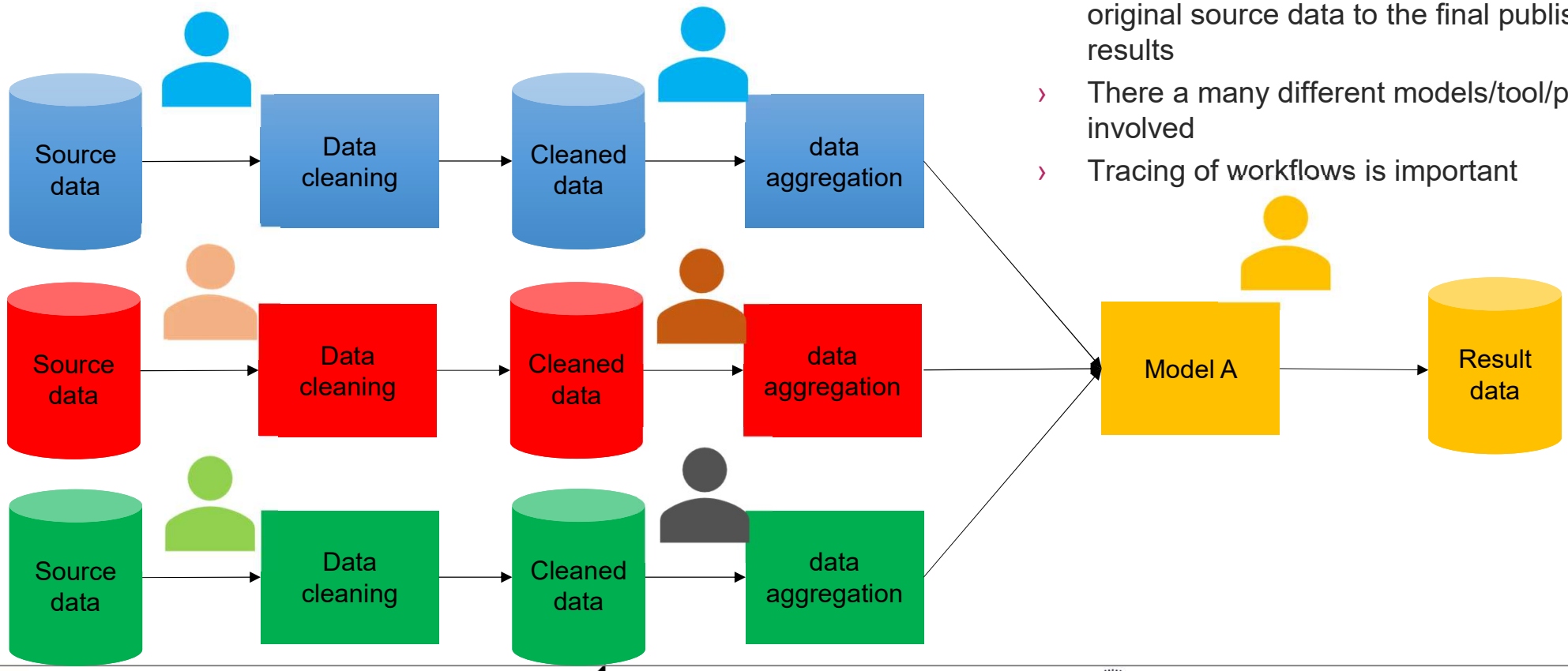
```
],
"resources": [
 {
  "profile": "tabular-data-resource",
  "name": "model_draft.wind_turbine_onshore_lod_geoss",
  "path": "http://openenergyplatform.org/dataedit/view/model_draft/wind_turbine_onshore_lod_geoss",
  "format": "PostgreSQL",
  "encoding": "UTF-8",
  "schema": {
   "primaryKey": [
    "id"
   ],
   "foreignKeys": [],
   "fields": [
    {
     "name": "id",
     "description": "",
     "type": "bigint",
     "unit": ""
    },
    {
     "name": "data",
     "description": "The column contains the different parameters of the wind turbine",
     "type": "string",
     "unit": ""
    },
    {
     "name": "_2015",
     "description": "The column contains the data for 2015",
     "type": "string",
     "unit": ""
    },
    {
     "name": "_2020",
     "description": "The column contains the data for 2020",
     "type": "string",
     "unit": ""
```


Schema description is inspired from the frictionless datapackage standard

https://specs.frictionlessdata.io/data-package/

Research Grant 03EI1005A-G

Challenge II: Who has done what to the data?



- There are many processing steps from the original source data to the final published results
- There a many different models/tool/persons involved
- Tracing of workflows is important

Research Grant 03EI1005A-G

PROV of the W3 Provenance Working Group

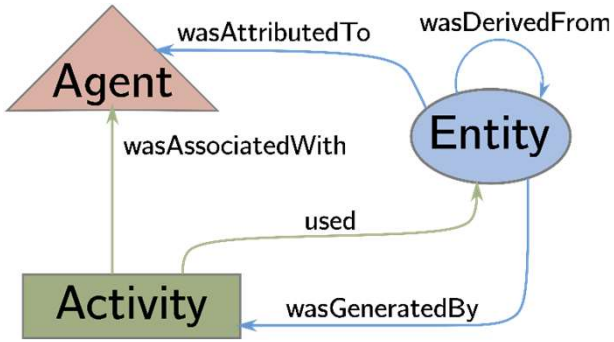
Entities	Digital, physical or conceptional (.e.g. data sets or files)
Activities	Use or generate entities, e.g. algorithms or workflows
Agents	Have control over activities or own entities, e.g. persons, organizations, software

Entities, **Activities** and **Agents** are related through **relations** with each other.

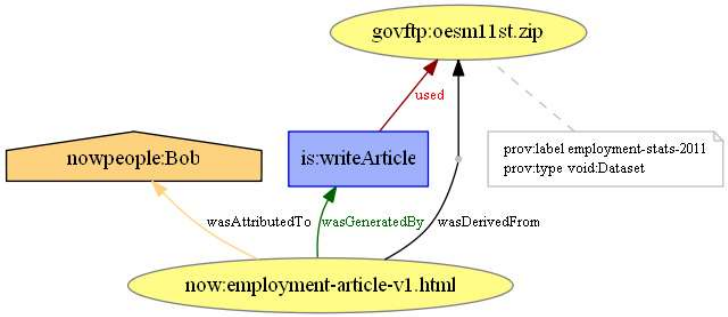
Provenance Information can be collected automatically and visualized in graphs, e.g. with ProvStore

A record **of who (Agent)** processed with which **Tools (Activity)** which **data (Entities, input and output)**

Schematic:



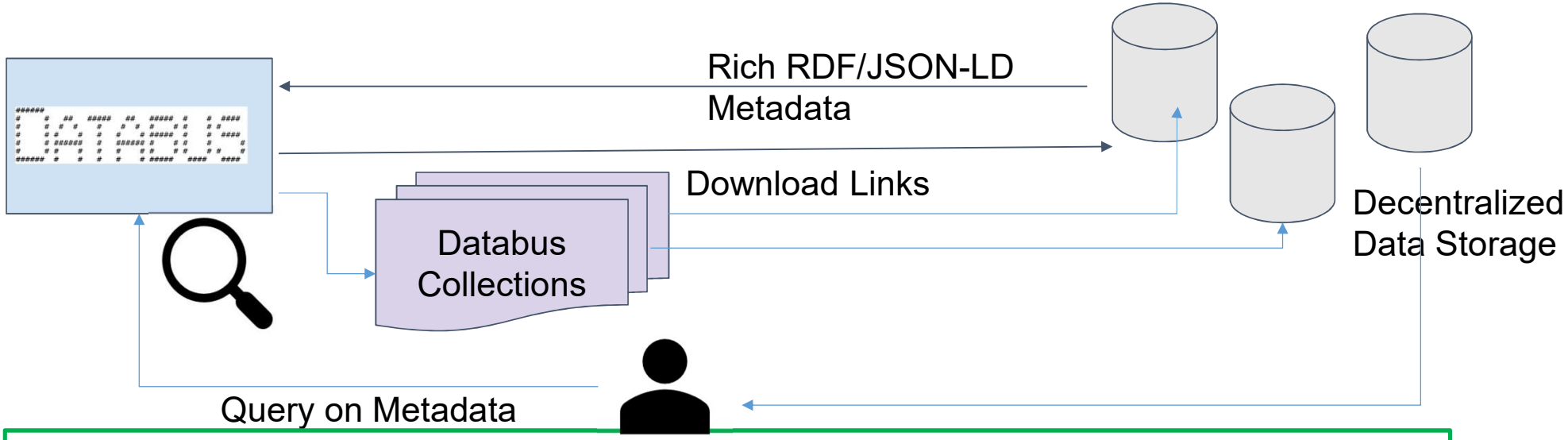
Example:



Research Grant 03EI1005A-G

The Databus Platform

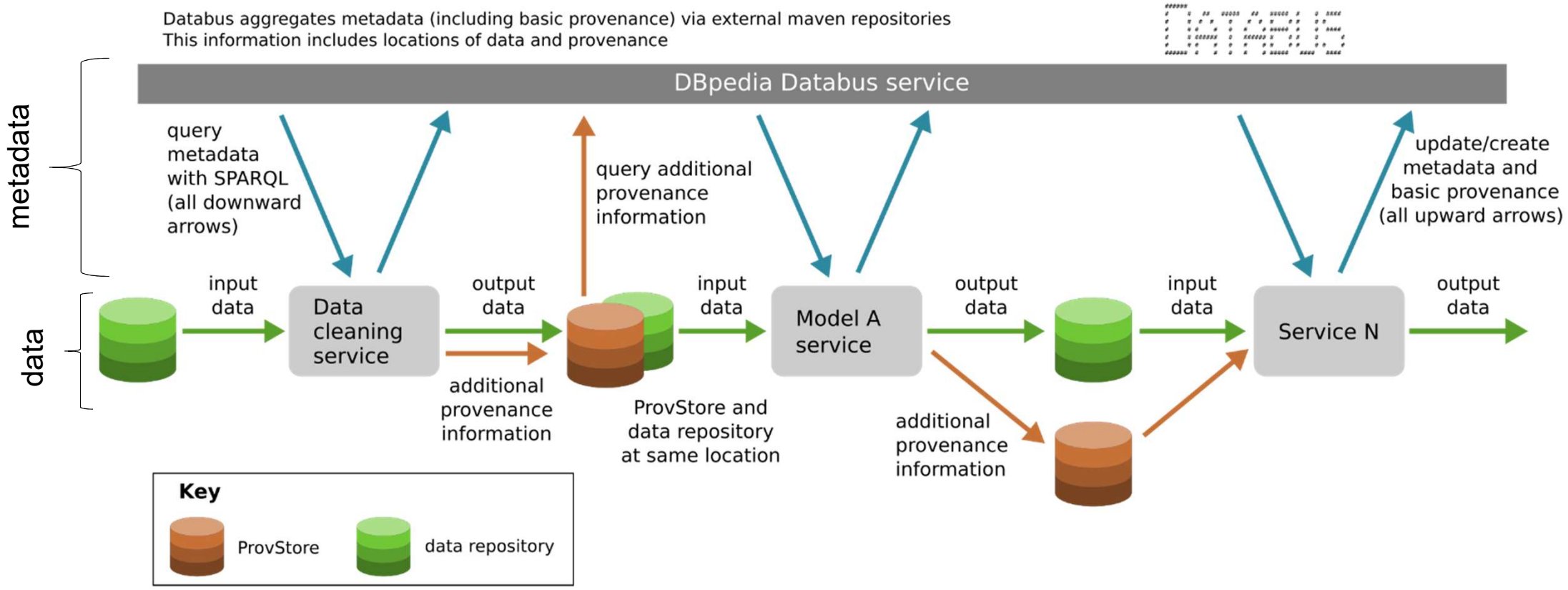
Databus is a virtual bus. It can address files on the web and coordinate dataflows based on DataID metadata. No actual data is uploaded to the bus.



- Unique and persistent data identifiers (PID) are created by the Databus
- Data sets are linked to their source data through the data ids
- Incremental modifications to data (e.g. people can reuse cleanings or aggregations someone else has done before)

Research Grant 03EI1005A-G

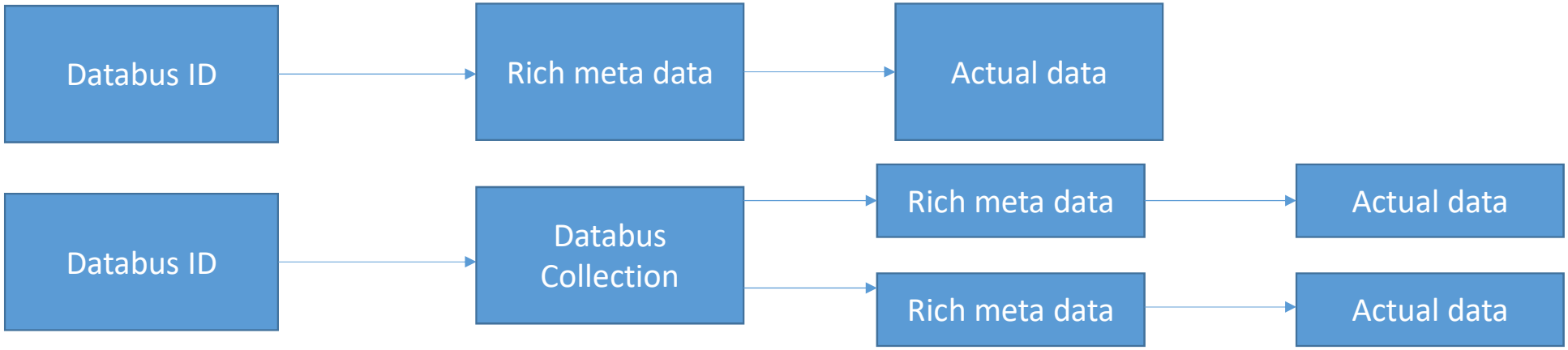
Architecture Concept



Research Grant 03EI1005A-G

Databus as PID / Fair Digital Object (FDO) provider

- Uploading meta data to the Databus creates a unique ID for the data set which can be used as a persistent data identifier (PID) to cite and share the data
  - Databus URI (PID): https://databus.dbpedia.org/account_name/group/artifact/version /
  - Databus collections can be used to group data
  - The actual data is linked within the meta data
- Databus ID can be used to access all necessary information
- Databus ID is a pointer to a (FAIR) digital object
- **Persistent ID to make data citable**



Research Grant 03EI1005A-G

Challenge III: What is the data about?

LOD-GEOSS

- Each data source comes with its own annotation

- Example from solar meteorology:
  - **GHI:** Global Horizontal Irradiation
  - **Global:** Could also be Global Horizontal Irradiation
  - **Surface downward irradiation:** The usual term in climate science for what we usually call GHI

- **Taxonomies** or **ontologies** create a **data language** to annotate data
- Ontologies can describe relations: *direct radiation* **is a part** of the *global irradiation reaching the surface*
- Ontologies make data interpretable, also by machines and algorithms

- **Good ontologies are created on a consensus building and open development process within the community.**
  - We use the 'Open Energy Ontology', https://openenergy-platform.org/ontology/, https://doi.org/10.1016/j.egyai.2021.100074

- New JASON-LD Meta data string can link meta data to an ontology

Supported by: Federal Ministry for Economic Affairs and Energy, on the basis of a decision by the German Bundestag

Research Grant 03EI1005A-G

Implementing the FAIR Principles

LOD-GEOSS

- **F**indable — Development and setup of the data bus as metadata catalog for data in energy systems analysis
- **A**ccessible — Descriptions of the data formats and interfaces to data bases, Best Practice Guides, links to the data sources in the metadata

  - Databus architecture and meta data
  - Citable PIDs

- **I**nteroperable — Enhancement of on Open Energy Ontology as a common data language, open data format descriptions

  - The Open Energy Ontology
  - Schema in the metadata

- **R**eusable — Improvement of data licenses, here with a special focus on GEOSS data, provenance information with the data

Supported by: Federal Ministry for Economic Affairs and Energy on the basis of a decision by the German Bundestag

Research Grant 03EI1005A-G

DLR Deutsches Zentrum für Luft- und Raumfahrt German Aerospace Center
REINER LEMOINE INSTITUT
InfAI Institute for Applied Informatics
Universität Stuttgart IER Institut für Energiewirtschaft und Rationelle Energieanwendung
PIK Potsdam Institute for Climate Impact Research
JÜLICH Forschungszentrum

Meta data version 1.5 JSON-LD

- Global:
  - @Context: Ontology links for the keys in the JSON
    `„@context“ : „https://raw.githubusercontent.com/LOD-GEOSS/databus-snippets/master/oep_metadata/context.jsonld“`
  - @id URI that links the resource via the dbpedia Databus / Persistent ID / PID
    `"@id": "https://databus.dbpedia.org/denis/lod-geoss-example/api-example/2021-05-10/api-example_type=turbineData.json"`
  - Subject: an OEO ID for the dataset
    `subject": "OEO:00000448"`
- Schema
  - Is_about: OEO links to the relenvat terms

```
{
     "name": "year",
     "description": "Describes the year the data applies to",
     "type": "int",
     "unit": "years",
     "isAbout": [
         {
             "name": "scenario year",
             "path": "http://openenergy-platform.org/ontology/oeo/OEO_00020097"
         }
      ],
}
```

Supported by: Federal Ministry for Economic Affairs and Energy, on the basis of a decision by the German Bundestag
Research Grant 03EI1005A-G

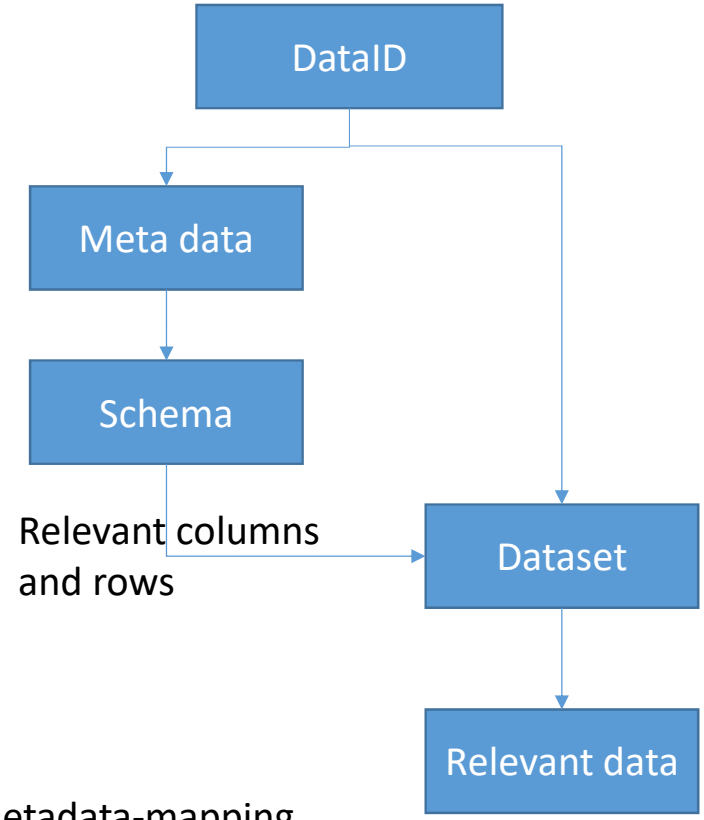
Metadata Tools for the Open Energy Platform

- Data-Wizzard for uploading data to the Open Energy Platform
  - https://openenergy-platform.org/dataedit/wizard/# -> Edit metadata
- Stand-alone tool for creating metadata
  - https://meta.rl-institut.de/meta_creator/
- Both tools are not supporting the 1.5 version yet, this still needs to be done.

Supported by: Federal Ministry for Economic Affairs and Energy on the basis of a decision by the German Bundestag
Research Grant 03EI1005A-G

FDO application example

- Idea: Use the DataID as sole reference to a dataset
- Download/data connector
  - Use the DataID to locate dataset and metadata
  - Use the schema in the metadata to derive column and row names for a data frame
  - Use the OEO for unified naming of columns and rows
  - Download the data and load it into a named data frame based on the metadata
- Named data frame can be used to extract relevant data e.g. as model input data



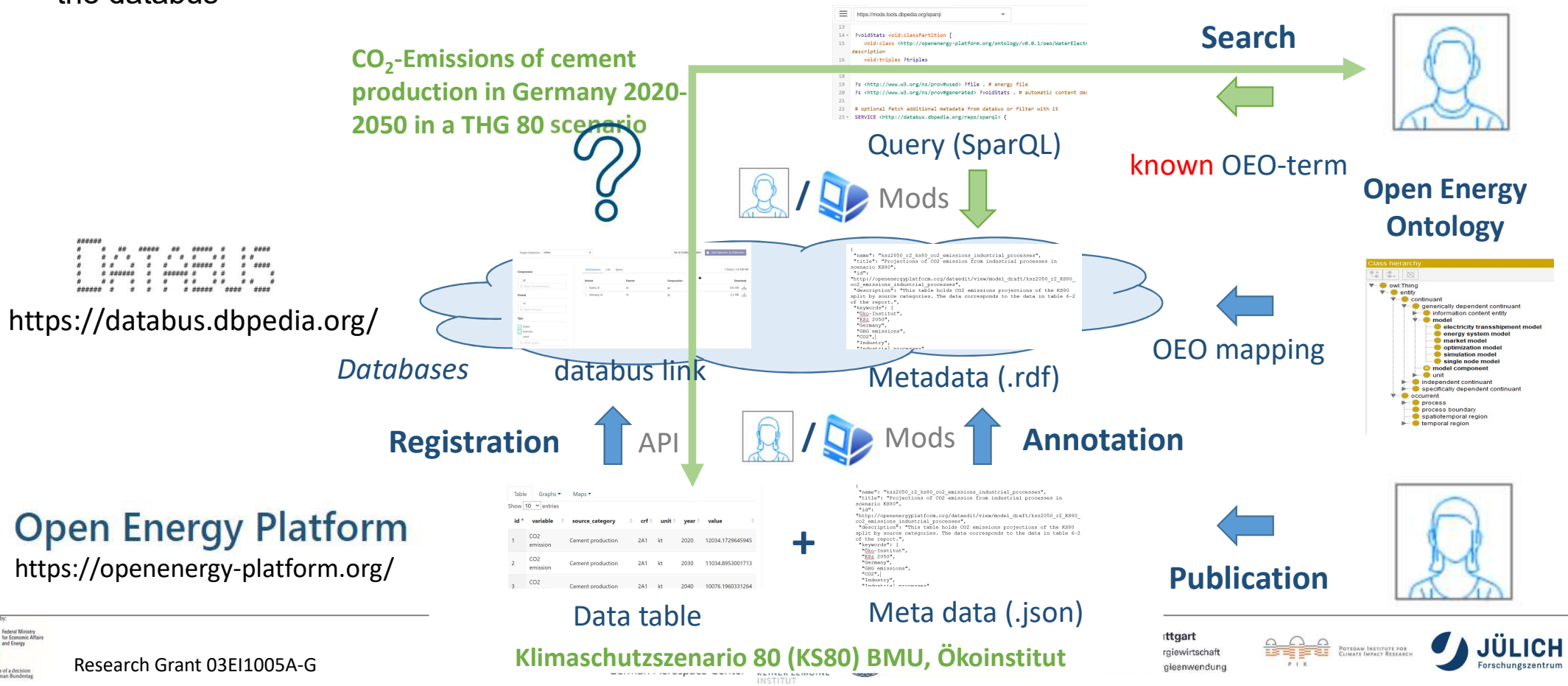
https://github.com/LOD-GEOSS/databus-snippets/tree/master/table-metadata-mapping

Research Grant 03EI1005A-G

Demonstrator: Publication of a Data Set Using the databus

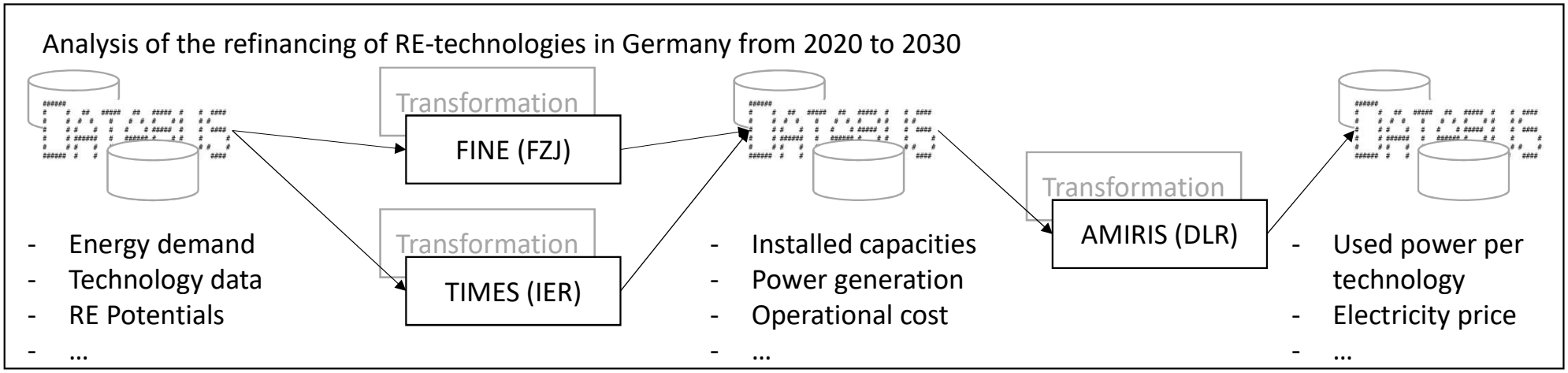
LOD-GEOSS

› **Goal**: Demonstration of the improved visibility and improved discovery of a data set through the registration in the databus



Research Grant 03EI1005A-G

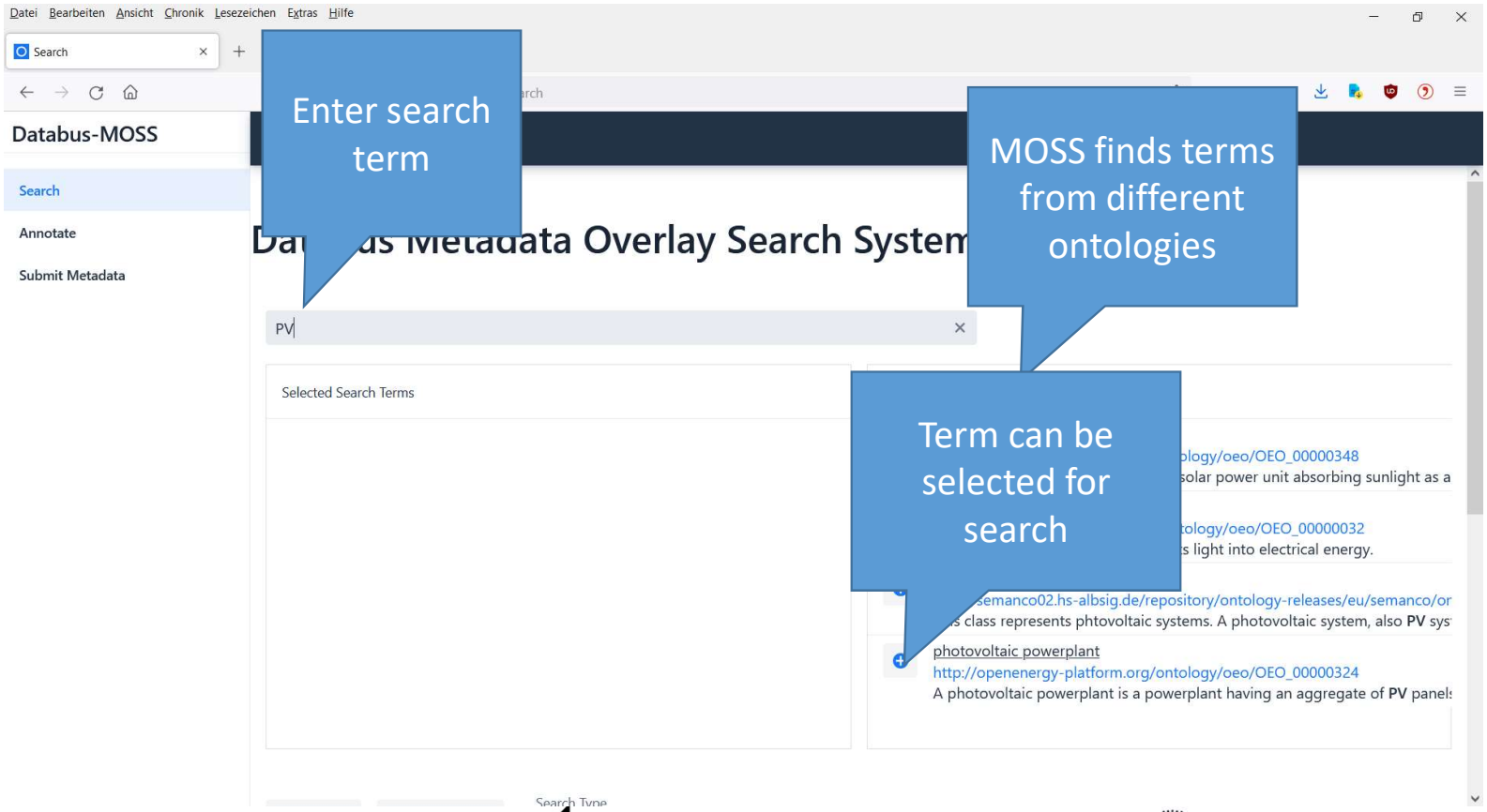
Demonstrator Model Coupling



- Goal of the demonstrator
  - Provision of a holistic application example of a databus based coupling of heterogeneous energy system models for answering a research question.
  - Databus and Metadata description as a standard interface for data between models
- Challenges in the handling of data
  - Preparation and description of data content (data and metadata standards)
  - Preparation and description of data formats (open licenses, machine readable formats, …)
  - Discovery, interpretation and further processing of the data

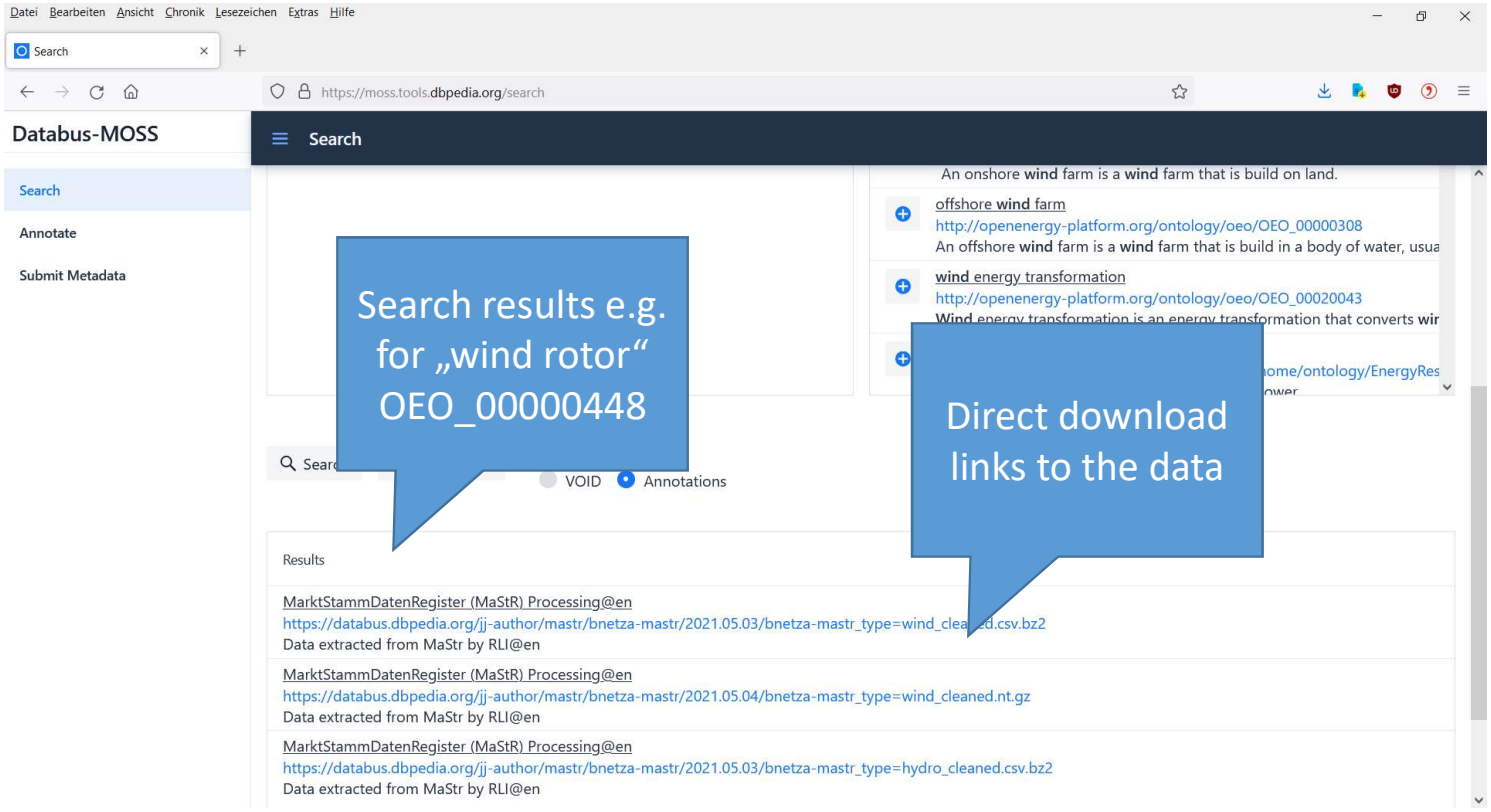
Supported by: Federal Ministry for Economic Affairs and Energy, on the basis of a decision by the German Bundestag
Research Grant 03EI1005A-G

Searching data with the MOSS



Research Grant 03EI1005A-G

Searching with MOSS



Research Grant 03EI1005A-G

Conclusion

- The databus offers
  - a service to manage and search registered metadata
  - Persistent identifies for tracing data processing and citing data
  - Databus as pointers to digital objects
- The databus can improve the data economy and scientific efficiency as we can build better on existing previous work, improve the data instead of repeating work others have done before.
- The databus supports the implementation of FAIR principles in the Domain of Energy Systems Analysis
- The developed architecture in conjunction with the use of the Open Energy Ontology enables semantic searches for data in the domain of energy systems analysis
- **The developed architecture can improve open data exchange, model coupling, tracing of workflows and collaboration for better scientific results**

- Further resources:
  - https://lod-geoss.github.io
  - https://databus.dbpedia.org
  - https://openenergy-platform.org/tutorials/jupyter/OEMetadata/
  - https://openenergy-platform.org/ontology/
  - https://doi.org/10.1016/j.egyai.2021.100074
  - https://moss.tools.dbpedia.org/search
  - https://github.com/LOD-GEOSS/databus-snippets

Supported by: Federal Ministry for Economic Affairs and Energy, on the basis of a decision by the German Bundestag
Research Grant 03EI1005A-G

LOD-GEOSS

Contact us

German Aerospace Center

Energy Systems Analysis

Carsten Hoyer-Klick

Email: carsten.hoyer-klick@dlr.de

License

Supported by:
Federal Ministry for Economic Affairs and Energy
on the basis of a decision by the German Bundestag

Research Grant 03EI1005A-G

Please cite as:

„Implementing FAIR through a distributed data infrastructure“ © Hoyer-Klick, et. Al. | CC BY 4.0

Supported by:
Federal Ministry for Economic Affairs and Energy
on the basis of a decision by the German Bundestag

Research Grant 03EI1005A-G

DLR Deutsches Zentrum für Luft- und Raumfahrt German Aerospace Center
REINER LEMOINE INSTITUT
InfAI Institute for Applied Informatics
Universität Stuttgart IER Institut für Energiewirtschaft und Rationelle Energieanwendung
PIK Potsdam Institute for Climate Impact Research
JÜLICH Forschungszentrum

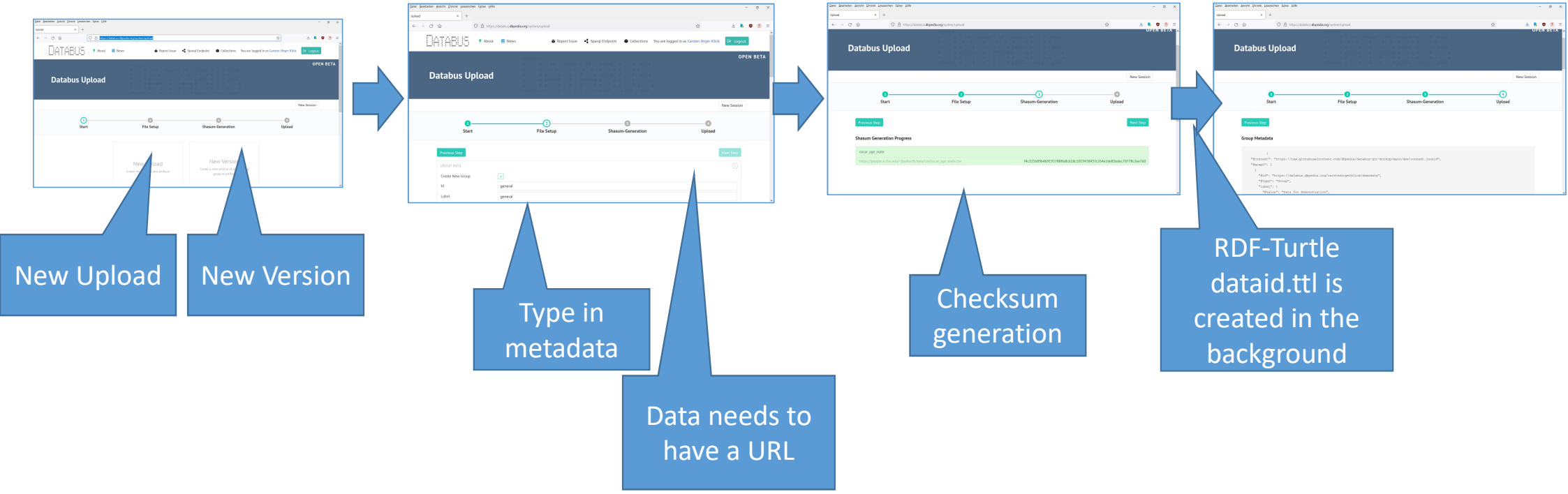
Uploading data to the Databus

https://github.com/LOD-GEOSS/databus-snippets

LOD-GEOSS

1. Using the webinterface

   https://databus.dbpedia.org/system/upload



Supported by:
Federal Ministry for Economic Affairs and Energy
on the basis of a decision by the German Bundestag

Research Grant 03EI1005A-G

DLR Deutsches Zentrum für Luft- und Raumfahrt German Aerospace Center

REINER LEMOINE INSTITUT

InfAI Institute for Applied Informatics

Universität Stuttgart IER Institut für Energiewirtschaft und Rationelle Energieanwendung

PIK Potsdam Institute for Climate Impact Research

JÜLICH Forschungszentrum

Uploading data to the Databus

https://github.com/LOD-GEOSS/databus-snippets

LOD-GEOSS

2. Using the API
    1. Create DataID in RDF-Turtle
        - › Create group-meatadata: Documentation for the group
        - › Create the actual dataID for the data with the URI to the data
    2. Fetch a Bearer Token from the databus, done by a http GET Requrest

```
TOKEN=$(curl -s -d 'client_id=upload-api' -d 'username=XXXXXX' -d
'password=XXXXXXXXXXX' -d 'grant_type=password'
https://databus.dbpedia.org/auth/realms/databus/protocol/openid-
connect/token | cut -d'"' -f 4).
```

    3. Use bearer token for deploying to the databus with http PUT Request

```
curl -H '{Authorization: Bearer $TOKEN}' -X PUT <GROUP-OR-
VERSION-URI> -d 'DATAID_JSONLD_CONTENT'
```

3. Using the python3 script
   use the `databus_api_example.py` and `dev_Databus_api_example.py` by modifying the variables below if__NAME==“__main__”

Supported by: Federal Ministry for Economic Affairs and Energy, on the basis of a decision by the German Bundestag
Research Grant 03EI1005A-G

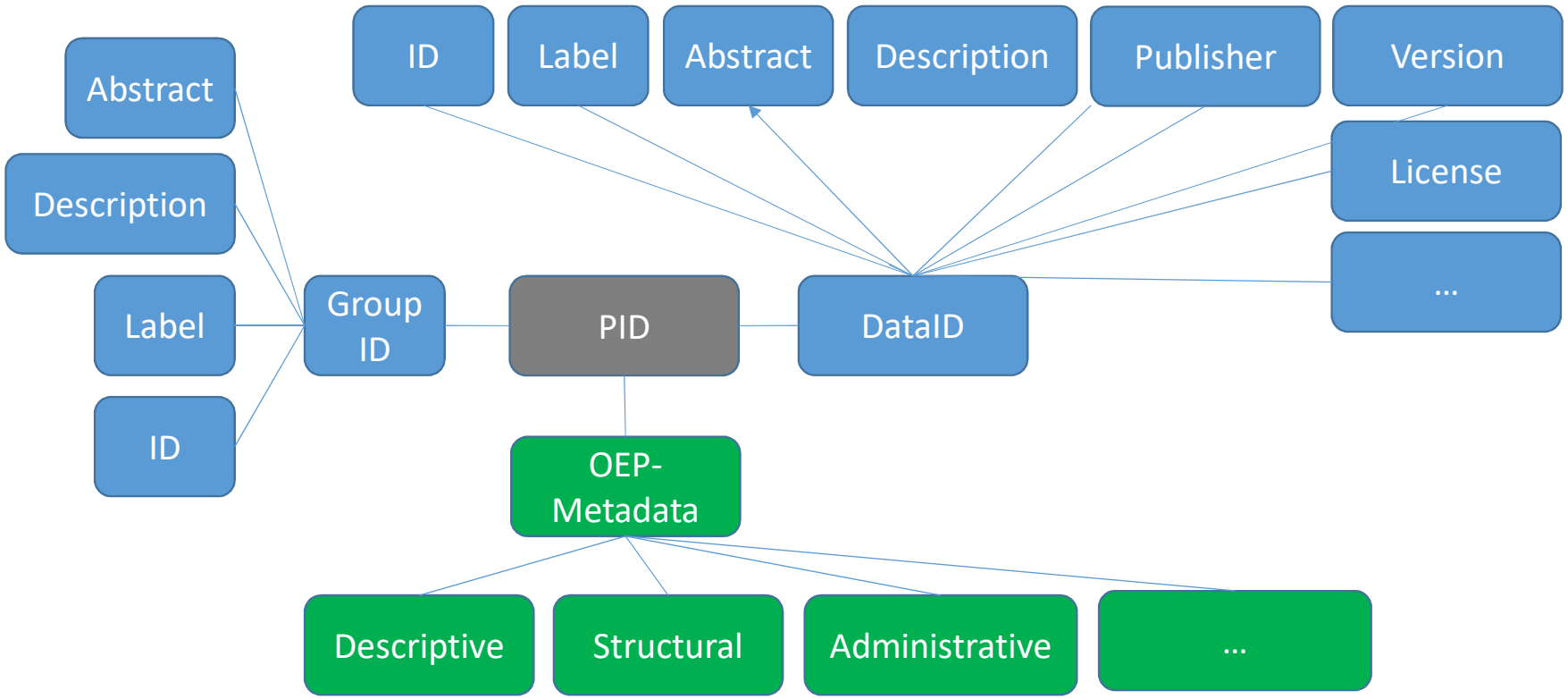
Contents of GroupID and DataID

- GroupID
  - ID
  - Label
  - Abstract
  - Description
- DataID
  - ID
  - Label
  - Title
  - Publisher (autogenerated)
  - Artifact (autogenerated)
  - Group (autogenerated)
  - Abstract
  - Version
  - License
  - License URI
  - License Documentation
  - Files (URIs to the data)

Supported by: Federal Ministry for Economic Affairs and Energy on the basis of a decision by the German Bundestag

Research Grant 03EI1005A-G

Extending the Graph by Additional Metadata



Research Grant 03EI1005A-G

Annotation of data: MOSS Metadata Overlay Search System

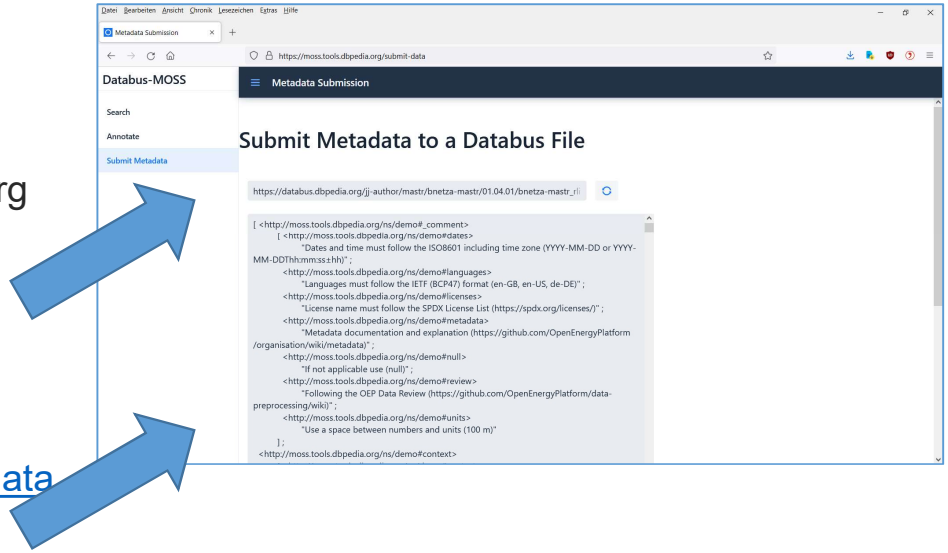
- Only basic metadata (license, publisher, etc) is used for the registration of data on the Databus
- **Databus Mods** (an additional software layer) are used to **add additional information** to a data set
- **Additional metadata can be added to any data set following different domain standards via databus mods**

- MOSS: Metadata Overlay Search System: moss.tools.dbpedia.org

1. Generate Metadata in JSON-LD (e.g. the updated OEP metadata string)

2. Find the URI of the dataset and enter it into the MOSS submission form at: https://moss.tools.dbpedia.org/submit-data

3. Submitt the Metadata by pasting it into the large textfield



Research Grant 03EI1005A-G