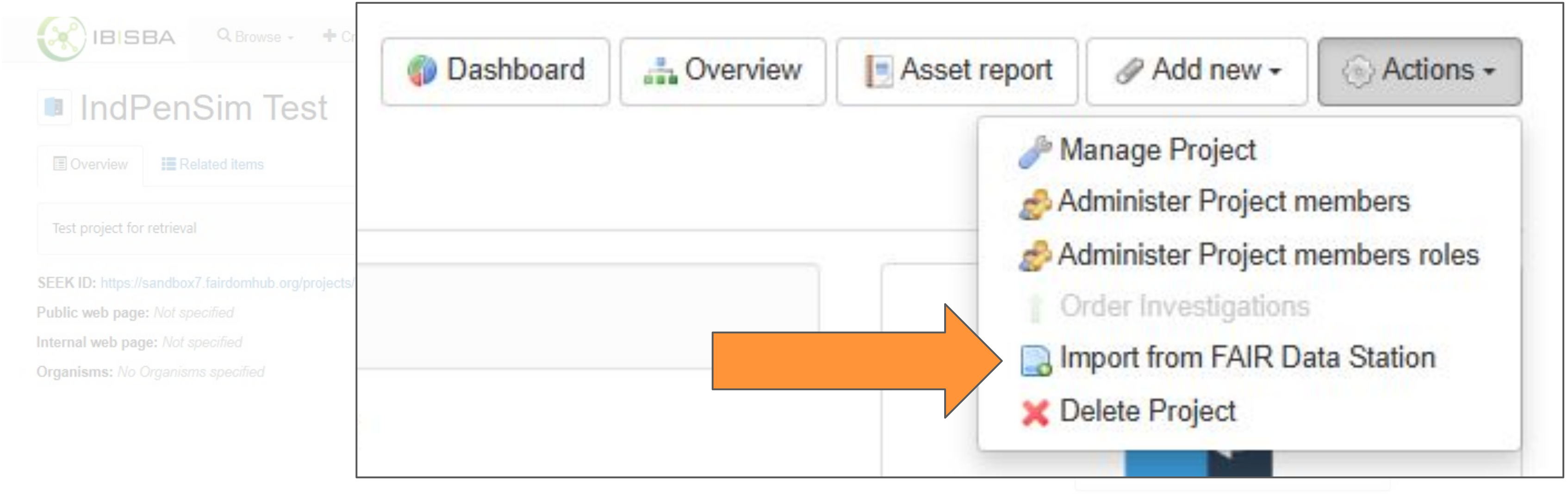
FAIR-DS for FAIRDOM-SEEK

Jasper Koehorst | Brett Metcalfe

FAIRDOM-SEEK Meeting 16th September 2025

Funded by the European Union

Aim



IndPenSim Test

Test project for retrieval

SEEK ID: https://sandbox7.fairdomhub.org/projects/

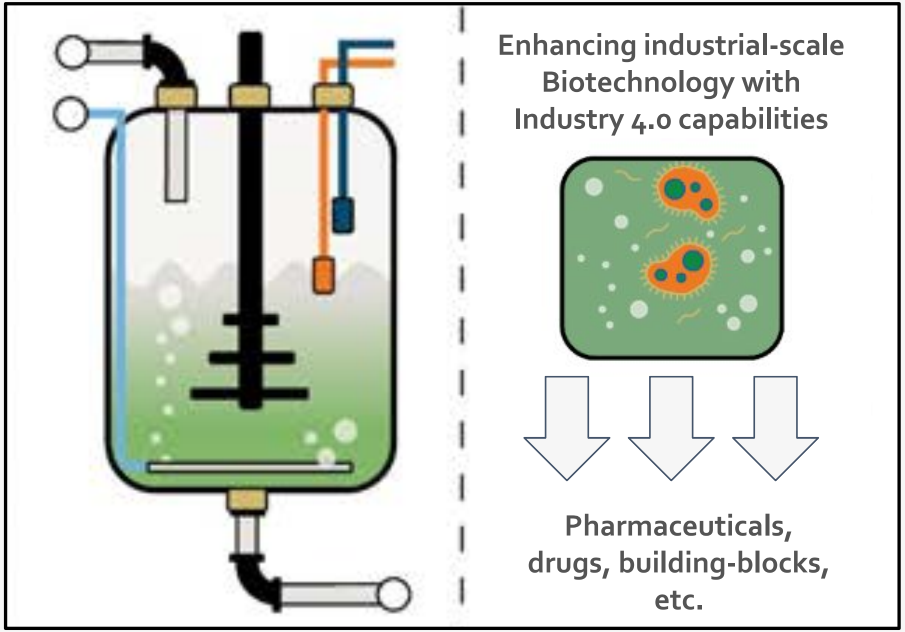
Public web page: *Not specified*

Internal web page: *Not specified*

Organisms: *No Organisms specified*

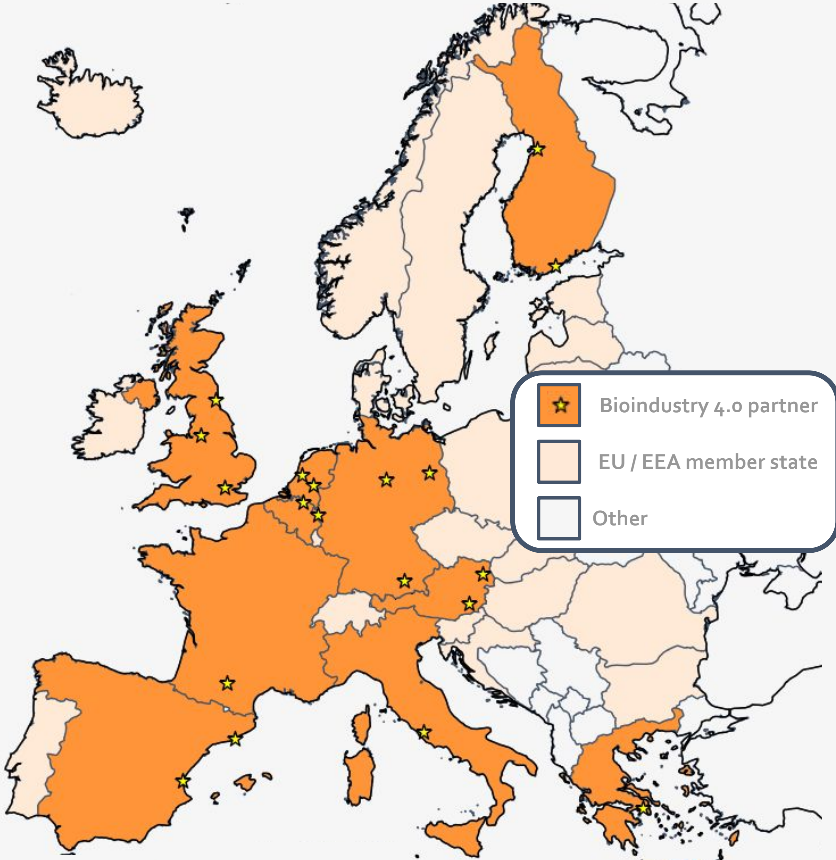
To talk about this new button: explain what FAIR Data Station is; and the integration of FAIR DS with FAIRDOM-SEEK. But first…

BioIndustry 4.0



Enhancing industrial-scale Biotechnology with Industry 4.0 capabilities

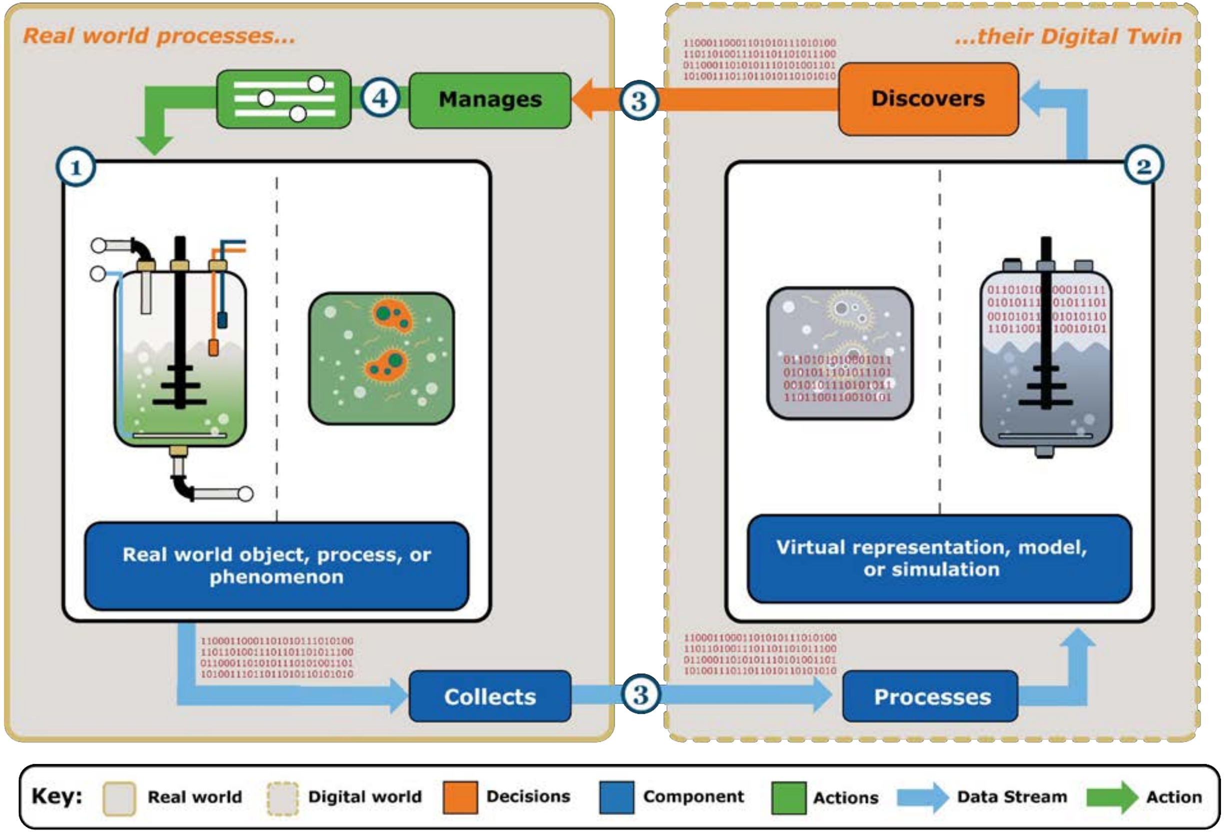
Pharmaceuticals, drugs, building-blocks, etc.



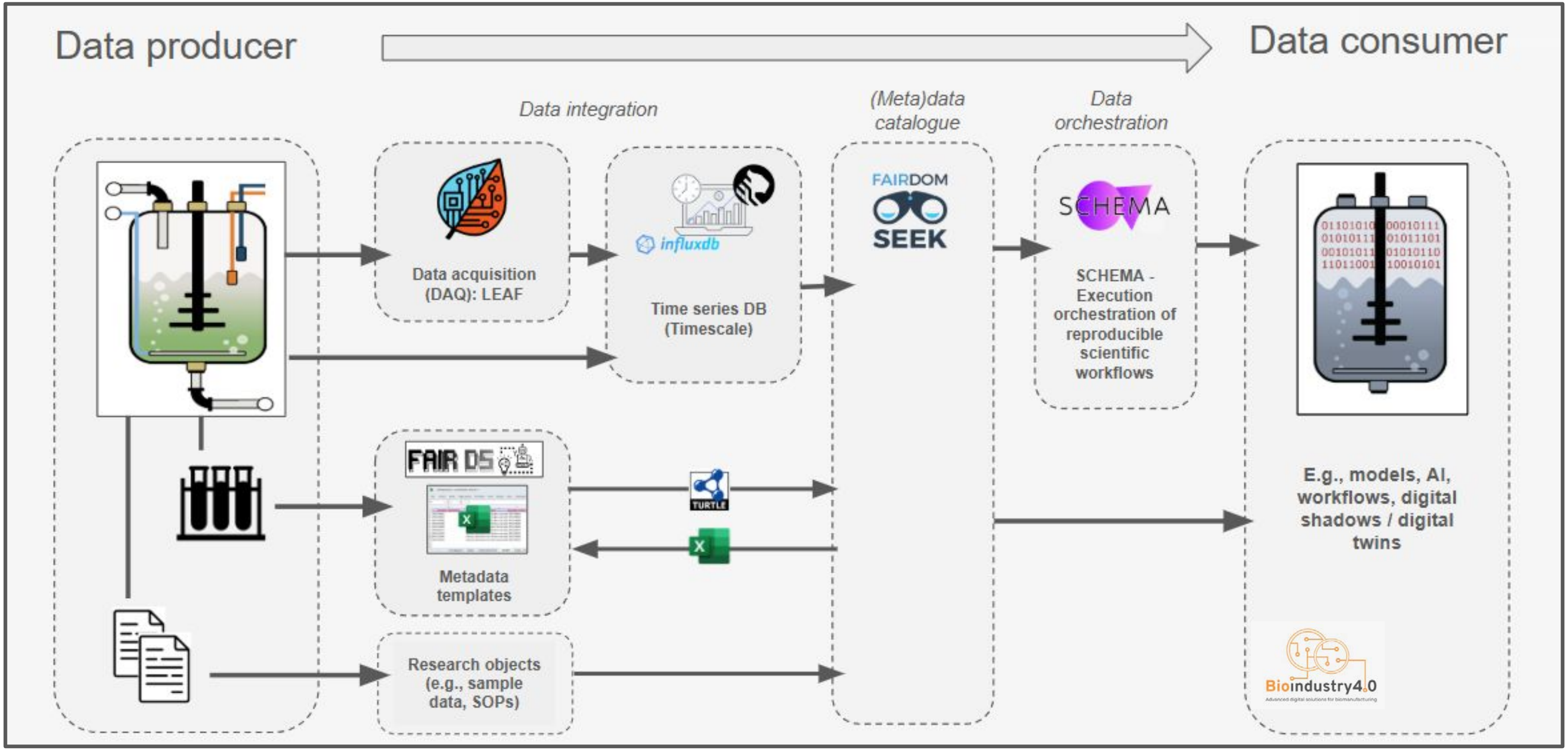
Bioindustry4.0
Advanced digital solutions for biomanufacturing

Funded by
the European Union

Digital twin



Data Fabric



Data producer

Data consumer

Data integration

(Meta)data catalogue

Data orchestration

Data acquisition (DAQ): LEAF

Time series DB (Timescale)

SCHEMA - Execution orchestration of reproducible scientific workflows

Metadata templates

Research objects (e.g., sample data, SOPs)

E.g., models, AI, workflows, digital shadows / digital twins

For a Digital Twin to work

- Live interaction with laboratory equipment
- Live stream of information
- **Data**
    - **Findable**
    - **Accessible**
    - **Interoperable**
    - **Reusable**

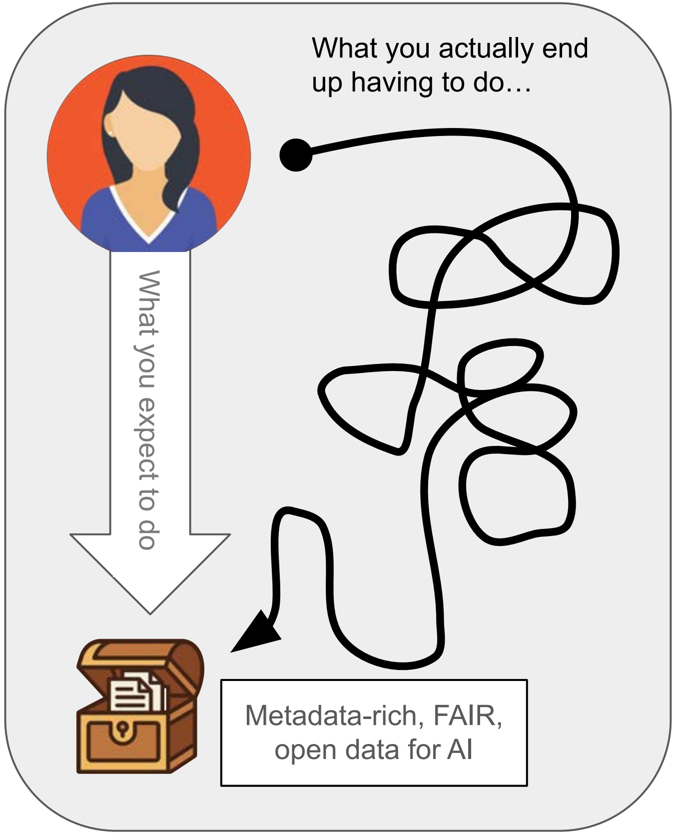
The problem: data management isn't seamless

Or 'why do I, a data producer, have to spend time…'

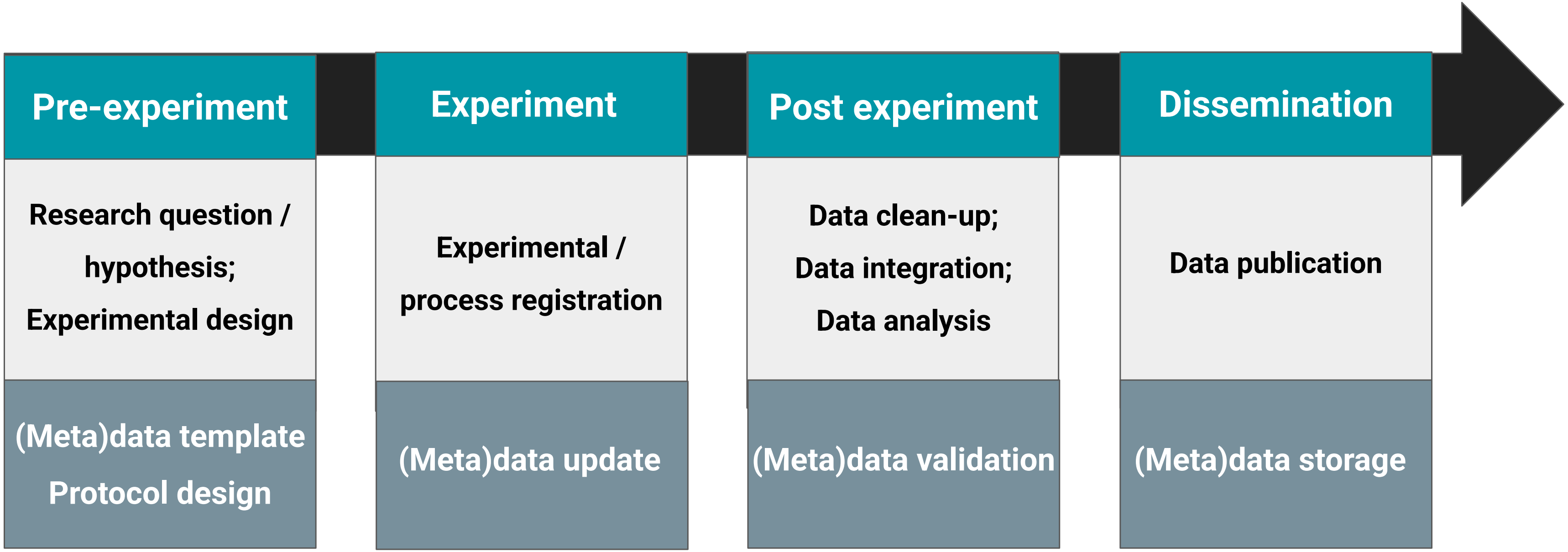
- '…cleaning data produced by my laboratory equipment.'
- '…integrating all the different datasets.'
- '…finding, formatting, standardising, uploading, and storing my data.'

I want…

- `…something accessible, easy to use, and not be bothered about FAIR'



FAIR by design approach



Pre-experiment	Experiment	Post experiment	Dissemination
Research question / hypothesis; Experimental design	Experimental / process registration	Data clean-up; Data integration; Data analysis	Data publication
(Meta)data template Protocol design	(Meta)data update	(Meta)data validation	(Meta)data storage

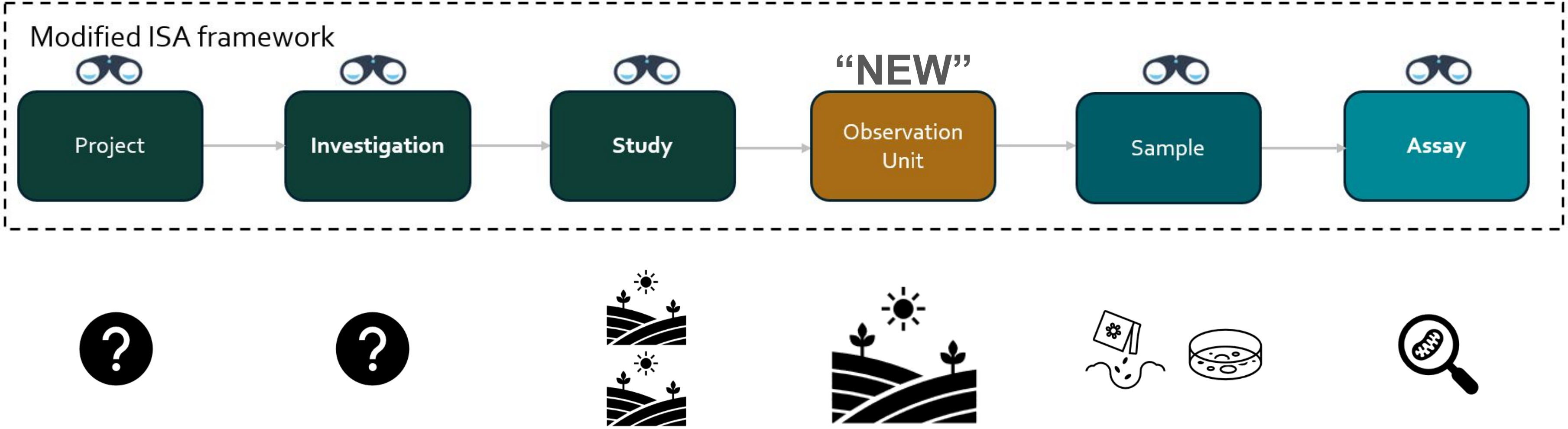
‘**FAIR by design**’ means recording, storing, and validating (meta)data as you go to facilitate the researcher rather than ‘**FAIR for publication**’, i.e., just before dissemination (for someone else).

FAIR by ‘accident’

Ideally we want it so that users ‘accidentally’ make their data FAIR from the outset with each step continuing this trend.

Pre-experiment	Experiment	Post experiment	Dissemination
Research question / hypothesis; Experimental design	Experimental / process registration	Data clean-up; Data integration; Data analysis	Data publication
(Meta)data template Protocol design	(Meta)data update	(Meta)data validation	(Meta)data storage

Data structure



Modified ISA framework

Project → Investigation → Study → Observation Unit → Sample → Assay

“NEW”

Research context

Research context and/or hypothesis

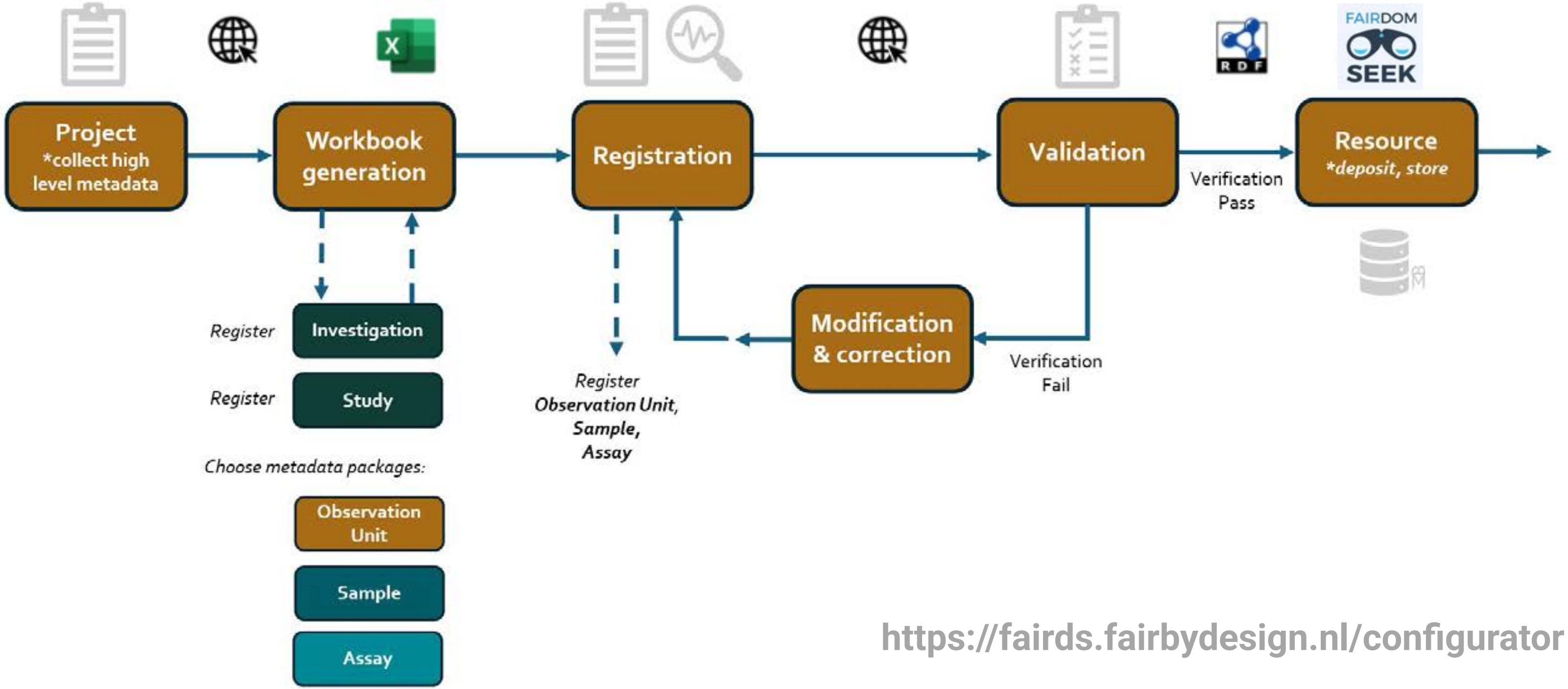
Unit of research (e.g., control vs contaminated sites)

Contaminated site

Soil sample

Analysis (e.g., microbial diversity)

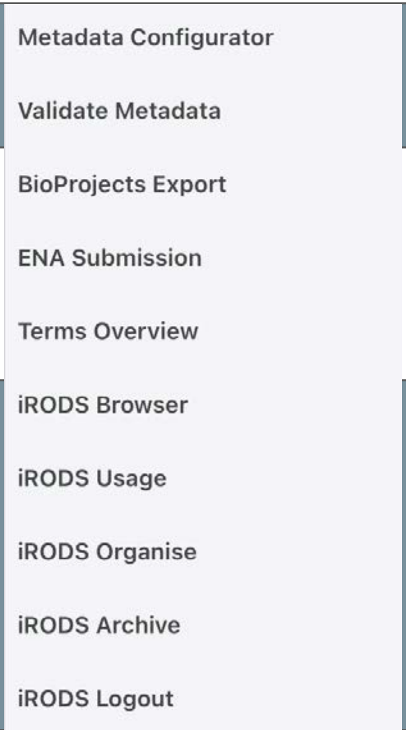
FAIR Data Station process



https://fairds.fairbydesign.nl/configurator

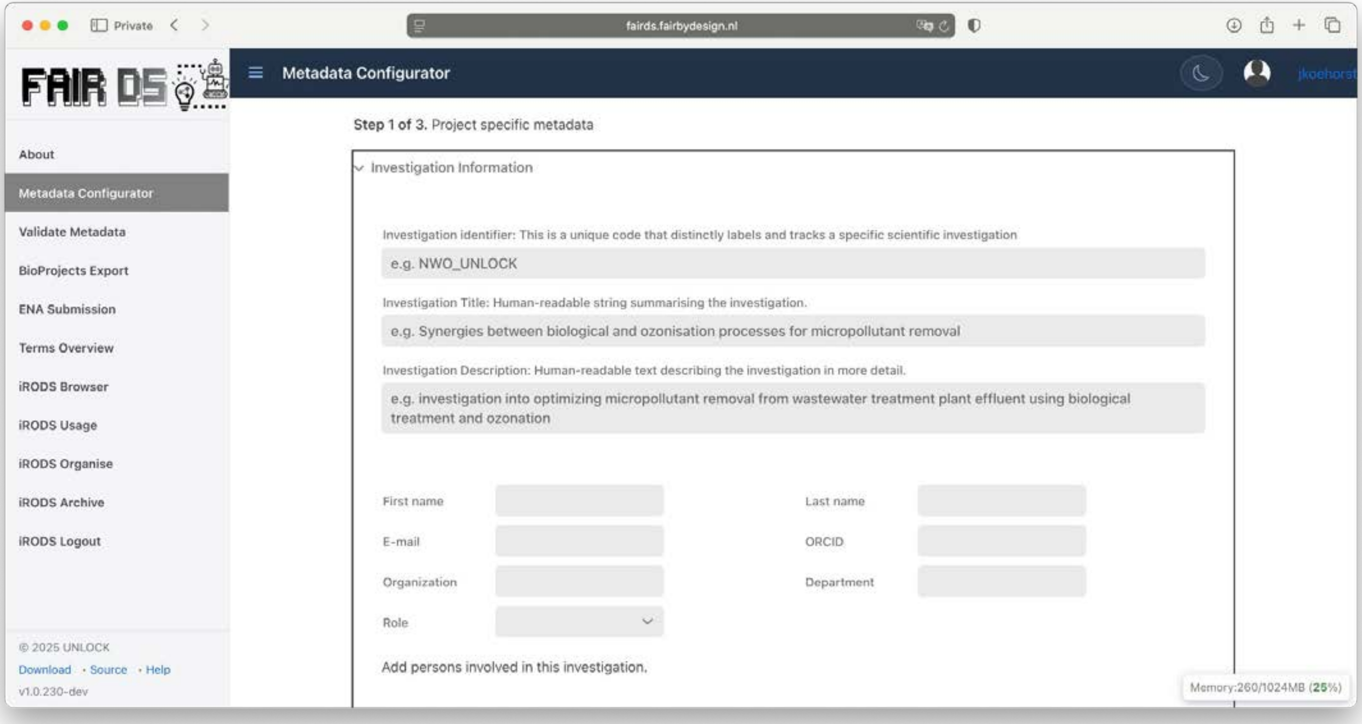
Features implemented

Metadata & validation



Data storage, organisation, and archiving

Metadata configurator: Investigation, Study



Step 1 of 3. Project specific metadata

Investigation Information

Investigation identifier: This is a unique code that distinctly labels and tracks a specific scientific investigation

e.g. NWO_UNLOCK

Investigation Title: Human-readable string summarising the investigation.

e.g. Synergies between biological and ozonisation processes for micropollutant removal

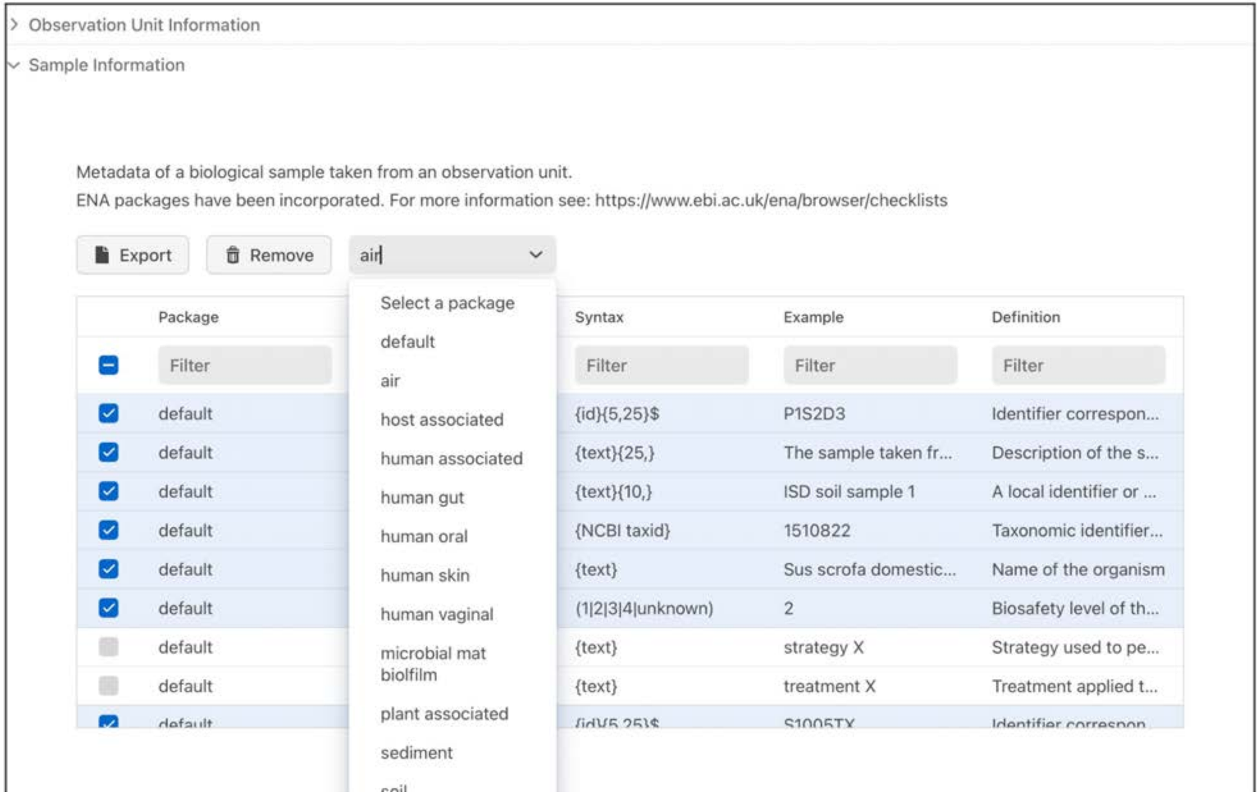
Investigation Description: Human-readable text describing the investigation in more detail.

e.g. investigation into optimizing micropollutant removal from wastewater treatment plant effluent using biological treatment and ozonation

First name		Last name	
E-mail		ORCID	
Organization		Department	
Role			

Add persons involved in this investigation.

Observation unit / Sample / Assay > Packages!

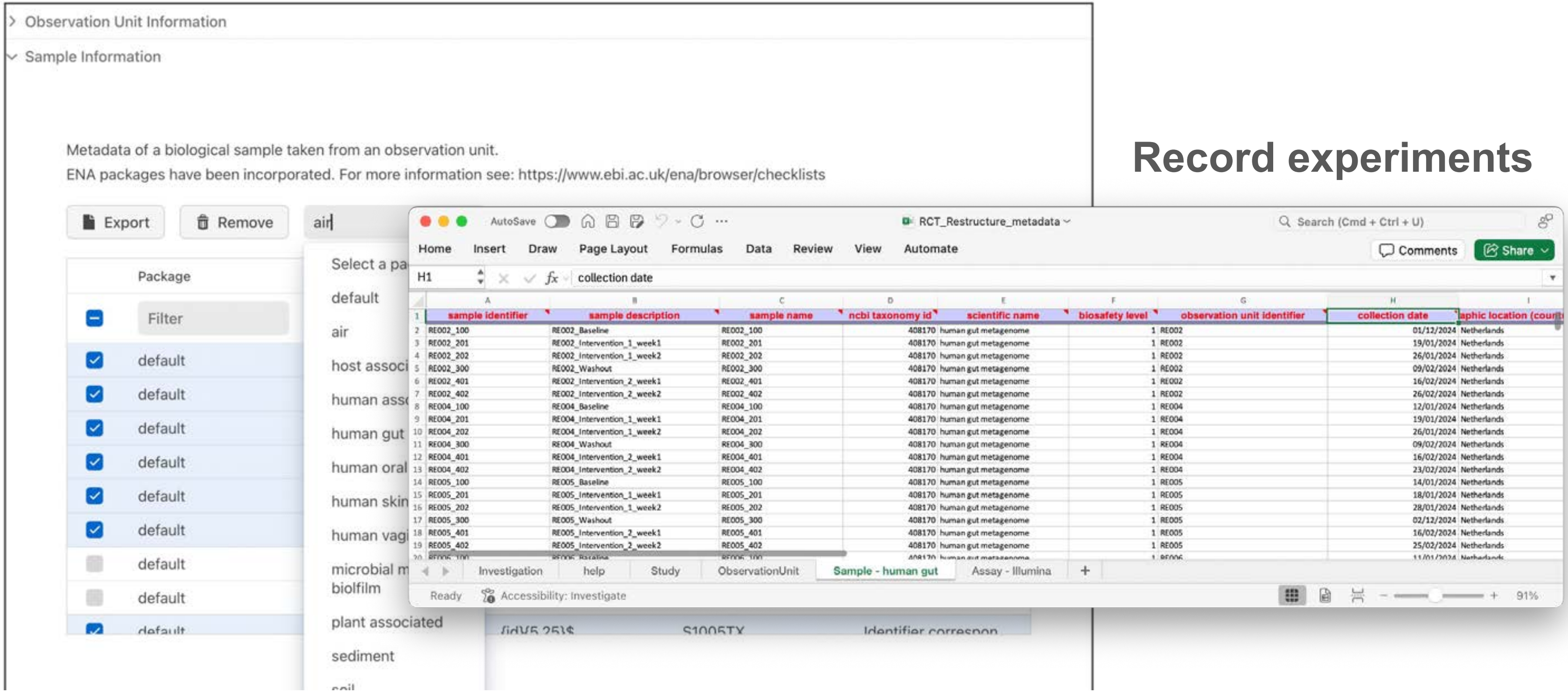


Metadata of a biological sample taken from an observation unit.
ENA packages have been incorporated. For more information see: https://www.ebi.ac.uk/ena/browser/checklists

	Package	Syntax	Example	Definition
⊟	Filter	Filter	Filter	Filter
☑	default	{id}{5,25}$	P1S2D3	Identifier correspon...
☑	default	{text}{25,}	The sample taken fr...	Description of the s...
☑	default	{text}{10,}	ISD soil sample 1	A local identifier or ...
☑	default	{NCBI taxid}	1510822	Taxonomic identifier...
☑	default	{text}	Sus scrofa domestic...	Name of the organism
☑	default	(1|2|3|4|unknown)	2	Biosafety level of th...
☐	default	{text}	strategy X	Strategy used to pe...
☐	default	{text}	treatment X	Treatment applied t...
☑	default	{id}{5,25}$	S1005TX	Identifier correspon

Observation unit / Sample / Assay > Packages!

Record experiments



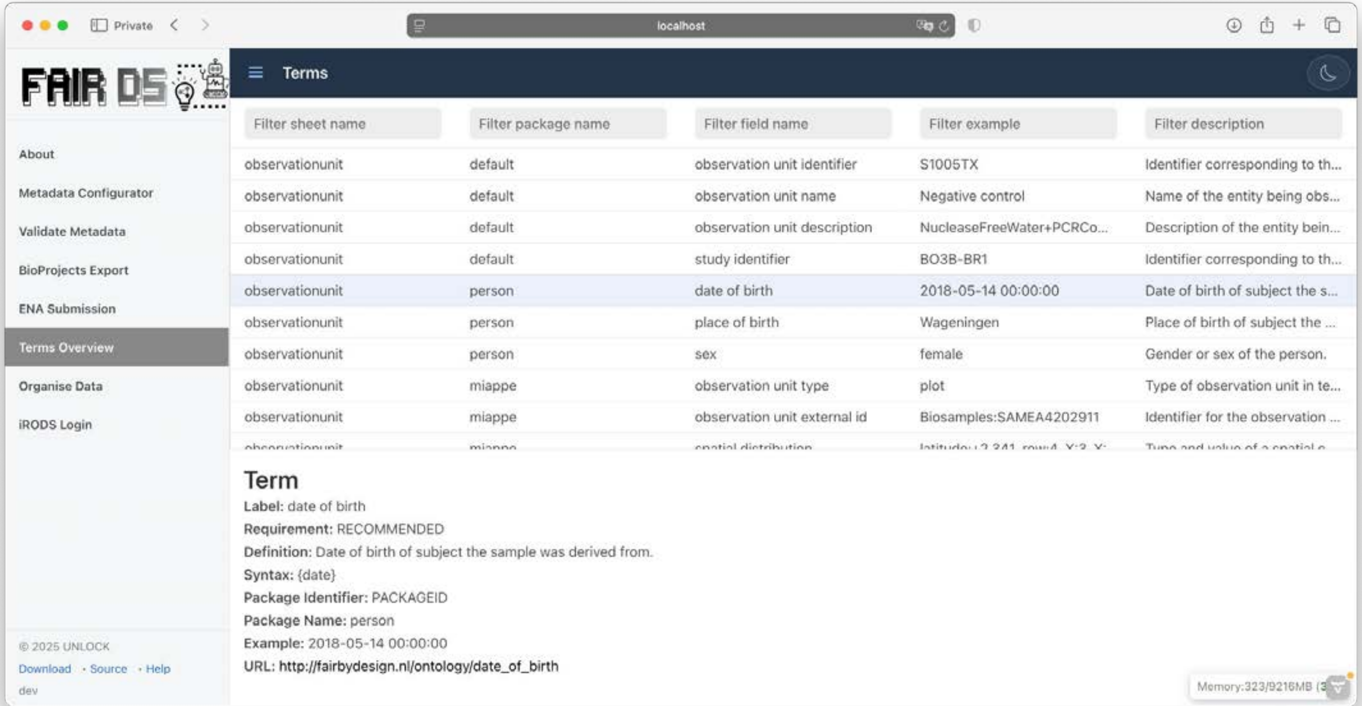
Metadata of a biological sample taken from an observation unit.
ENA packages have been incorporated. For more information see: https://www.ebi.ac.uk/ena/browser/checklists

	Package
☑	default
☑	default
☑	default
☑	default
☑	default
☑	default
☐	default
☐	default

collection date

sample identifier	sample description	sample name	ncbi taxonomy id	scientific name	biosafety level	observation unit identifier	collection date	aphic location (count
RE002_100	RE002_Baseline	RE002_100	408170	human gut metagenome	1	RE002	01/12/2024	Netherlands
RE002_201	RE002_Intervention_1_week1	RE002_201	408170	human gut metagenome	1	RE002	19/01/2024	Netherlands
RE002_202	RE002_Intervention_1_week2	RE002_202	408170	human gut metagenome	1	RE002	26/01/2024	Netherlands
RE002_300	RE002_Washout	RE002_300	408170	human gut metagenome	1	RE002	09/02/2024	Netherlands
RE002_401	RE002_Intervention_2_week1	RE002_401	408170	human gut metagenome	1	RE002	16/02/2024	Netherlands
RE002_402	RE002_Intervention_2_week2	RE002_402	408170	human gut metagenome	1	RE002	26/02/2024	Netherlands
RE004_100	RE004_Baseline	RE004_100	408170	human gut metagenome	1	RE004	12/01/2024	Netherlands
RE004_201	RE004_Intervention_1_week1	RE004_201	408170	human gut metagenome	1	RE004	19/01/2024	Netherlands
RE004_202	RE004_Intervention_1_week2	RE004_202	408170	human gut metagenome	1	RE004	26/01/2024	Netherlands
RE004_300	RE004_Washout	RE004_300	408170	human gut metagenome	1	RE004	09/02/2024	Netherlands
RE004_401	RE004_Intervention_2_week1	RE004_401	408170	human gut metagenome	1	RE004	16/02/2024	Netherlands
RE004_402	RE004_Intervention_2_week2	RE004_402	408170	human gut metagenome	1	RE004	23/02/2024	Netherlands
RE005_100	RE005_Baseline	RE005_100	408170	human gut metagenome	1	RE005	14/01/2024	Netherlands
RE005_201	RE005_Intervention_1_week1	RE005_201	408170	human gut metagenome	1	RE005	18/01/2024	Netherlands
RE005_202	RE005_Intervention_1_week2	RE005_202	408170	human gut metagenome	1	RE005	28/01/2024	Netherlands
RE005_300	RE005_Washout	RE005_300	408170	human gut metagenome	1	RE005	02/12/2024	Netherlands
RE005_401	RE005_Intervention_2_week1	RE005_401	408170	human gut metagenome	1	RE005	16/02/2024	Netherlands
RE005_402	RE005_Intervention_2_week2	RE005_402	408170	human gut metagenome	1	RE005	25/02/2024	Netherlands

Metadata lookup service



Filter sheet name	Filter package name	Filter field name	Filter example	Filter description
observationunit	default	observation unit identifier	S1005TX	Identifier corresponding to th...
observationunit	default	observation unit name	Negative control	Name of the entity being obs...
observationunit	default	observation unit description	NucleaseFreeWater+PCRCo...	Description of the entity bein...
observationunit	default	study identifier	BO3B-BR1	Identifier corresponding to th...
observationunit	person	date of birth	2018-05-14 00:00:00	Date of birth of subject the s...
observationunit	person	place of birth	Wageningen	Place of birth of subject the ...
observationunit	person	sex	female	Gender or sex of the person.
observationunit	miappe	observation unit type	plot	Type of observation unit in te...
observationunit	miappe	observation unit external id	Biosamples:SAMEA4202911	Identifier for the observation ...
observationunit	miappe	spatial distribution	latitude:+2.341, row:4, X:3, Y:	Type and value of a spatial c

Term

Label: date of birth

Requirement: RECOMMENDED

Definition: Date of birth of subject the sample was derived from.

Syntax: {date}

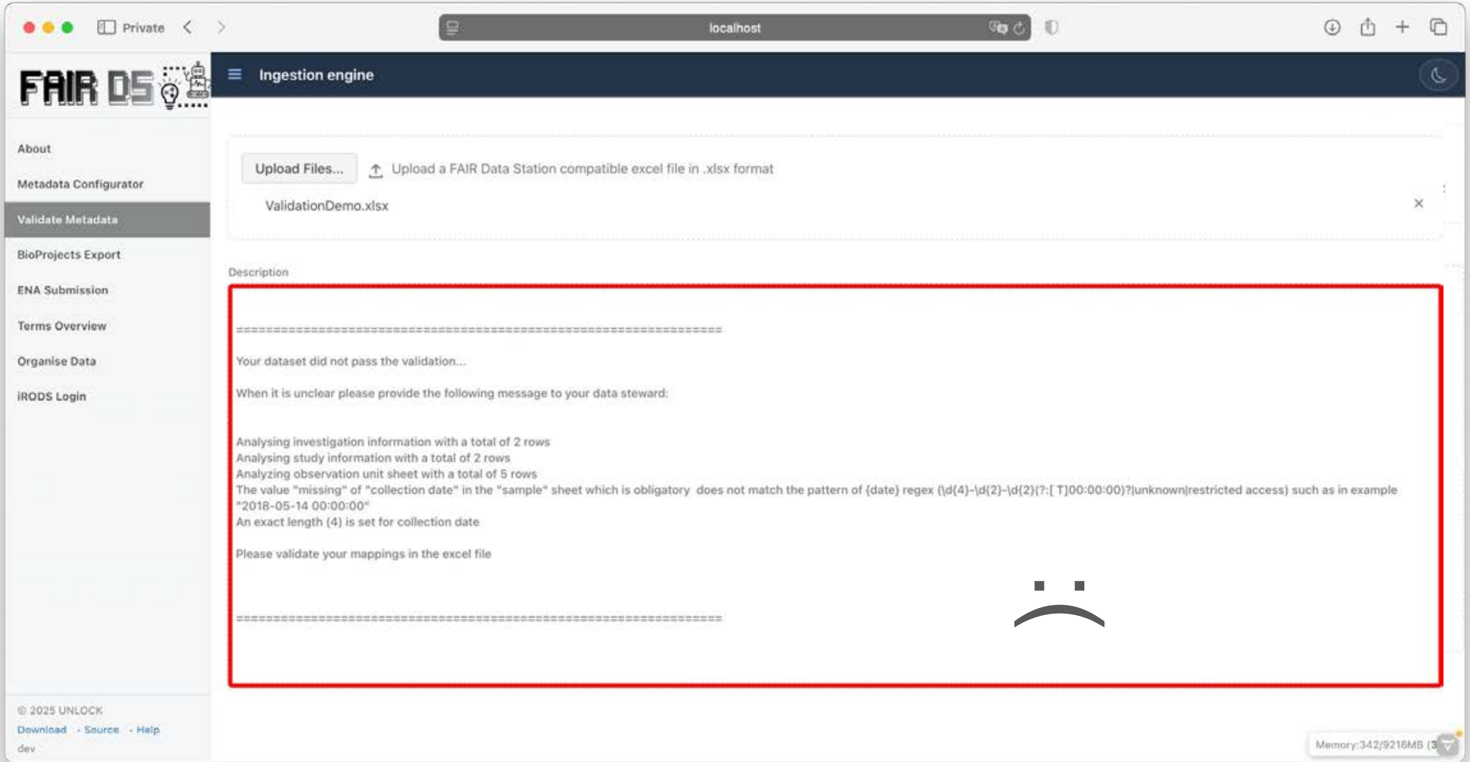
Package Identifier: PACKAGEID

Package Name: person

Example: 2018-05-14 00:00:00

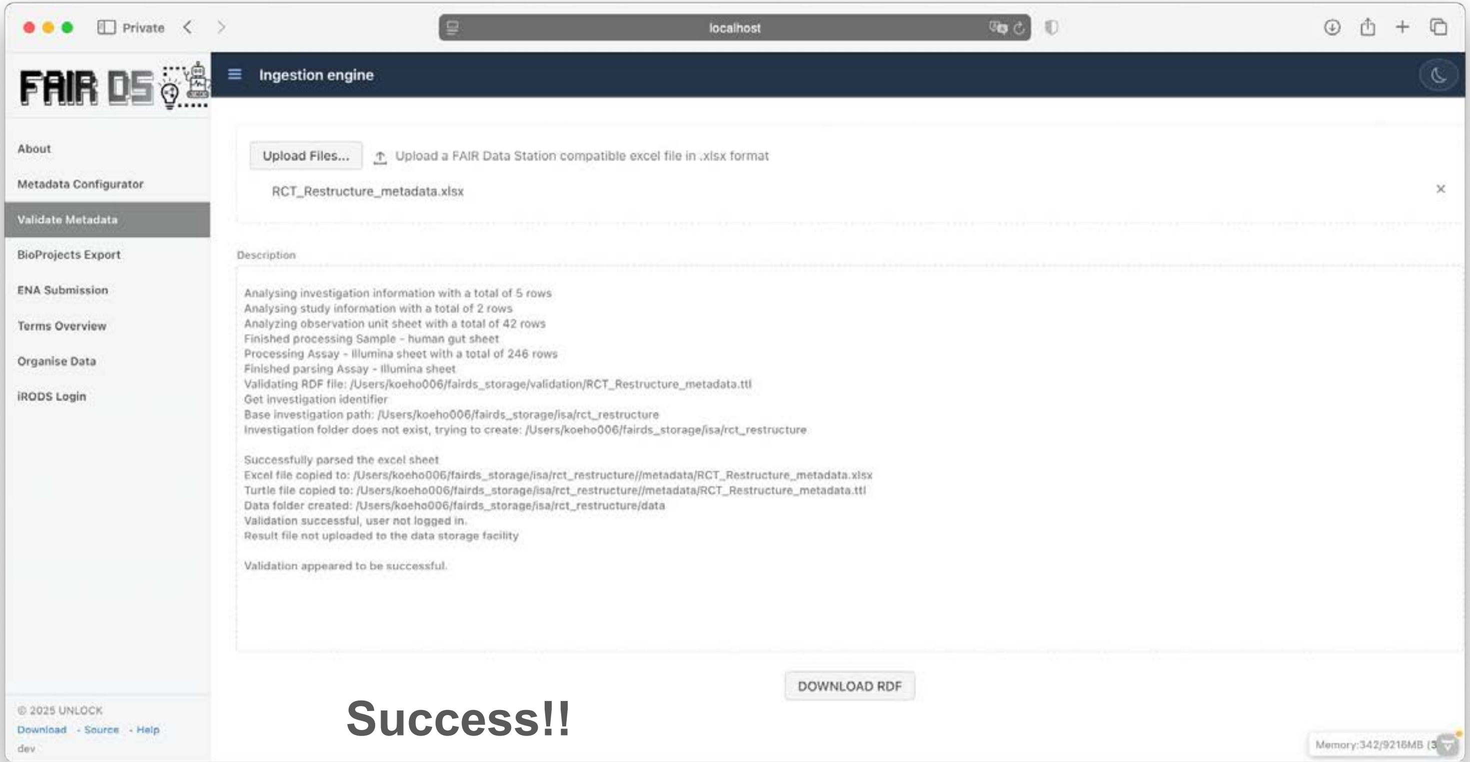
URL: http://fairbydesign.nl/ontology/date_of_birth

Metadata validation



```
====================================================================

Your dataset did not pass the validation...

When it is unclear please provide the following message to your data steward:


Analysing investigation information with a total of 2 rows
Analysing study information with a total of 2 rows
Analyzing observation unit sheet with a total of 5 rows
The value "missing" of "collection date" in the "sample" sheet which is obligatory  does not match the pattern of {date} regex (\d{4}-\d{2}-\d{2}(?:[ T]00:00:00)?|unknown|restricted access) such as in example "2018-05-14 00:00:00"
An exact length (4) is set for collection date

Please validate your mappings in the excel file


====================================================================
```

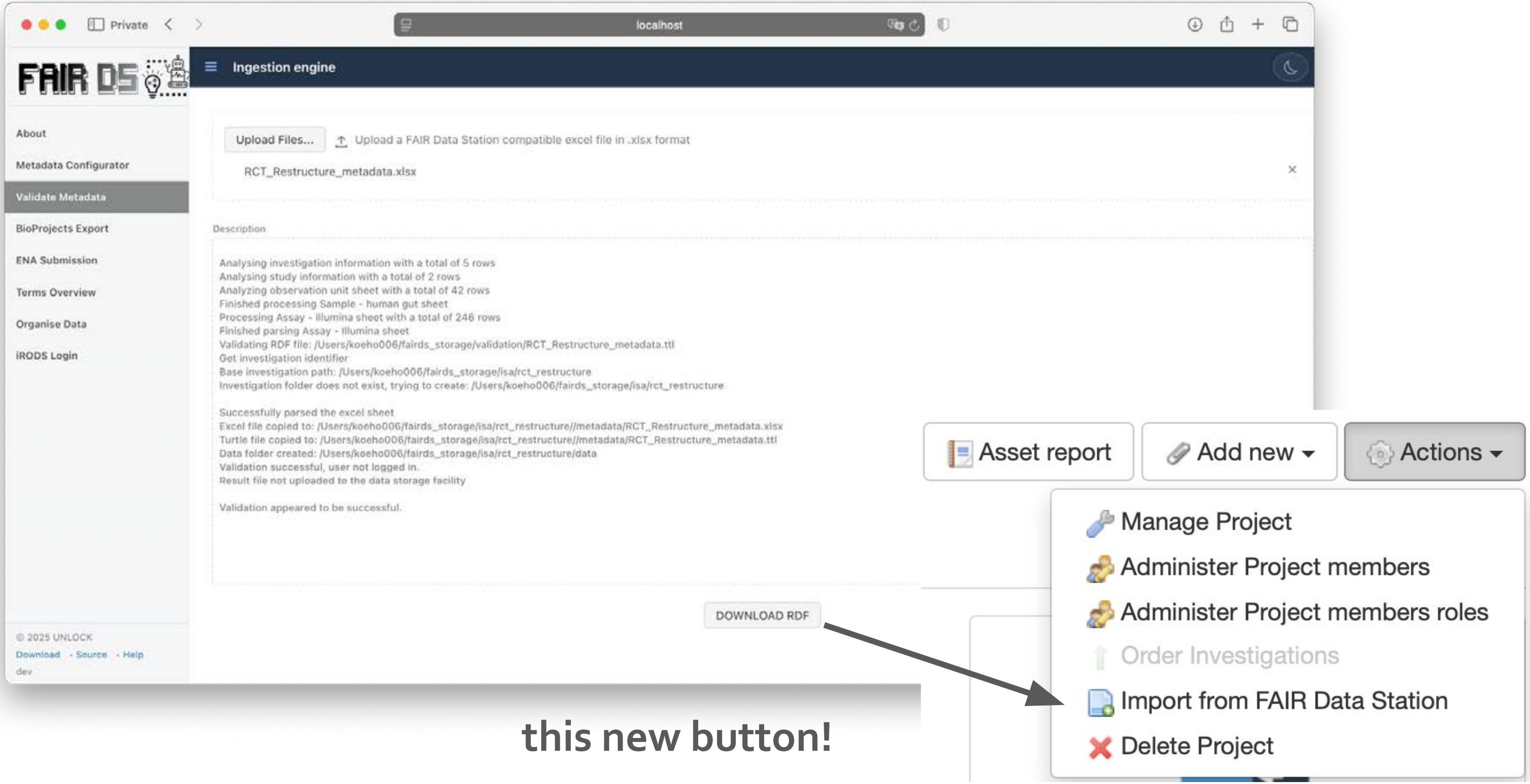
Metadata validation



```
Analysing investigation information with a total of 5 rows
Analysing study information with a total of 2 rows
Analyzing observation unit sheet with a total of 42 rows
Finished processing Sample - human gut sheet
Processing Assay - Illumina sheet with a total of 246 rows
Finished parsing Assay - Illumina sheet
Validating RDF file: /Users/koeho006/fairds_storage/validation/RCT_Restructure_metadata.ttl
Get investigation identifier
Base investigation path: /Users/koeho006/fairds_storage/isa/rct_restructure
Investigation folder does not exist, trying to create: /Users/koeho006/fairds_storage/isa/rct_restructure

Successfully parsed the excel sheet
Excel file copied to: /Users/koeho006/fairds_storage/isa/rct_restructure//metadata/RCT_Restructure_metadata.xlsx
Turtle file copied to: /Users/koeho006/fairds_storage/isa/rct_restructure//metadata/RCT_Restructure_metadata.ttl
Data folder created: /Users/koeho006/fairds_storage/isa/rct_restructure/data
Validation successful, user not logged in.
Result file not uploaded to the data storage facility

Validation appeared to be successful.
```

Success!!

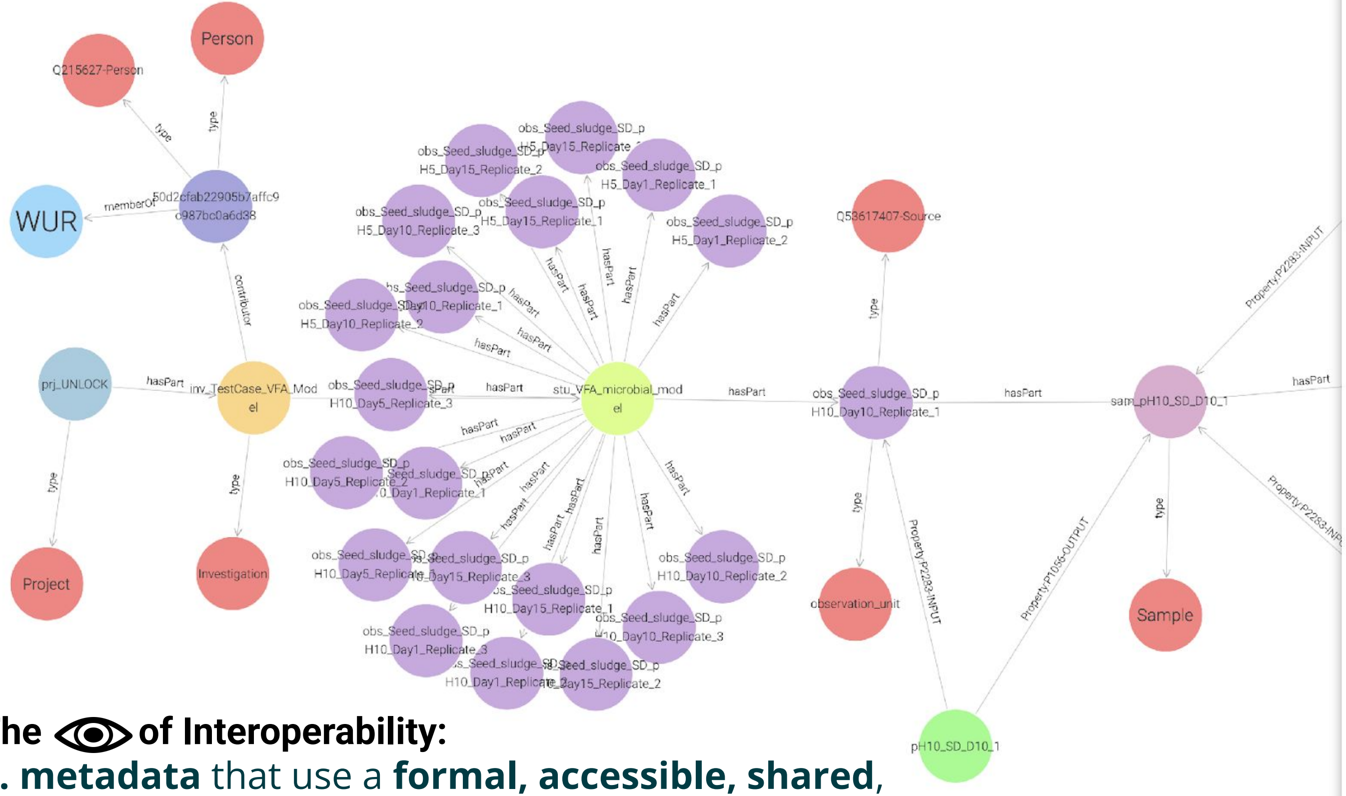
Metadata validation

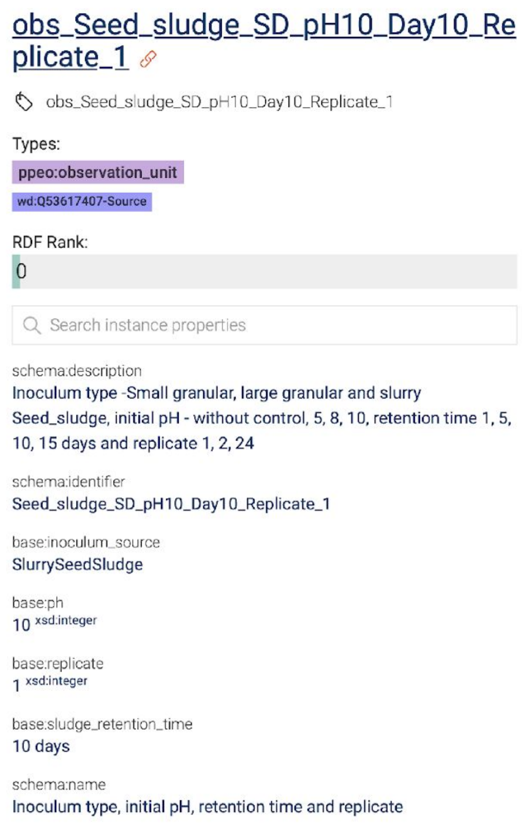


```
Analysing investigation information with a total of 5 rows
Analysing study information with a total of 2 rows
Analyzing observation unit sheet with a total of 42 rows
Finished processing Sample - human gut sheet
Processing Assay - Illumina sheet with a total of 246 rows
Finished parsing Assay - Illumina sheet
Validating RDF file: /Users/koeho006/fairds_storage/validation/RCT_Restructure_metadata.ttl
Get investigation identifier
Base investigation path: /Users/koeho006/fairds_storage/isa/rct_restructure
Investigation folder does not exist, trying to create: /Users/koeho006/fairds_storage/isa/rct_restructure

Successfully parsed the excel sheet
Excel file copied to: /Users/koeho006/fairds_storage/isa/rct_restructure//metadata/RCT_Restructure_metadata.xlsx
Turtle file copied to: /Users/koeho006/fairds_storage/isa/rct_restructure//metadata/RCT_Restructure_metadata.ttl
Data folder created: /Users/koeho006/fairds_storage/isa/rct_restructure/data
Validation successful, user not logged in.
Result file not uploaded to the data storage facility

Validation appeared to be successful.
```

this new button!

So what is this “RDF” File?





obs_Seed_sludge_SD_pH10_Day10_Replicate_1

obs_Seed_sludge_SD_pH10_Day10_Replicate_1

Types:

ppeo:observation_unit

wd:Q53617407-Source

RDF Rank:

0

Search instance properties

schema:description
Inoculum type -Small granular, large granular and slurry Seed_sludge, initial pH - without control, 5, 8, 10, retention time 1, 5, 10, 15 days and replicate 1, 2, 24

schema:identifier
Seed_sludge_SD_pH10_Day10_Replicate_1

base:inoculum_source
SlurrySeedSludge

base:ph
10 xsd:integer

base:replicate
1 xsd:integer

base:sludge_retention_time
10 days

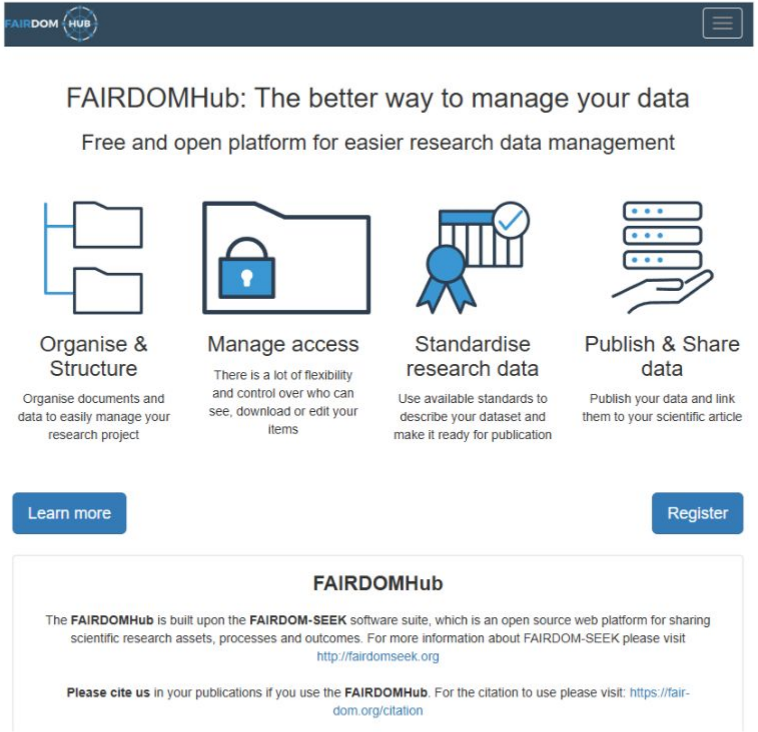
schema:name
Inoculum type, initial pH, retention time and replicate

The 👁 of Interoperability:

… **metadata** that use a **formal, accessible, shared**, and **broadly applicable language** for knowledge representation

Why FAIRDOM-SEEK?

- an open-source web platform designed to <u>share</u> heterogeneous scientific research assets, processes, and outcomes
- a metadata catalogue making your data findable and accessible
- can store small datasets or link to larger datasets (e.g., zenodo)
- **New: Validated *FAIR data station files* can be uploaded for quick population**



FAIRDOMHub: The better way to manage your data

Free and open platform for easier research data management

Organise & Structure

Organise documents and data to easily manage your research project

Manage access

There is a lot of flexibility and control over who can see, download or edit your items

Standardise research data

Use available standards to describe your dataset and make it ready for publication

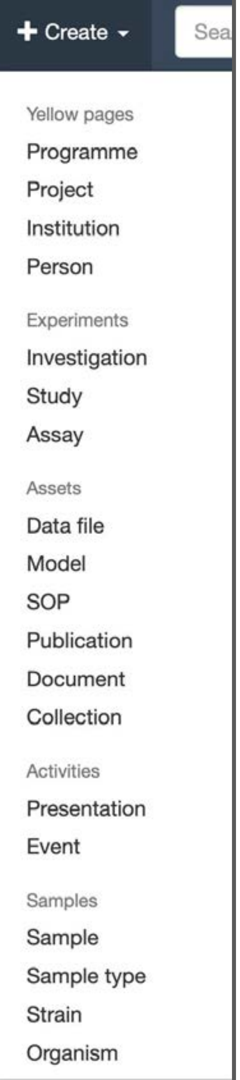
Publish & Share data

Publish your data and link them to your scientific article

FAIRDOMHub

The **FAIRDOMHub** is built upon the **FAIRDOM-SEEK** software suite, which is an open source web platform for sharing scientific research assets, processes and outcomes. For more information about FAIRDOM-SEEK please visit http://fairdomseek.org

Please cite us in your publications if you use the **FAIRDOMHub**. For the citation to use please visit: https://fair-dom.org/citation


+ Create

Sea

Yellow pages
- Programme
- Project
- Institution
- Person

Experiments
- Investigation
- Study
- Assay

Assets
- Data file
- Model
- SOP
- Publication
- Document
- Collection

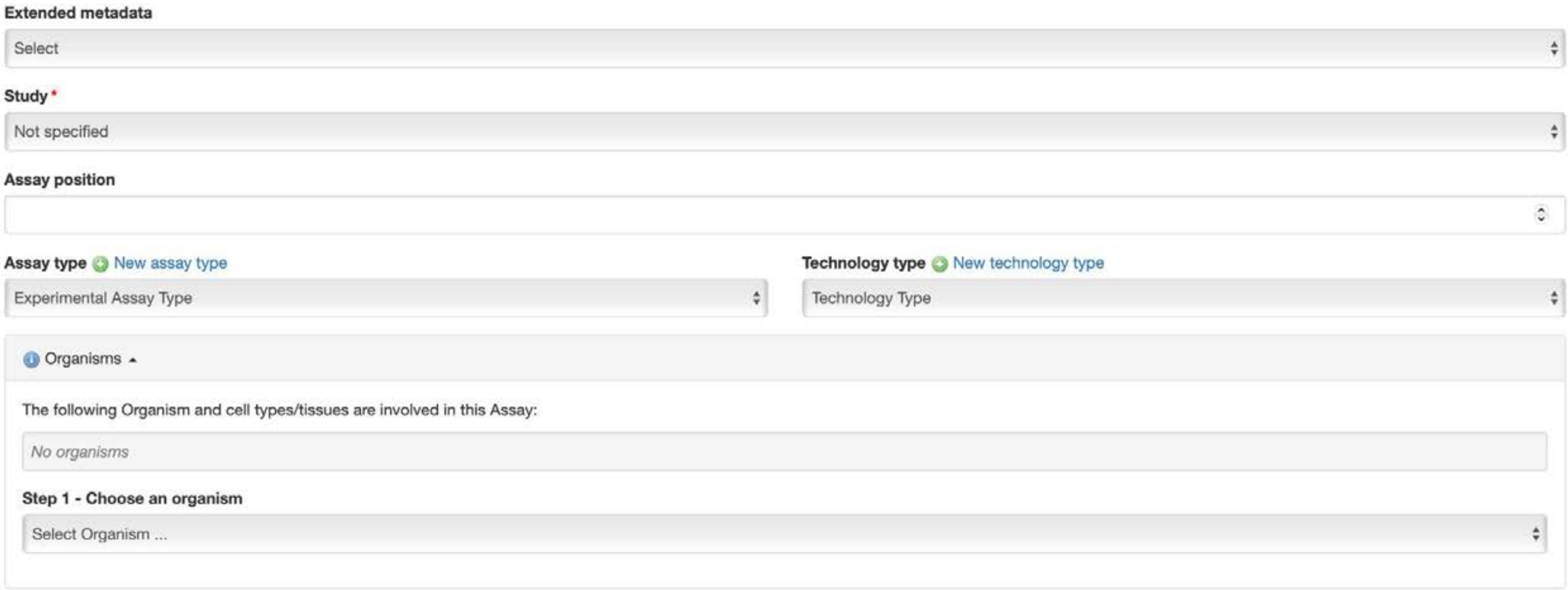
Activities
- Presentation
- Event

Samples
- Sample
- Sample type
- Strain
- Organism

Ability to record many different types of information

- Follows ISA and more!
- Mostly for individual recording of information
- High control on information
- Flexible on user input

What if you want to record a multitude of standardised observation units, assays, samples in one go?


Extended metadata
Select

Study *
Not specified

Assay position

Assay type New assay type
Experimental Assay Type

Technology type New technology type
Technology Type

Organisms

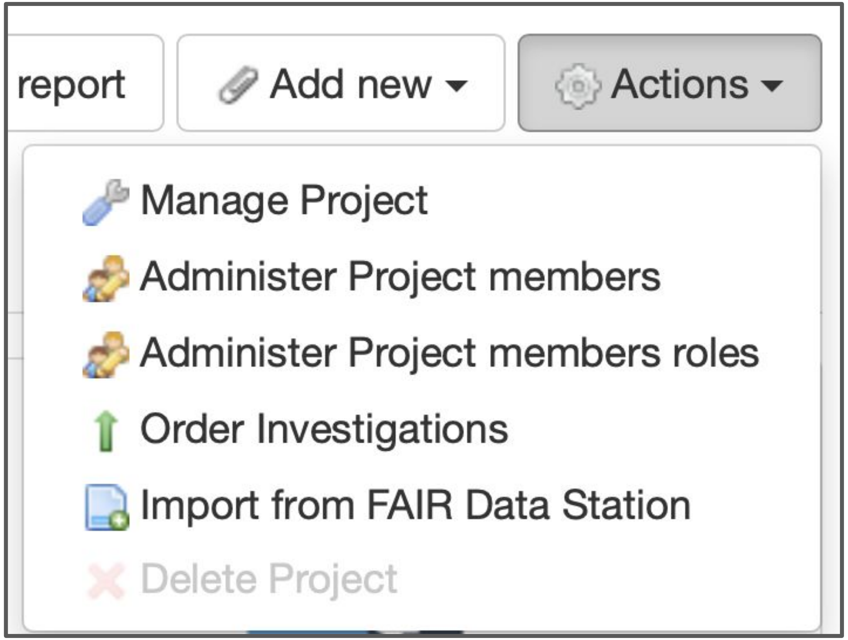
The following Organism and cell types/tissues are involved in this Assay:

No organisms

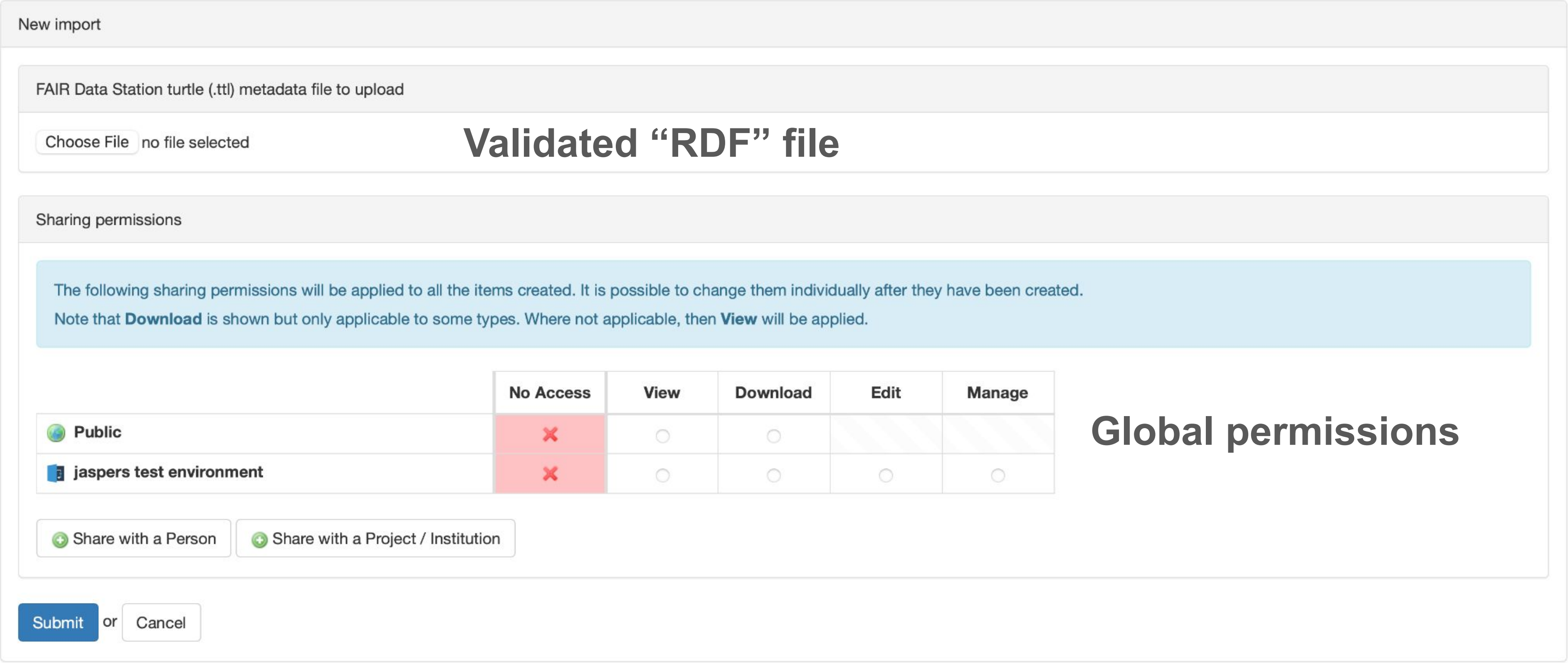
Step 1 - Choose an organism
Select Organism ...

Where is this FAIR Data Station?

At project level (when enabled)



Import ISA from FAIR Data Station for Project ...



New import

FAIR Data Station turtle (.ttl) metadata file to upload

Validated “RDF” file

Sharing permissions

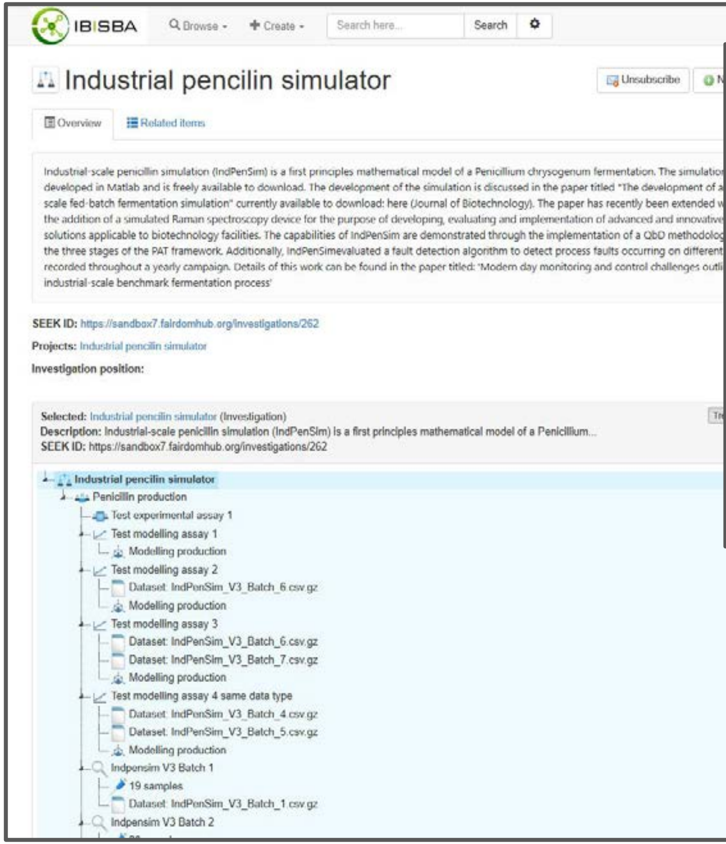
The following sharing permissions will be applied to all the items created. It is possible to change them individually after they have been created.

Note that **Download** is shown but only applicable to some types. Where not applicable, then **View** will be applied.

	No Access	View	Download	Edit	Manage
Public	✗	○	○		
jaspers test environment	✗	○	○	○	○

Global permissions

Result of the FAIR Data Station



Industrial pencilin simulator

Industrial-scale penicillin simulation (IndPenSim) is a first principles mathematical model of a Penicillium chrysogenum fermentation. The simulatio
developed in Matlab and is freely available to download. The development of the simulation is discussed in the paper titled "The development of a
scale fed-batch fermentation simulation" currently available to download: here (Journal of Biotechnology). The paper has recently been extended w
the addition of a simulated Raman spectroscopy device for the purpose of developing, evaluating and implementation of advanced and innovative
solutions applicable to biotechnology facilities. The capabilities of IndPenSim are demonstrated through the implementation of a QbD methodolog
the three stages of the PAT framework. Additionally, IndPenSimevaluated a fault detection algorithm to detect process faults occurring on different
recorded throughout a yearly campaign. Details of this work can be found in the paper titled: 'Modem day monitoring and control challenges outli
industrial-scale benchmark fermentation process'

SEEK ID: https://sandbox7.fairdomhub.org/investigations/262

Projects: Industrial pencilin simulator

Investigation position:

Selected: Industrial pencilin simulator (Investigation)
Description: Industrial-scale penicillin simulation (IndPenSim) is a first principles mathematical model of a Penicillium...
SEEK ID: https://sandbox7.fairdomhub.org/investigations/262

- **Industrial pencilin simulator**
  - Penicillin production
    - Test experimental assay 1
    - Test modelling assay 1
      - Modelling production
    - Test modelling assay 2
      - Dataset: IndPenSim_V3_Batch_6.csv.gz
      - Modelling production
    - Test modelling assay 3
      - Dataset: IndPenSim_V3_Batch_6.csv.gz
      - Dataset: IndPenSim_V3_Batch_7.csv.gz
      - Modelling production
    - Test modelling assay 4 same data type
      - Dataset: IndPenSim_V3_Batch_4.csv.gz
      - Dataset: IndPenSim_V3_Batch_5.csv.gz
      - Modelling production
    - Indpensim V3 Batch 1
      - 19 samples
      - Dataset: IndPenSim_V3_Batch_1.csv.gz
    - Indpensim V3 Batch 2

Dataset: IndPenSim_V3_Batch_1.csv.gz Version 1

file

External datasets

SEEK ID: https://sandbox7.fairdomhub.org/data_files/6591?version=1

Link: https://data.yoda.wur.nl/research-bioindustry/Use%20cases/Simulation%20datasets/IndPenSim/IndPenSim_V3_Batch_1.csv.gz

Selected: Dataset: IndPenSim_V3_Batch_1.csv.gz (Datafile)
Description: file
SEEK ID: https://sandbox7.fairdomhub.org/data_files/6591

- Dataset: IndPenSim_V3_Batch_1.csv.gz

Updates?

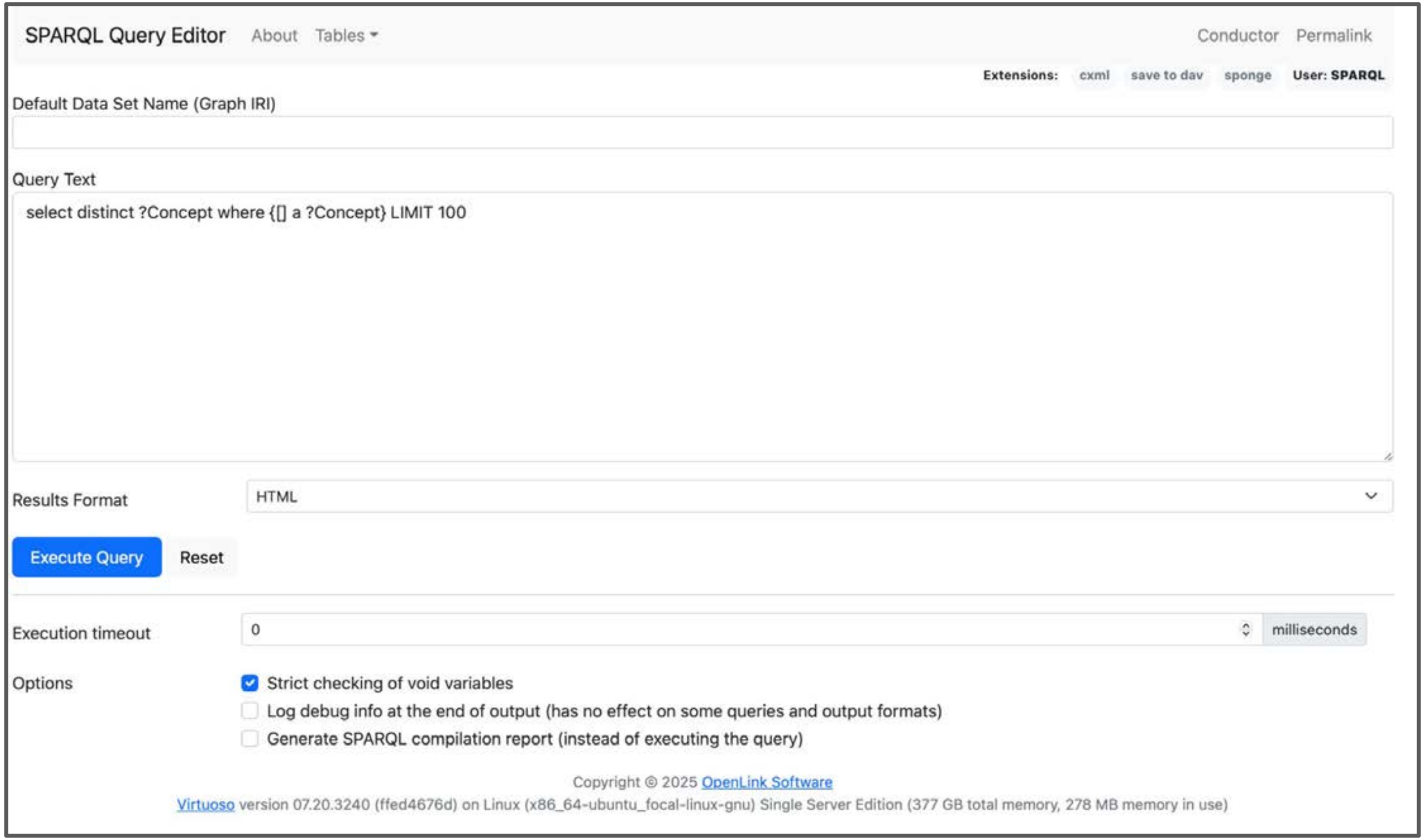
Beyond the browser / API

Using the FAIR Data Station:

- Registry of 100+ samples, assays, etc.. becomes easier
- Enforced consistency of term requirements (even outside SEEK)
- When made public
    - **Ease of cross data retrieval using SPARQL**

https://sparql.fairdomhub.org/

Current query interface (did you know it exists?)



Default Data Set Name (Graph IRI)

Query Text

```
select distinct ?Concept where {[] a ?Concept} LIMIT 100
```

Results Format: HTML

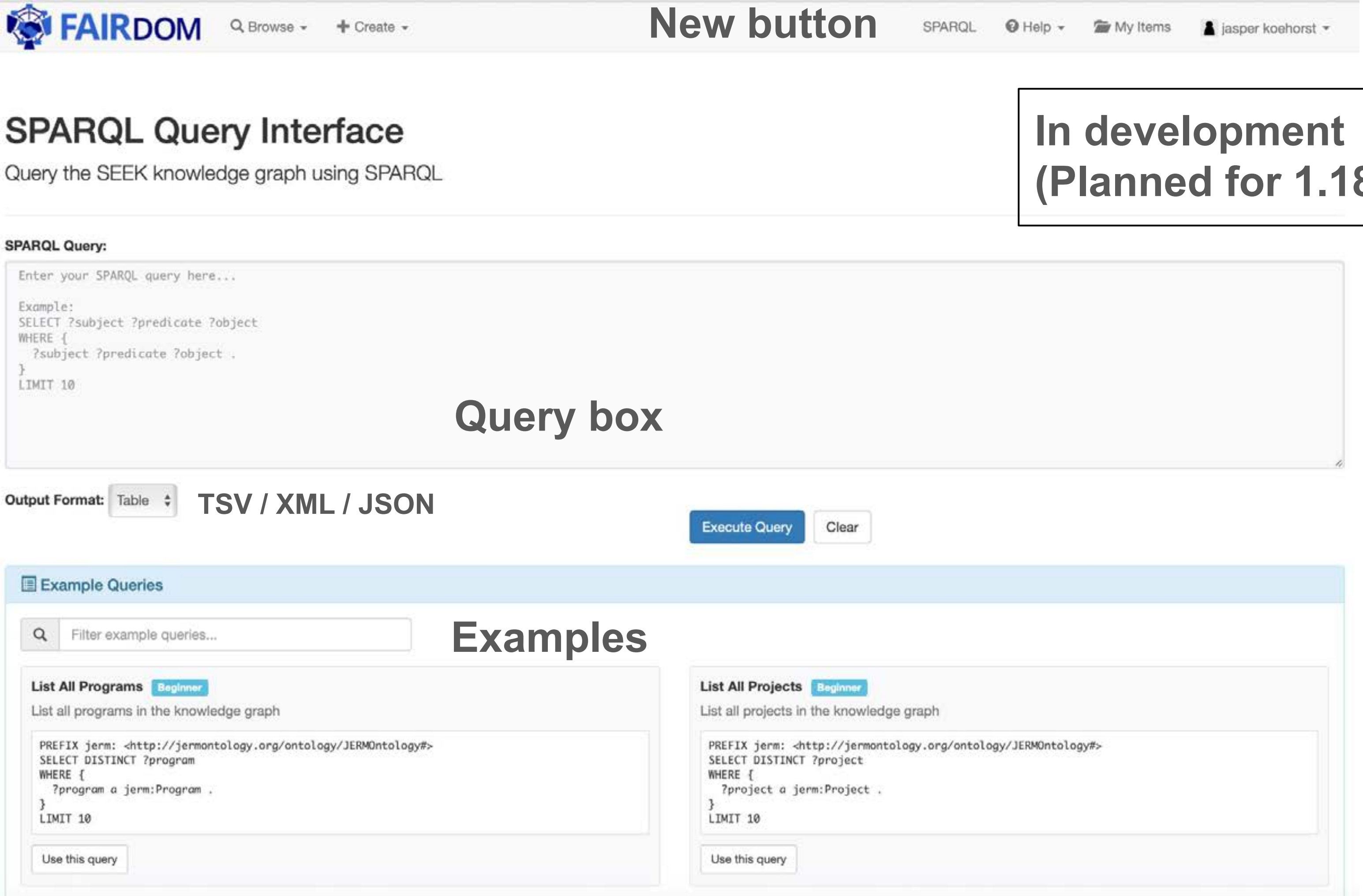
Execution timeout: 0 milliseconds

Options:
- ☑ Strict checking of void variables
- ☐ Log debug info at the end of output (has no effect on some queries and output formats)
- ☐ Generate SPARQL compilation report (instead of executing the query)

Copyright © 2025 OpenLink Software

Virtuoso version 07.20.3240 (ffed4676d) on Linux (x86_64-ubuntu_focal-linux-gnu) Single Server Edition (377 GB total memory, 278 MB memory in use)

In development (Planned for 1.18.0)



New button

SPARQL Query Interface

Query the SEEK knowledge graph using SPARQL

SPARQL Query:

```
Enter your SPARQL query here...

Example:
SELECT ?subject ?predicate ?object
WHERE {
  ?subject ?predicate ?object .
}
LIMIT 10
```

Query box

Output Format: Table

TSV / XML / JSON

Example Queries

Examples

List All Programs Beginner

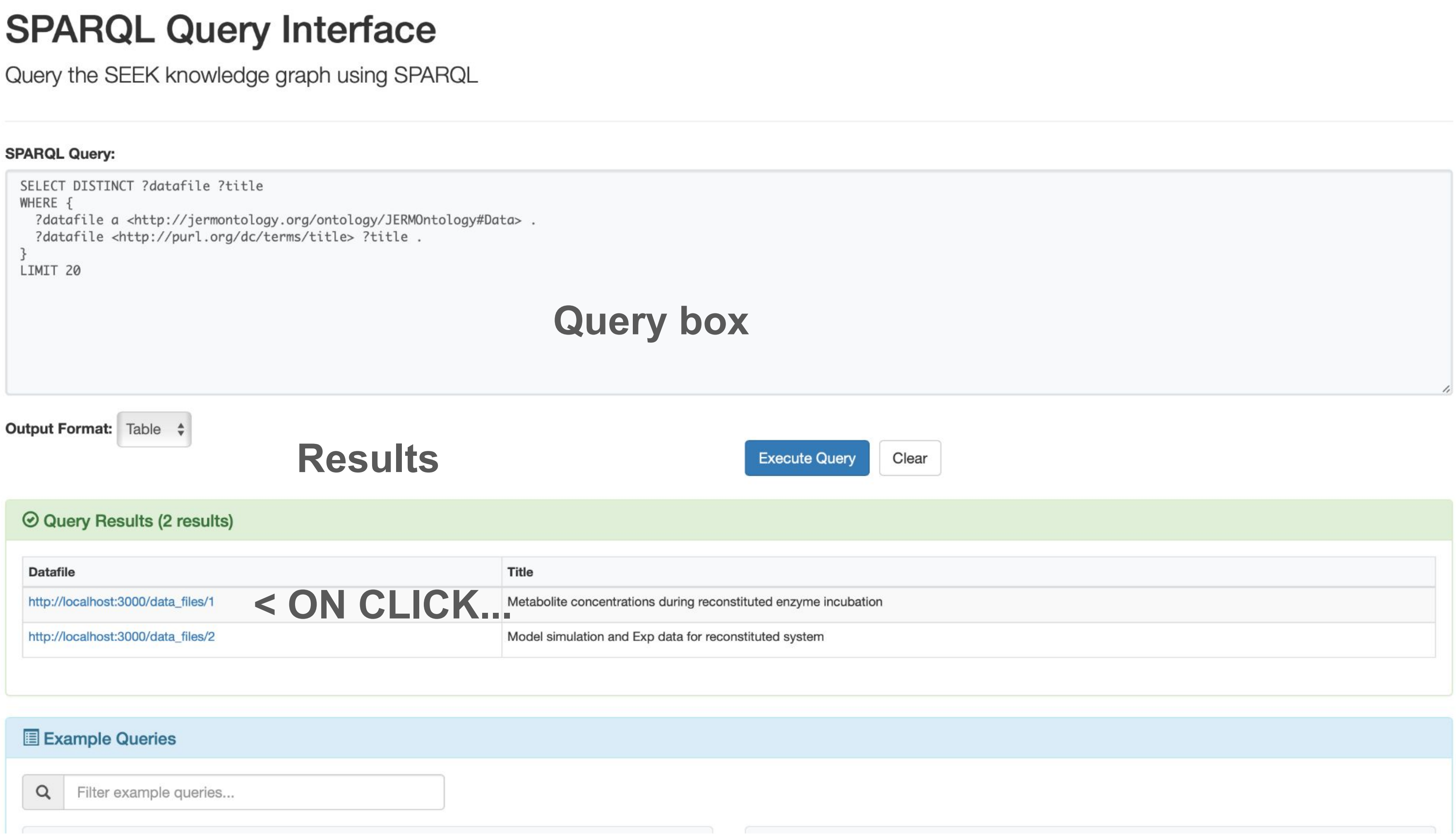
List all programs in the knowledge graph

```
PREFIX jerm: <http://jermontology.org/ontology/JERMOntology#>
SELECT DISTINCT ?program
WHERE {
  ?program a jerm:Program .
}
LIMIT 10
```

List All Projects Beginner

List all projects in the knowledge graph

```
PREFIX jerm: <http://jermontology.org/ontology/JERMOntology#>
SELECT DISTINCT ?project
WHERE {
  ?project a jerm:Project .
}
LIMIT 10
```



SPARQL Query Interface

Query the SEEK knowledge graph using SPARQL

SPARQL Query:

```
SELECT DISTINCT ?datafile ?title
WHERE {
  ?datafile a <http://jermontology.org/ontology/JERMOntology#Data> .
  ?datafile <http://purl.org/dc/terms/title> ?title .
}
LIMIT 20
```

Query box

Output Format: Table

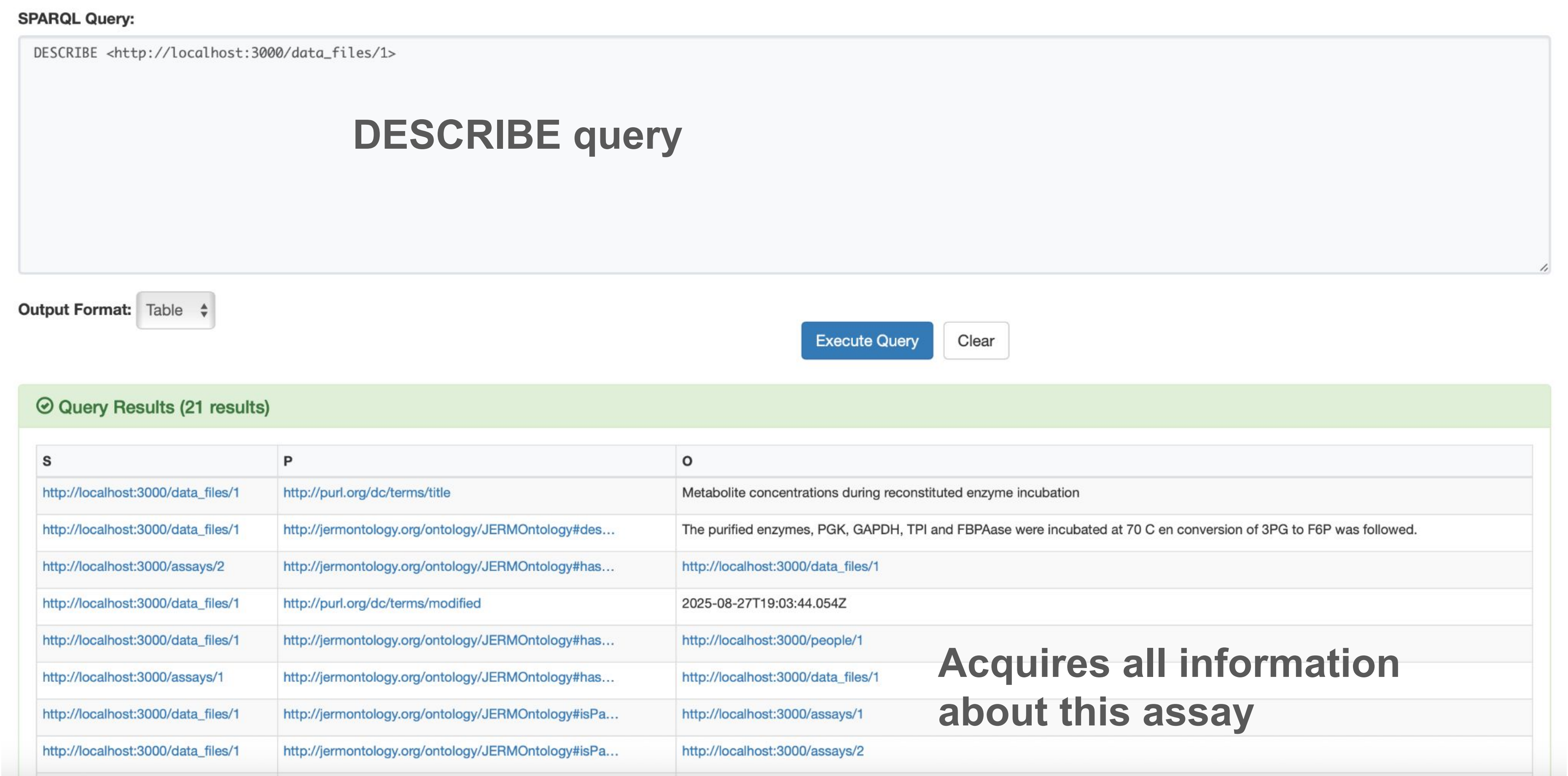
Results

Query Results (2 results)

Datafile	Title
http://localhost:3000/data_files/1	Metabolite concentrations during reconstituted enzyme incubation
http://localhost:3000/data_files/2	Model simulation and Exp data for reconstituted system

< ON CLICK...

Example Queries



DESCRIBE query

SPARQL Query:

```
DESCRIBE <http://localhost:3000/data_files/1>
```

Output Format: Table

Query Results (21 results)

S	P	O
http://localhost:3000/data_files/1	http://purl.org/dc/terms/title	Metabolite concentrations during reconstituted enzyme incubation
http://localhost:3000/data_files/1	http://jermontology.org/ontology/JERMOntology#des…	The purified enzymes, PGK, GAPDH, TPI and FBPAase were incubated at 70 C en conversion of 3PG to F6P was followed.
http://localhost:3000/assays/2	http://jermontology.org/ontology/JERMOntology#has…	http://localhost:3000/data_files/1
http://localhost:3000/data_files/1	http://purl.org/dc/terms/modified	2025-08-27T19:03:44.054Z
http://localhost:3000/data_files/1	http://jermontology.org/ontology/JERMOntology#has…	http://localhost:3000/people/1
http://localhost:3000/assays/1	http://jermontology.org/ontology/JERMOntology#has…	http://localhost:3000/data_files/1
http://localhost:3000/data_files/1	http://jermontology.org/ontology/JERMOntology#isPa…	http://localhost:3000/assays/1
http://localhost:3000/data_files/1	http://jermontology.org/ontology/JERMOntology#isPa…	http://localhost:3000/assays/2

Acquires all information about this assay

Acknowledgements

Contributors:

Jasper J. Koehorst

Brett Metcalfe

Stuart Owen (University of Manchester)

Check-out the FAIR Data station!:

- https://fairds.fairbydesign.nl/
- https://doi.org/10.1093/gigascience/giad014

Funding:

WAGENINGEN UNIVERSITY & RESEARCH

UNLOCK MICROBIAL POTENTIAL

NWO

Funded by the European Union

Bioindustry4.0
Advanced digital solutions for biomanufacturing