ARCify your research project

fdm.nrw – FDM Werkstatt

Dominik Brilhaus, CEPLAS

April 2nd, 2025

CEPLAS
Cluster of Excellence on Plant Sciences

DataPLANT

Training Materials

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

Introduce yourself

- Science / RDM
- Used code / programming language before
- Experience with Git / GitLab / GitHub?
- My motivation / expectation

Resources

Info & materials

- DataPLANT Website: https://nfdi4plants.org/
- ARC website: https://arc-rdm.org
- Knowledge Base: https://nfdi4plants.org/nfdi4plants.knowledgebase/

Continuous support

- HelpDesk: https://helpdesk.nfdi4plants.org
- Matrix for ad hoc support: https://matrix.to/#/%23arc-user-support:matrix.org
- User support meeting (2nd Friday of the month | 1 - 2pm): https://nfdi4plants.github.io/events/arc-user-support/

Tools and Services

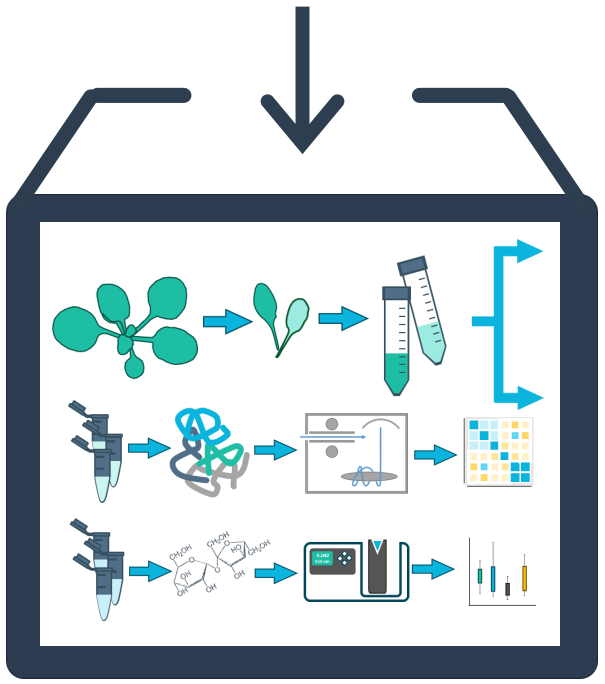
- ARCitect: https://github.com/nfdi4plants/arcitect
- DataHUB: https://git.nfdi4plants.org

Open Source Development

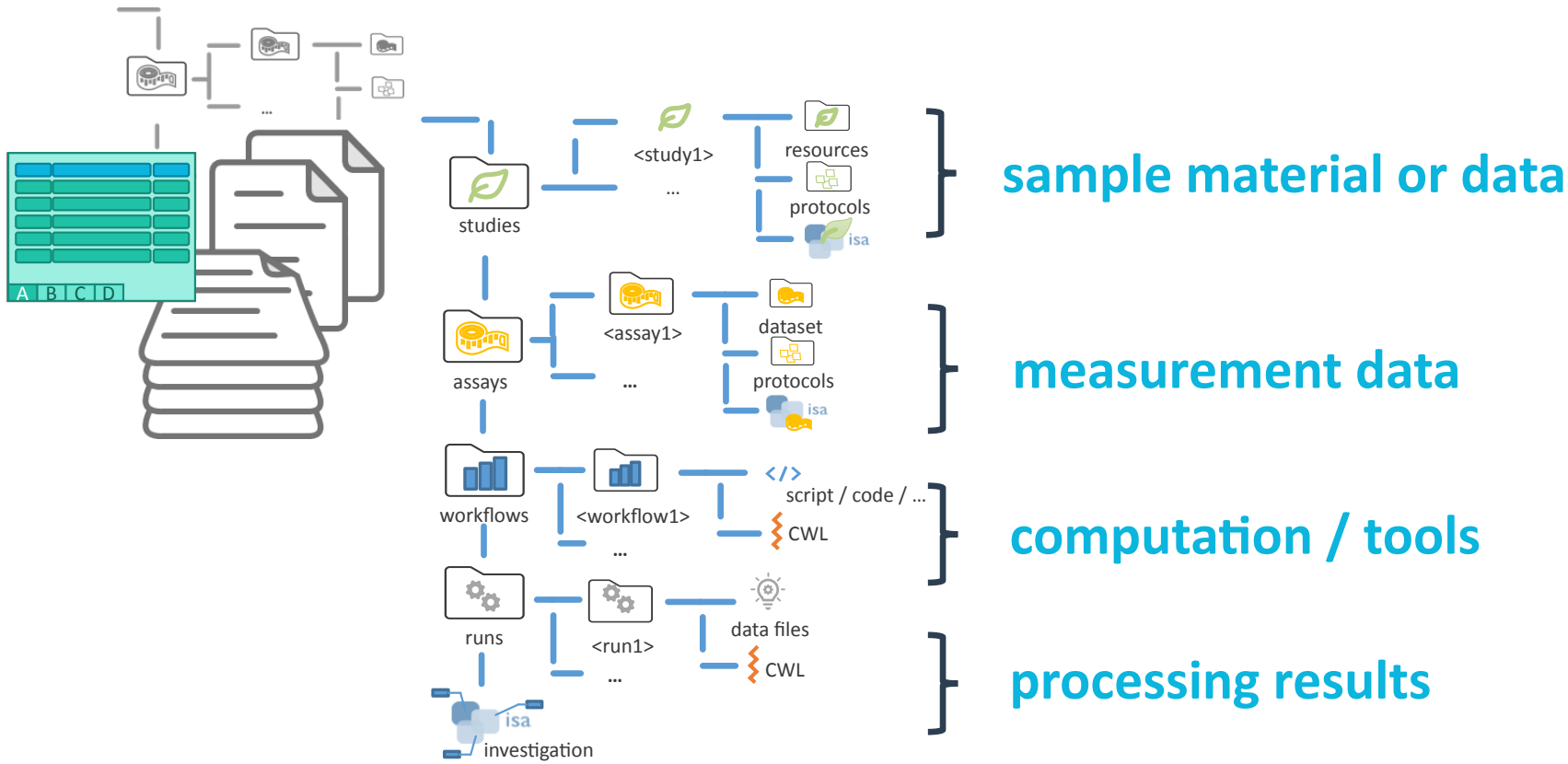
- GitHub: https://github.com/nfdi4plants

Annotated Research Context (ARC)

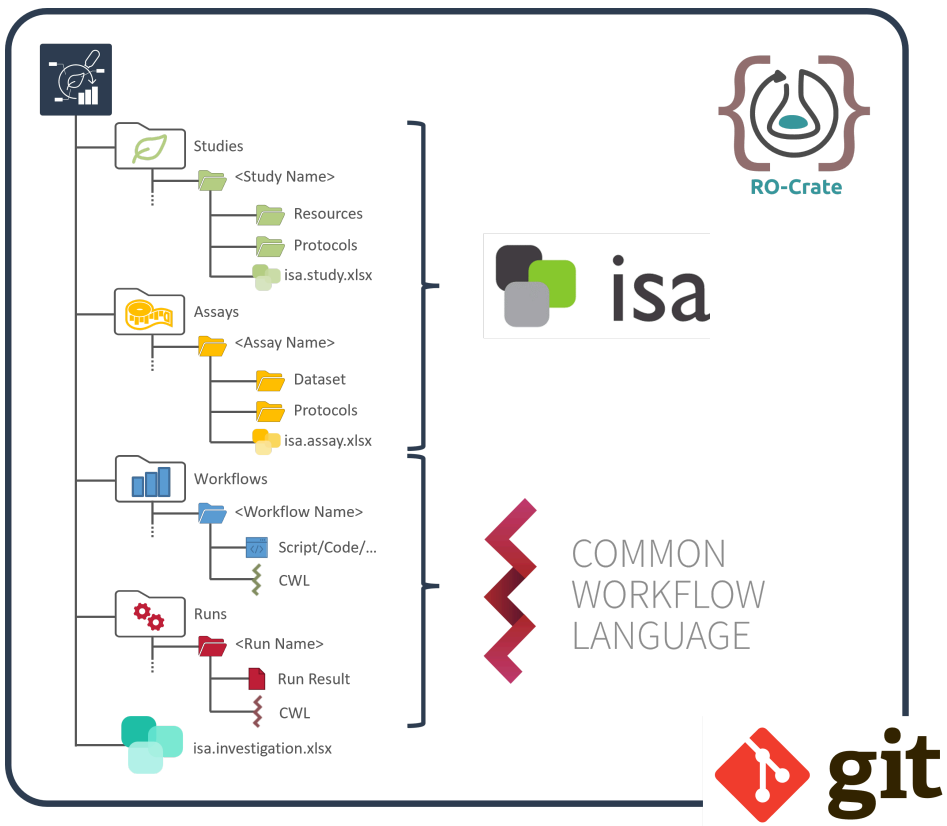
FAIR
digital (research) objects



The ARC scaffold structure



ARC builds on standards



ISA

- Structured, machine-readable experimental metadata
- https://isa-tools.org/

CWL

- Reproducible, re-usable data analysis
- https://www.commonwl.org/

Git

- Version control
- https://git-scm.com

RO-Crate

- Standardized exchange
- https://www.researchobject.org/ro-crate/

ISA abstract model in a nutshell

isa

Investigation
administrative (meta)data

- Summary
  - Titel
  - Description
- Person
- Organisation
- Publication reference

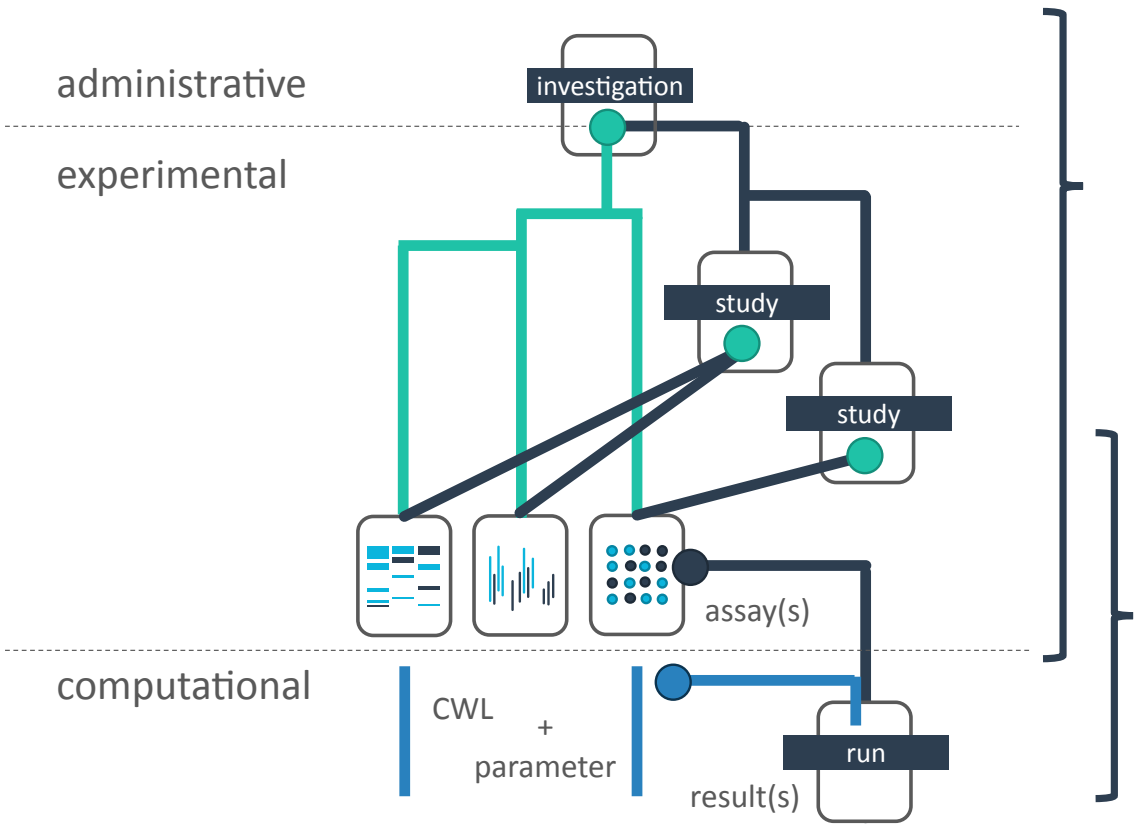
Study
descriptive (meta)data
information on the subject

- Characteristics
- Parameters
- Components
- Factors

Assay
descriptive (meta)data
information on the measurement

- Characteristics
- Parameters
- Components
- Factors

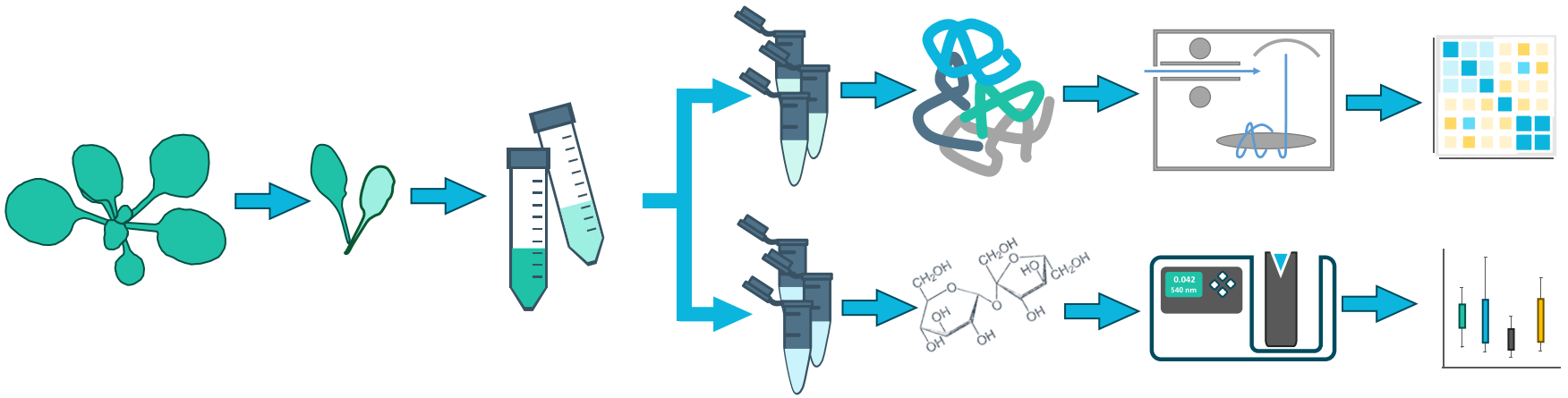
ISA and CWL – Connected by similarity



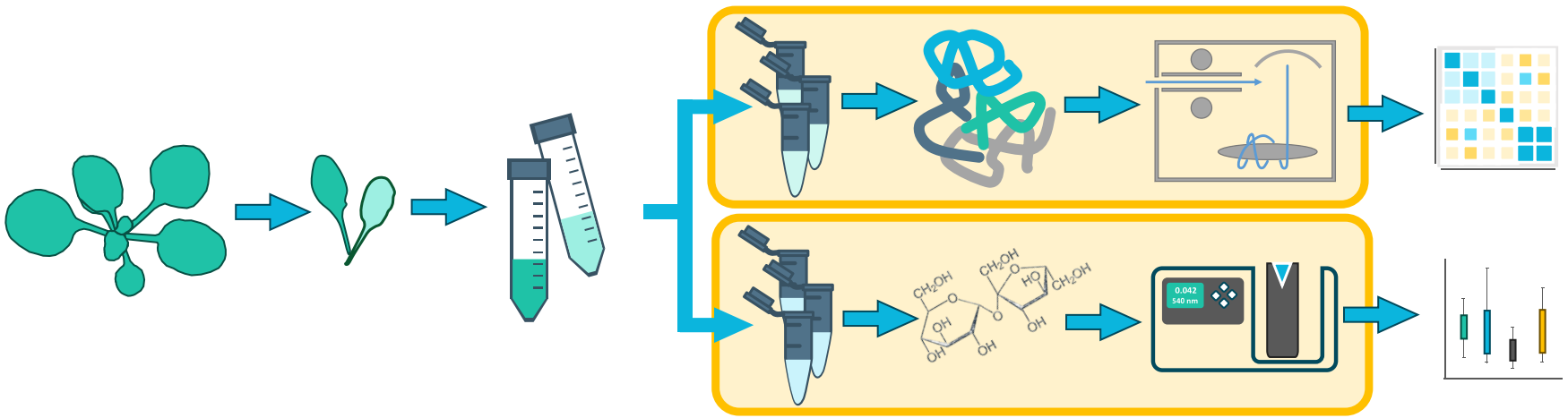
isa

COMMON
WORKFLOW
LANGUAGE

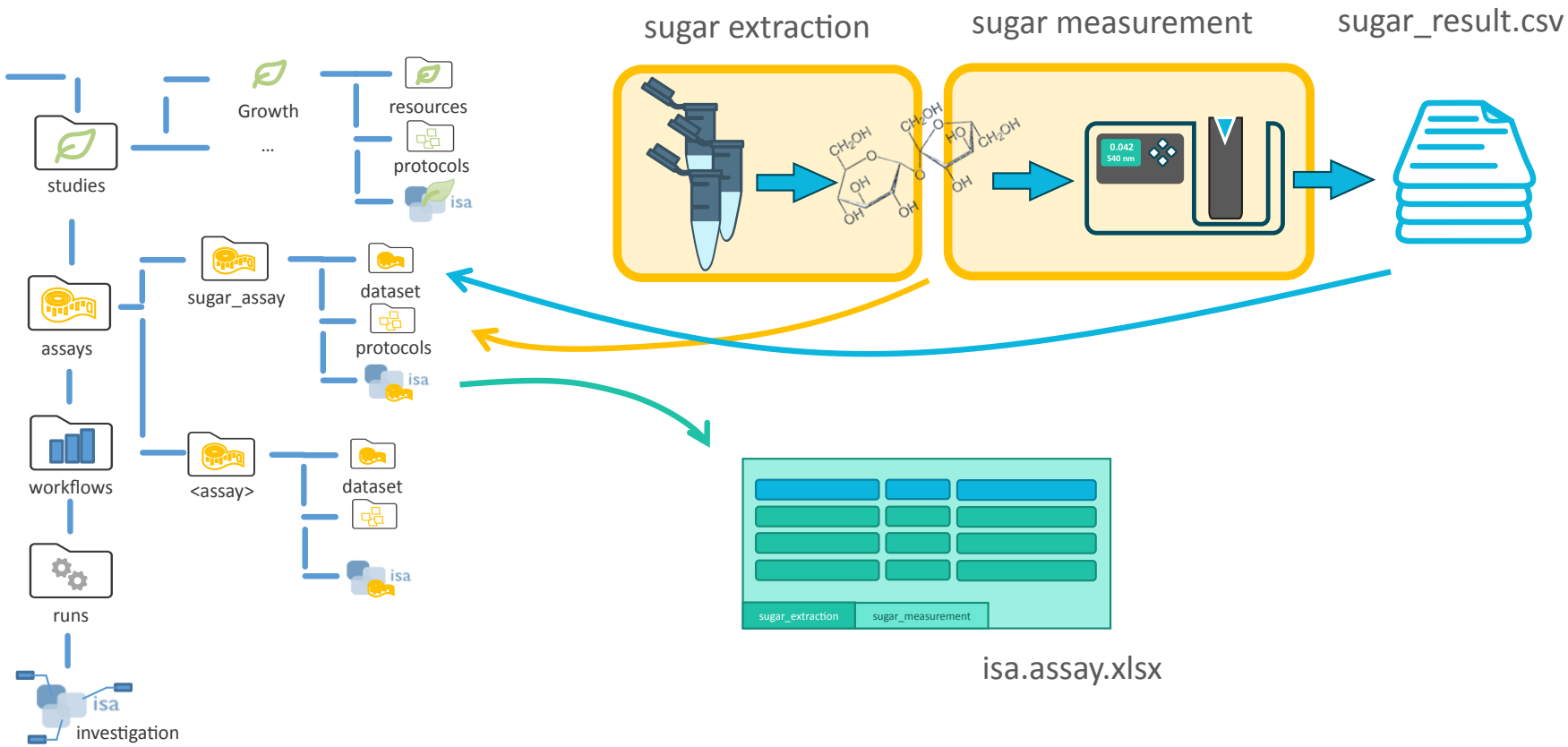
Metadata annotation – from sample to data



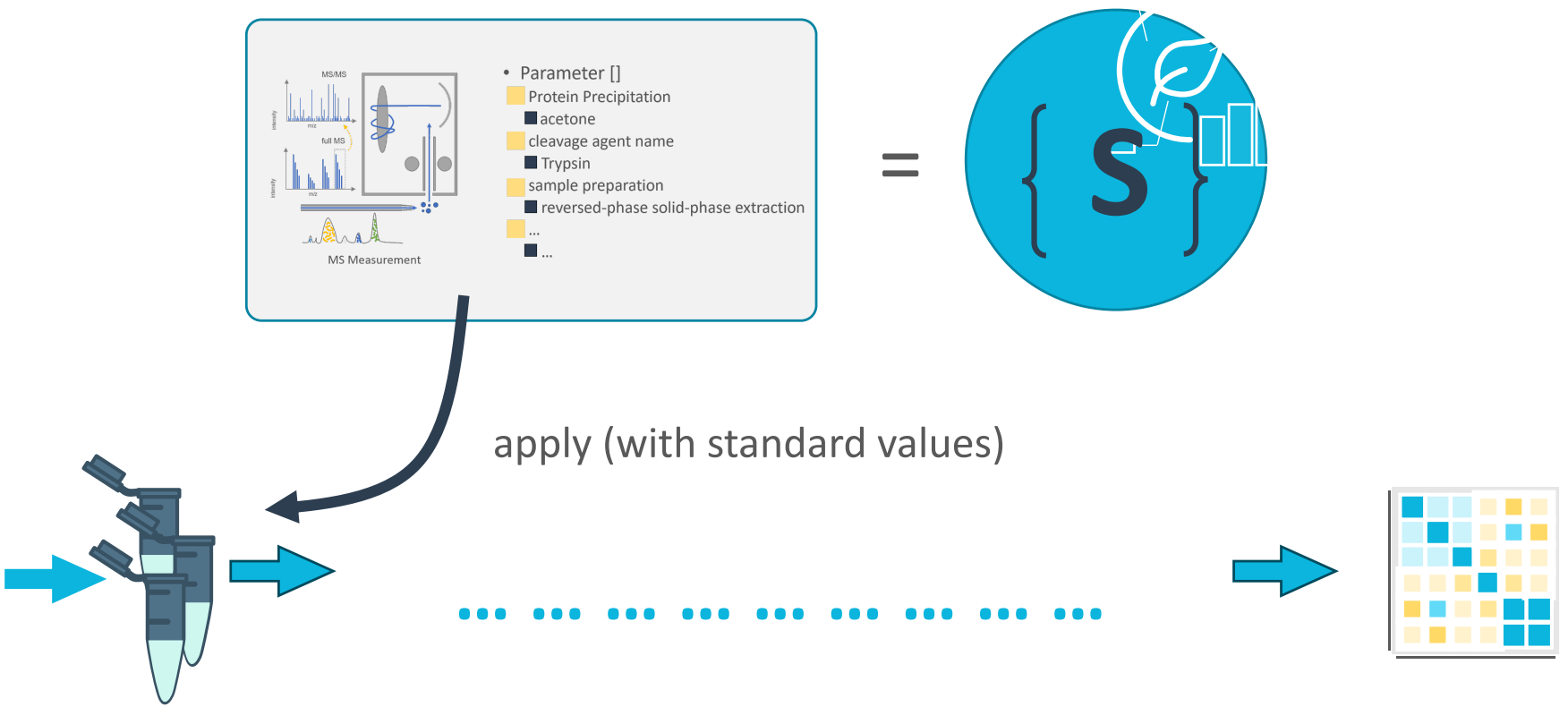
Modular separation of experimental processes



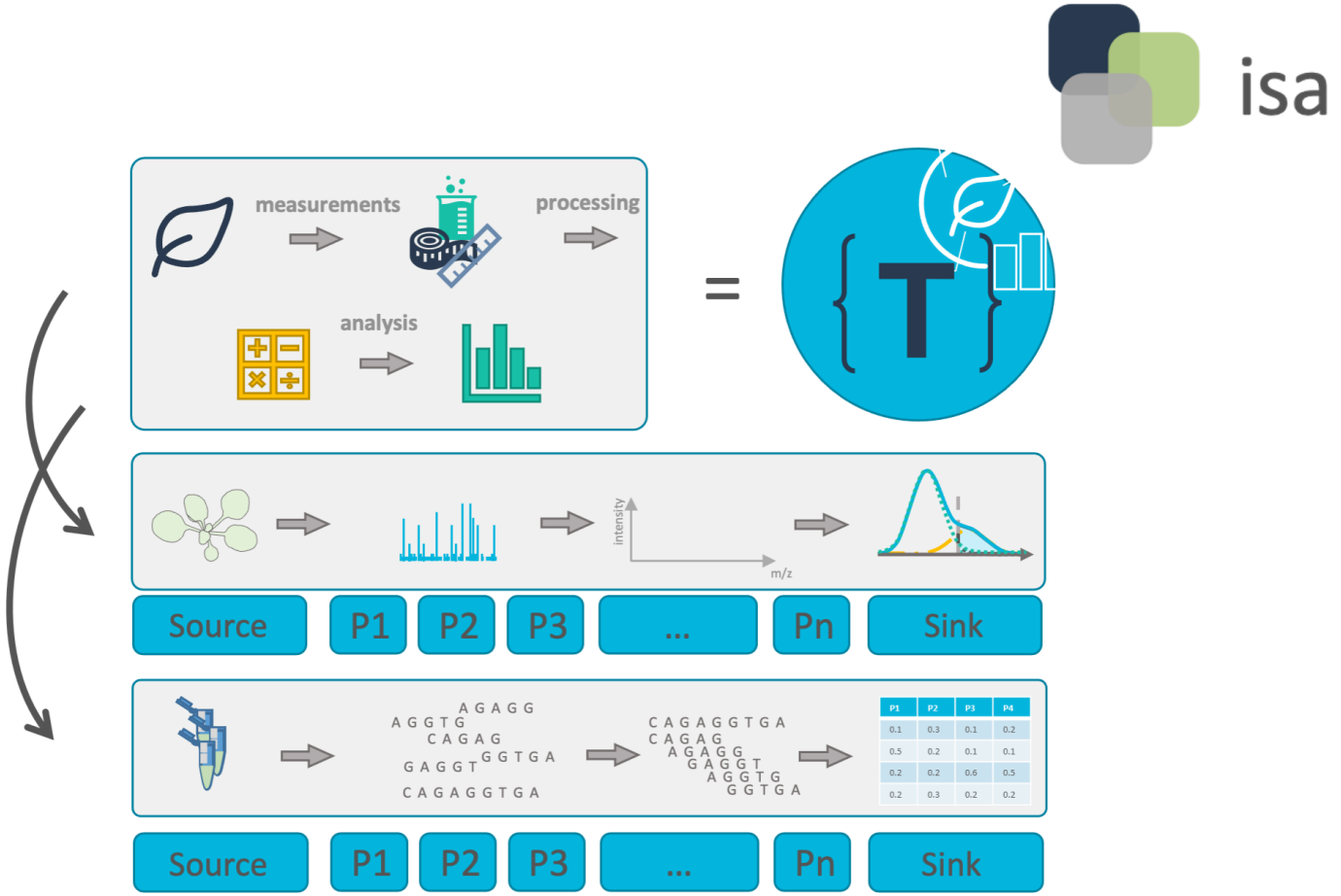
Modular separation of experimental processes



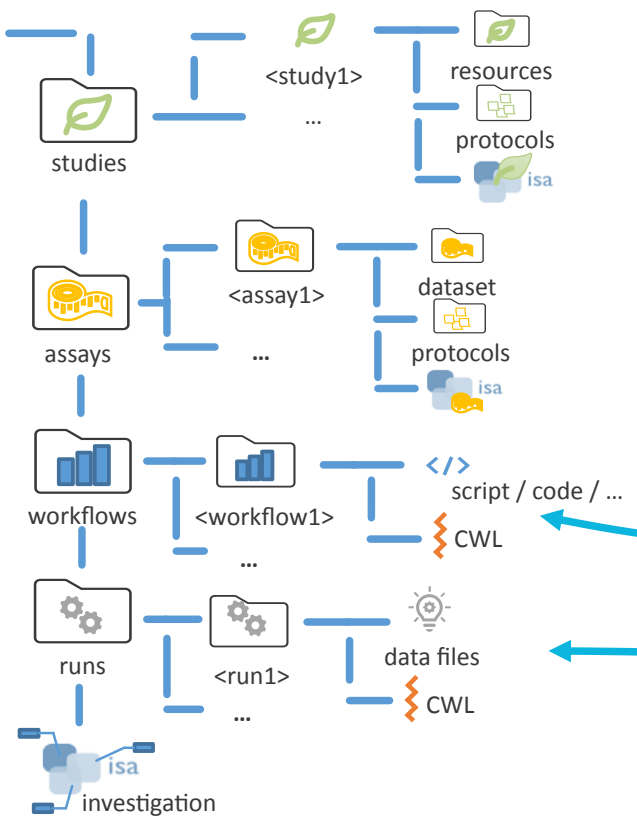
Applying standard procedures to sample record



Realization of lab-specific metadata with templates

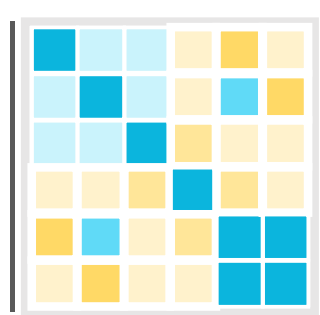


Data analysis: CWL workflows and runs




```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()
```

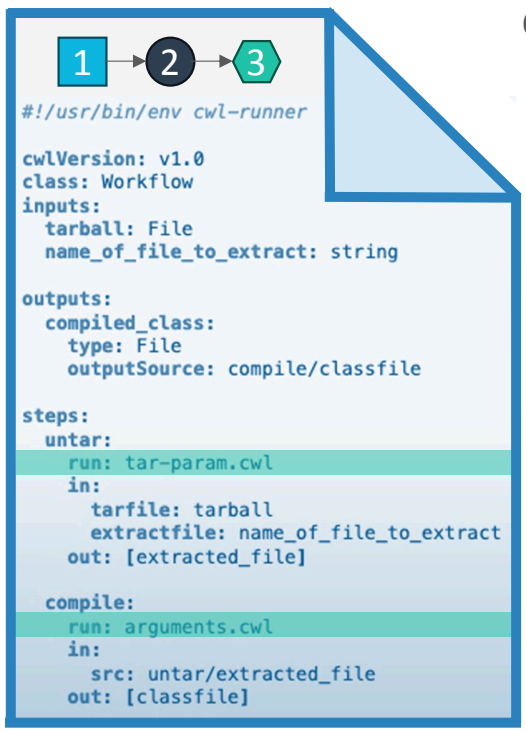
heatmap.py



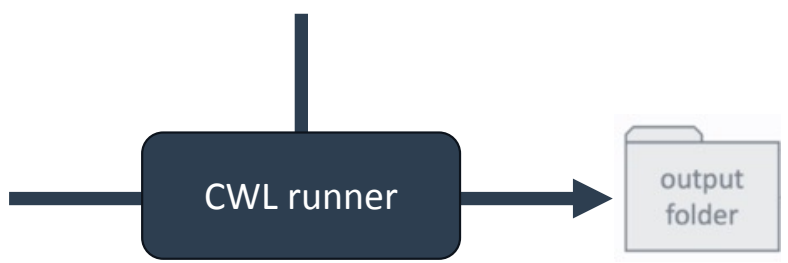
Metadata annotation – from data to result

CWL workflow document (*.cwl)

CWL
tool descriptors (*.cwl)

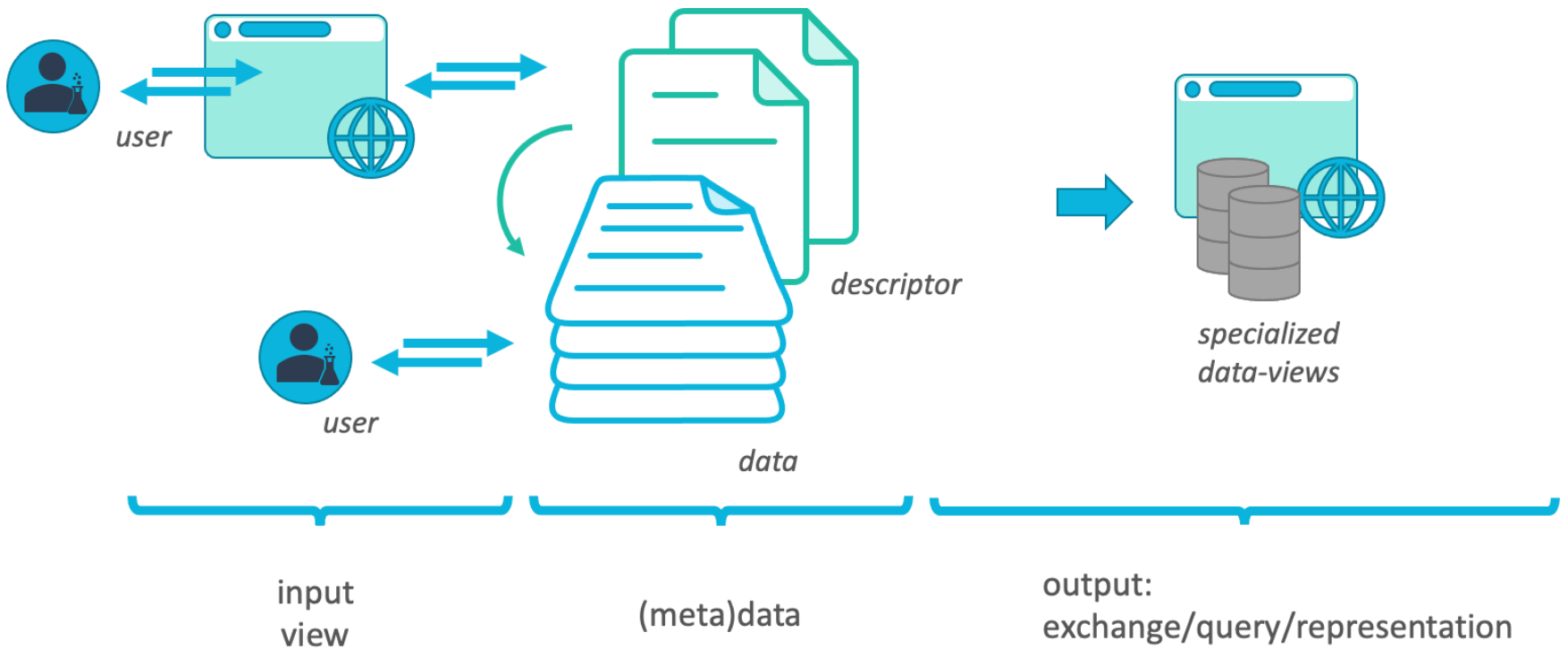


CWL job parameter (*.yaml)

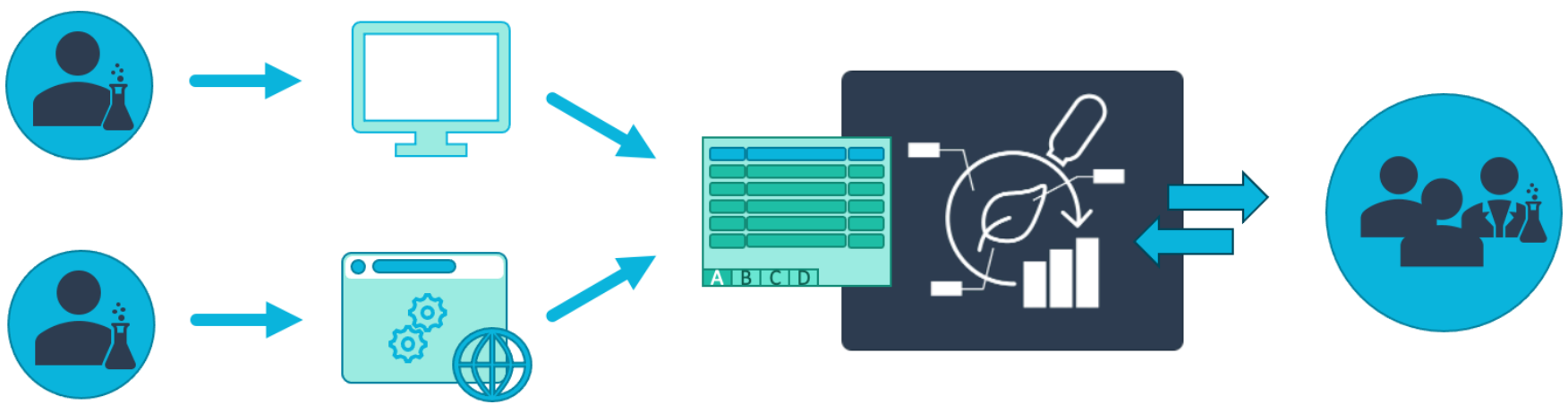


Everything is a file

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

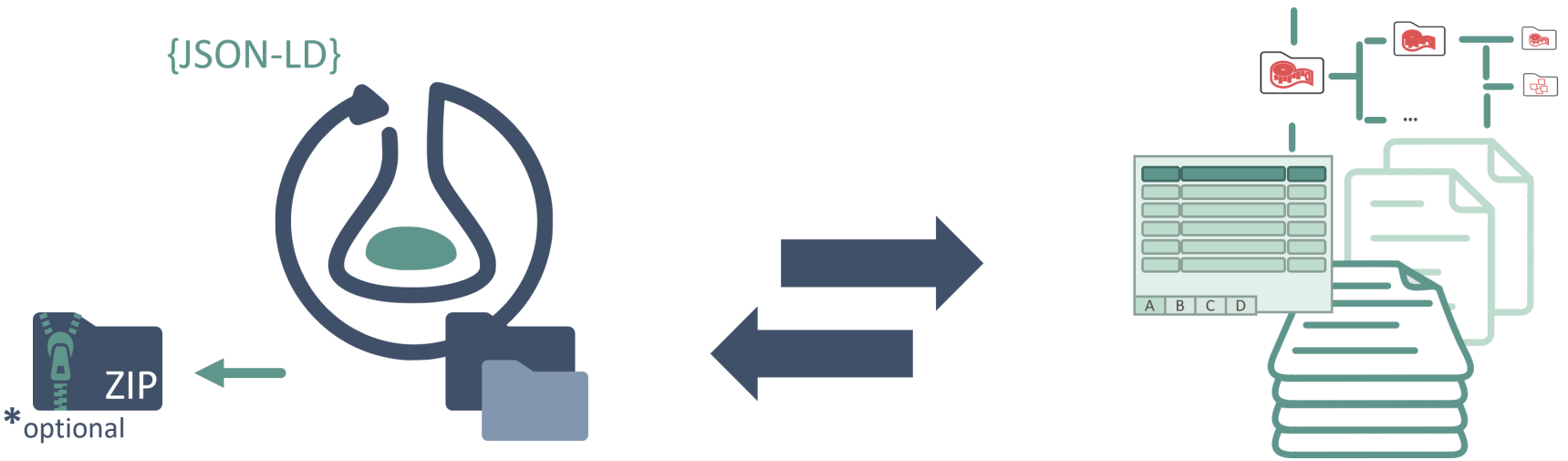


No technical lock-in



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

Two representations of the ARC



ARC RO-Crate

ARC Scaffold

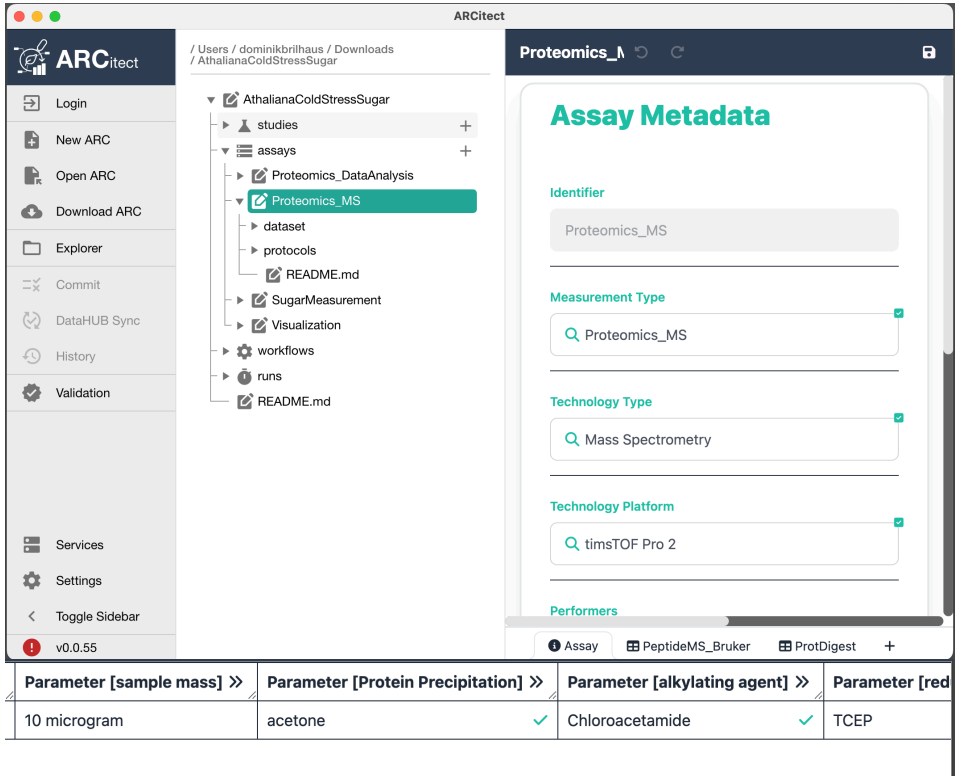
Two sides of the same coin

"Developer View": RO-Crate

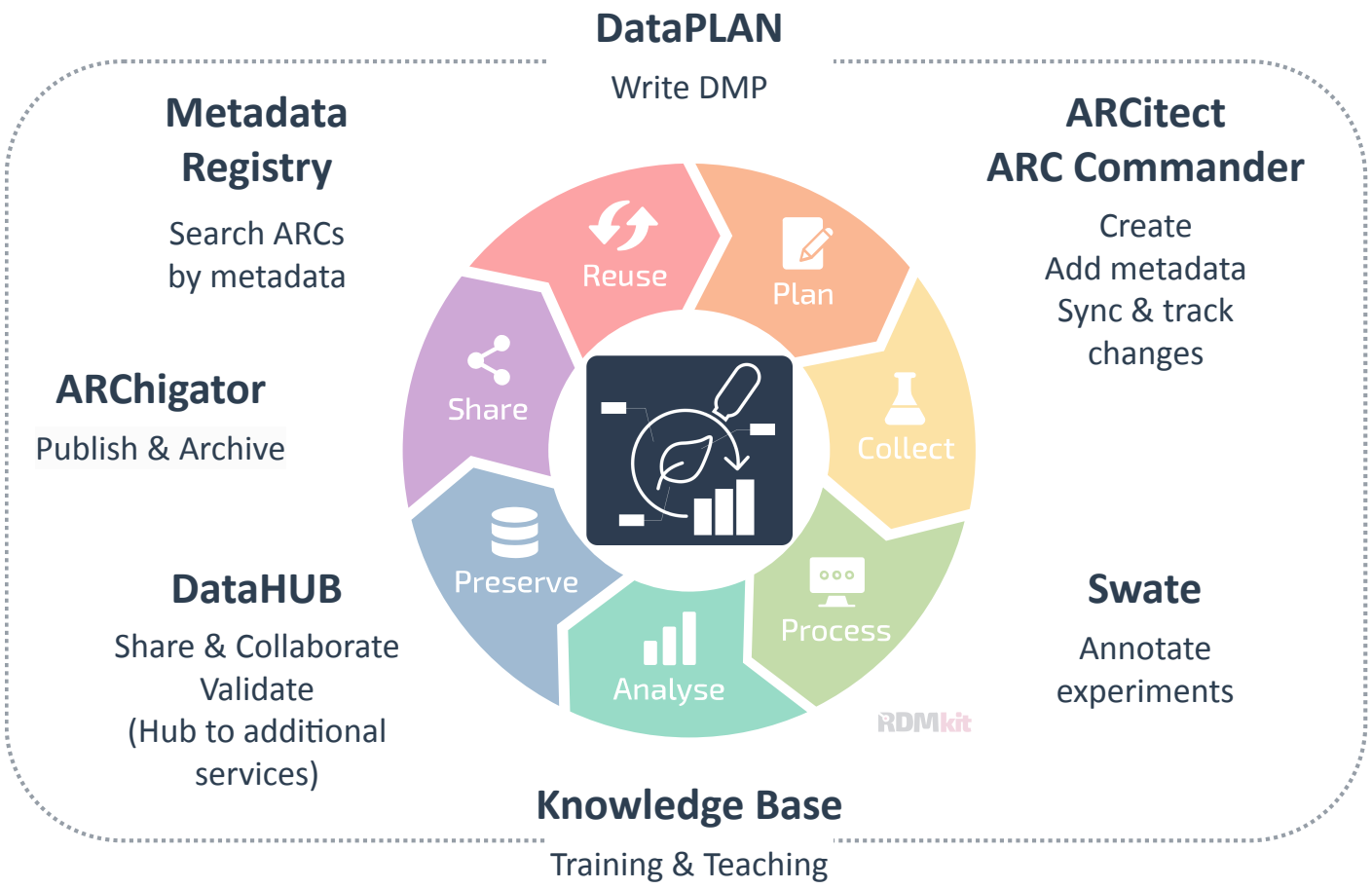
```
{
"Identifier": "Proteomics_MS",
"MeasurementType": {
  "annotationValue": "Proteomics_MS",
  "termSource": "MS",
  "termAccession": "https://purl.obolibrary.org/obo/FMS_1003348"
},
"TechnologyType": {
  "annotationValue": "Mass Spectrometry",
  "termSource": "NCIT",
  "termAccession": "https://bioregistry.io/NCIT:C17156"
},
```

```
...
"Tables": [
  {
    "name": "ProtDigest",
    "header": [
      {
        "headertype": "Parameter",
        "values": [
          {
            "annotationValue": "sample mass",
            "termSource": "MS",
            "termAccession": "MS:1000004"
```

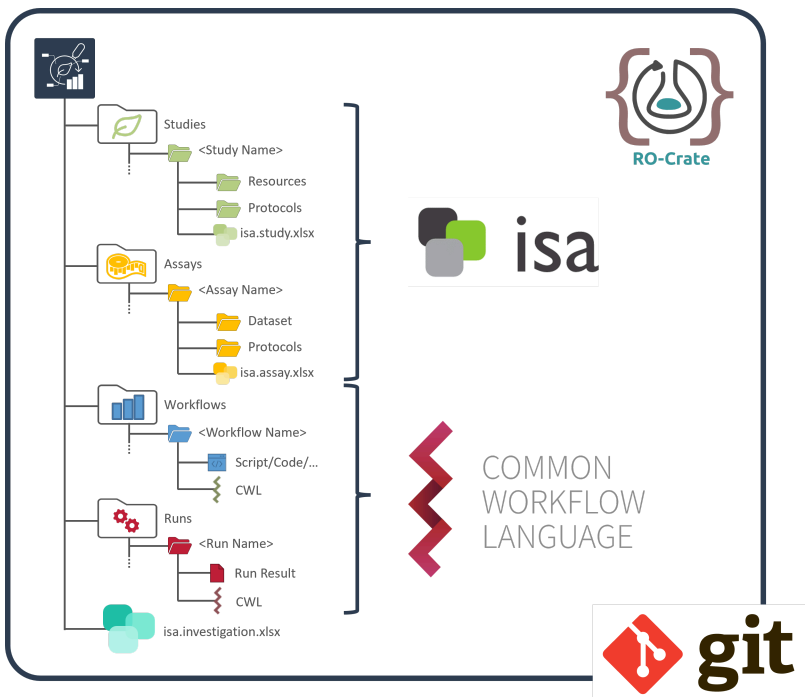
"User View": ARC Scaffold and metadata tables



The ARC ecosystem



ARC builds on standards



RO-Crate

- Standardized exchange
- https://www.researchobject.org/ro-crate/

ISA

- Structured, machine-readable metadata
- https://isa-tools.org/

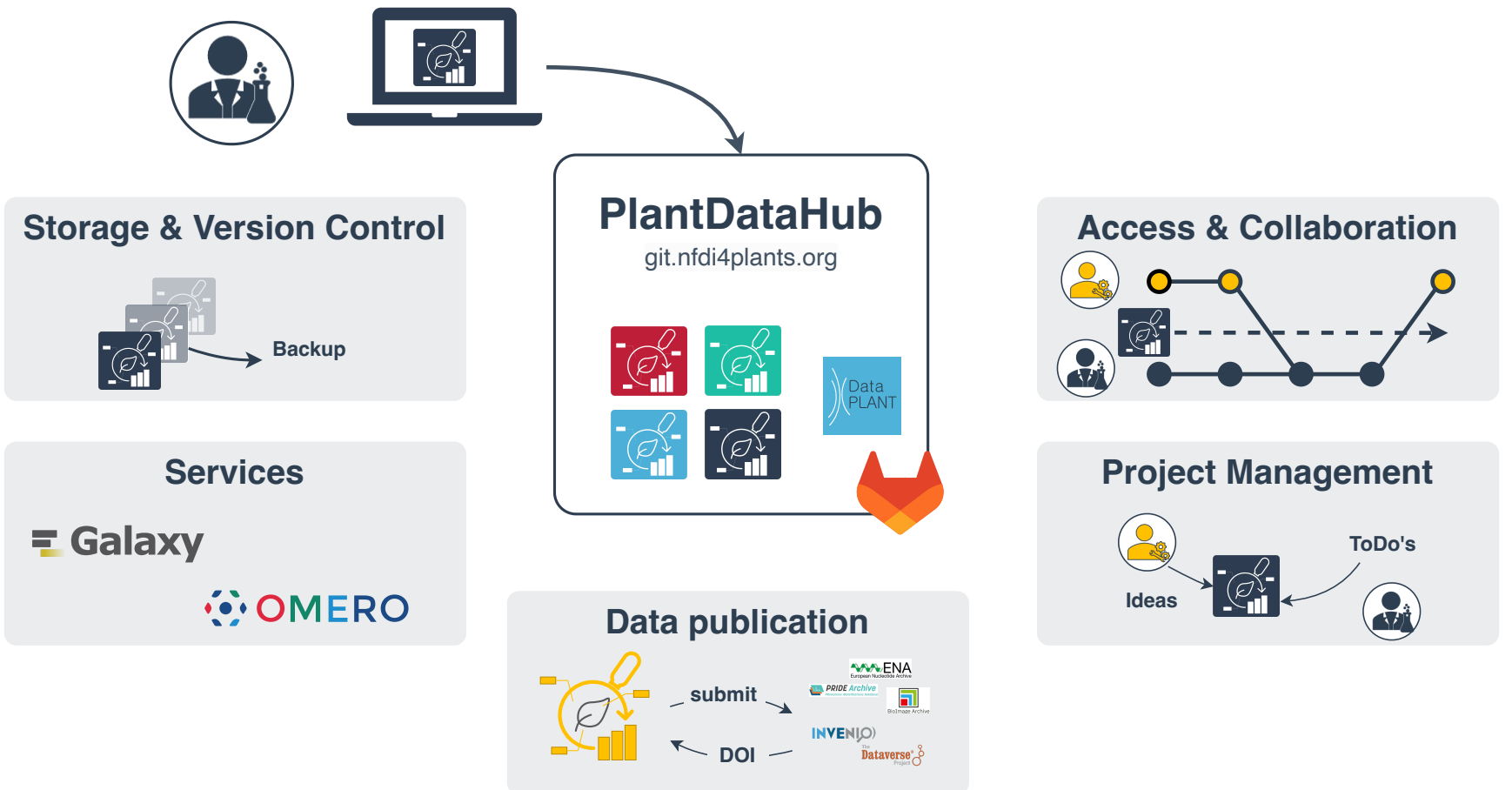
CWL

- Reproducible, re-usable data analysis
- https://www.commonwl.org/

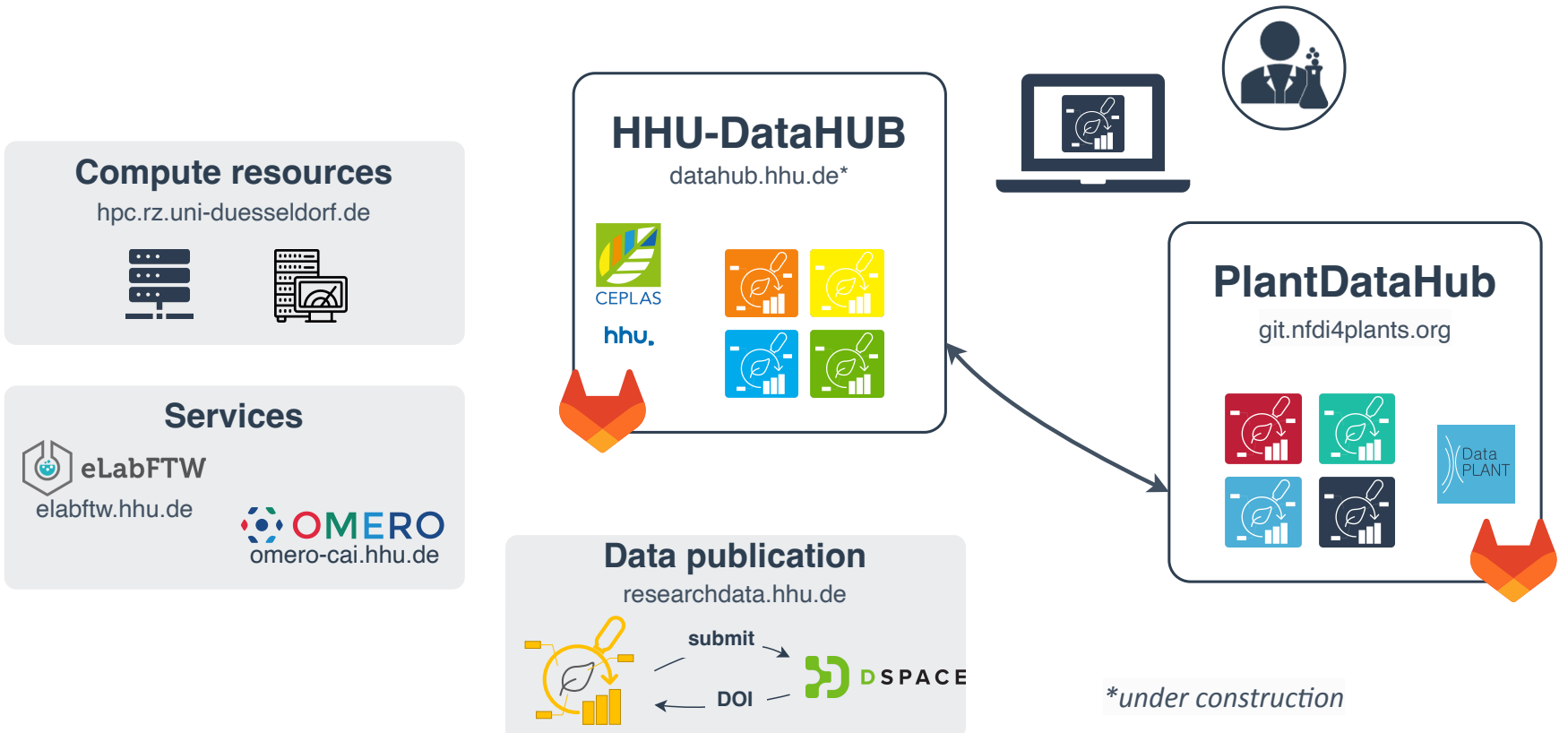
Git

- Version control

ARC and DataHUB as entry point



HHU-DataHUB – On-premise DataHUB node



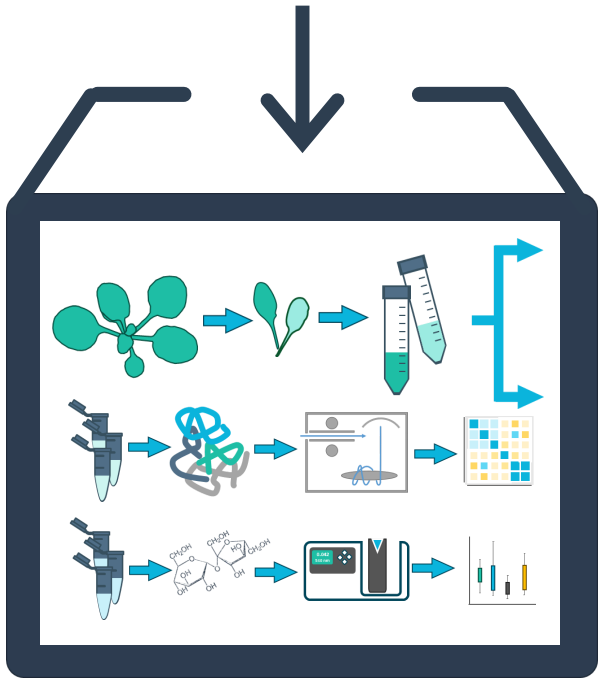
Hands-on part 1: Setup and ARCitect

Follow the Start Here guide in the DataPLANT knowledge base.

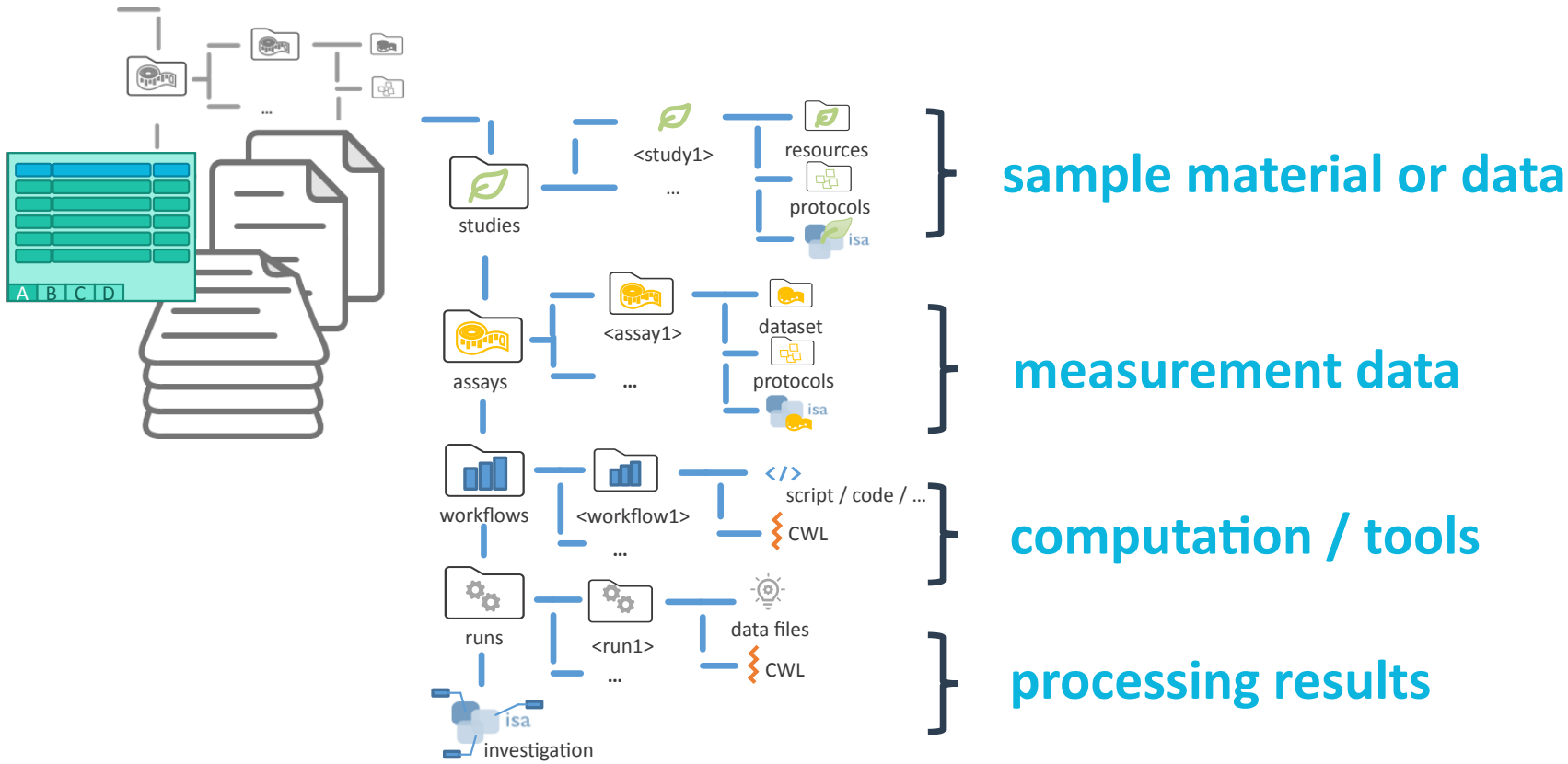
✏️ Until step **Add a study**

ARC: Annotated research context

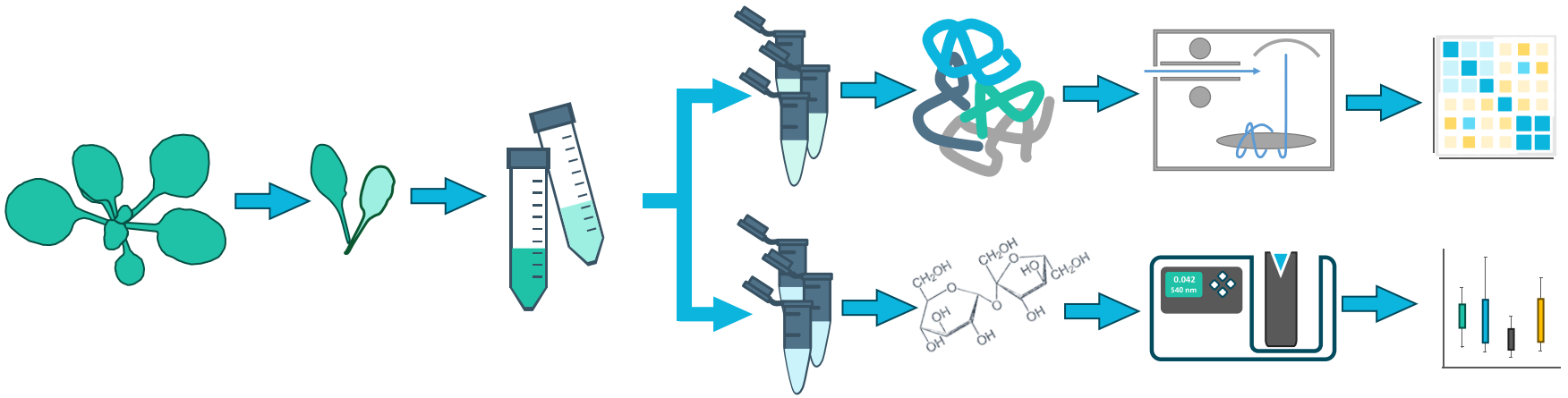
FAIR
digital (research) objects



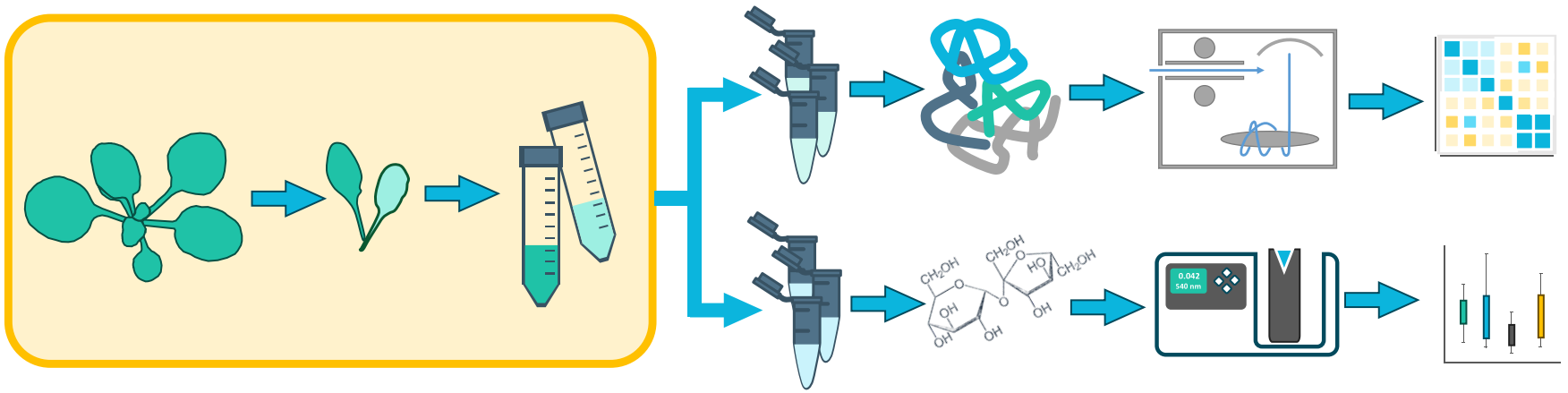
The ARC scaffold structure



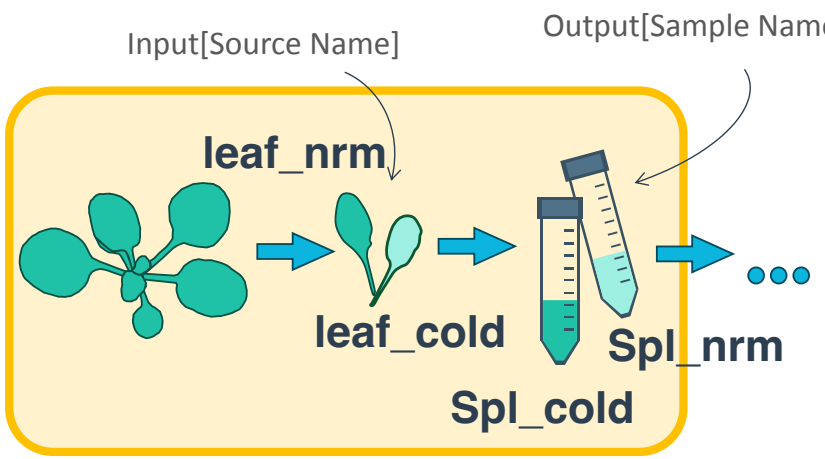
A small prototypic project



Divide and conquer for reproducibility

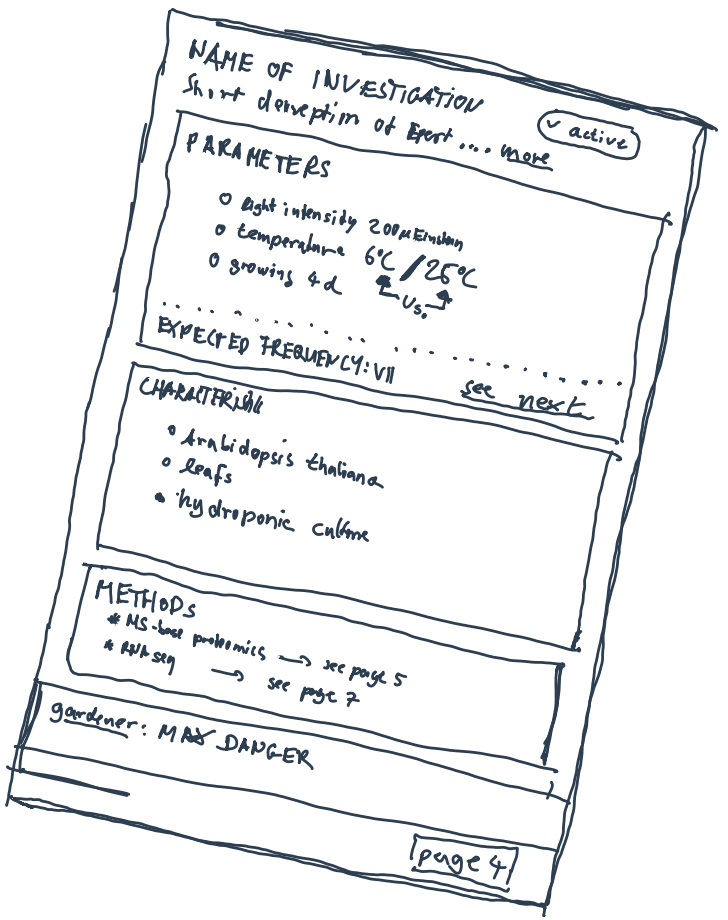


Identifying the ‘study’ part

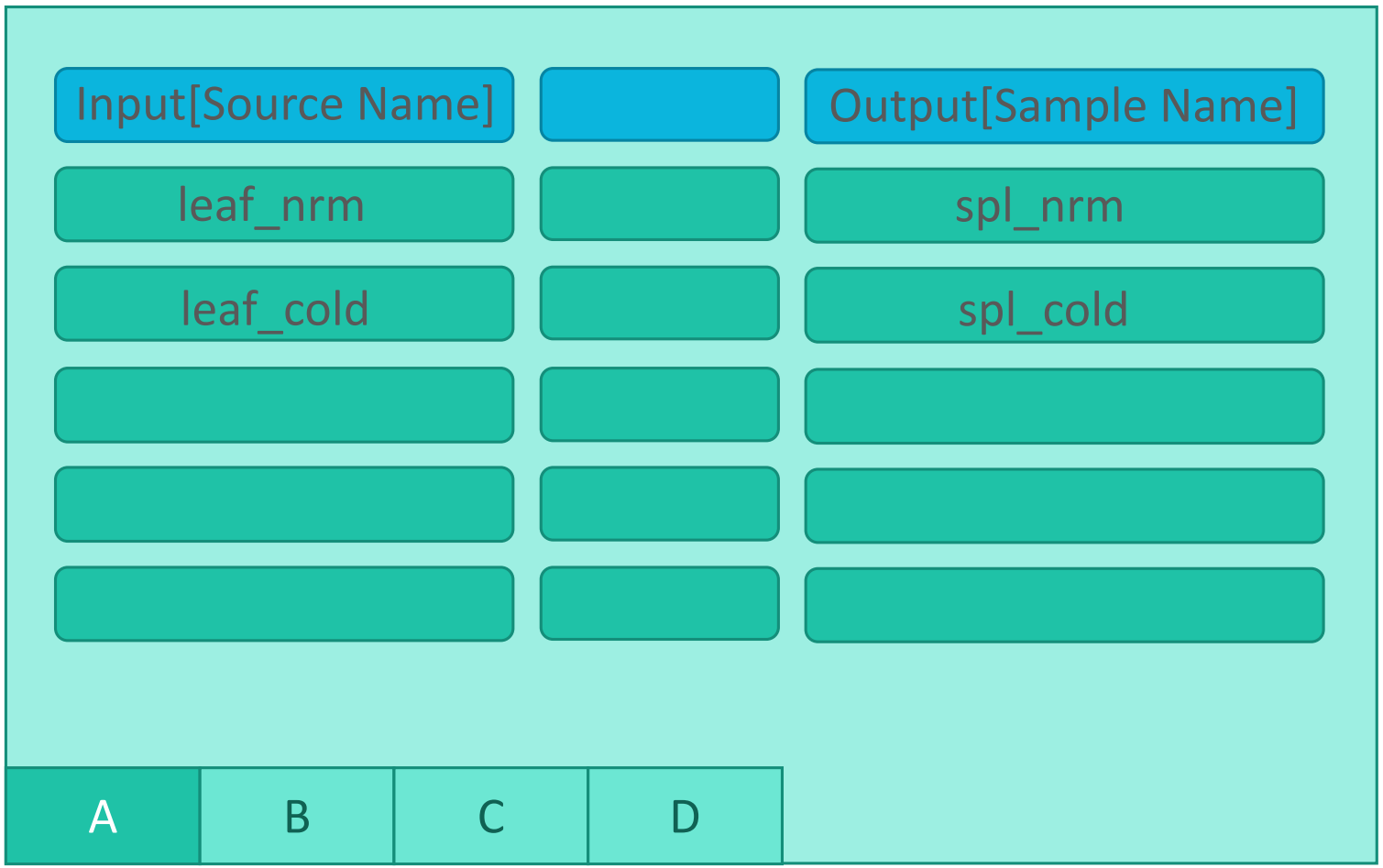


Protocol REF

growth_protocol.txt



A table-based organization schema

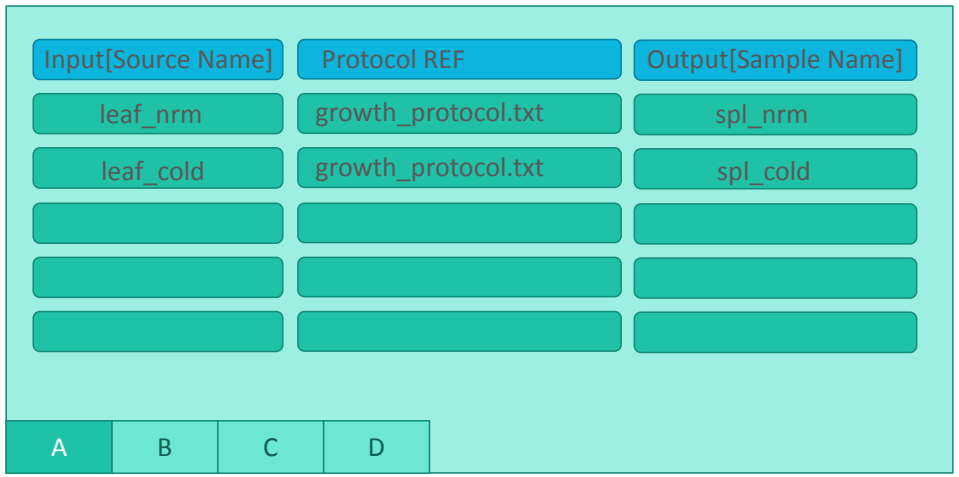


Referencing a protocol

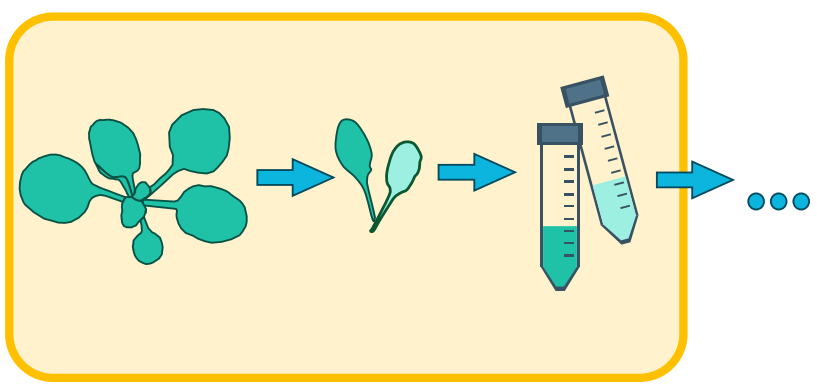
This allows you to reference the free-text, human-readable protocol.

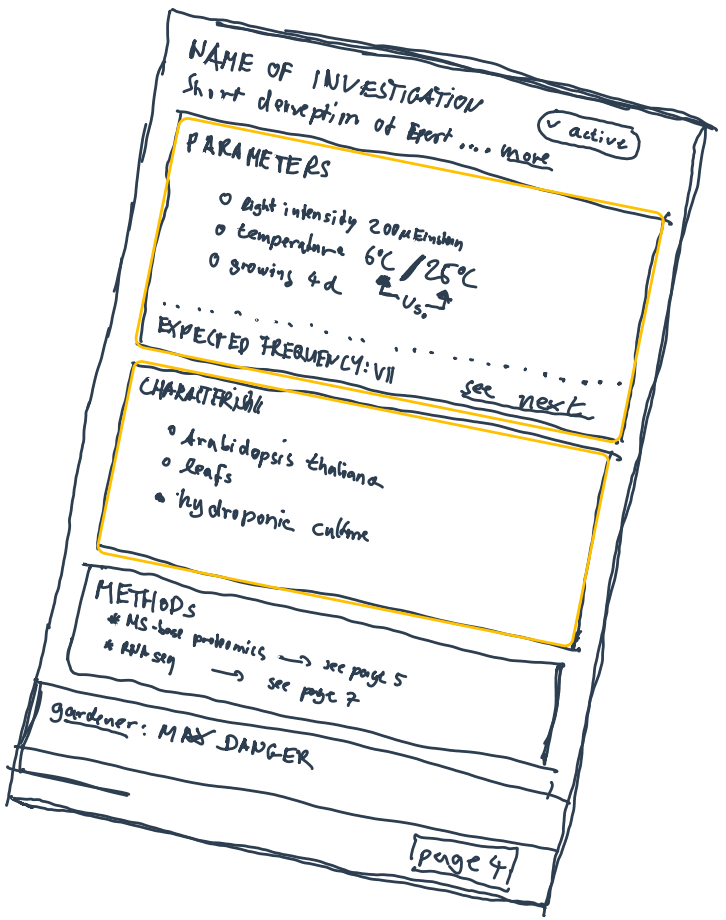
💡 It is recommended that the protocol is in an open format (.md|.txt|.docx|…)

💡 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`

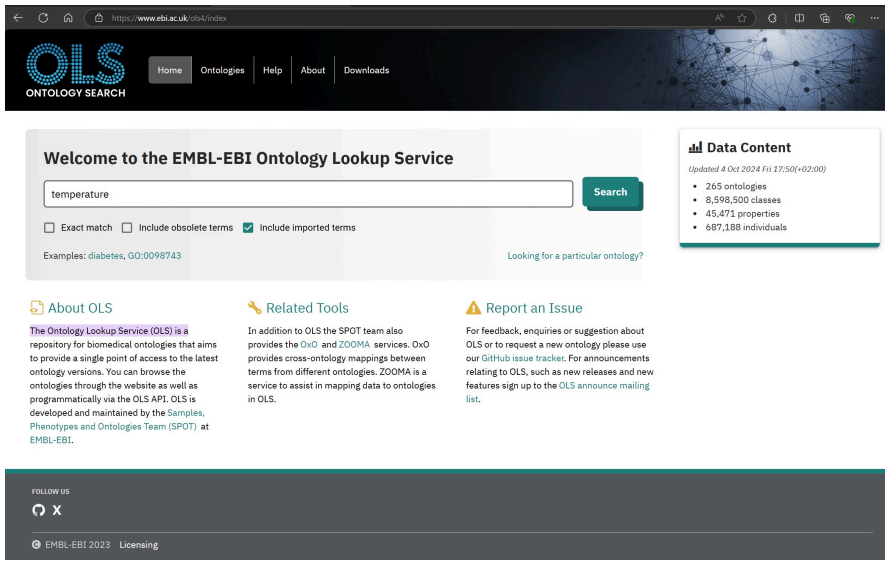
temperature `PATO:0000146`

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

A physical quality of the thermal energy of a system.

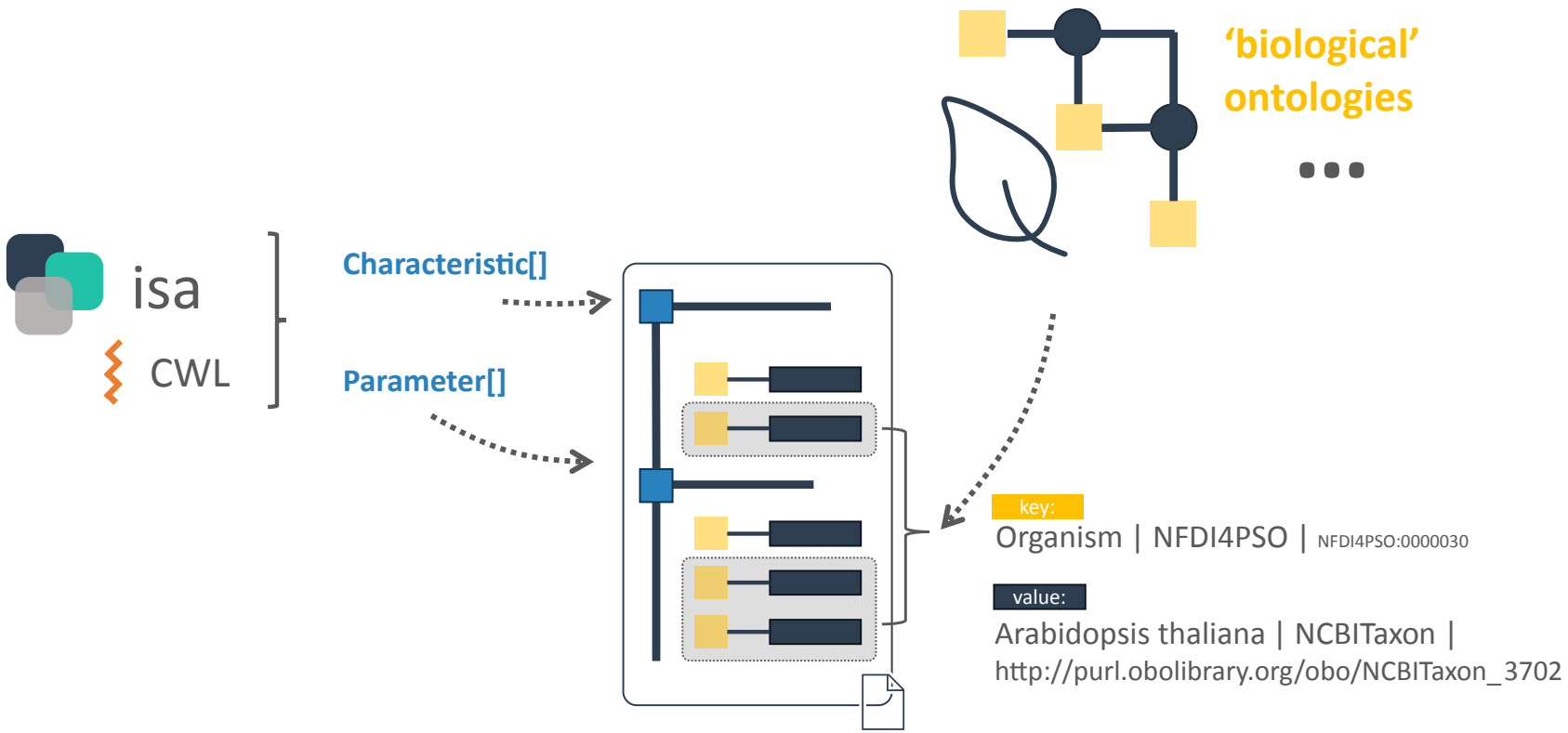
Ontology: `PATO`

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



Ontology Lookup Sevice (OLS): https://www.ebi.ac.uk/ols4/

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

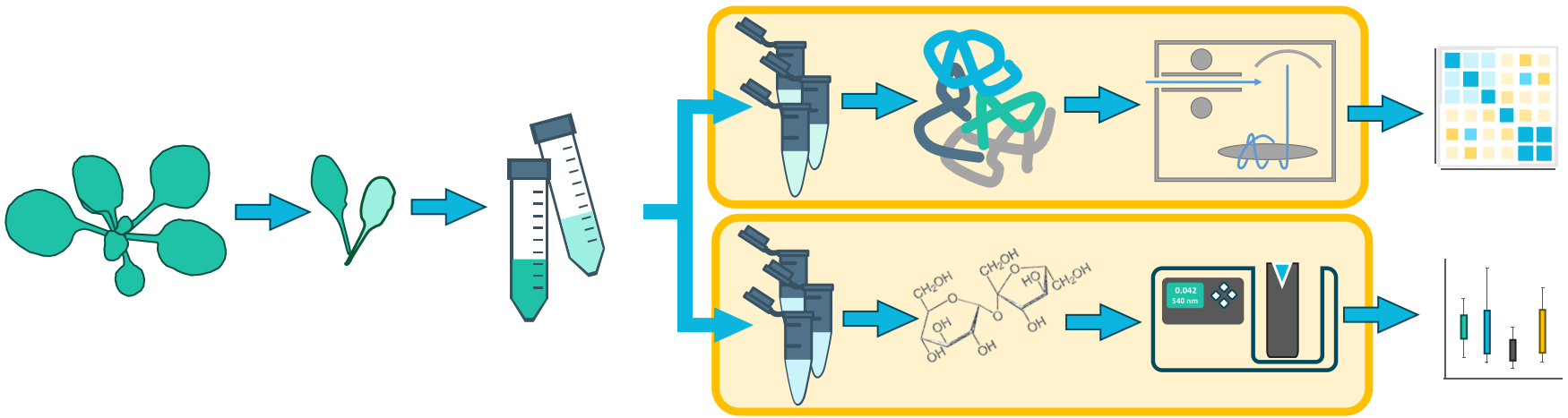
Hands-on part 1: Setup and ARCitect

Follow the Start Here guide in the DataPLANT knowledge base.

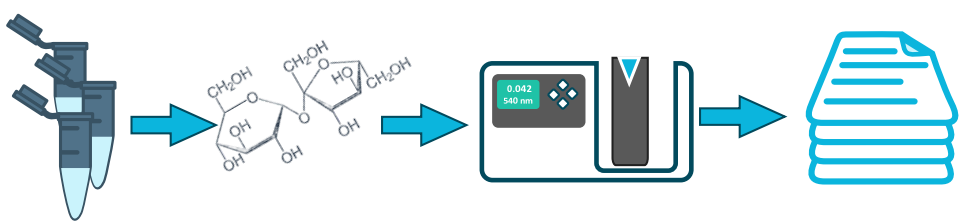
✏️ Until step **Add a study**

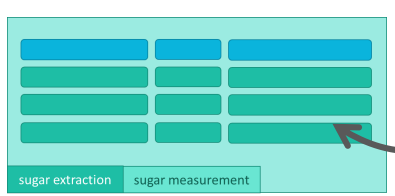
Hands-on part 2: ARCitect (and Swate)

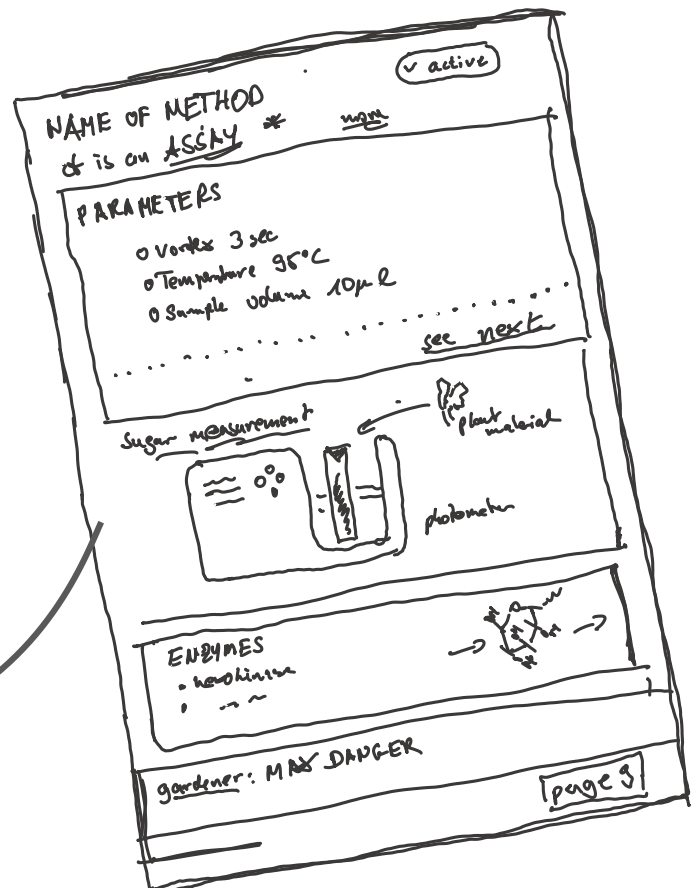
Identifying assays



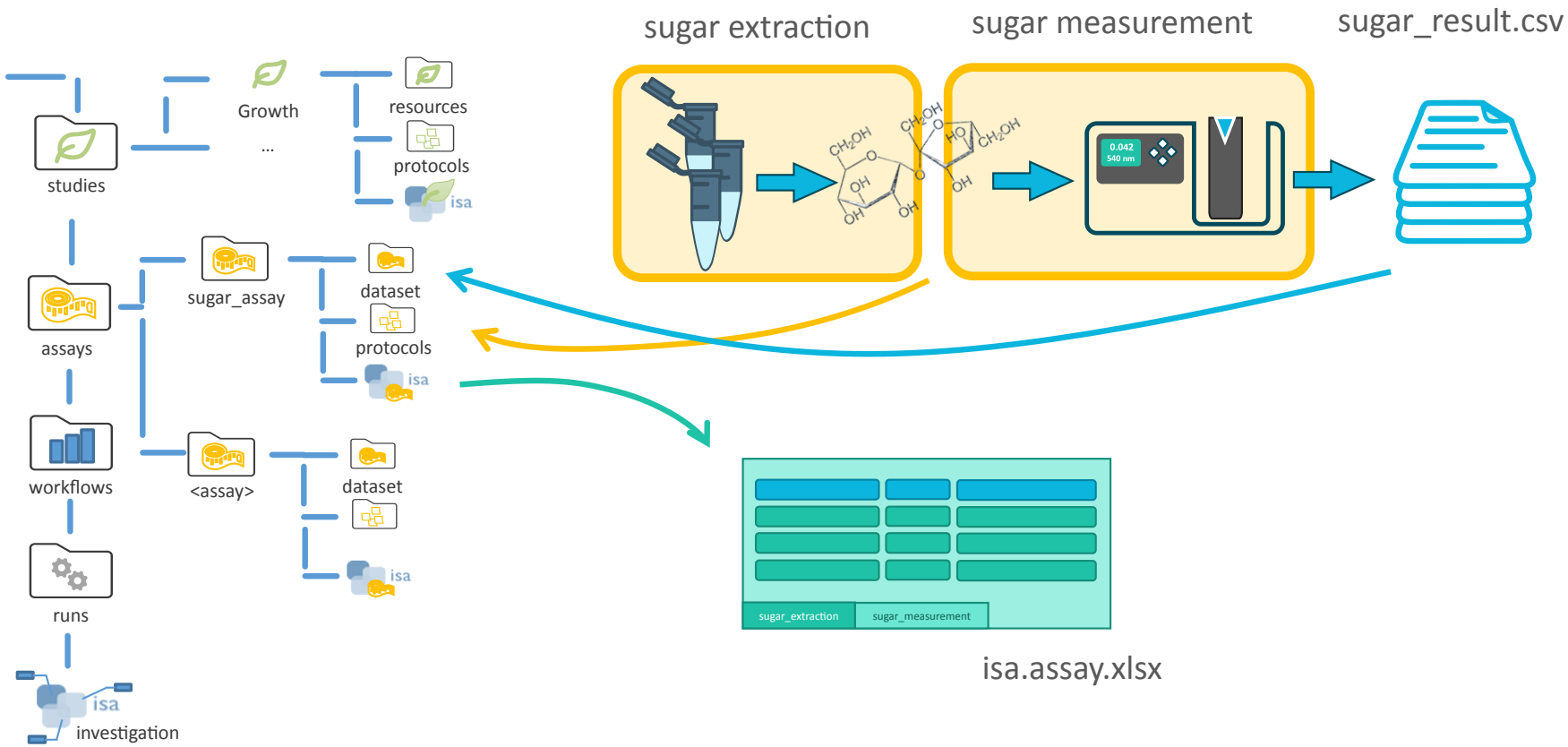
Assay for sugar measurement



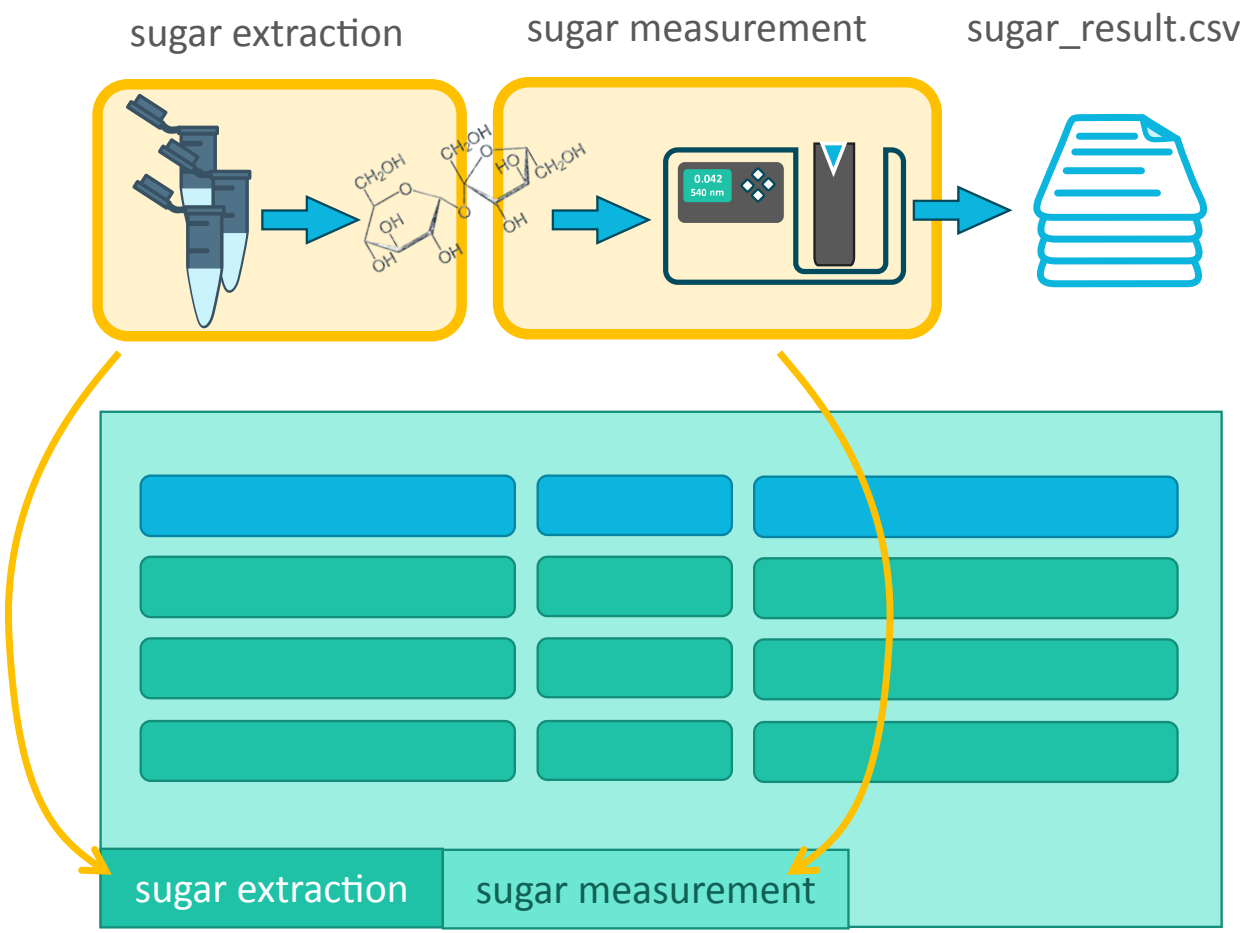




Separating different assay elements



Isolating the lab processes in an assay

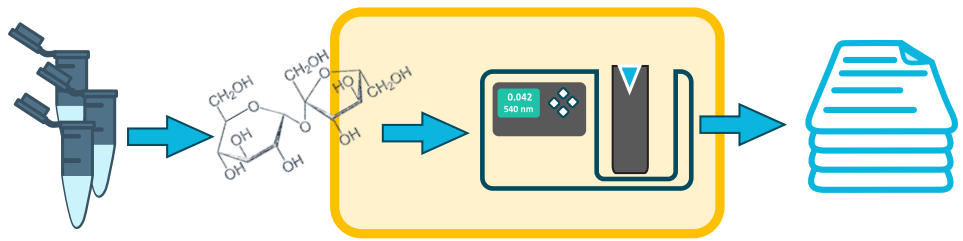


Parameteterization: sugar extraction

- Vortex Mixer
  - 3 seconds
- Temperature
  - 95 degree celsius

Parameteterization: sugar measurement

- technical replicate
  - 1,2,3,⋯
- sample volume
  - 10 `microliter`
- buffer volume
  - 190 `microliter`
- cycle count
  - 5



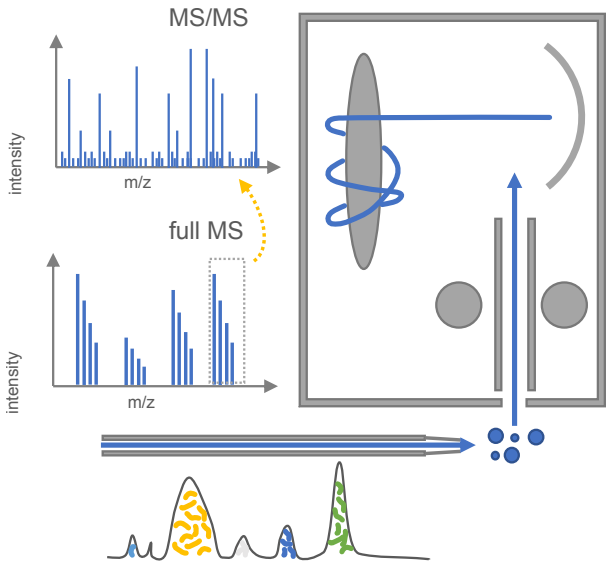
Save time using standard methods and SOPs

Parameter []

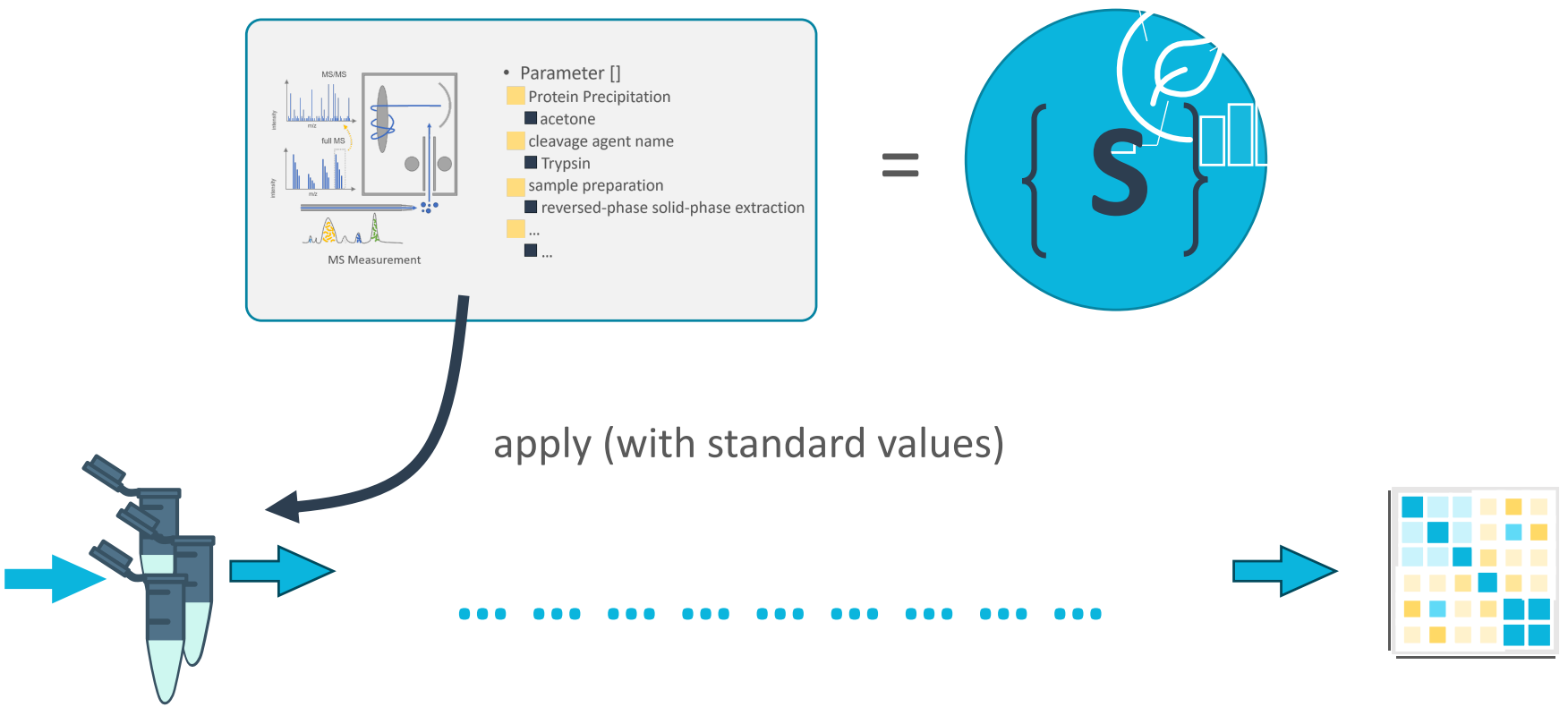
- Protein Precipitation
  - acetone
- cleavage agent name
  - Trypsin
- sample preparation
  - reversed-phase solid-phase extraction
- ...

Component []

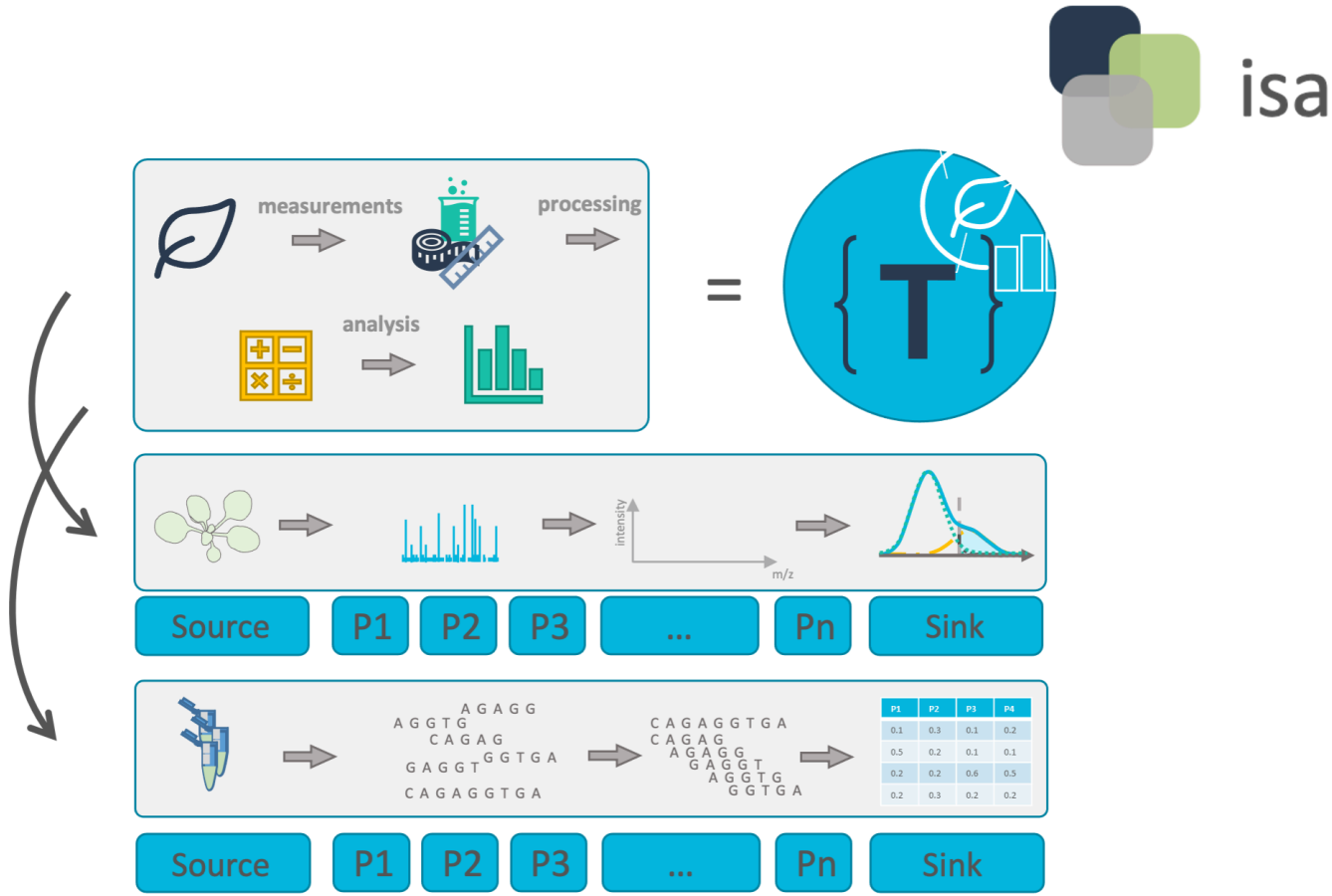
- chromatography instrument model
  - nanoElute2
- chromatography column model
  - PepSep C18 1.9u, 25cm x 75um ...



Applying standard procedures to sample record



Realization of lab-specific metadata with templates



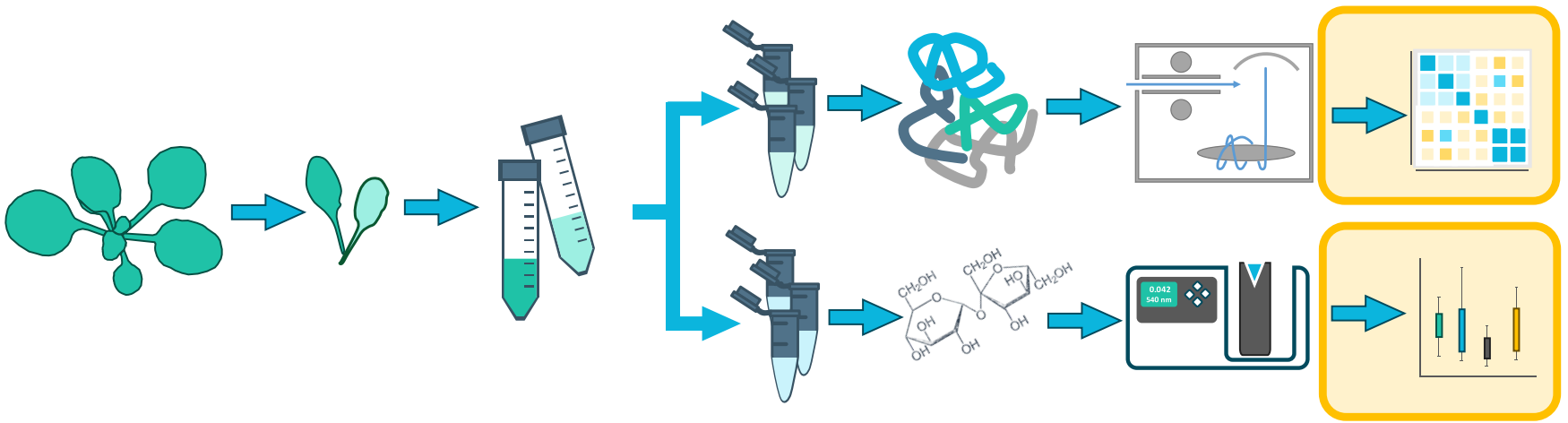
Hands-on part 2: ARCitect (and Swate)

Continue the Start Here guide in the DataPLANT knowledge base.

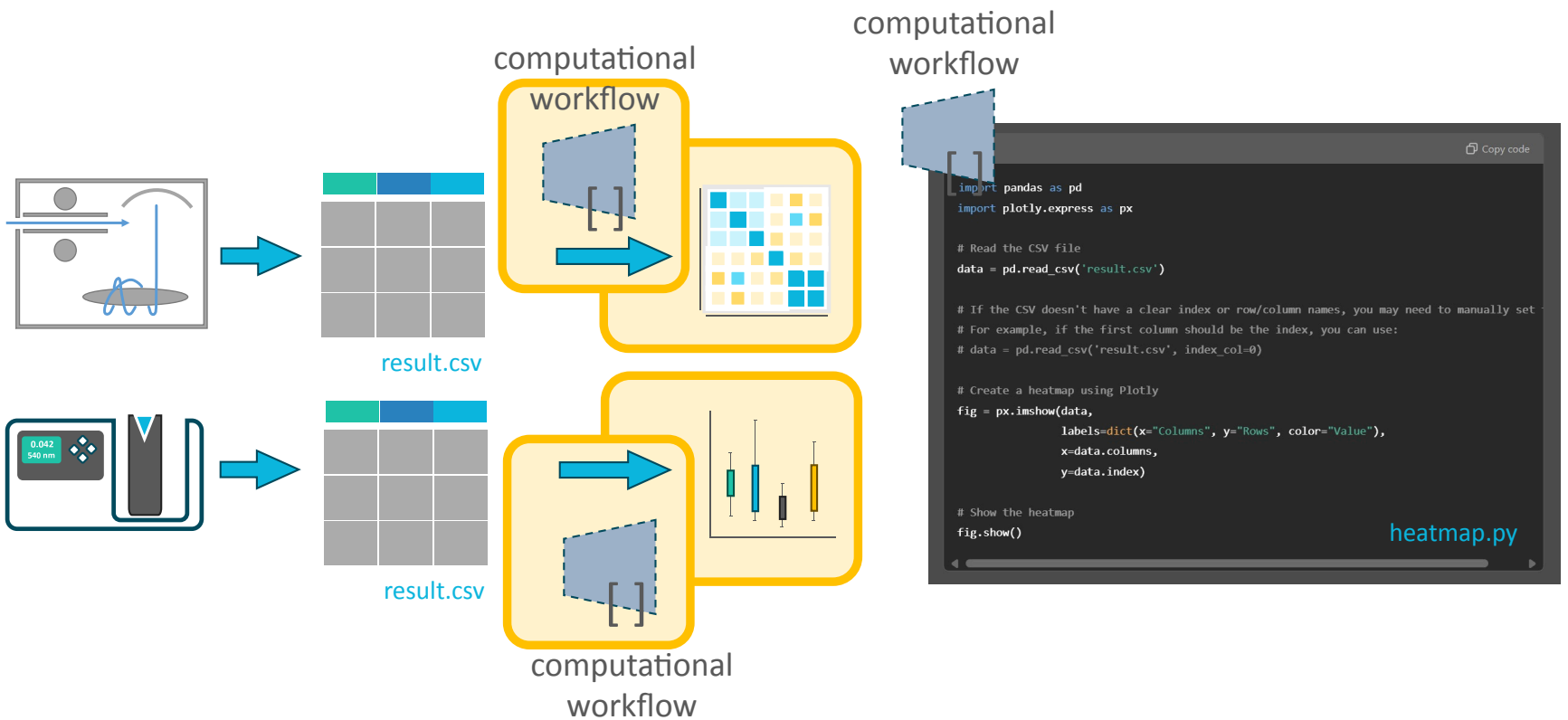
✏️ Until step **Add assay data**

Hands-on part 3: Data

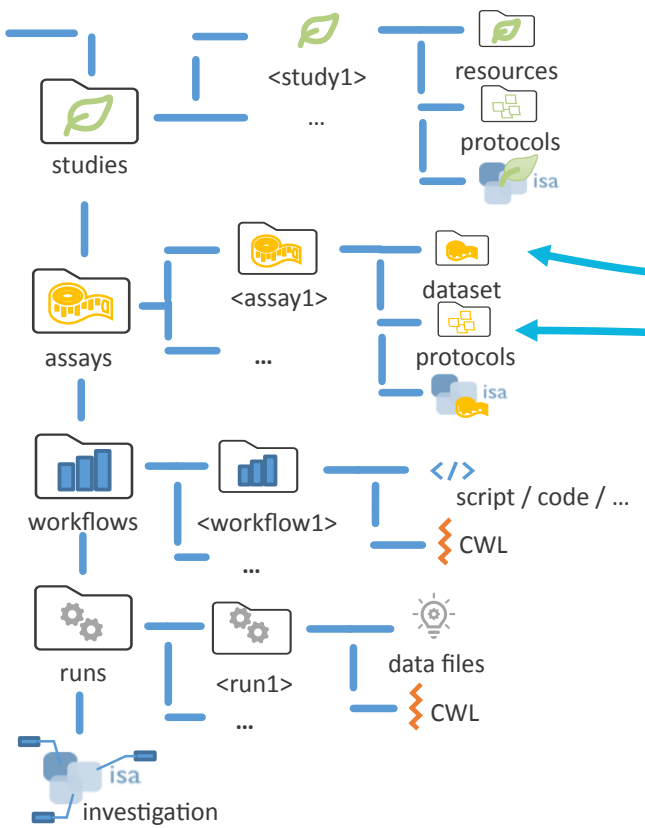
Options to annotate the data analysis



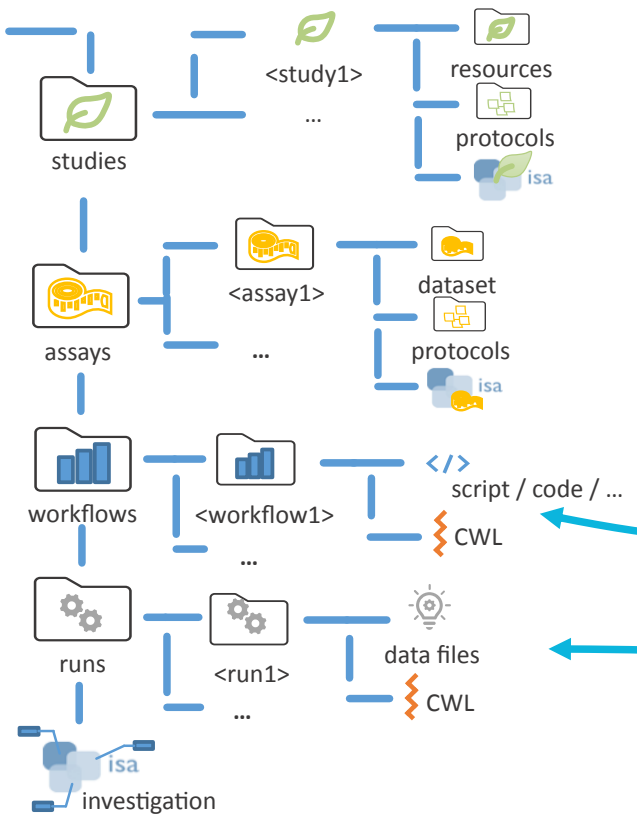
A computational workflow is like a protocol

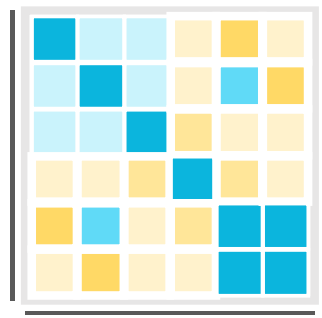


Option I: Create a virtual assay



Option II: Create a workflow and run



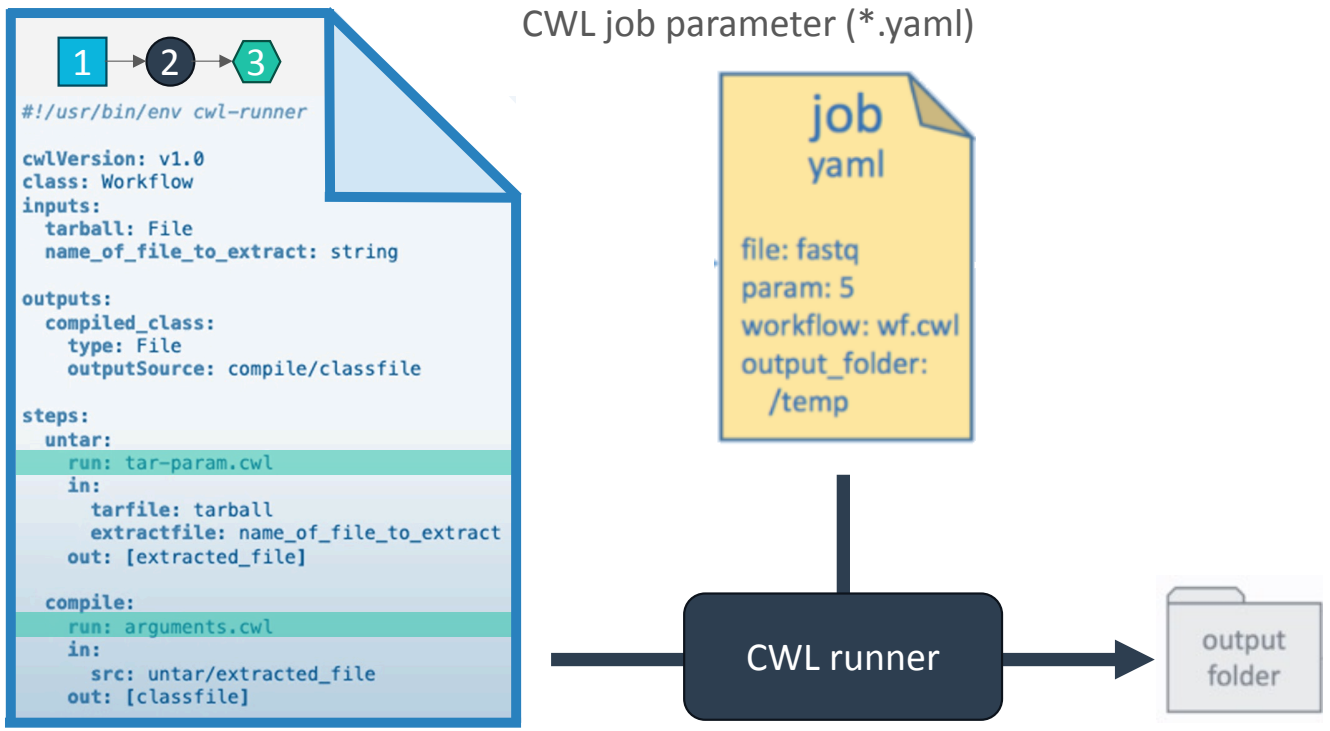


Use CWL to wrap your workflow

CWL workflow document (*.cwl)

CWL job parameter (*.yaml)

CWL
tool descriptors (*.cwl)



Hands-on part 3: Data

Continue the Start Here guide in the DataPLANT knowledge base.

✏️ Until step **Data analysis**

Hands-on part 4: DataHUB

Continue the Start Here guide in the DataPLANT knowledge base.

✏️ Until the final **complete** step

DataHUB terminology and data sharing

Choosing the proper role

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

Permissions & Roles

Roles are assigned when adding a user to an ARC or to a group. This is a very simplified summary.

Guest – Can only see the ARC's wiki

Reporter – Can read, but not add or edit data

Developer – Reporter permissions **+** can read, add, and edit data

Maintainer – Developer permissions **+** can add new members

Owner – Maintainer **+** can delete ARC, manage memberships and permissions

💡 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

Projects and Groups are not the same

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

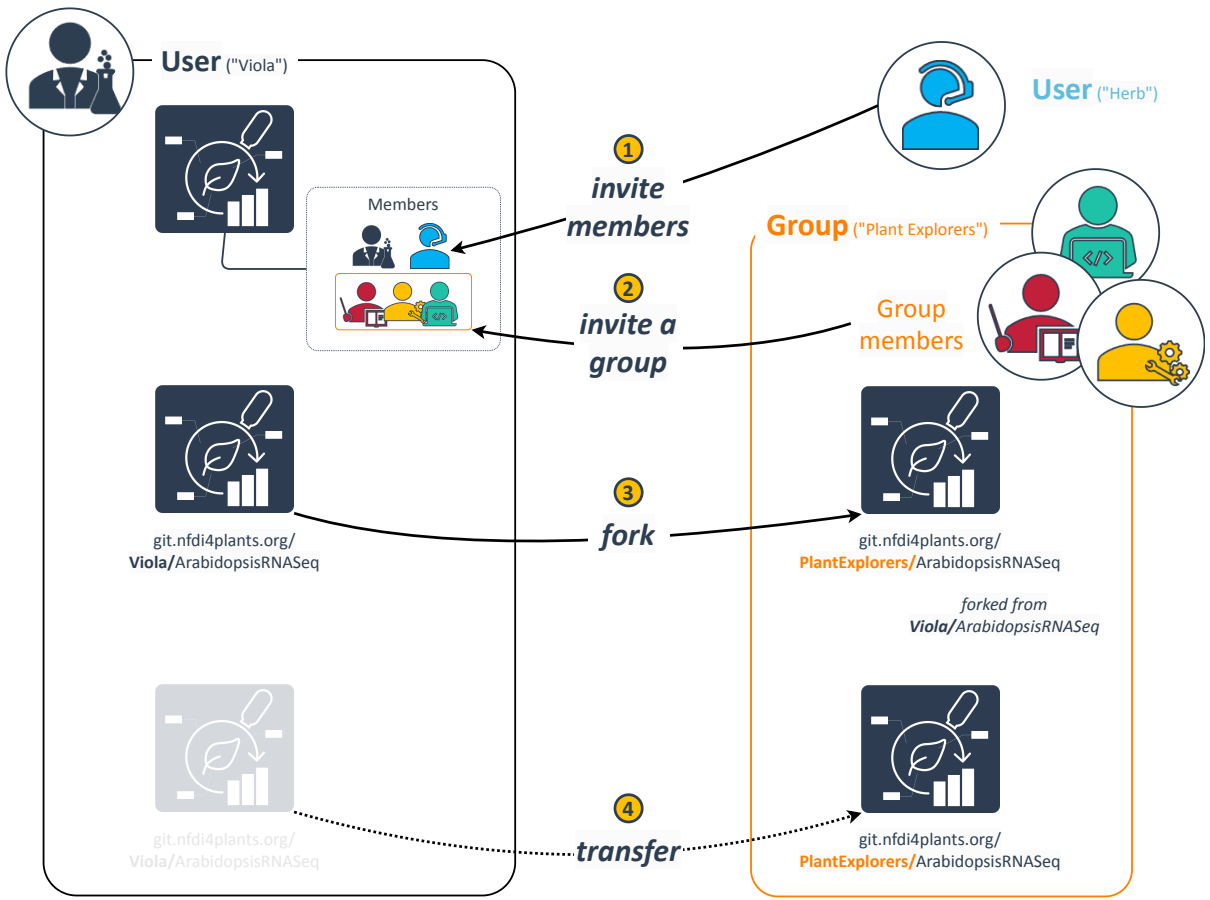
Project = ARC

- 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



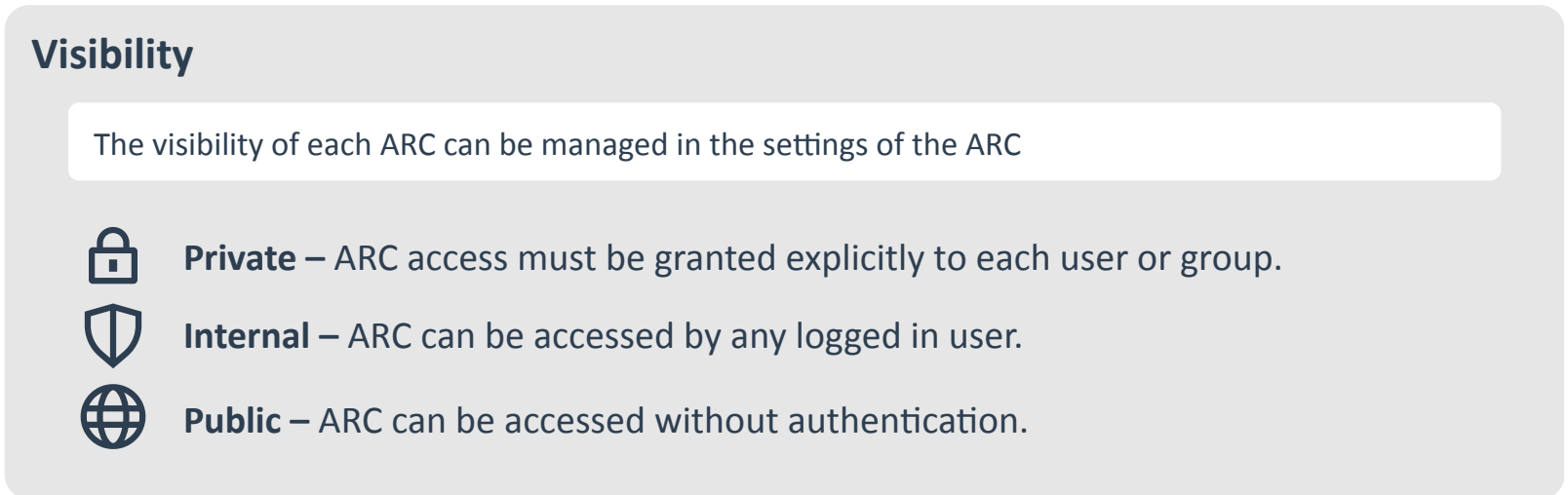
Namespaces

- 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

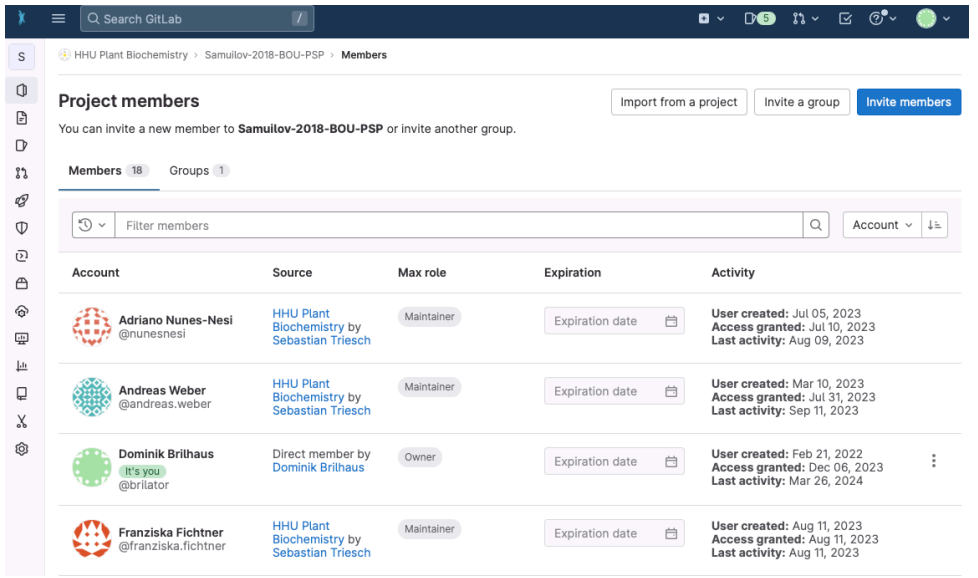
Visibility

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



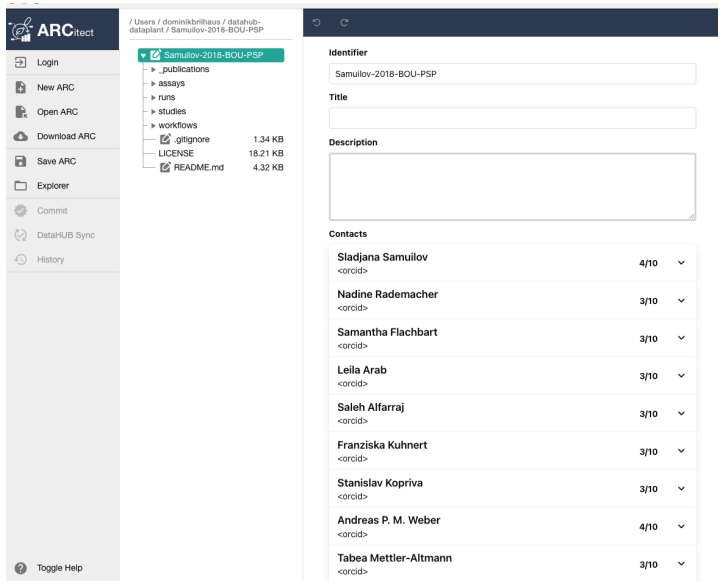
💡 By default every ARC and every group is set to **private**.

ARC DataHUB members // ARC Investigation contacts



DataHUB: ARC members

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

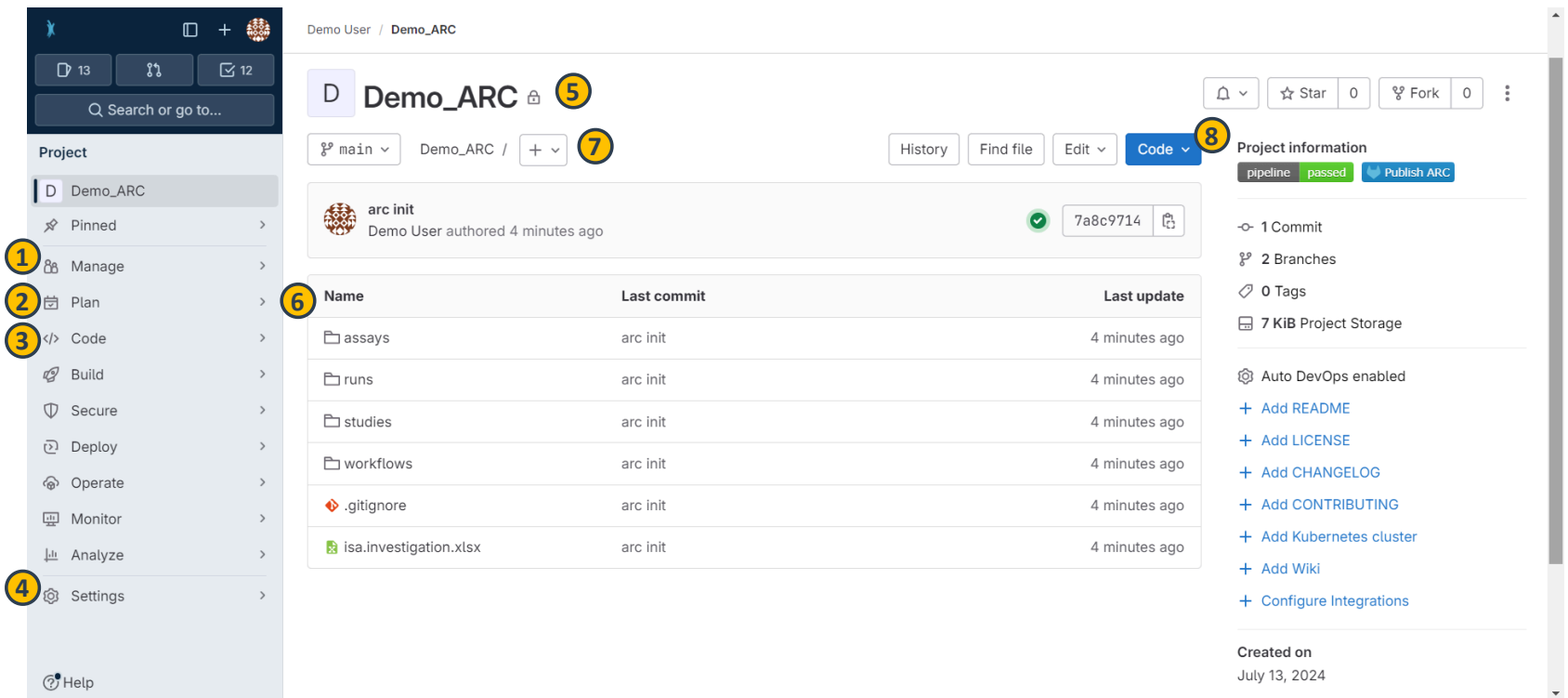


ARCitect: Investigation Contacts

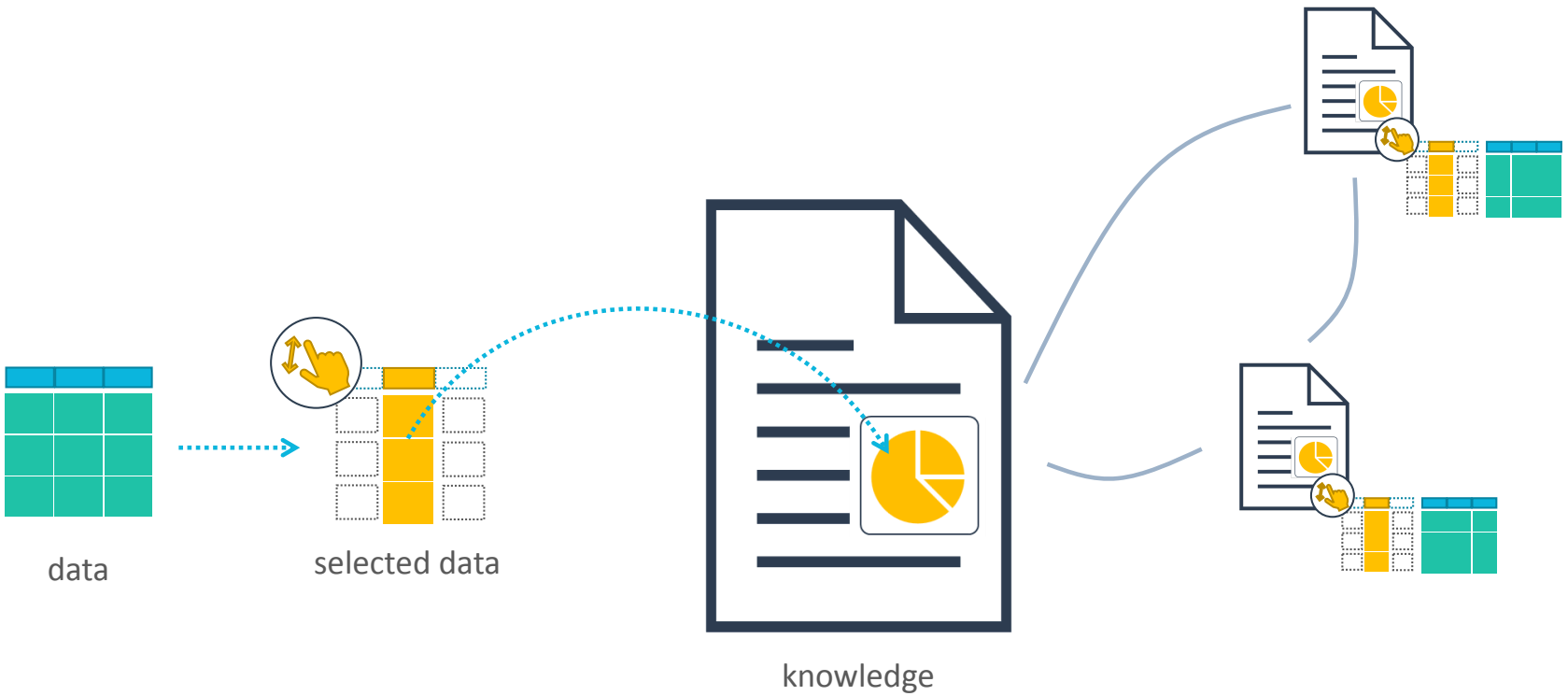
💡 Investigation contacts are not automatically invited as members to the ARC.

Version control

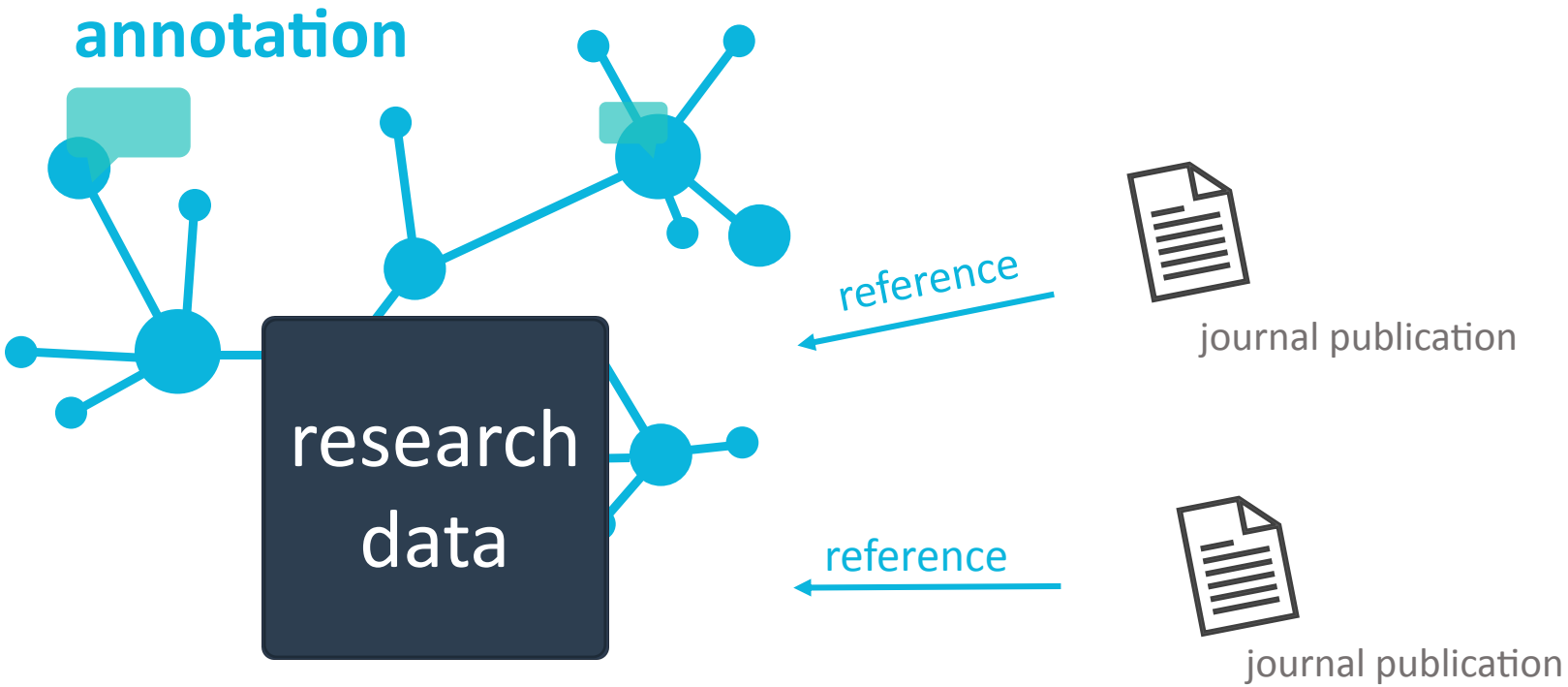
Check out the **commit history** of your ARC via Repository (2) or directly via commits (7)



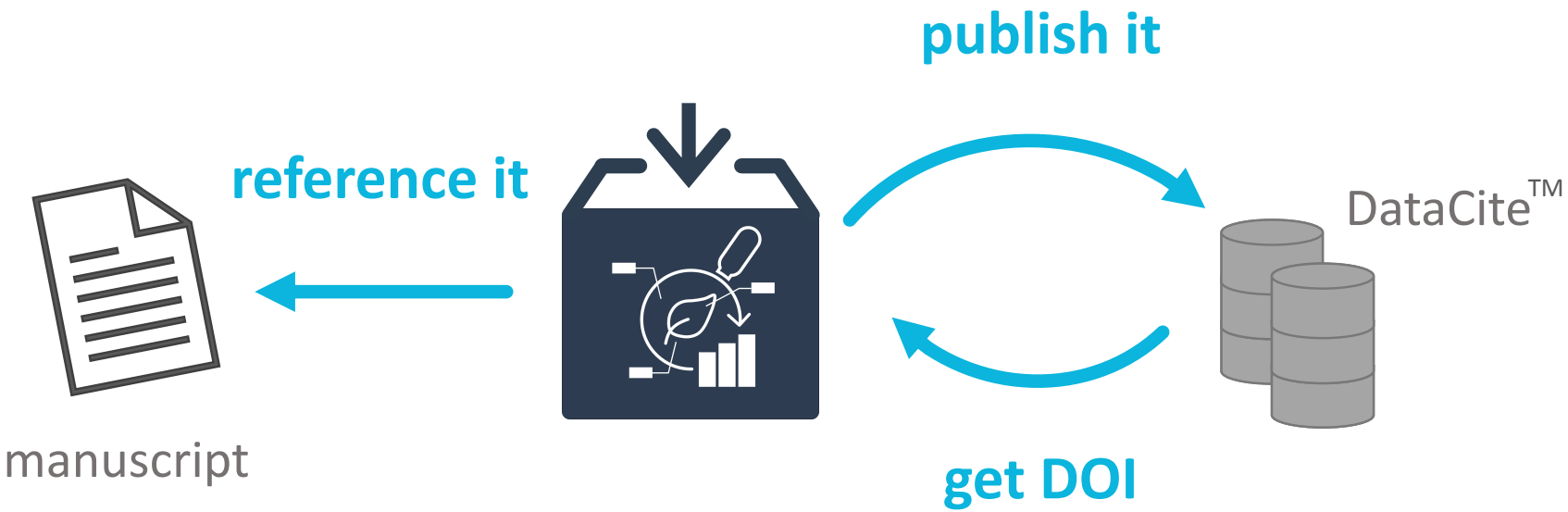
Hands-on part 5: ARC data publication



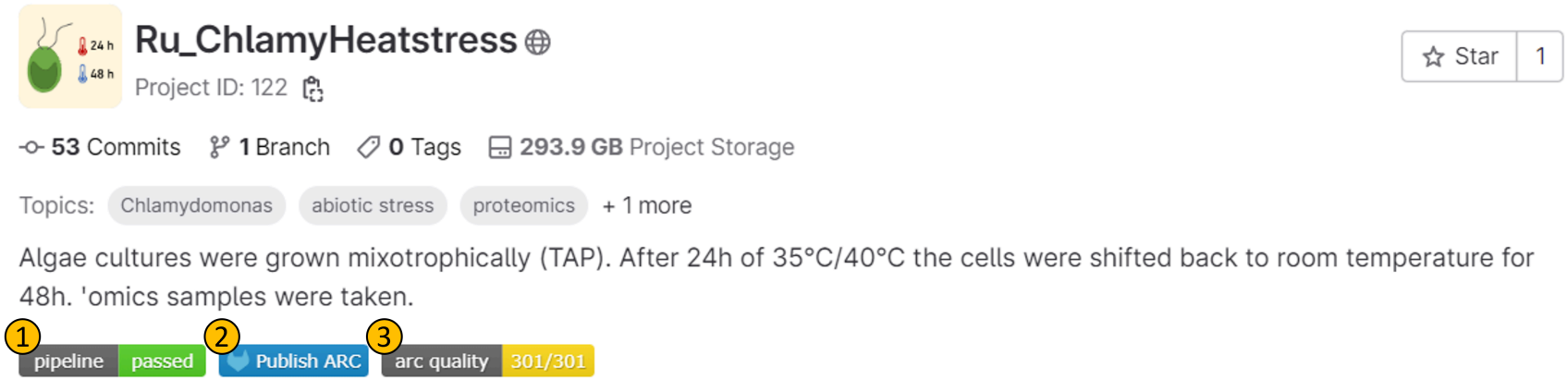
Moving from paper to data publications



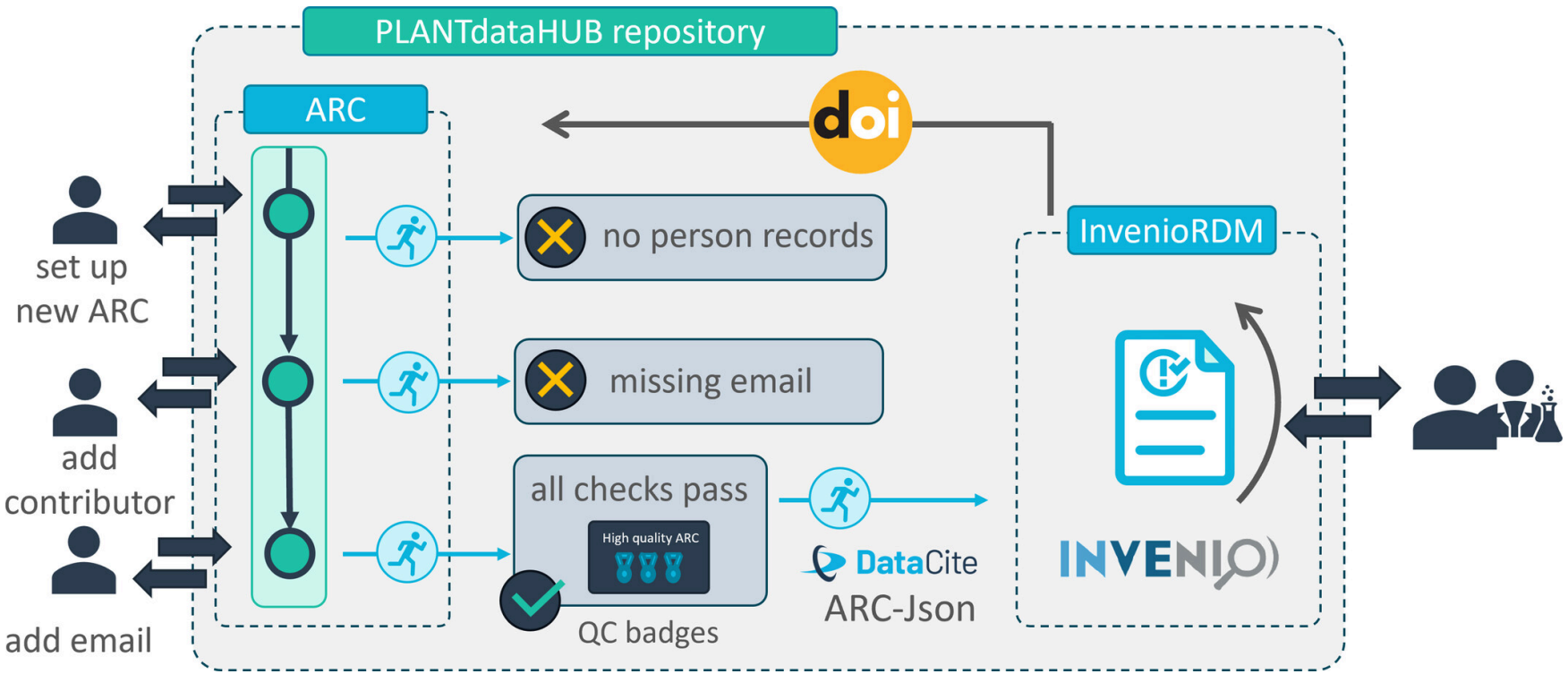
Publish your ARC, get a DOI



Publish your ARC with a few clicks

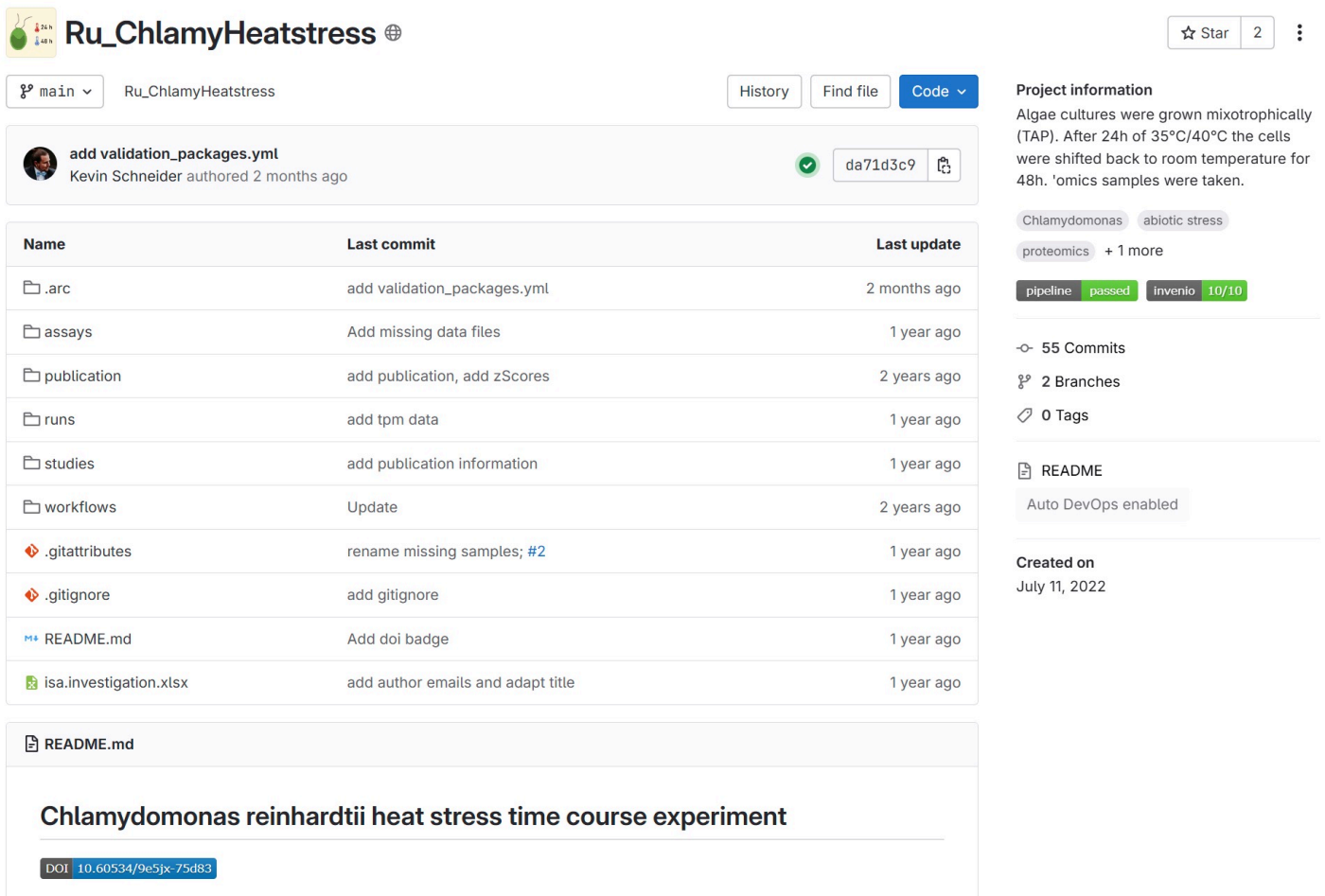


Validate & publish



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

DOI 10.60534/9e5jx-75d83

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

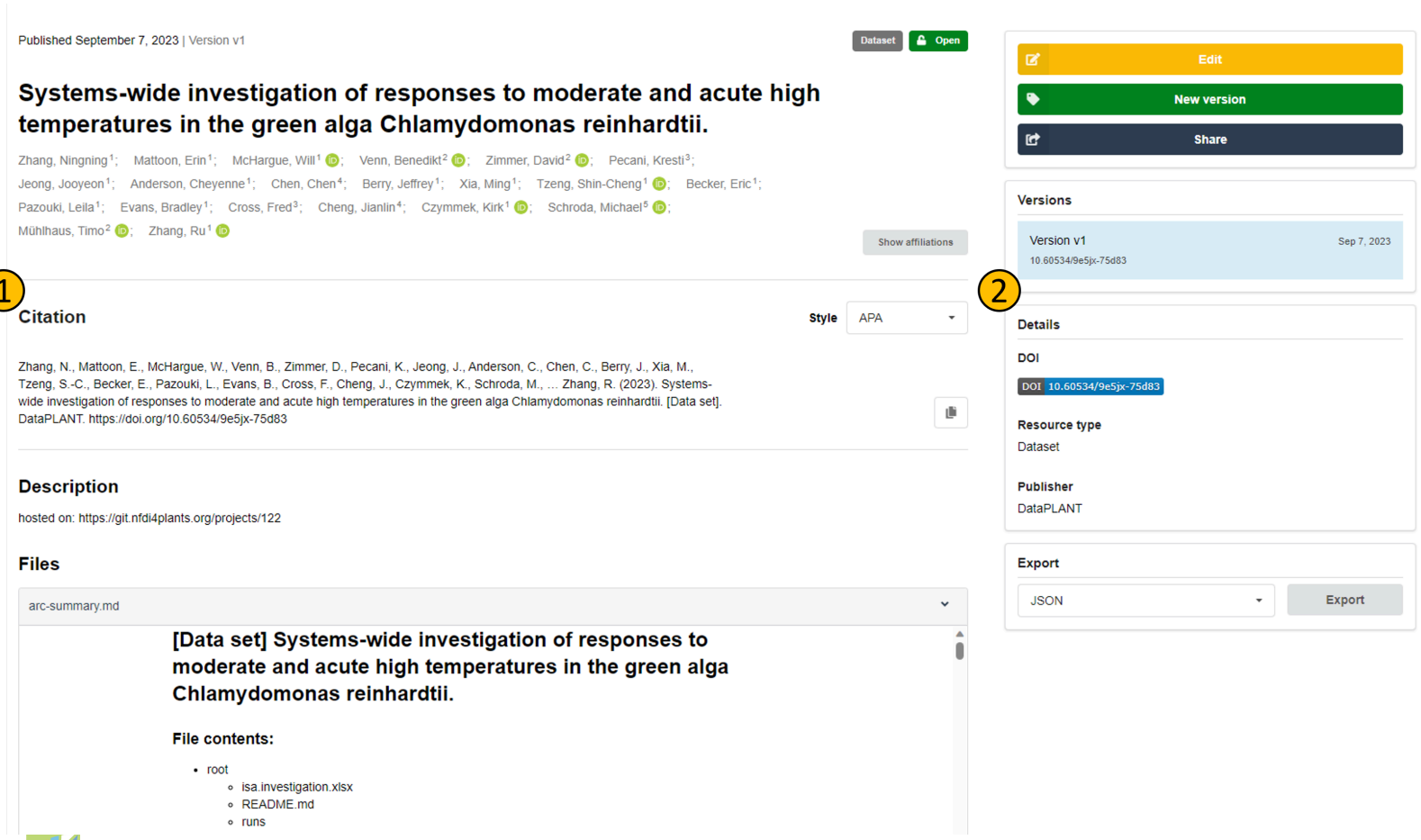
55 Commits
2 Branches
0 Tags

README
Auto DevOps enabled

Created on
July 11, 2022

Data PLANT
CEPLAS

Receive a DOI



Hands-on part 5: Data publication

1. ARC validation
2. (Towards) ARC publication

Resources

Info & materials

- DataPLANT Website: https://nfdi4plants.org/
- ARC website: https://arc-rdm.org
- Knowledge Base: https://nfdi4plants.org/nfdi4plants.knowledgebase/

Continuous support

- HelpDesk: https://helpdesk.nfdi4plants.org
- Matrix for ad hoc support: https://matrix.to/#/%23arc-user-support:matrix.org
- User support meeting (2nd Friday of the month | 1 - 2pm): https://nfdi4plants.github.io/events/arc-user-support/

Tools and Services

- ARCitect: https://github.com/nfdi4plants/arcitect
- DataHUB: https://git.nfdi4plants.org

Open Source Development

- GitHub: https://github.com/nfdi4plants

Acknowledgements

Team Kaiserslautern

- Timo Mühlhaus
- Lukas Weil
- Kevin Frey
- Kevin Schneider
- Jonas Lukasczyk

Team Freiburg

- Dirk von Suchodeletz
- Jonathan Bauer
- Marcel Tschöpe
- Julian Weidhase

CEPLAS
Cluster of Excellence on Plant Sciences

- Björn Usadel
- Vittorio Tracanna
- Yaser Alashloo

MibiNet

- Sabrina Zander