

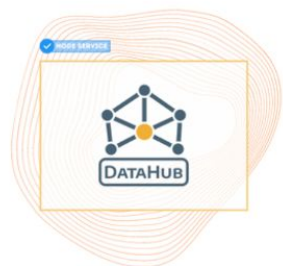


DataHub: focusing on sample metadata

Rafael Andrade Bueno, PhD

FAIRDOM User Meeting, 17th October 2022





A (meta)data management platform

- Primary users: Researchers
- Secondary users: Support staff and service facilities
- Tertiary users: Repositories and schema creators



+



FAIRDOM

<https://fair-dom.org/>

Flora D'Anna (Data Steward)

Rafael Andrade Buono (Data Steward)

Vahid Kiani (Software Developer)



FAIR-by-design approach: *start with the end in mind*



Multi-omics data
submission/brokering

Repositories' metadata
requirements

Local archive



FAIR-by-design approach: *start with the end in mind*



Custom metadata fields

- Admin
- Study-level



Sample templates

- Repositories' checklists, ontologies, controlled vocabularies
- Sample-level
- Customisable
- 1 step at a time



Link to data storage

- Access to external storage (e.g., iRODS based storage service)

Multi-omics data
submission/brokering

Repositories' metadata
requirements

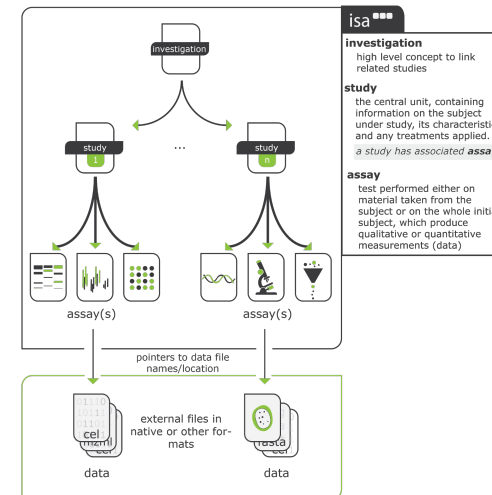
Local archive



DataHub method: FAIRDOM-SEEK and ISA model



Powered by
FAIRDOM-SEEK
software



Implementation
of **ISA model** in
FAIRDOM-SEEK

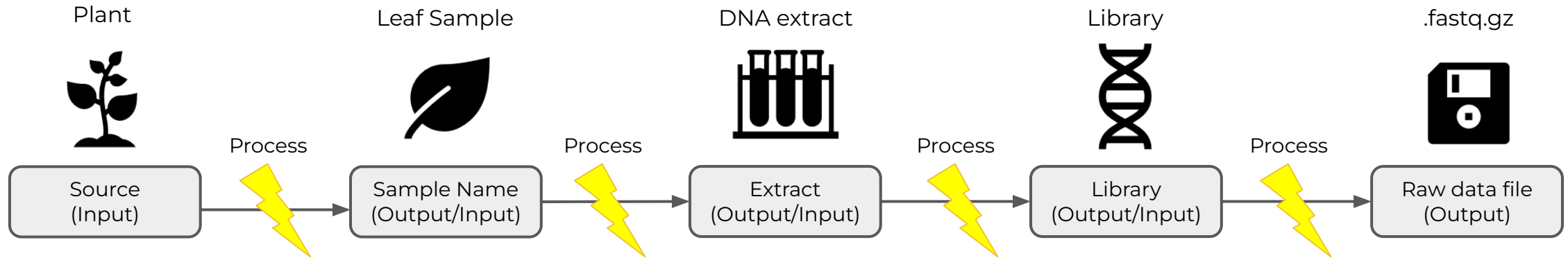
Multi-omics data
submission/brokering

Repositories' metadata
requirements

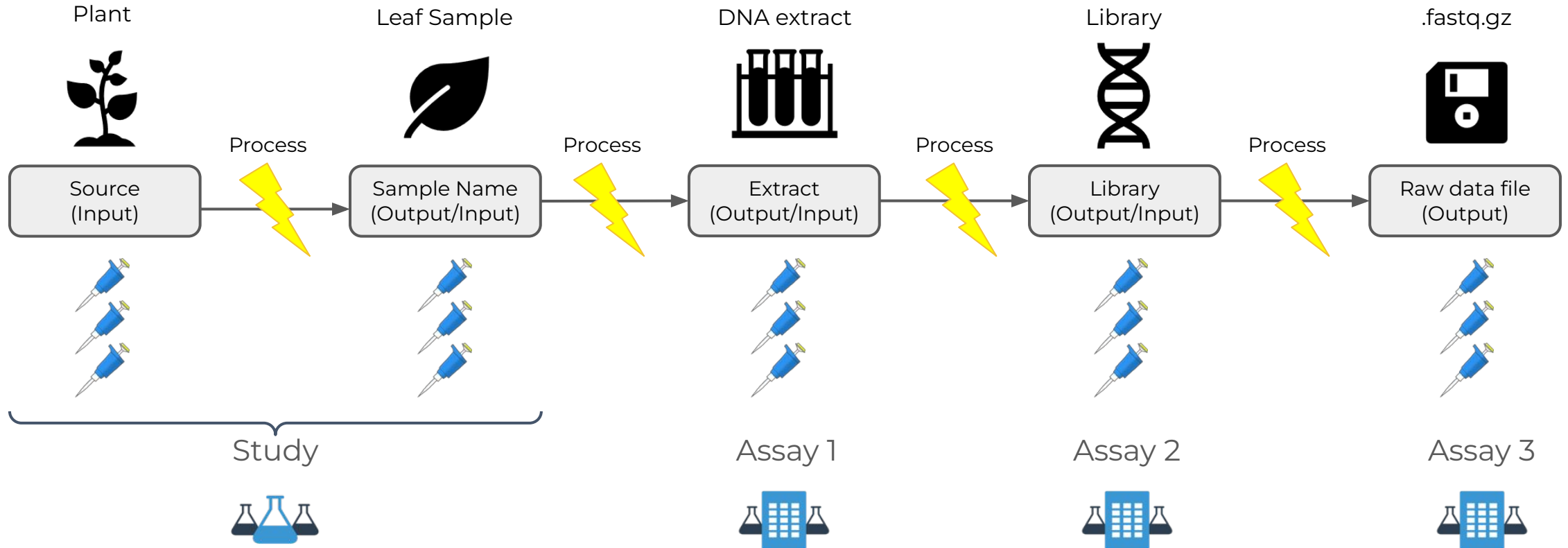
Local archive



Building a full experimental flow in DataHub



Building a full experimental flow in DataHub



Interacting with a Project in a single page

 Browse ▾ Create ▾ Search ⚙️ About ▾ Help ▾ My Items 👤 claire s ▾

Project

Investigation

Study

Assay

Toggle: Investigations Studies Assays

Project

Mutant re-sequencing

Project items 0

Data files 0

Models 0

Published 0

Presentations 0

Publications 0

SOPs 0

Unsorted items 5

Investigation - hi-per characterization Investigation

Study - Mutant re-sequencing Study

Source material

samples 2

Sample collection

Study sample

samples 2

Study table

Assay - DNA extraction Assay

samples 2

DNA extraction

Assay - Library preparation Assay

samples 2

Library preparation

Assay - Sequencing Assay

samples 4

DNA sequencing

Investigation - hi-per characterization

Unsubscribe New Investigation based on this one Design Study Export ISA Actions ▾

Overview

Related items

Characterizing the hi-per mutant line

SEEK ID: <http://localhost:3000/investigations/4>

Projects: [Mutant re-sequencing](#)

Investigation position:

Creators and Submitter

Creator

👤 claire s

Submitter

👤 claire s

Activity

Views: 48

Created: 19th Jul 2022 at 09:41



[Funding](#) | [Terms & Conditions](#) | [Privacy Policy](#)

Copyright © 2008 - 2022 The University of Manchester and HITS gGmbH

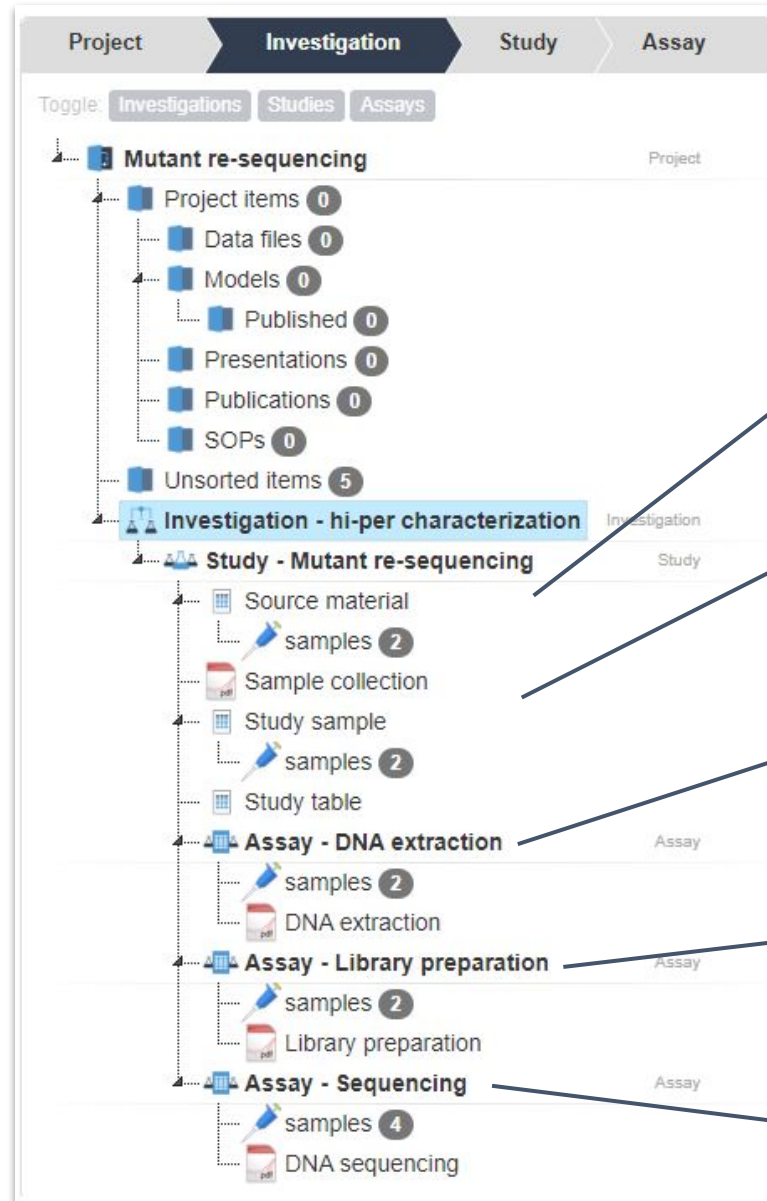
<https://datahub.test.elixir-belgium.org/>

Powered by



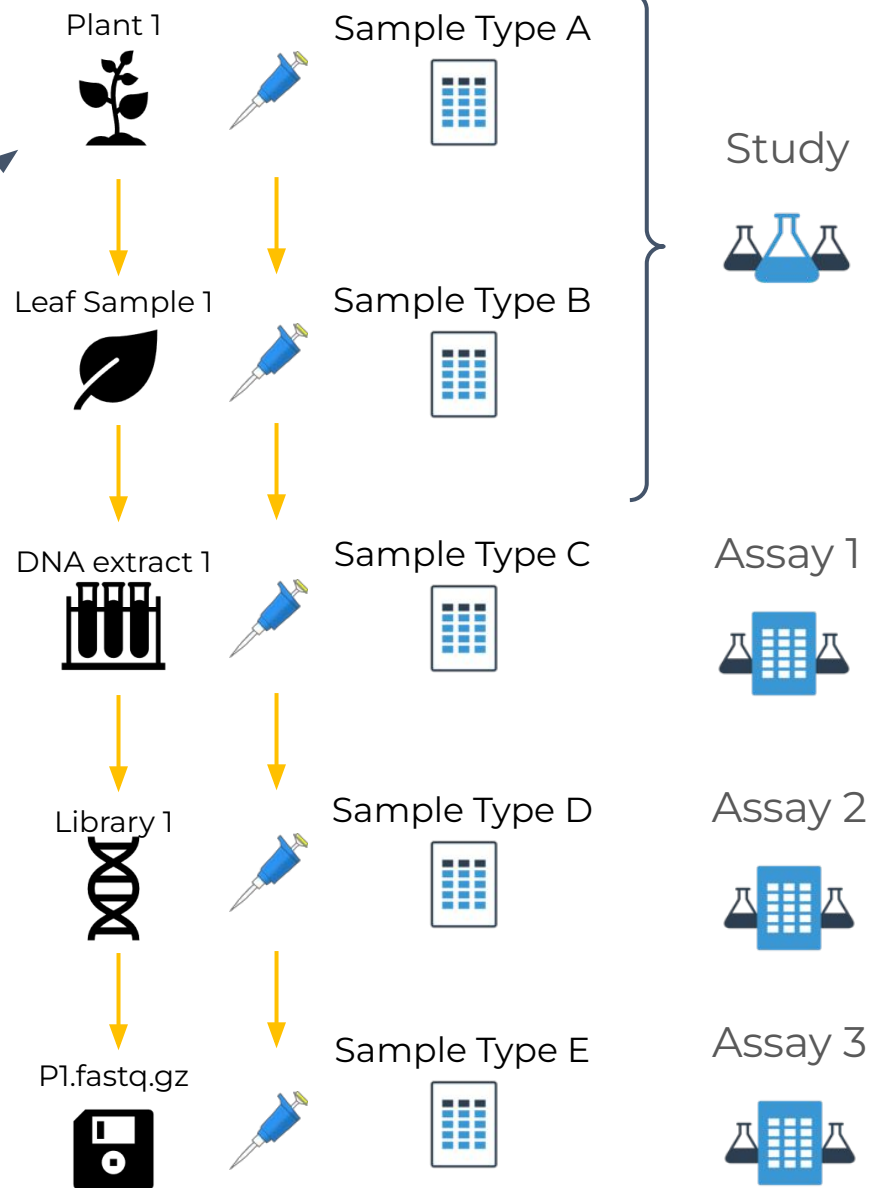
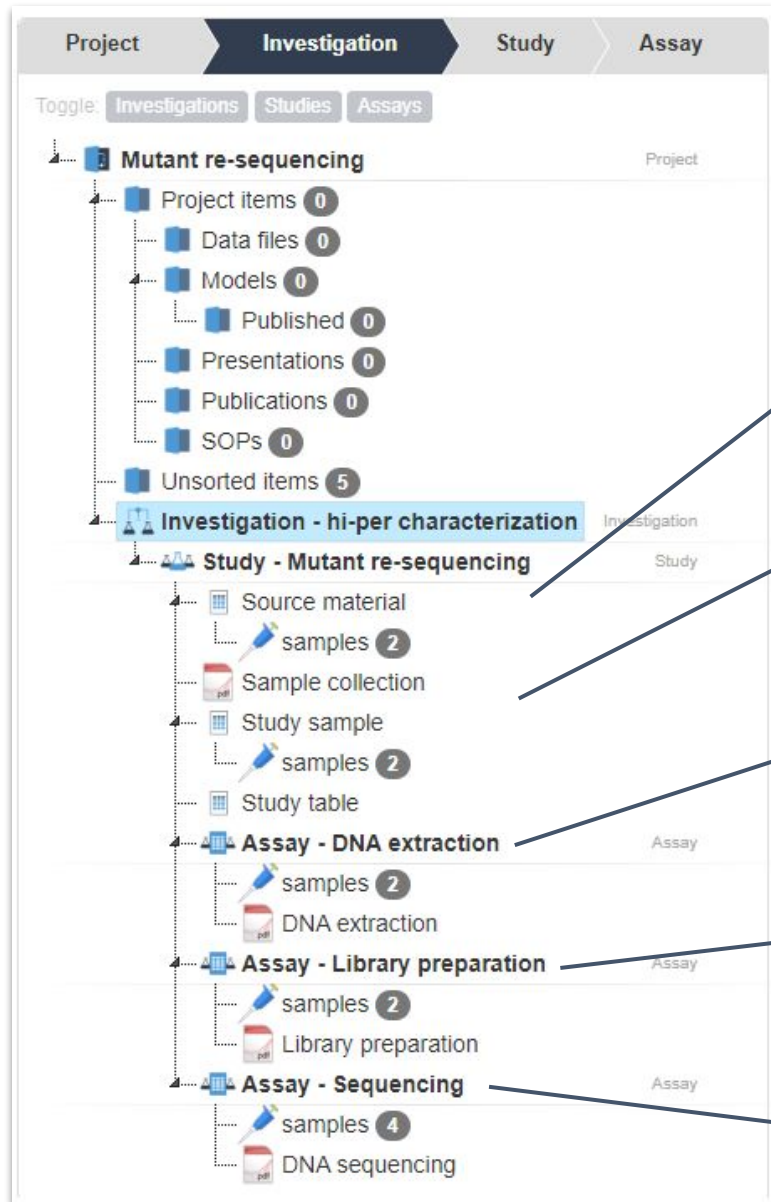
(v.1.13.0-master)

Interacting with a experimental flow in DataHub



SEEK Samples
reference their input

Interacting with a experimental flow in DataHub



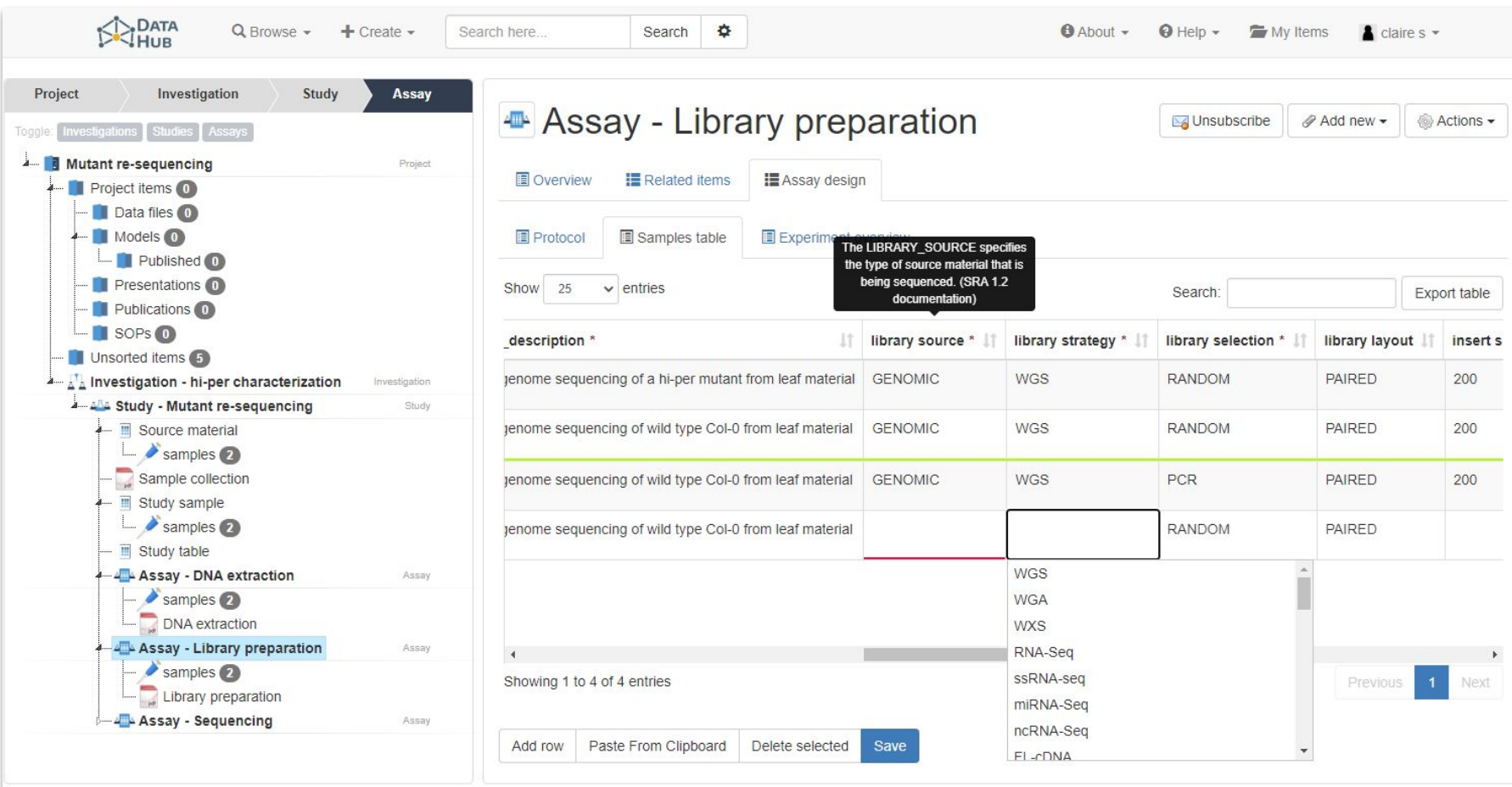
SEEK Samples
reference their input

Dynamic Table for Sample metadata

- Sample input in a table format
- Headers contain descriptions
- Displays validation errors
- Controlled vocabulary
- Ontology terms choices
- Search box for input values
- Copy/export and paste content

Future

- Batch upload/update of Samples
- Being able to add/remove/edit Attributes after the table is defined



Assay - Library preparation

Unsubscribe Add new Actions

Overview Related items Assay design

Protocol Samples table Experimental design

Show 25 entries Search: Export table

description *	library source *	library strategy *	library selection *	library layout	insert s
genome sequencing of a hi-per mutant from leaf material	GENOMIC	WGS	RANDOM	PAIRED	200
genome sequencing of wild type Col-0 from leaf material	GENOMIC	WGS	RANDOM	PAIRED	200
genome sequencing of wild type Col-0 from leaf material	GENOMIC	WGS	PCR	PAIRED	200
genome sequencing of wild type Col-0 from leaf material			RANDOM	PAIRED	

The LIBRARY_SOURCE specifies the type of source material that is being sequenced. (SRA 1.2 documentation)

Showing 1 to 4 of 4 entries

Add row Paste From Clipboard Delete selected Save

WGS
WGA
WXS
RNA-Seq
ssRNA-seq
miRNA-Seq
ncRNA-Seq
EL-cDNA

Templates: easing adoption and exchange of standards

Filter templates ▾

Repository name: Type: Organism:

Choose a template:

Show entries

Search:

Required	Attribute name	Description	Unit	Data type	Is title	IRI	Remove
☒	Source Name	Sources are considered as the starting biological material used in a study.	-	String	☒	-	
☒	Title	Short text that can be used to call out sample records in search results or in displays.	-	String	☐	-	
☒	Description	Free-form text describing the sample, its origin, and its method of isolation.	-	String	☐	-	
☒	tax_id	NCBI Taxonomy Identifier number ID. This is appropriate for individual organisms and some environmental samples.	-	String	☐	-	
☐	organism	Provide a taxonomic information associated to the Source Biological material (for example, Species, genus, strain,...). Ideally, provide NCBI TaxID or NEWT identifier	-	String	☐	-	Remove
☐	cell_type	cell type from which the sample was obtained	-	String	☐	-	Remove
☐	dev_stage	if the sample was obtained from an organism in a specific developmental stage, it is specified with this qualifier	-	String	☐	-	Remove
☐	germline	the sample described presented in the entry has not undergone somatic genomic rearrangement as part of an adaptive immune response; it is the unrearranged molecule that was inherited from the parental germline	-	String	☐	-	Remove
☐	tissue_lib	tissue library from which sample was obtained	-	String	☐	-	Remove
☐	tissue_type	tissue type from which the sample was obtained	-	String	☐	-	Remove

Showing 1 to 10 of 38 entries

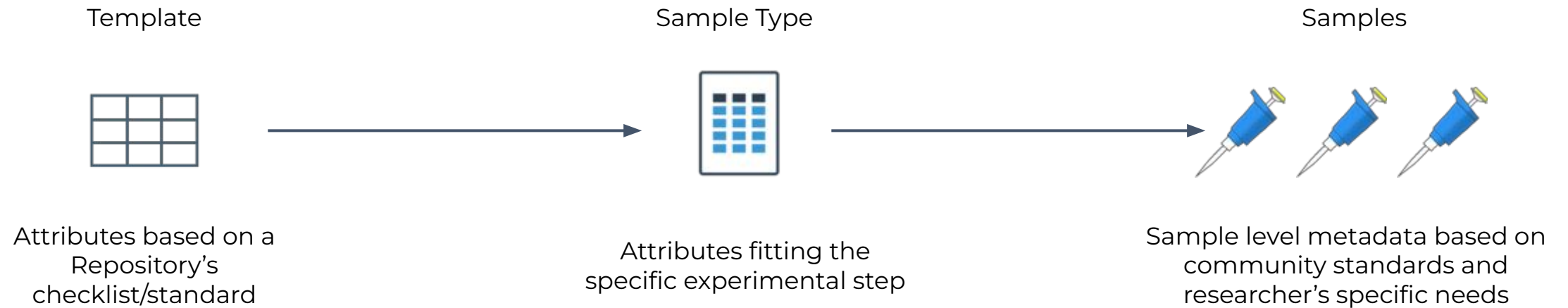
Previous Next

or

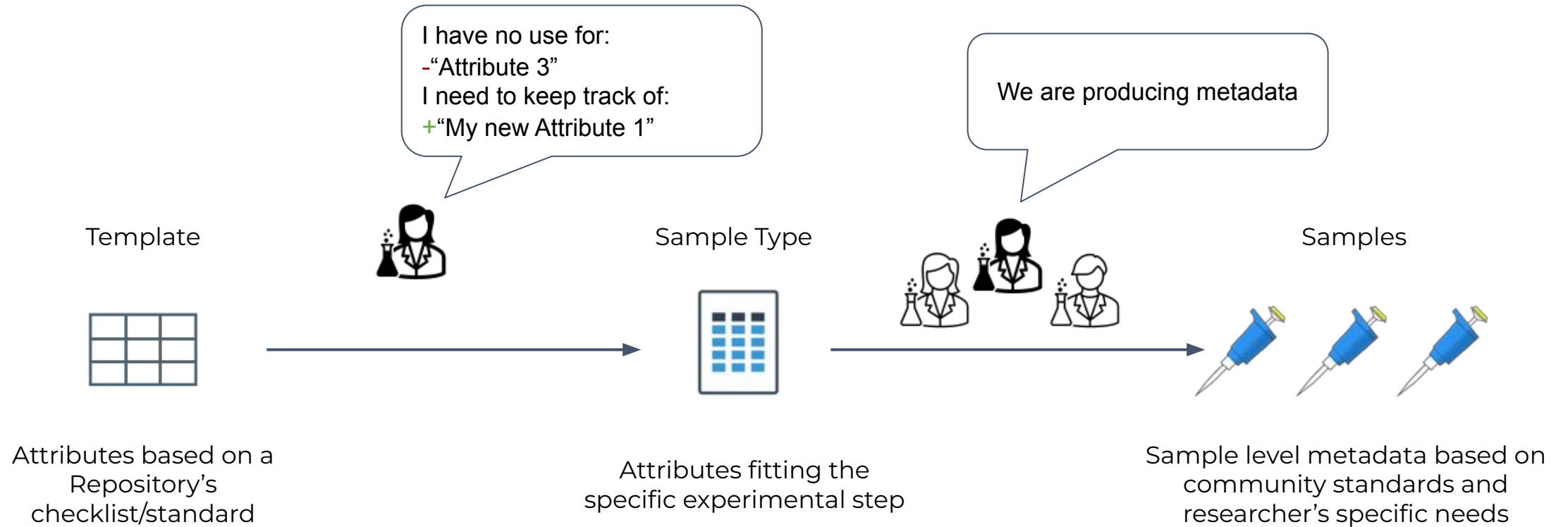
- Sets of attributes to provide **Sample level (meta)data**
 - Helps balancing flexibility and compliance
- DataHub provides Templates containing metadata attributes required by EBI data repositories (e.g., ENA, ArrayExpress, MetaboLights, etc.)
- Attributes can be defined by ontology terms (or URLs)
- Users will be able to make custom Templates
- Templates have sharing permission
- (Some) Future goals:
 - One Template that covers many Repositories
 - Template versioning
 - Being able to reuse Attributes (defined by an IRI)
 - Being able to assemble new Templates by picking & choosing existing Attributes

Template based on ENA's ERC000011 Sample Checklist

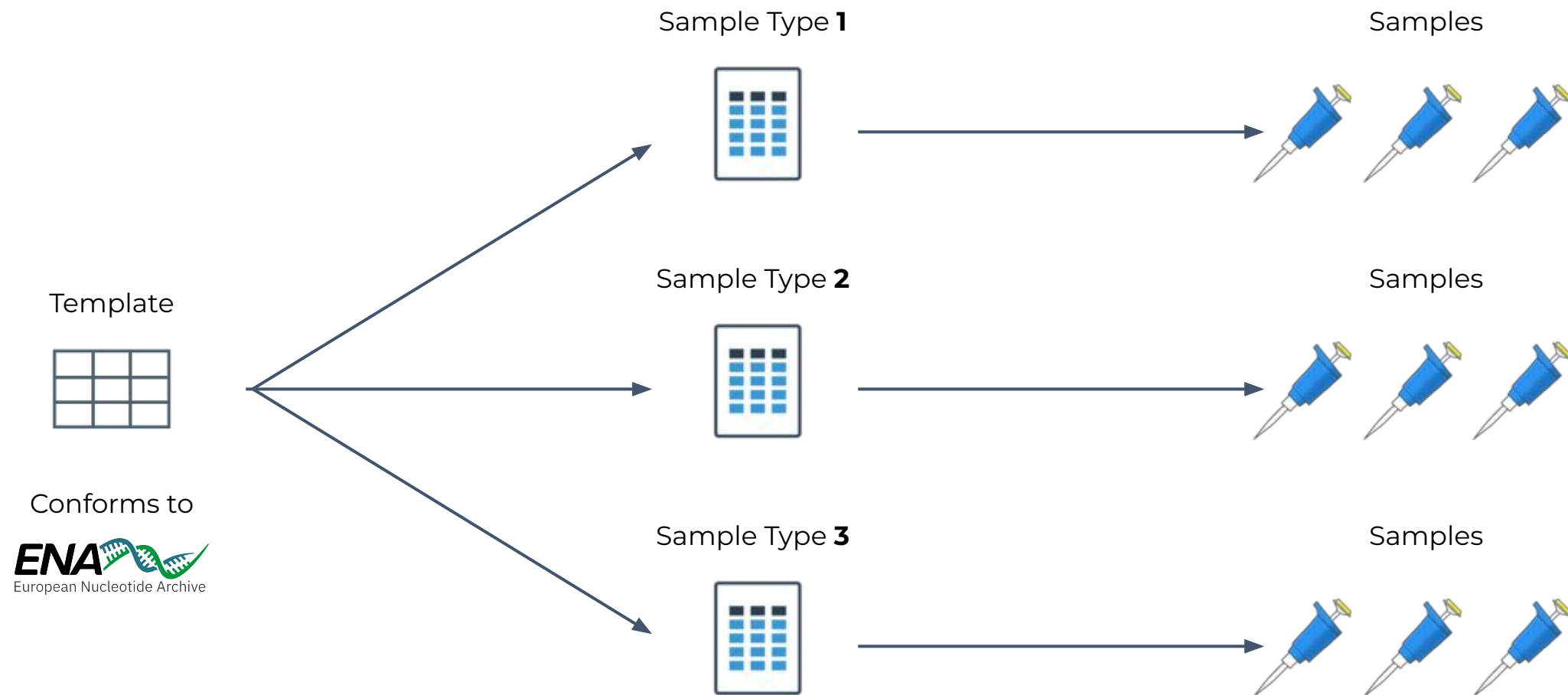
DataHub features: Templates to Sample Types



DataHub features: Templates to Sample Types



DataHub features: Templates



Querying Samples across experimental steps

Query Samples

Select Project(s)

× Environmental stimuli on yeast

Select a Template*

nucleic acid sequencing

Select an Attribute

sequencing instrument

Enter value

Illumina MiSeq

Advanced filtering ⌵

Input sample(s)

Select a Template

source_plant

Select an Attribute

Source Name

Enter value

Yeast culture 1

Output sample(s)

Select a Template

Not selected

Select an Attribute

Not selected

Enter value

Query

6 sample(s) visible to you

Show

10

 entries

Export table

Input	nucleic acid sequencing	sequencing instrument	Assay Name	Technology type	Raw Data File
library 1-1	RNA sequencing protocol	Illumina MiSeq	yeast sequencing assay 1	sequencing assay	1-1_R1.fastq.gz
library 1-1	RNA sequencing protocol	Illumina MiSeq	yeast sequencing assay 1	sequencing assay	1-1_R2.fastq.gz
library 1-2	RNA sequencing protocol	Illumina MiSeq	yeast sequencing assay 1	sequencing assay	1-2_R1.fastq.gz
library 1-2	RNA sequencing protocol	Illumina MiSeq	yeast sequencing assay 1	sequencing assay	1-2_R2.fastq.gz
library 1-3	RNA sequencing protocol	Illumina MiSeq	yeast sequencing assay 1	sequencing assay	1-3_R1.fastq.gz
library 1-3	RNA sequencing protocol	Illumina MiSeq	yeast sequencing assay 1	sequencing assay	1-3_R2.fastq.gz

Search Input

Search nucleic acid sequen

Search sequencing instrum

Search Assay Name

Search Technology type

Search Raw Data File

Showing 1 to 6 of 6 entries

Previous

1

Next

“In a Project, show me the sequencing files originated from a Illumina MiSeq instrument and that come from Yeast culture 1”



Metadata rich submissions: ISA-JSON and RO-Crate



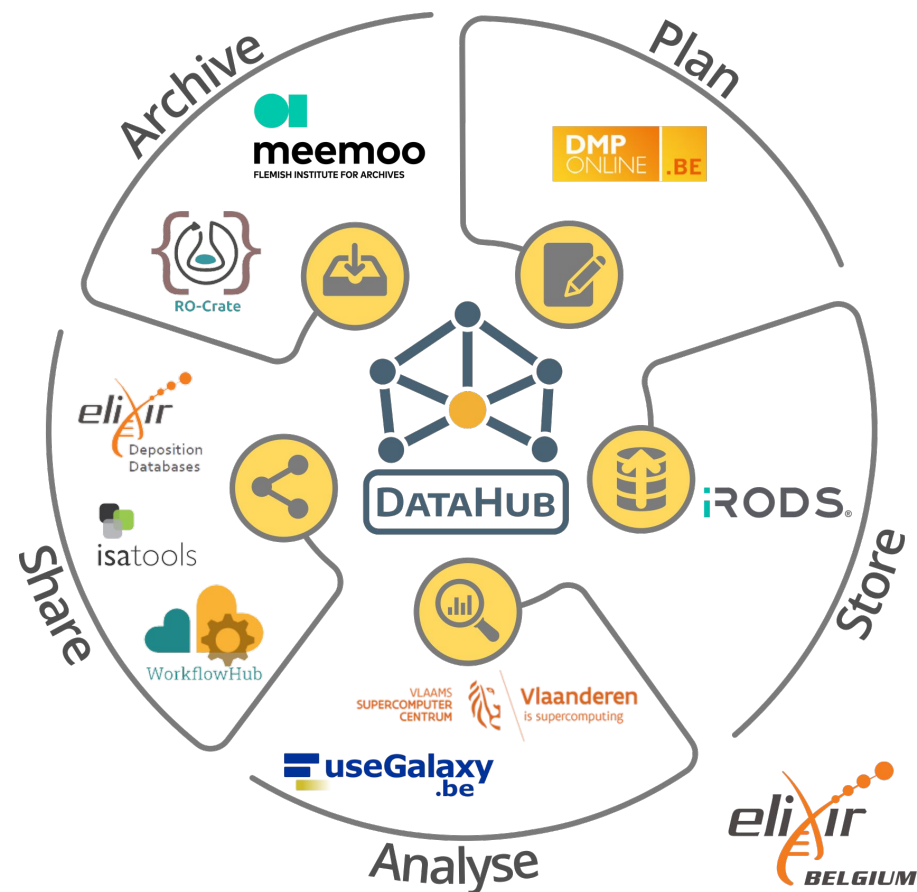
Direct submission



Conversion



DataHub features: view for the future



- Improvements and streamlining the GUI
- Integration with external storage systems (e.g., iRODS)
- Integration with useGalaxy.be, WorkflowHub
- Features for data submission/brokering to ELIXIR Deposition Databases
- Packaging of Data and Metadata in RO-Crates
- Integration with external long term archival solutions
- Explore possible link with DMP tools (e.g. DMPonline)

<https://datahub.elixir-belgium.org/>

Acknowledgements



Vlaanderen
is supercomputing

DataHub is hosted at the Flemish Supercomputer Center



DataHub development receives support from the
Research Foundation – Flanders (FWO) under project No I002819N



Data brokering efforts have received support from European Union's Horizon
2020 programme under grant agreement No 871075 (ELIXIR-CONVERGE).



This work is licensed under the Creative Commons Attribution 4.0 International License,
except where otherwise noted.

To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.





Get in touch
with us.



datahub@elixir-belgium.org



<https://www.elixir-belgium.org/>

