

# ARCify your research project

fdm.nrw – FDM Werkstatt

Dominik Brilhaus, [CEPLAS](#)

April 2nd, 2025



**CEPLAS**  
Cluster of Excellence on Plant Sciences

Data)(PLANT

# 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/>

## Tools and Services

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

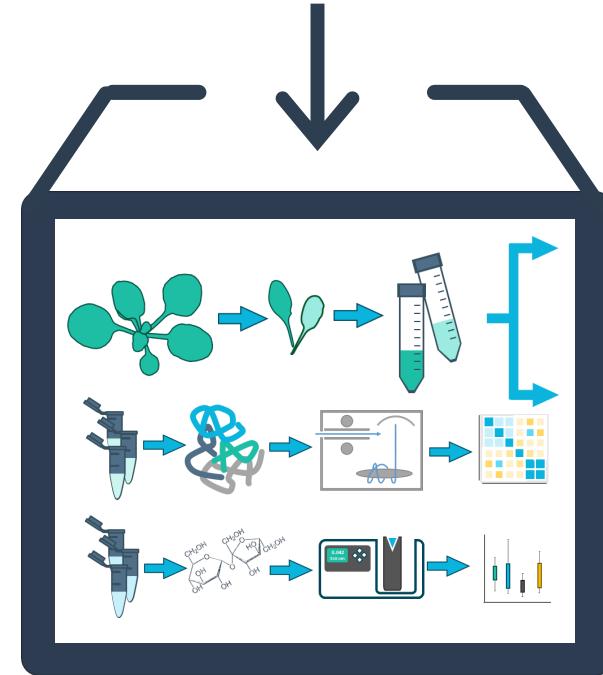
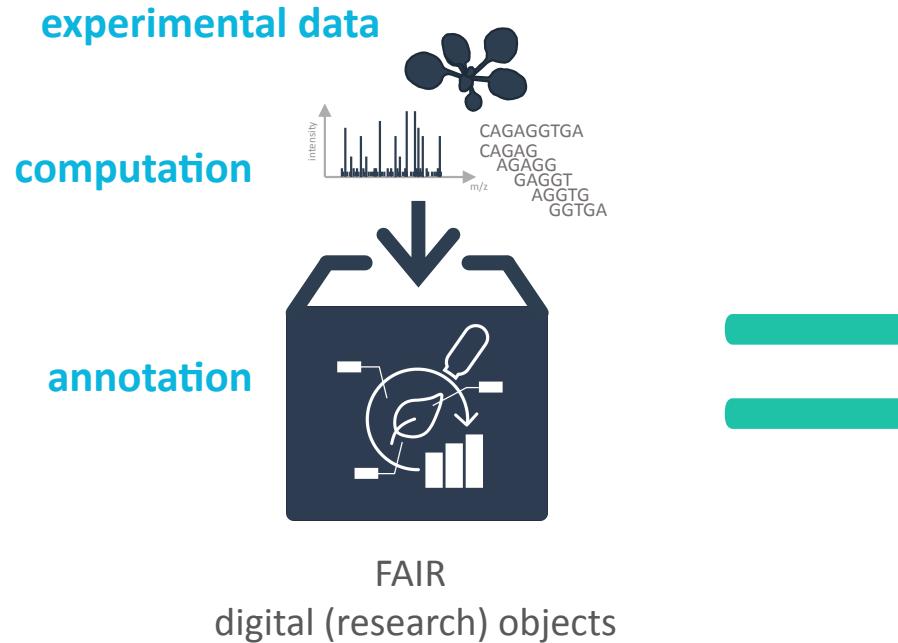
## Continuous support

- 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/>

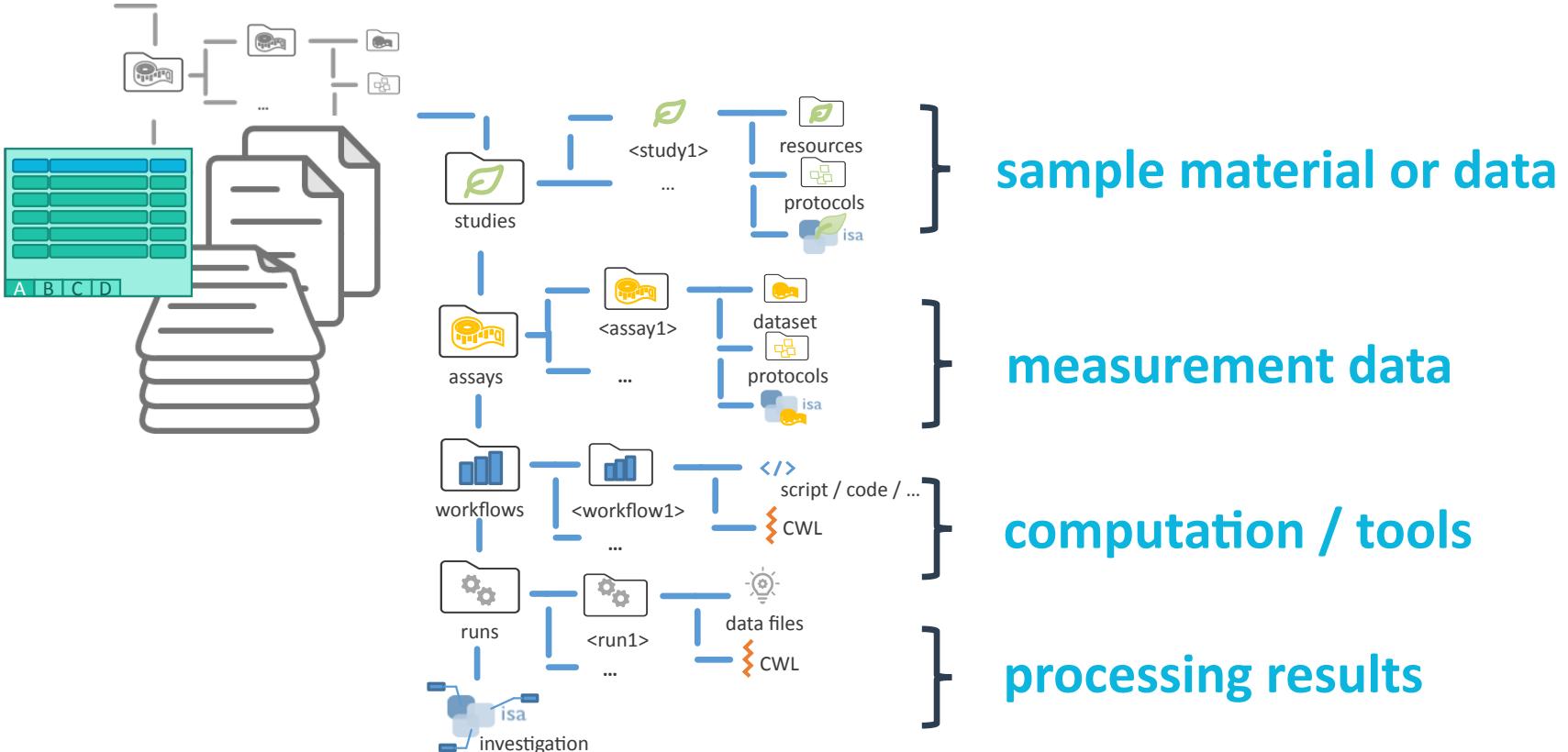
## Open Source Development

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

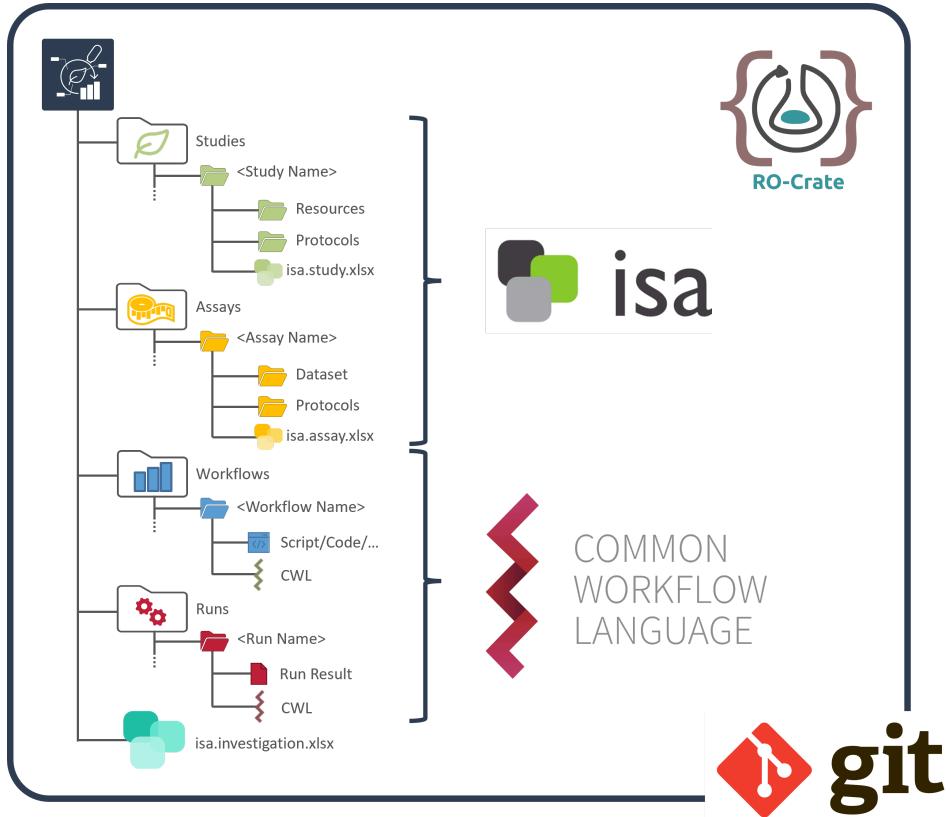
# Annotated Research Context (ARC)



# The ARC scaffold structure



# ARC builds on standards



# 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

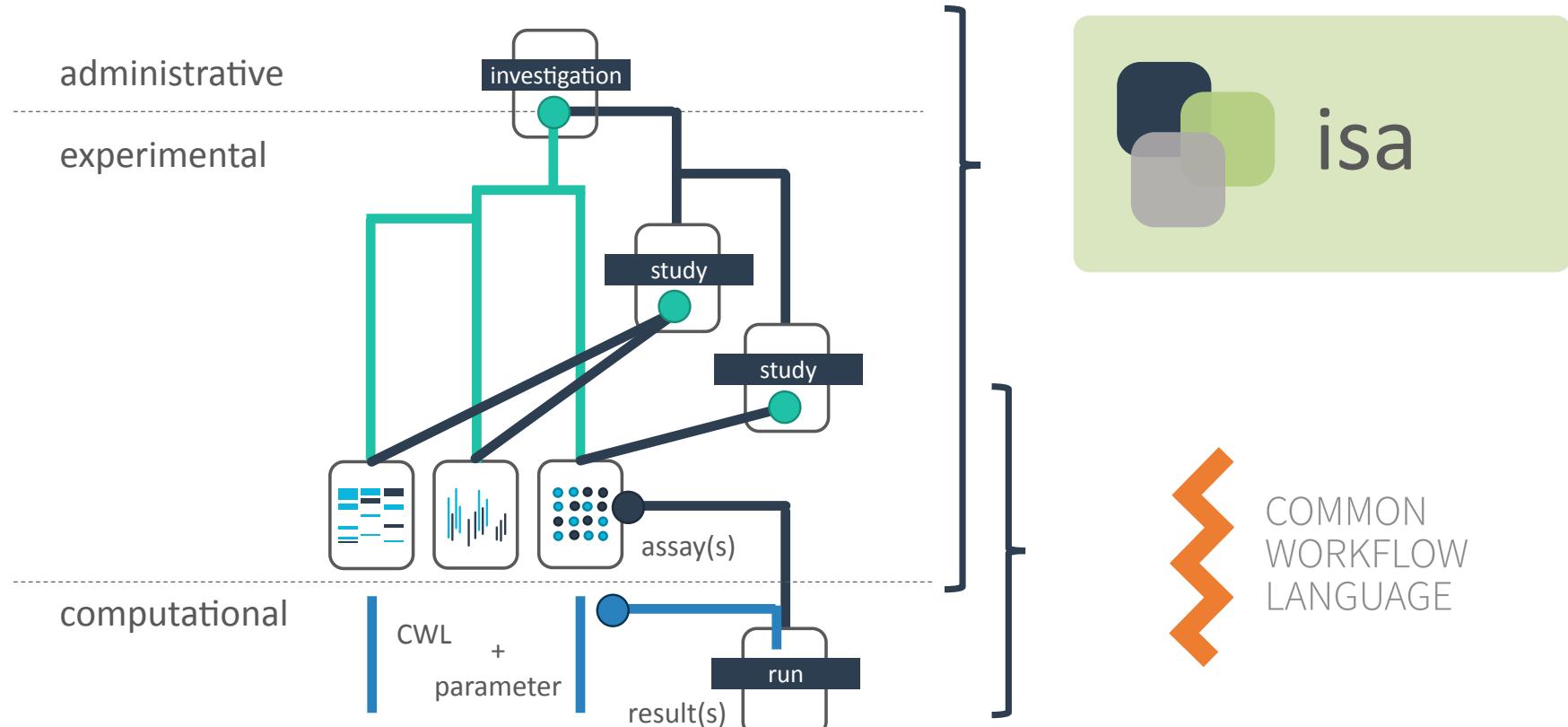
- Characteristics
- Parameters
- Components
- Factors

## Assay

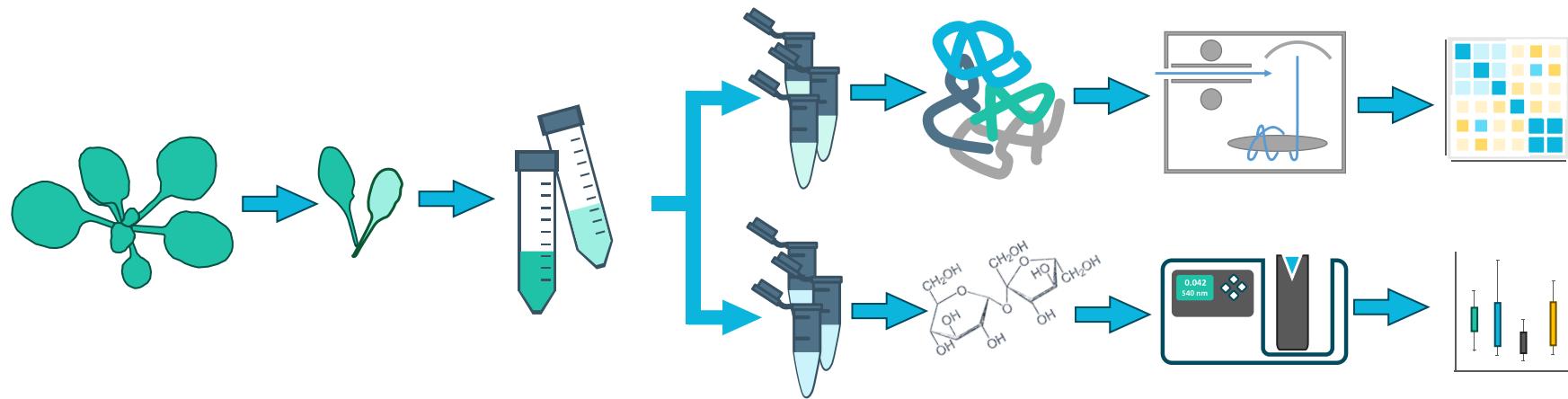
descriptive (meta)data  
information on the measurement

- Characteristics
- Parameters
- Components
- Factors

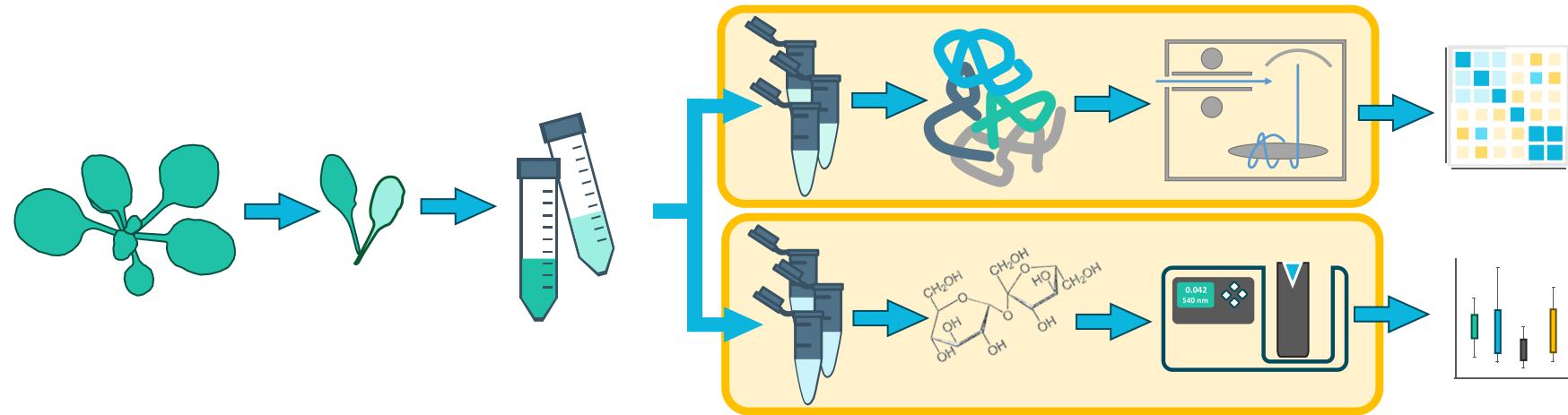
# ISA and CWL – Connected by similarity



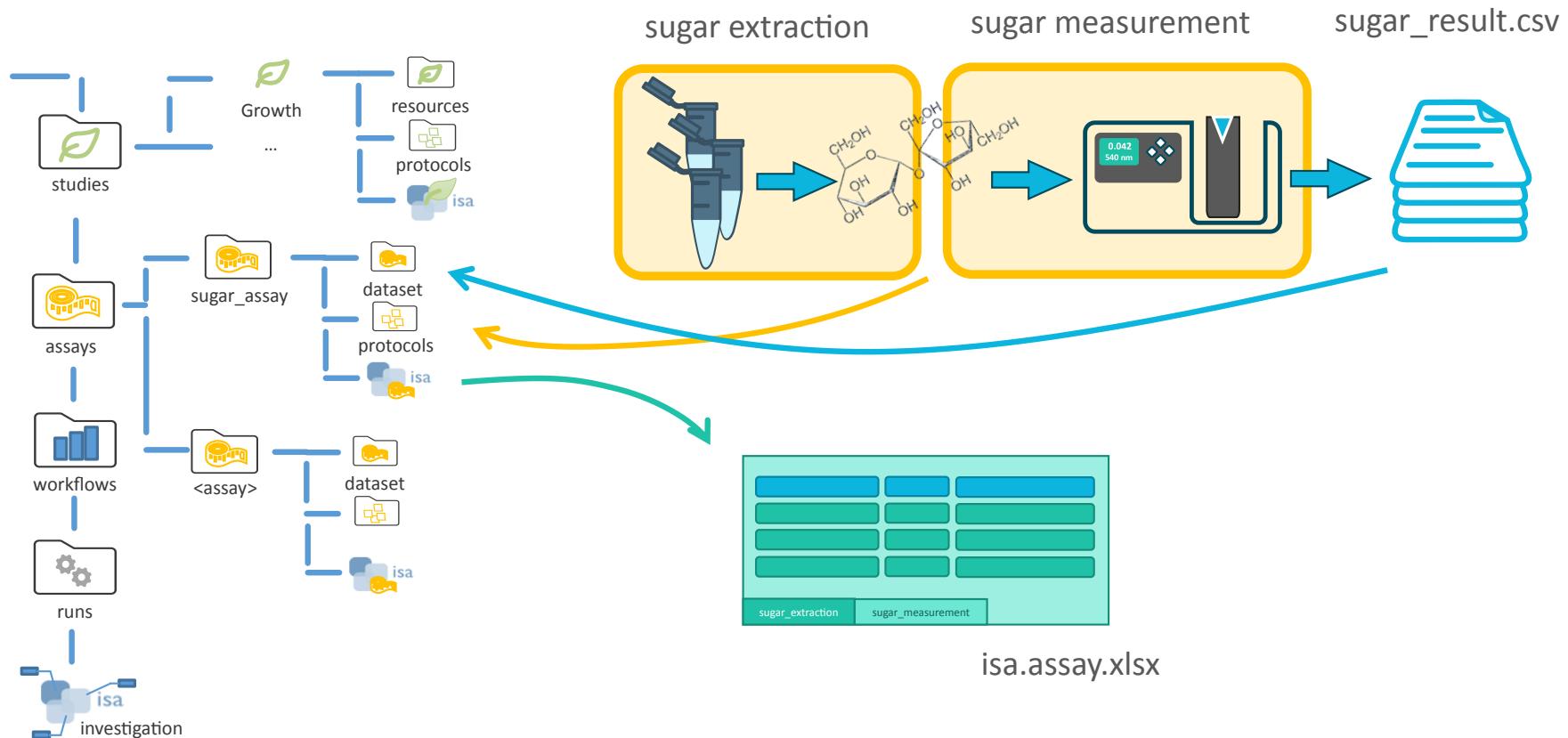
## Metadata annotation – from sample to data



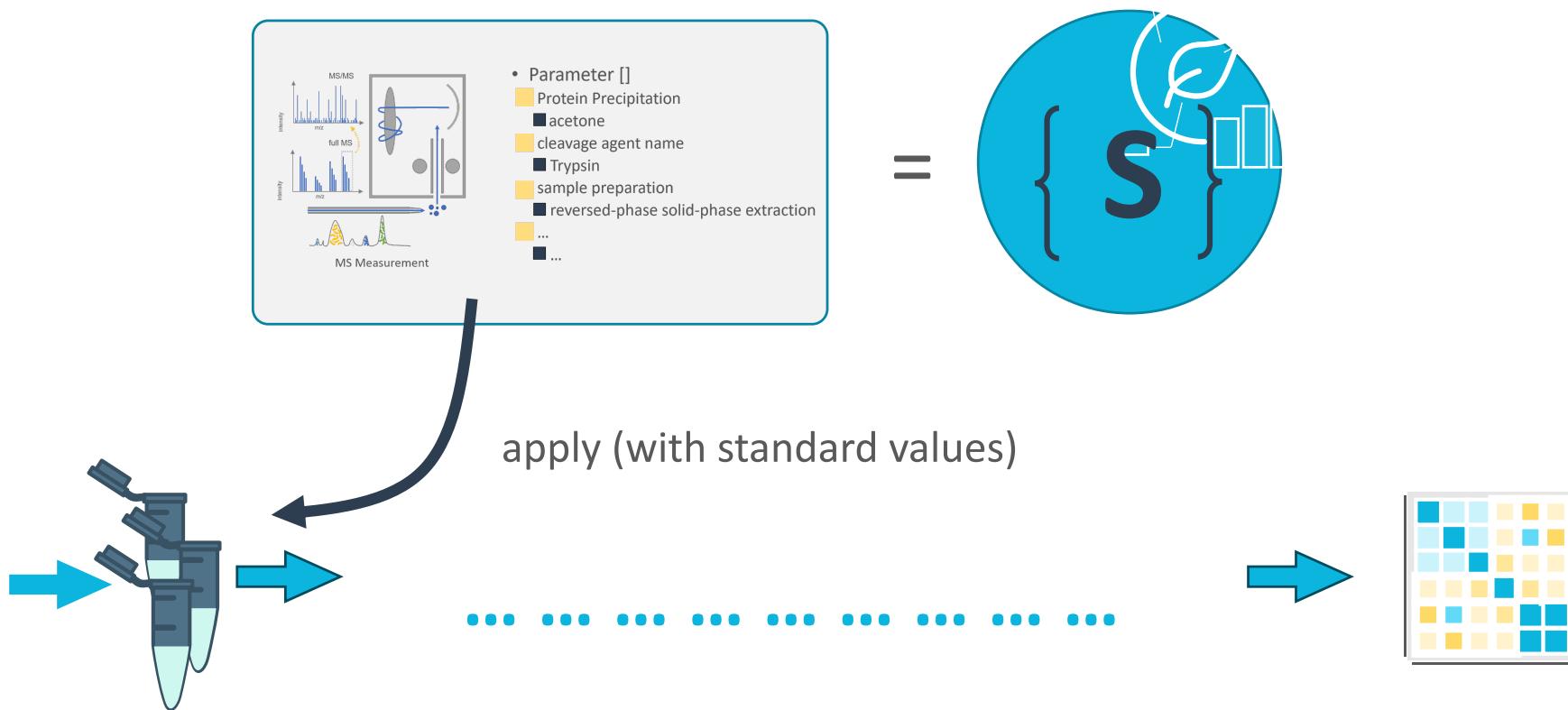
# Modular separation of experimental processes



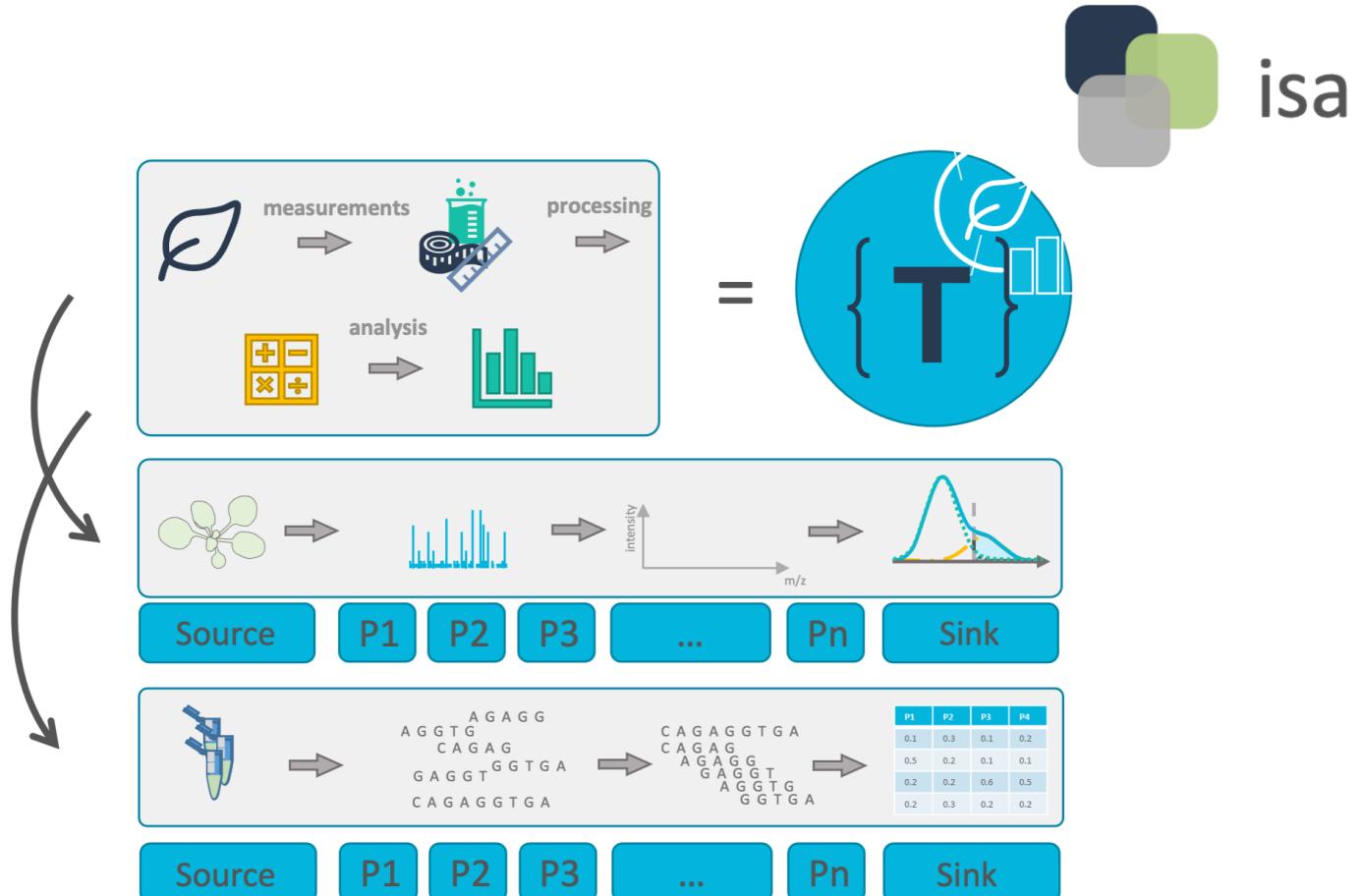
# Modular separation of experimental processes



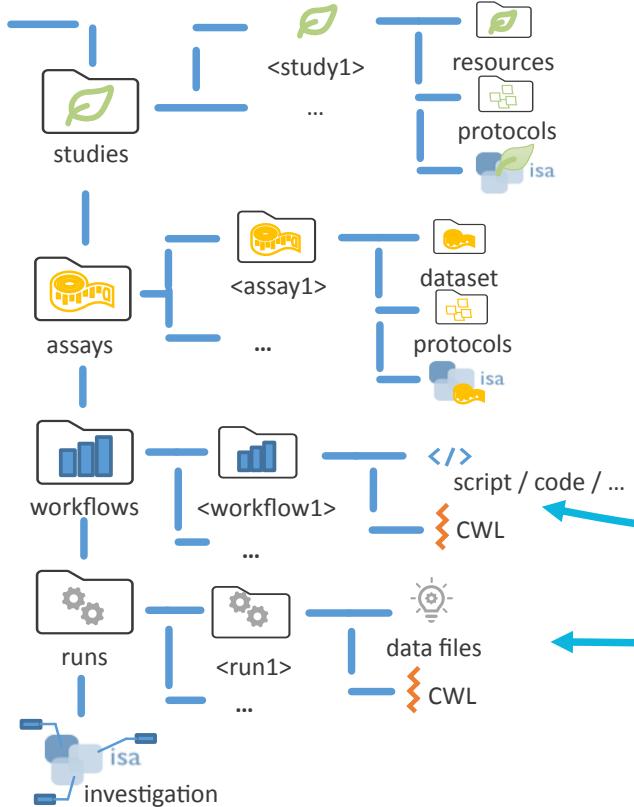
# Applying standard procedures to sample record



# Realization of lab-specific metadata with templates



# Data analysis: CWL workflows and runs



```
python
Copy code

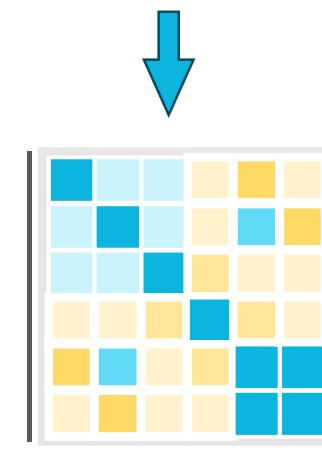
import pandas as pd
import plotly.express as px

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

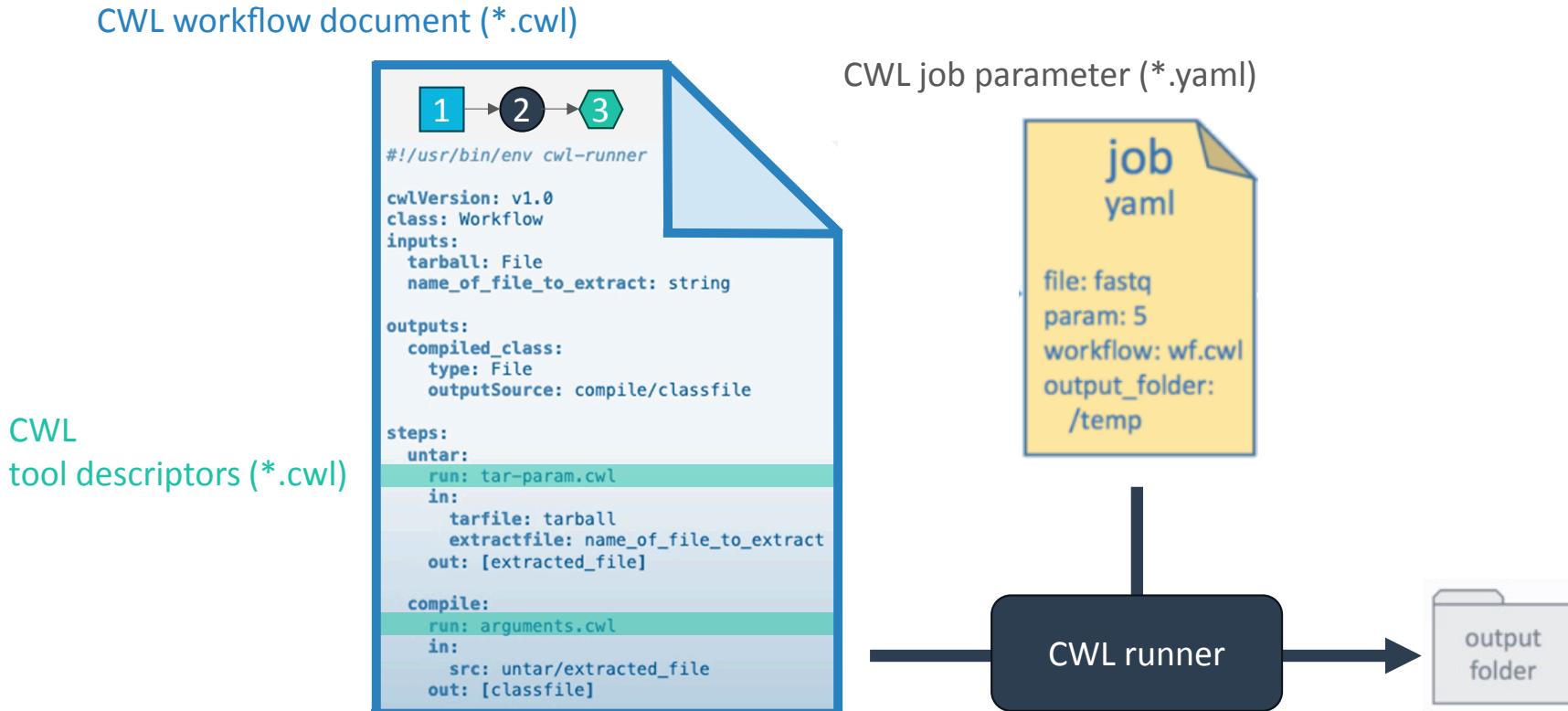
# If the CSV doesn't have a clear index or row/column names, you may need to manually set
# For example, if the first column should be the index, you can use:
# data = pd.read_csv('result.csv', index_col=0)

# 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()
```

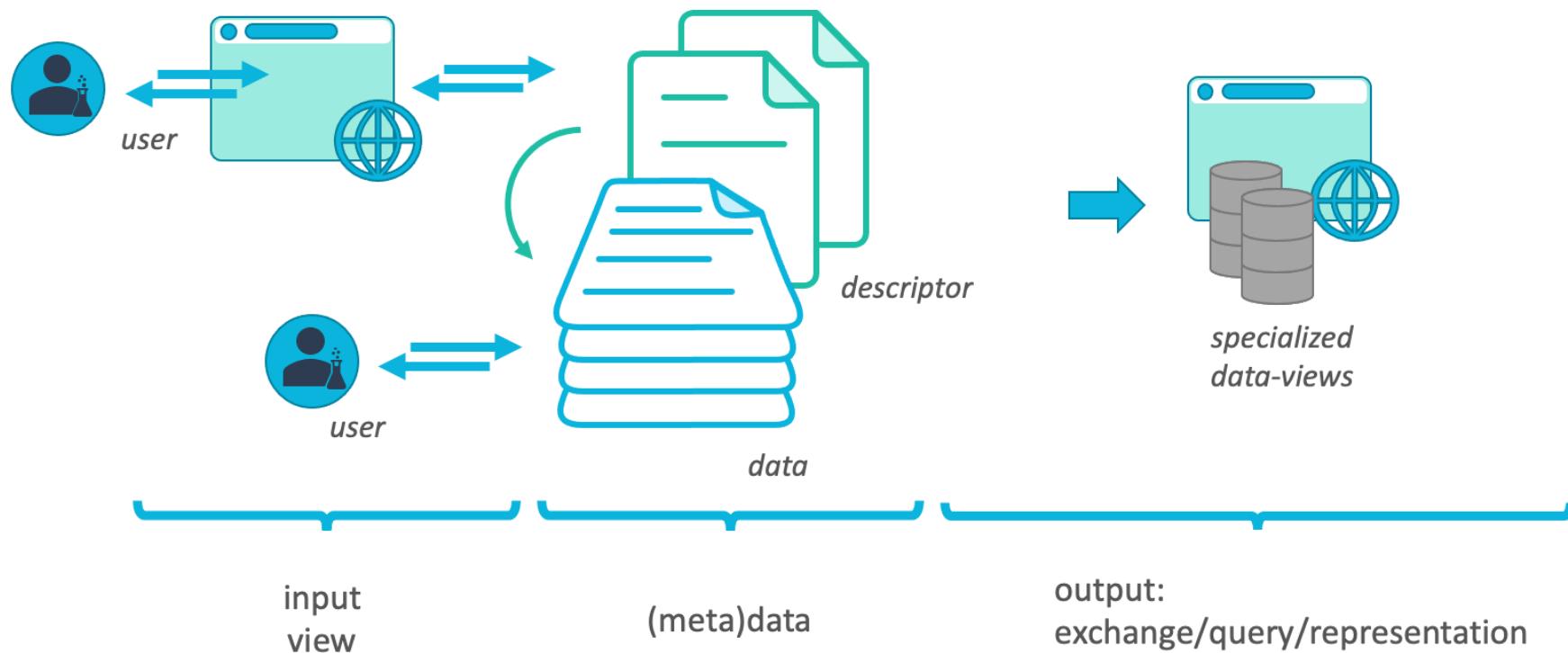


# Metadata annotation – from data to result

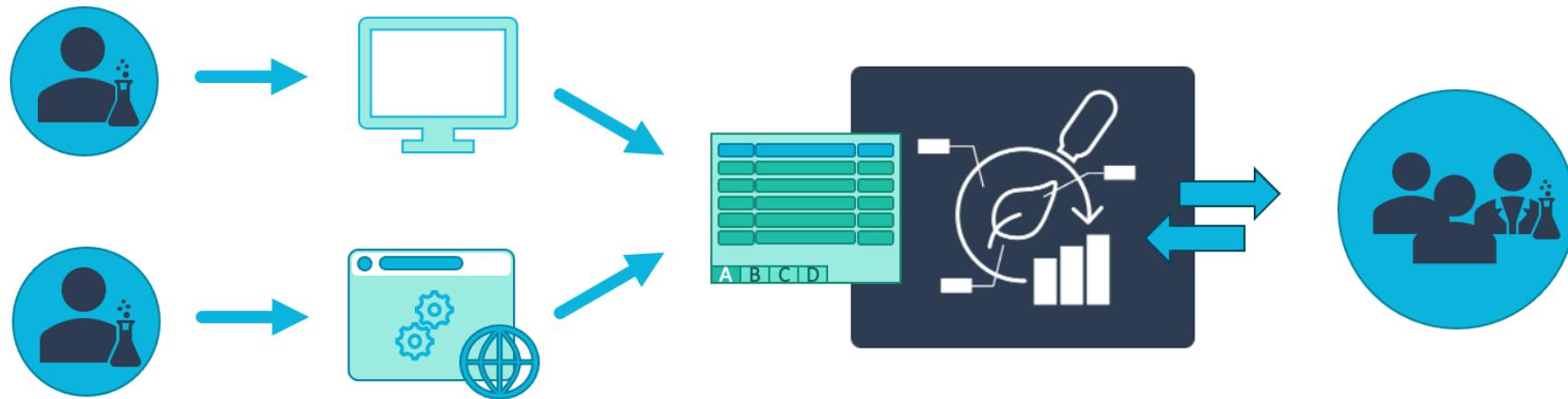


# Everything is a file

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

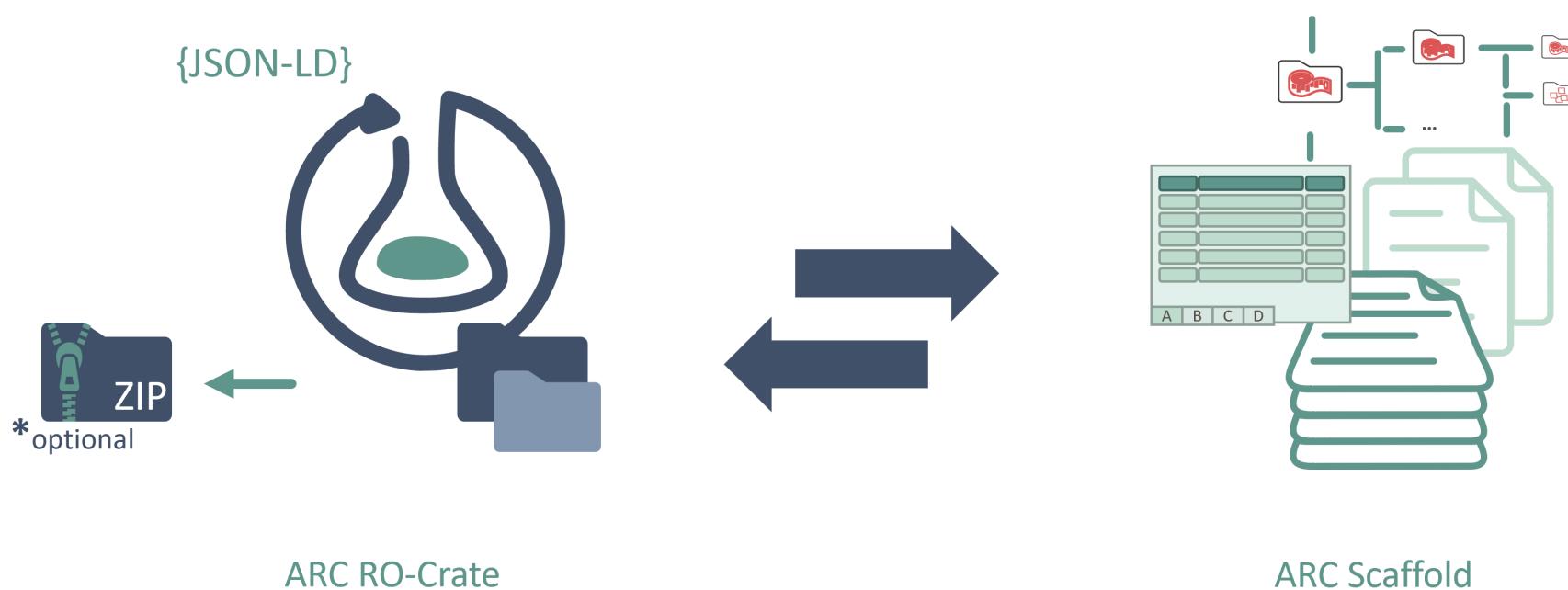


## No technical lock-in



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

## Two representations of the ARC



# 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 screenshot shows the ARCitect application interface. On the left, the ARC Scaffold is displayed as a hierarchical tree structure:

- / Users / dominikbrilhause / Downloads / AthalianaColdStressSugar
  - studies
  - assays
    - Proteomics\_DataAnalysis
      - Proteomics\_MS
      - dataset
      - protocols
      - README.md
      - SugarMeasurement
      - Visualization
    - workflows
    - runs
    - README.md

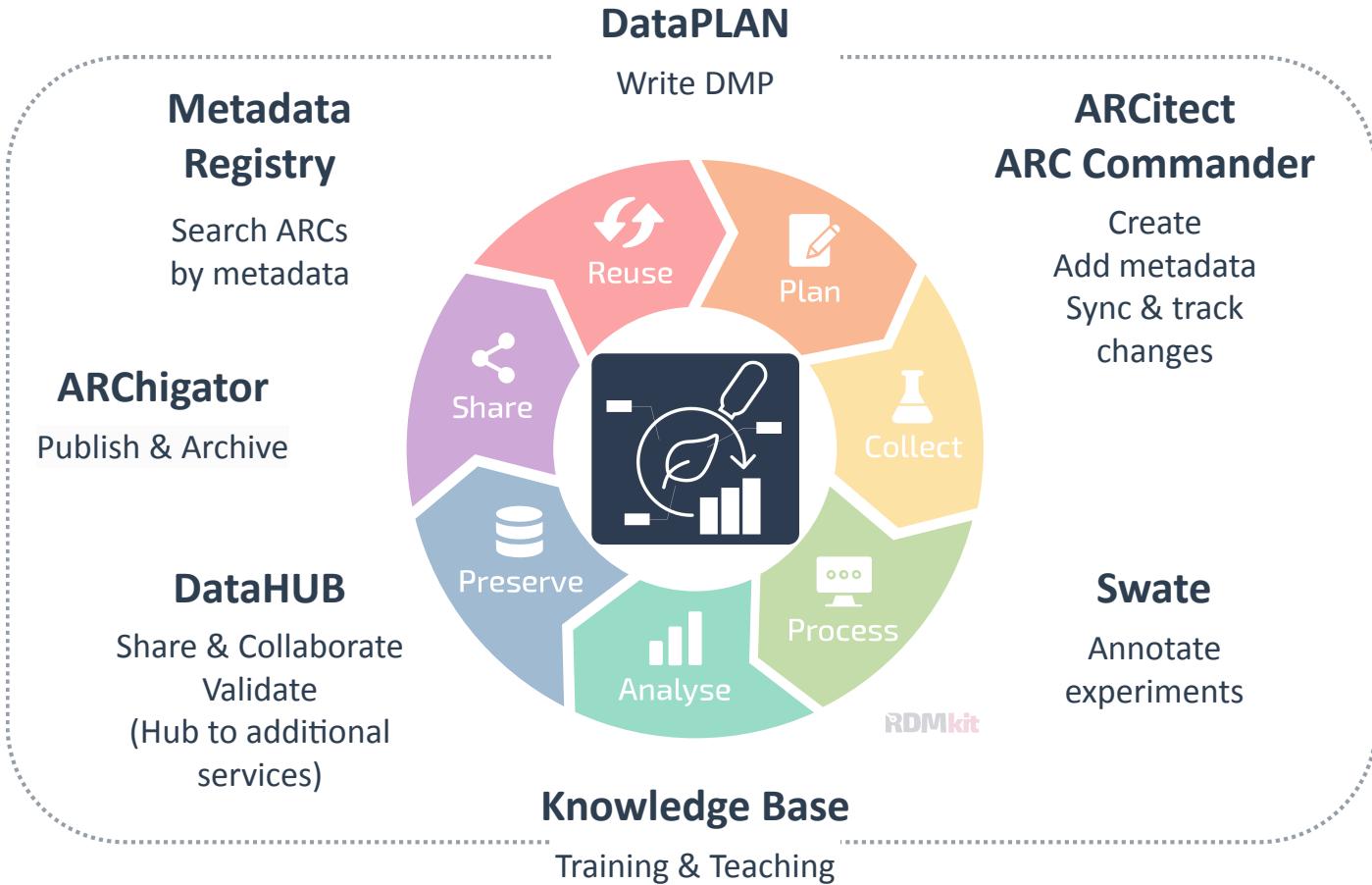
On the right, the "Assay Metadata" panel is open, showing the following fields:

Identifier	Proteomics_MS
Measurement Type	Proteomics_MS
Technology Type	Mass Spectrometry
Technology Platform	timsTOF Pro 2
Performers	Assay, PeptideMS_Bruker, ProtDigest

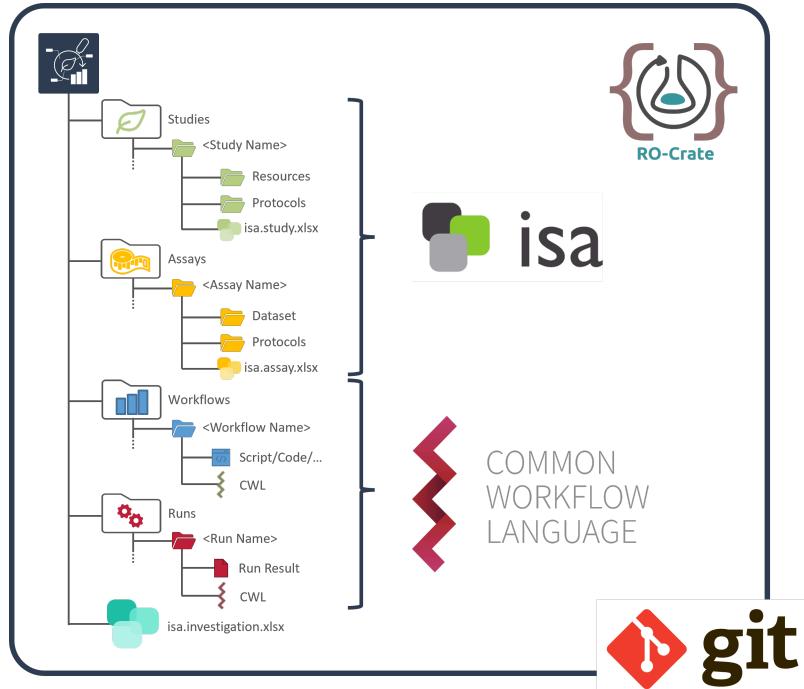
Below the metadata panel, there is a table with four columns:

Parameter [sample mass] >>	Parameter [Protein Precipitation] >>	Parameter [alkylating agent] >>	Parameter [red]
10 microgram	acetone	✓ Chloroacetamide	✓ TCEP

# 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/>

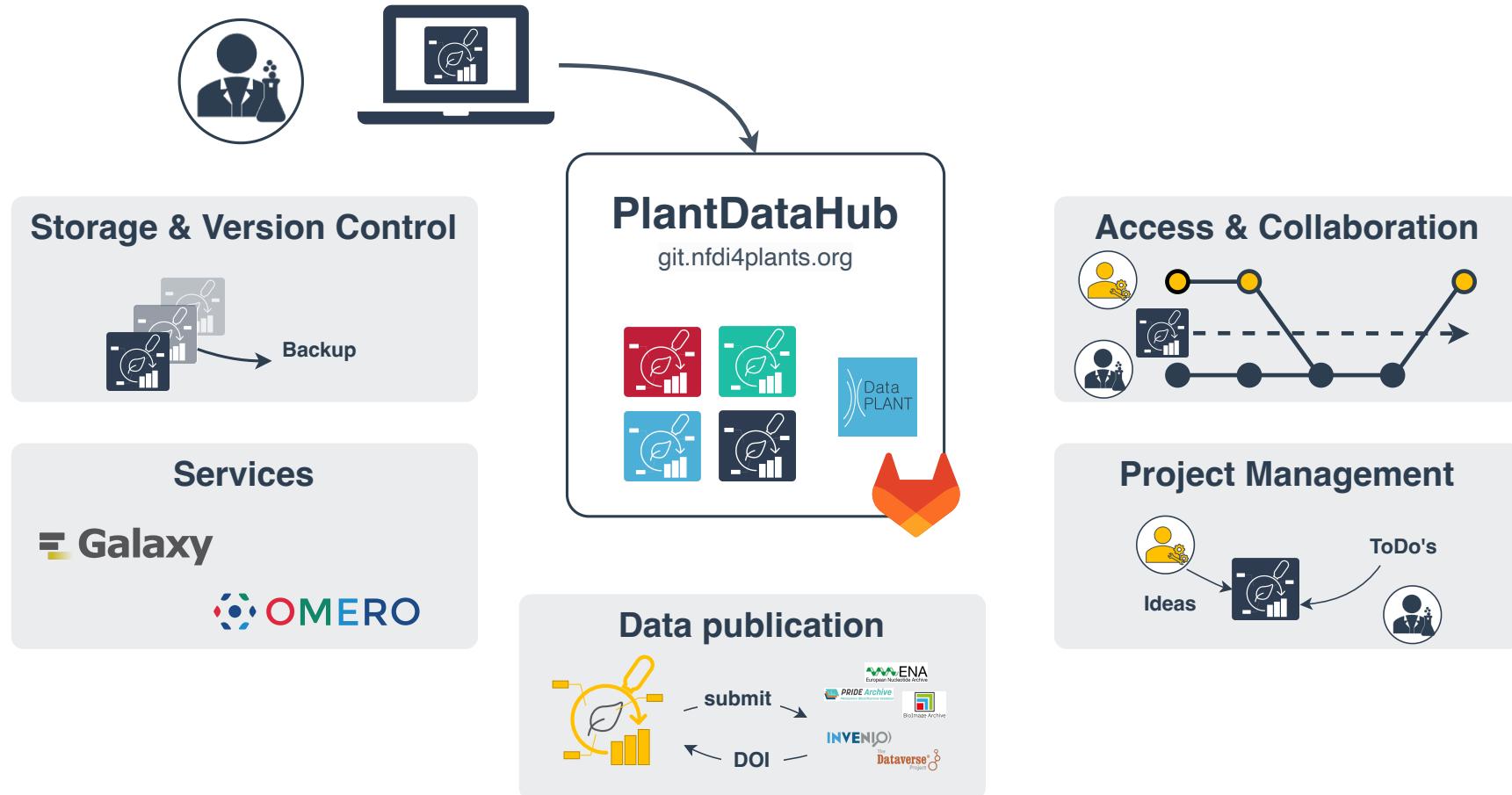
## CWL

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

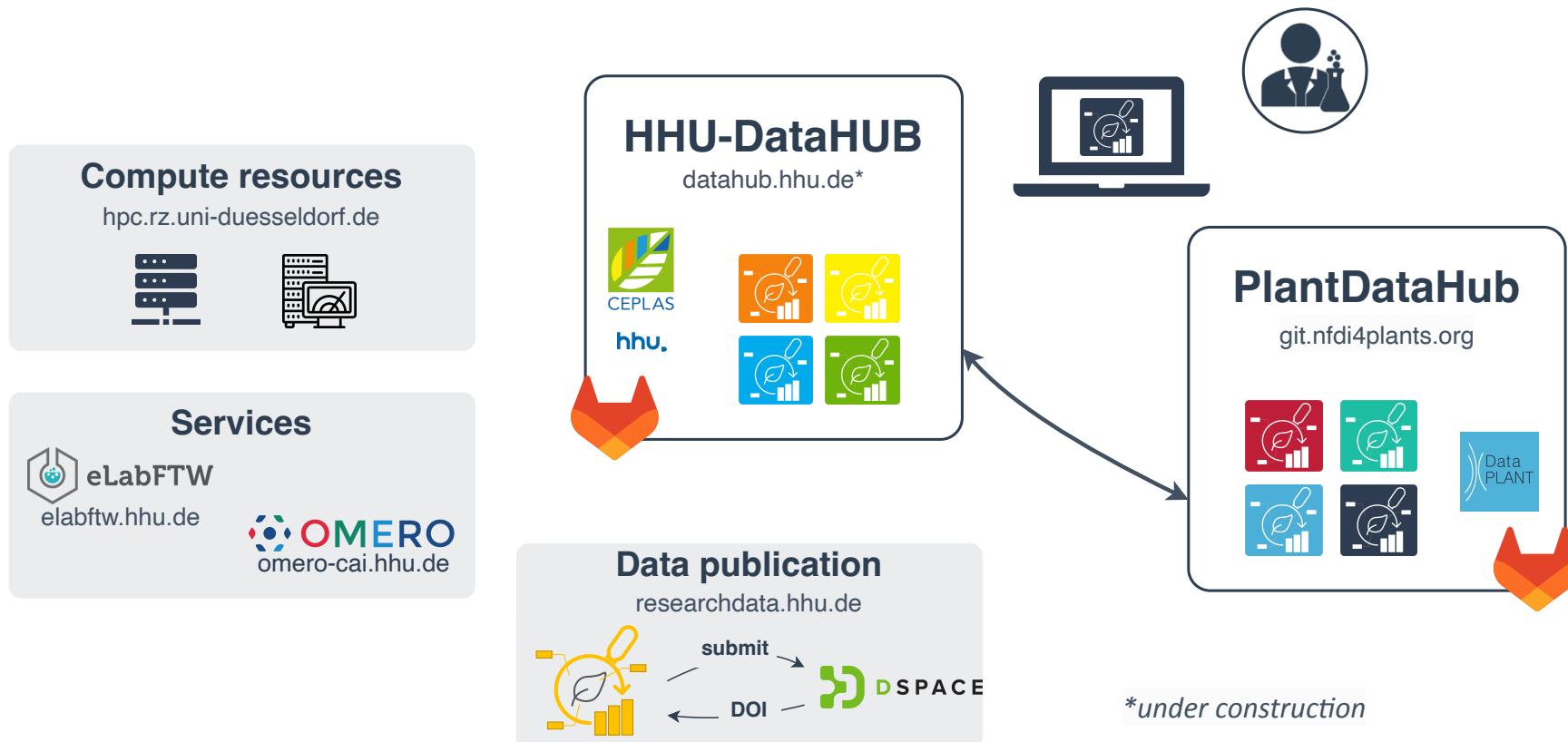
## Git

- Version control

# ARC and DataHUB as entry point



# HHU-DataHUB – On-premise DataHUB node

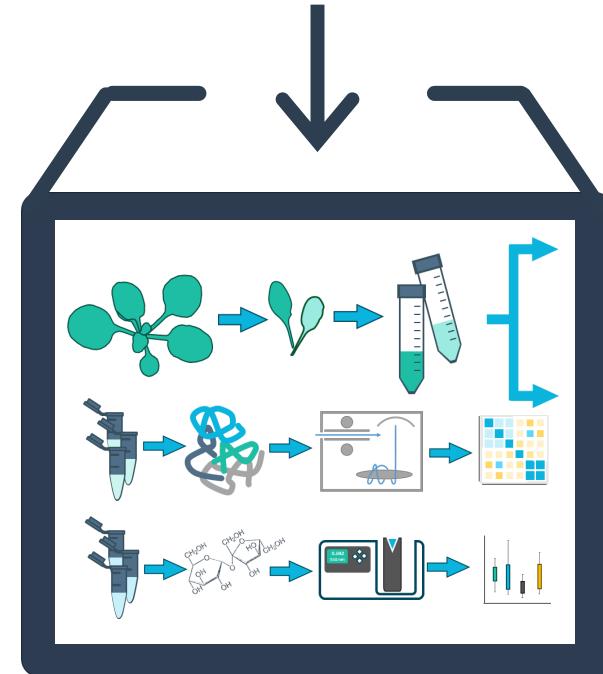
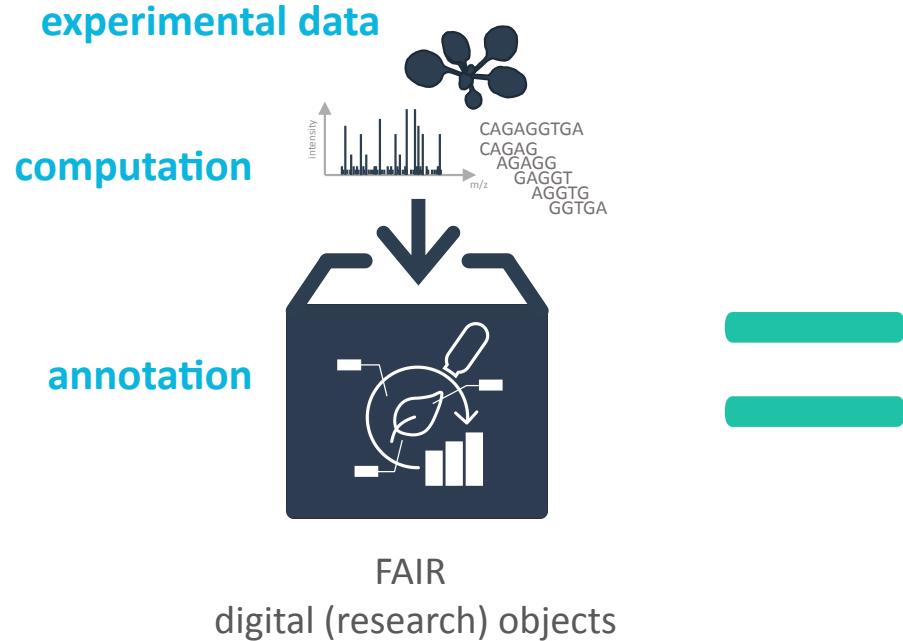


## Hands-on part 1: Setup and ARCitect

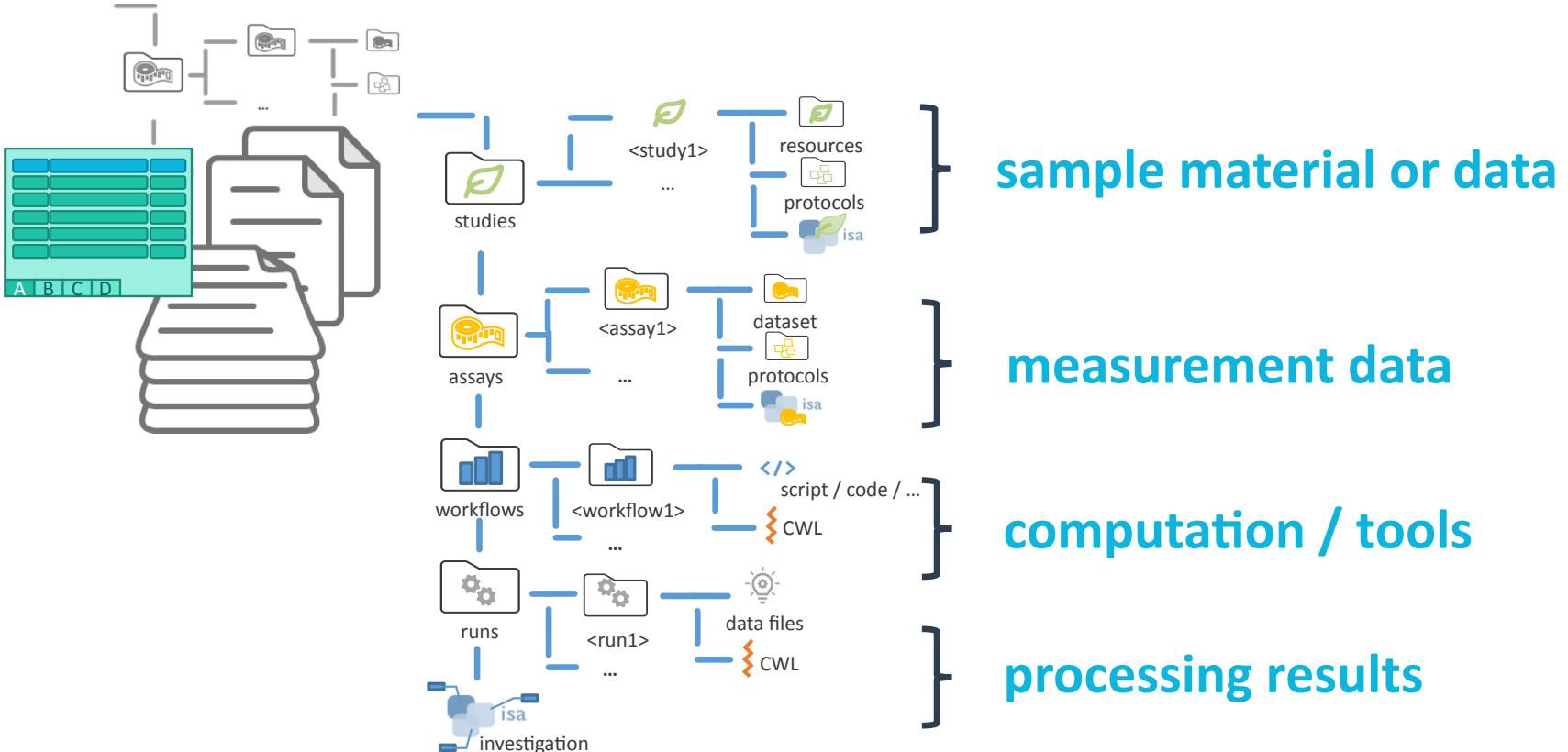
Follow the [Start Here guide](#) in the DataPLANT knowledge base.

- 💡 Until step Add a study

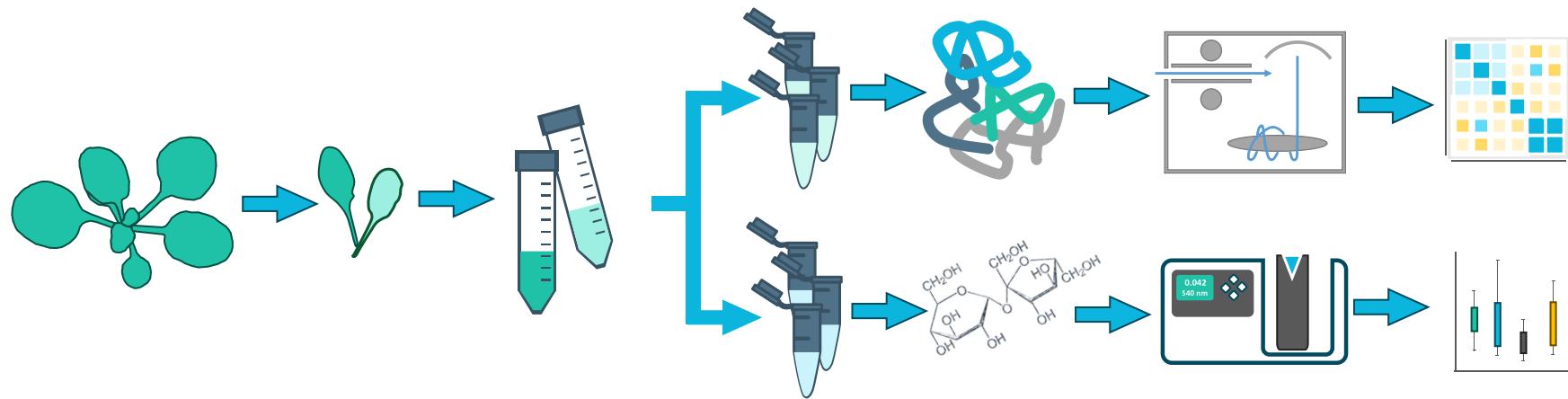
# ARC: Annotated research context



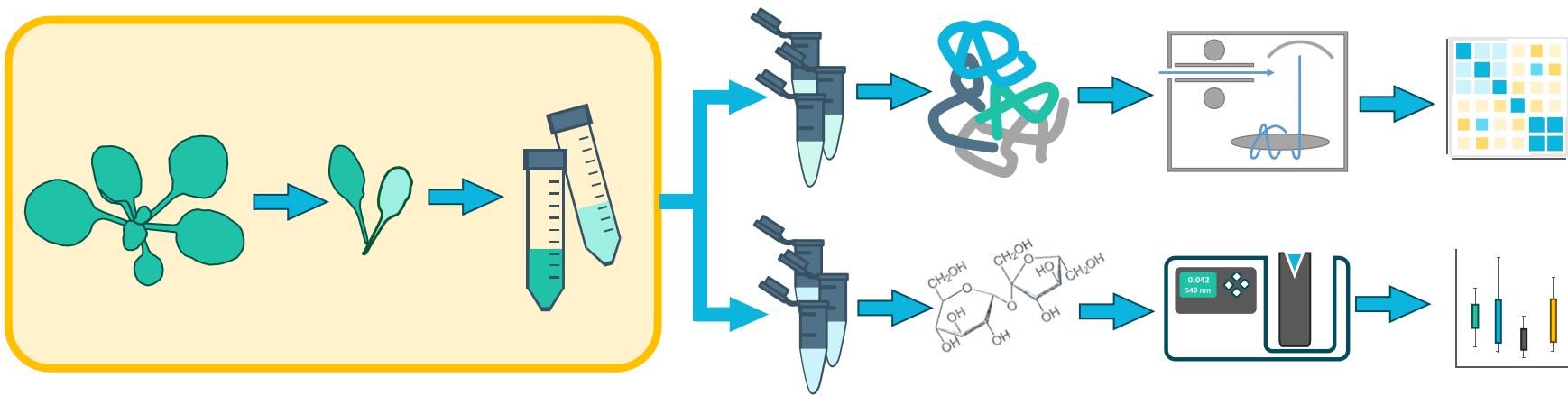
# The ARC scaffold structure



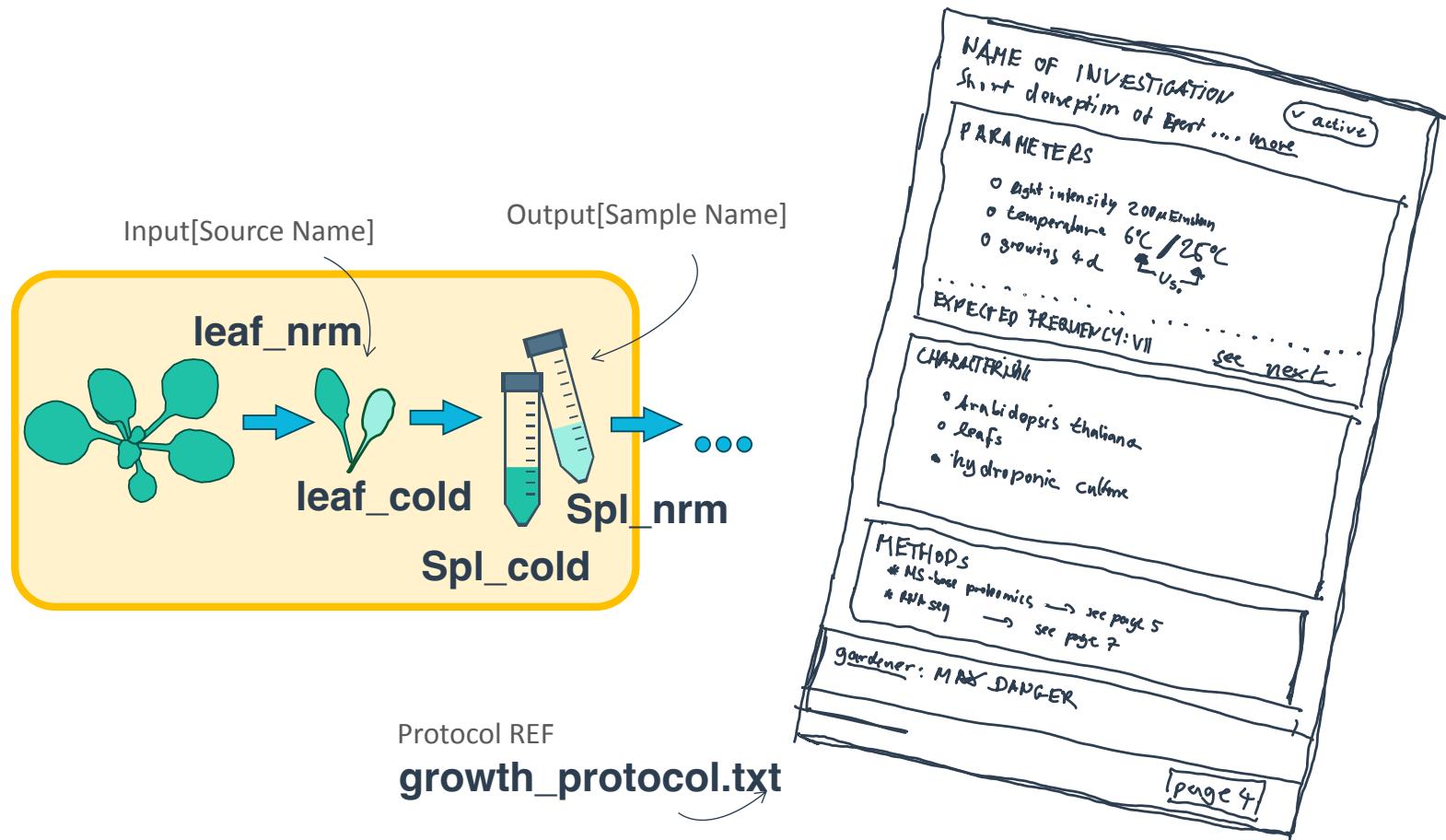
# A small prototypic project



# Divide and conquer for reproducibility



# Identifying the ‘study’ part



## A table-based organization schema

Input[Source Name]		Output[Sample Name]
leaf_nrm		spl_nrm
leaf_cold		spl_cold
A	B	C
D		

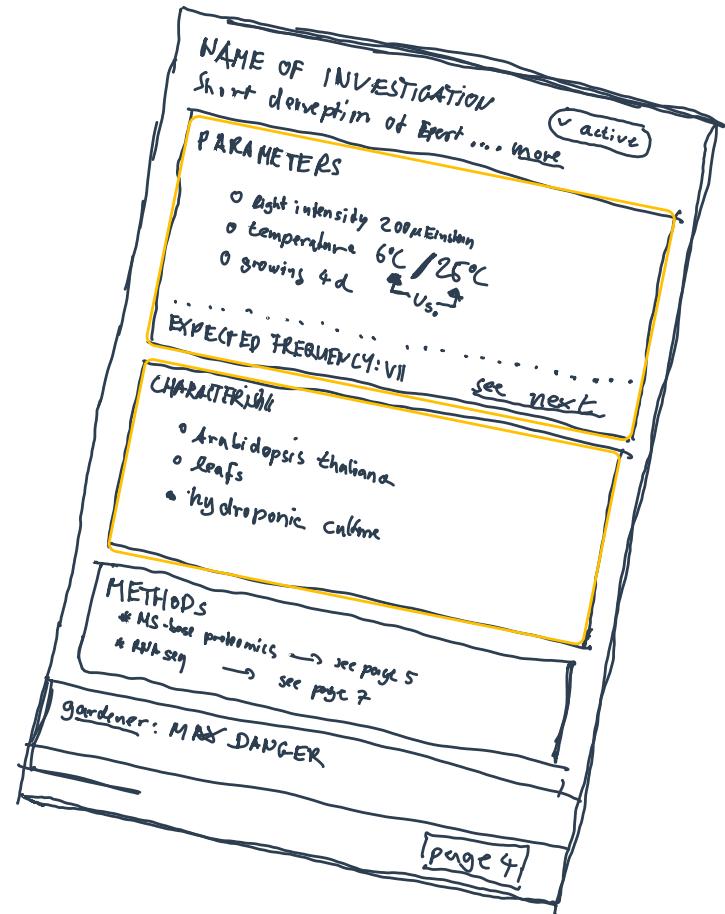
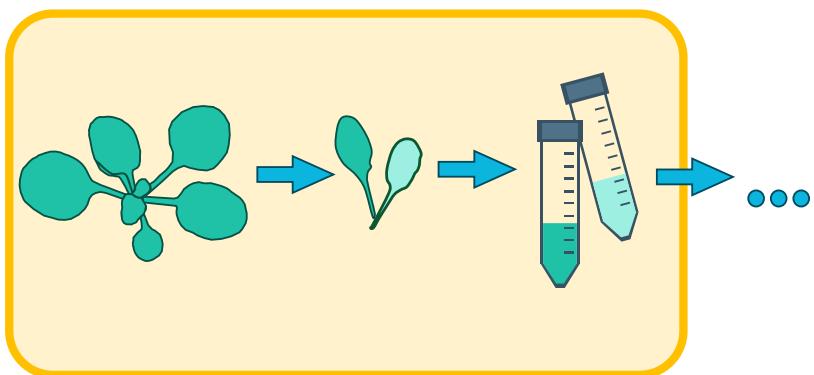
# Referencing a protocol

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

Input[Source Name]	Protocol REF	Output[Sample Name]
leaf_nrm	growth_protocol.txt	spl_nrm
leaf_cold	growth_protocol.txt	spl_cold
A	B	C
D		

# 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](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: UNIPROT RDFS

## temperature

### AFO:/result#AFR\_0001584



[http://purl.allotrope.org/ontologies/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](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](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

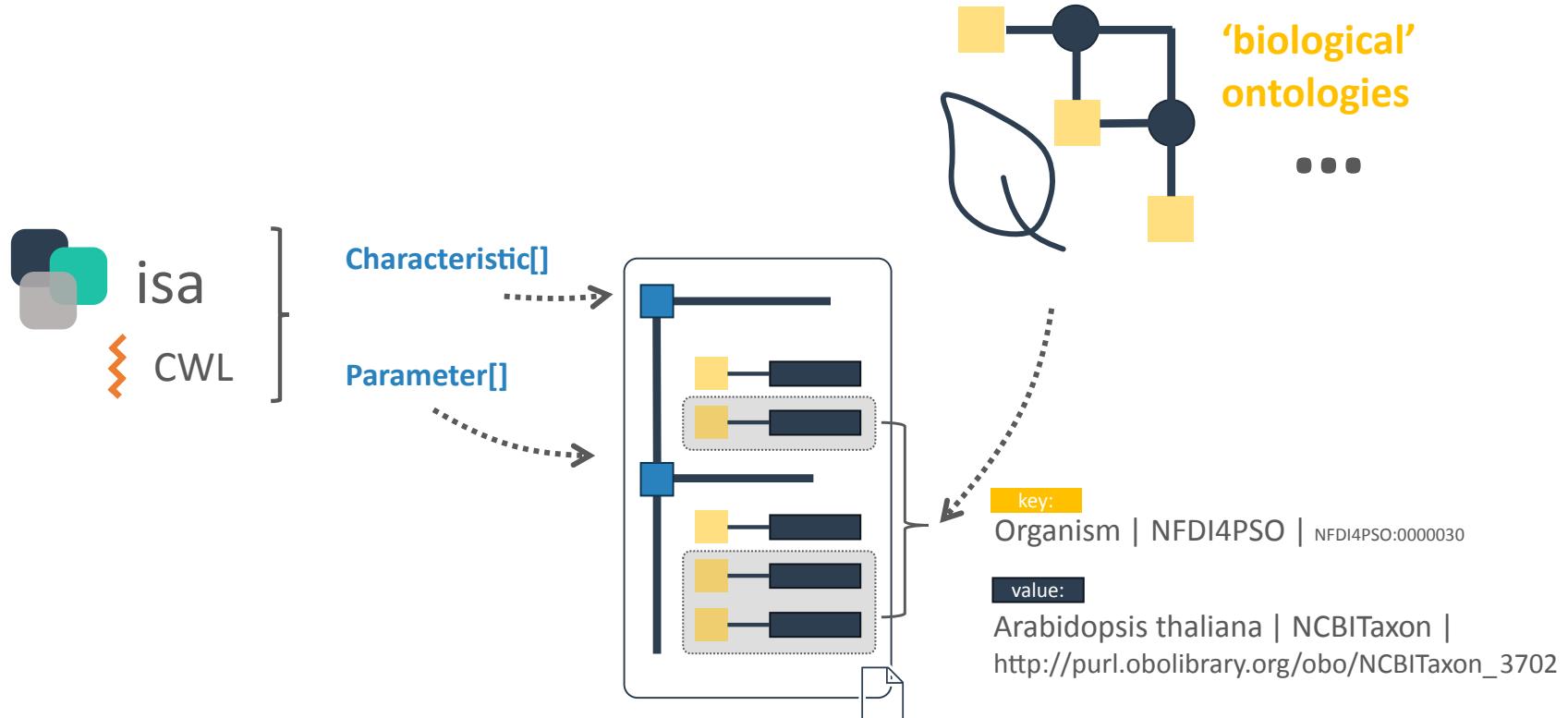
TXPO

MCO

+

Ontology Lookup Service (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

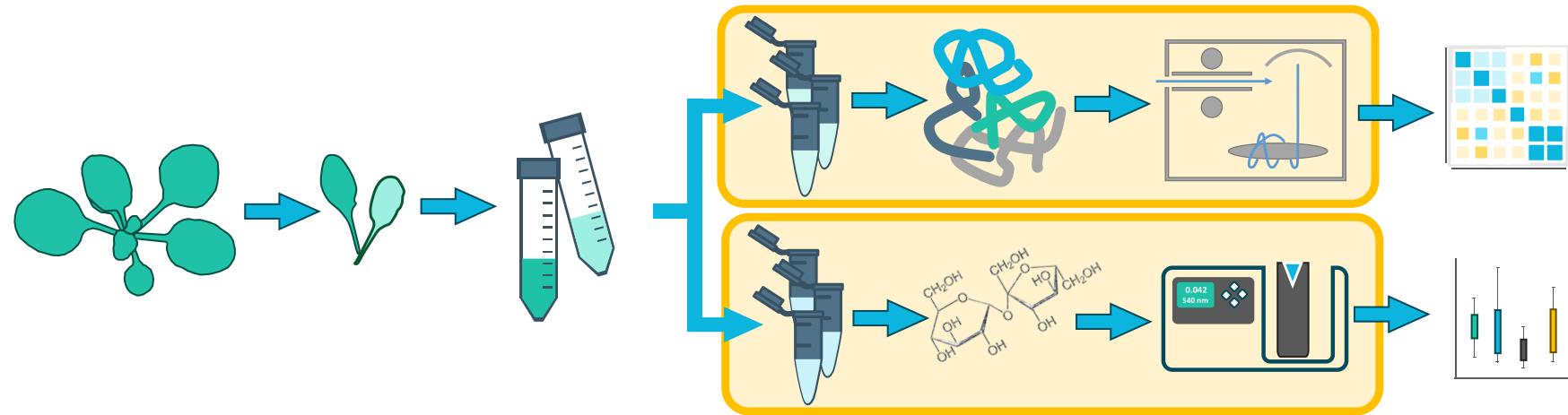
## Hands-on part 1: Setup and ARCitect

Follow the [Start Here guide](#) in the DataPLANT knowledge base.

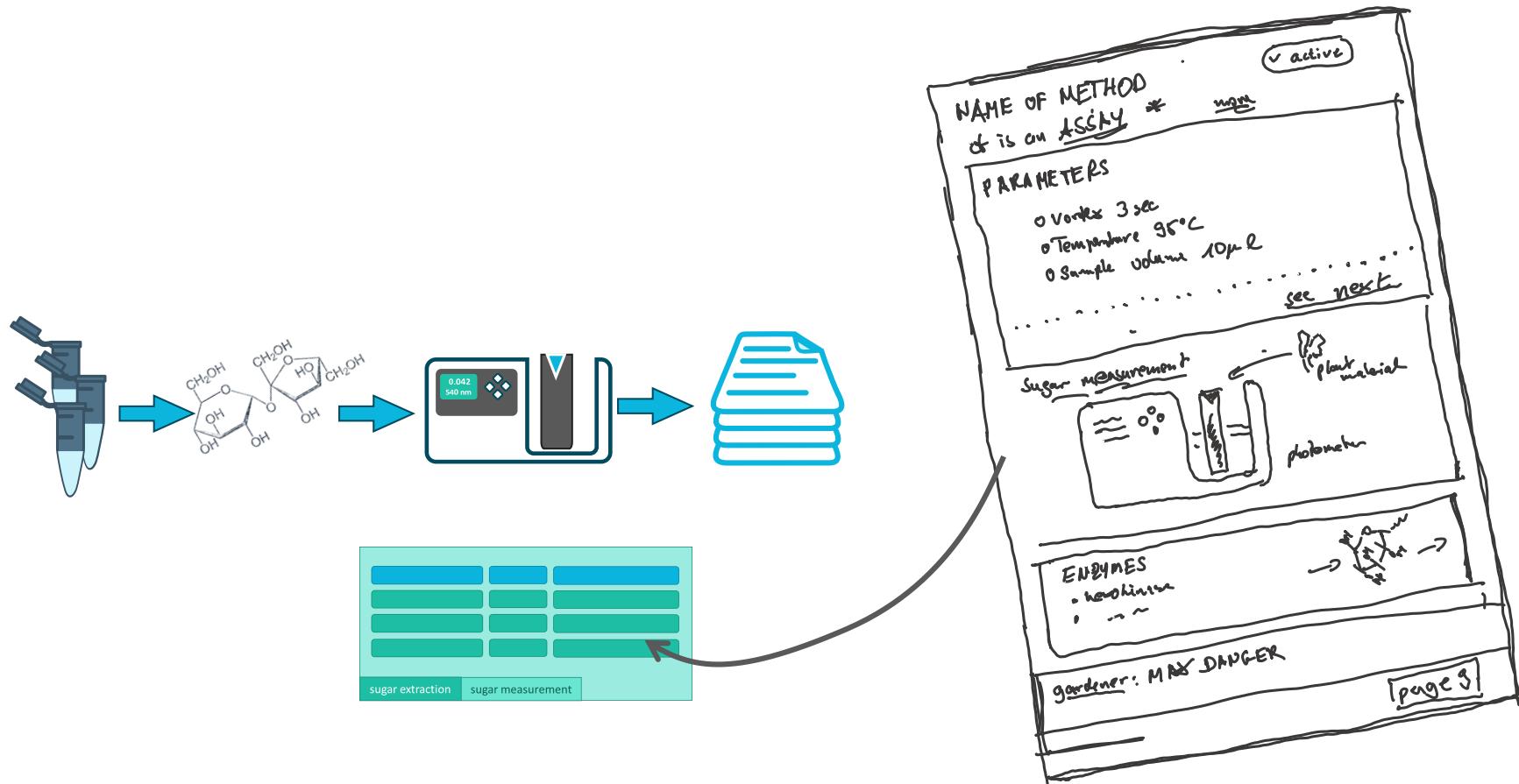
- 💡 Until step Add a study

## Hands-on part 2: ARCitect (and Swate)

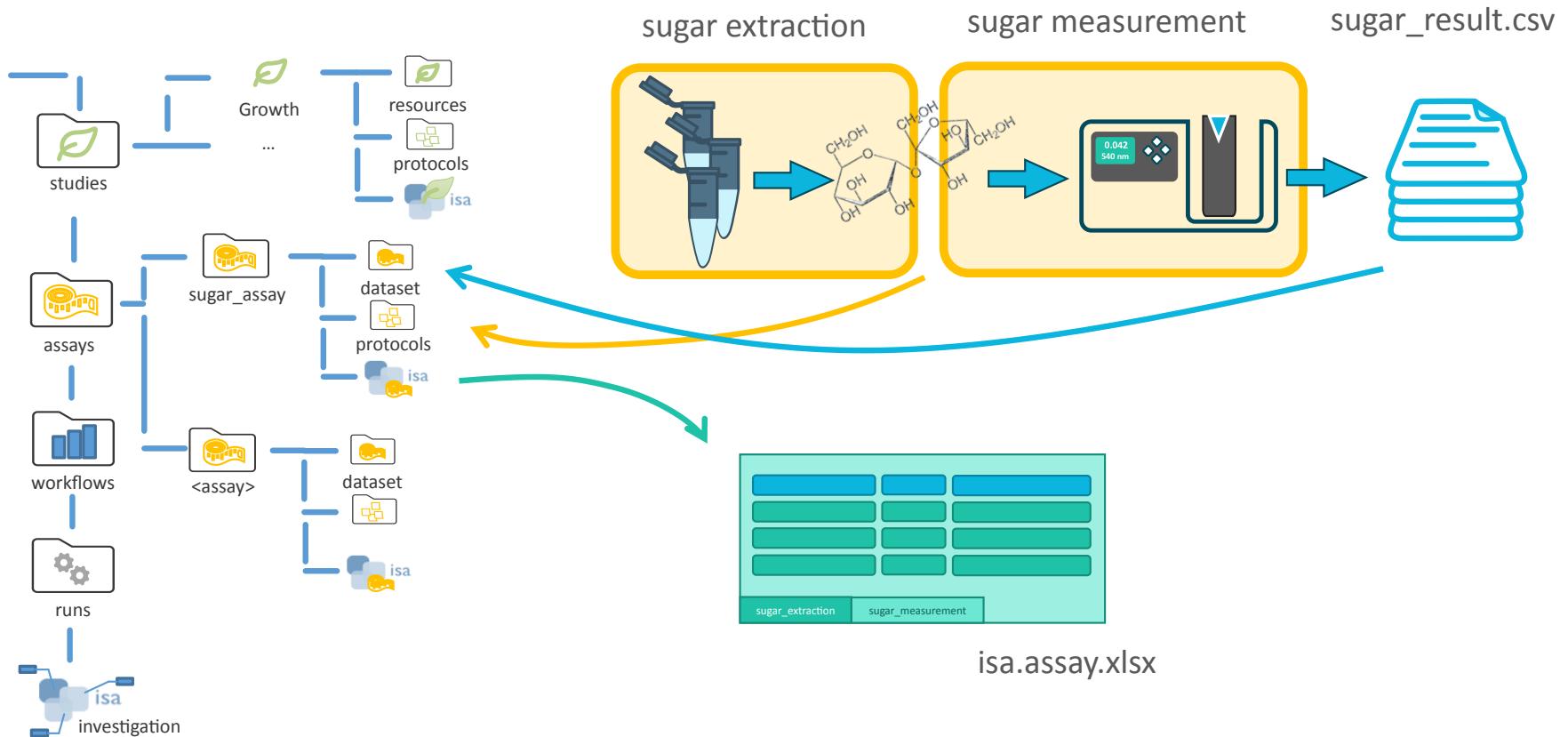
# Identifying assays



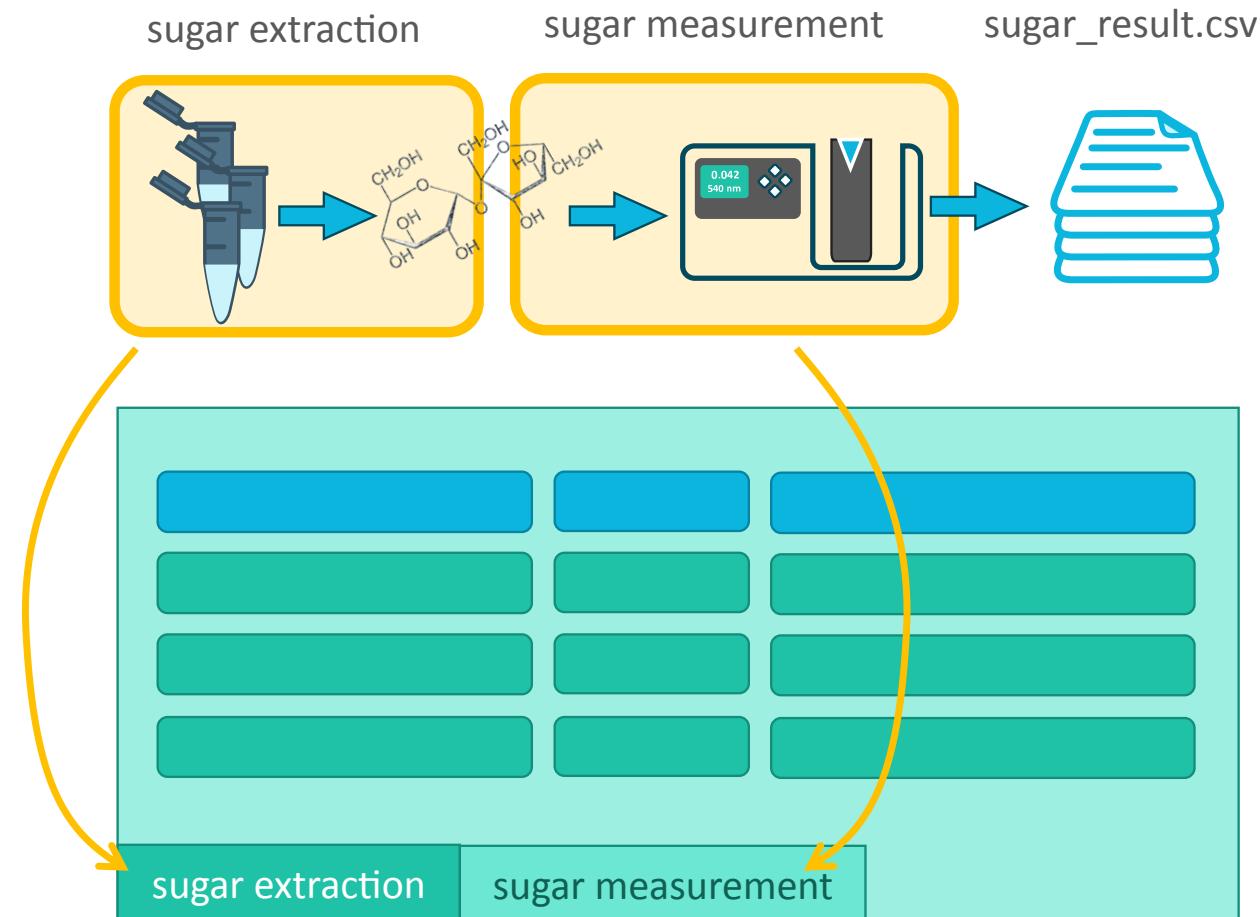
# Assay for sugar measurement



# Separating different assay elements



# Isolating the lab processes in an assay



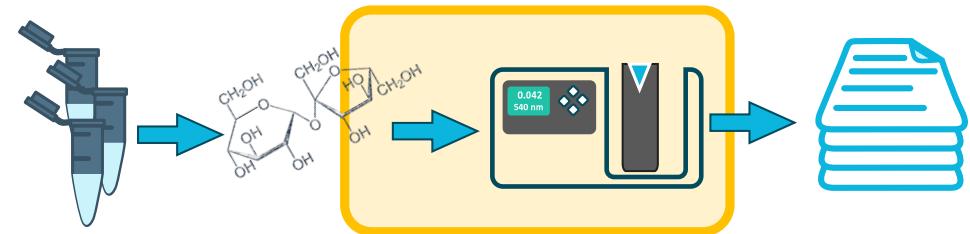
# Parameteterization: sugar extraction

- 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



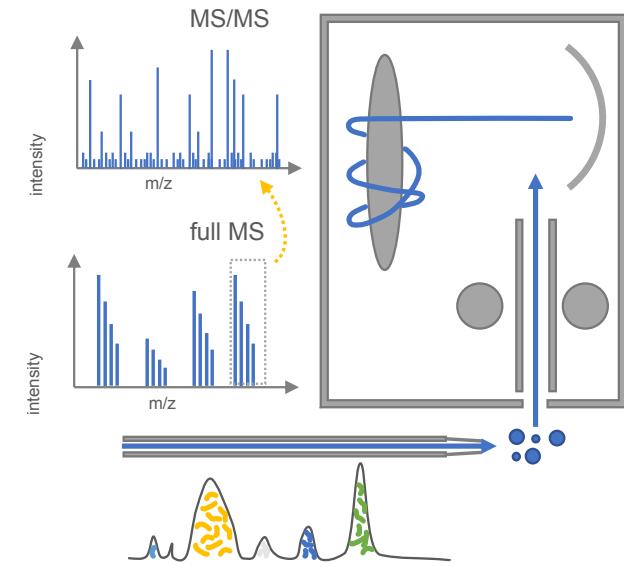
# Save time using standard methods and SOPs

## Parameter []

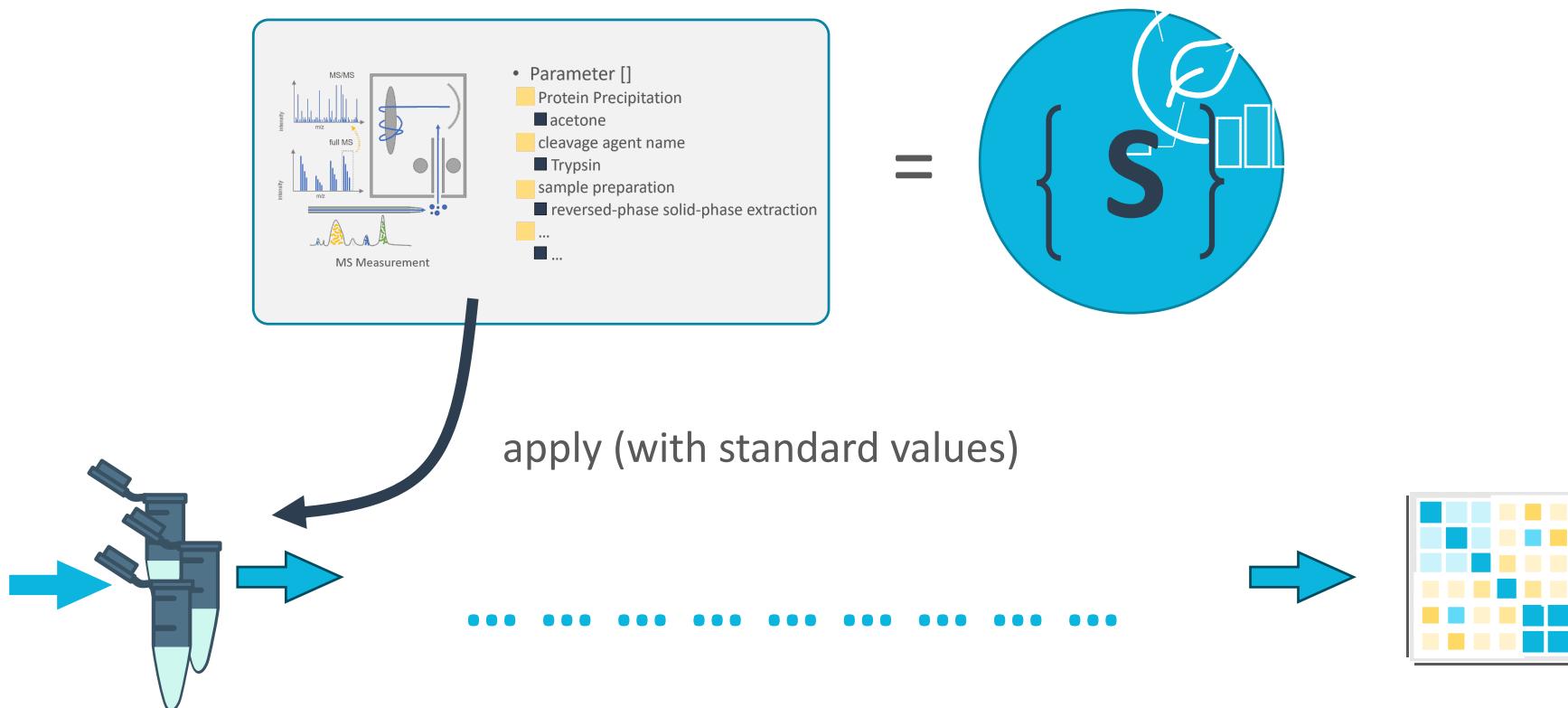
- Protein Precipitation
  - acetone
- cleavage agent name
  - Trypsin
- sample preparation
  - reversed-phase solid-phase extraction
- ...

## Component []

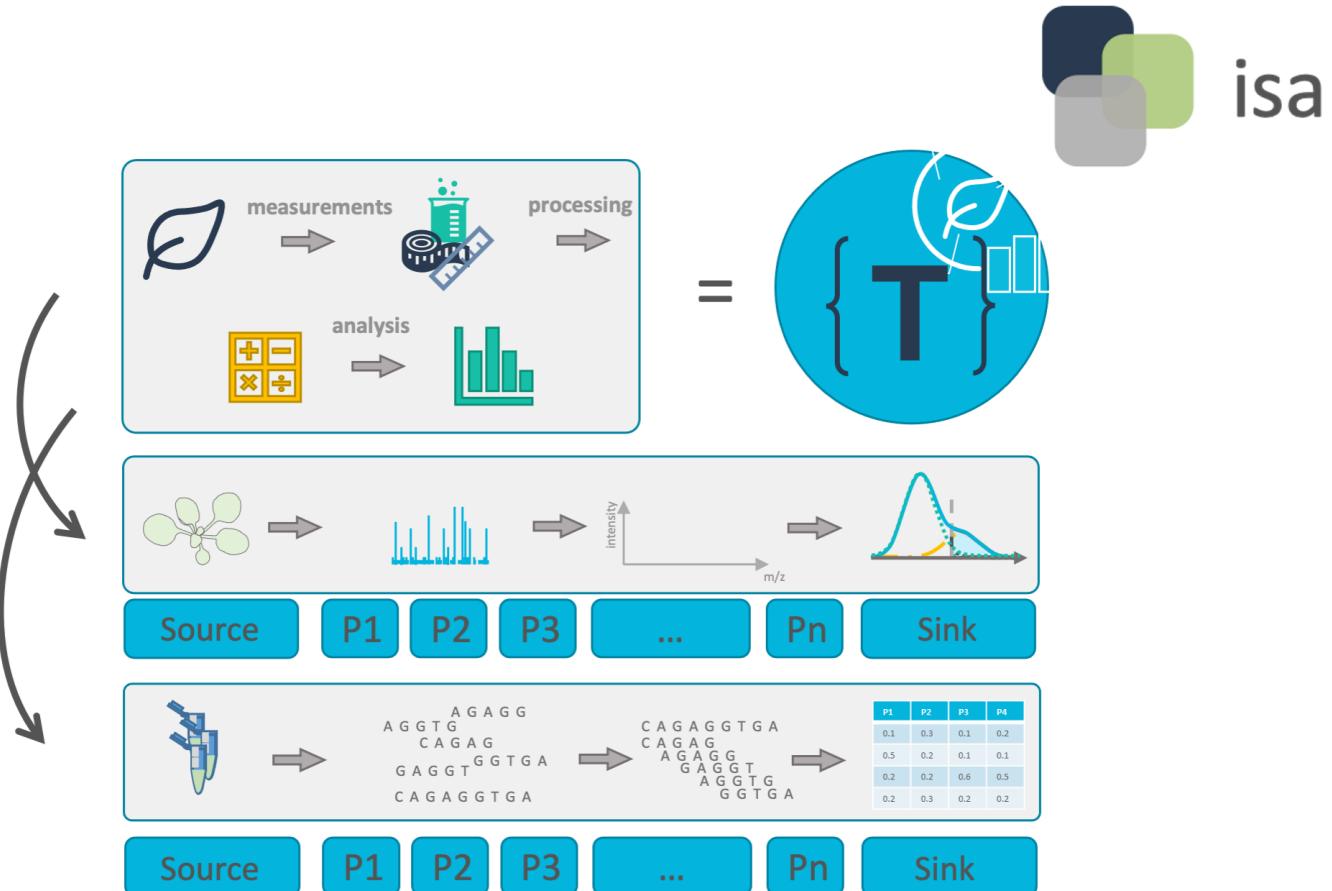
- chromatography instrument model
  - nanoElute2
- chromatography column model
  - PepSep C18 1.9u, 25cm x 75um ...



# Applying standard procedures to sample record



# Realization of lab-specific metadata with templates



## Hands-on part 2: ARCitect (and Swate)

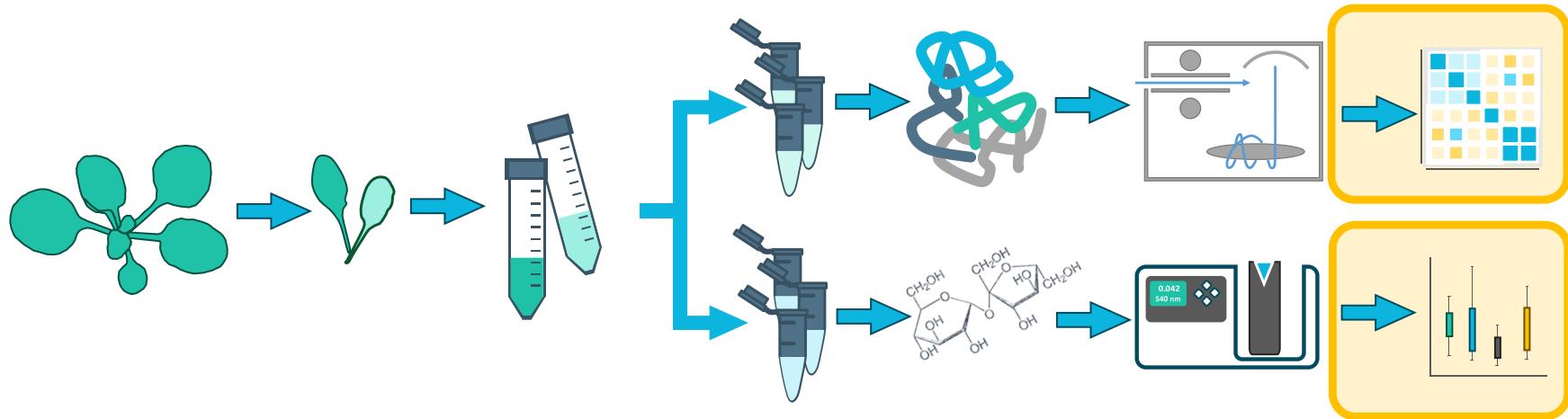
Continue the [Start Here guide](#) in the DataPLANT knowledge base.



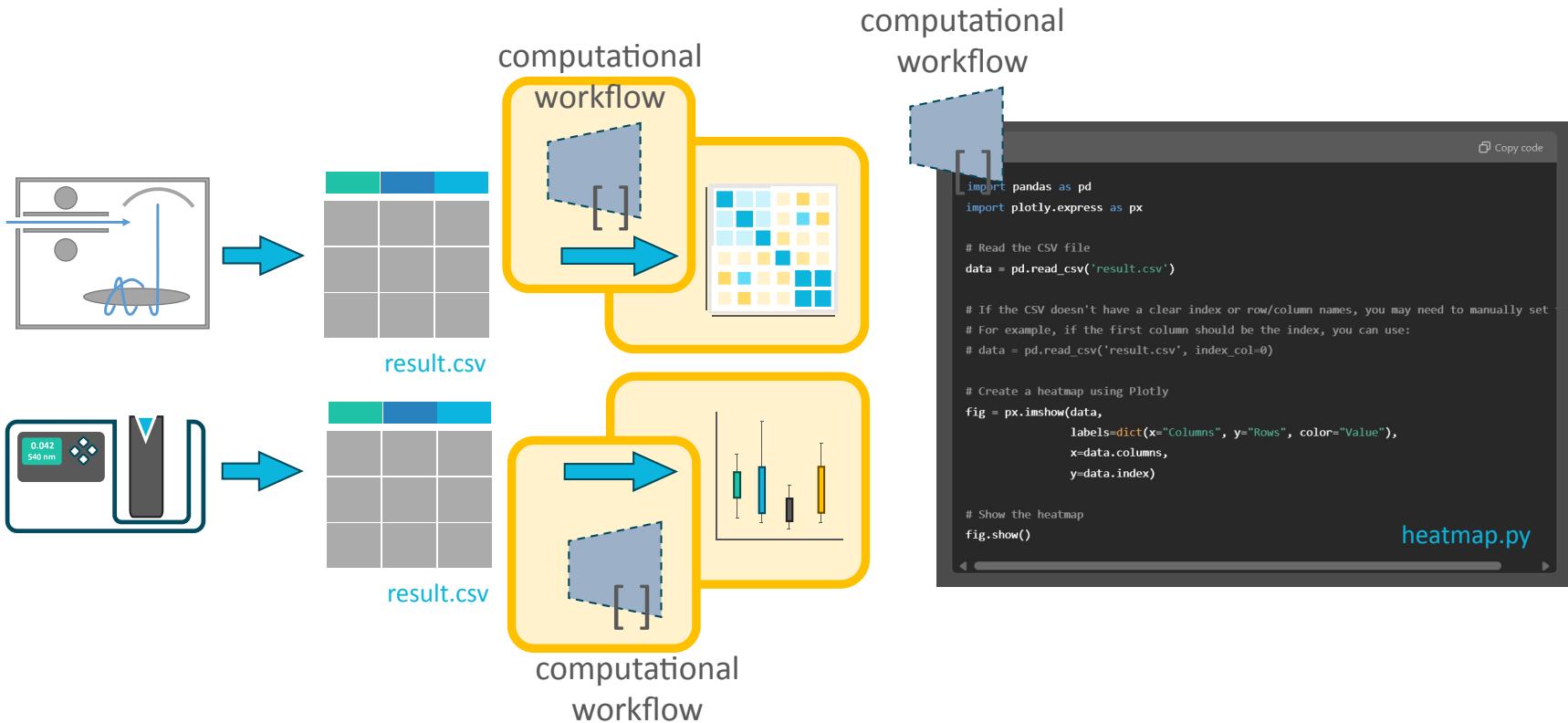
Until step Add assay data

## Hands-on part 3: Data

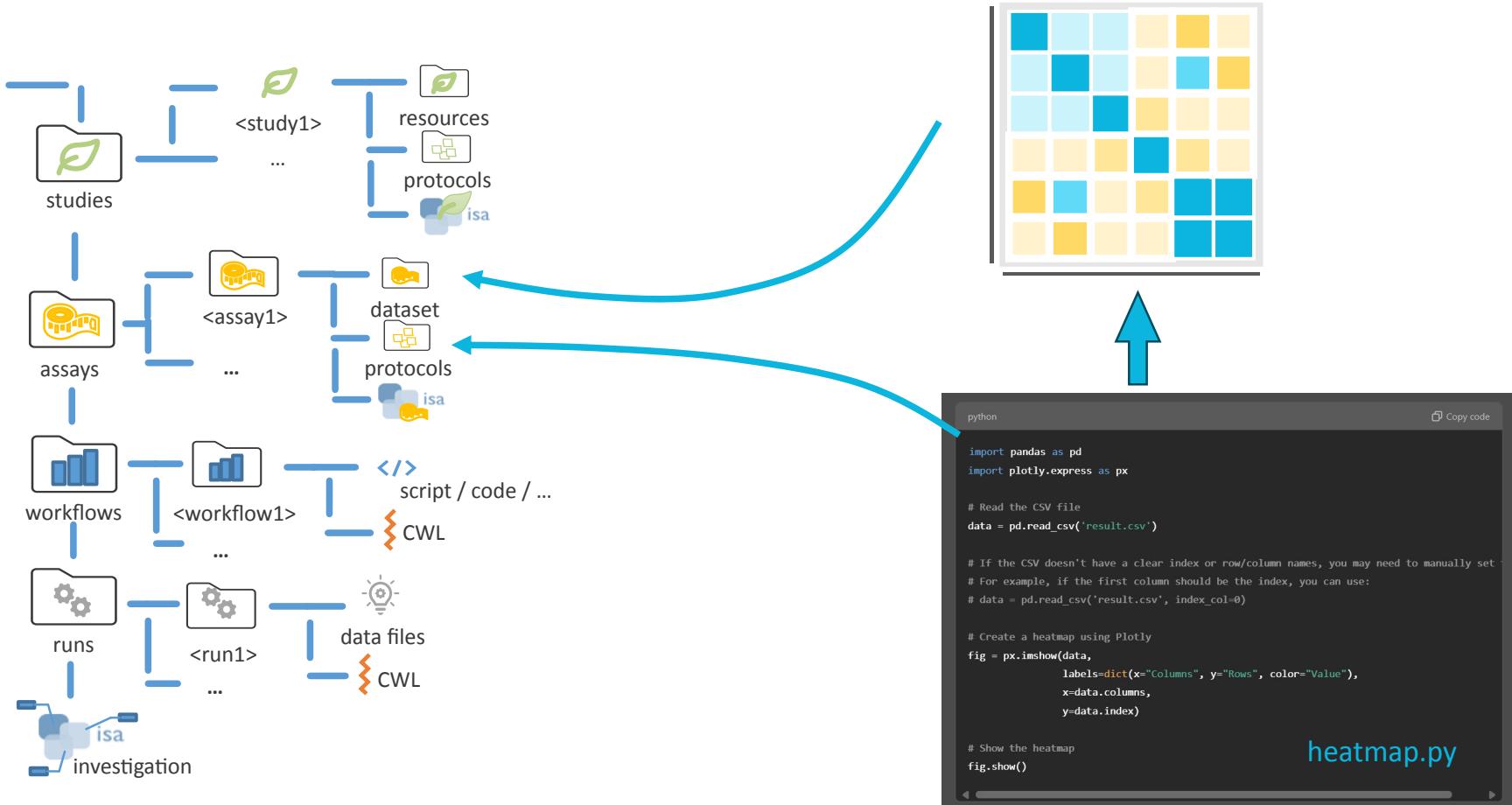
# Options to annotate the data analysis



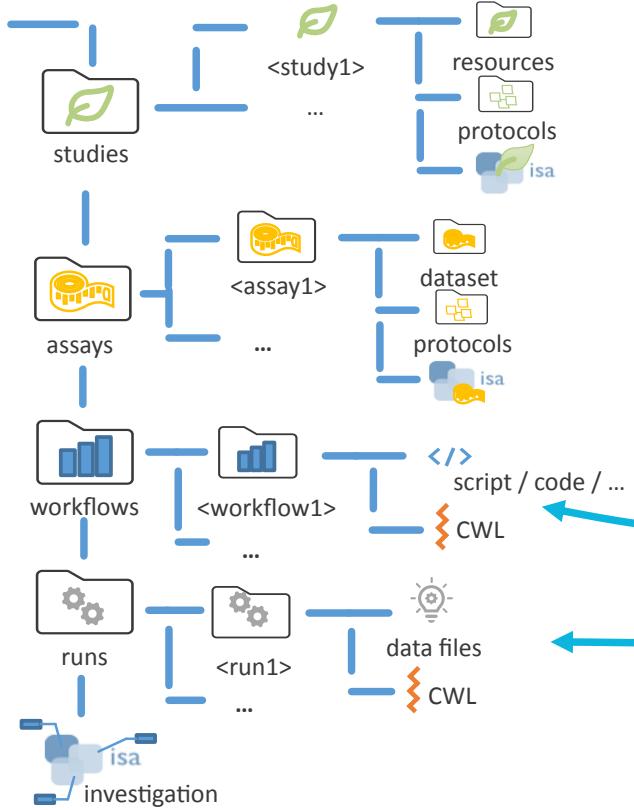
# A computational workflow is like a protocol



# Option I: Create a virtual assay



## Option II: Create a workflow and run



A screenshot of a code editor window titled "heatmap.py" containing the following Python code:

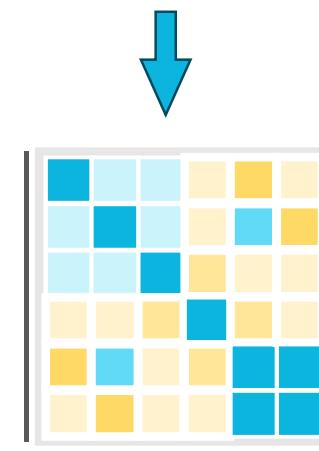
```
python
import pandas as pd
import plotly.express as px

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

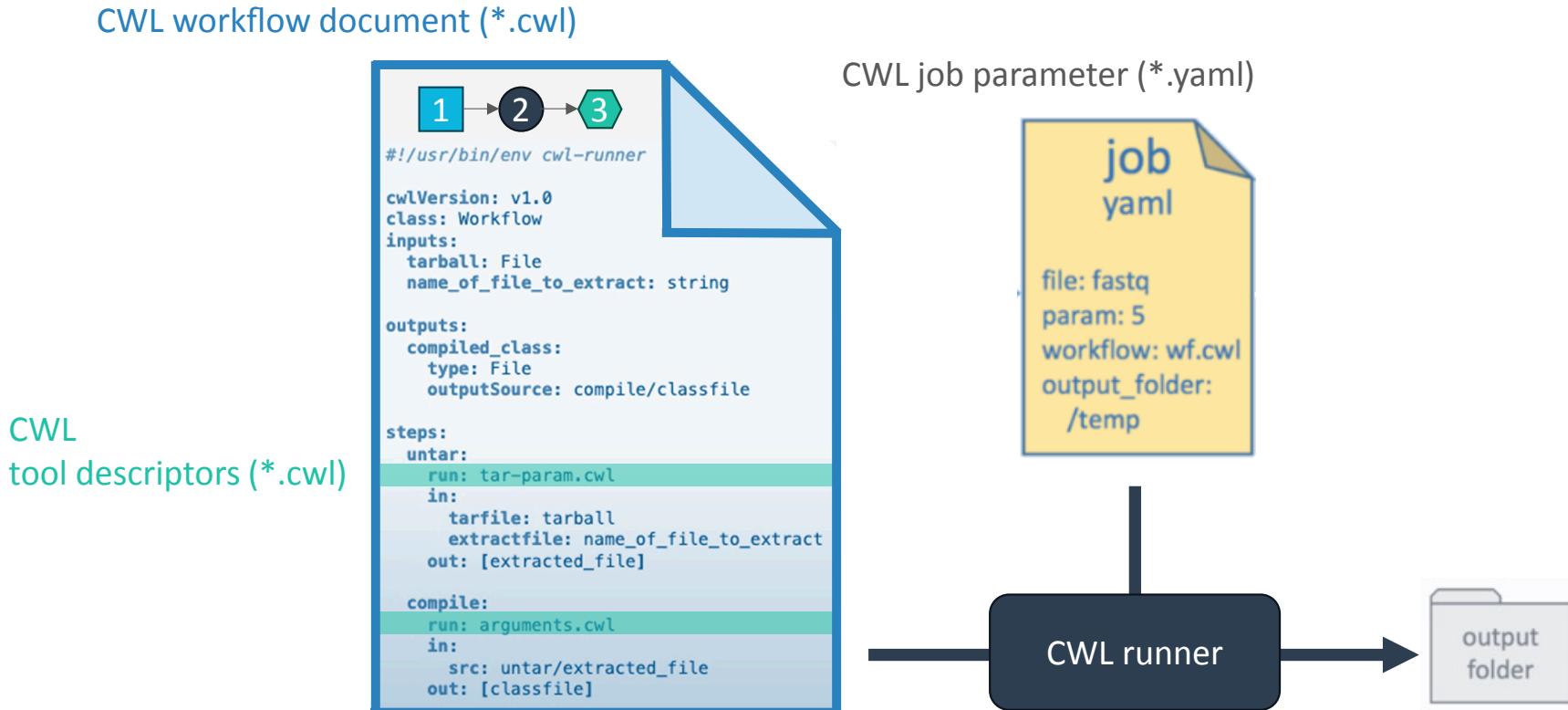
# If the CSV doesn't have a clear index or row/column names, you may need to manually set
# For example, if the first column should be the index, you can use:
# data = pd.read_csv('result.csv', index_col=0)

# 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()
```



# Use CWL to wrap your workflow



## 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

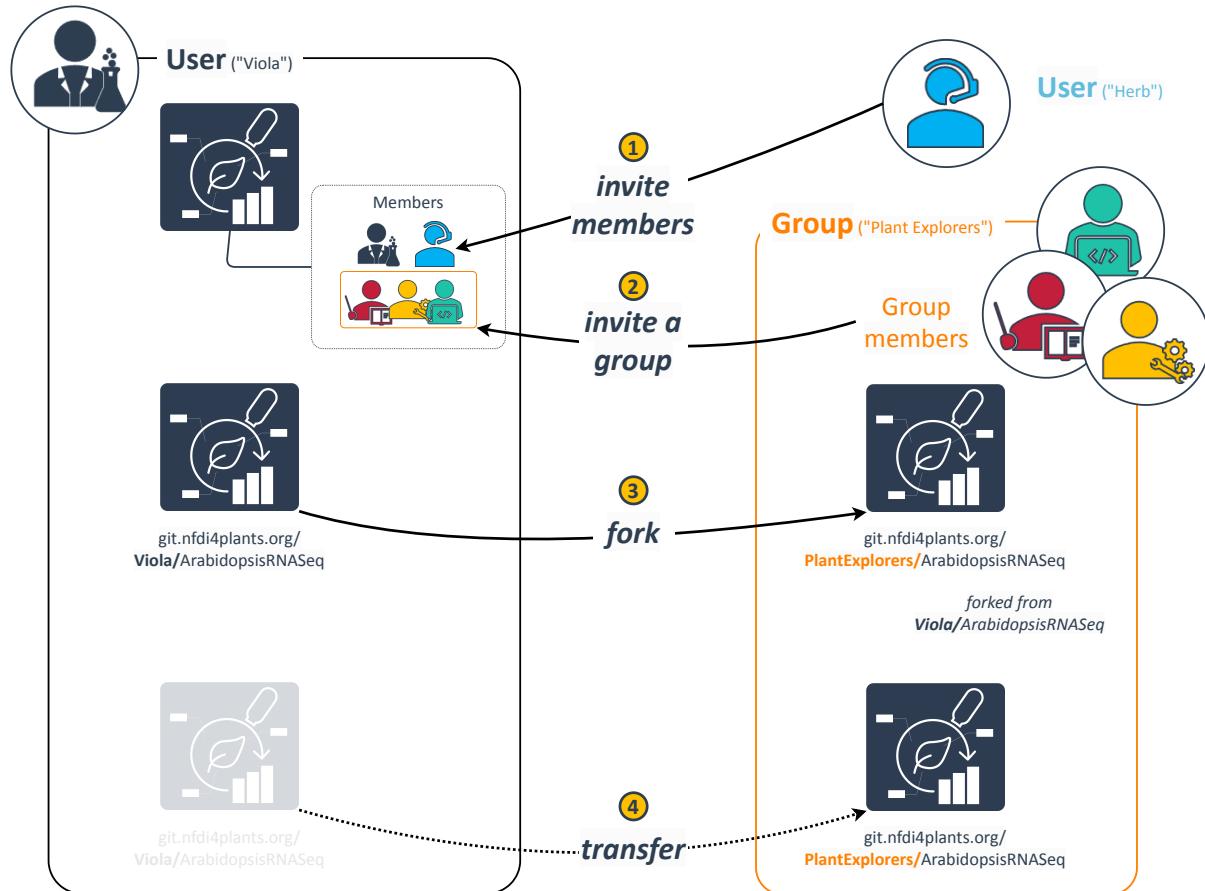
# Project = ARC

- 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 <b>personal</b> ARC	<a href="https://git.nfdi4plants.org/brilator/Facultative-CAM-in-Talinum">https://git.nfdi4plants.org/brilator/Facultative-CAM-in-Talinum</a>	brilator	Dominik Brilhaus
An <b>group-shared</b> ARC	<a href="https://git.nfdi4plants.org/hhu-plant-biochemistry/Samuilov-2018-BOU-PSP">https://git.nfdi4plants.org/hhu-plant-biochemistry/Samuilov-2018-BOU-PSP</a>	hhu-plant-biochemistry	HHU Plant Biochemistry

# Visibility

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

## Visibility

The visibility of each ARC can be managed in the settings of the ARC



**Private** – ARC access must be granted explicitly to each user or group.



**Internal** – ARC can be accessed by any logged in user.



**Public** – ARC can be accessed without authentication.

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

# ARC DataHUB members // ARC Investigation contacts

The screenshot shows the 'Project members' section of the DataHUB interface. It includes a search bar, navigation links for 'Import from a project', 'Invite a group', and 'Invite members'. Below this, there are tabs for 'Members' (18) and 'Groups' (1). A filter bar allows users to 'Filter members'. The main table lists four members:

Account	Source	Max role	Expiration	Activity
Adriano Nunes-Nesi @nunesnesi	HHU Plant Biochemistry by Sebastian Triesch	Maintainer	Expiration date	User created: Jul 05, 2023 Access granted: Jul 10, 2023 Last activity: Aug 09, 2023
Andreas Weber @andreas.weber	HHU Plant Biochemistry by Sebastian Triesch	Maintainer	Expiration date	User created: Mar 10, 2023 Access granted: Jul 31, 2023 Last activity: Sep 11, 2023
Dominik Brilhaus It's you @brilator	Direct member by Dominik Brilhaus	Owner	Expiration date	User created: Feb 21, 2022 Access granted: Dec 06, 2023 Last activity: Mar 26, 2024
Franziska Fichtner @franziska.fichtner	HHU Plant Biochemistry by Sebastian Triesch	Maintainer	Expiration date	User created: Aug 11, 2023 Access granted: Aug 11, 2023 Last activity: Aug 11, 2023

DataHUB: ARC members

[https://git.nfdi4plants.org/hhu-plant-biochemistry/Samuilov-2018-BOU-PSP/-/project\\_members](https://git.nfdi4plants.org/hhu-plant-biochemistry/Samuilov-2018-BOU-PSP/-/project_members)

The screenshot shows the 'ARCIctect' interface for the 'Samuilov-2018-BOU-PSP' project. It includes a sidebar with options like 'Login', 'New ARC', 'Open ARC', 'Download ARC', 'Save ARC', 'Explorer', 'Commit', 'DataHUB Sync', and 'History'. The main area shows the project's identifier as 'Samuilov-2018-BOU-PSP' and its title. There is a 'Description' field and a 'Contacts' section listing several individuals with their ORCID IDs and scores:

Contact	Score
Sladjana Samuilov <orcid>	4/10
Nadine Rademacher <orcid>	3/10
Samantha Flachbart <orcid>	3/10
Leila Arab <orcid>	3/10
Saleh Alfarraj <orcid>	3/10
Franziska Kuhnert <orcid>	3/10
Stanislav Kopriva <orcid>	3/10
Andreas P. M. Weber <orcid>	4/10
Tabea Mettler-Altmann <orcid>	3/10

ARCIctect: 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)

The screenshot shows the Data PLANT CEPLAS interface with the following numbered callouts:

- Manage
- Plan
- Code
- Build
- Secure
- Deploy
- Operate
- Monitor
- Analyze
- Settings
- Help

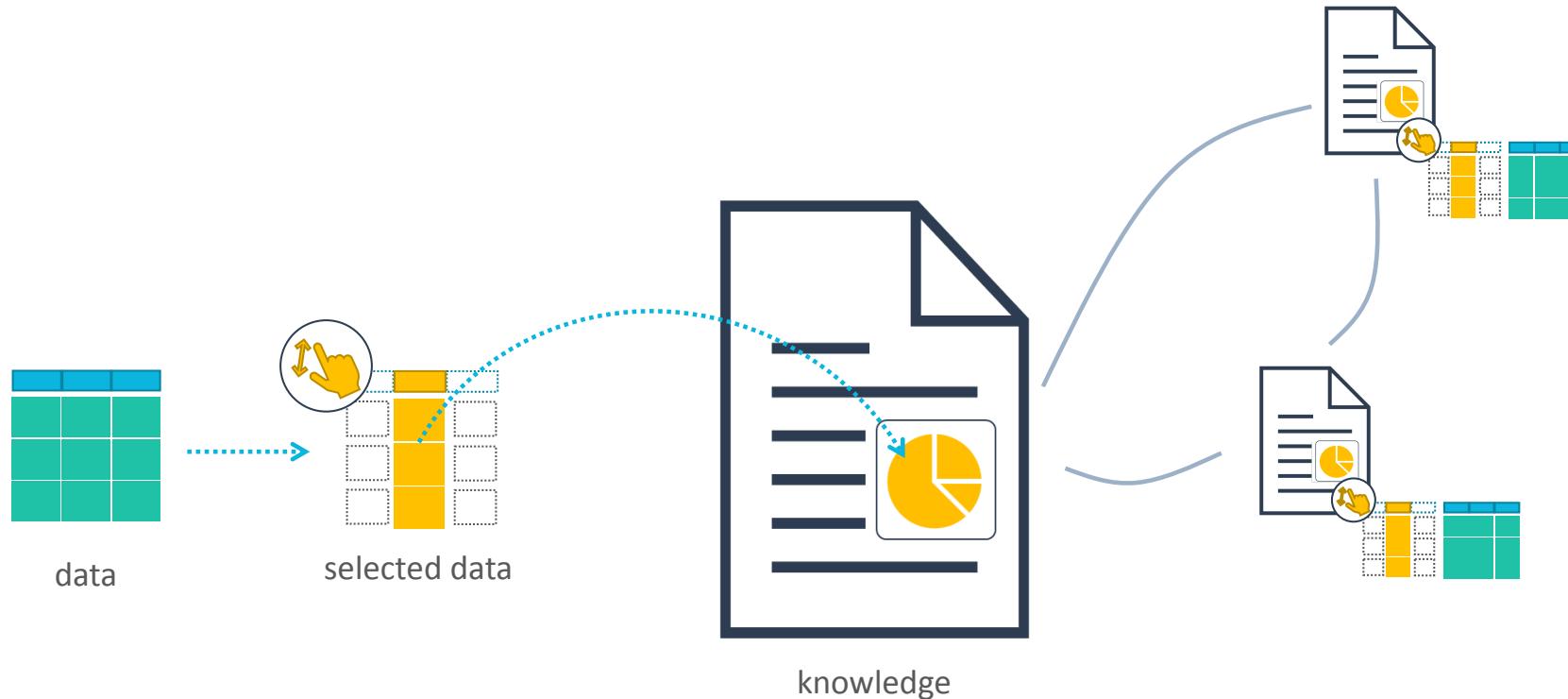
Key UI elements include:

- D Demo\_ARC**: Project name in the header.
- Code**: Selected tab in the top navigation bar.
- History**: Tab in the top navigation bar.
- Find file**: Tab in the top navigation bar.
- Edit**: Tab in the top navigation bar.
- Project information**: Pipeline status (passed), Star count (0), Fork count (0), and a three-dot menu.
- 1 Commit**: Summary of commit activity.
- 2 Branches**: Summary of branch activity.
- 0 Tags**: Summary of tag activity.
- 7 KIB Project Storage**: Project storage usage.
- Auto DevOps enabled**: Project status.
- Add README**, **Add LICENSE**, **Add CHANGELOG**, **Add CONTRIBUTING**, **Add Kubernetes cluster**, **Add Wiki**, **Configure Integrations**: Project configuration options.
- Created on**: Project creation date (July 13, 2024).

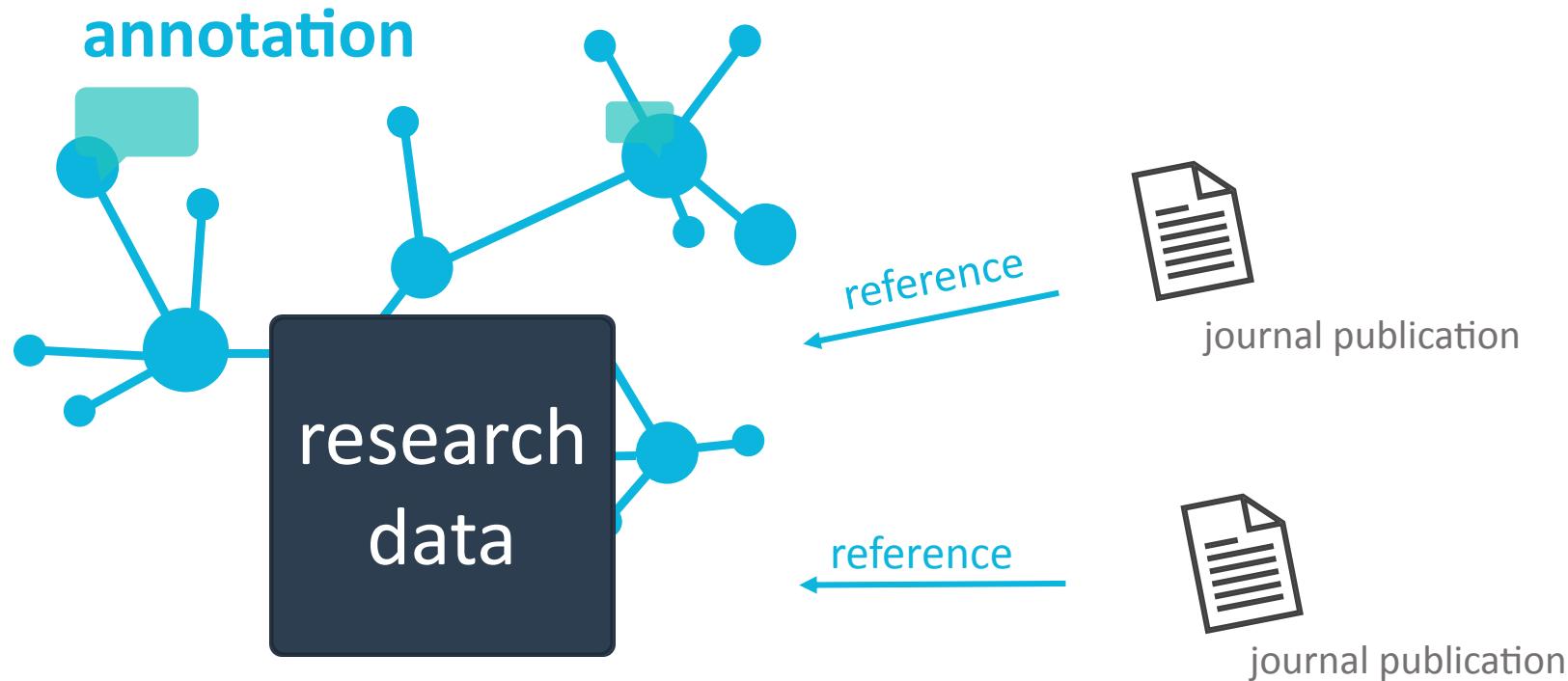
The main content area displays a list of files and their last commit details:

Name	Last commit	Last update
assays	arc init	4 minutes ago
runs	arc init	4 minutes ago
studies	arc init	4 minutes ago
workflows	arc init	4 minutes ago
.gitignore	arc init	4 minutes ago
isa.investigation.xlsx	arc init	4 minutes ago

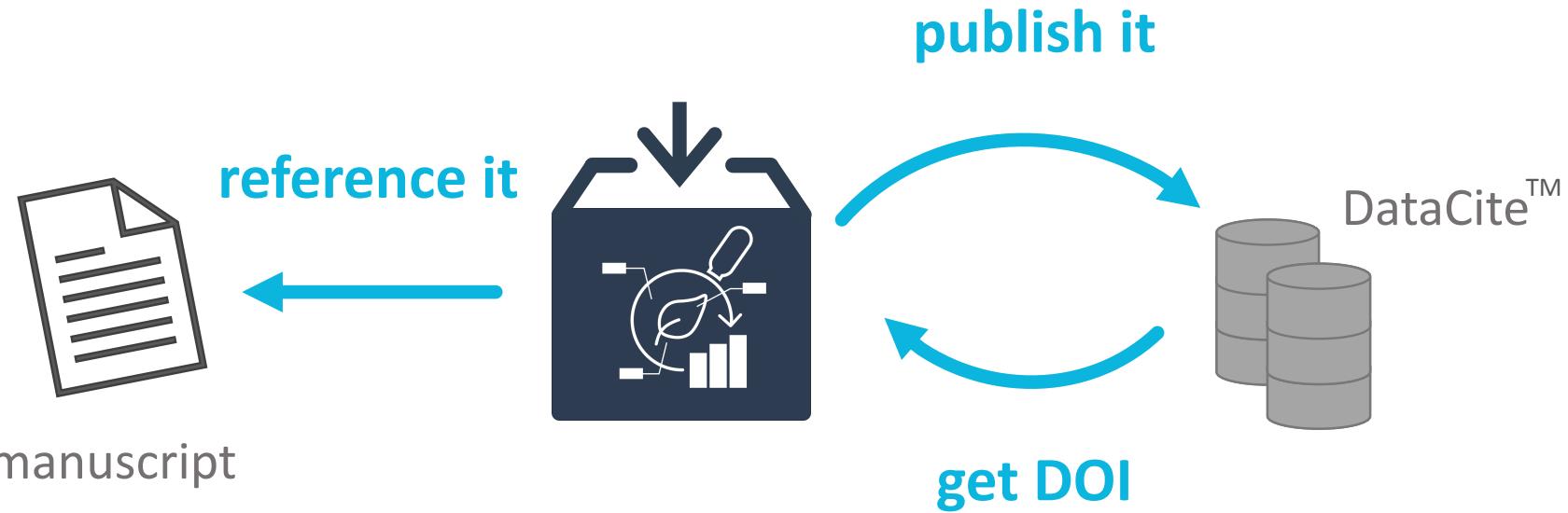
# Hands-on part 5: ARC data publication



# Moving from paper to data publications



# Publish your ARC, get a DOI



# Publish your ARC with a few clicks



## Ru\_ChlamyHeatstress

Project ID: 122

24 h  
48 h

Star 1

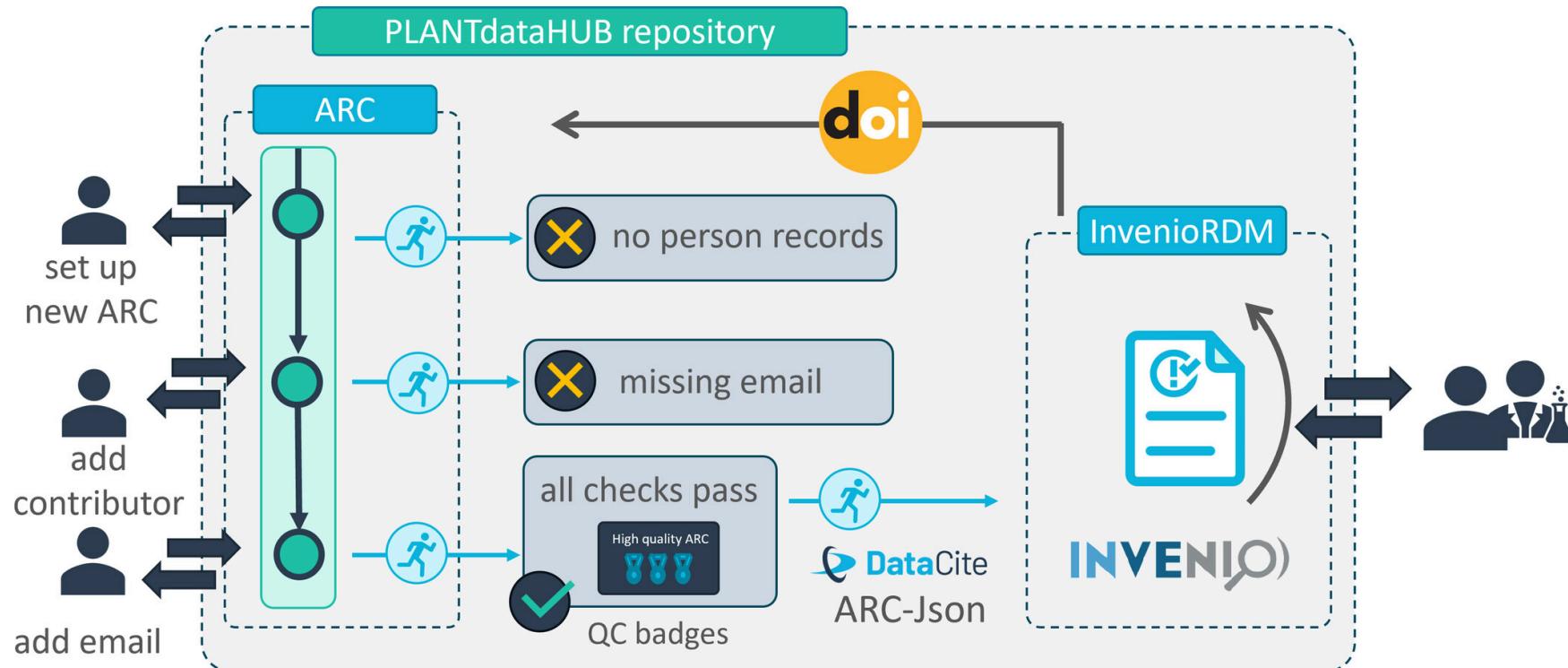
-o 53 Commits 1 Branch 0 Tags 293.9 GB Project Storage

Topics: Chlamydomonas, abiotic stress, proteomics + 1 more

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.

1 pipeline passed 2 Publish ARC 3 arc quality 301/301

# 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** 

[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](https://doi.org/10.60534/9e5jx-75d83)

[Star](#) 2 [⋮](#)

**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   invenio 10/10

-o- 55 Commits  
2 Branches  
0 Tags

 README  
Auto DevOps enabled

**Created on**  
July 11, 2022

# Receive a DOI

Published September 7, 2023 | Version v1

Dataset 

## Systems-wide investigation of responses to moderate and acute high temperatures in the green alga *Chlamydomonas reinhardtii*.

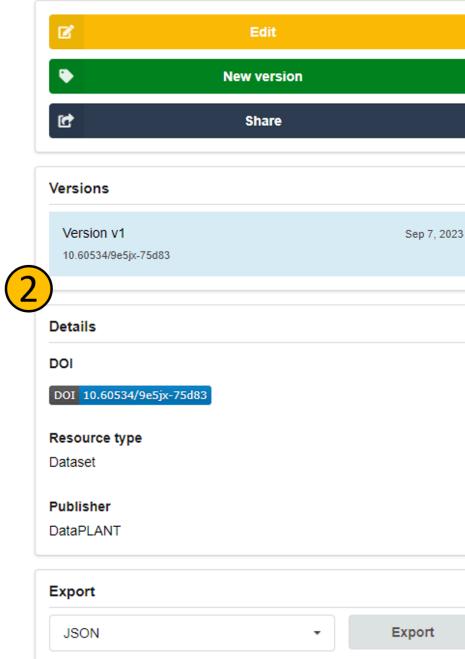
Zhang, Ningning<sup>1</sup>; Mattoon, Erin<sup>1</sup>; McHargue, Will<sup>1</sup> ; Venn, Benedikt<sup>2</sup> ; Zimmer, David<sup>2</sup> ; Pecani, Kresti<sup>3</sup>; Jeong, Jooyeon<sup>1</sup>; Anderson, Cheyenne<sup>4</sup>; Chen, Chen<sup>4</sup>; Berry, Jeffrey<sup>1</sup>; Xia, Ming<sup>1</sup>; Tzeng, Shin-Cheng<sup>1</sup> ; Becker, Eric<sup>1</sup>; Pazouki, Leila<sup>1</sup>; Evans, Bradley<sup>1</sup>; Cross, Fred<sup>3</sup>; Cheng, Jianlin<sup>4</sup>; Czymmek, Kirk<sup>1</sup> ; Schröder, Michael<sup>5</sup> ; Mühlhaus, Timo<sup>2</sup> ; Zhang, Ru<sup>1</sup> 

Show affiliations

Style

APA

1  
2



The screenshot shows a dataset page with various sections and a sidebar. The main content includes a citation, a description, and a file section. The sidebar contains buttons for 'Edit', 'New version', and 'Share', along with sections for 'Versions', 'Details', 'Resource type', 'Publisher', and 'Export'.

**Citation**

Zhang, N., Mattoon, E., McHargue, W., Venn, B., Zimmer, D., Pecani, K., Jeong, J., Anderson, C., Chen, C., Berry, J., Xia, M., Tzeng, S.-C., Becker, E., Pazouki, L., Evans, B., Cross, F., Cheng, J., Czymmek, K., Schröder, M., ... Zhang, R. (2023). Systems-wide investigation of responses to moderate and acute high temperatures in the green alga *Chlamydomonas reinhardtii*. [Data set]. DataPLANT. <https://doi.org/10.60534/9e5jx-75d83>

**Description**

hosted on: <https://git.nfd4plants.org/projects/122>

**Files**

arc-summary.md

[Data set] Systems-wide investigation of responses to moderate and acute high temperatures in the green alga *Chlamydomonas reinhardtii*.

**File contents:**

- root
  - isa.investigation.xlsx
  - README.md
  - runs

**Details**

DOI: [10.60534/9e5jx-75d83](https://doi.org/10.60534/9e5jx-75d83)

Resource type: Dataset

Publisher: DataPLANT

Export: JSON 

## 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/>

## Tools and Services

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

## 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/>

## Open Source Development

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

# Acknowledgements



## Team Kaiserslautern

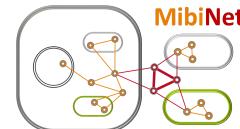
- Timo Mühlhaus
- Lukas Weil
- Kevin Frey
- Kevin Schneider
- Jonas Lukasczyk

## Team Freiburg

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



- Björn Usadel
- Vittorio Tracanna
- Yaser Alashloo



- Sabrina Zander