CEPLAS
Cluster of Excellence on Plant Sciences

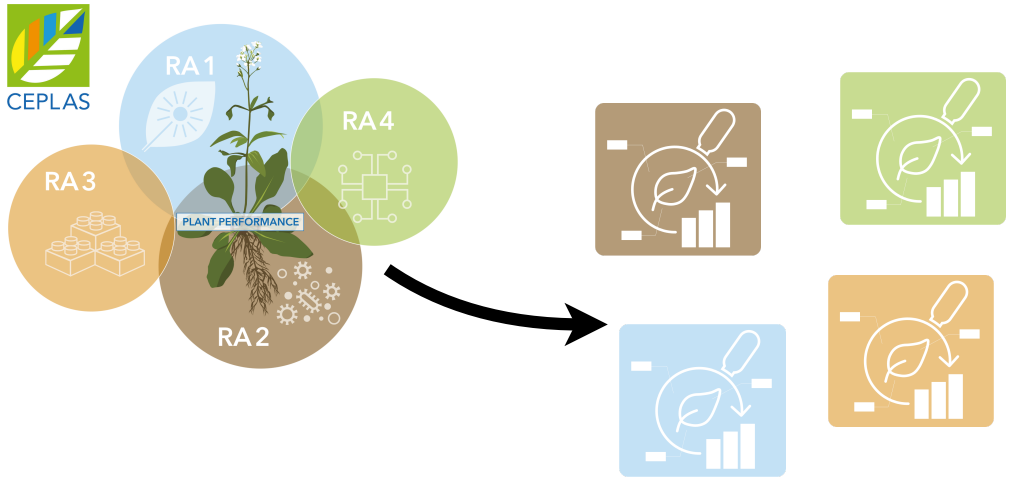
ARCify your research project

October, 2024

Dominik Brilhaus (CEPLAS)
Sabrina Zander (MibiNet)

Goals

- First few steps into ARC ecosystem
- Move existing datasets into ARCs
- Share them via the DataHUB
- Annotate with metadata



House-keeping

- Use the pad to raise questions and feedback
- Copy / paste links (hands-on)

Training Materials

Slides are shared via DataPLANT knowledge base

Resources – join the open source movement

DataPLANT (nfdi4plants)

DataPLANT Website: https://nfdi4plants.org/

Knowledge Base: https://nfdi4plants.org/nfdi4plants.knowledgebase/

DataHUB: https://git.nfdi4plants.org

GitHub: https://github.com/nfdi4plants

HelpDesk: https://helpdesk.nfdi4plants.org

💡 You can help us by raising issues, bugs, ideas...

🚀 NEW! ARC website: https://arc-rdm.org

Continuous support

Data managers in Düsseldorf, Cologne, Jülich and close by (CEPLAS, MibiNet, TRR341) offer support.

1. Slack Workspace for ad hoc support
2. Monthly user meeting (2nd Friday of the month)

→ Details.

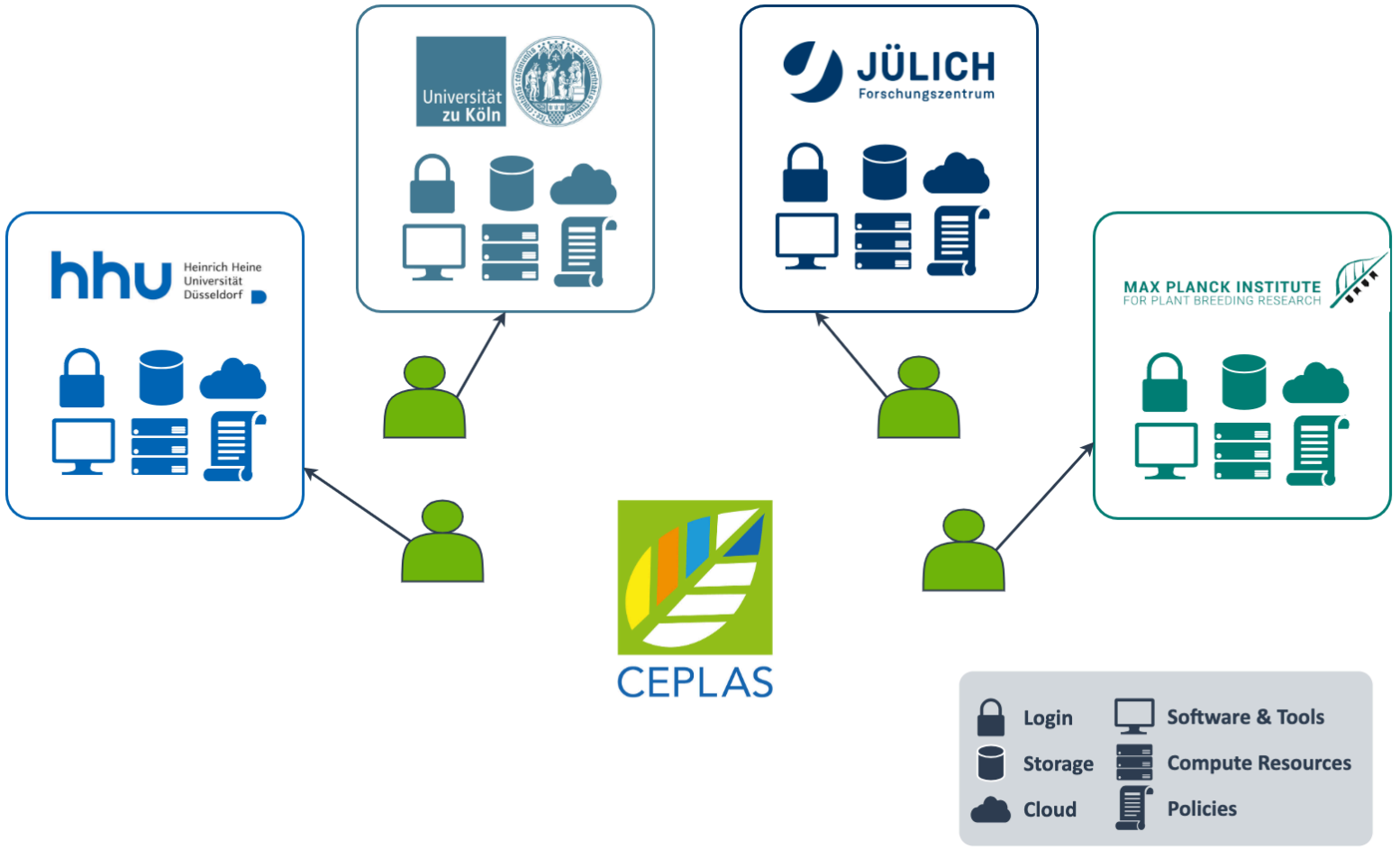
CEPLAS
Cluster of Excellence on Plant Sciences

trr_341 plant ecological genetics

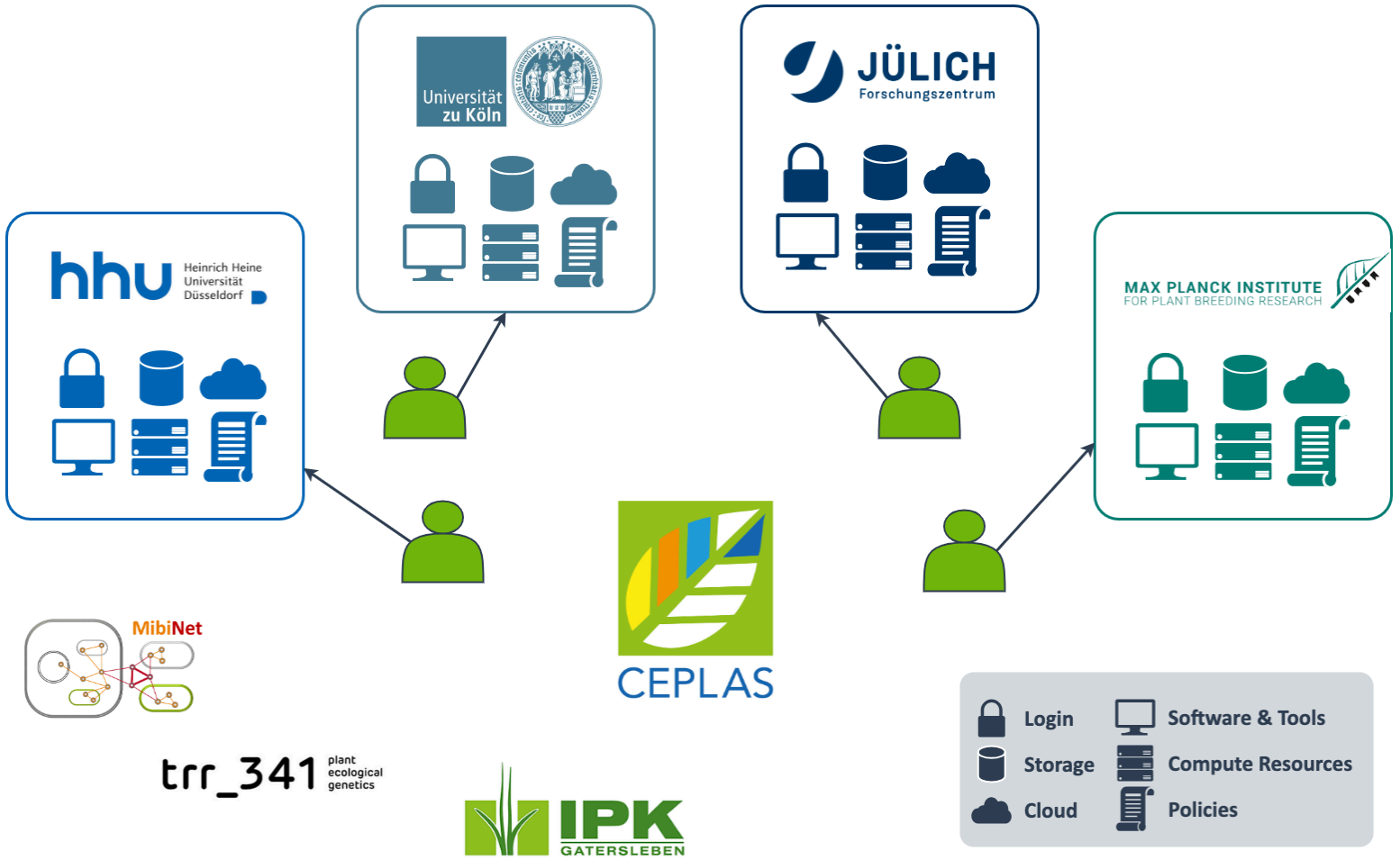
MibiNet

CEPLAS – One cluster, four locations

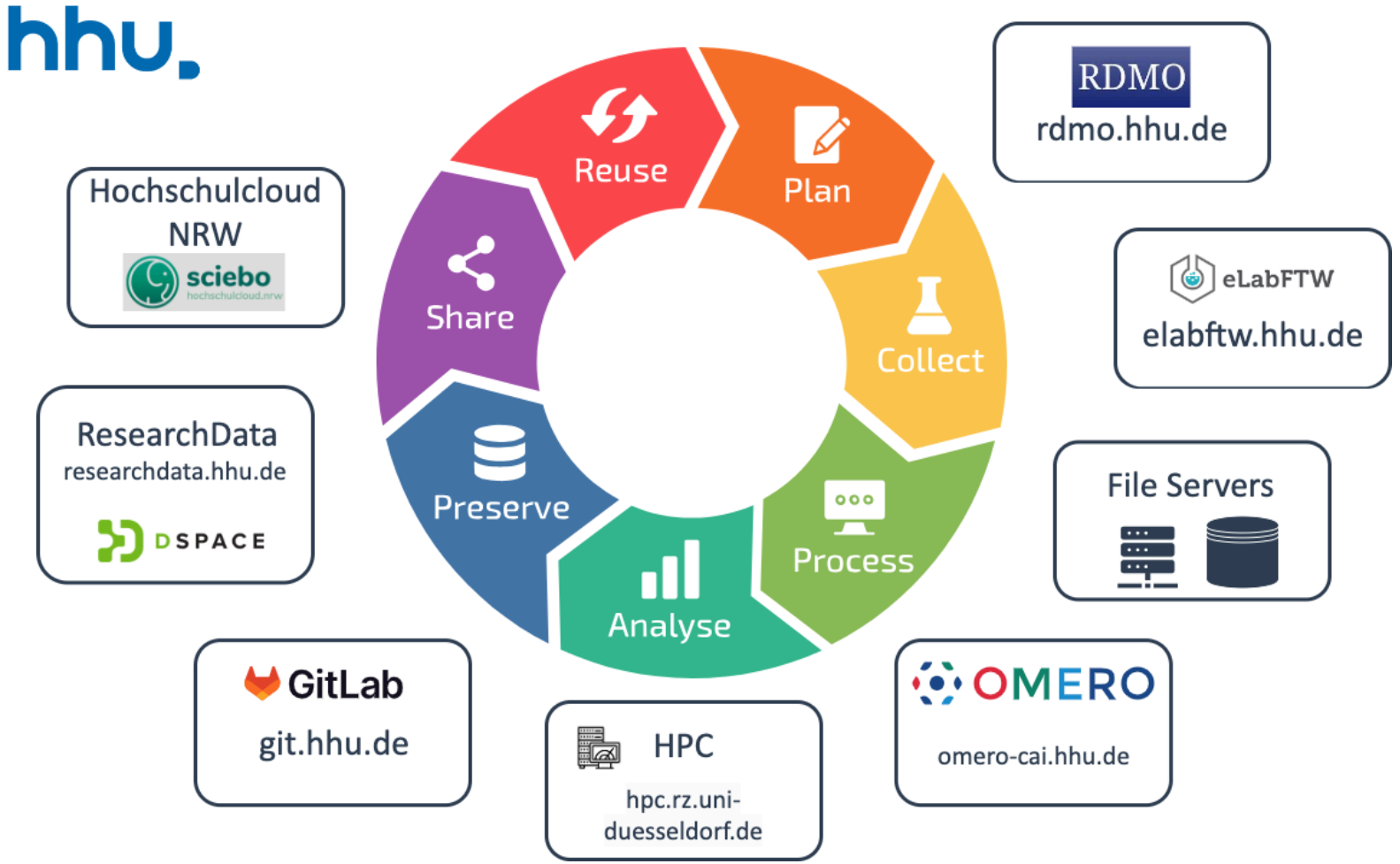
Data silos impede collaboration



Data silos impede collaboration



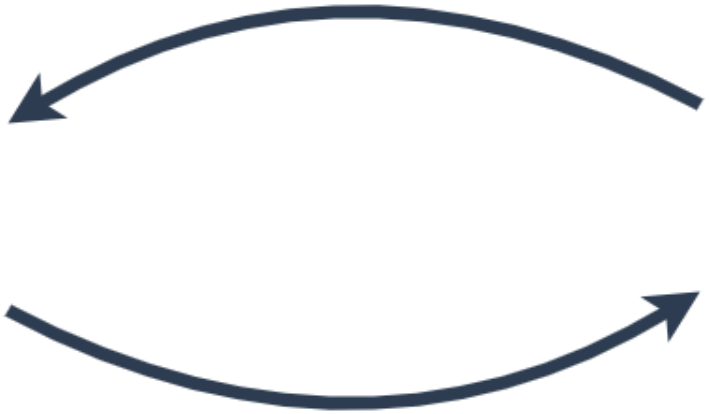
Missing interfaces impede collaboration



Data Stewardship between DataPLANT and the community

Community

nfdi4plants

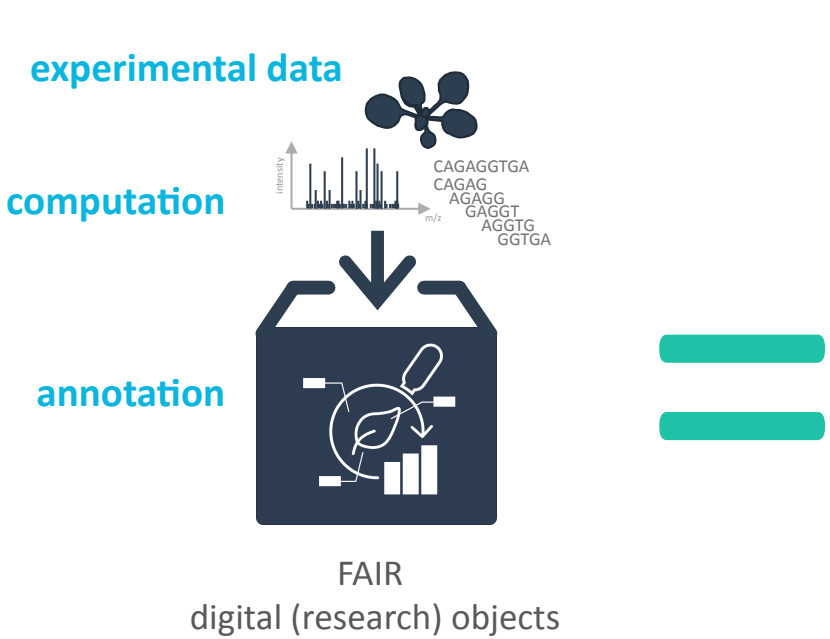


CEPLAS

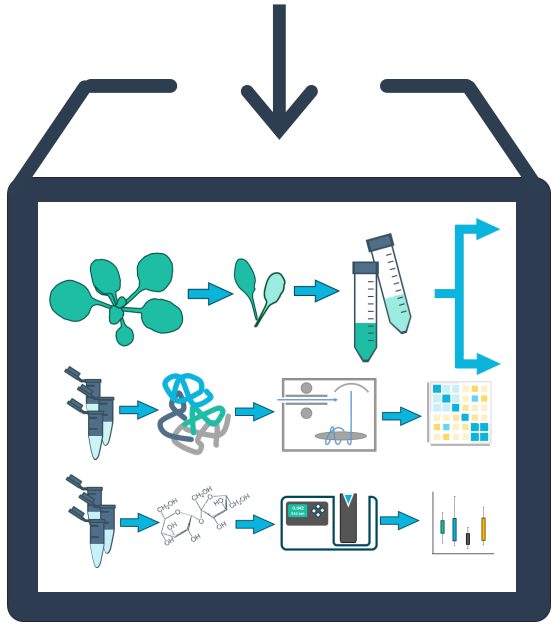
Domain experts
User experience
Training

Service provider
Developers
Tech experts

Annotated Research Context (ARC)

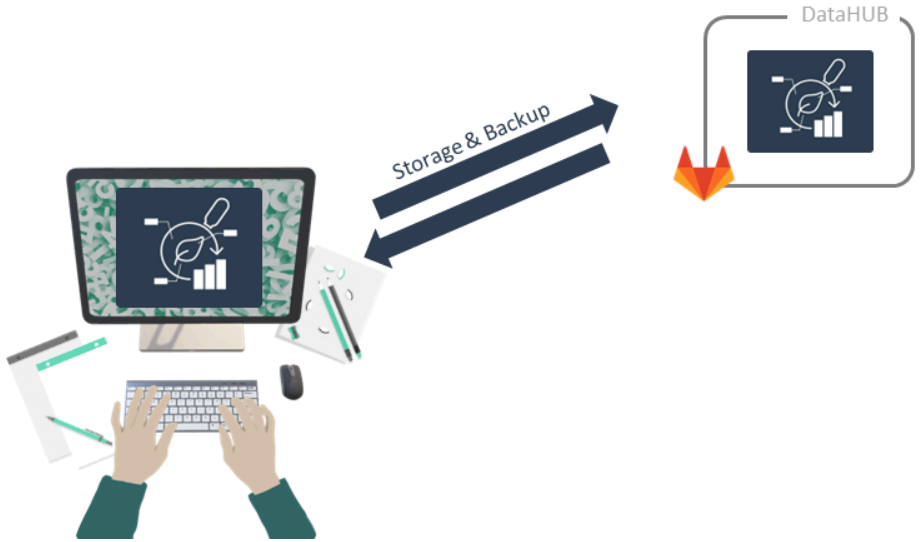


FAIR
digital (research) objects

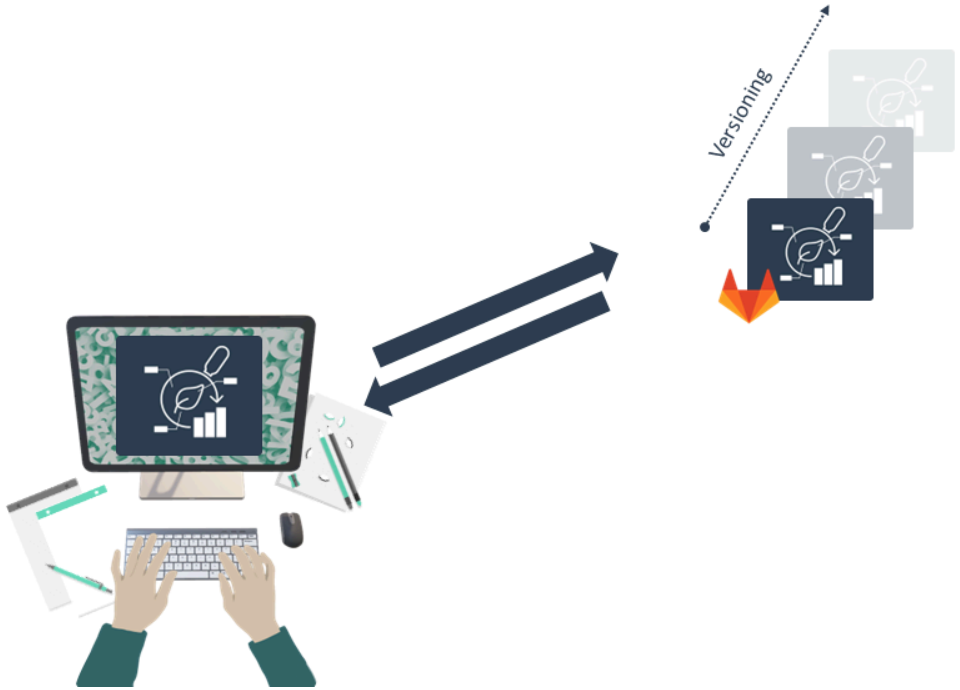


Your entire investigation in a single unified bag

You can store your ARC in the DataHUB

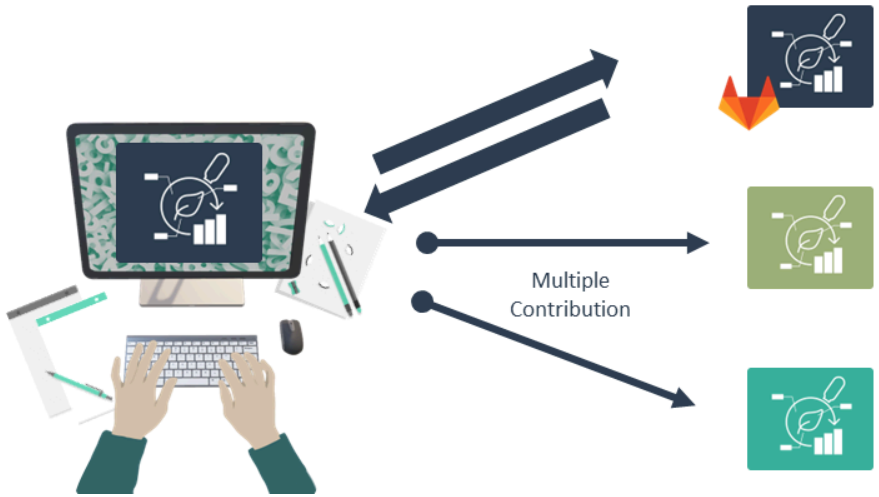


ARCs are versioned

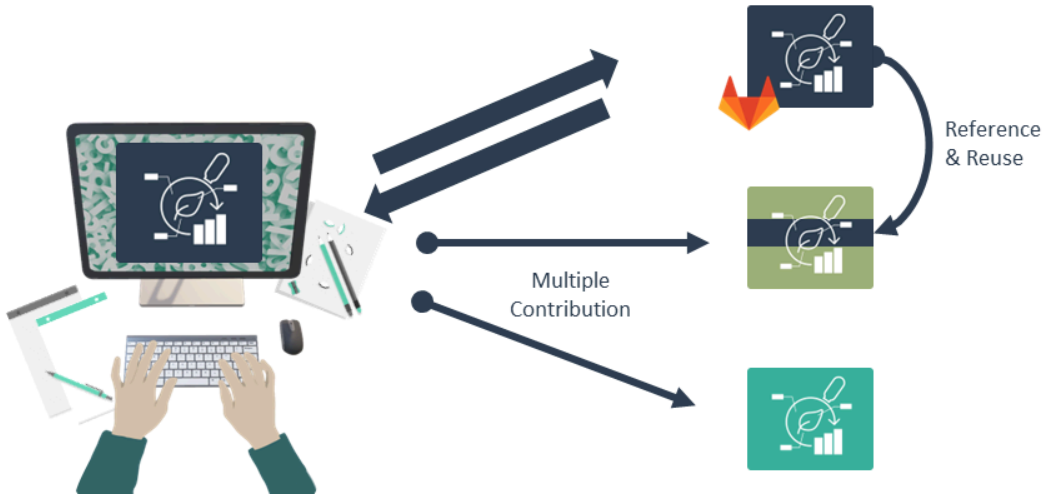


You can invite collaborators

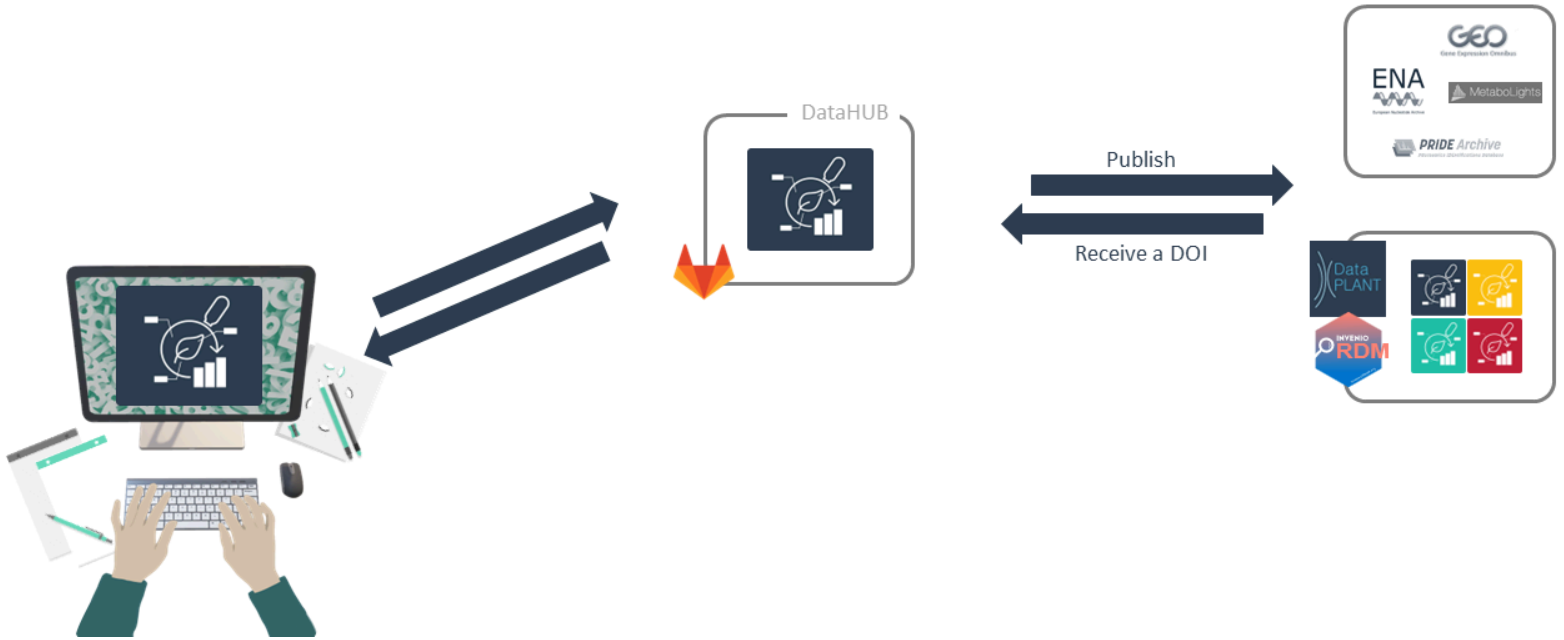
Collaborate and contribute



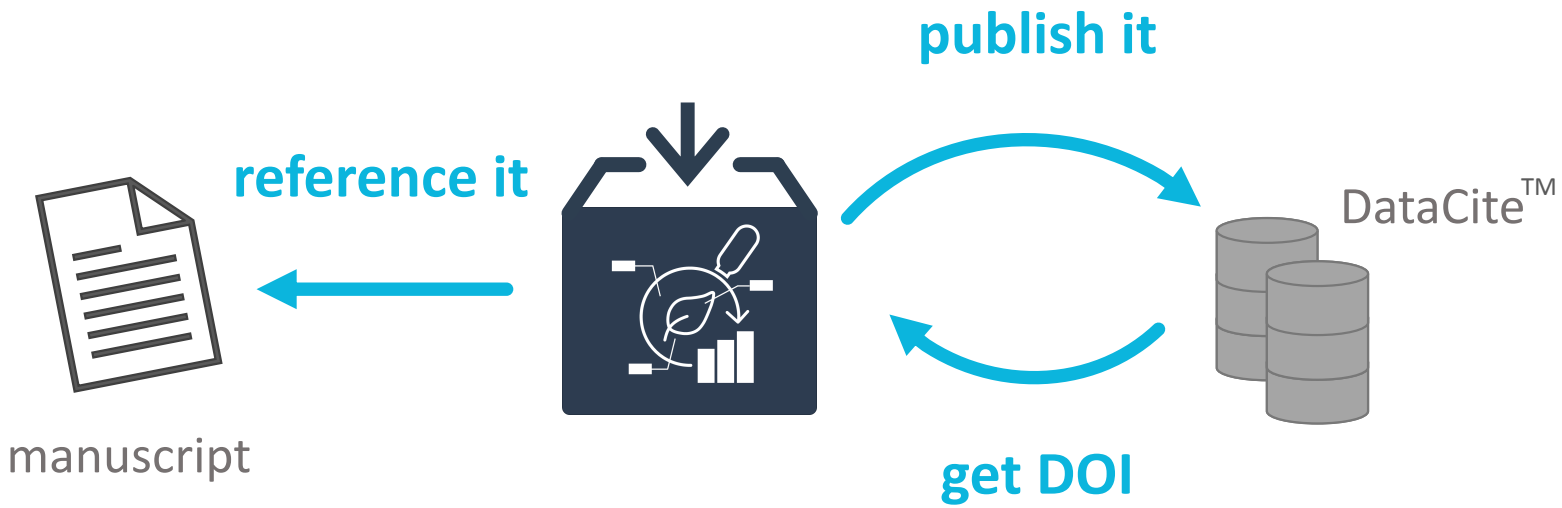
Reuse data in ARCs



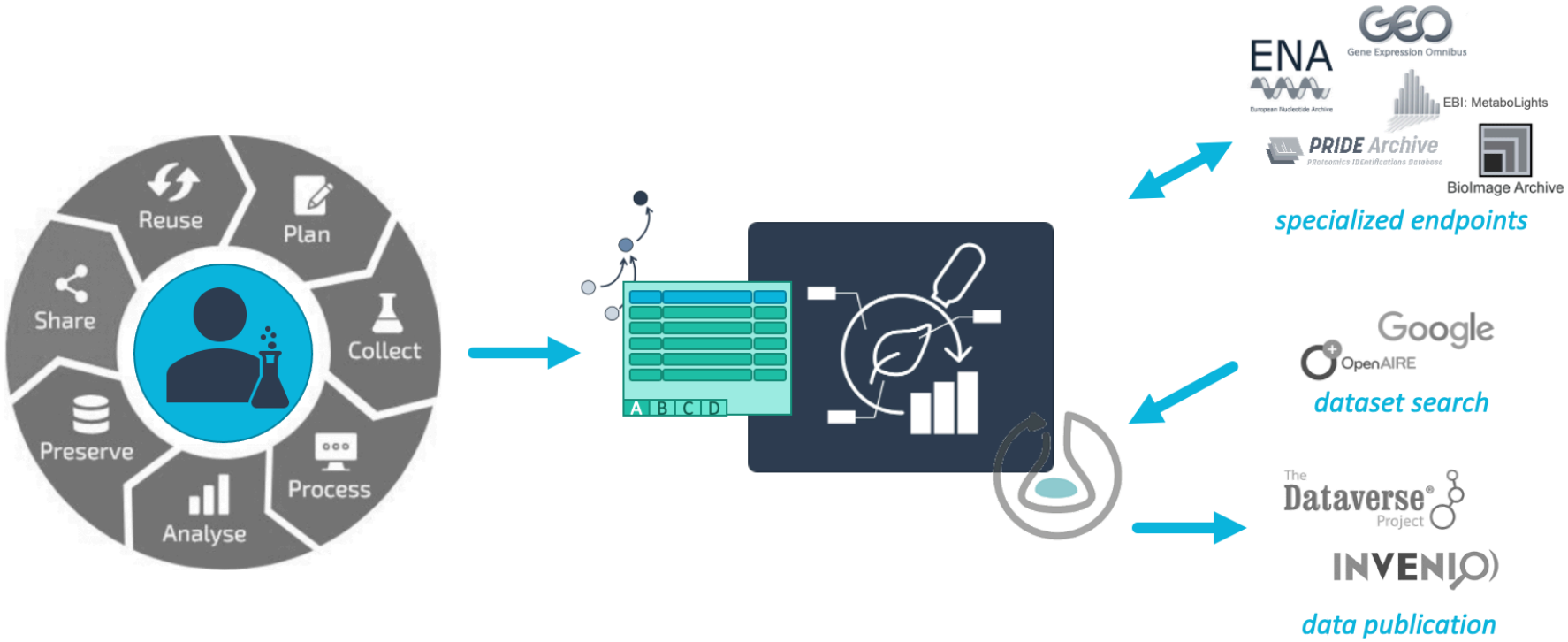
Publish your ARC



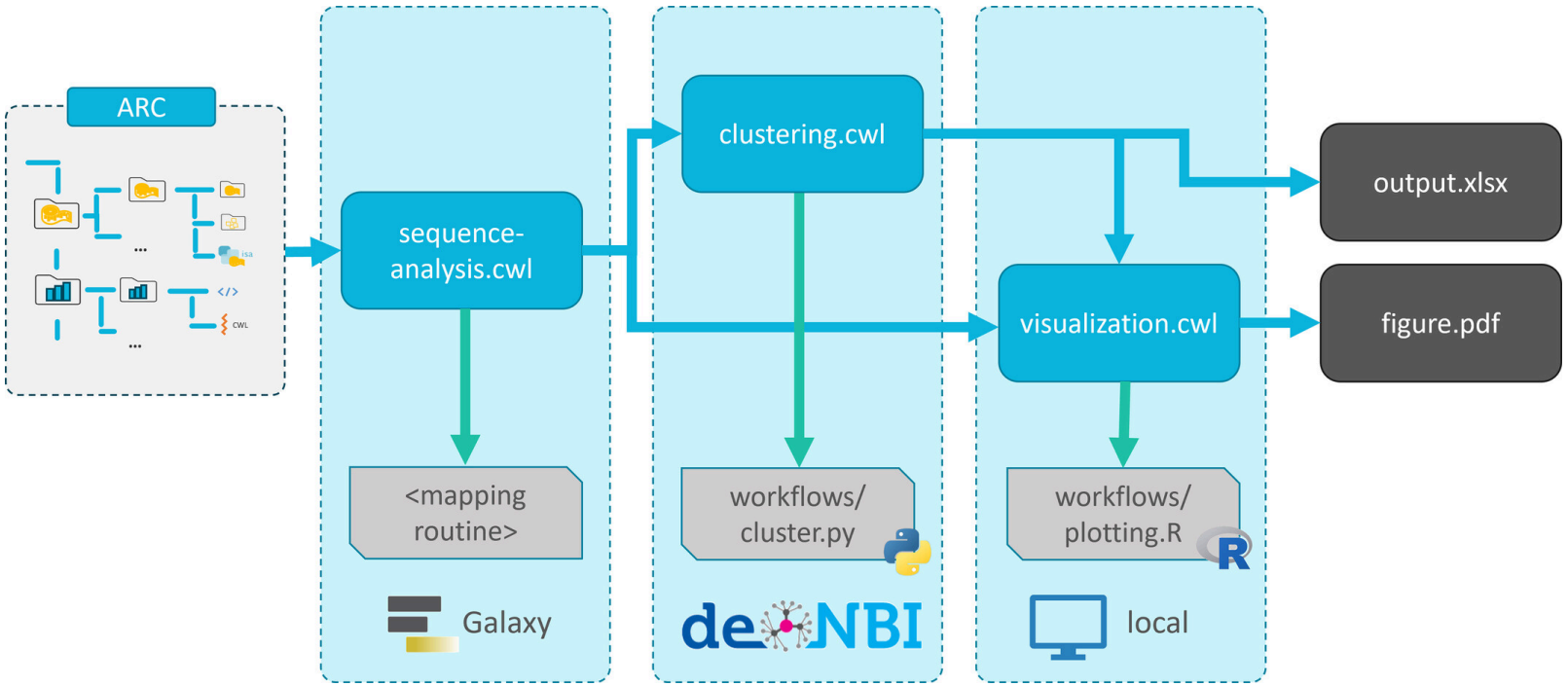
Publish your ARC, get a DOI



ARC as single-entry point



Data analysis and workflows



Weil, H.L., Schneider, K., et al. (2023), PLANTdataHUB: a collaborative platform for continuous FAIR data sharing in plant research. Plant J. https://doi.org/10.1111/tpj.16474

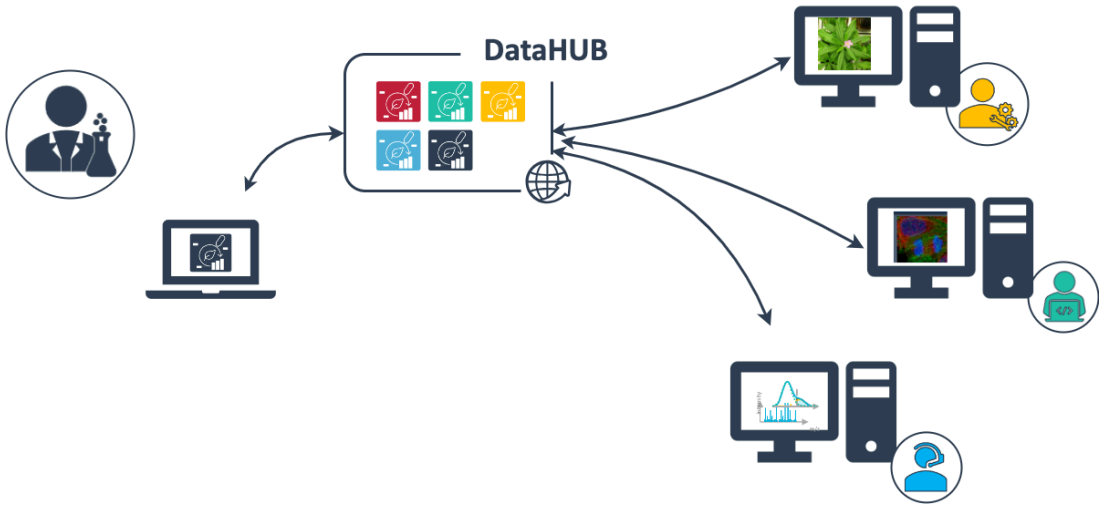
Galaxy integration: Extra value for plant research

e.g. https://plants.usegalaxy.eu

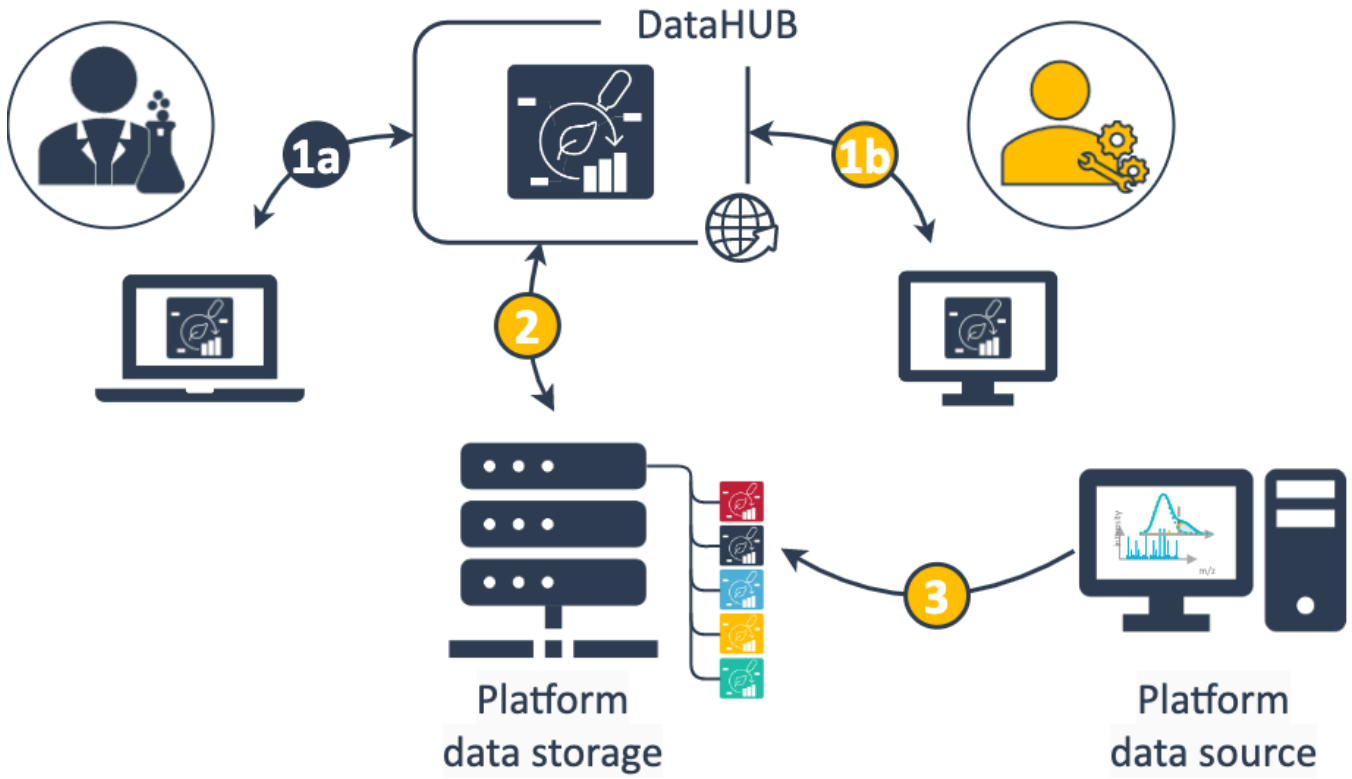
- Full ARC compatibility
- Automated metadata generation
- Specialized tools and workflows for ‘omics processing and analysis
- Public repository compatibility
- Galaxy teaching resource for data analysis

Enabling platforms

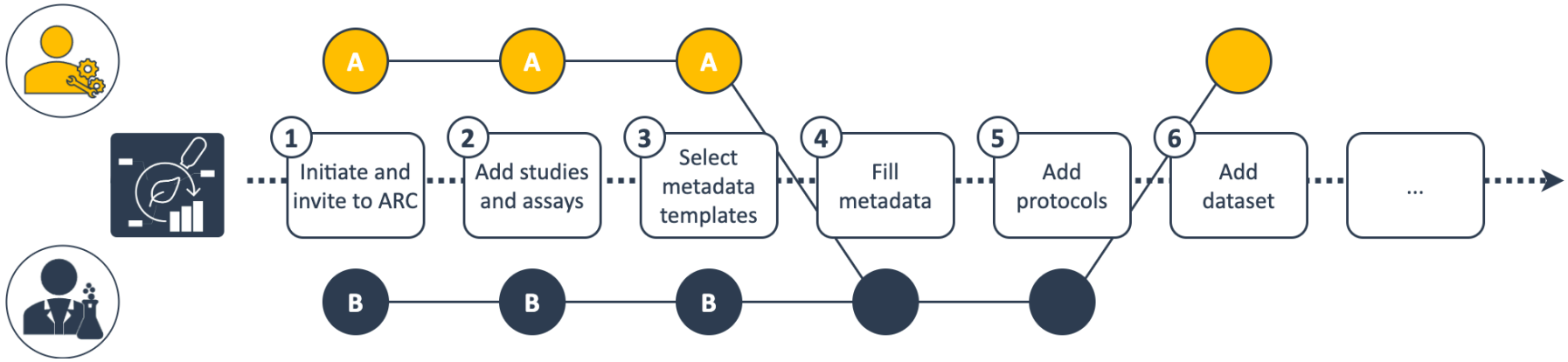
- Streamlined exchange of (meta)data
- Communication and project management



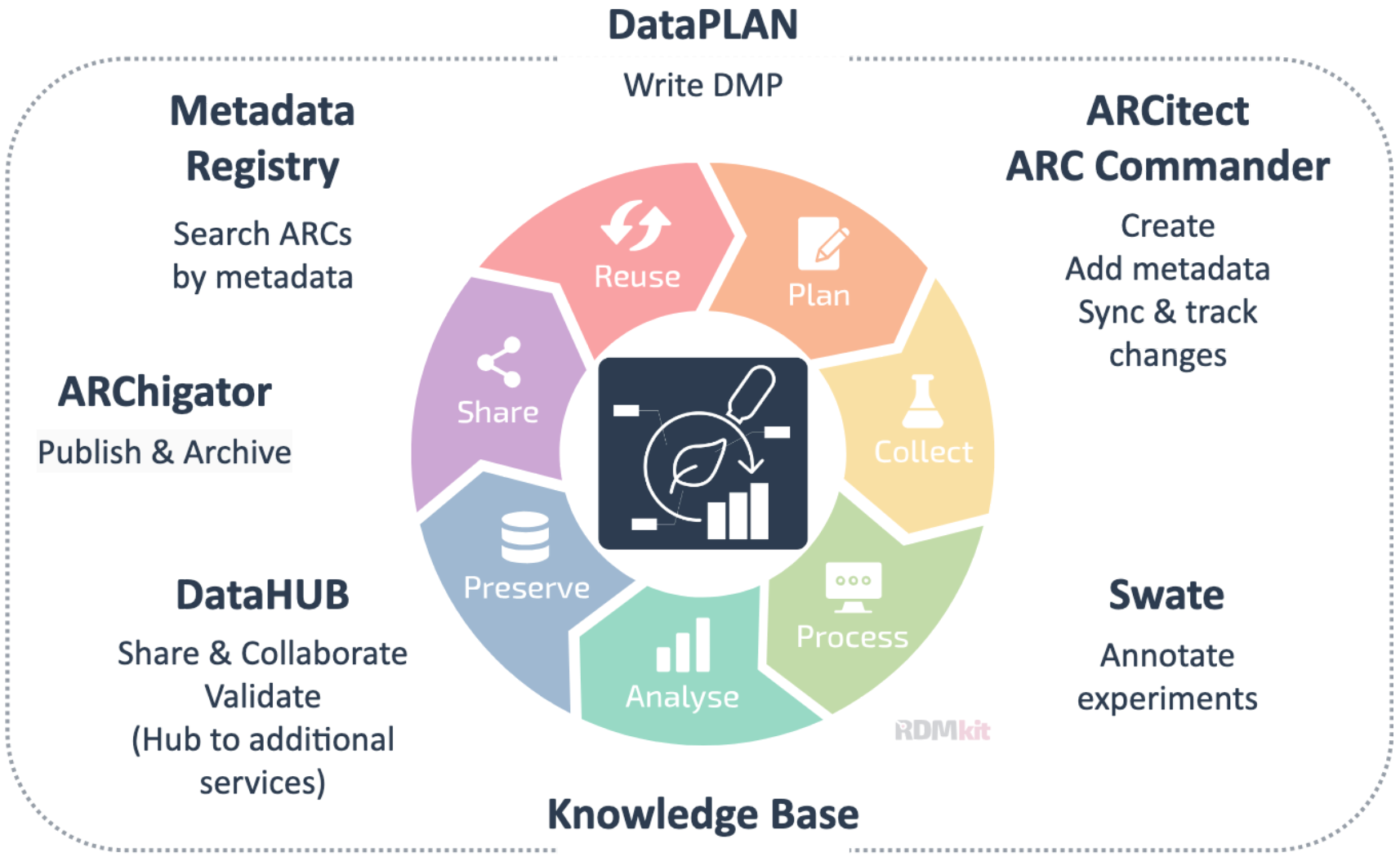
Streamlined data exchange



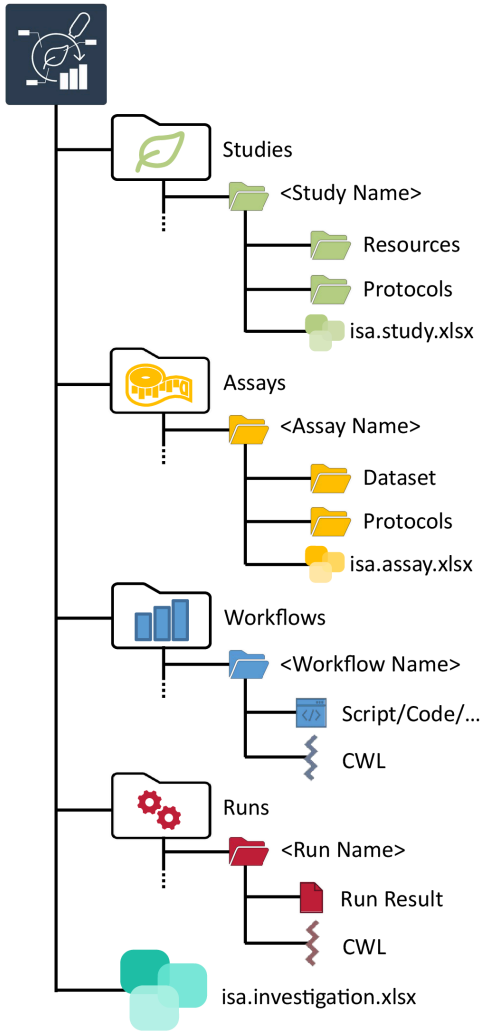
Meet your collaborators in an ARC



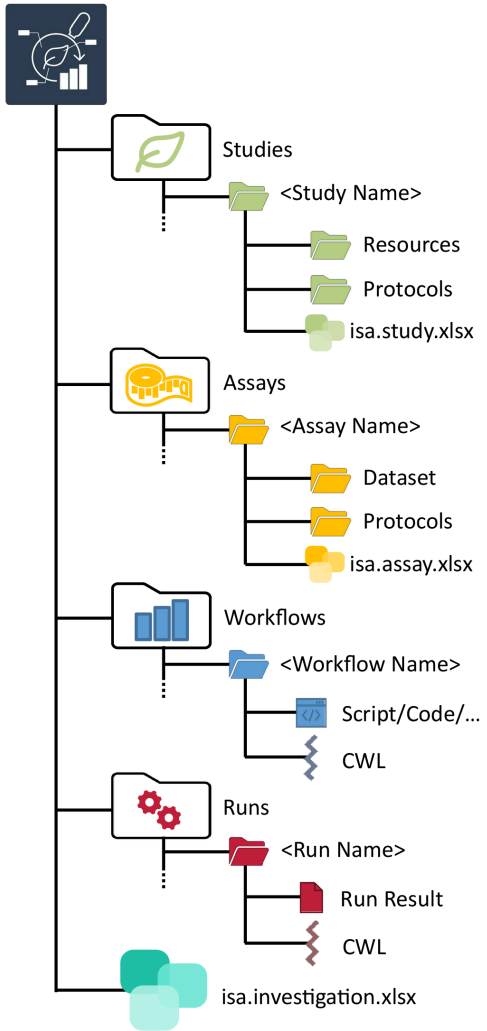
The ARC ecosystem

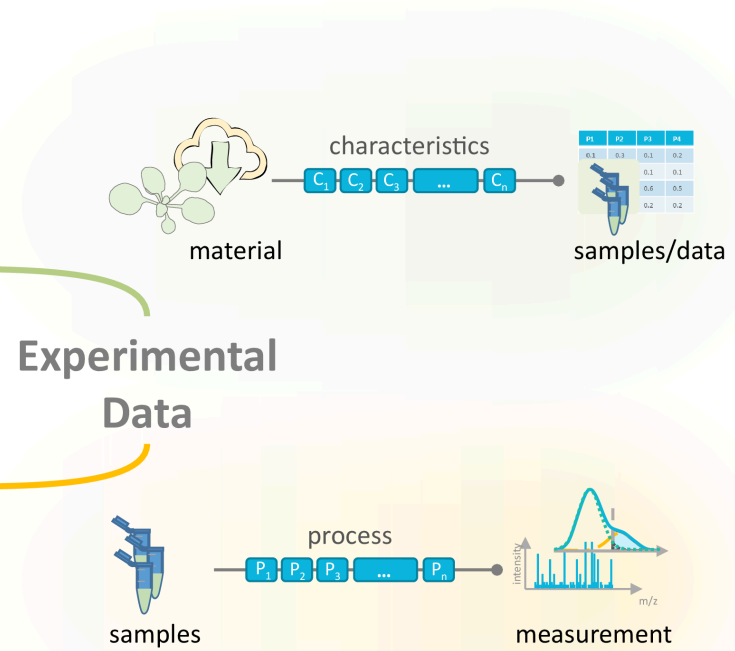


What does an ARC look like?

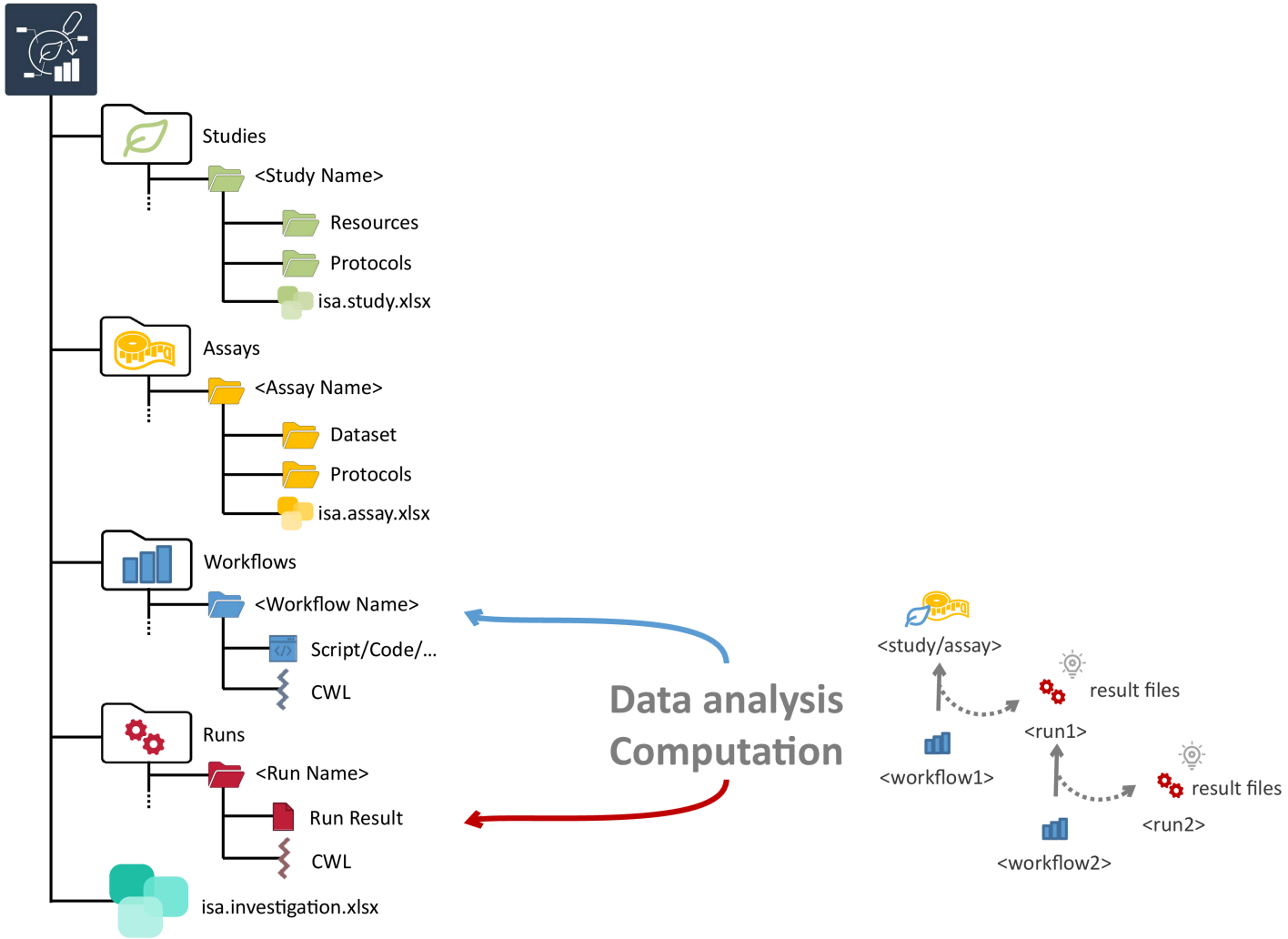


ARCs store experimental data

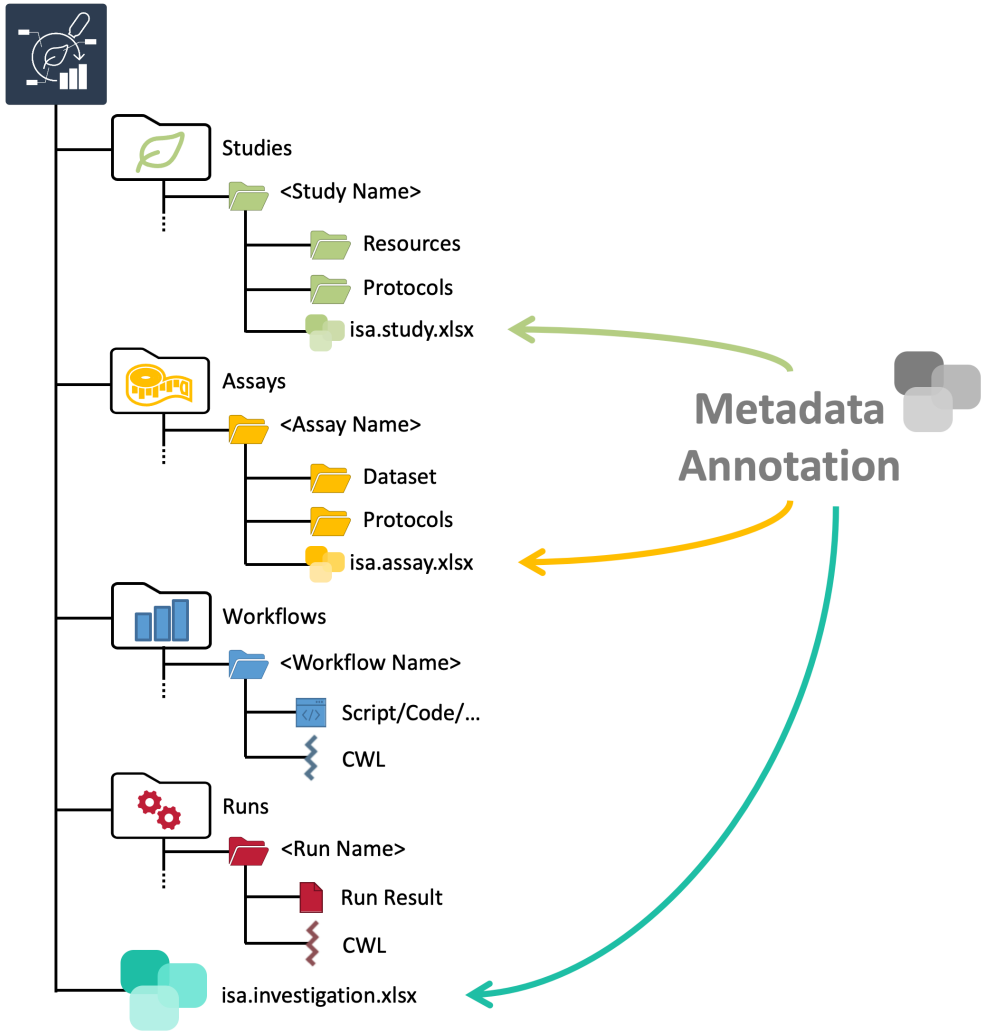




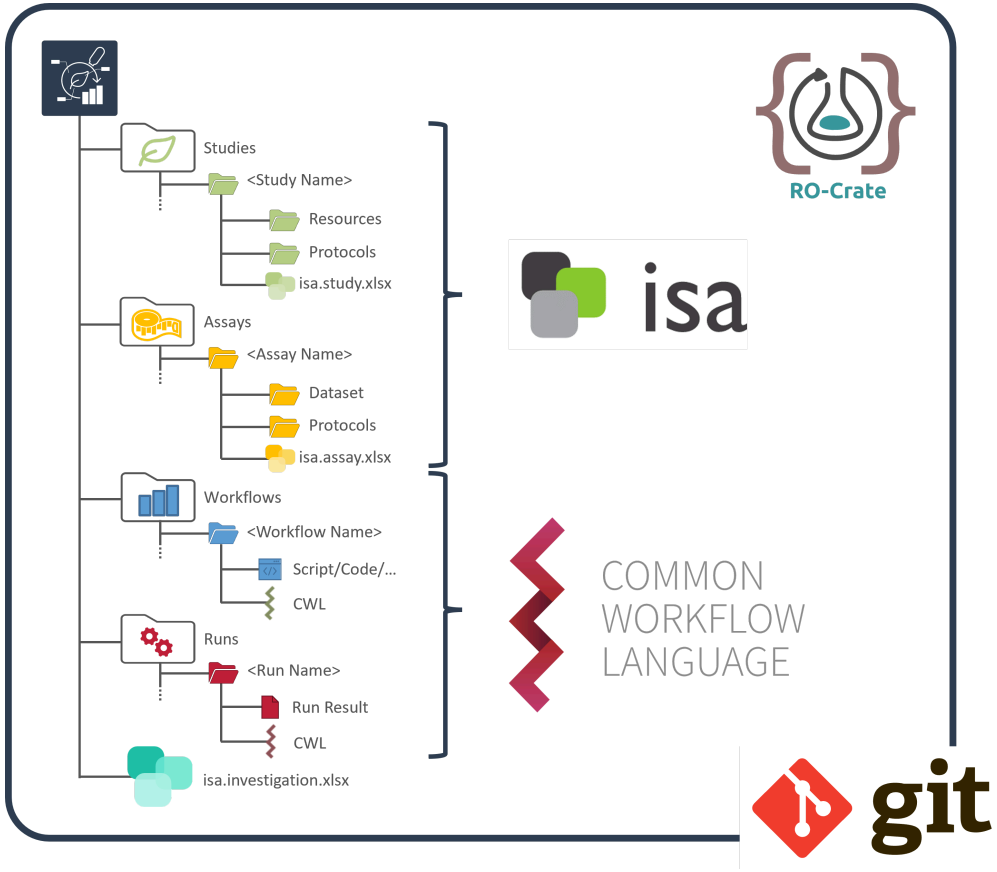
Computations can be run inside ARCs



ARCs come with comprehensive metadata



ARC builds on standards



- RO-Crate: standardized exchange
- ISA: structured, machine-readable metadata
- CWL: reproducible, re-usable data analysis
- Git: version control

https://isa-tools.org/ | https://www.commonwl.org/
https://www.researchobject.org/ro-crate/ | https://git-scm.com

ARCitect Hands-on

ARCitect installation

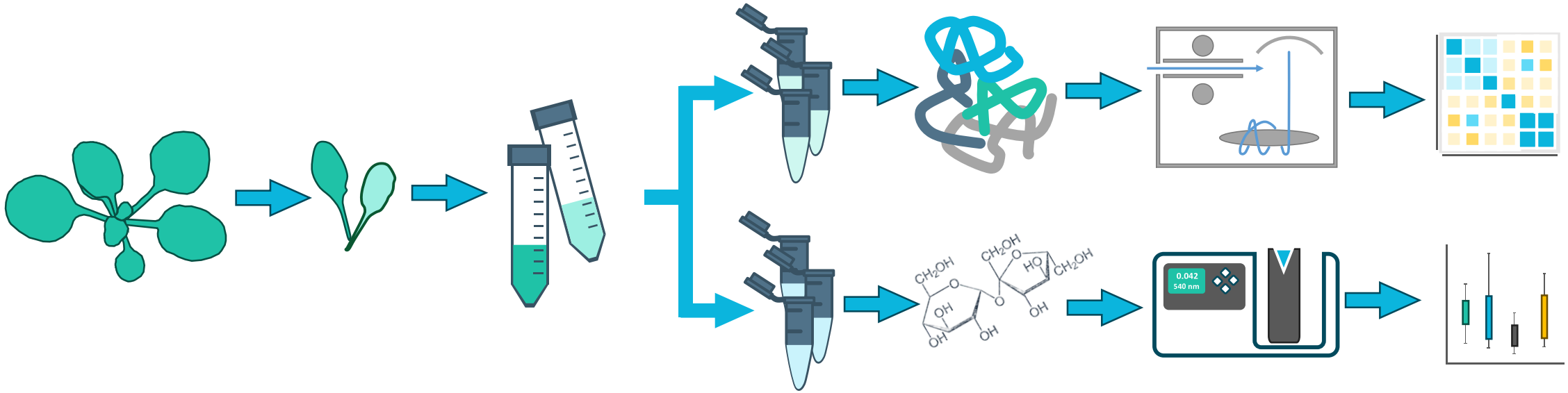
Please install version **v0.0.48** (or later) of the ARCitect:

https://github.com/nfdi4plants/ARCitect/releases/latest

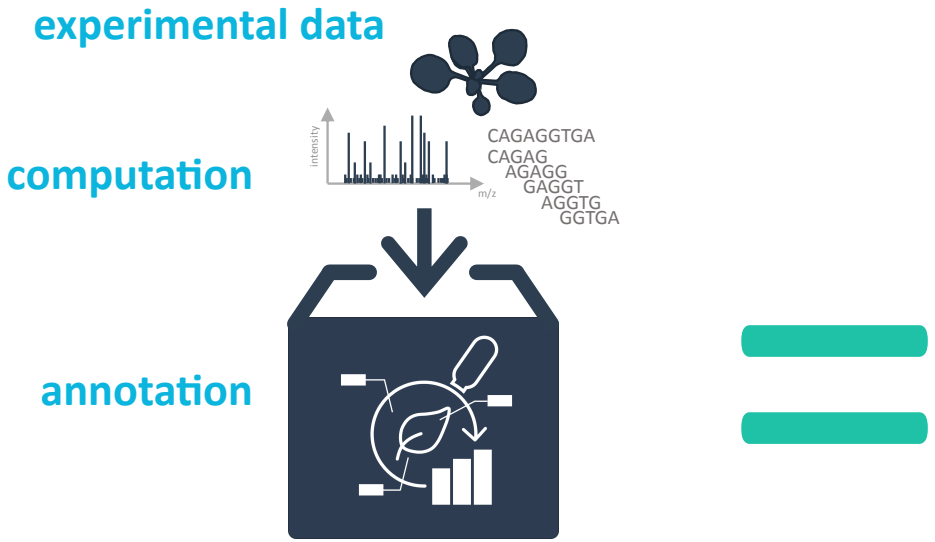
Download the demo data

https://uni-duesseldorf.sciebo.de/s/C7ms3QA6q7OZnU2

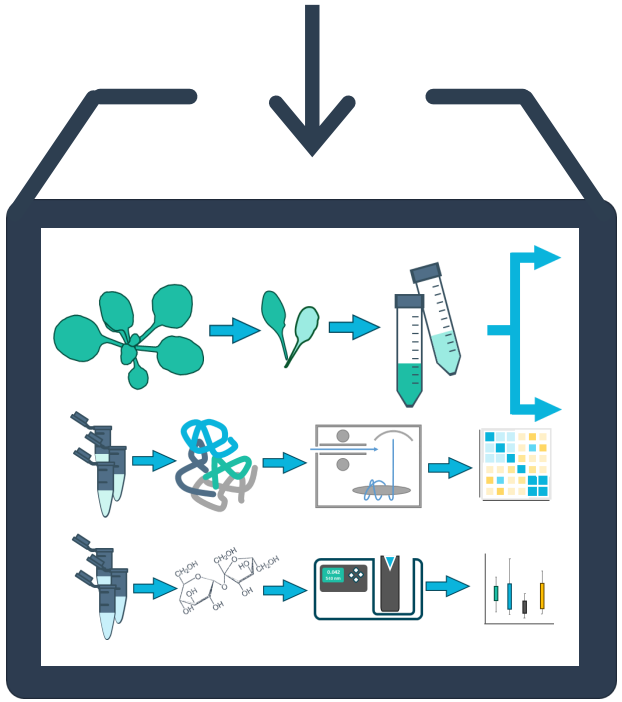
A small prototypic project



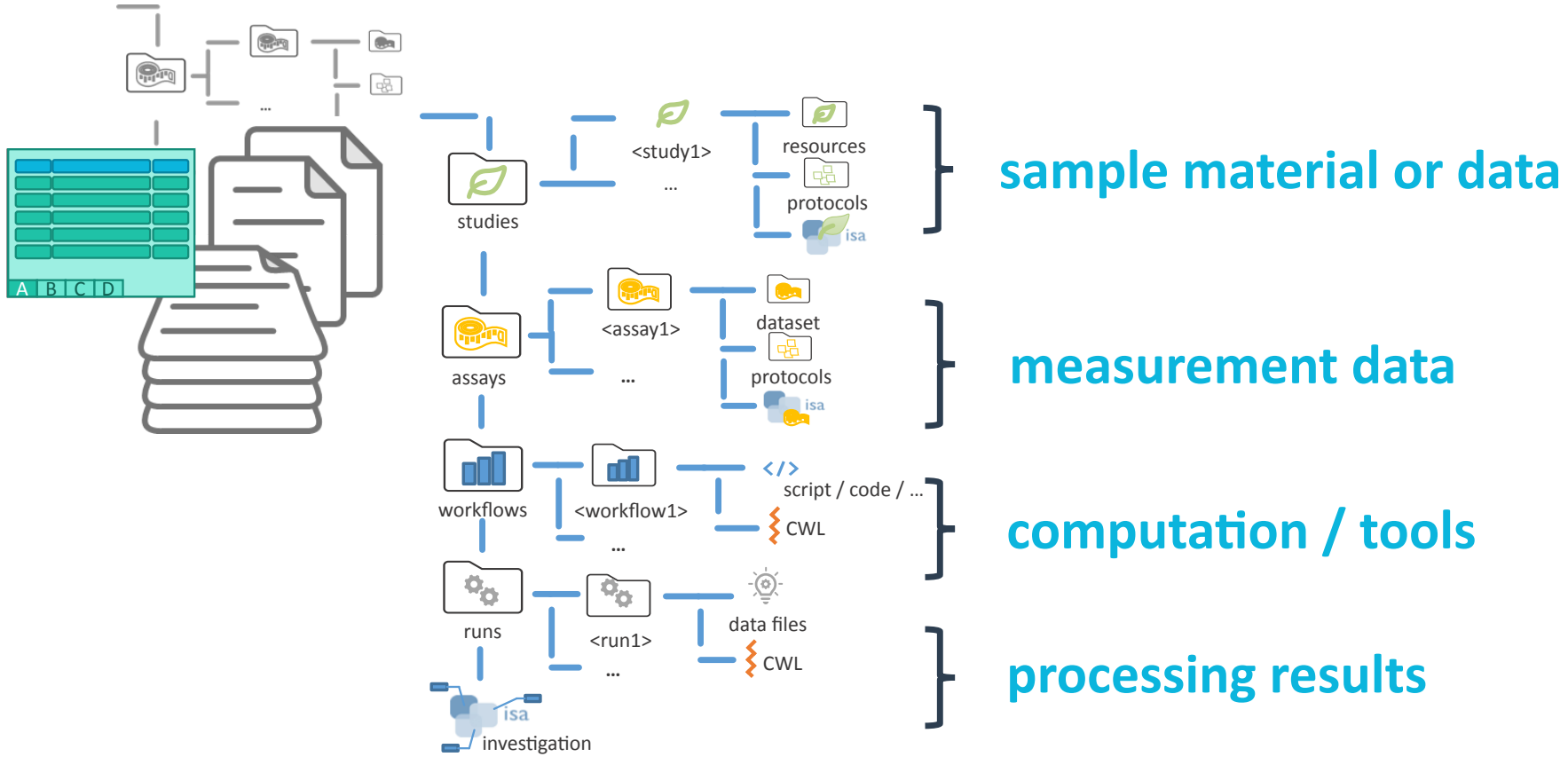
ARC: Annotated research context



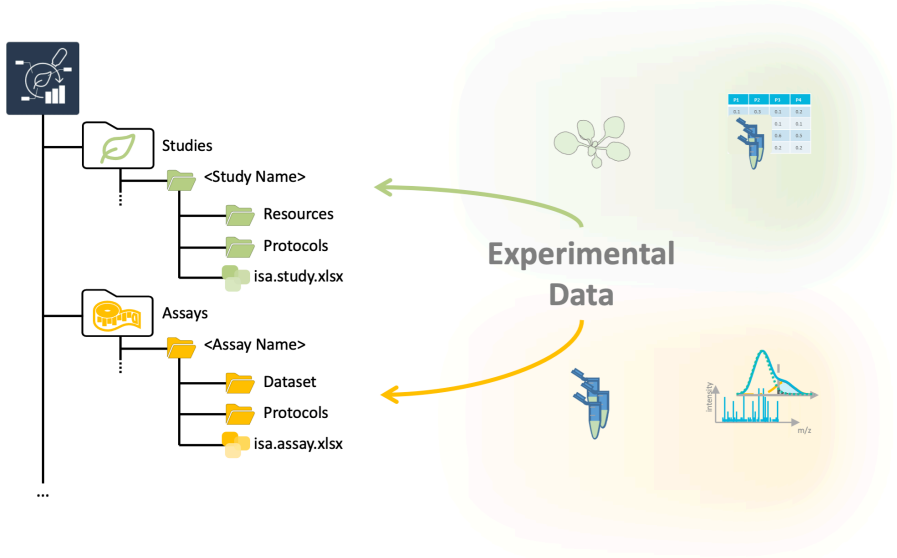
FAIR
digital (research) objects



The ARC scaffold structure



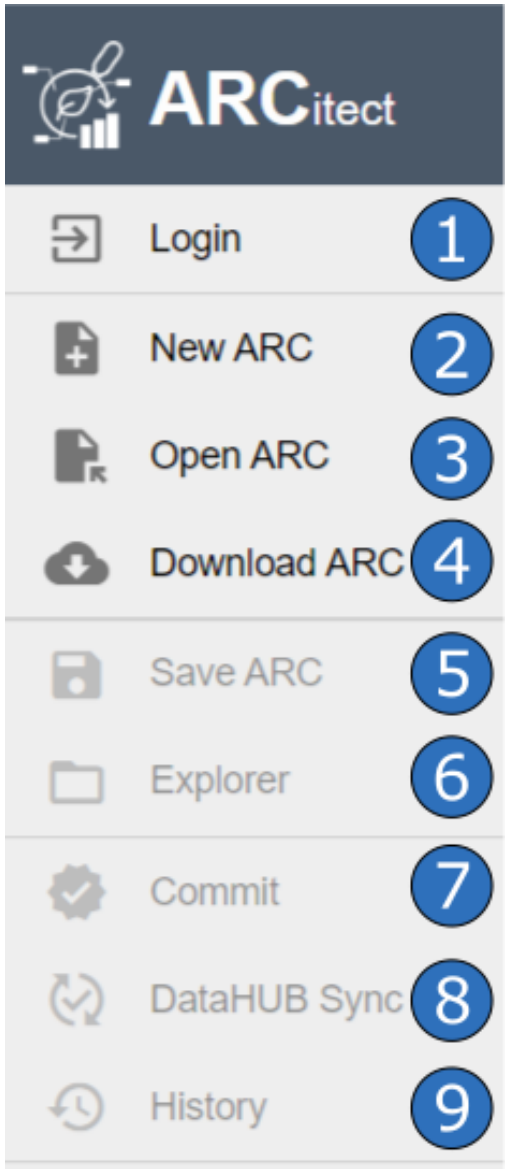
Sort Demo data in an ARC



- metabolomics_data
  - 150112_56.D
  - 150112_62.D
  - 150112_66.D
  - 150115_12.D
  - 150115_14.D
  - 150115_16.D
  - gcms_samplelist.tsv
  - method_gcms.txt
  - sample_submission_gcms.csv
- methods
  - Illumina_libraries.txt
  - metabolite_extraction.txt
  - plant_material.txt
  - RNA_extraction.txt
- rnaseq_data
  - DB_097_CAGATC_L001_R1_001.fastq.gz
  - DB_099_CTTGTA_L001_R1_001.fastq.gz
  - DB_103_AGTCAA_L001_R1_001.fastq.gz
  - DB_161_GTCCGC_L001_R1_001.fastq.gz
  - DB_163_GTGAAA_L001_R1_001.fastq.gz
  - DB_165_GTGAAA_L002_R1_001.fastq.gz
  - NGS_SampleSheet.xlsx

Initiate the ARC folder structure

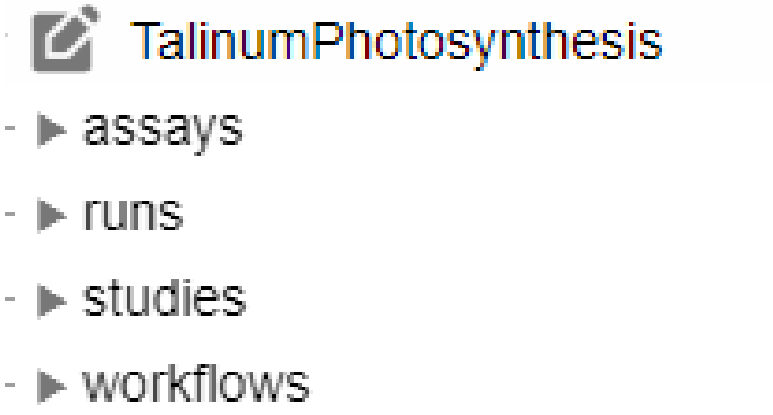
1. Open ARCitect
2. Create a **New ARC** (2)
3. Select a location and name it
   TalinumPhotosynthesis



Your ARC's name

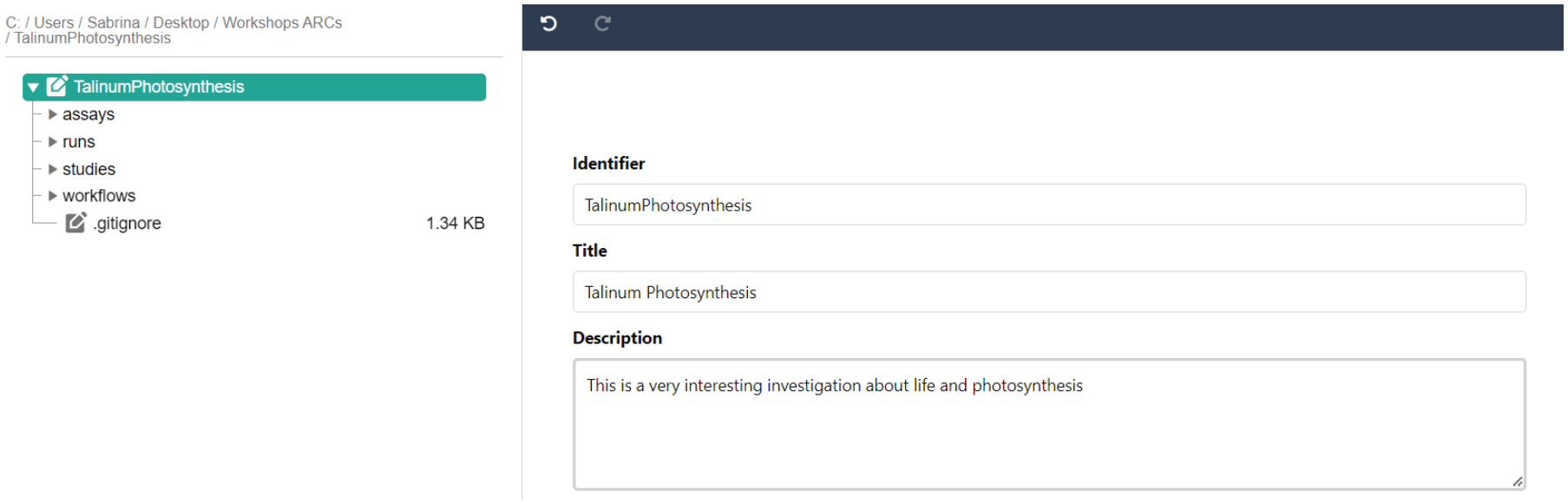
💡 By default, your ARC's name will be used

- for the ARC folder on your machine
- to create your ARC in the DataHUB at `https://git.nfdi4plants.org/<YourUserName>/<YourARC>` (see next steps)
- as the identifier for your investigation



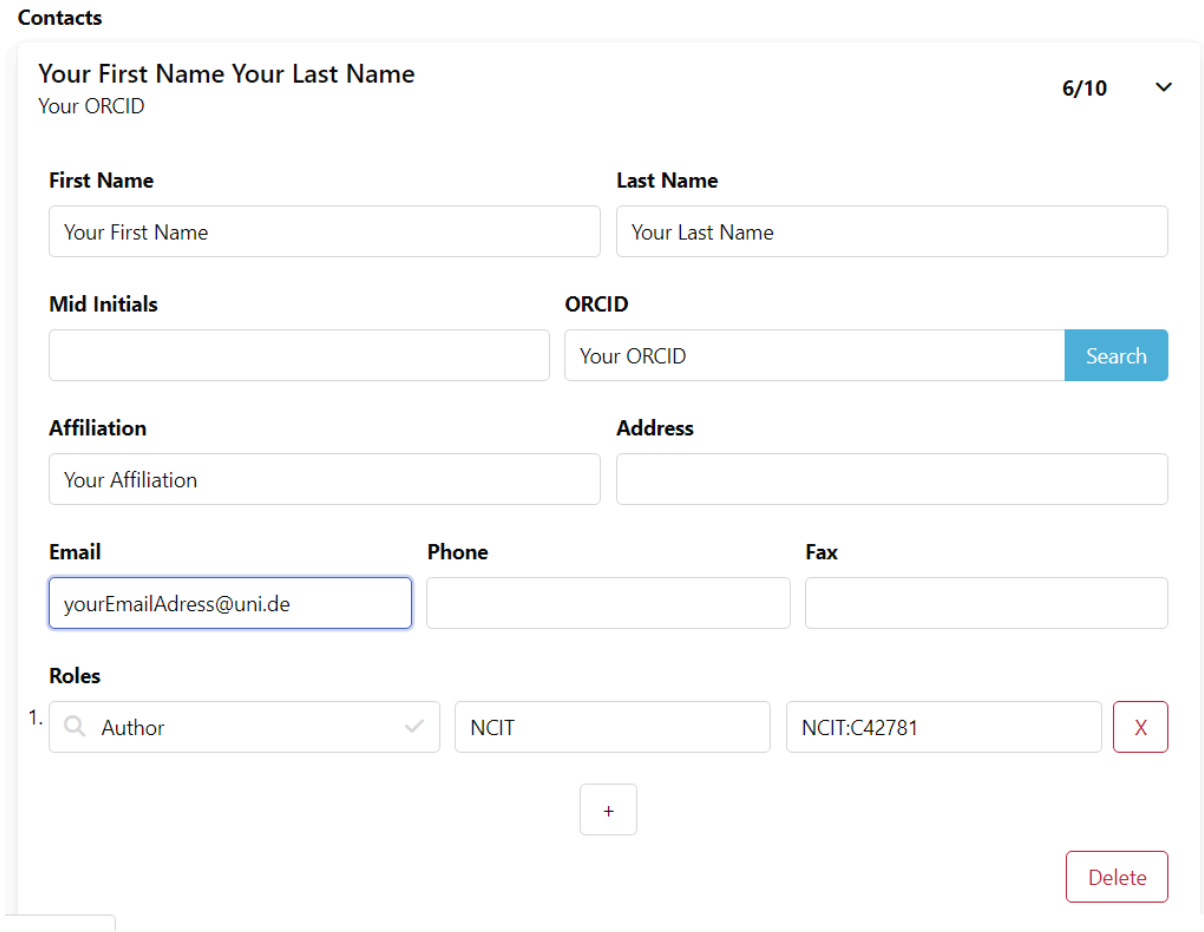
💡 Make sure that no ARC exists at `https://git.nfdi4plants.org/<YourUserName>/<YourARC>`.
Otherwise you will sync to that ARC.

Add a description to your investigation

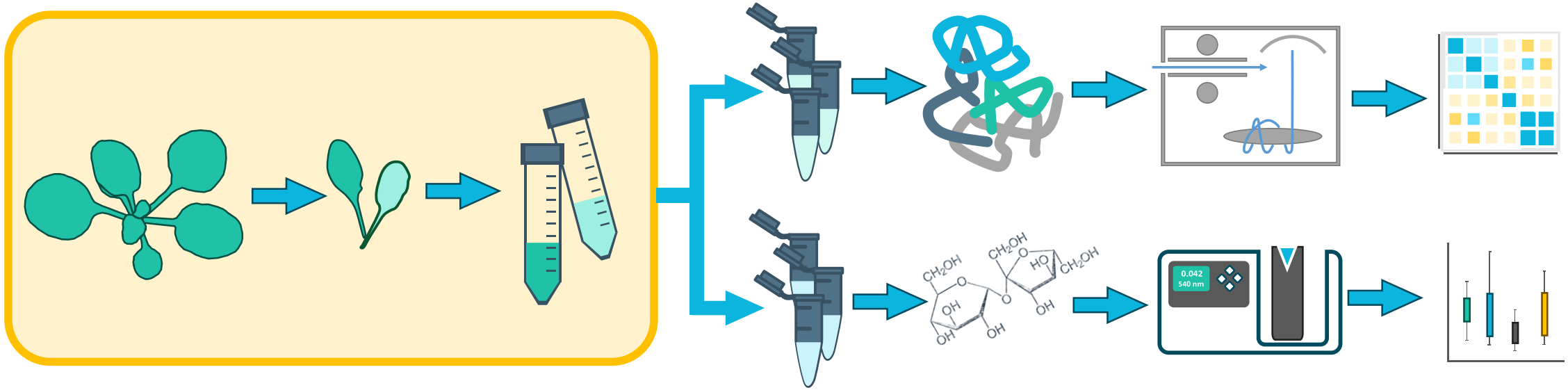


Add contributors

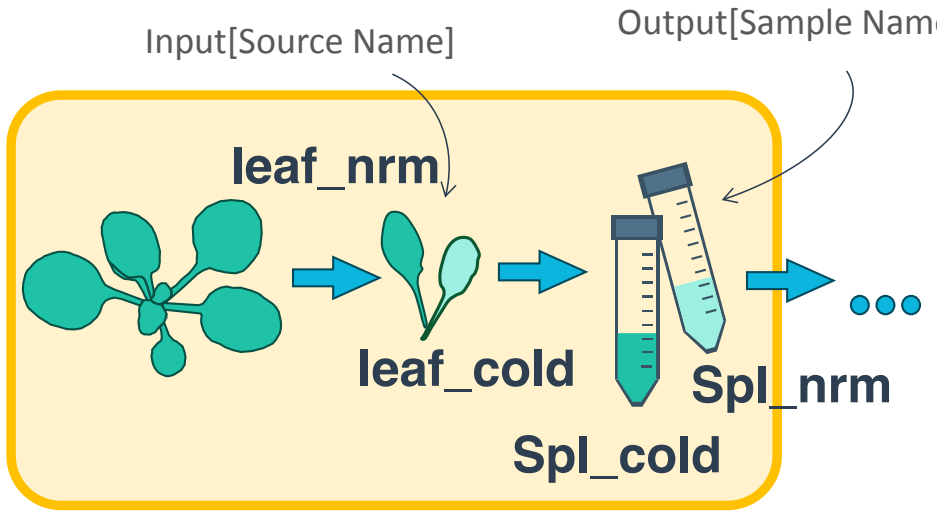
- you could also add the contacts directly via ORCID

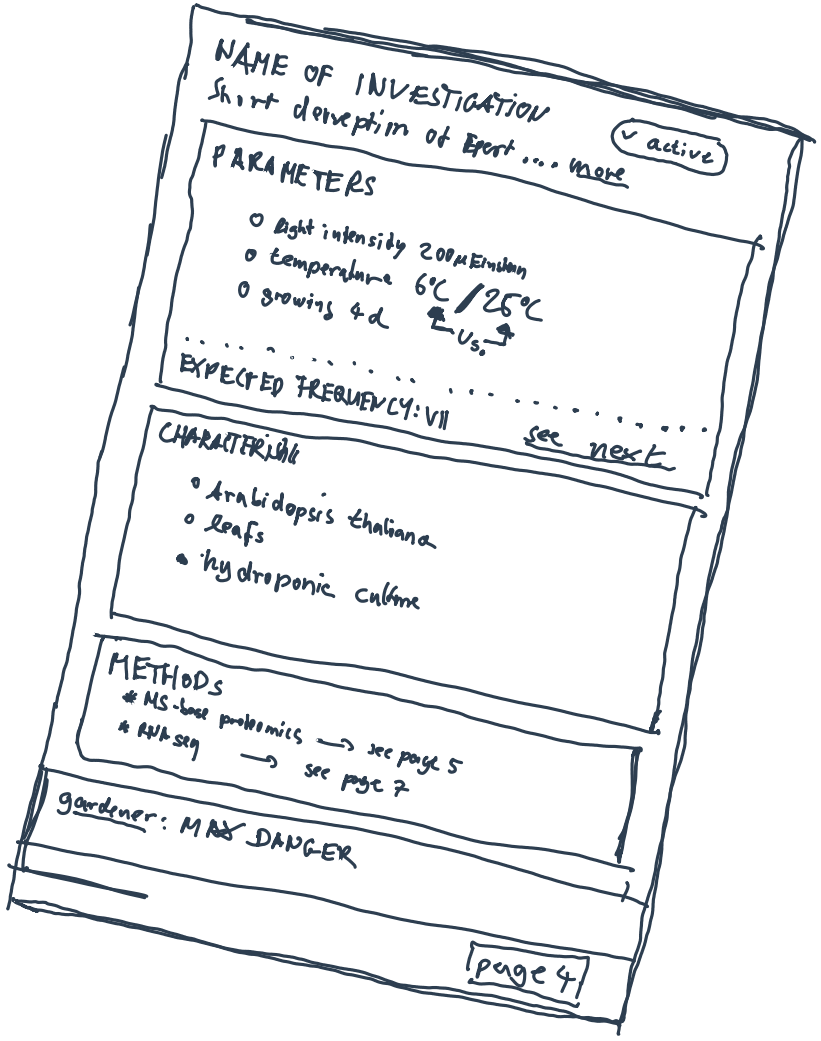


Divide and conquer for reproducibility



Identifying the ‘study’ part





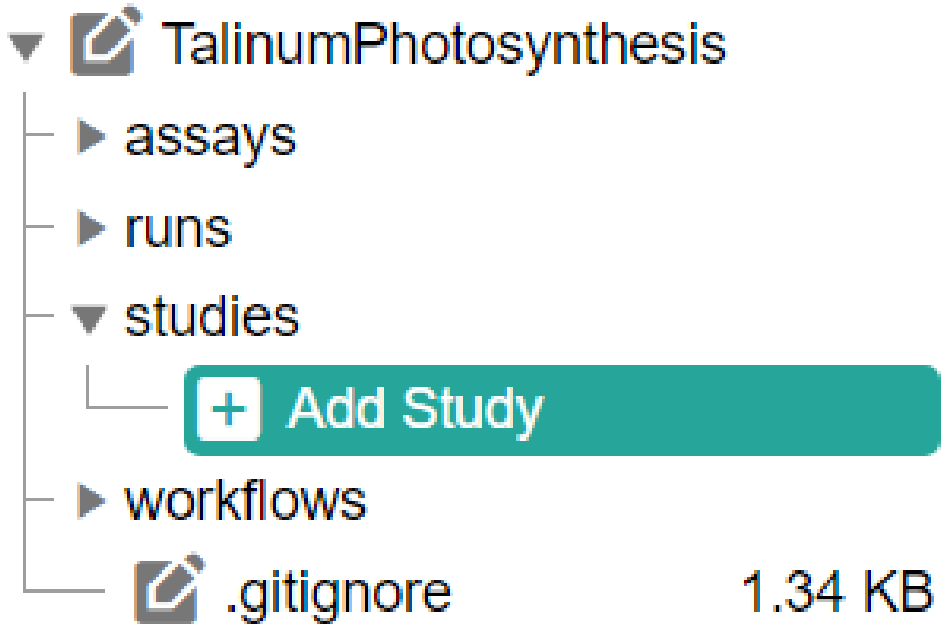
Protocol REF

growth_protocol.txt

Add a study

by clicking "Add Study" and entering an identifier for your study

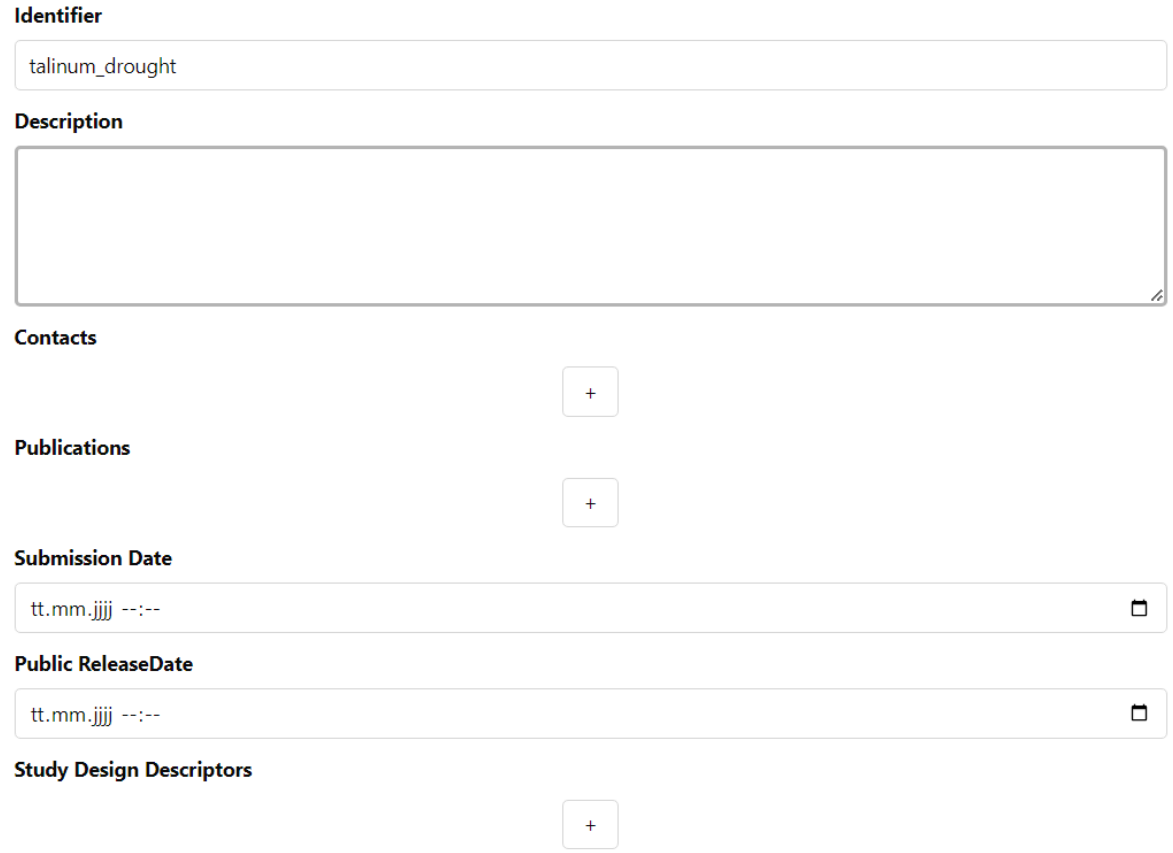
Use **talinum_drought** as an identifier



Study panel

In the study panel you can add

- general metadata,
- people, and
- publications
- data process information

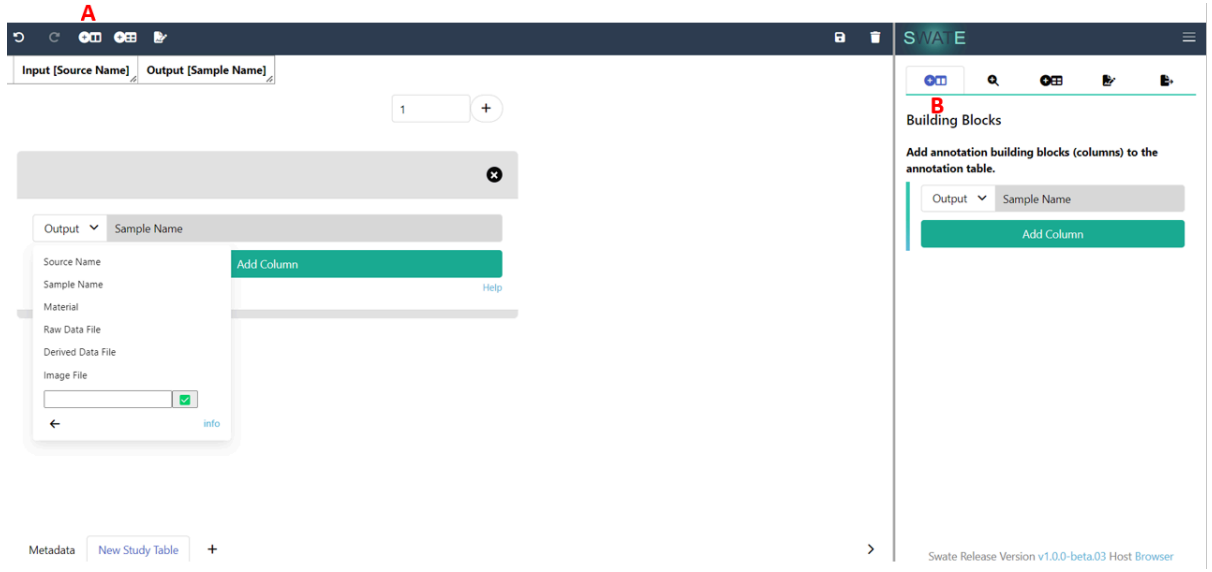


Let's annotate the plant samples

1. Check out the lab notes
   `studies/talinum_drought/protocols/plant_material.txt`
2. Select the study `talinum_drought`
3. Add a new table sheet at the bottom

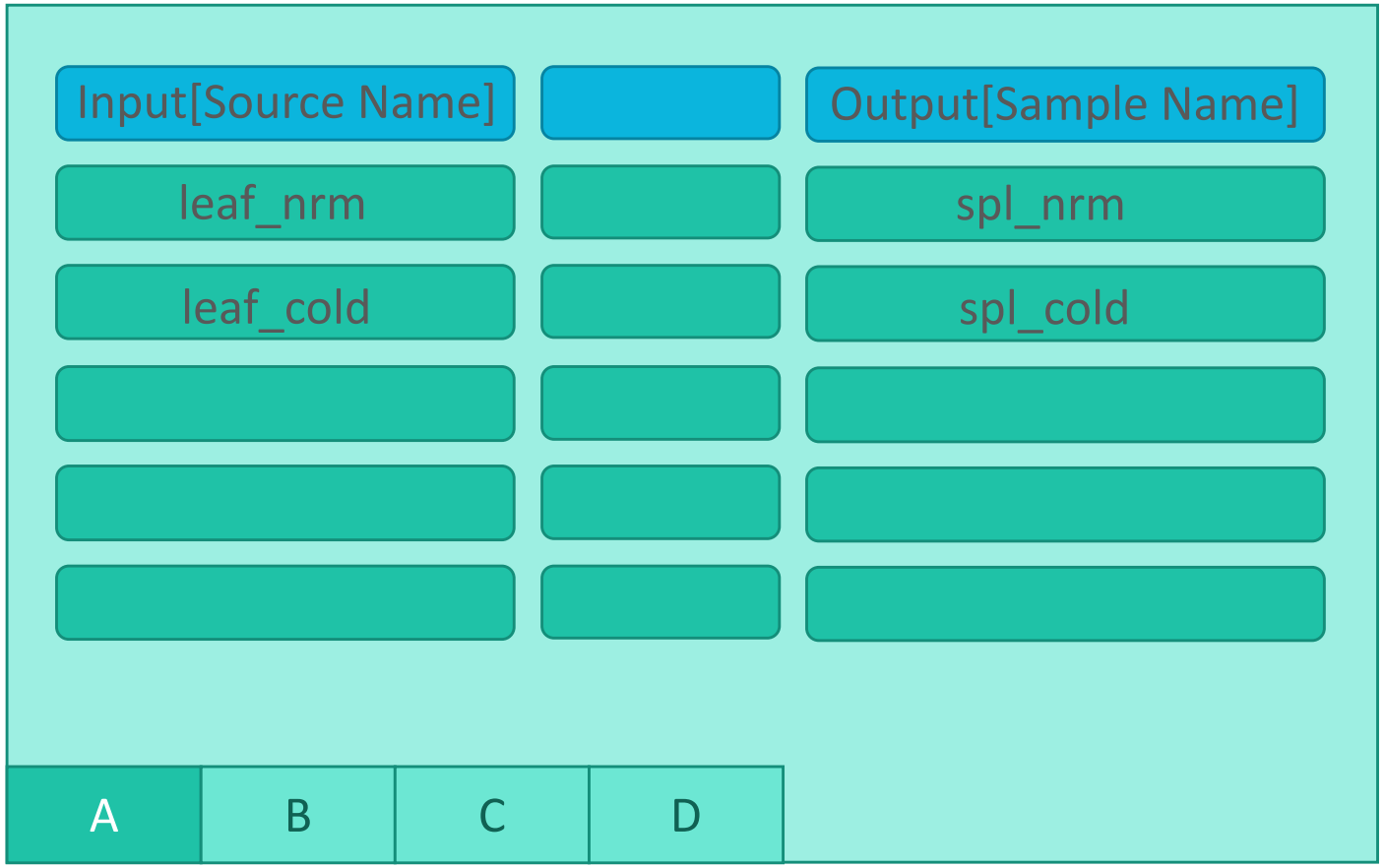
Create an annotation table

1. Create a annotation table by adding **Building blocks** via the widget (A) or the sidebar (B)
2. Add an *Input* (`Source Name`) and *Output* (`Sample Name`) column



💡 Each table can contain only one *Input* and one *Output* column

A table-based organization schema



Fill out source name and sample name

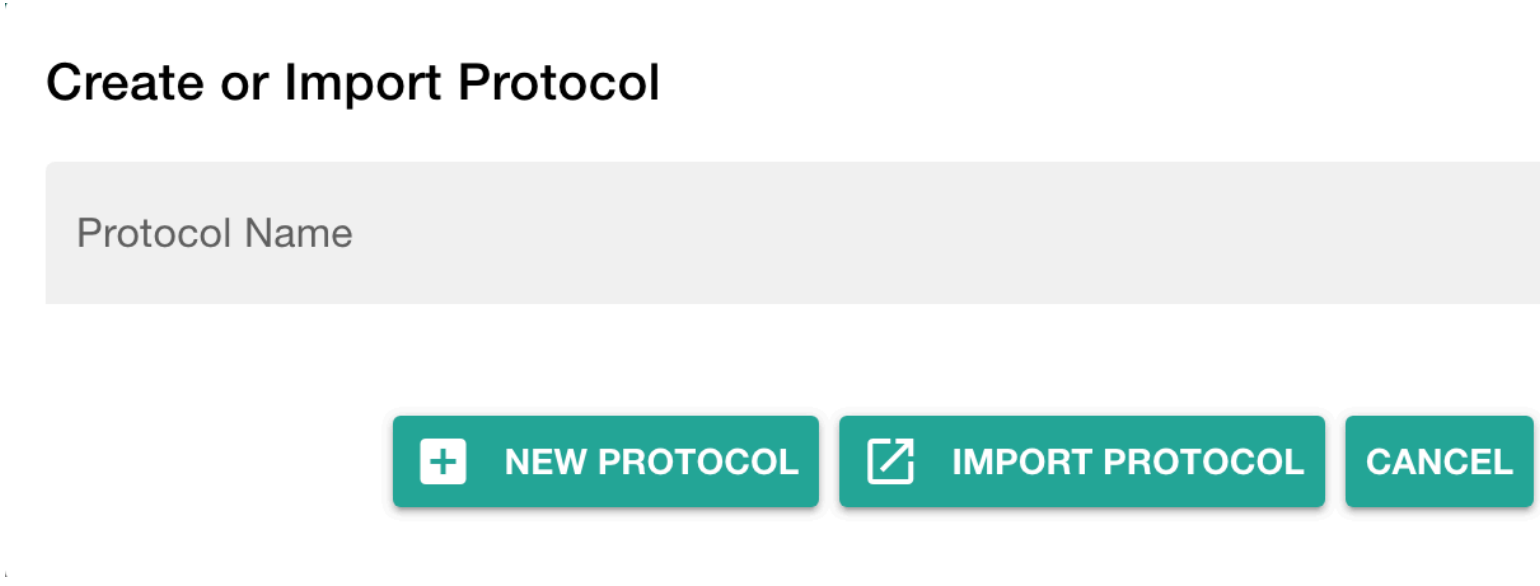
Transfer the sample ids from the protocol.

1. Invent names for `Source Name` (we do not have this information)
2. Use the sample names from the protocol (DB_*) as `Sample Name`

Add protocols

You can either

- directly write a **new protocol** within the ARCitect or
- import an existing one from your computer

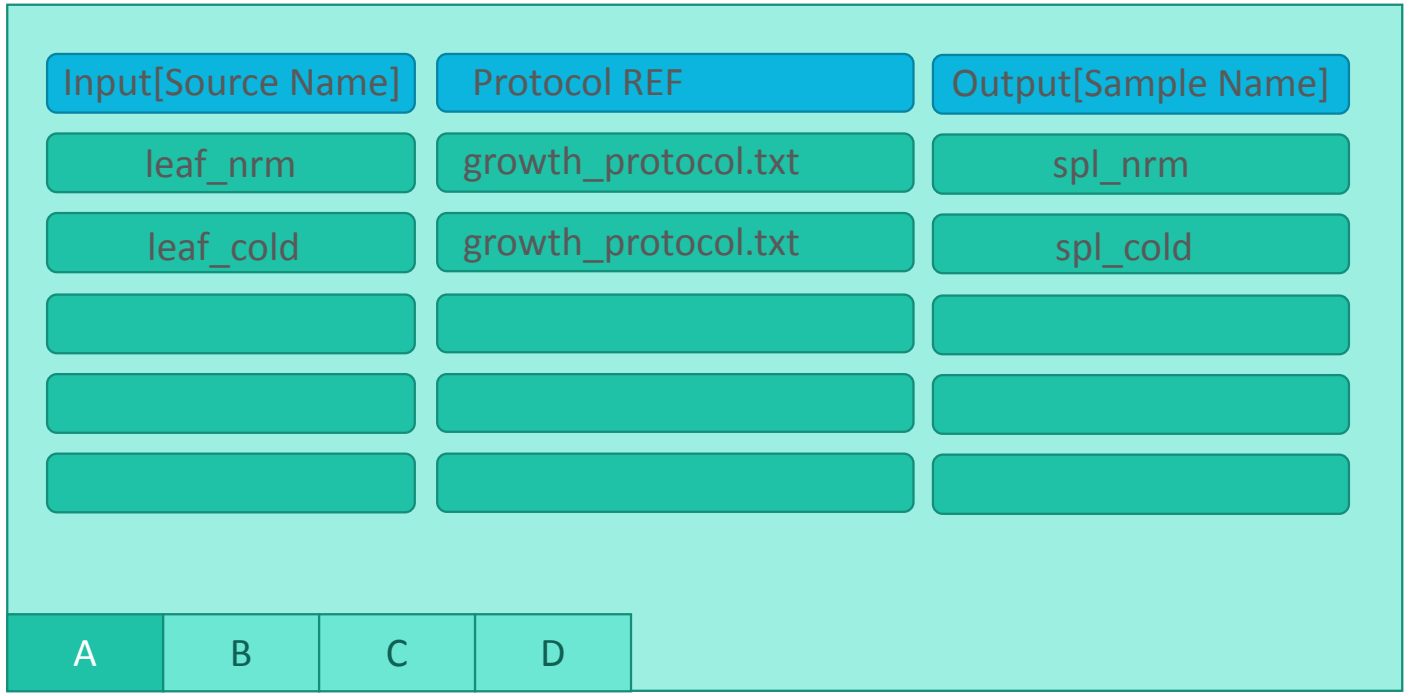


Link the protocol to the isa table

1. In the *Building Blocks* tab, select *More -> Protocol REF*.
2. Click `Add Column` .
3. Add the name of the protocol file (plant_material.txt) to the *Protocol REF* column.

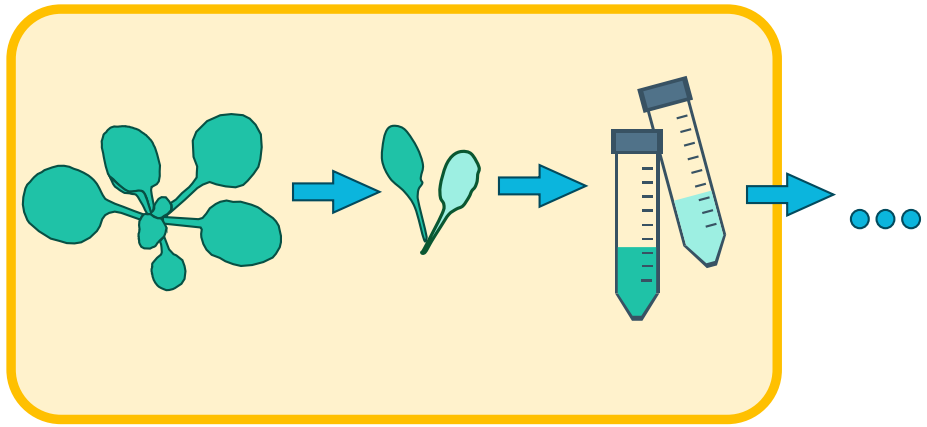
Referencing a protocol

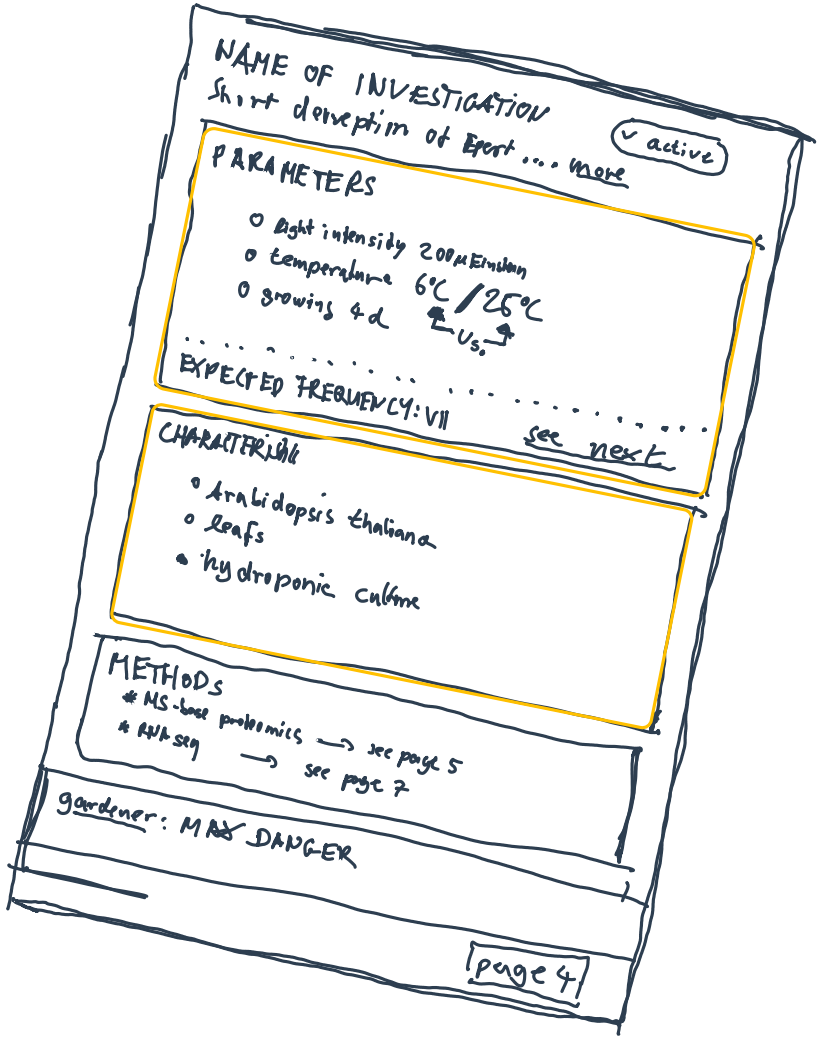
This allows you to reference the free-text, human-readable protocol.



- It is recommended that the protocol is in an open format (.md|.txt|.docx|…)
- But everything is possible also an URI to an electronic lab notebook

Parameterizing the ‘study’





Finding the right metadata vocabulary

Parameters []

- Light intensity 200 µEinstein
- Temperature 6°C / 25°C
- Growing 4d

Characteristics []

- Arabidopsis thaliana
- Leaf
- Hydroponic culture
- Columbia

OLS: Finding the right metadata vocabulary

Temperature Dependence `Temperature:Dependence_Annotation`

http://purl.uniprot.org/core/Temperature_Dependence_Annotation

Indicates the optimum temperature for enzyme activity and/or the variation of enzyme activity with temperature variation; the thermostability/thermolability of the enzyme is also mentioned when it is known.

Ontology: `UNIPROTRDFS`

temperature `AFO:/result#AFR_0001584`

http://purl.allotrope.org/ontologies/result#AFR_0001584

A temperature (datum) is a quantity facet that quantifies some temperature. [Allotrope]

Ontology: `AFO`

temperature `FBcv:0000466`

http://purl.obolibrary.org/obo/FBcv_0000466

Mutation caused by exposure to a temperature that is higher or lower than 25 degrees Celsius.

Ontology: `FBCV`

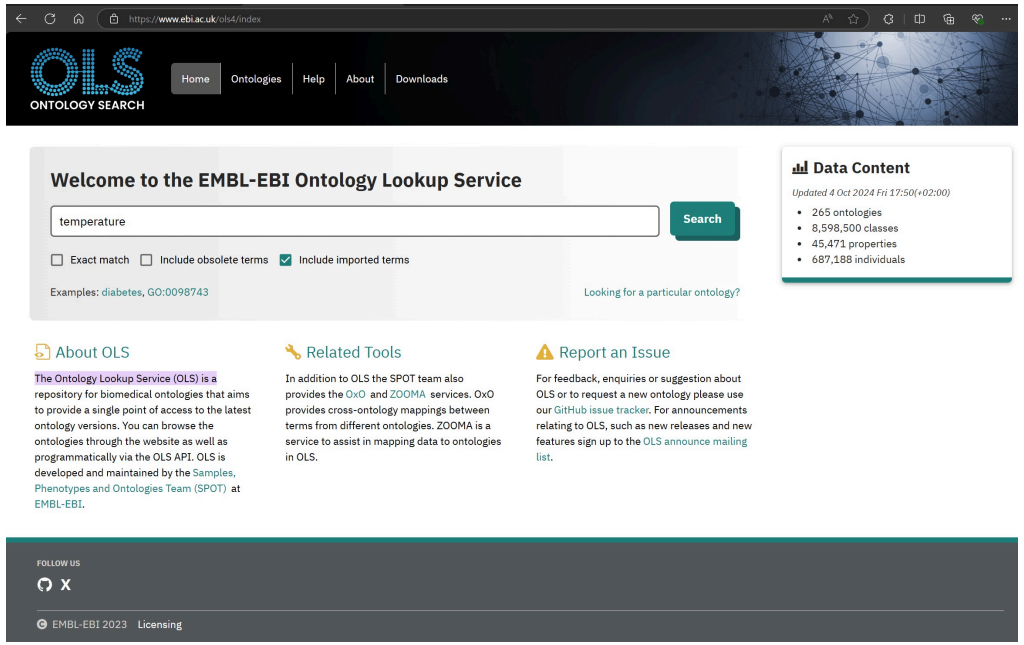
temperature `PATO:0000146`

http://purl.obolibrary.org/obo/PATO_0000146

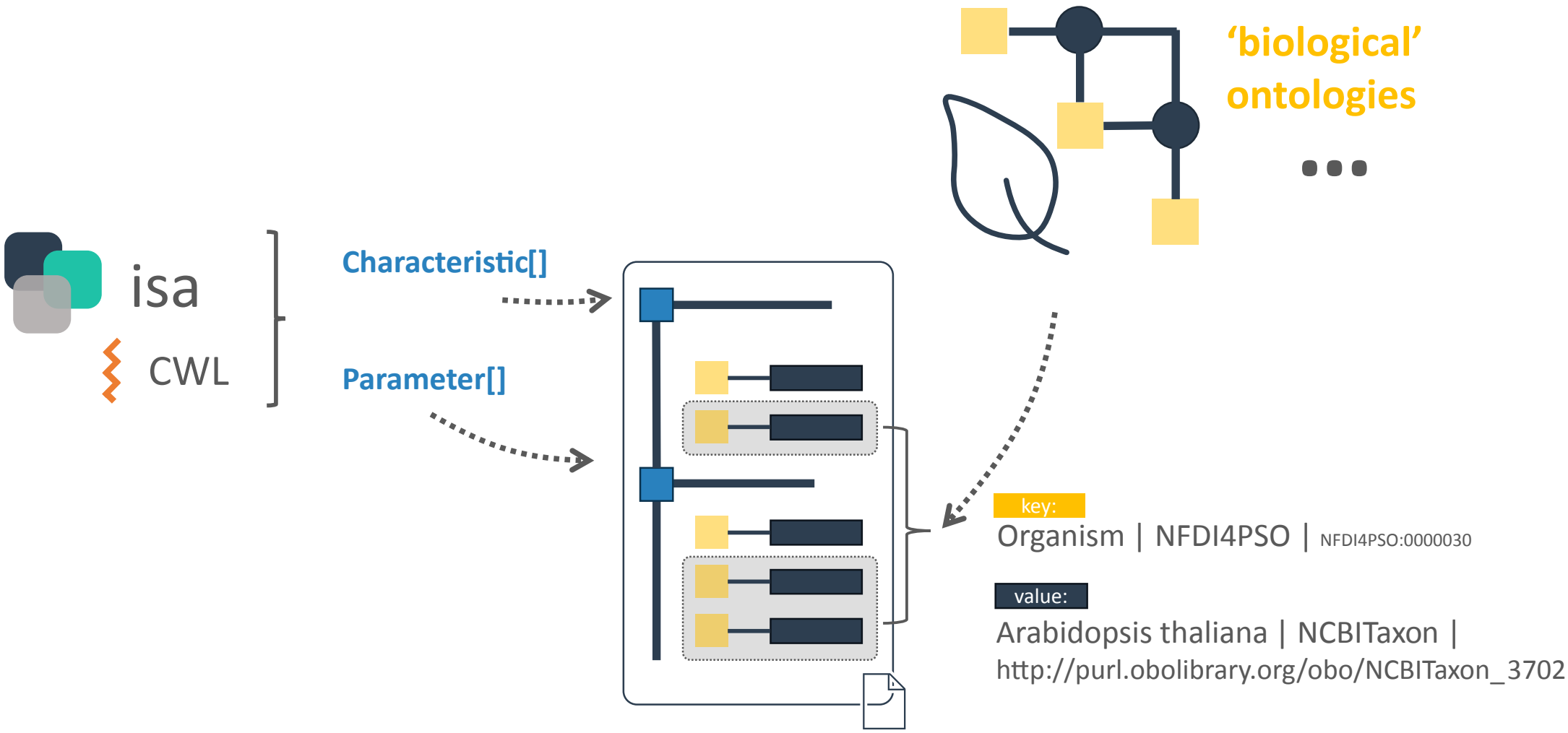
A physical quality of the thermal energy of a system.

Ontology: `PATO`

Also appears in: `NGBO` `HTN` `CAO` `ZP` `AGRO` `OMIABIS` `OBIB` `MONDO` `TXPO` `MCO` +



Finding the metadata vocabulary and descriptors



Finding the metadata vocabulary and descriptors

Parameters []

-
  - Light intensity 200 µEinstein
-
  - Temperature 6°C / 25°C
-
  - Growing 4d

Characteristics []

-
  - Arabidopsis thaliana
-
  - Leaf
-
  - Hydroponic culture
-
  - Columbia

Finding the metadata vocabulary and descriptors

Parameters []

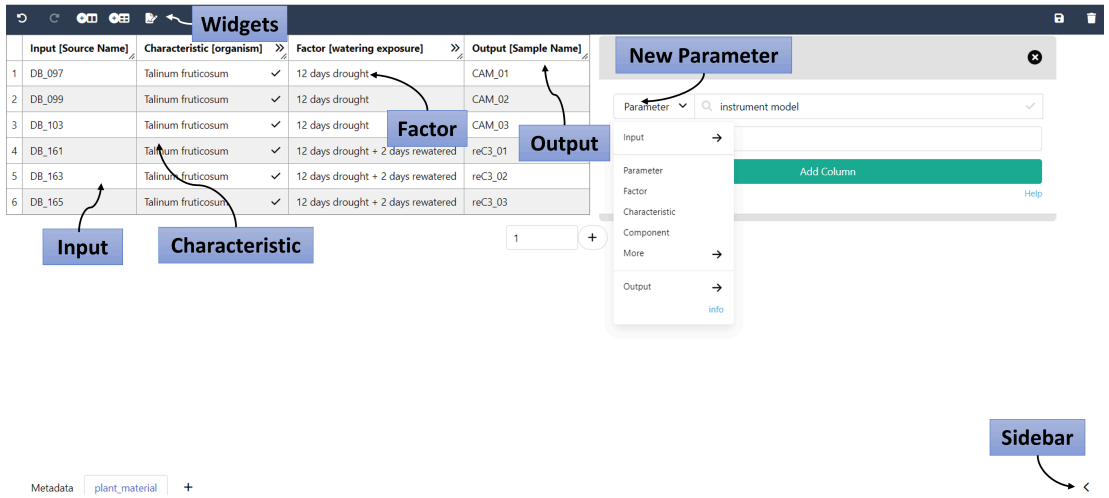
- Light intensity
  - 200 µEinstein
- Temperature
  - 6°C / 25°C
- Growth time
  - 4d

Characteristics []

- Organism
  - Arabidopsis thaliana
- Tissue
  - Leaf
- Growth medium
  - Hydroponic culture
- Ecotype
  - Columbia

Annotation Building Block types

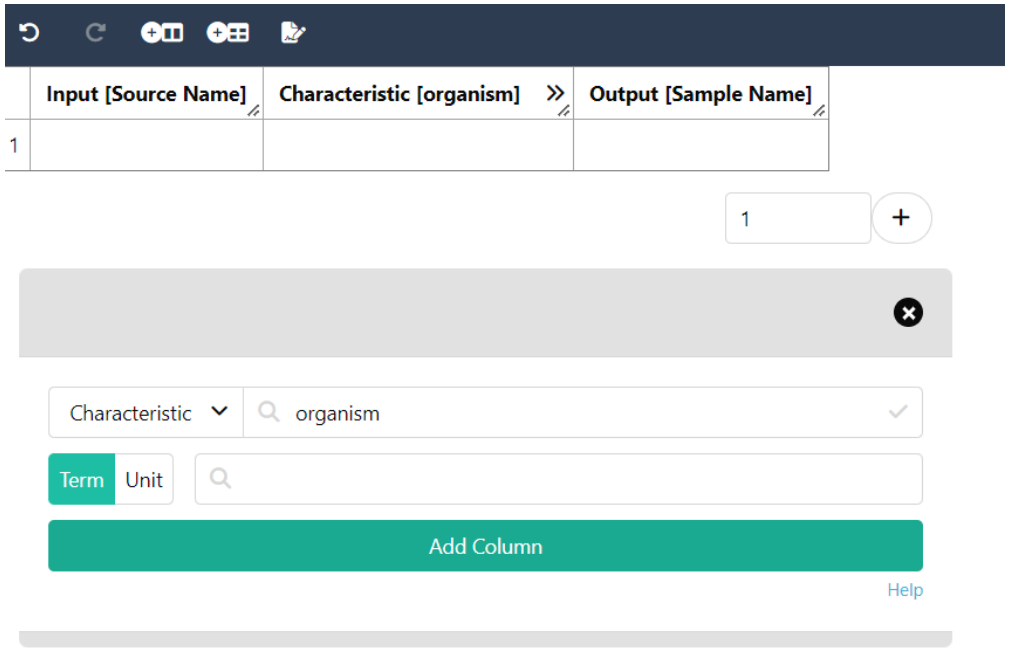
- Input (e.g. Source Name, Sample Name)
- Protocol columns
- Characteristic // Parameter // Factor // Component
- Output (e.g. Sample Name, Raw Data File, Derived Data File)



Let's take a detour on Annotation Principles

Add Characteristics

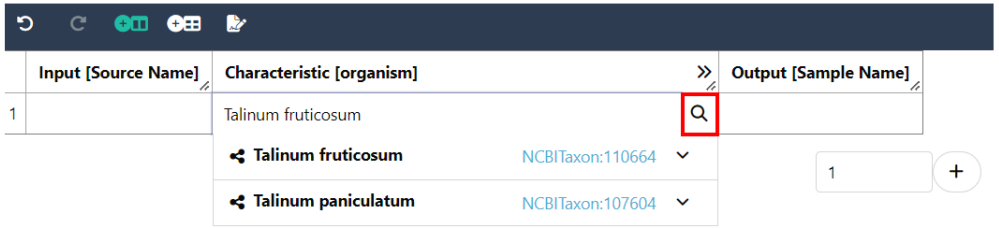
1. Select *Characteristic* from the drop-down menu
2. Enter `organism` in the search bar. This search looks for suitable *Terms* in our *Ontology* database.
3. Select the Term with the id `OBI:0100026` and,
4. Click Add Column



💡 This adds four columns to your table, one visible and **three** hidden.

Insert ontology terms

1. Insert values by selecting any cell below
   `Characteristic [organism]`
2. Use free text or use the magnifying glass to activate *Term* search
3. Write "Talinum fruticosum" and enable *Term* search
4. Select the hit



Add a Parameter building block with a unit

1. In the *Building Blocks* widget, select *Parameter*, search for `light intensity exposure` and select the term with id `PECO:0007224`.
2. Check the box for *Unit* and search for `microeinstein per square meter per second` in the adjacent search bar.
3. Select `UO:0000160`.
4. Click Add Column.

💡 This also adds four columns to your table, one visible and **three** hidden.

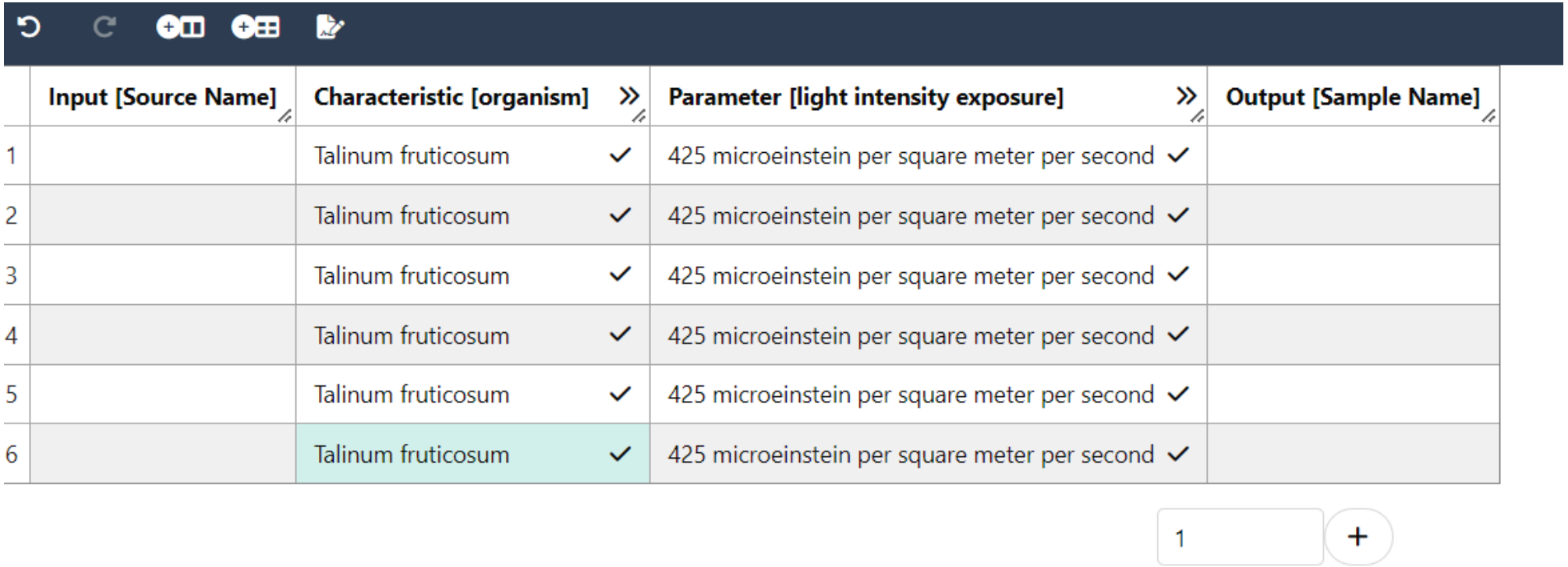
Insert unit-values

In the annotation table, select any cell below `Parameter [light intensity exposure]` and add "425" as light intensity.

💡 You can see the numbers being complemented with the chosen unit, e.g. `425 microeinstein per square meter per second`

Your annotation table is growing

At this point. Your table should look similar to this:



	Input [Source Name]	Characteristic [organism]	Parameter [light intensity exposure]	Output [Sample Name]
1		Talinum fruticosum ✓	425 microeinstein per square meter per second ✓	
2		Talinum fruticosum ✓	425 microeinstein per square meter per second ✓	
3		Talinum fruticosum ✓	425 microeinstein per square meter per second ✓	
4		Talinum fruticosum ✓	425 microeinstein per square meter per second ✓	
5		Talinum fruticosum ✓	425 microeinstein per square meter per second ✓	
6		Talinum fruticosum ✓	425 microeinstein per square meter per second ✓	

1 +

Exercise 📝

Try to add suitable *Building Blocks* for other pieces of metadata from the plant growth protocol (`studies/talinum_drought/protocols/plant_material.txt`).

Add a factor building block

1. In the *Building Blocks* widget, select *Factor*, search for `watering exposure` and select the term with id `PECO:0007383`.
2. Click `Add Column`.
3. Add the drought treatment ("no water for 12 days", "re-water for 2 days") to the respective samples

💡 There are different options to add the drought treatment.

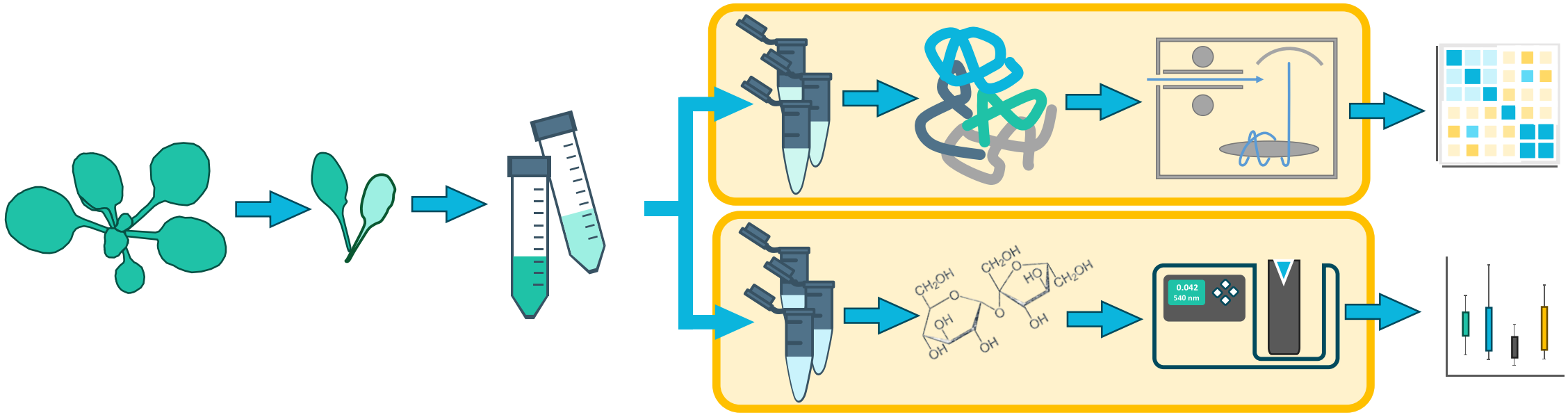
Showing ontology reference columns

Use double pointed quotation mark to un-hide hidden columns.

💡 You can see that your organism of choice was added with id and source Ontology in the reference (hidden) columns.

Characteristic [organism] «	Unit	TSR (OBI:0100026)	TAN (OBI:0100026)
Talinum fruticosum ✓	-	NCBITaxon	NCBITaxon:110664

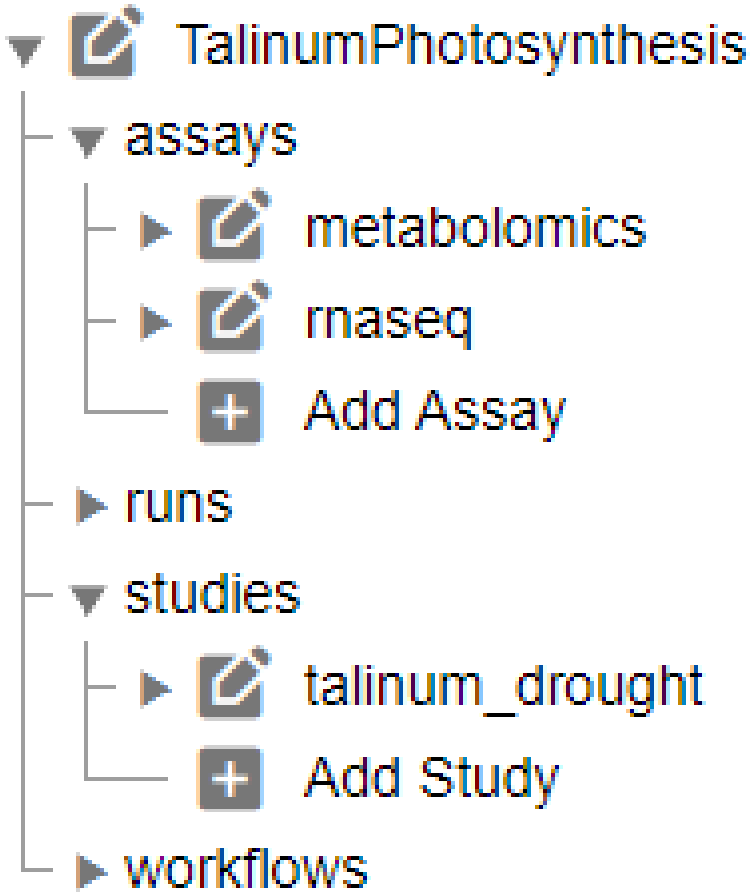
Identifying assays



Add an assay

by clicking "Add Assay" and entering an identifier for your assay

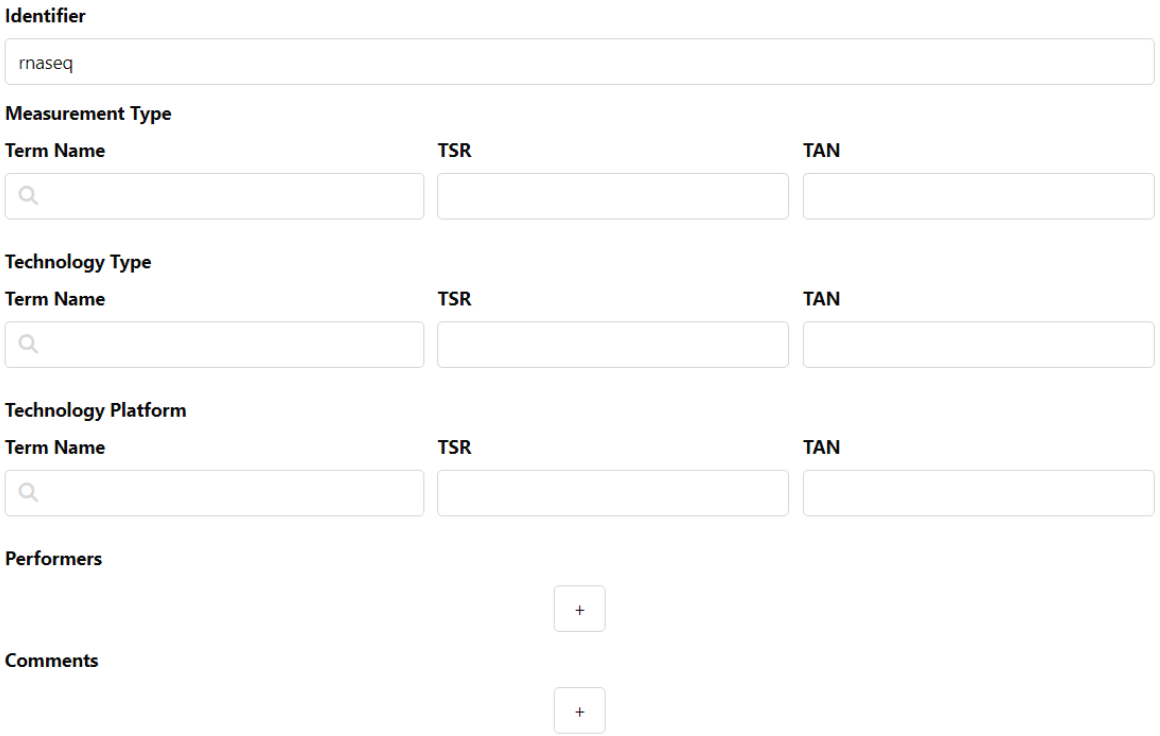
Add two assays with **rnaseq** and **metabolomics** as an identifier



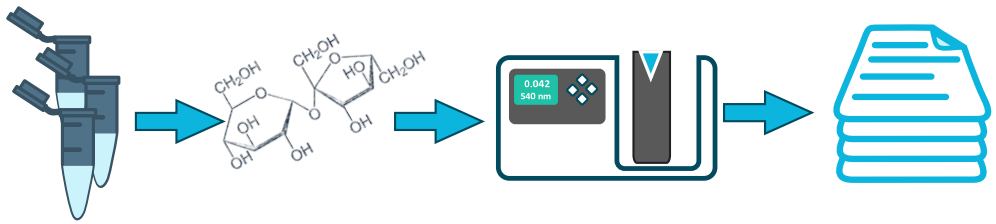
Add information about your assay

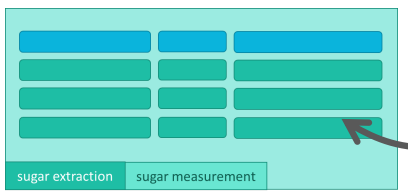
In the assay panel you can define the assay's

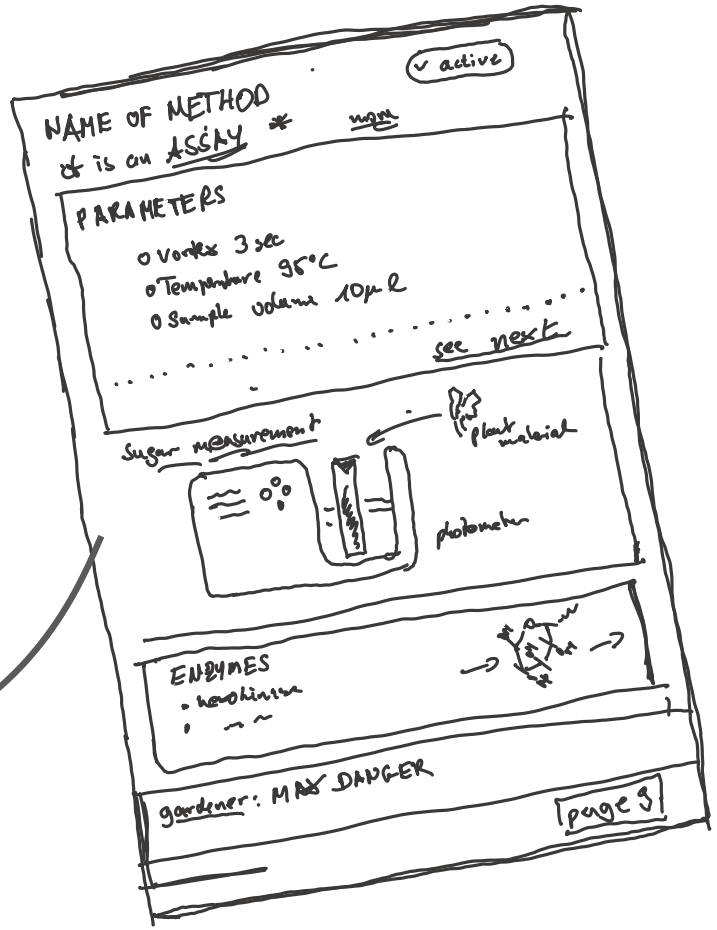
- measurement type
- technology type, and
- technology platform



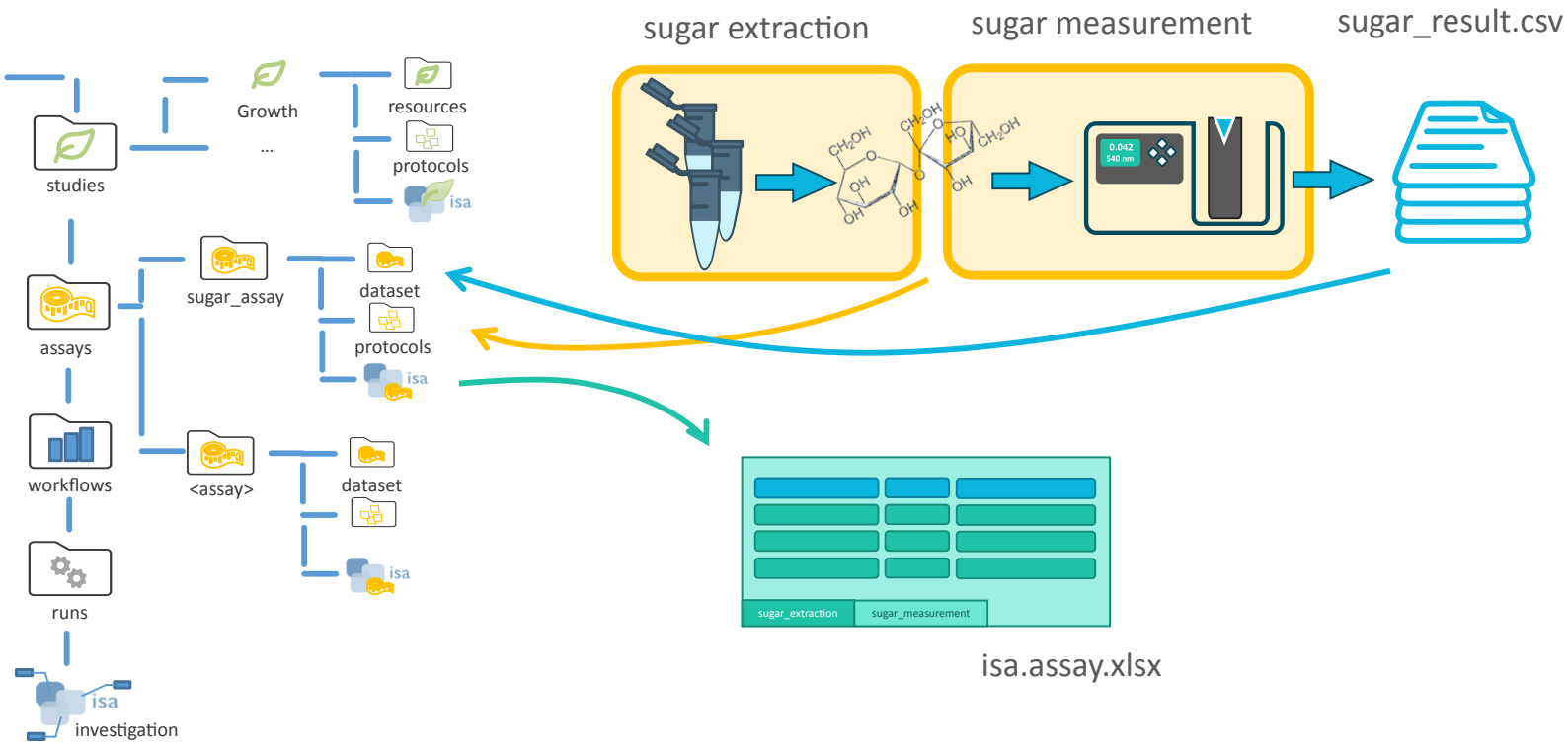
Assay for sugar measurement







Separating different assay elements



Parametric description of the lab process

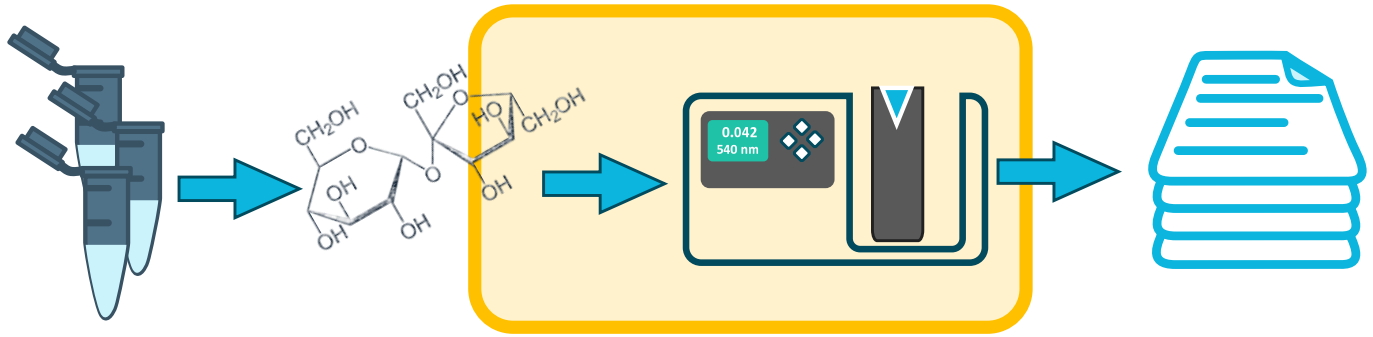
- Parameter []
  - Temperature
    - 95 °C
  - Vortex Mixer
    - 3 s

sugar extraction

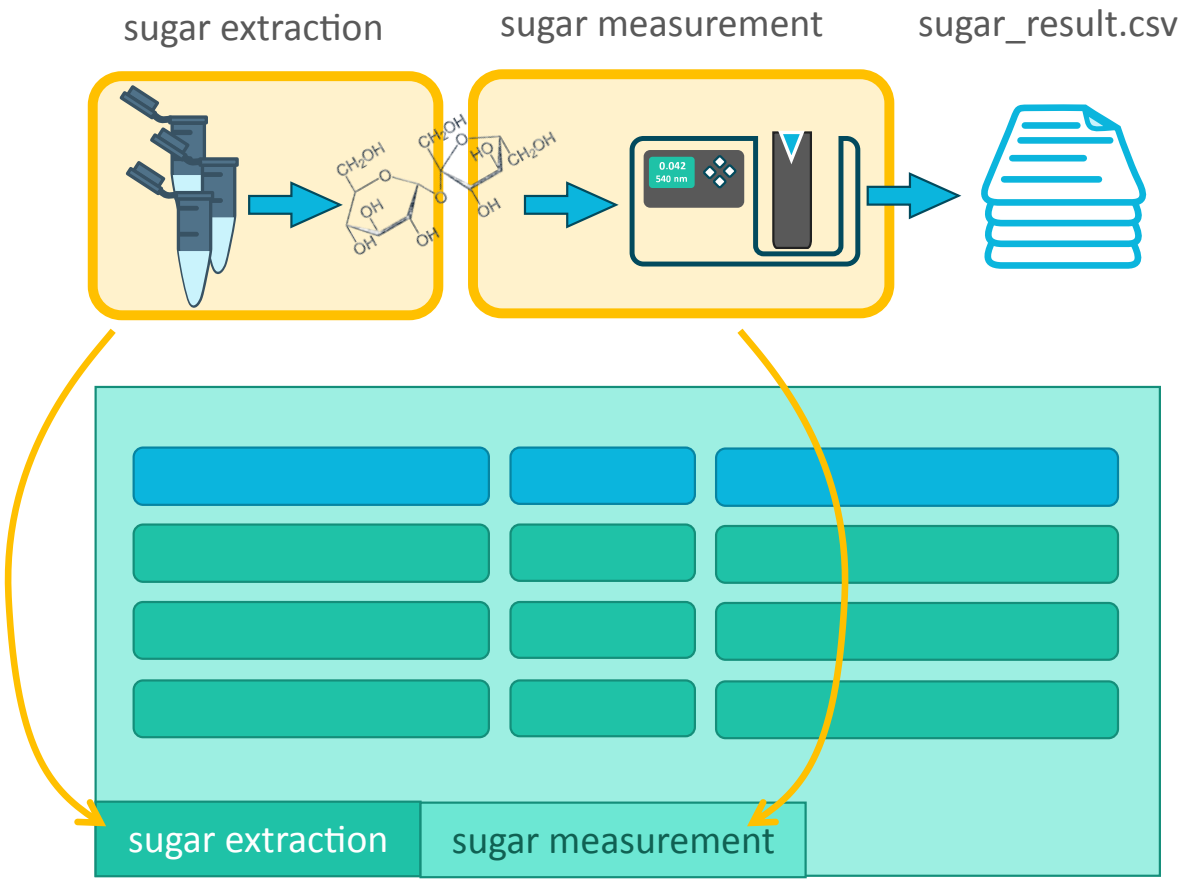
Parametric description of the lab process

- Parameter []
  - technical replicate
    - 1-3
  - sample volume
    - 10 µl
  - Buffer
    - 190 µl
  - Cycle
    - 5

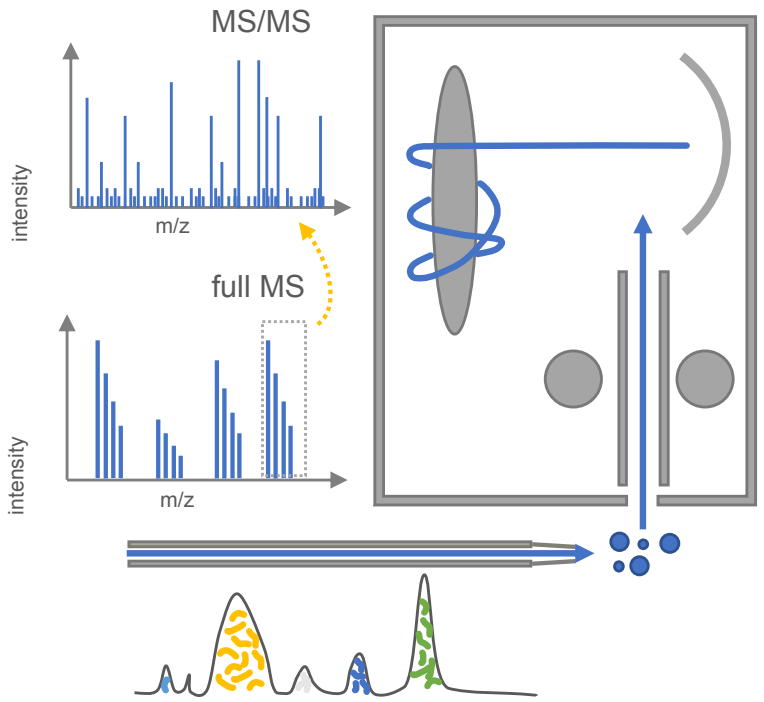
sugar measurement



Isolating the lab processes in an assay



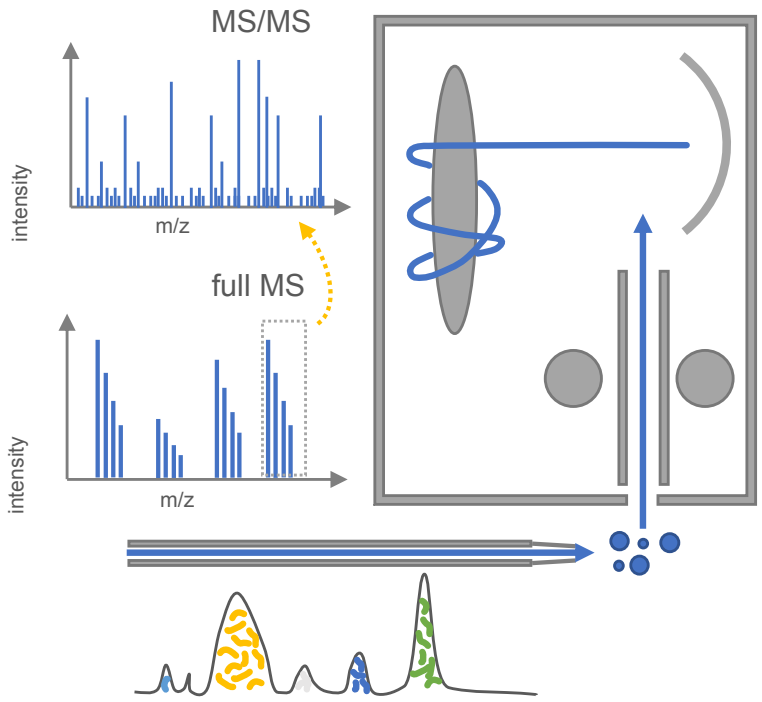
Save time using standard methods and SOPs



MS Measurement

- Parameter []
  - Protein Precipitation
    - acetone
  - cleavage agent name
    - Trypsin
  - sample preparation
    - reversed-phase solid-phase extraction
  - …
    - …

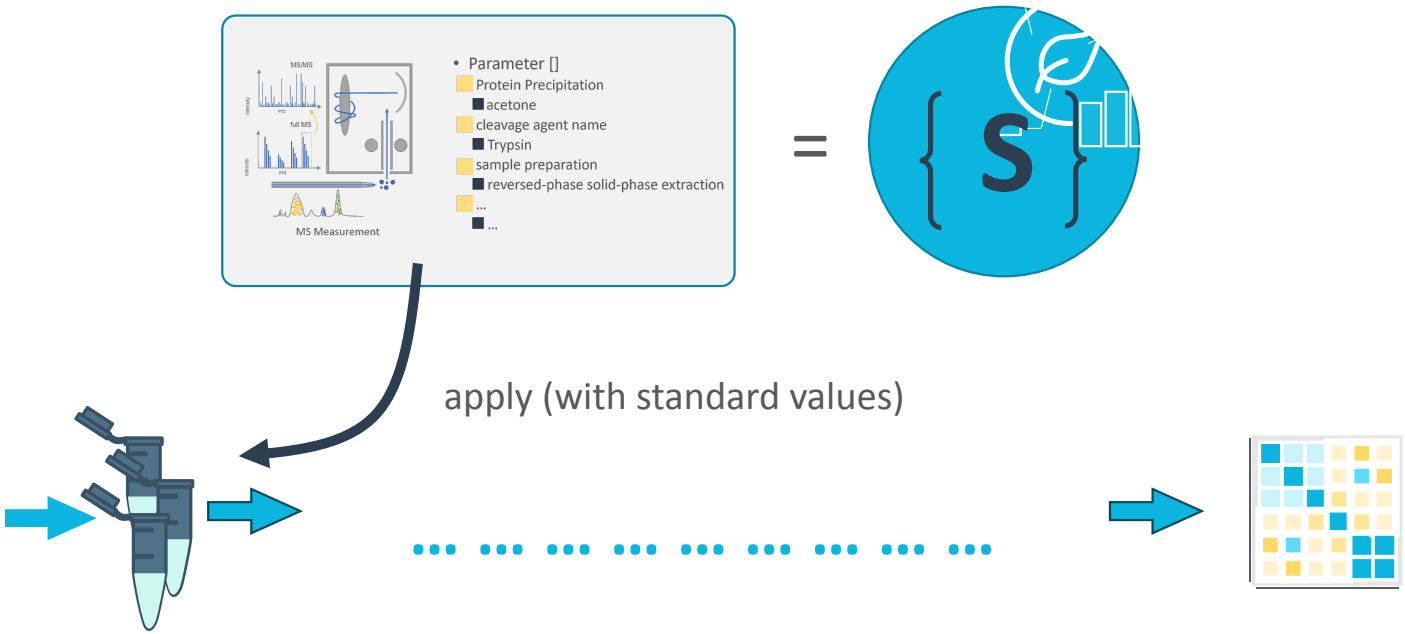
Save time using standard methods and SOPs



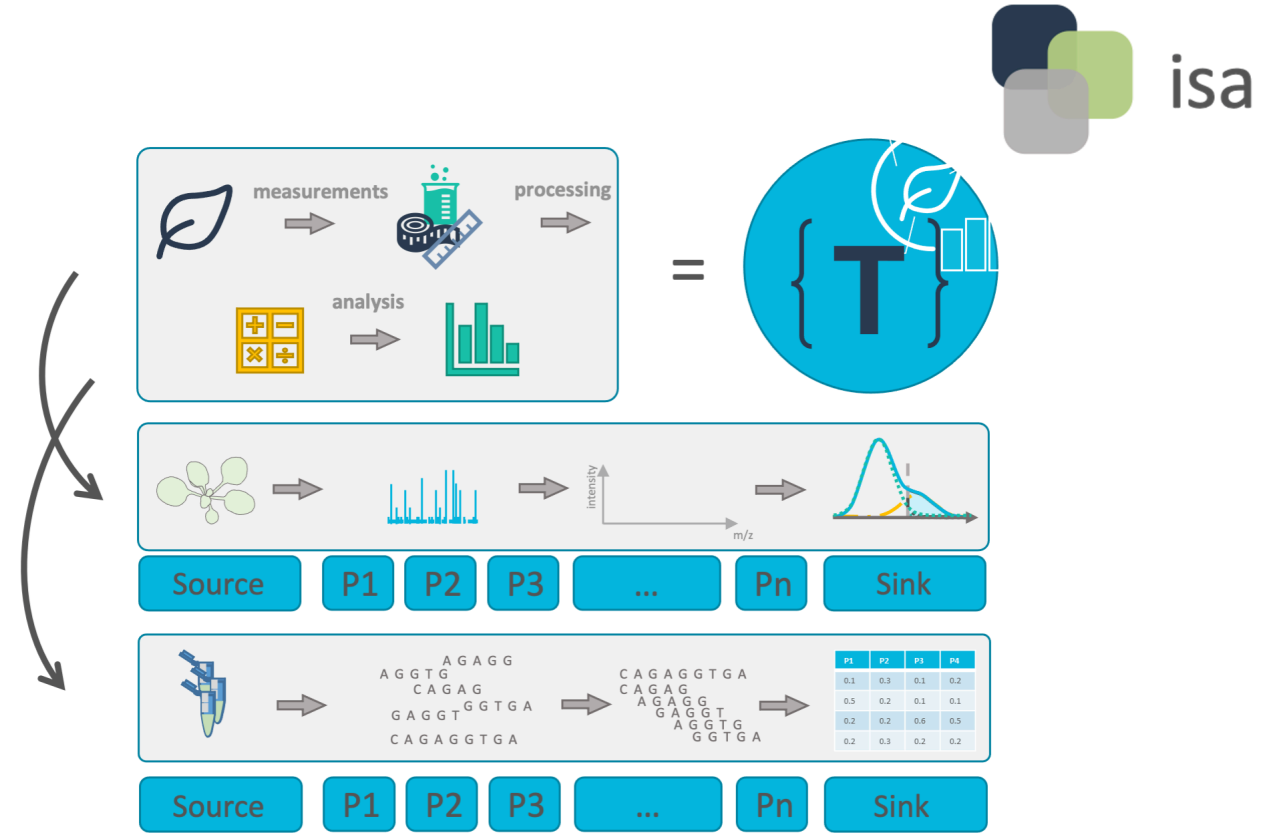
MS Measurement

- Component []
  - chromatography instrument model
    - nanoElute2
  - chromatography column model
    - PepSep C18 1.9u, 25cm x 75um
  - …
    - …
  - …
    - …

Applying standard procedures to sample record



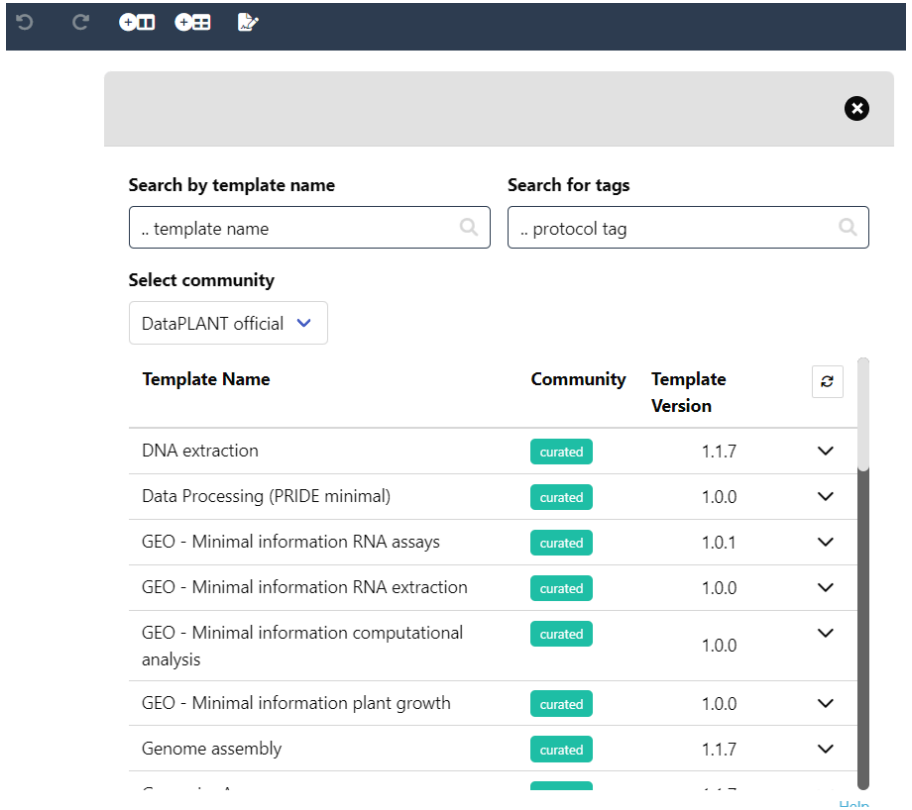
Realization of lab-specific metadata templates



Facilities can define their most common workflows as templates

Import templates from a database

- DataPLANT curated
- Community templates



Let's annotate the RNA Seq assay

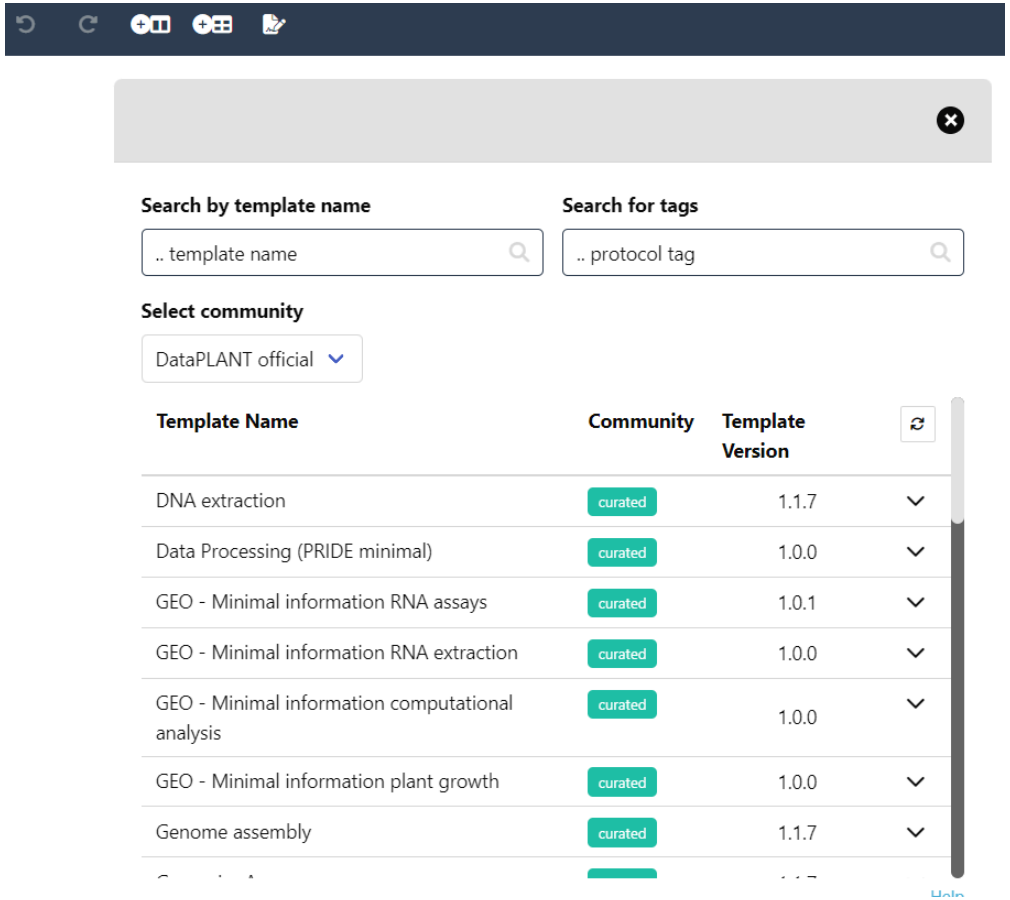
Open the lab notes `assays/rnaseq/protocols/`

Use a template

1. Open the *Templates* widget in the Bar

💡 Here you can find DataPLANT and community created workflow annotation templates

2. Search for `RNA extraction` and click `select`
   - You will see a preview of all *Building Blocks* which are part of this template.
3. Click `Add template` to add all *Building Blocks* from the template to your table

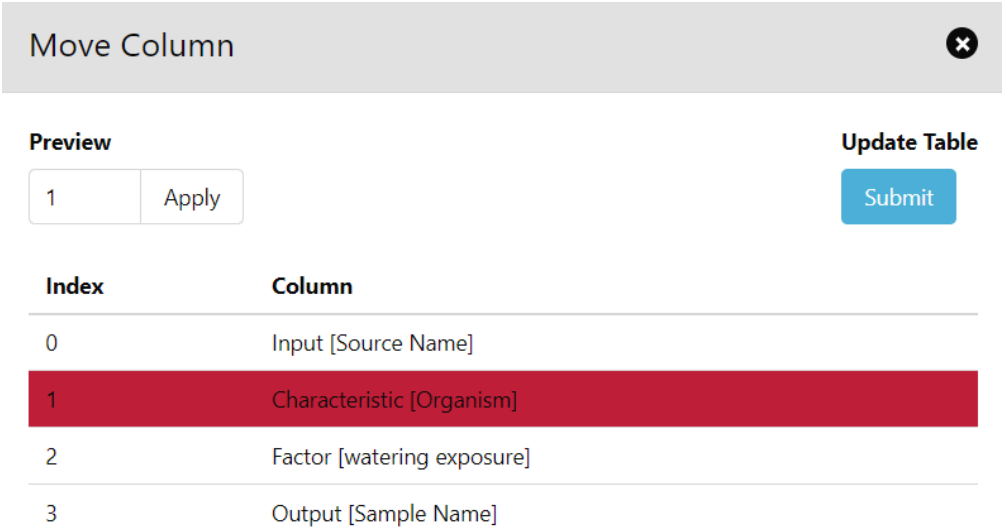


Remove Building blocks

If there are any *Building Blocks* which do not fit to your experiment you can use right click --> "Delete Column" to remove it including all related (hidden) reference columns.

Move Building blocks

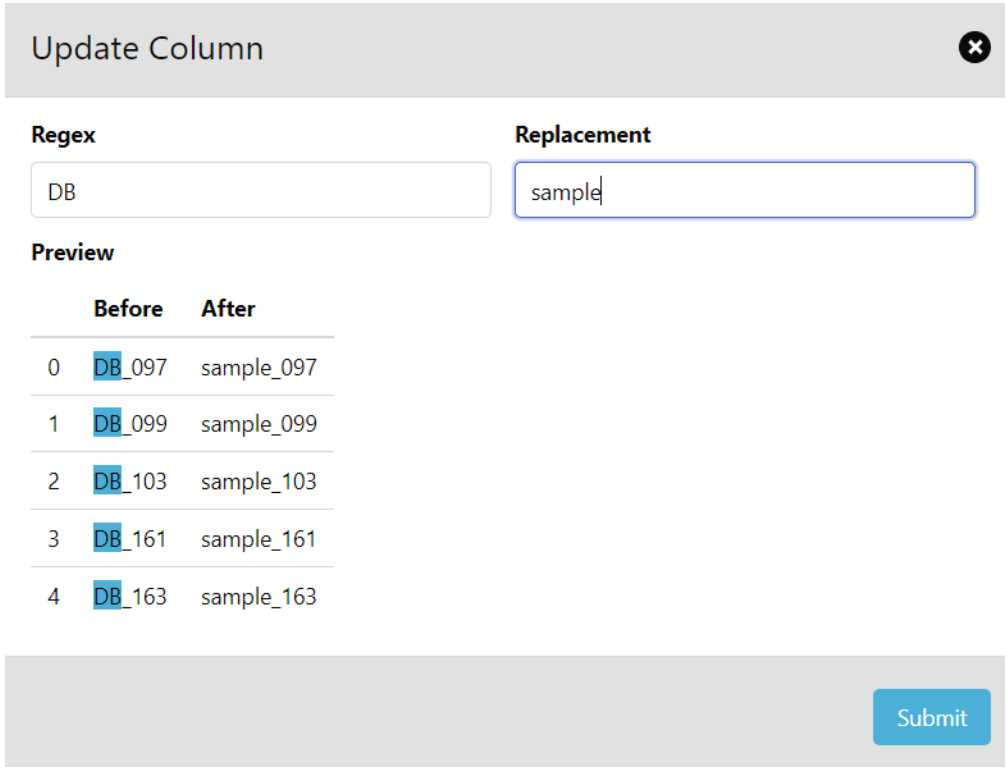
If the order of the *Building Blocks* should be adjusted you can use right click --> "Move Column"



Replace multiple names

Right click --> "Update Column" can be used to replace names in batches

💡 this only works on Input columns



New process, new worksheet

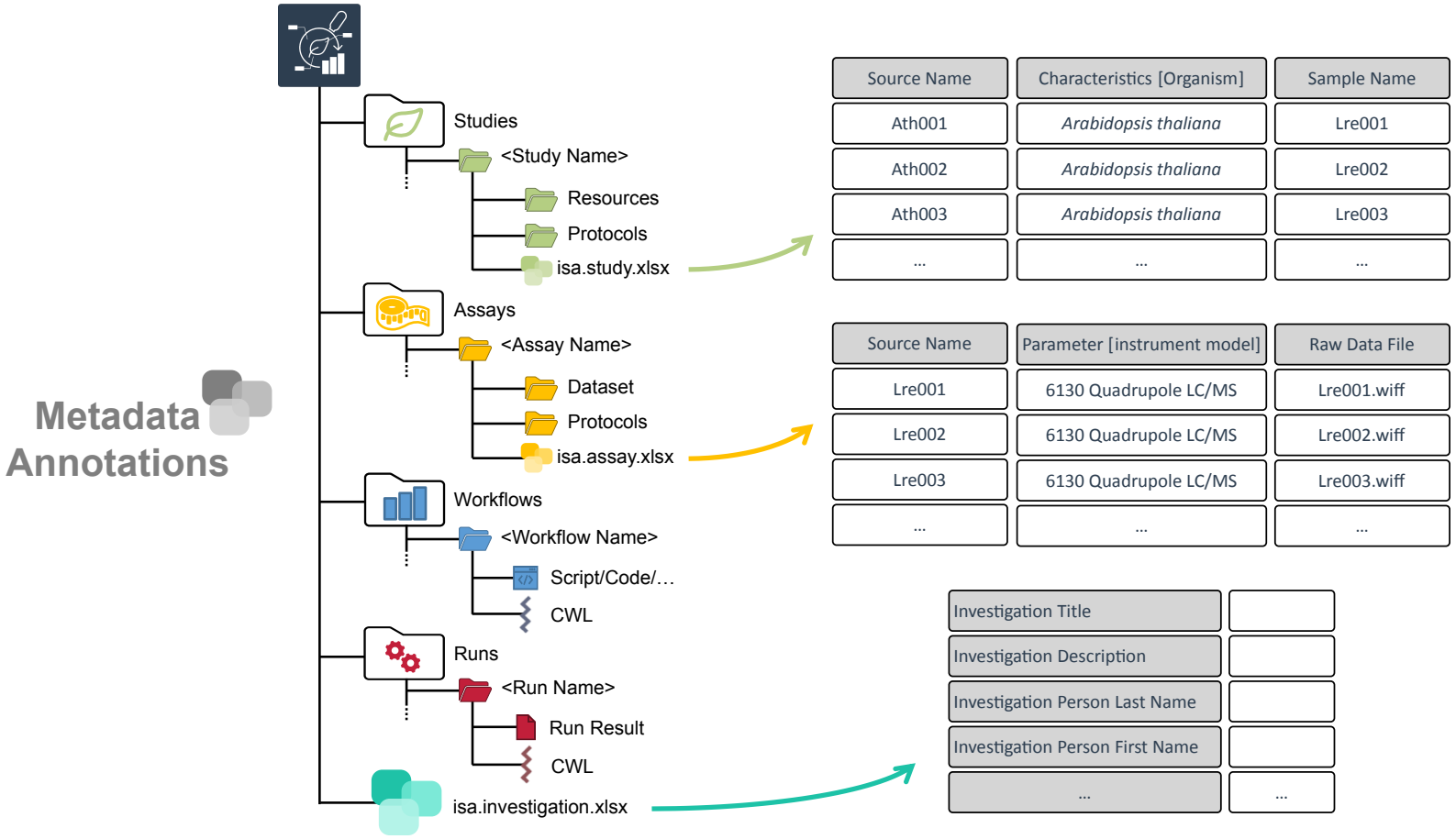
1. Add a new sheet to the `assays/rnaseq/isa.assay.xlsx` workbook.
2. Add the template "RNASeq Assay"

Exercise 📝

Try to fill the two sheets with the protocol details:

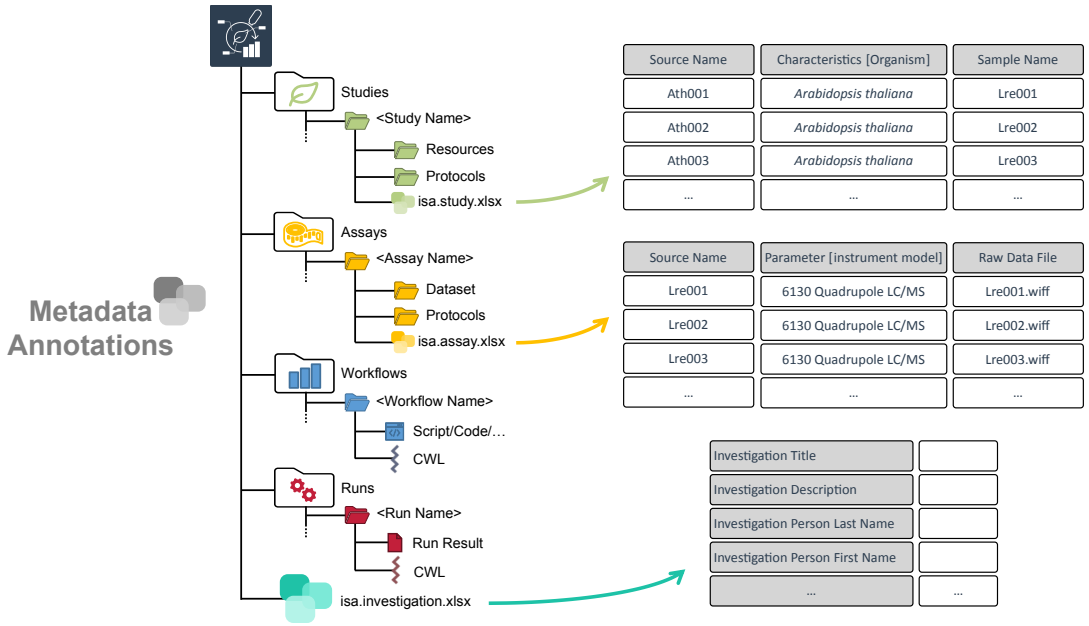
- `assays/rnaseq/protocols/RNA_extraction.txt` and
- `assays/rnaseq/protocols/Illumina_libraries.txt`

ARC builds on ISA to connect data



ARC builds on ISA to link data

- Samples are linked study-to-assay, assay-to-assay
- Raw data is linked to assays
- Protocols can be referenced
- ...



Link samples across studies and assays

1. Use the `Output [Sample Name]` of studies/talinum_drought/isa.study.xlsx as the `Input [Sample Name]` to **rna-extraction**.
2. Use the `Output [Sample Name]` of **rna-extraction** as the `Input [Sample Name]` to **illumina-libraries**.

Seeds —*Plant growth*→ **Leaves** —*RNA Extraction*→ **RNA** —*Illumina*→ **fastq files**

Link dataset files to samples

1. In the *Building Blocks* widget select *Output -> Raw Data File*.
2. Click `Add Column`.

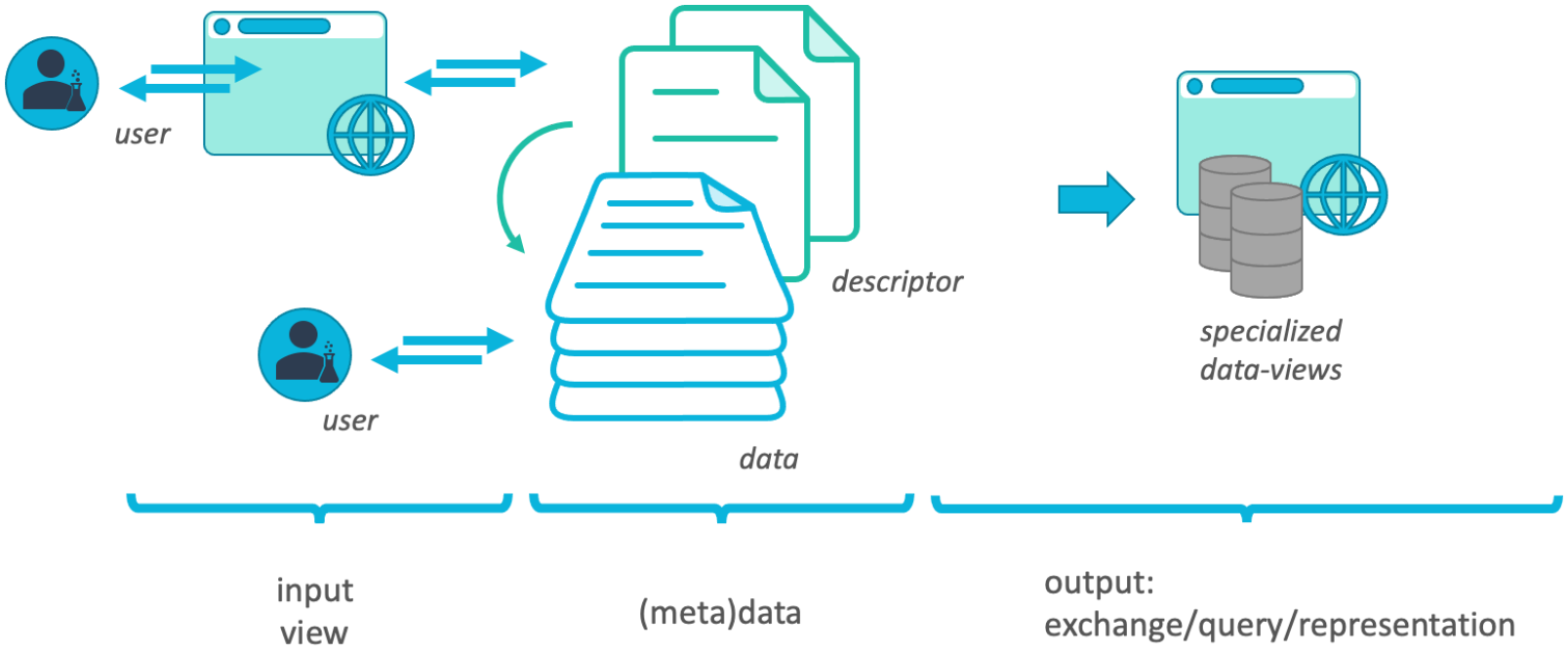
💡 You see a warning about a changed output column.

3. Click `Continue`.
4. Go to the *File Picker* tab and click `Pick file names`.
5. Select and open the *fastq.gz files from the dataset folder.
6. Copy / paste them to the `Raw Data File`.

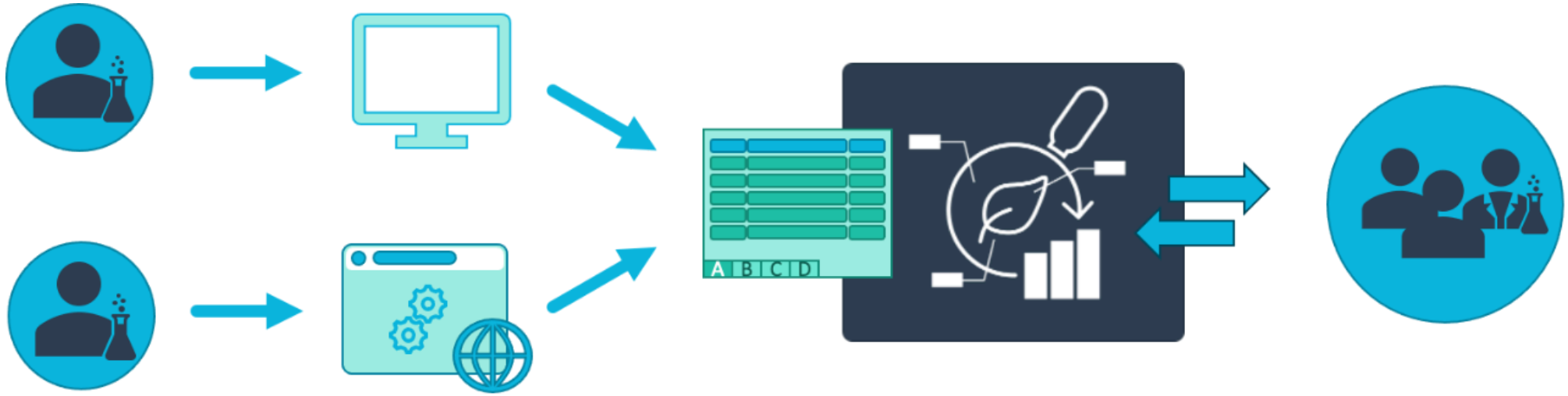
💡 This allows you to link your samples to the resulting raw data files.

Everything is a file

The ARC is a **data-centric** approach to RDM



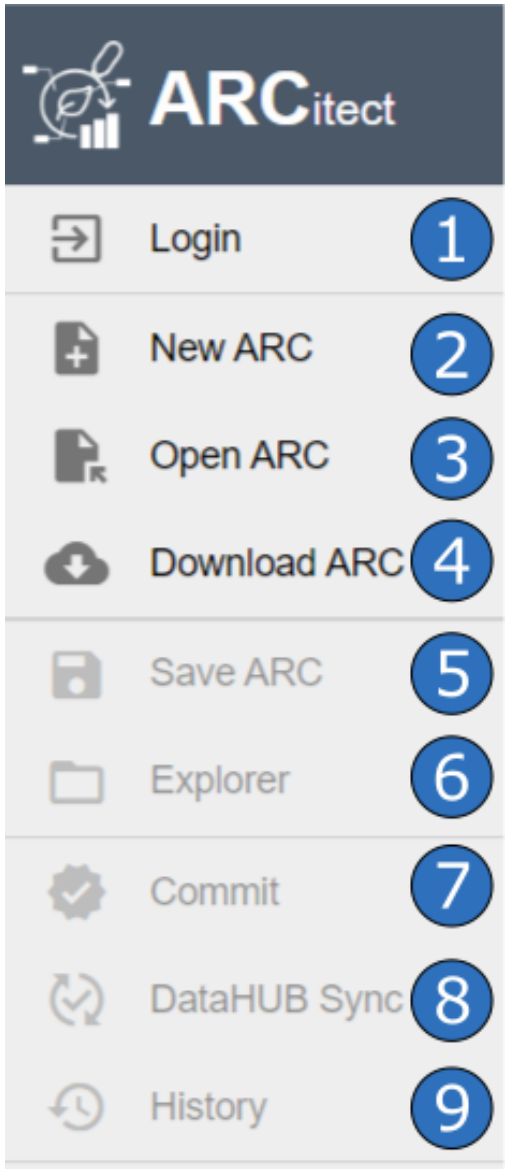
No technical lock-in



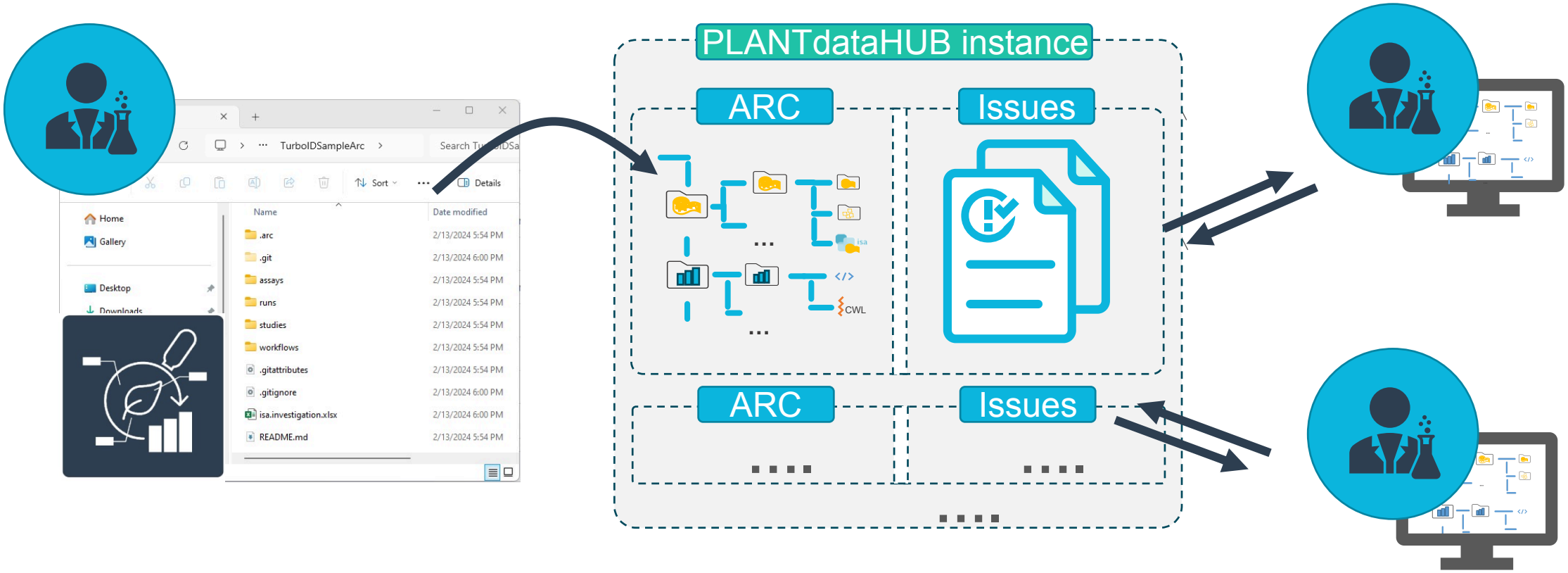
(Meta)data transparency with tool assistance but **no technical lock-in**

Explorer

The **Explorer (6)** button directly opens your ARC locally

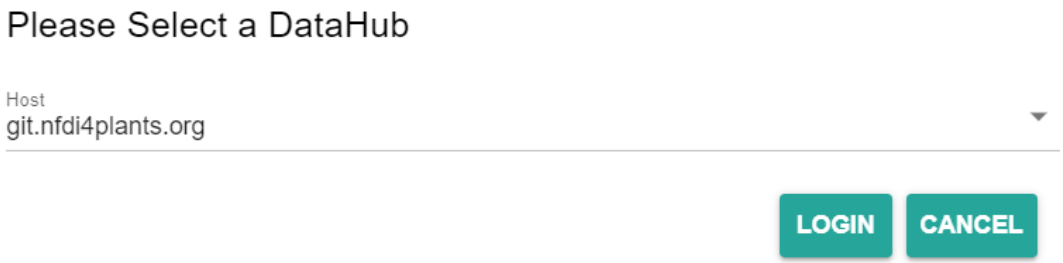


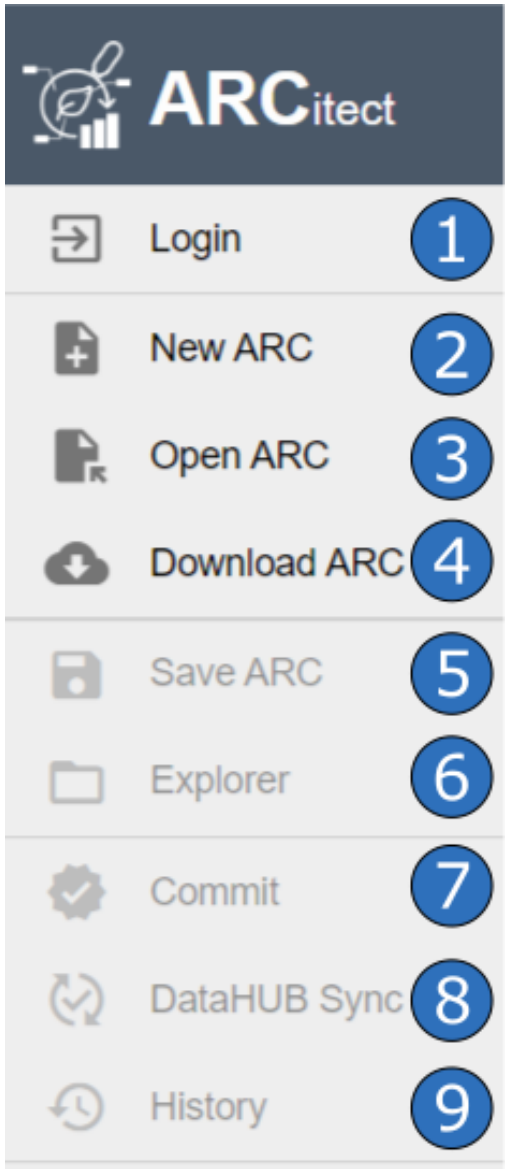
Using the DataHUB to collaborate



Login to DataHUB

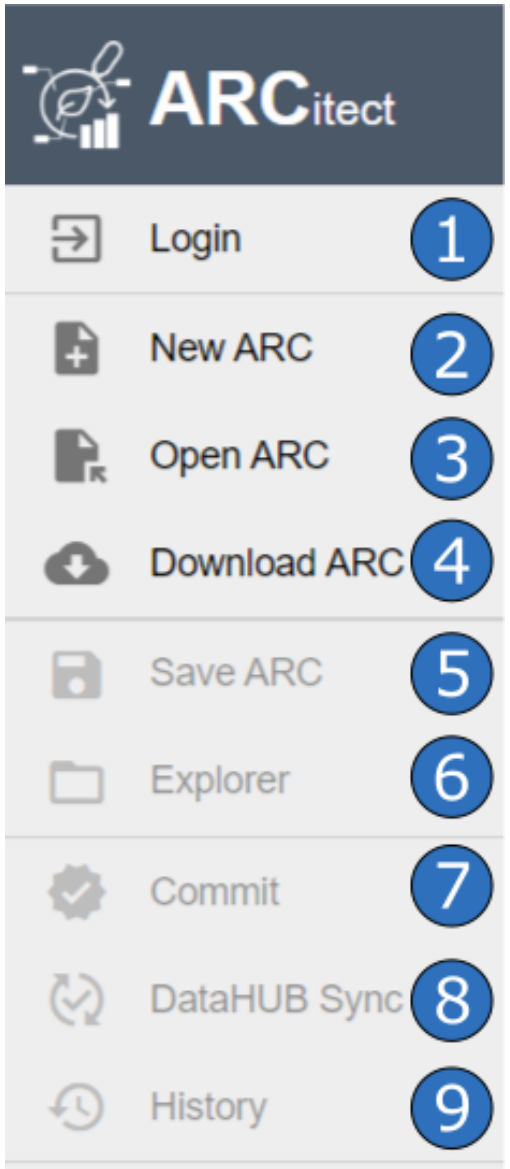
1. Login to DataHUB (1)
2. Select `git.nfdi4plants.org` as Host





Commit panel (7)

You have to commit changes before you can upload to the DataHUB



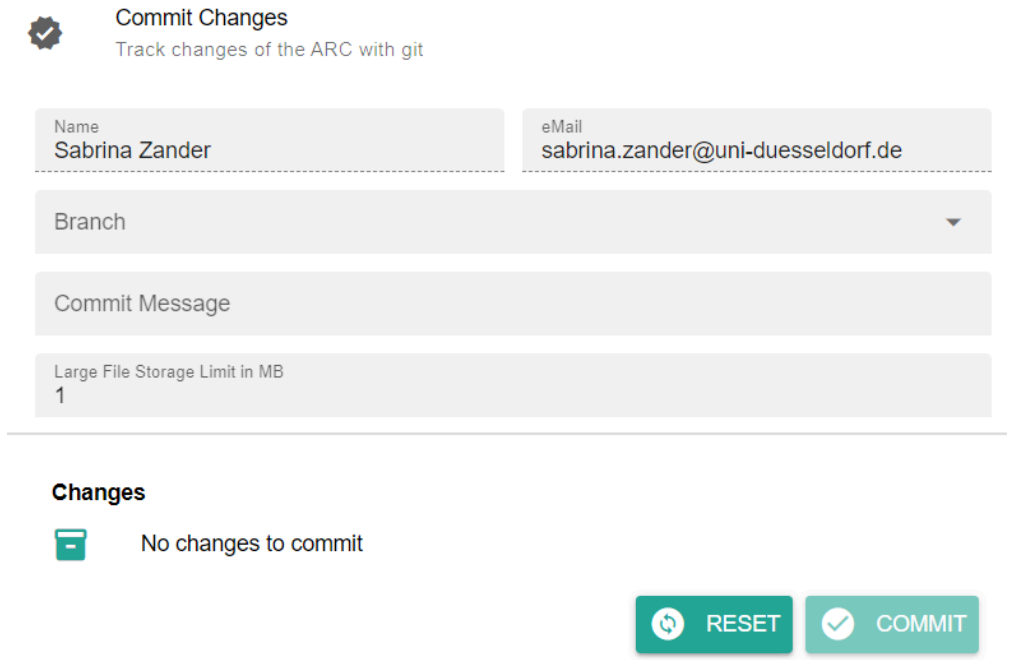
Commit your changes

If you are logged in, the **Commit panel** shows

- your DataHUB's *Full Name* and *eMail*

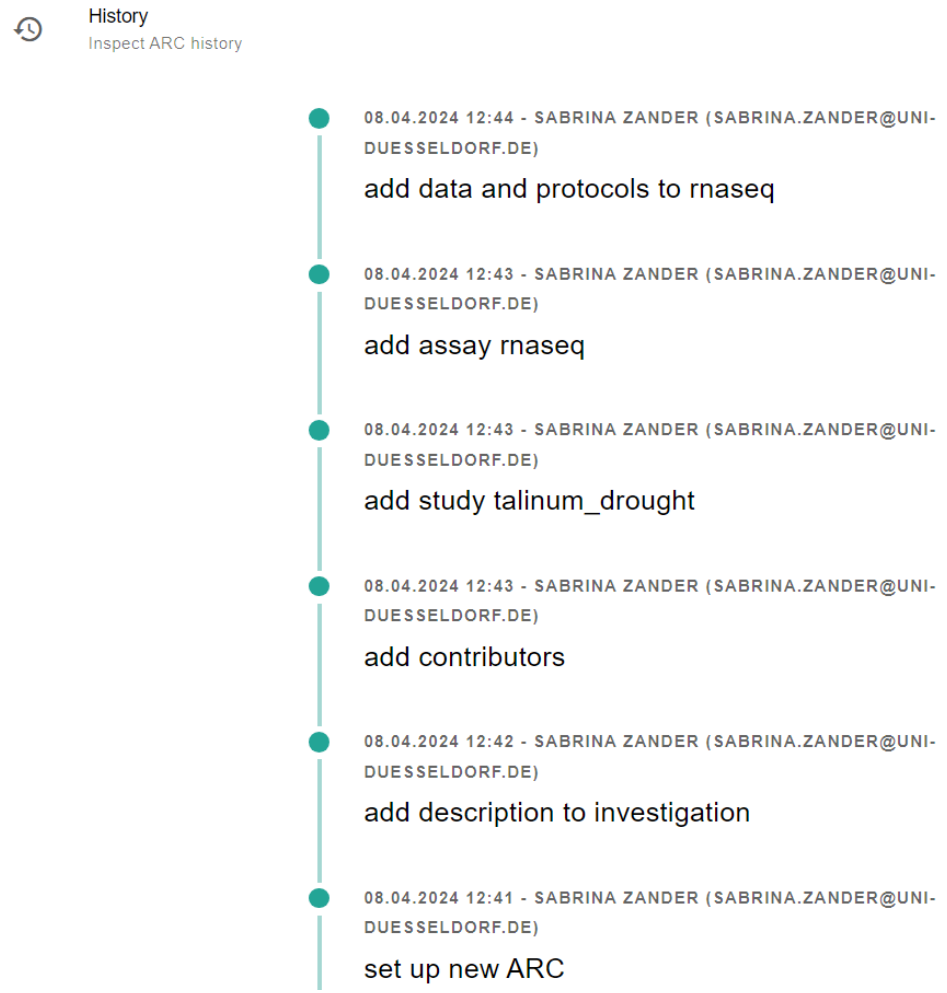
It allows you to

- track changes of the ARC with git
- add a commit message
- use different branches



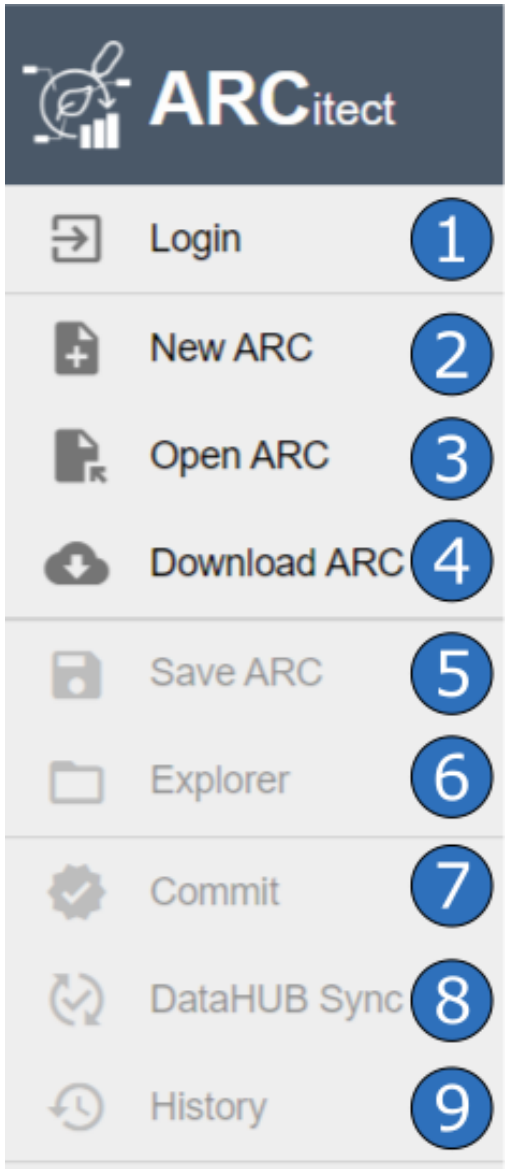
History panel (9)

In the History panel you can inspect your ARCs history with all commits



Upload your local ARC to the DataHUB

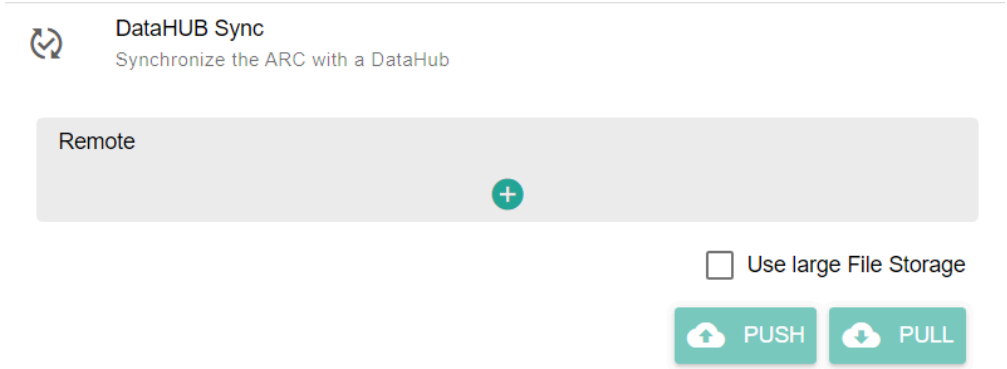
From the sidebar, navigate to **DataHUB Sync** (8)



DataHUB Sync

The DataHUB Sync panel allows you to

- sync the changes to the DataHUB: **Push**
- sync from the DataHUB: **Pull**, and
- change the Remote for the synchronization

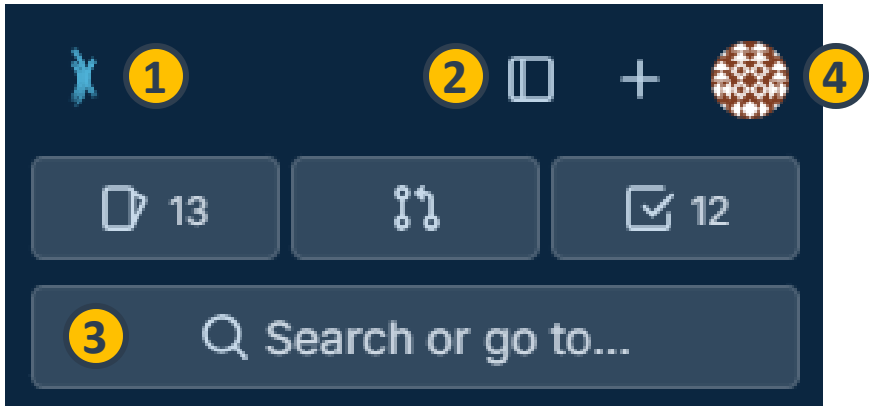


Check if your ARC is successfully uploaded

1. sign in to the DataHUB
2. Check your projects

DataHub Hands-On

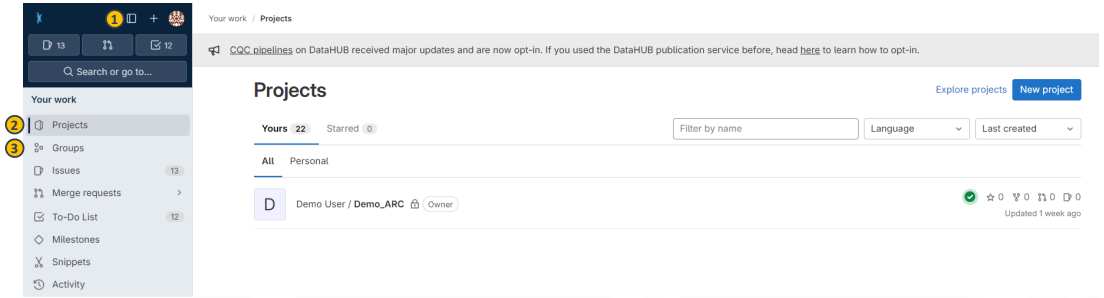
Navigation Bar



1. navigate directly to the projects panel via the icon in the top-left (1)
2. open the hamburger Menu (2)
3. use the search field (3) to find ARCs, users and groups
4. open the avatar Menu (4)

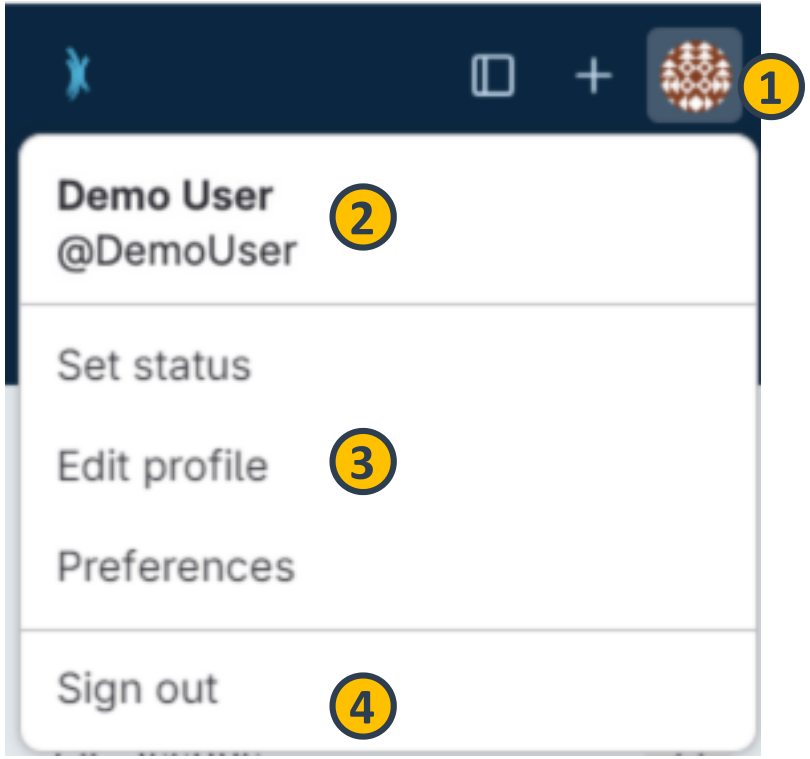
Hamburger Menu

1. From the hamburger menu (1) you can
2. navigate to the projects (2)
3. or groups (3) panels



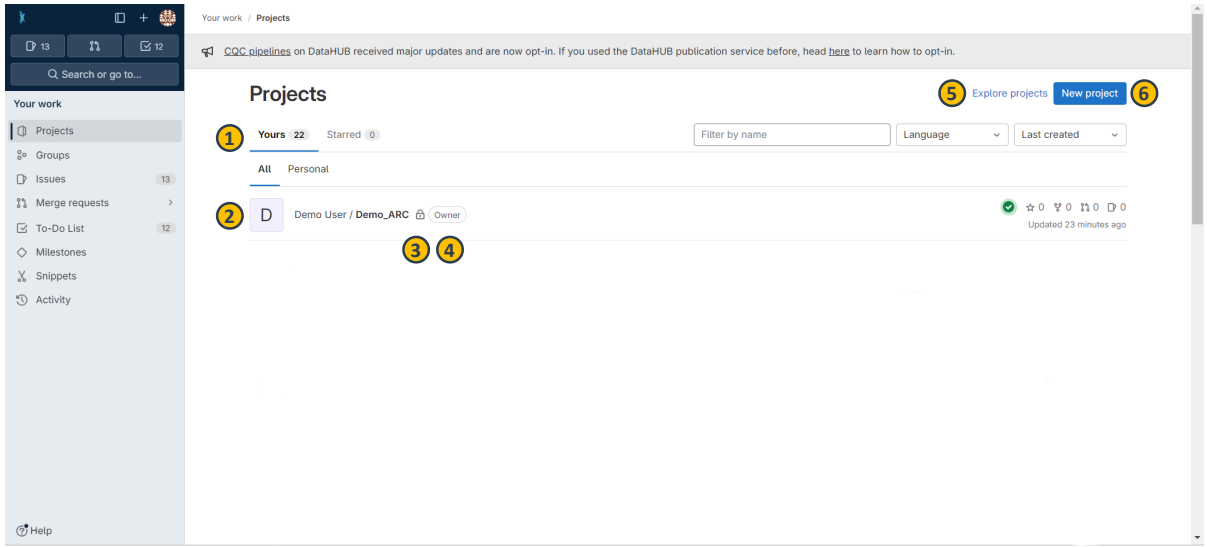
Avatar Menu

1. In the avatar menu (1) you can
2. find your profile name and user name (2),
3. navigate to the user settings (3)
4. or sign out (4) of the DataHUB.



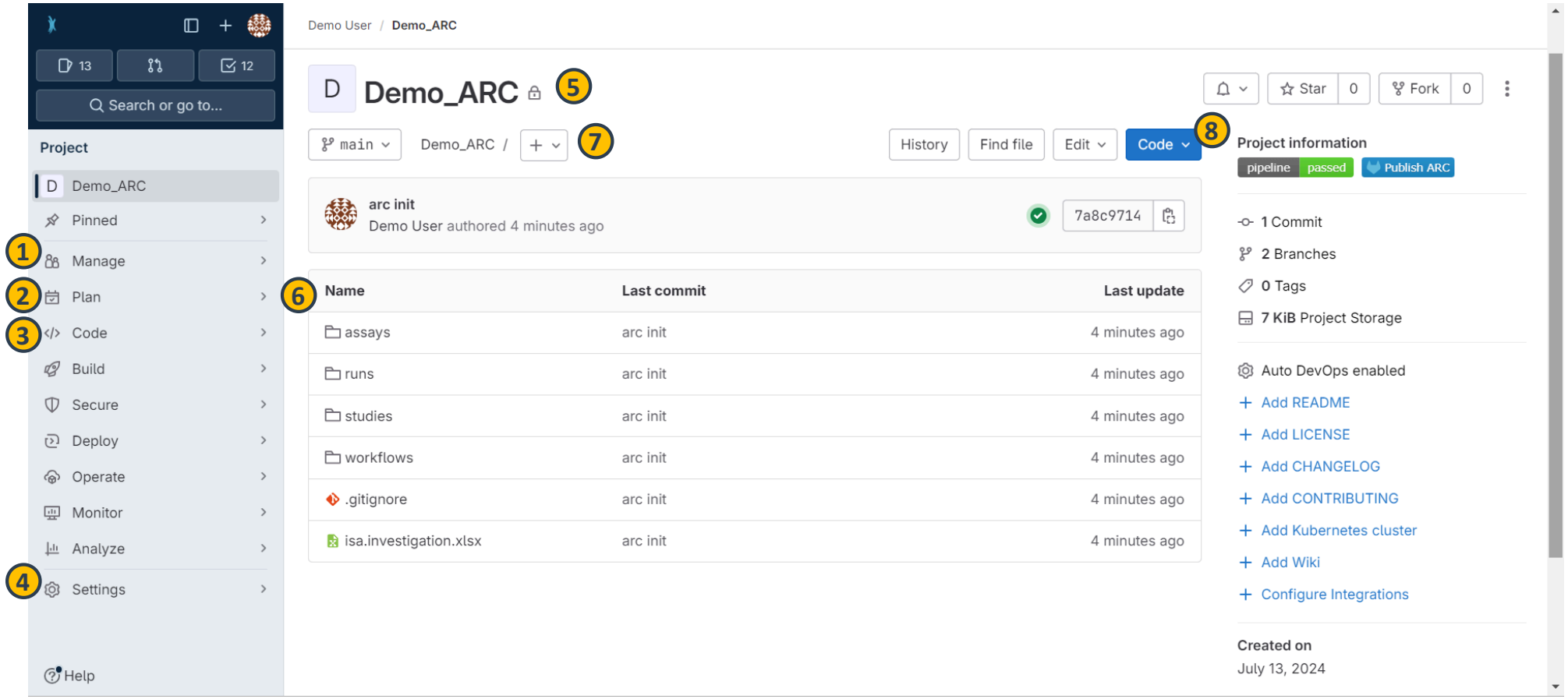
Projects Panel

1. Choose a tab (1) to see only your ARCs, or explore other publicly available ARCs.
2. The main panel (2) lists all ARCs
3. Here you can also see, the visibility level (3), and
4. your permission or role (4) for the listed ARC.



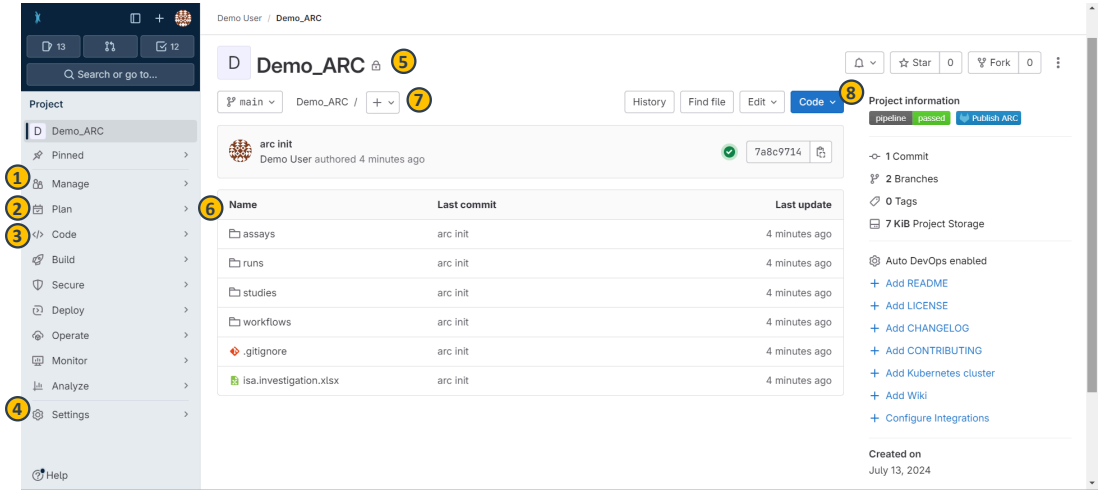
ARC Panel

The ARC Panel is the main working area for your ARC.



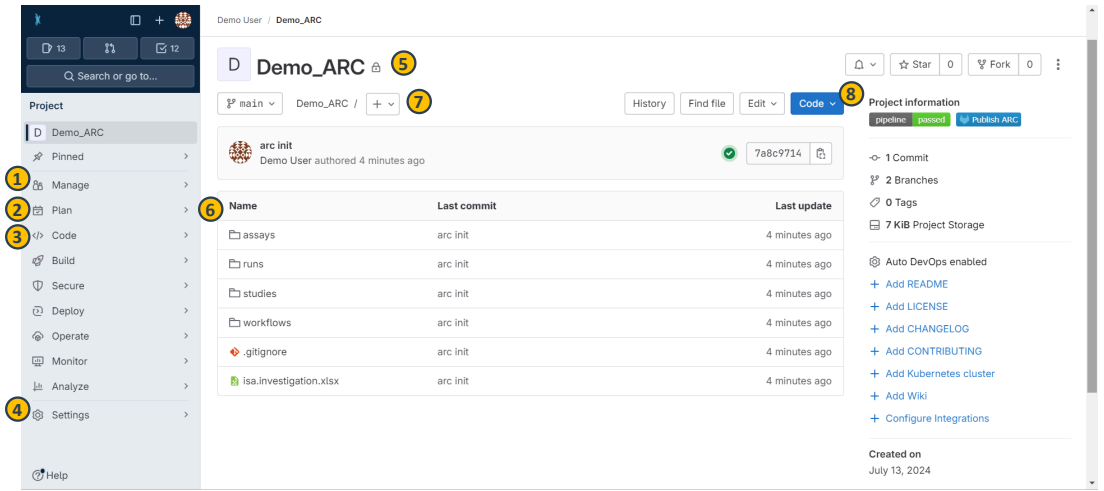
ARC Panel – sidebar

1. access the project information (1), e.g. invite members to the ARC
2. follow the progress of your ARC repository (2),
3. organize tasks in issue lists and boards (3),
4. take notes in a wiki to your ARC (4),
5. adapt the settings (5) of the ARC.

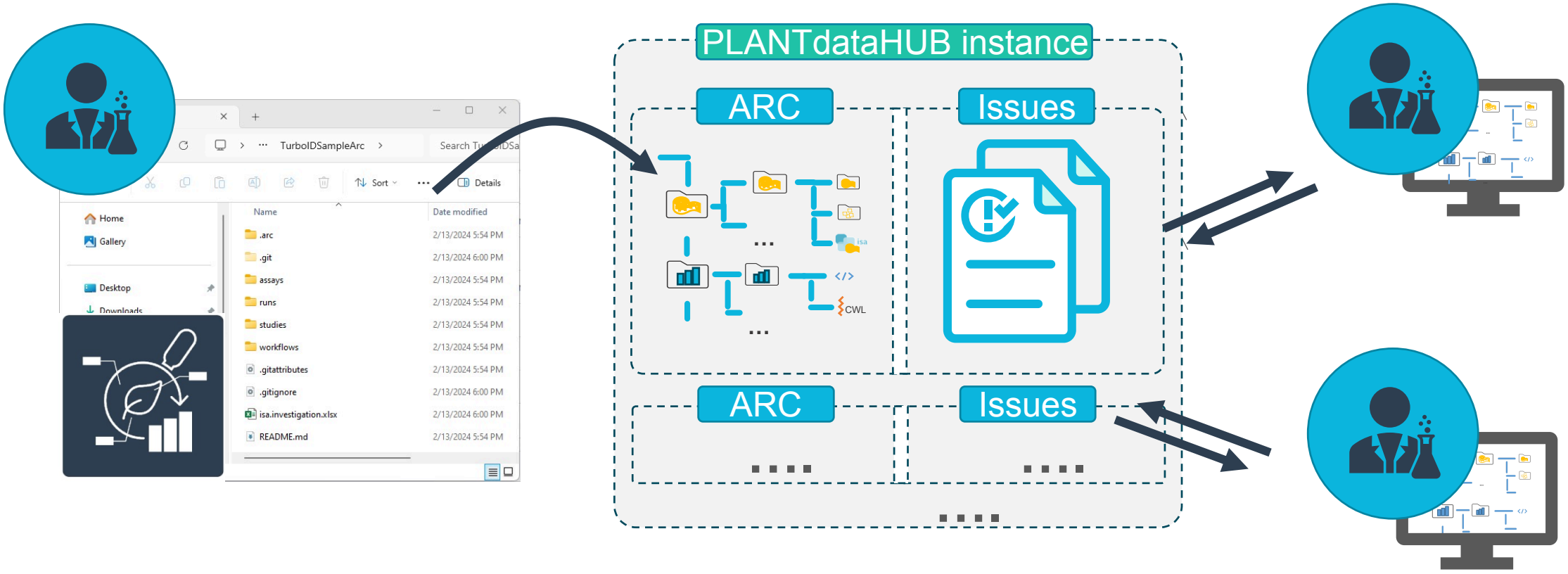


ARC Panel – main panel

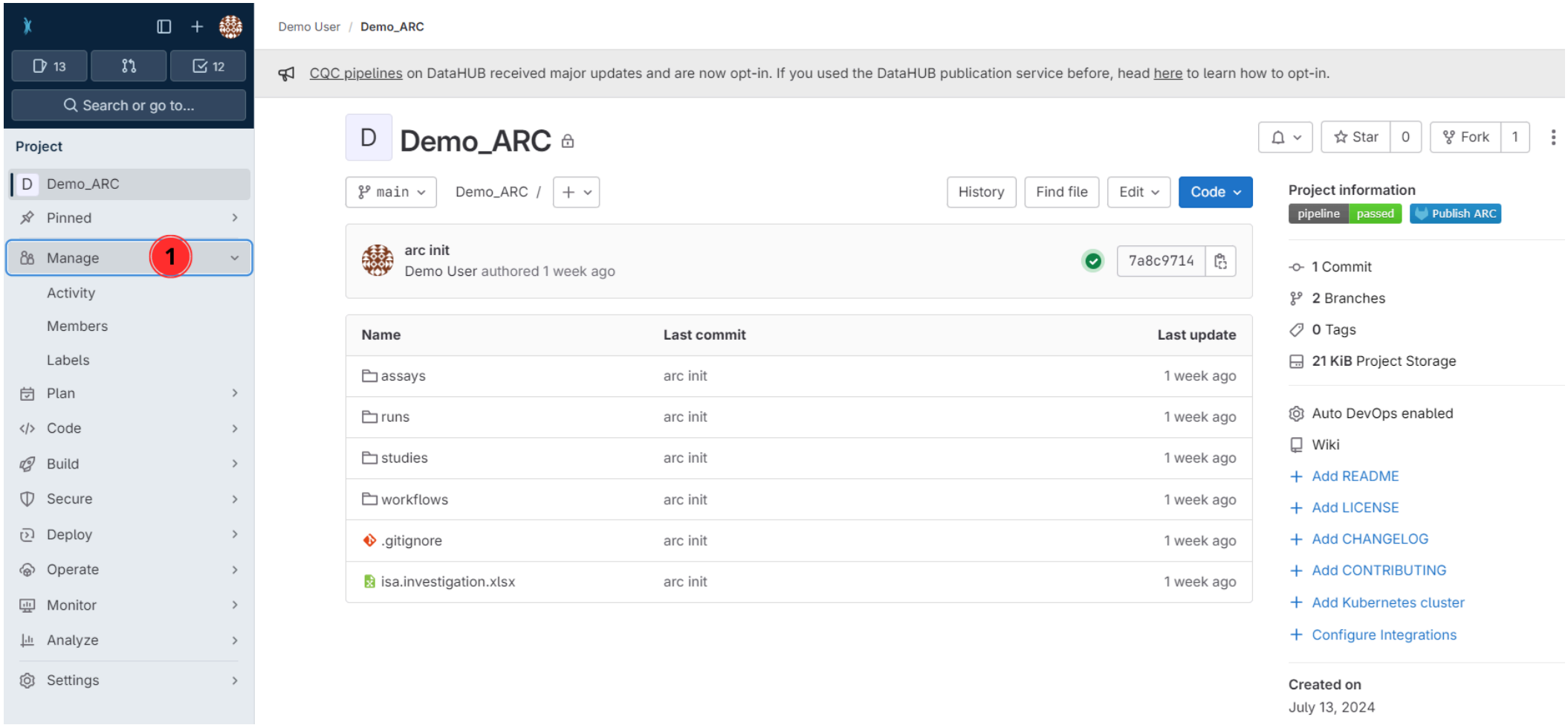
6. see the ARC's name and visibility level (6),
7. follow the ARC's commit history (7),
8. see files contained in your ARC just like on your computer (8),
9. add new files or directories (9), and
10. download or clone your ARC (10).



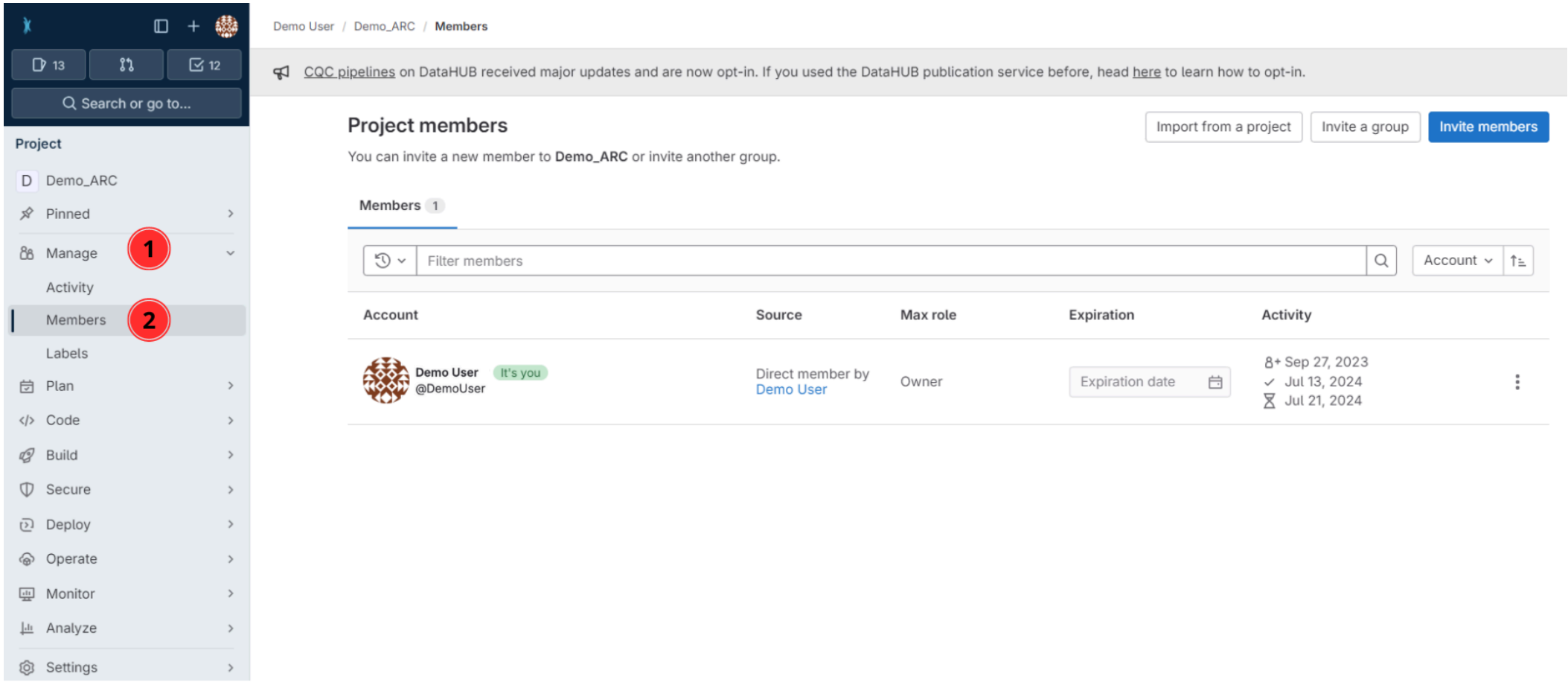
Collaborate and share



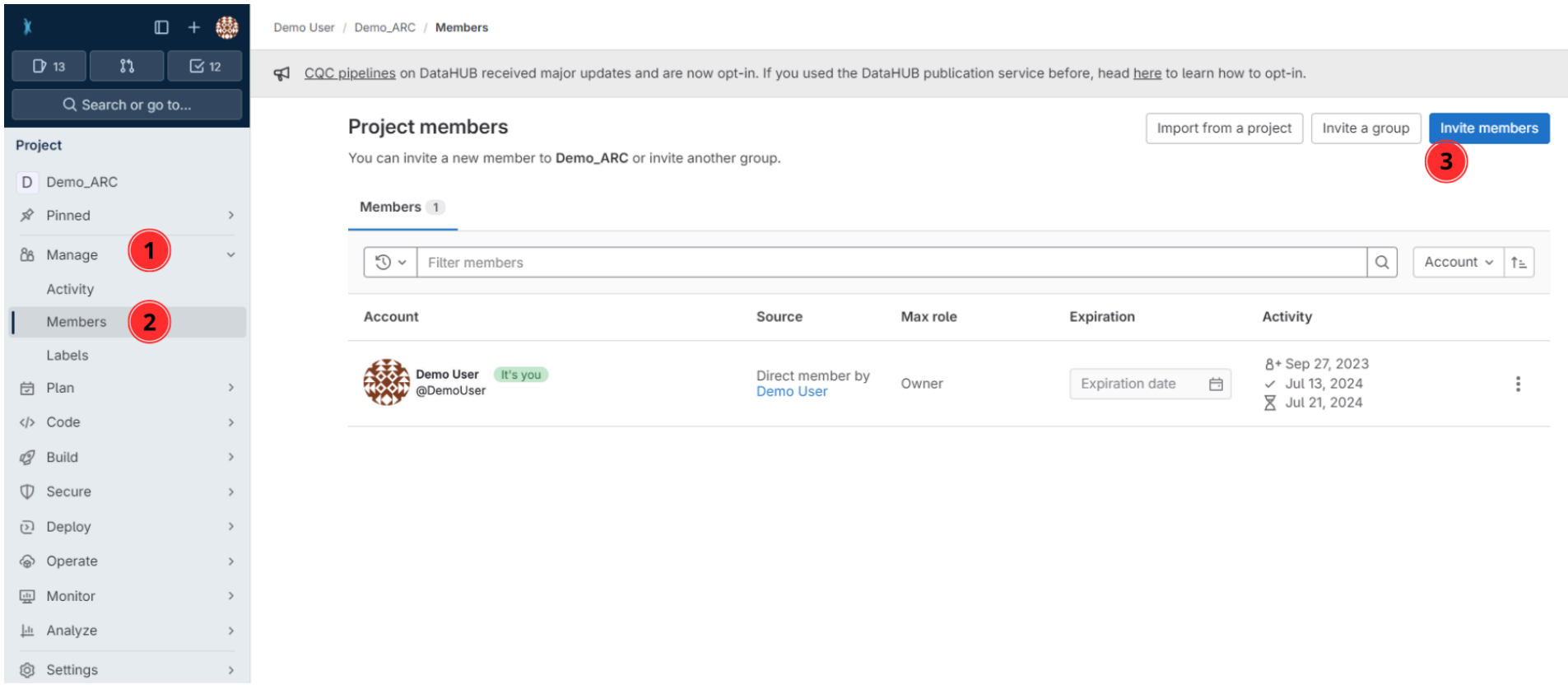
1. Click on **Project Information** in the left navigation panel



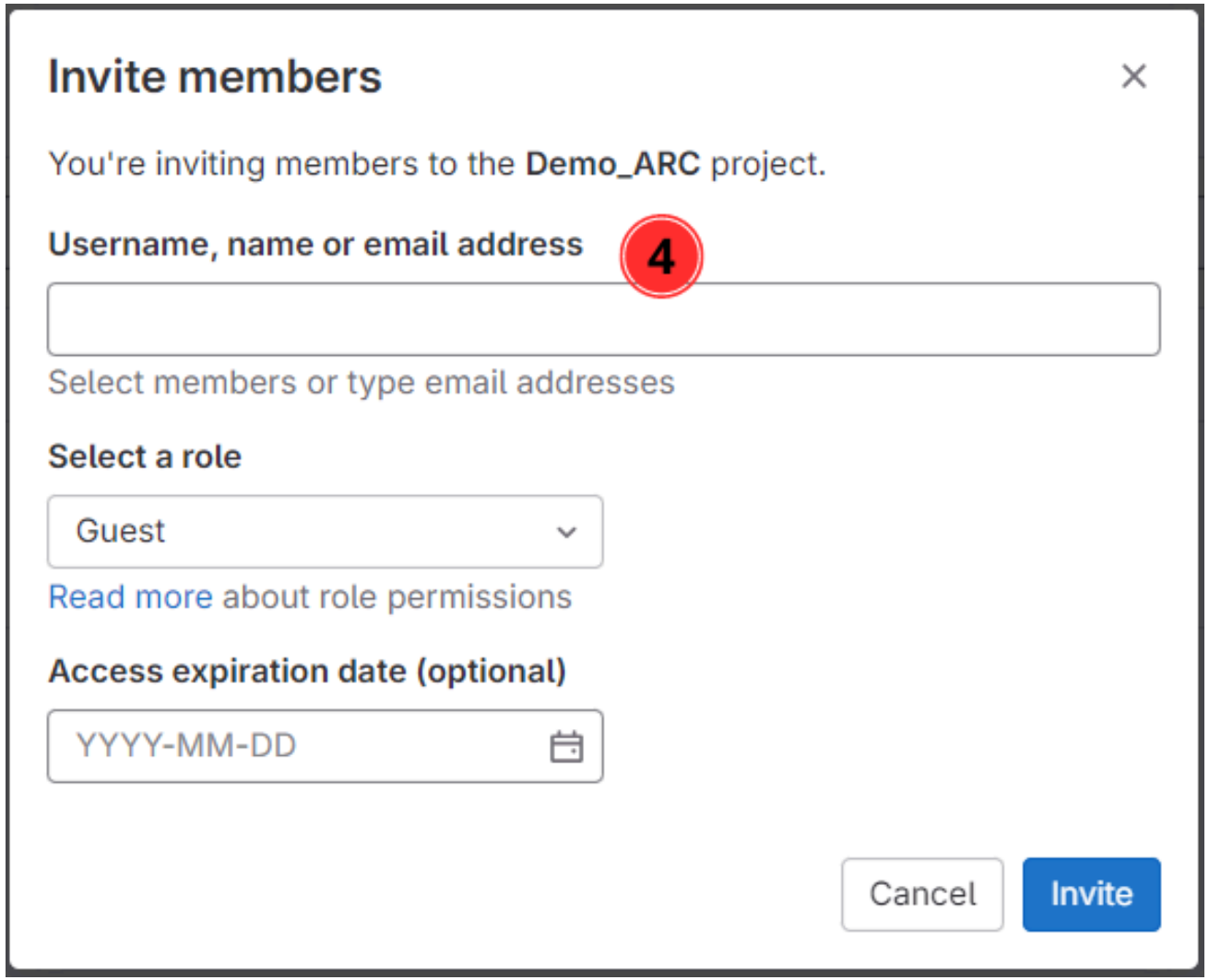
2. Click on **Members**



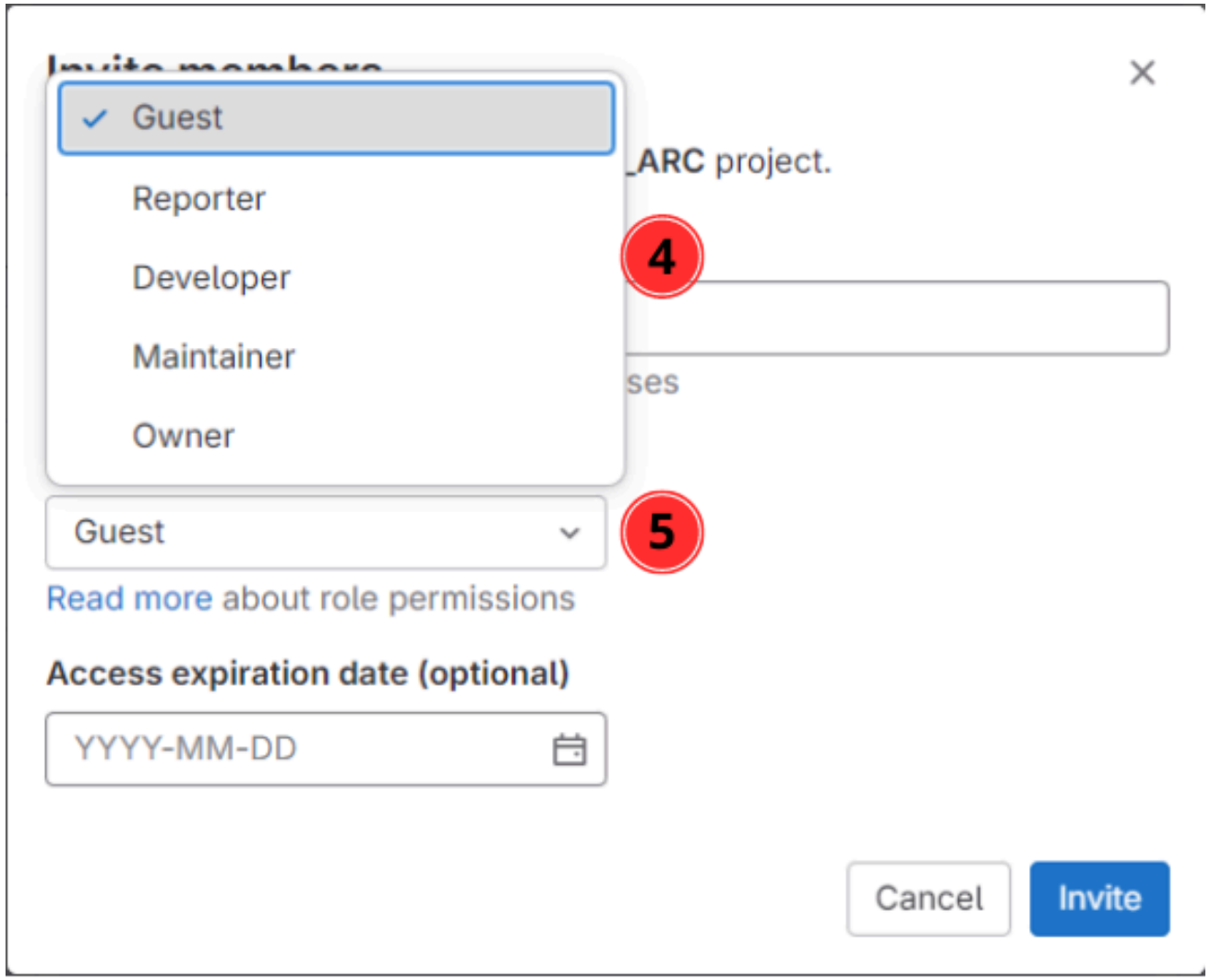
3. Click on **Invite members**



4. Search for potential collaborators

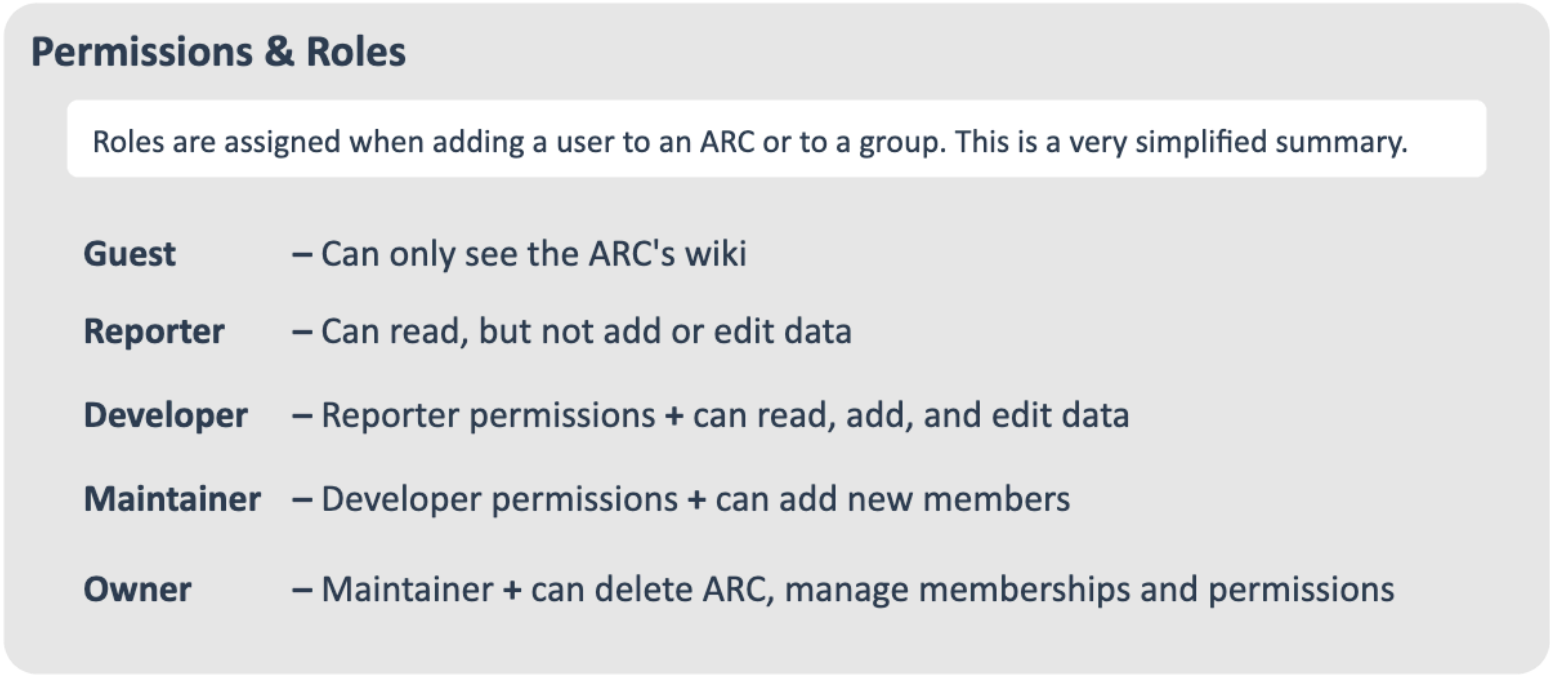


5. Select a role



Choosing the proper role

When inviting new members to an ARC or group, you can choose between different levels.



💡 By default you are **Owner** of an ARC you create or upload to the DataHUB.

Projects and Groups are not the same

- "Project" = ARC
- "Groups" = Group of users

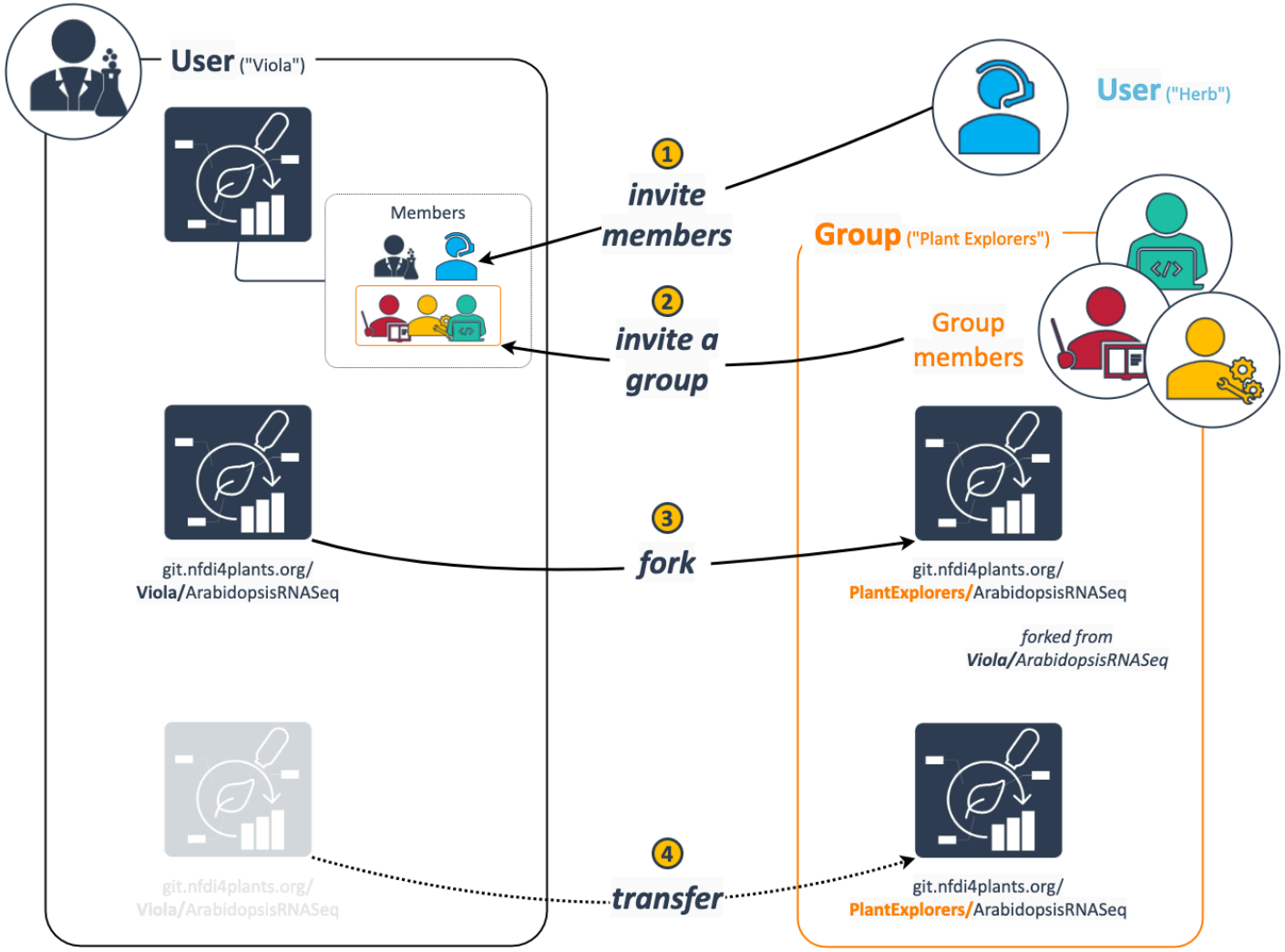
Project = ARC

- In the DataHUB, ARCs are called "projects"; they are the same.
- An ARC can be shared with individual users (invited as "members") or a group.

DataHUB Groups = Group of members (e.g. lab)

- A "Group" is a group of users with specific permissions
- A group can share ARCs
- A group can be invited to an ARC
- Groups can have subgroups

Options to share an ARC via the DataHUB



Namespaces

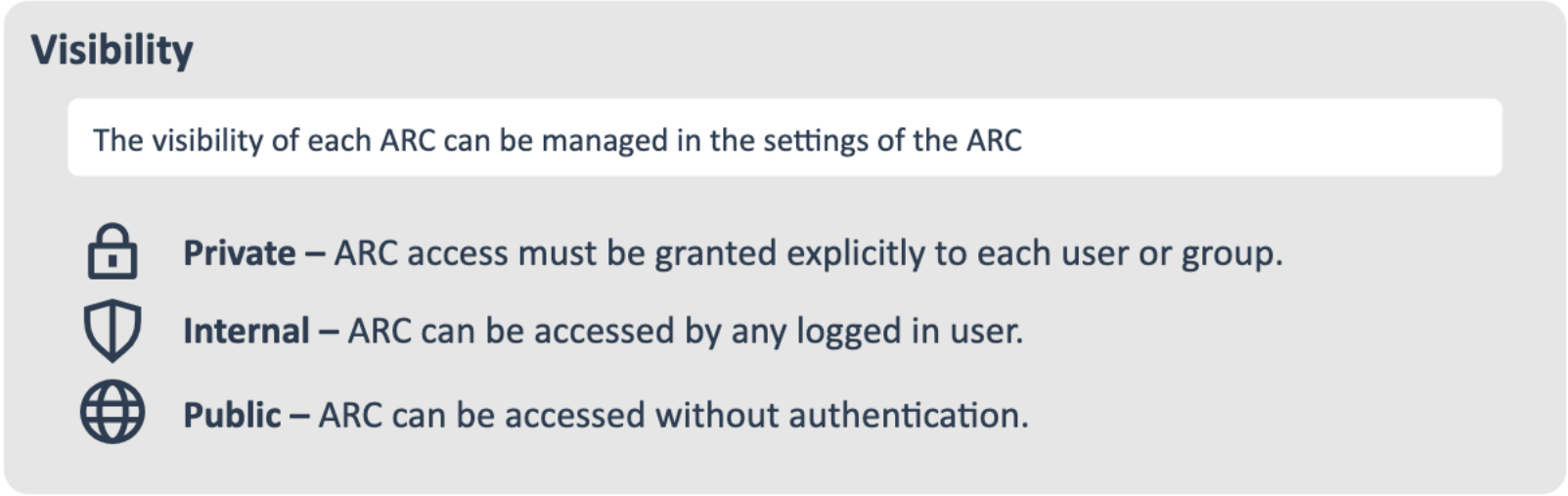
- Every user has a personal namespace, where they can upload or create new ARCs
- Every group and subgroup has its own namespace

Type	URL	Namespace	Name
A **personal** ARC	https://git.nfdi4plants.org/brilator/Facultative-CAM-in-Talinum	`brilator`	Dominik Brilhaus
An **group**-shared ARC	https://git.nfdi4plants.org/hhu-plant-biochemistry/Samuilov-2018-BOU-PSP	`hhu-plant-biochemistry`	HHU Plant Biochemistry

💡 **Personal** is not the same as **private**

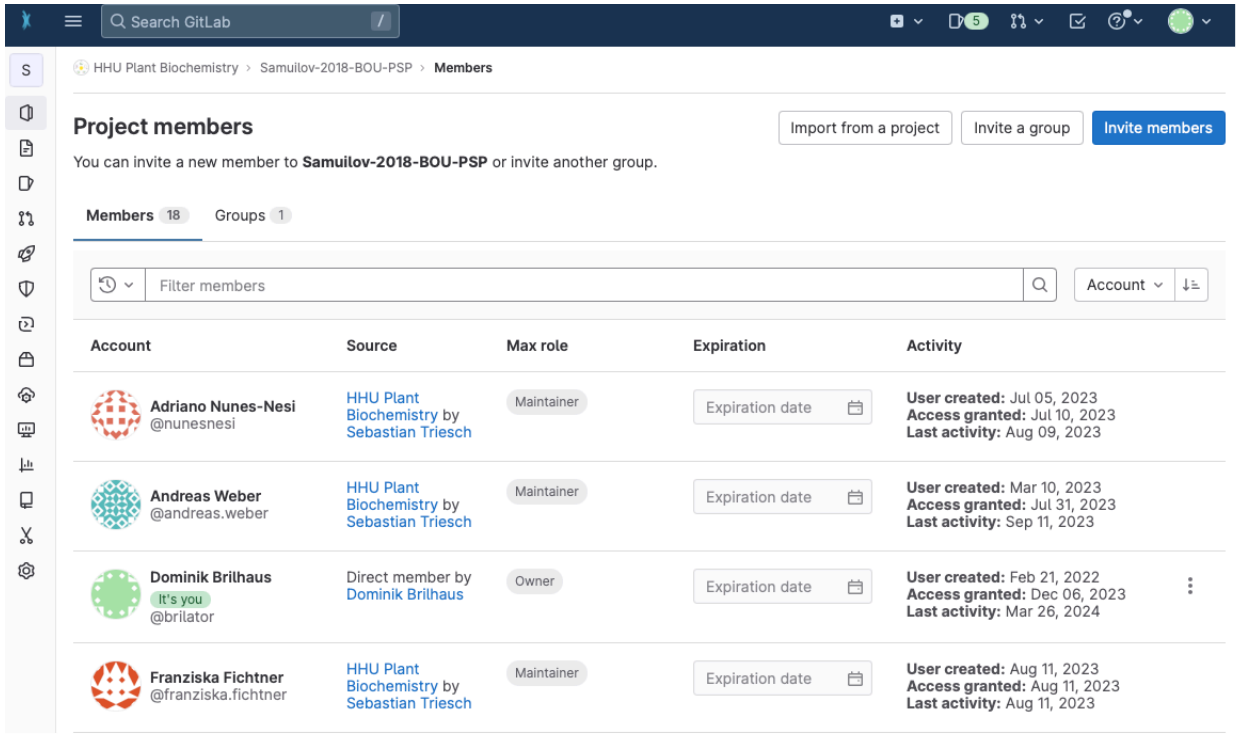
Visibility

The visibility of ARCs and groups can be managed individually for each ARC or group

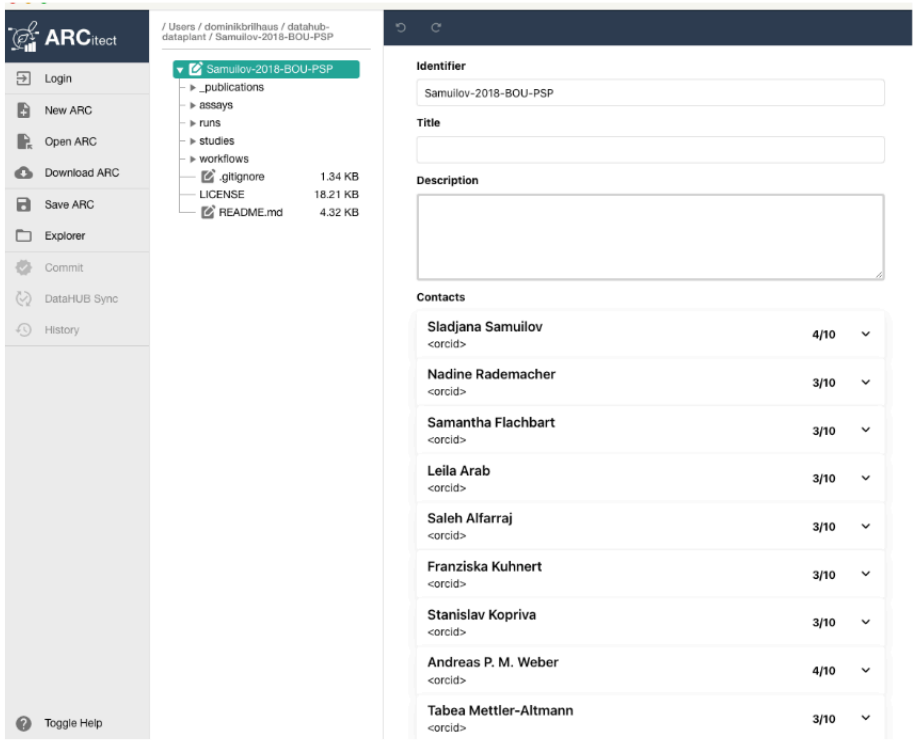


💡 By default every ARC and every group is set to **private**.

ARC DataHUB members // ARC Investigation contacts



DataHUB: ARC members



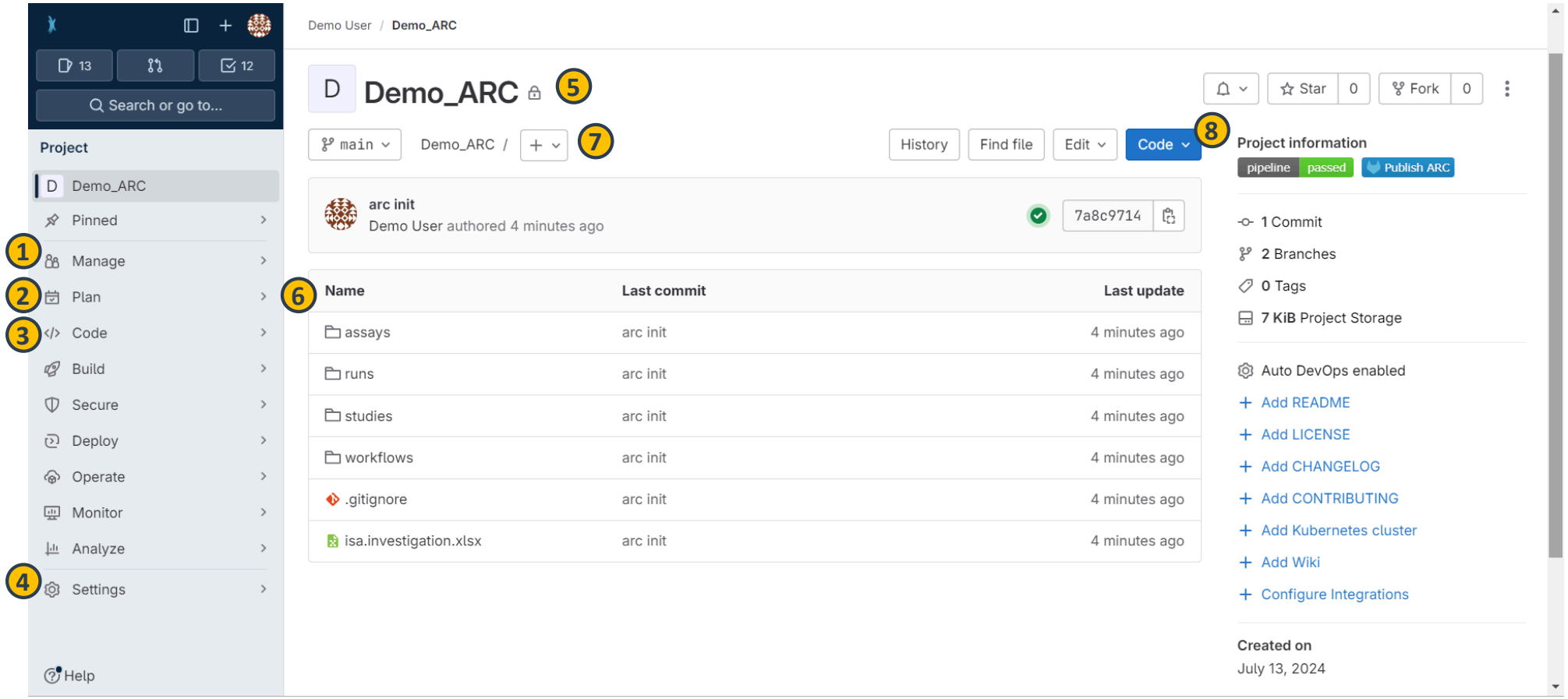
ARCitect: Investigation Contacts

https://git.nfdi4plants.org/hhu-plant-biochemistry/Samuilov-2018-BOU-PSP/-/project_members

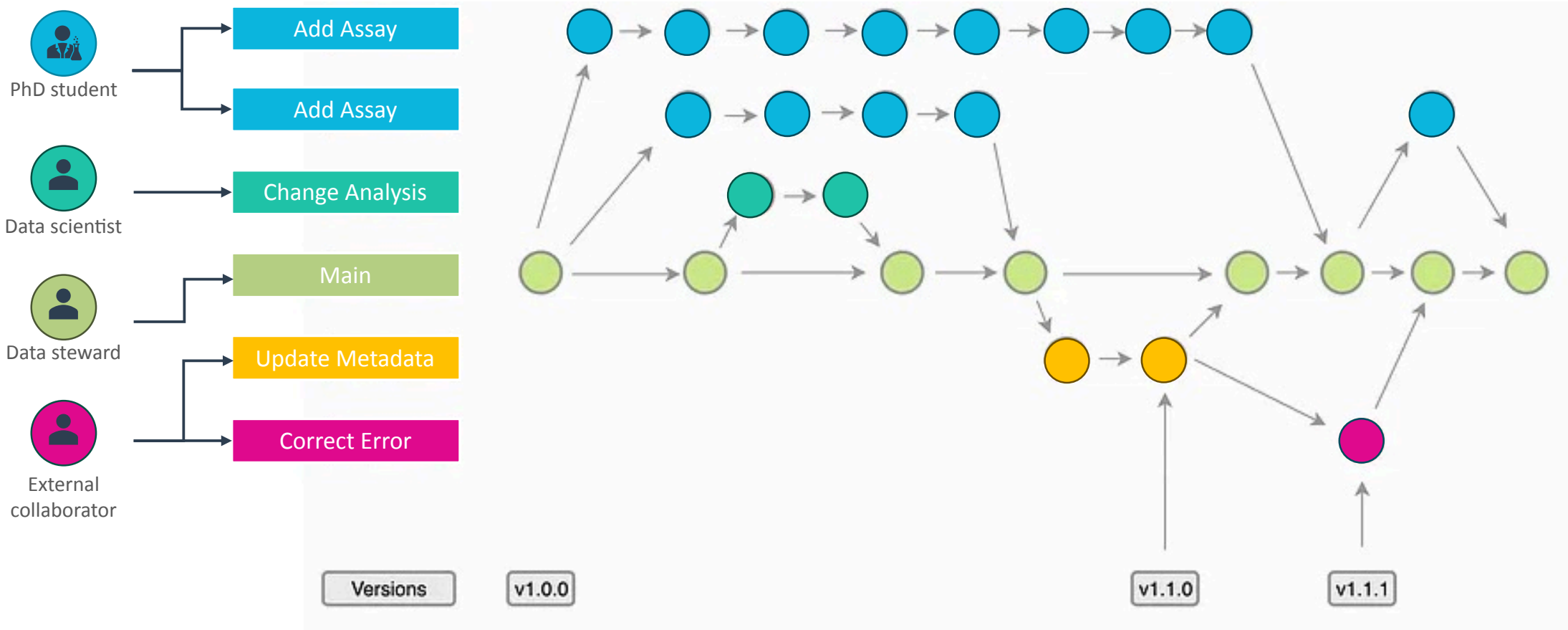
Investigation contacts are not automatically invited as members to the ARC.

Version control

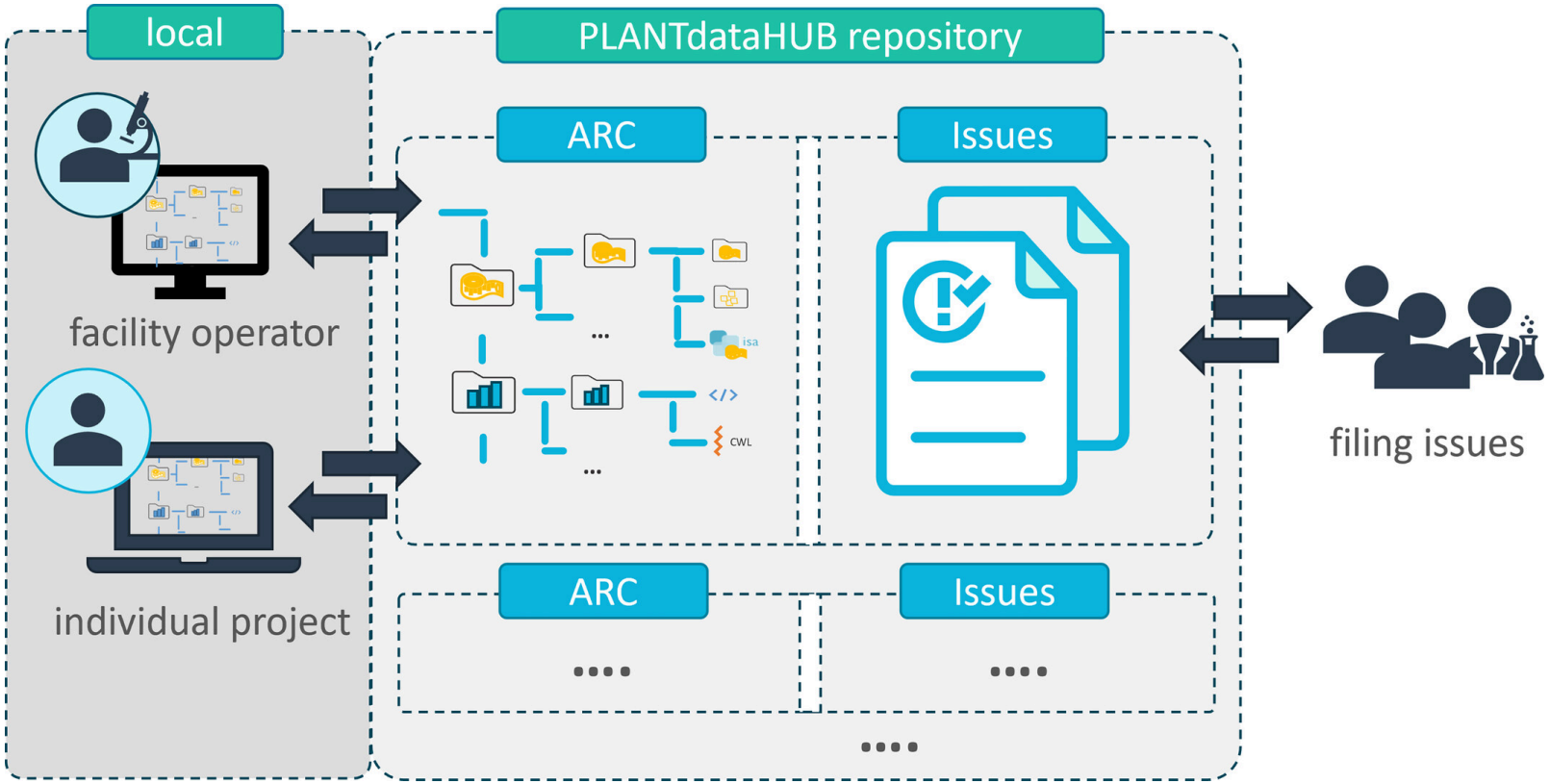
Check out the **commit history** of your ARC via Repository (2) or directly via commits (7)



Collaboration Workflows



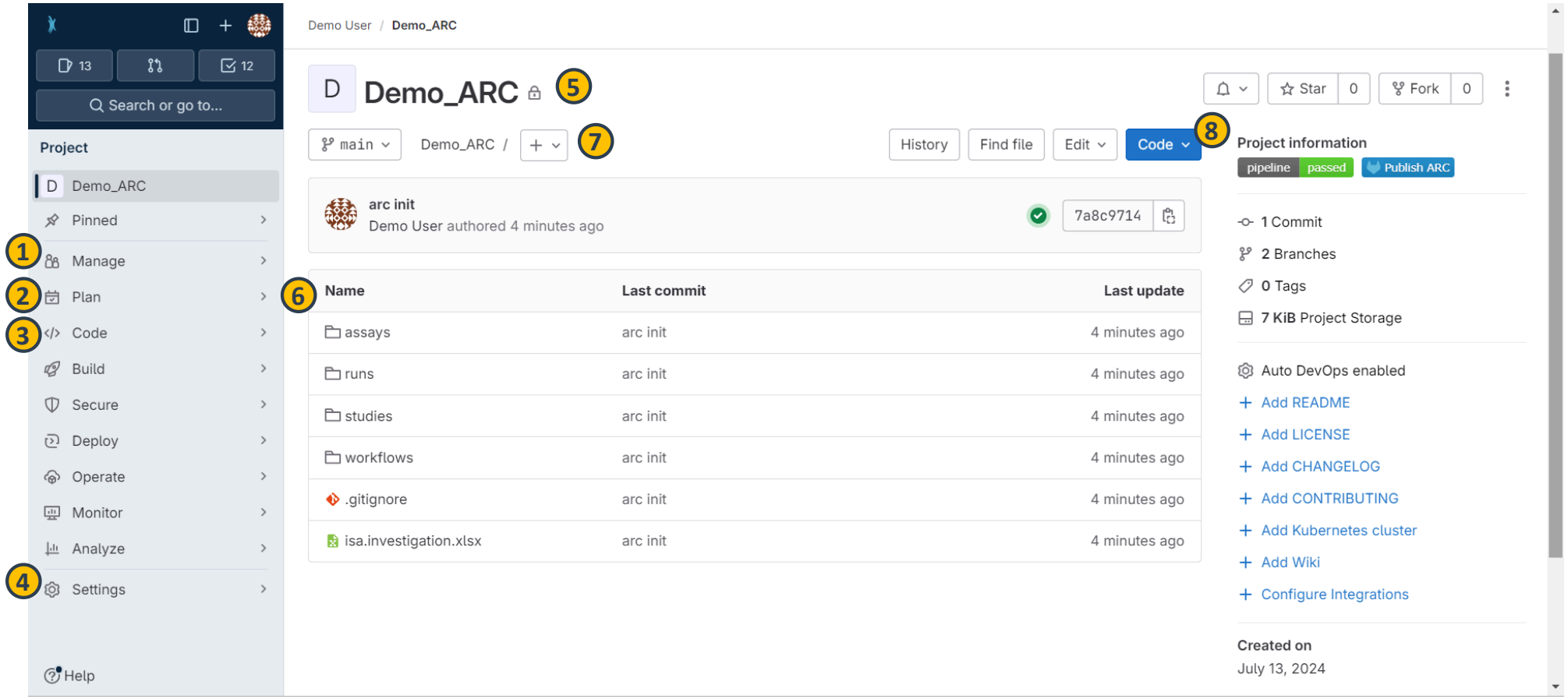
Project management



Weil, H.L., Schneider, K., et al. (2023), PLANTdataHUB: a collaborative platform for continuous FAIR data sharing in plant research. Plant J. https://doi.org/10.1111/tpj.16474

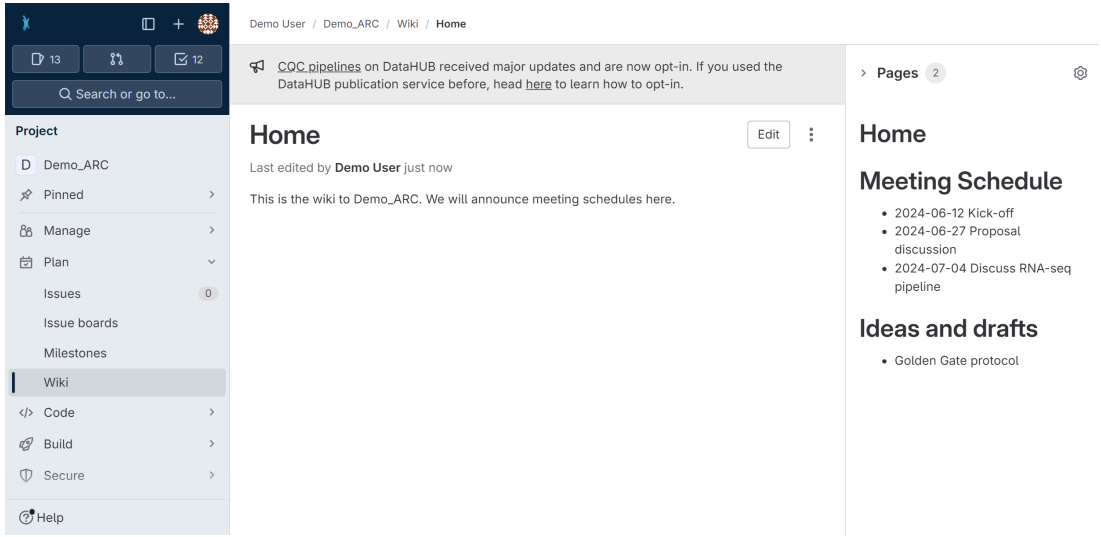
Project Management

Add issues to your ARC (3)

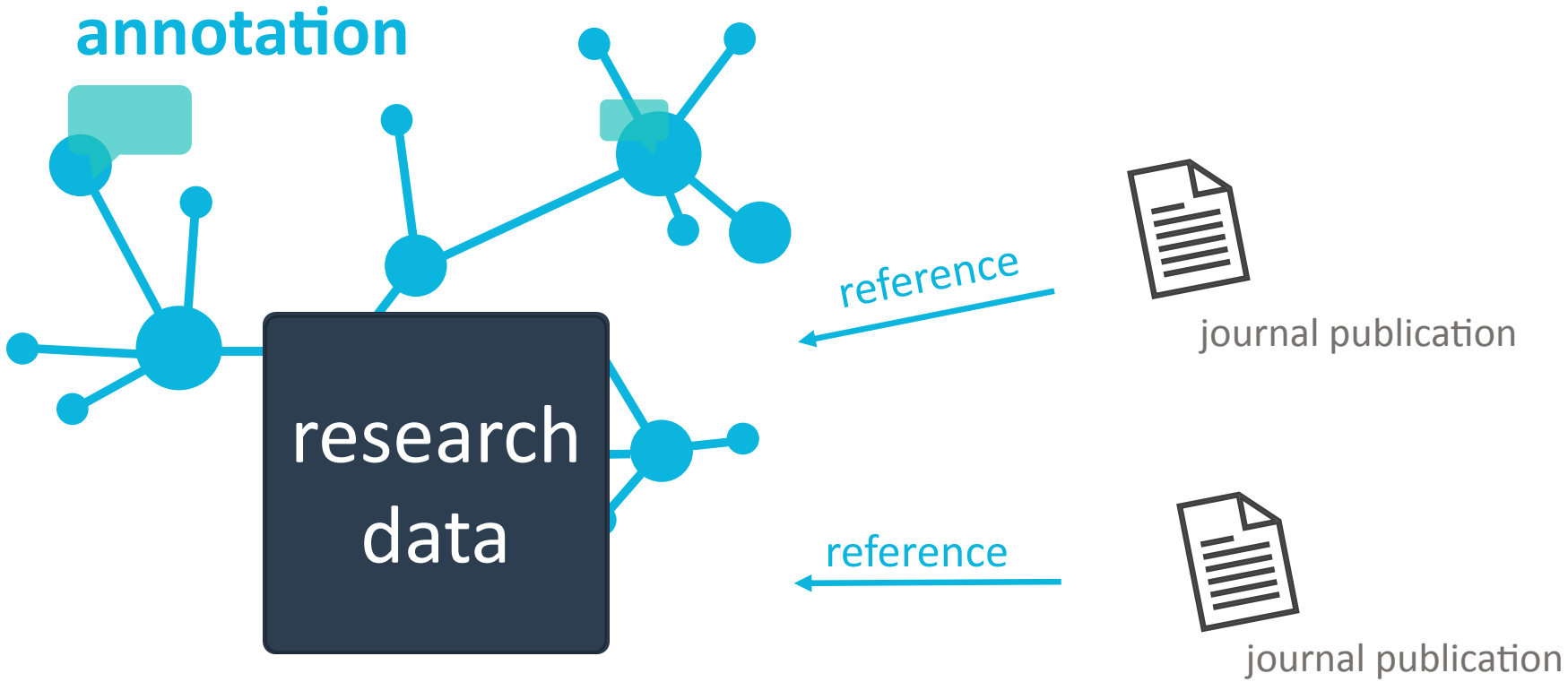


ARCs come with their own wiki space

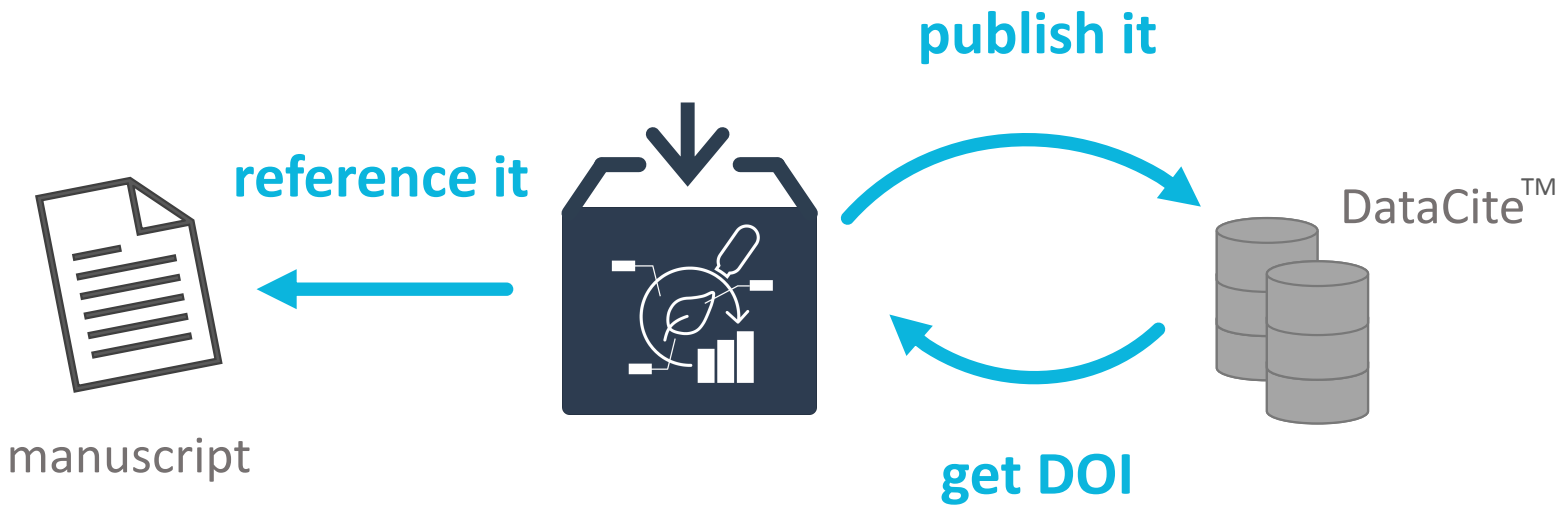
- directly associated to your ARC
- same access rights as your ARC
- share meeting minutes or ideas with collaboration partners
- keep ARC clean of files that are not considered "research data"



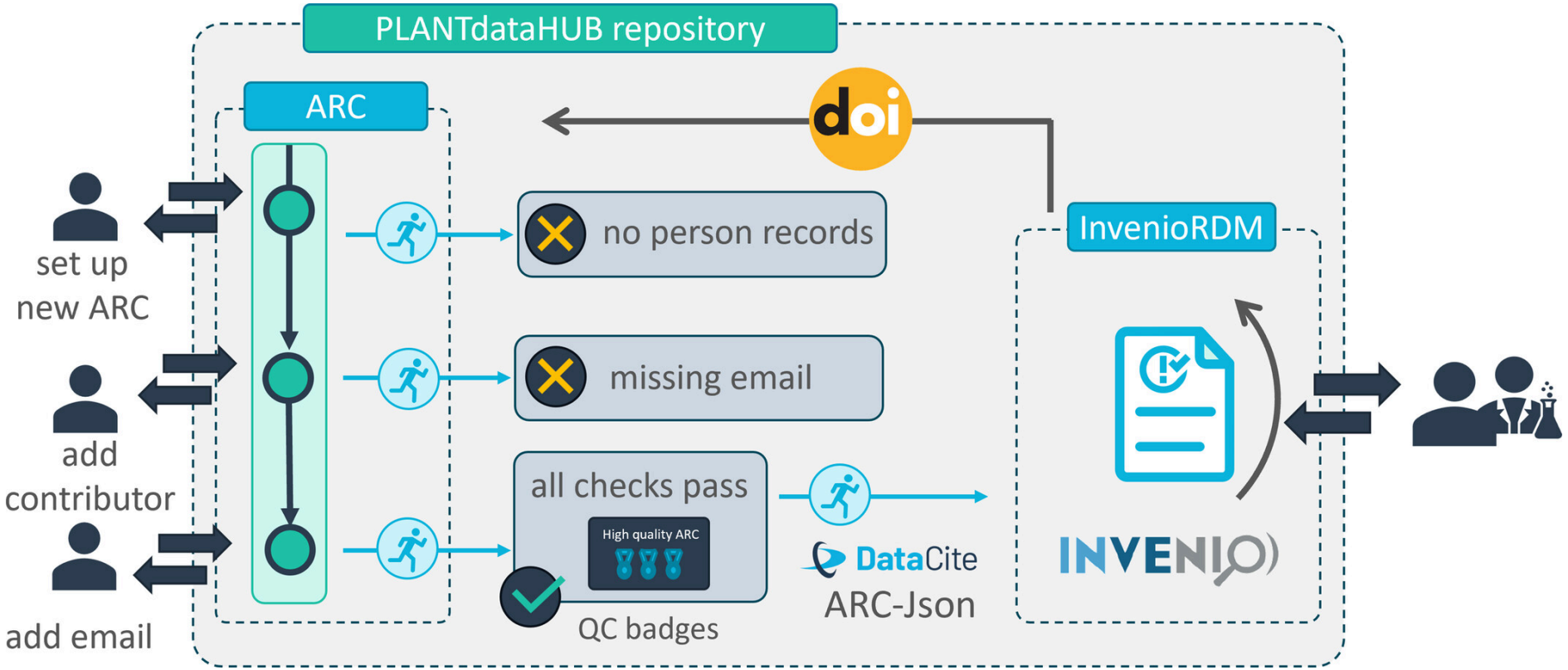
Moving from paper to FAIR data publication



Publish your ARC, get a DOI

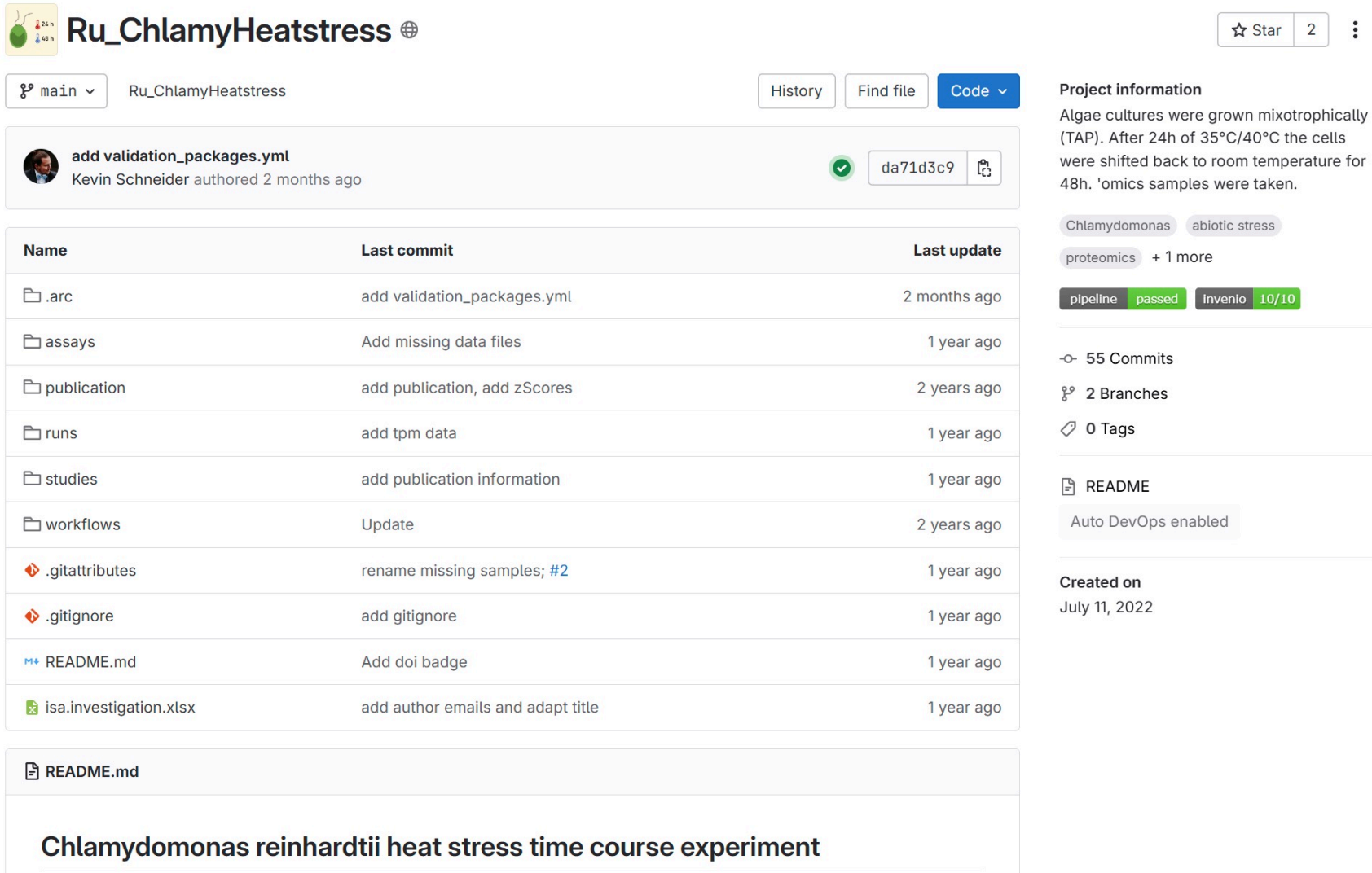


Validate & publish



Weil, H.L., Schneider, K., et al. (2023), PLANTdataHUB: a collaborative platform for continuous FAIR data sharing in plant research. Plant J. https://doi.org/10.1111/tpj.16474

Validation towards publication



Ru_ChlamyHeatstress

Star 2

main | Ru_ChlamyHeatstress | History | Find file | Code

add validation_packages.yml
Kevin Schneider authored 2 months ago — da71d3c9

Name	Last commit	Last update
.arc	add validation_packages.yml	2 months ago
assays	Add missing data files	1 year ago
publication	add publication, add zScores	2 years ago
runs	add tpm data	1 year ago
studies	add publication information	1 year ago
workflows	Update	2 years ago
.gitattributes	rename missing samples; #2	1 year ago
.gitignore	add gitignore	1 year ago
README.md	Add doi badge	1 year ago
isa.investigation.xlsx	add author emails and adapt title	1 year ago

README.md

Chlamydomonas reinhardtii heat stress time course experiment

Project information
Algae cultures were grown mixotrophically (TAP). After 24h of 35°C/40°C the cells were shifted back to room temperature for 48h. 'omics samples were taken.

Chlamydomonas | abiotic stress | proteomics | + 1 more

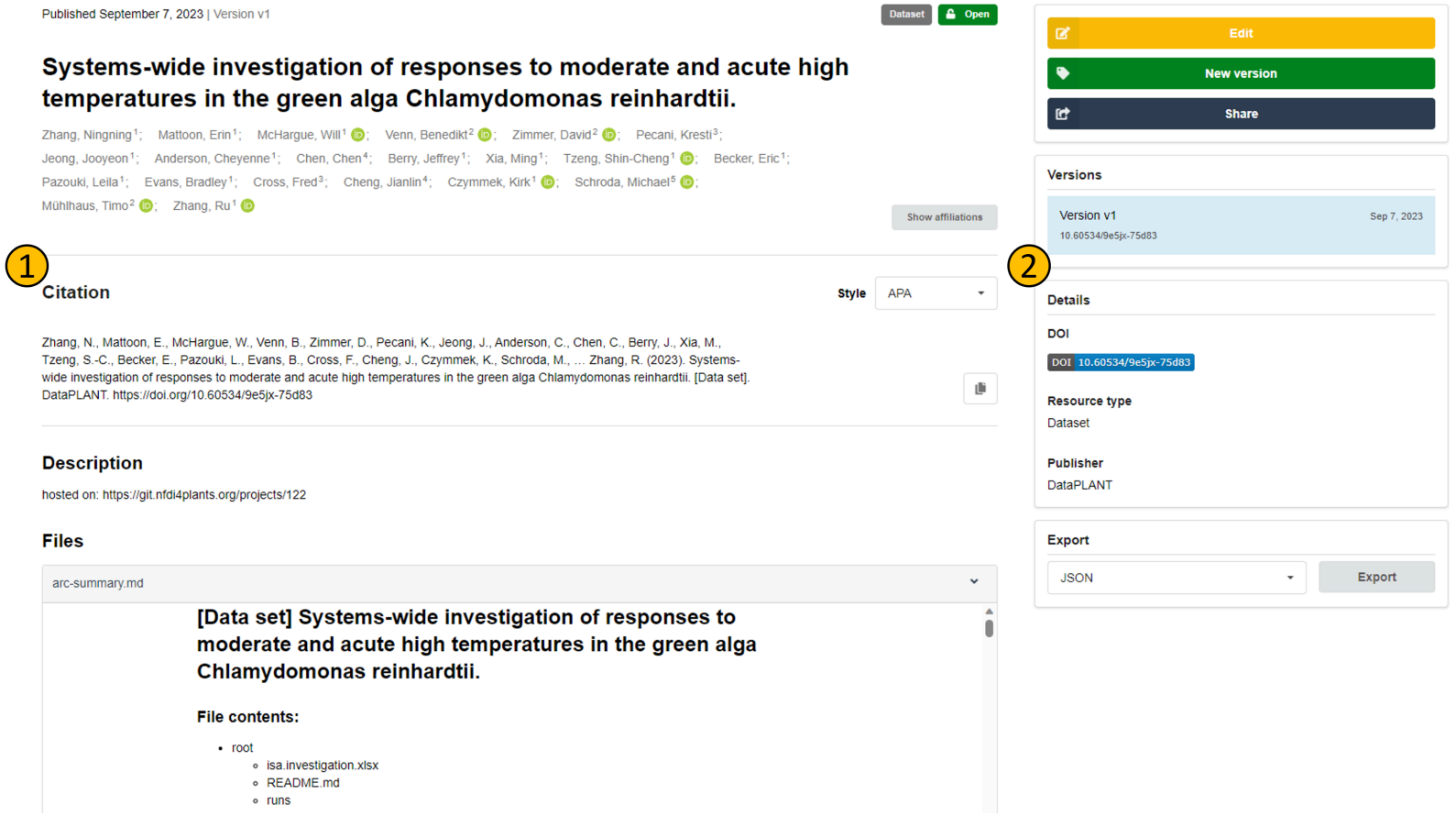
pipeline passed | invenio 10/10

55 Commits
2 Branches
0 Tags

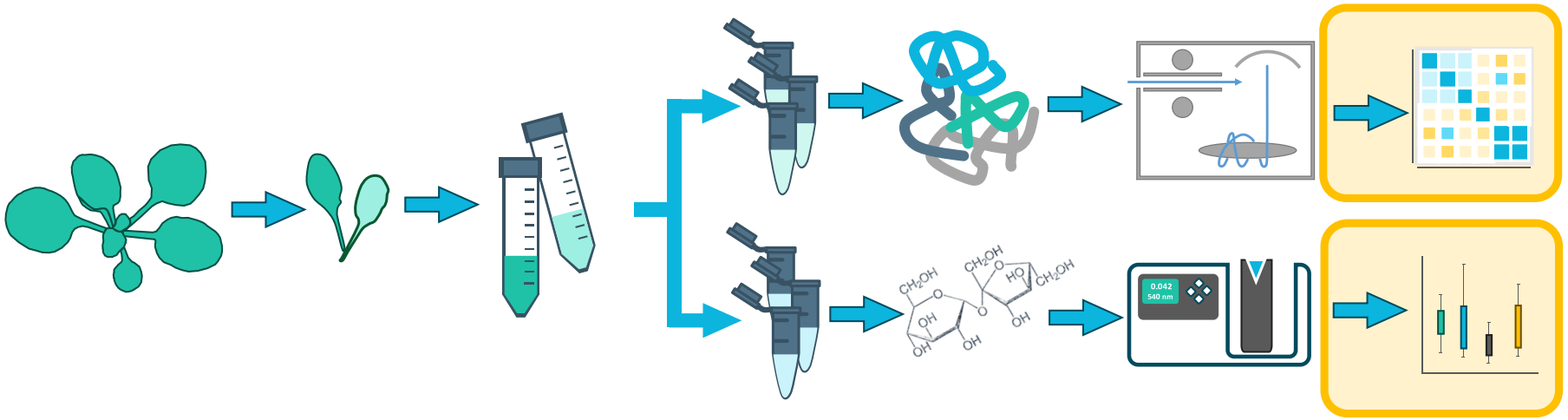
README
Auto DevOps enabled

Created on
July 11, 2022

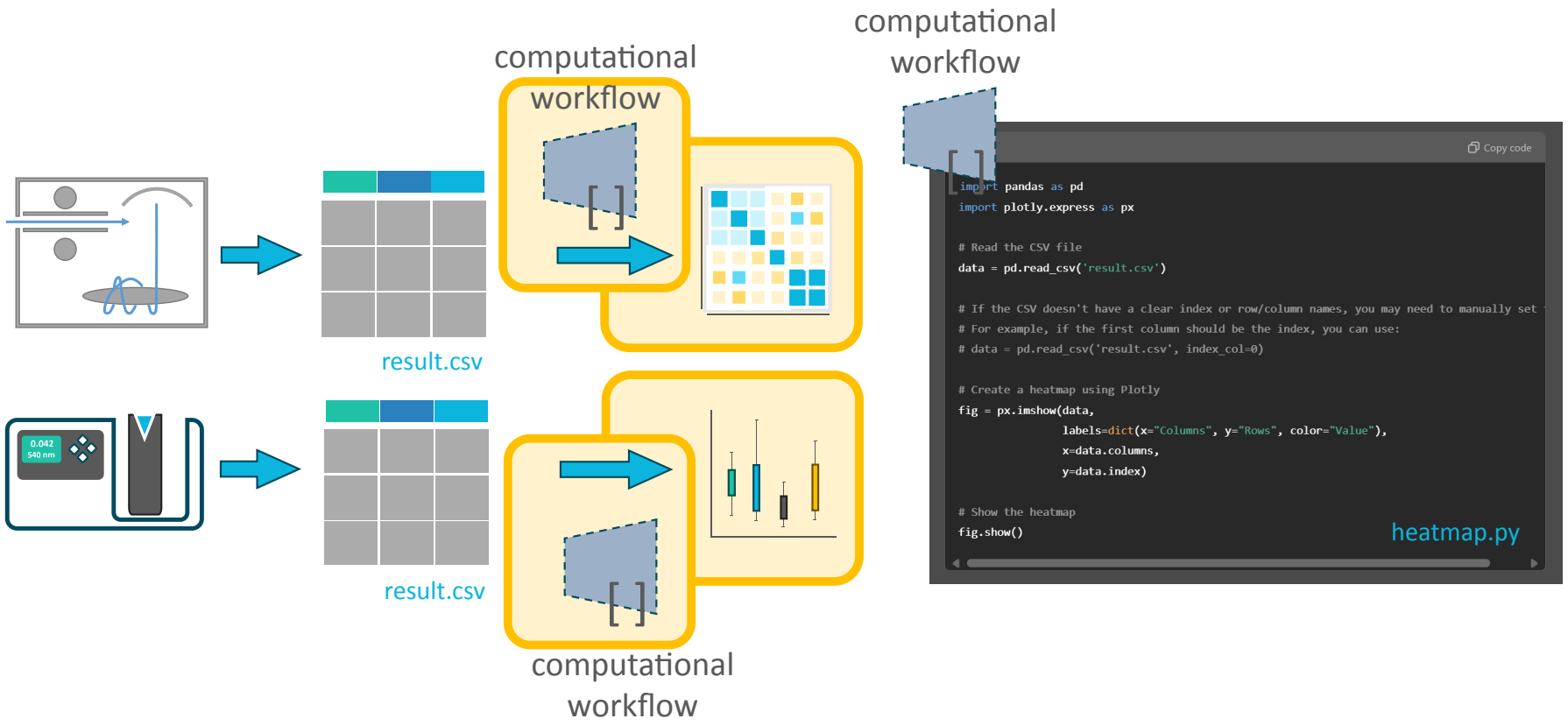
Receive a DOI



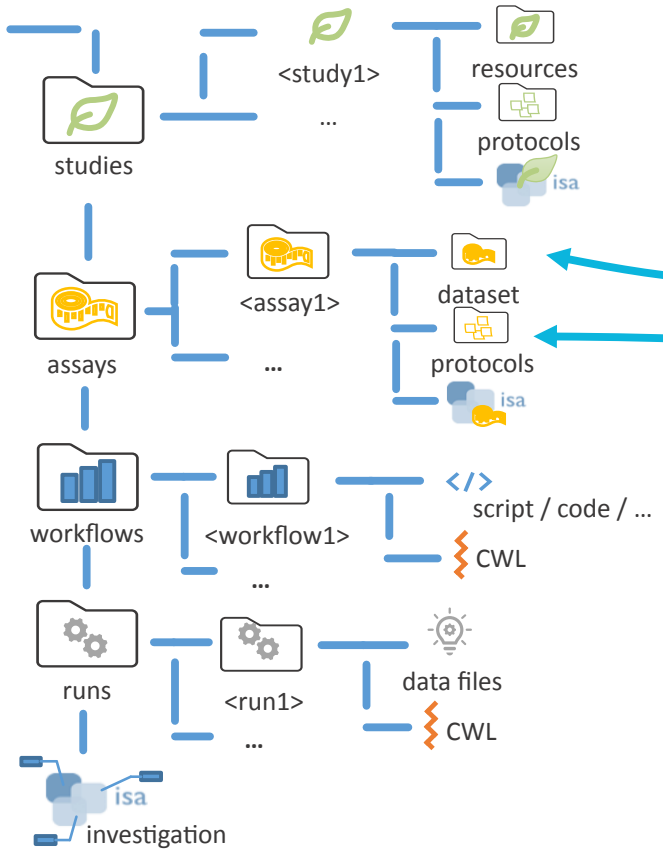
Options to annotate the data analysis



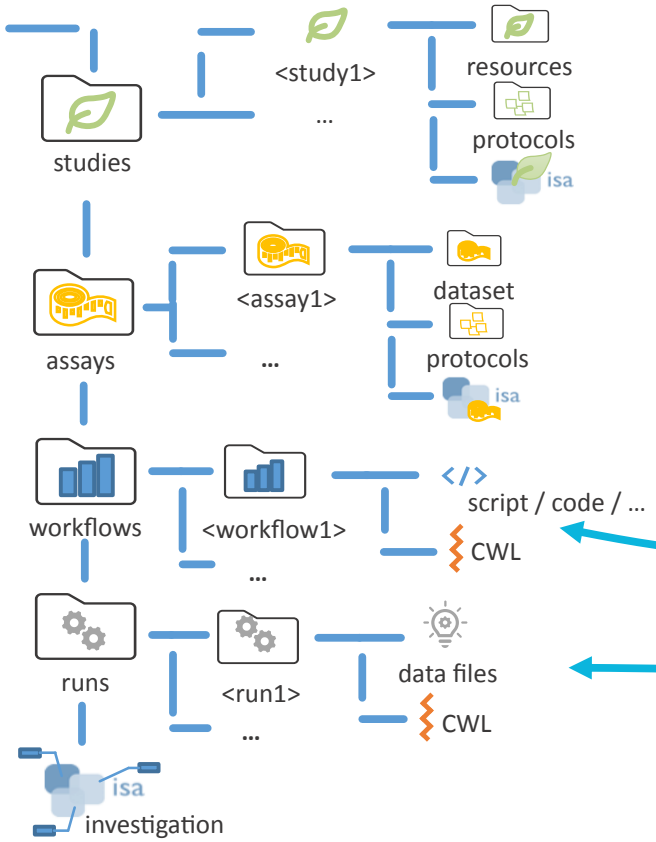
A computational workflow is like a protocol



Option I: Create a virtual assay

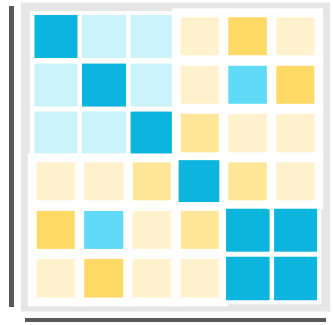


Option II: Create a workflow and run




```python
import pandas as pd
import plotly.express as px

# Read the CSV file
data = pd.read_csv('result.csv')

# If the CSV doesn't have a clear index or row/column names, you may need to manually set
# For example, if the first column should be the index, you can use:
# data = pd.read_csv('result.csv', index_col=0)

# Create a heatmap using Plotly
fig = px.imshow(data,
                labels=dict(x="Columns", y="Rows", color="Value"),
                x=data.columns,
                y=data.index)

# Show the heatmap
fig.show()
```

heatmap.py

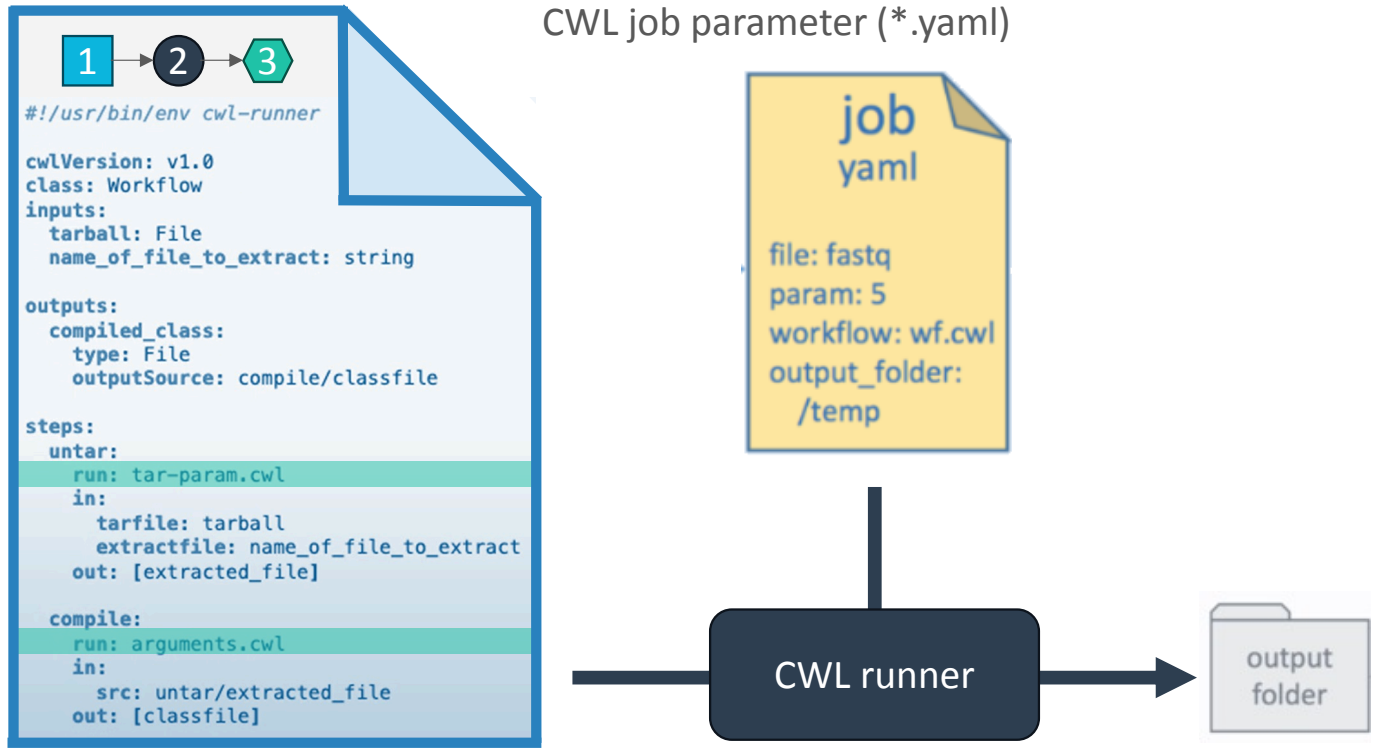


Use CWL to wrap your workflow

CWL workflow document (*.cwl)

CWL job parameter (*.yaml)

CWL
tool descriptors (*.cwl)



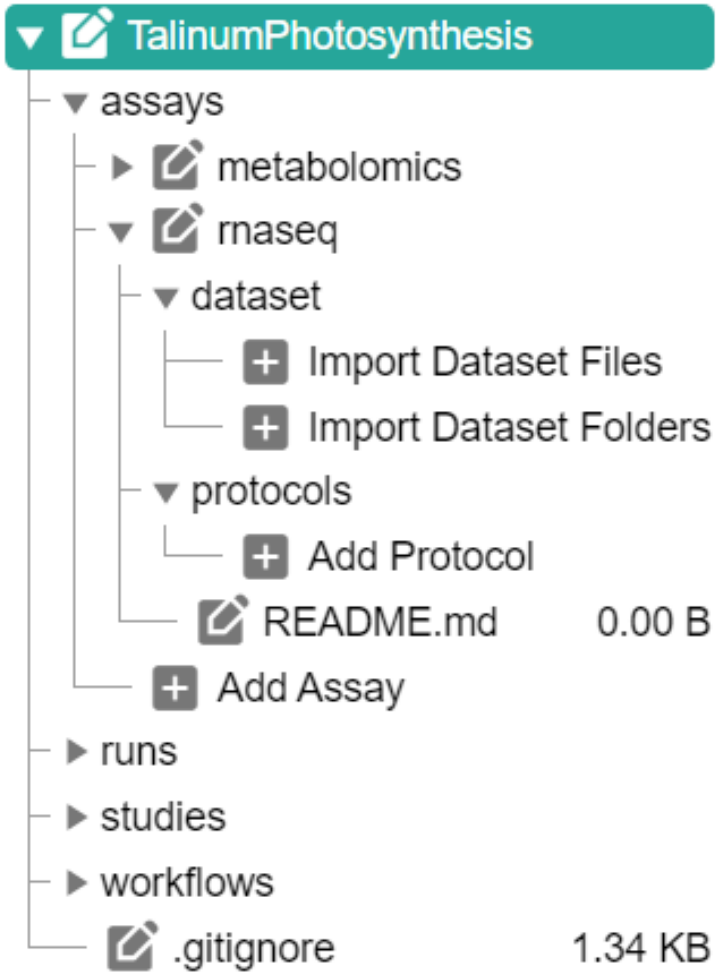
Add datasets

In the file tree you can

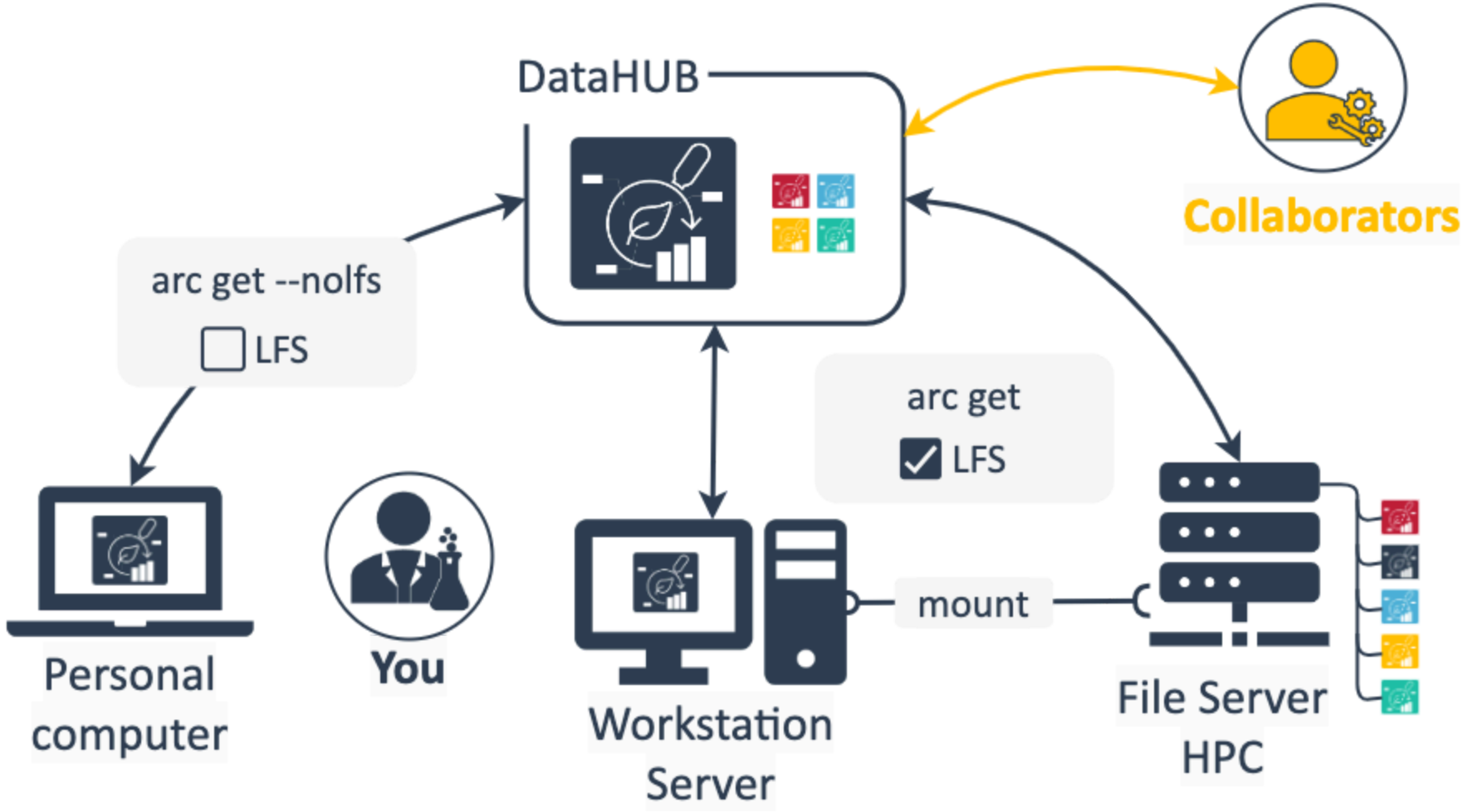
- **import dataset files or folders** and
- **protocols** associated to that dataset.

💡 **Import Dataset** allows to import data from any location on your computer into the ARC.

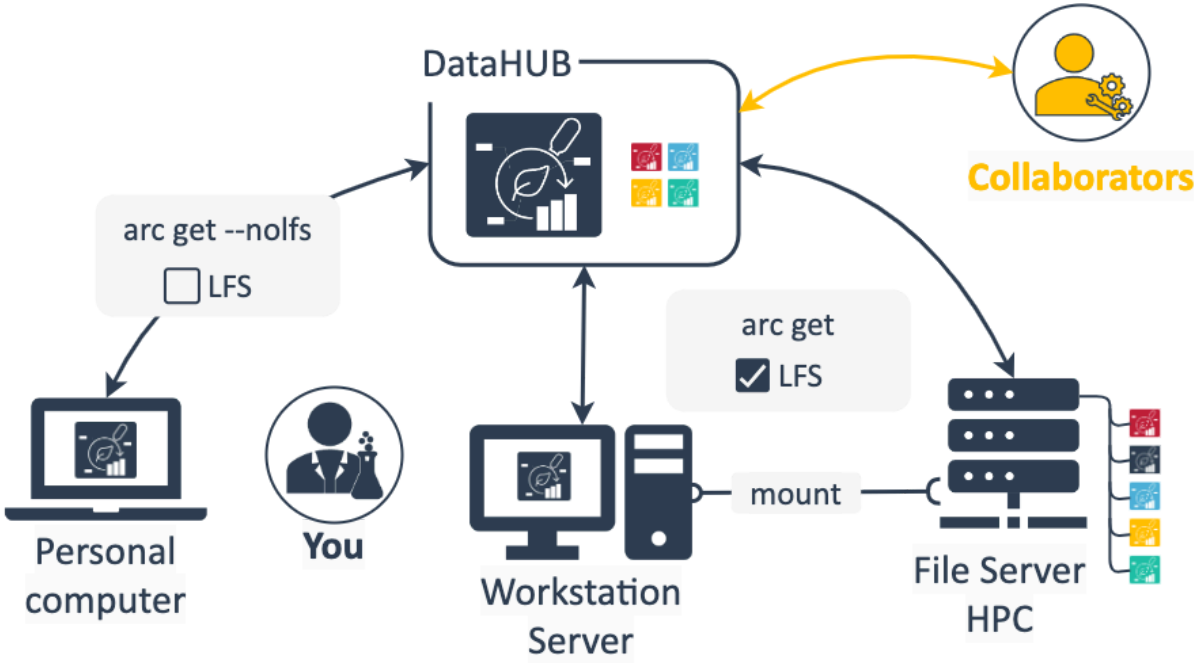
⚠️ Depending on the file size, this may take a while. Test this with a small batch of files first.



Where do I store my ARC?



ARC storage and sharing



- DataHUB as "ground truth" / original clone
- You can sync and communicate all changes to your ARC via the DataHUB

ARCitect and ARC commander provide options to avoid syncing large files (LFS = Large file storage)

Example setup to store and use ARCs

Personal computer

- work on small files
- annotate metadata
- add scripts, protocols

Workstation / Server

- work on large files
- run computations

FileShare

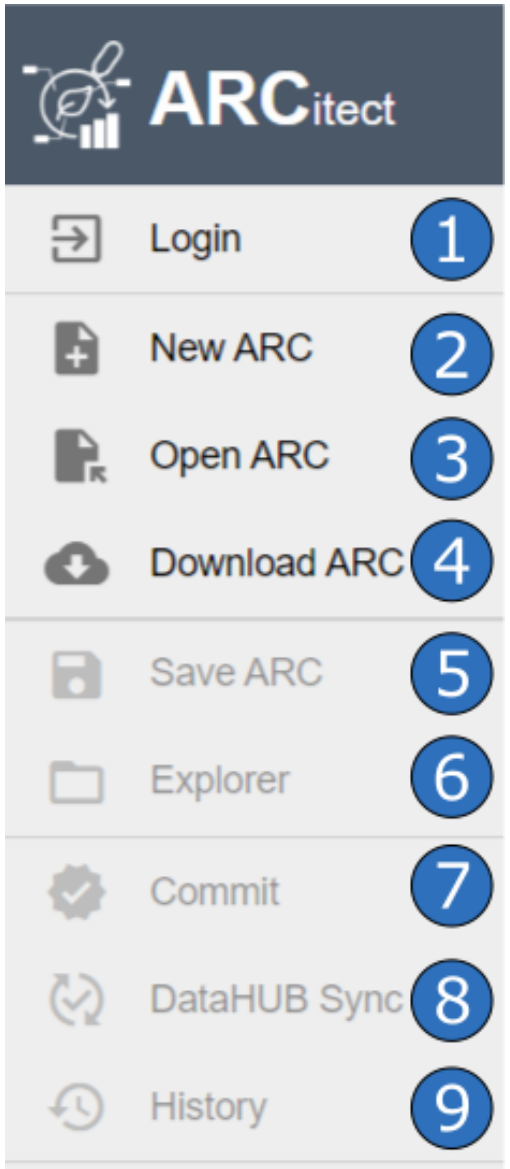
- mount to local machine, sync ARC from there

HPC

- direct connection HPC to DataHUB (depends on security settings)
- or mount to local machine and sync, ARC from there

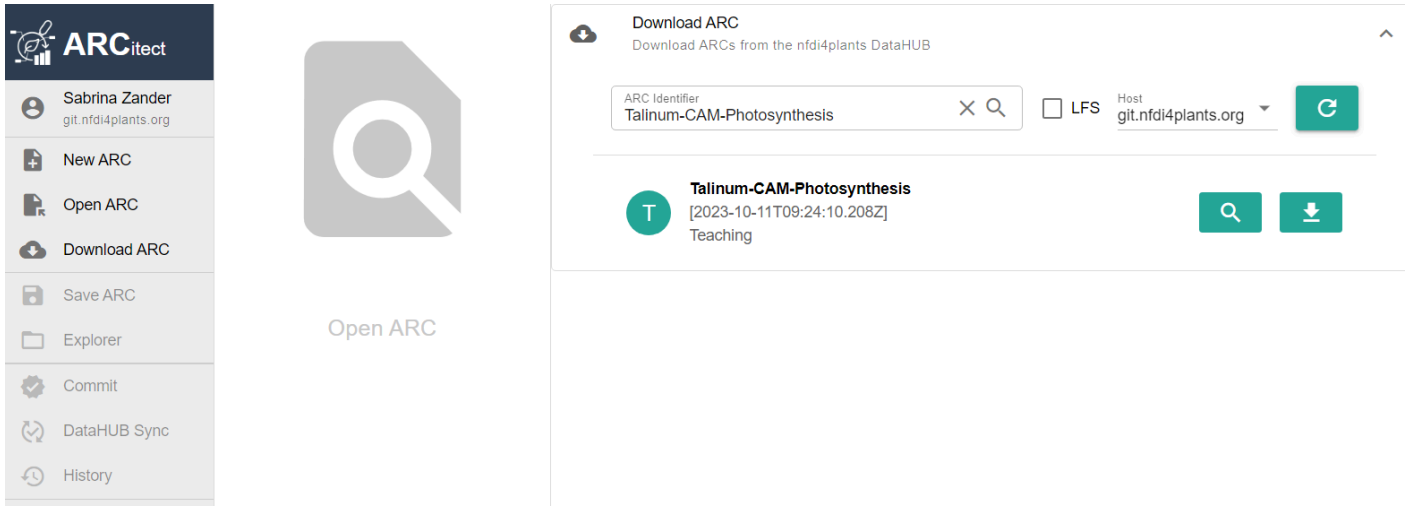
Download the demo data

1. Open the ARCitect
2. Login (1) to your DataHUB account
3. Navigate to **Download ARC** (4)



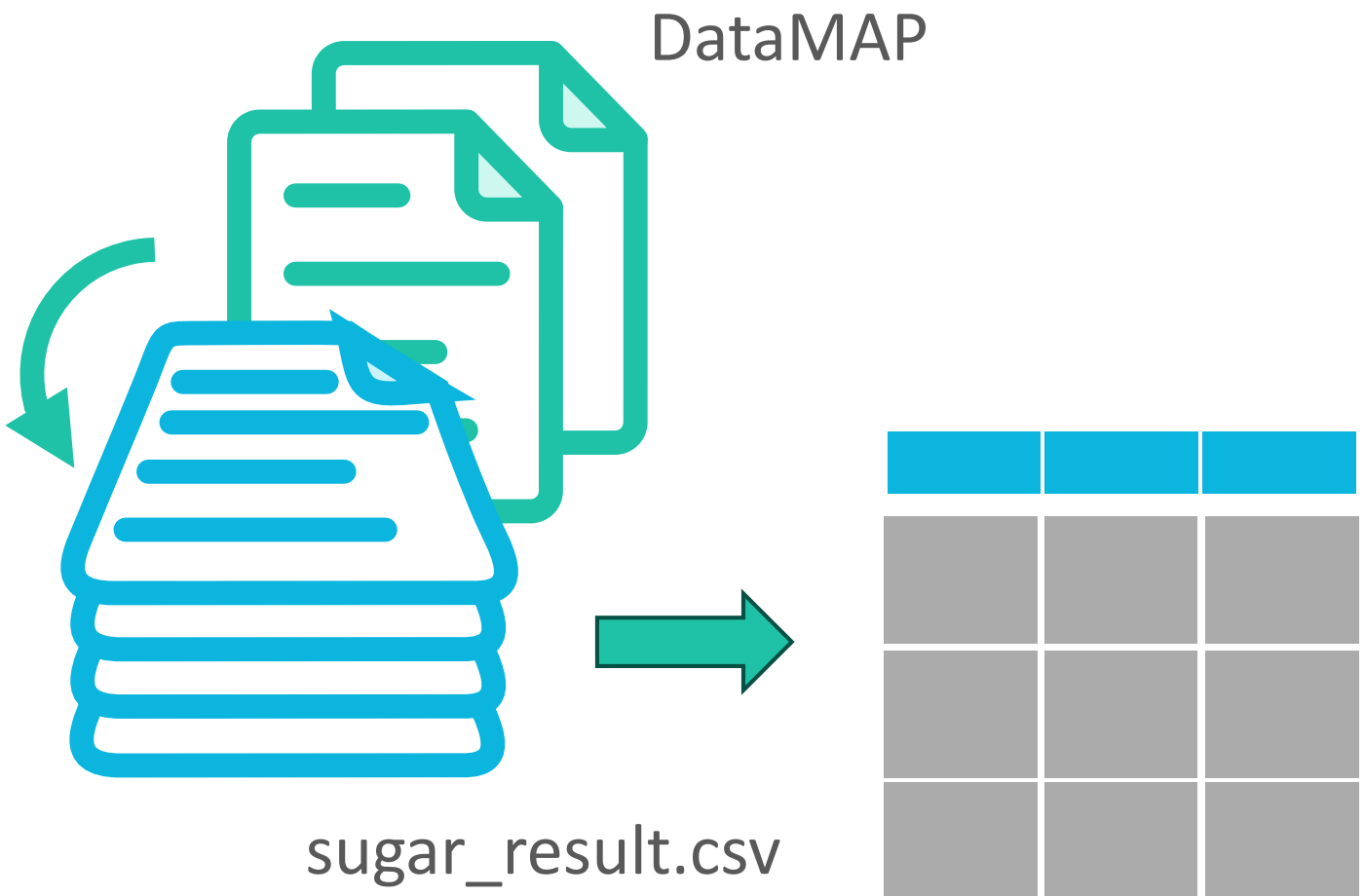
Download the demo data

4. Search for **Talinum-CAM-Photosynthesis**
5. Click the download button, select a location and open the ARC.
6. Open the downloaded ARC



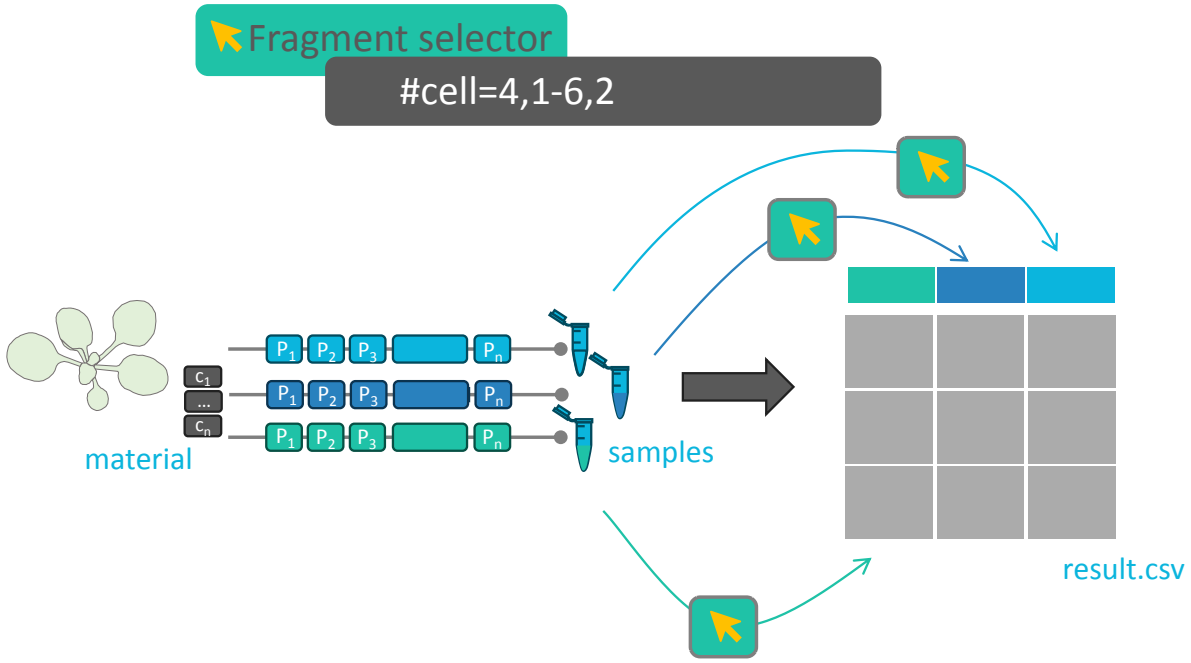
💡 This is basically the ARC we created in the last session.

Annotation of the result data points

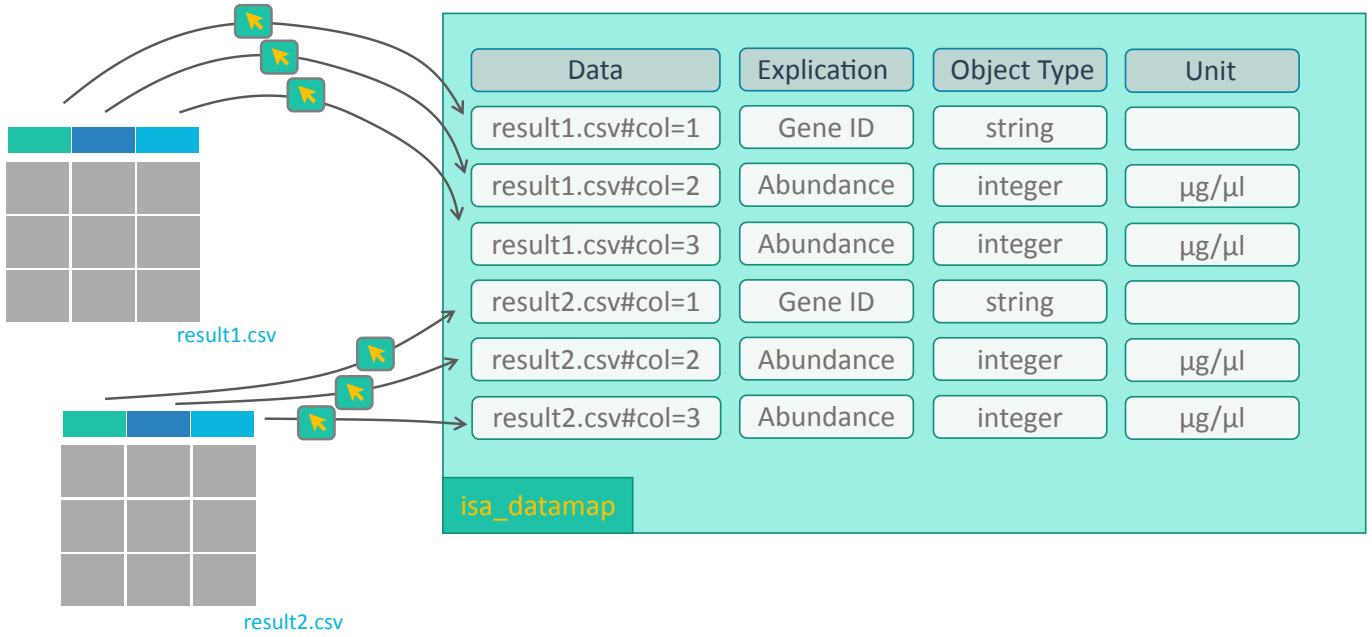


Persistence from sample to data point using fragment selectors

Point into results - Fragment selectors



DataMAP: Annotation for the fragment selectors



Contributors

If not referenced otherwise, figures and slides presented here were created by members of DataPLANT (https://nfdi4plants.org). Additional slides were contributed by

- name: Dominik Brilhaus
  github: https://github.com/brilator
  orcid: https://orcid.org/0000-0001-9021-3197
- name: Cristina Martins Rodrigues
  github: https://github.com/CMR248
  orcid: https://orcid.org/0000-0002-4849-1537
- name: Kevin Frey
  github: https://github.com/Freymaurer
  orcid: https://orcid.org/0000-0002-8493-1077
- name: Sabrina Zander
  orcid: https://orcid.org/0009-0000-4569-6126
- name: Martin Kuhl
  github: https://github.com/Martin-Kuhl
  orcid: https://orcid.org/0000-0002-8493-1077