

ARCify your research project

Plant2030 Spring School – Introduction to bioinformatics

Dominik Brilhaus, [CEPLAS](#)

May 15th, 2025



Training Materials

- Use the pad to raise questions and feedback
- Provide links for hands-on



Introduce yourself

- Used code / programming language before?
- Experience with Git / GitLab / GitHub?
- Operating system: Linux / Windows / Mac?

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>

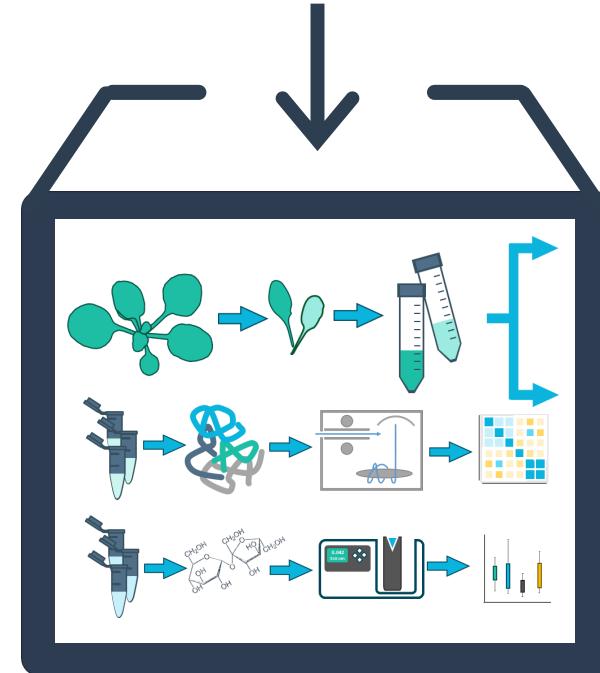
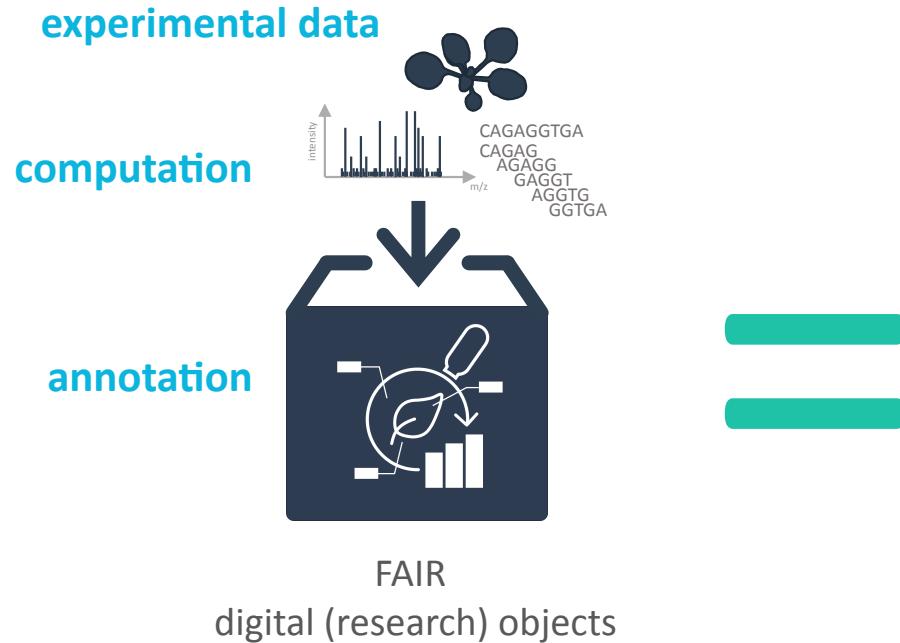
Continuous support

- 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/>
- User Support Mailing List: [Click here to subscribe](#)

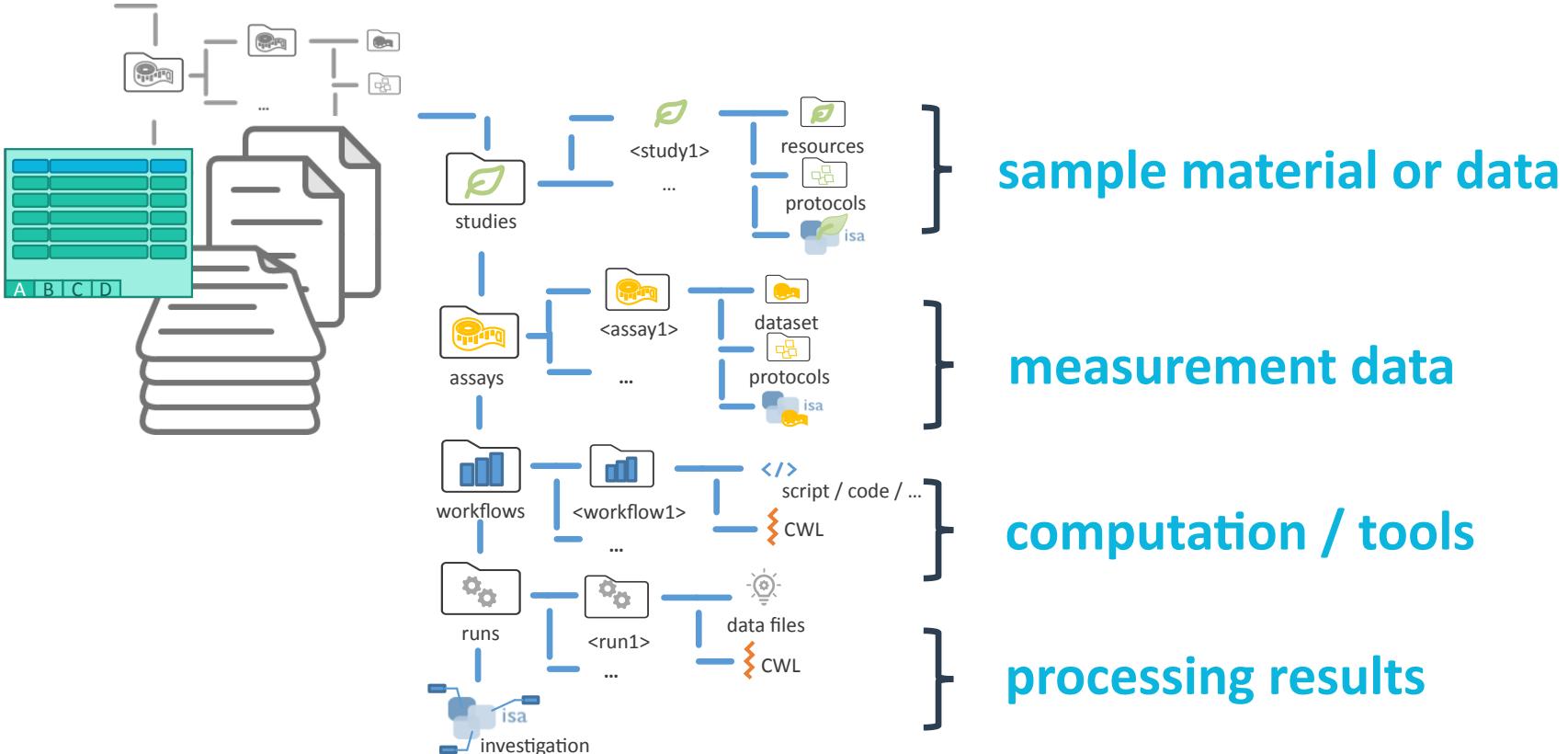
Open Source Development

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

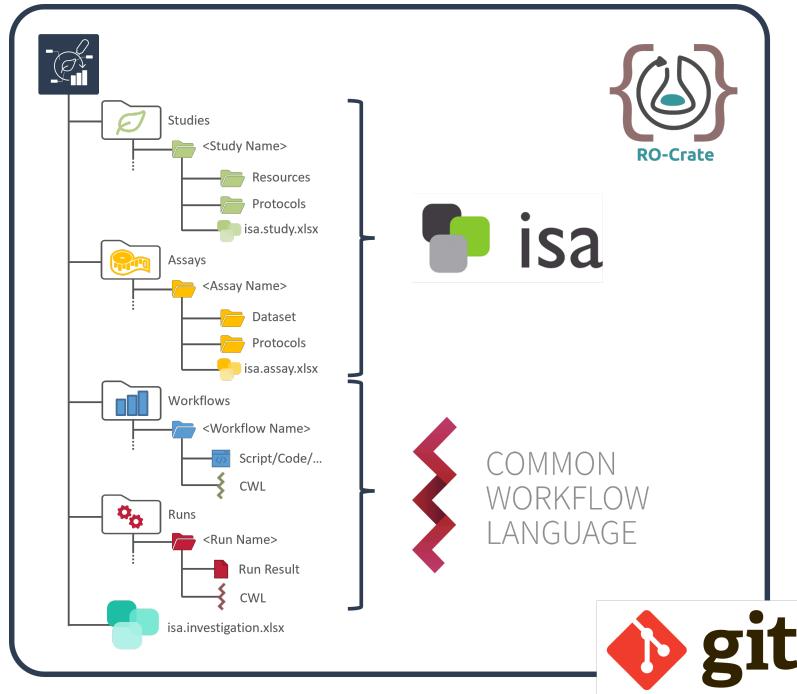
Annotated Research Context (ARC)



The ARC scaffold structure



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
- <https://git-scm.com>

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

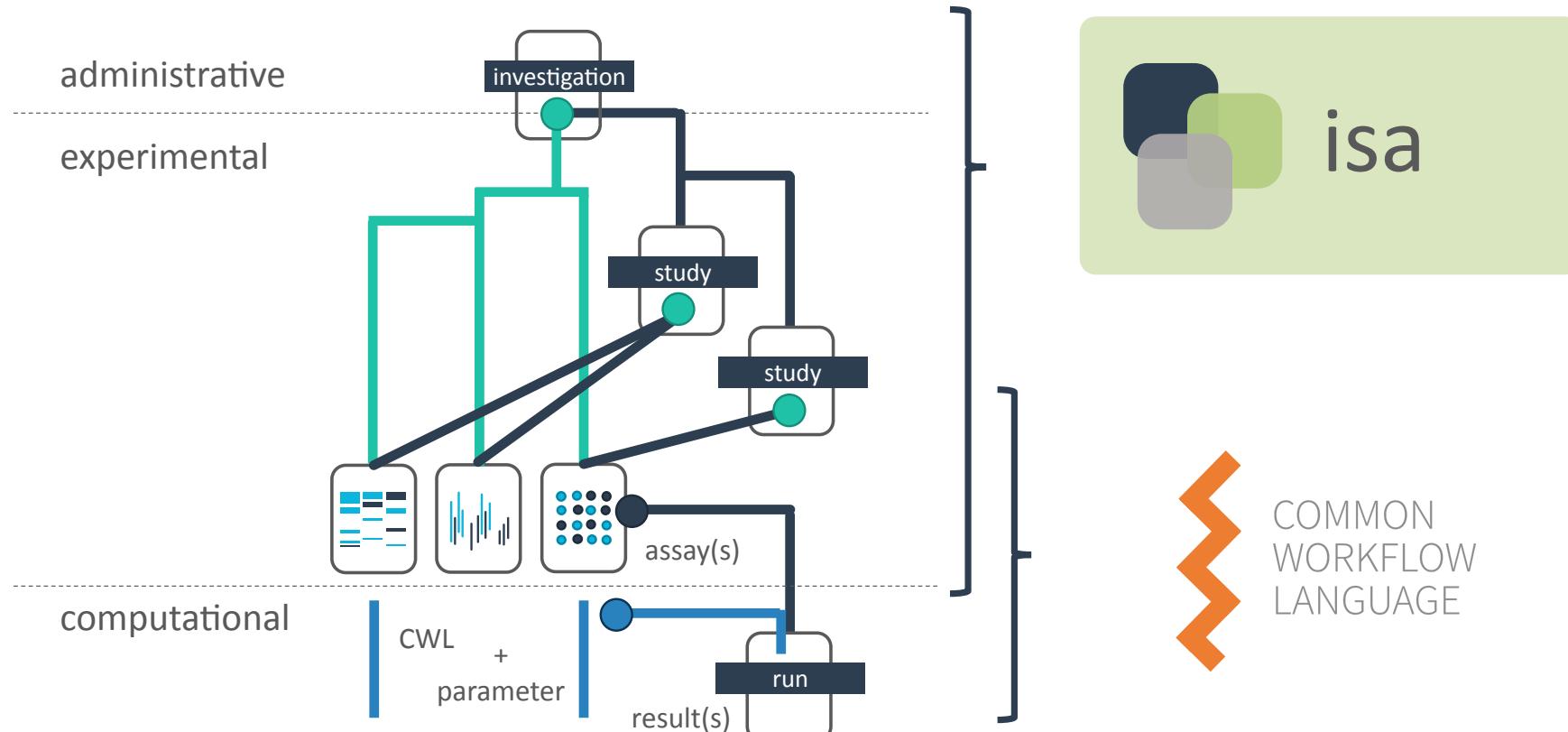
- Characteristics
- Parameters
- Components
- Factors

Assay

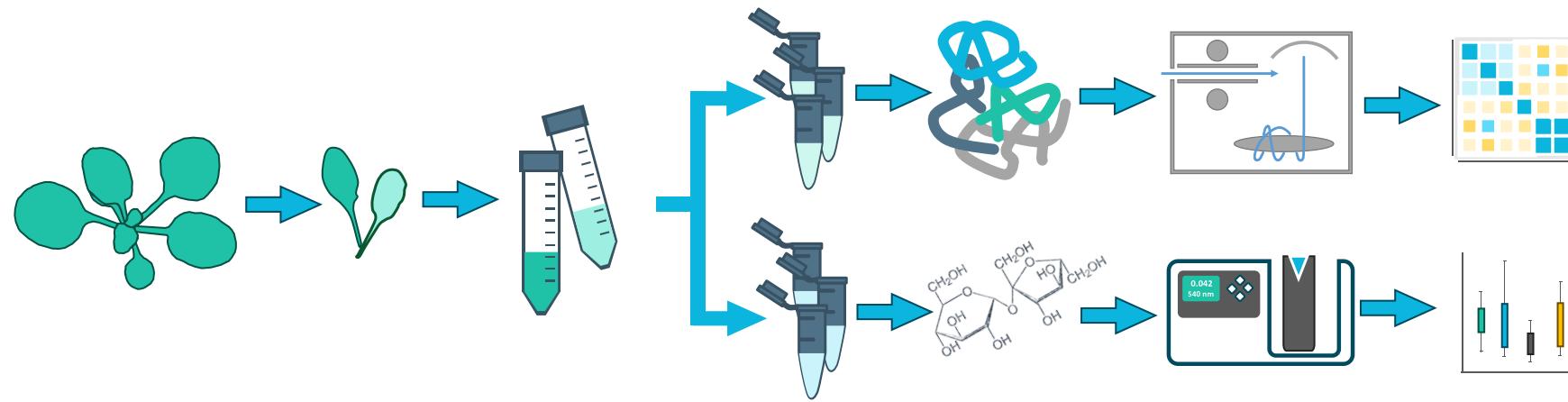
descriptive (meta)data
information on the measurement

- Characteristics
- Parameters
- Components
- Factors

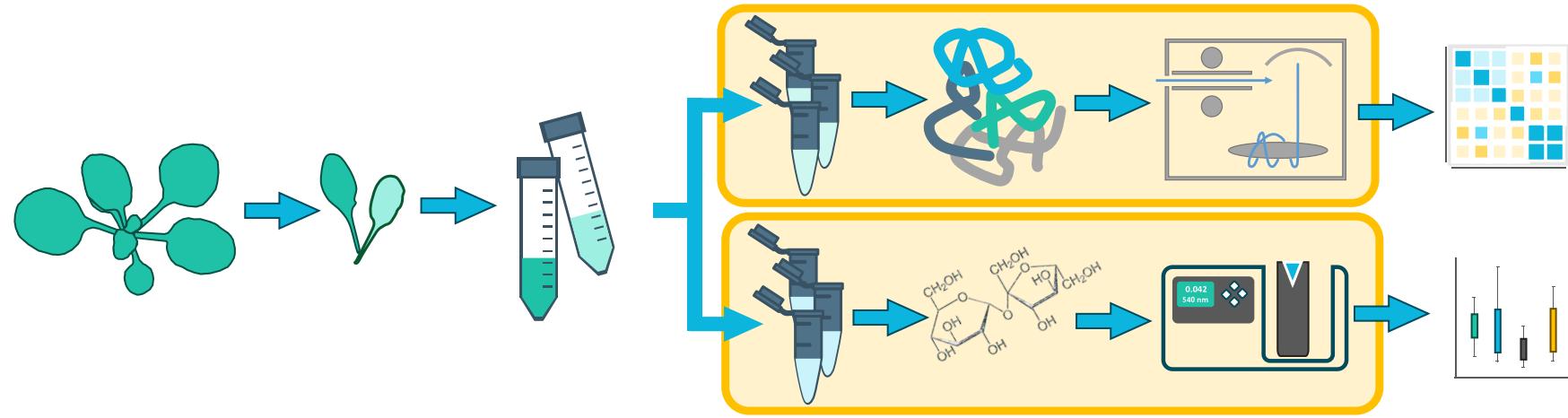
ISA and CWL – Connected by similarity



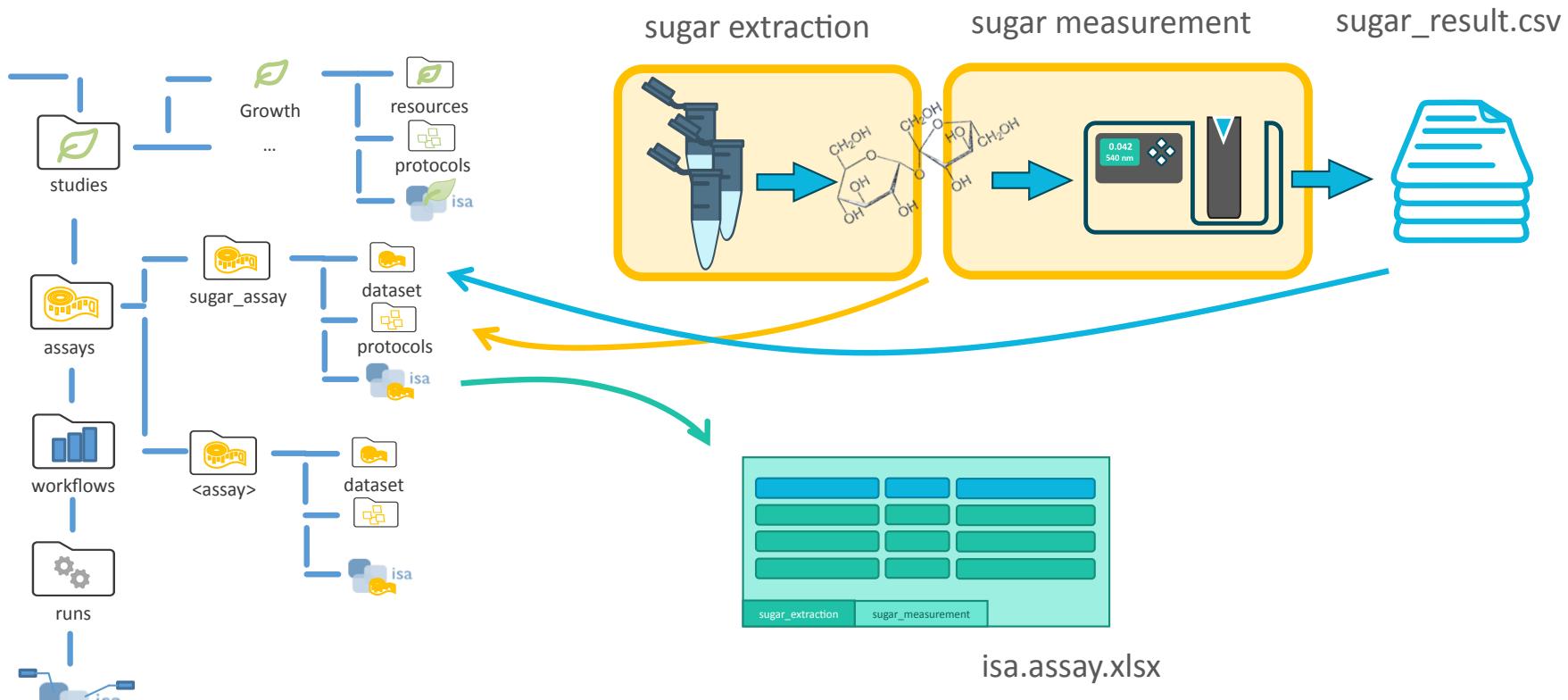
Metadata annotation – from sample to data



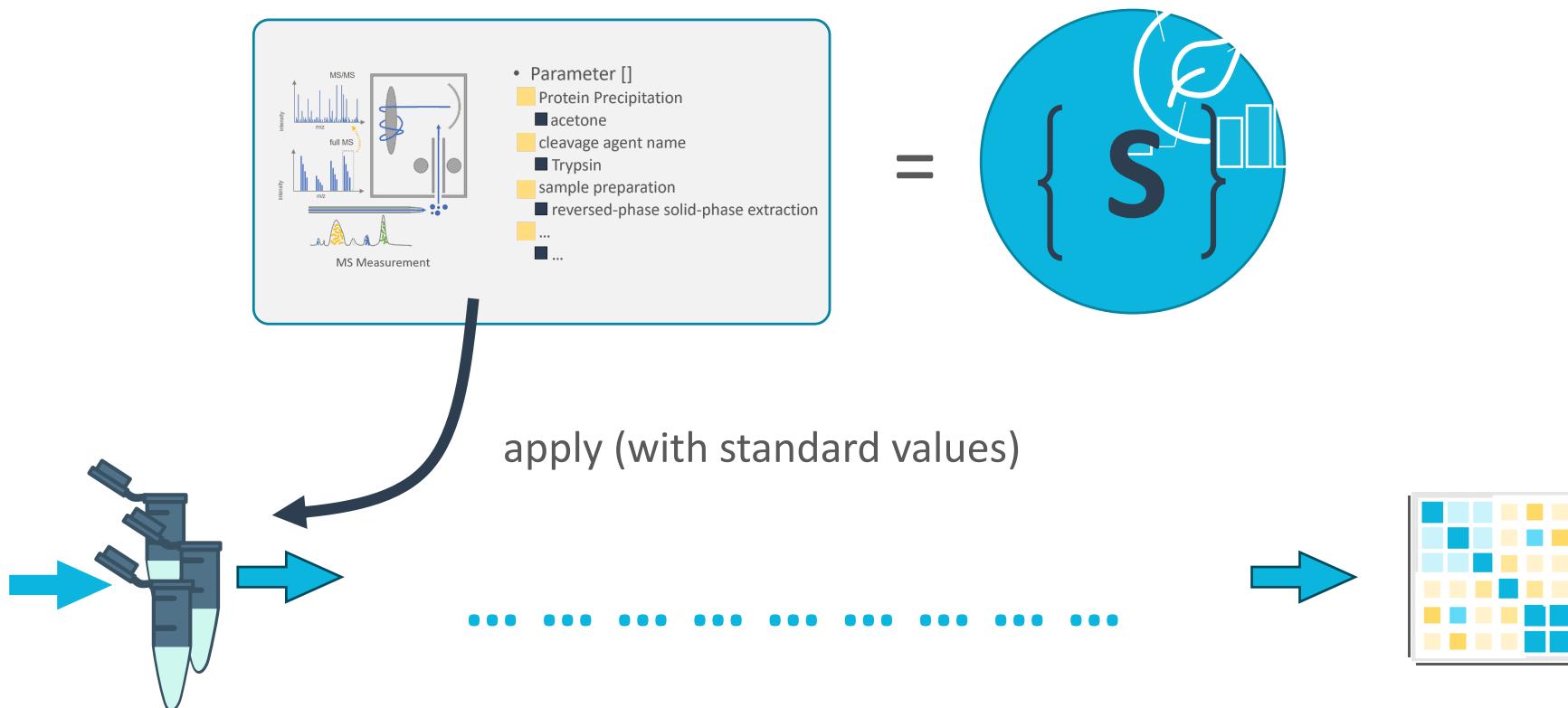
Modular separation of experimental processes



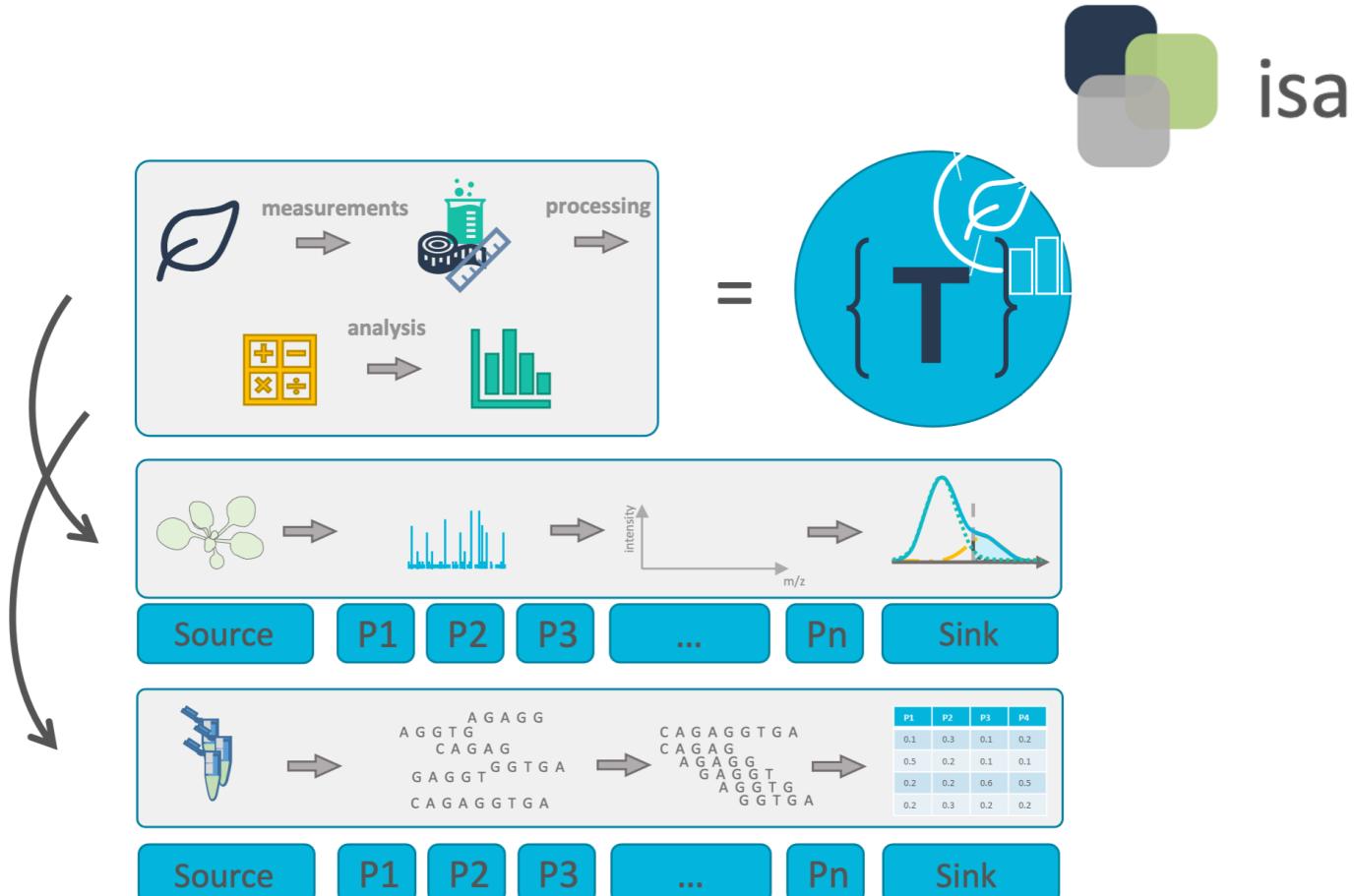
Modular separation of experimental processes



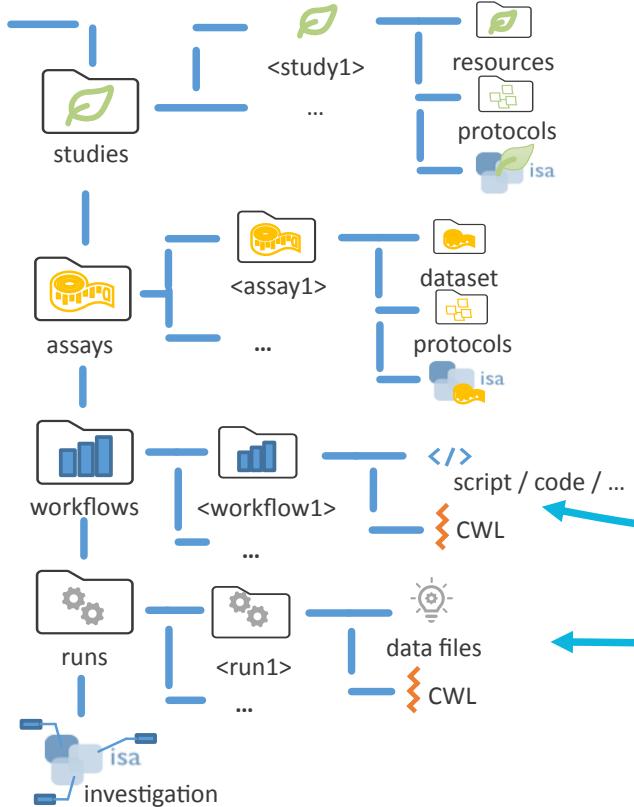
Applying standard procedures to sample record



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

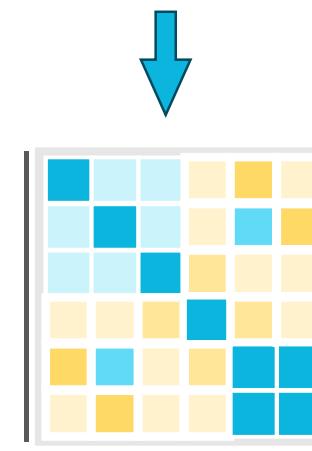
# Read the CSV file
data = pd.read_csv('result.csv')

# 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)

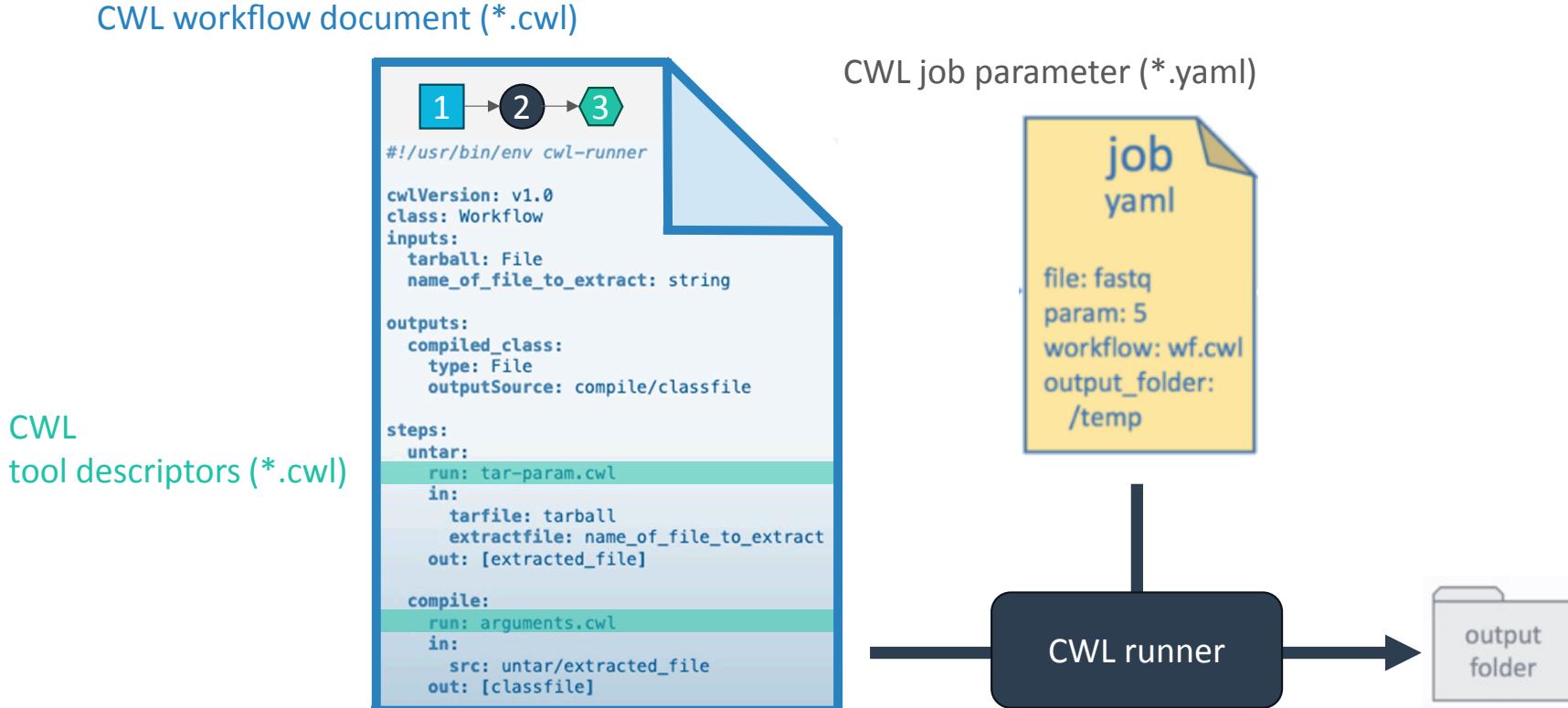
# Create a heatmap using Plotly
fig = px.imshow(data,
                 labels=dict(x="Columns", y="Rows", color="Value"),
                 x=data.columns,
                 y=data.index)

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

heatmap.py

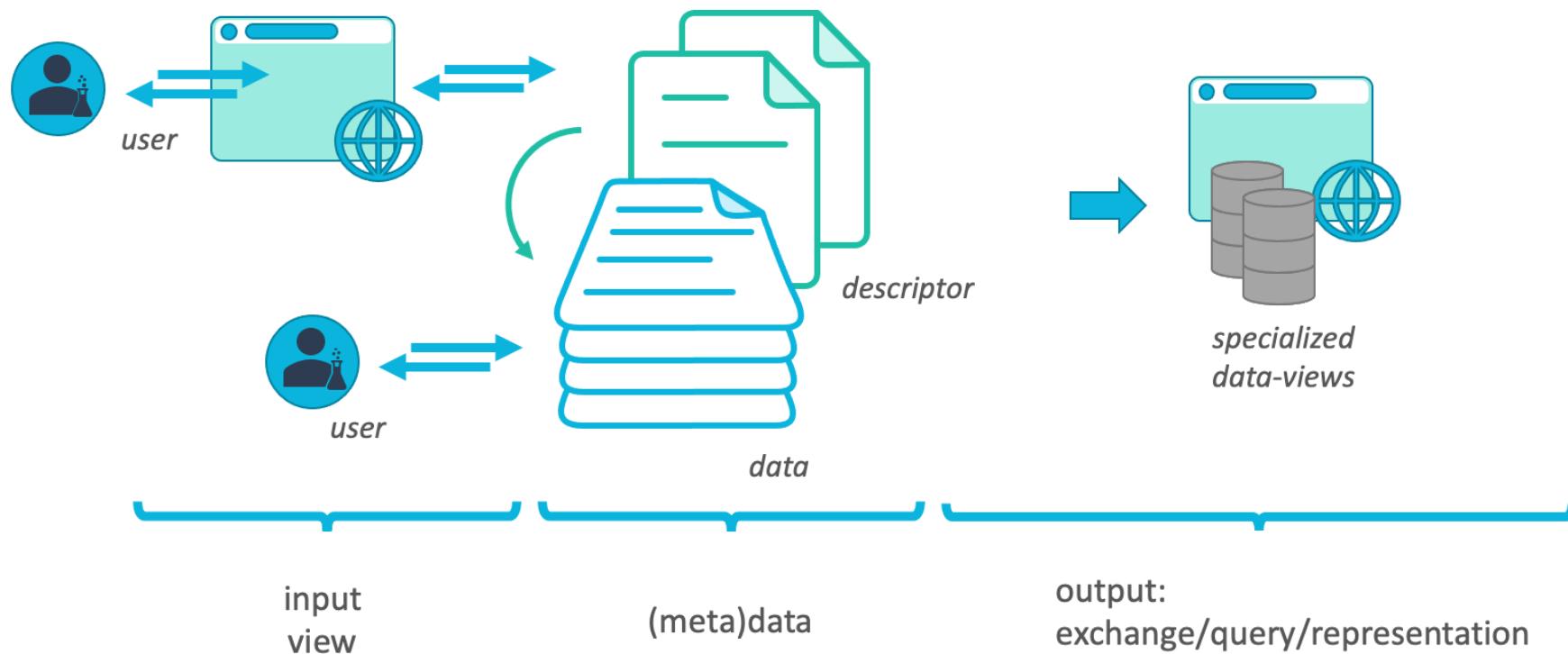


Metadata annotation – from data to result

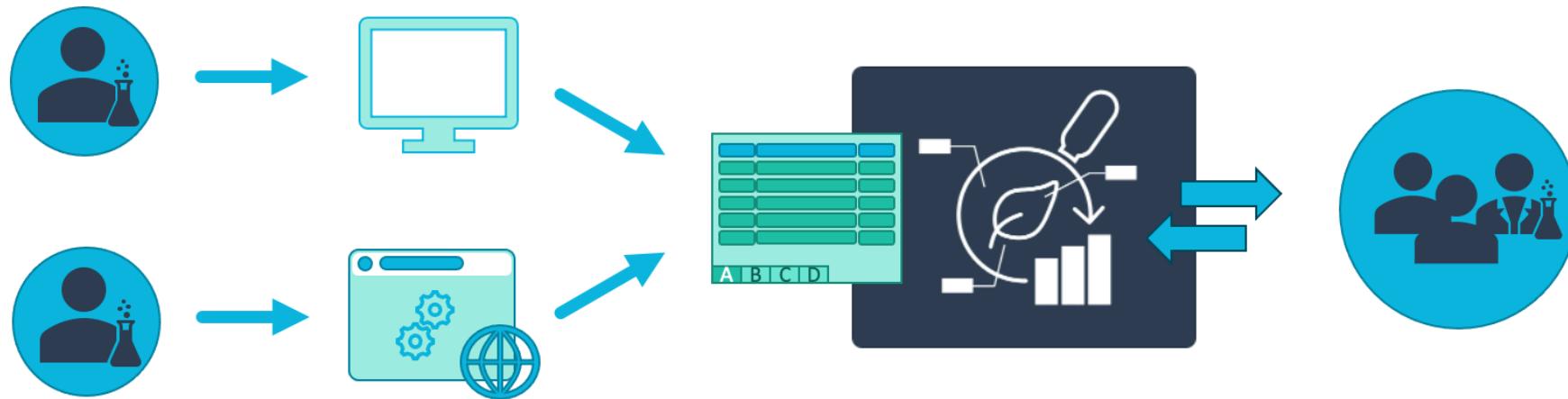


Everything is a file

The ARC is a data-centric approach to RDM

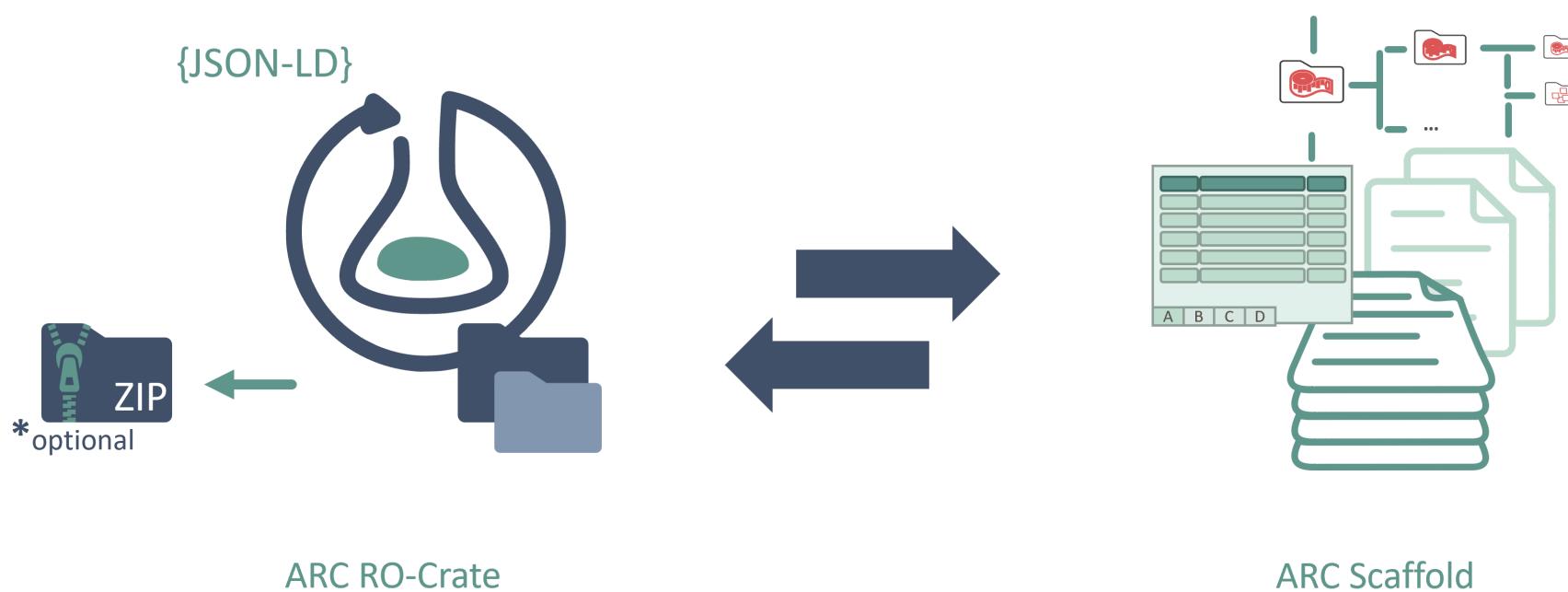


No technical lock-in



(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

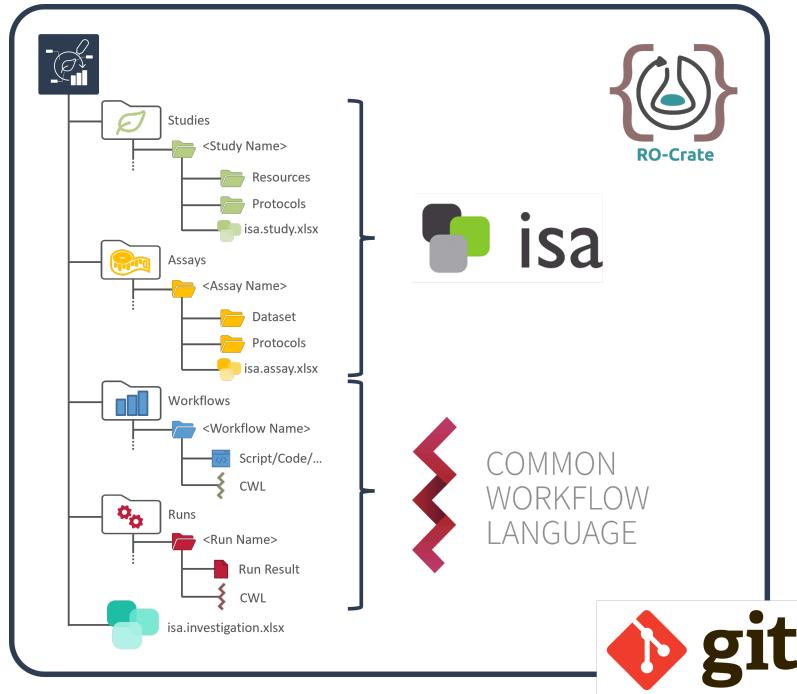
```
{  
  "Identifier": "Proteomics_MS",  
  "MeasurementType": {  
    "annotationValue": "Proteomics_MS",  
    "termSource": "MS",  
    "termAccession": "https://purl.obolibrary.org/obo/FMS_1003348"  
  },  
  "TechnologyType": ...  
  ...  
  "Tables": [  
    {  
      "name": "ProtDigest",  
      "header": [  
        {  
          "headertype": "Parameter",  
          "values": [  
            {  
              "annotationValue": "sample mass",  
              "termSource": "MS",  
              "termAccession": "https://purl.obolibrary.org/obo/FMS_1003348"  
            }  
          ]  
        ]  
      ]  
    }  
  ]  
}
```

"User View": ARC Scaffold and metadata tables

The screenshot shows the ARCIct application interface. On the left, the ARCIct sidebar displays a file structure for a project named 'AthalianaColdStressSugar'. The 'assays' folder contains sub-folders like 'Proteomics_DataAnalysis' and 'SugarMeasurement', each with its own README.md file. On the right, a detailed view of the 'Assay Metadata' for the 'Proteomics_MS' assay is shown. This view includes fields for Identifier (Proteomics_MS), Measurement Type (Proteomics_MS), Technology Type (Mass Spectrometry), Technology Platform (timsTOF Pro 2), and Performers (PeptideMS_Bruker, ProtDigest). Below this, a table lists parameters: sample mass (10 microgram), Protein Precipitation (acetone), alkylating agent (Chloroacetamide), and redox agent (TCEP).

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

ARC builds on standards



RO-Crate

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

ISA

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

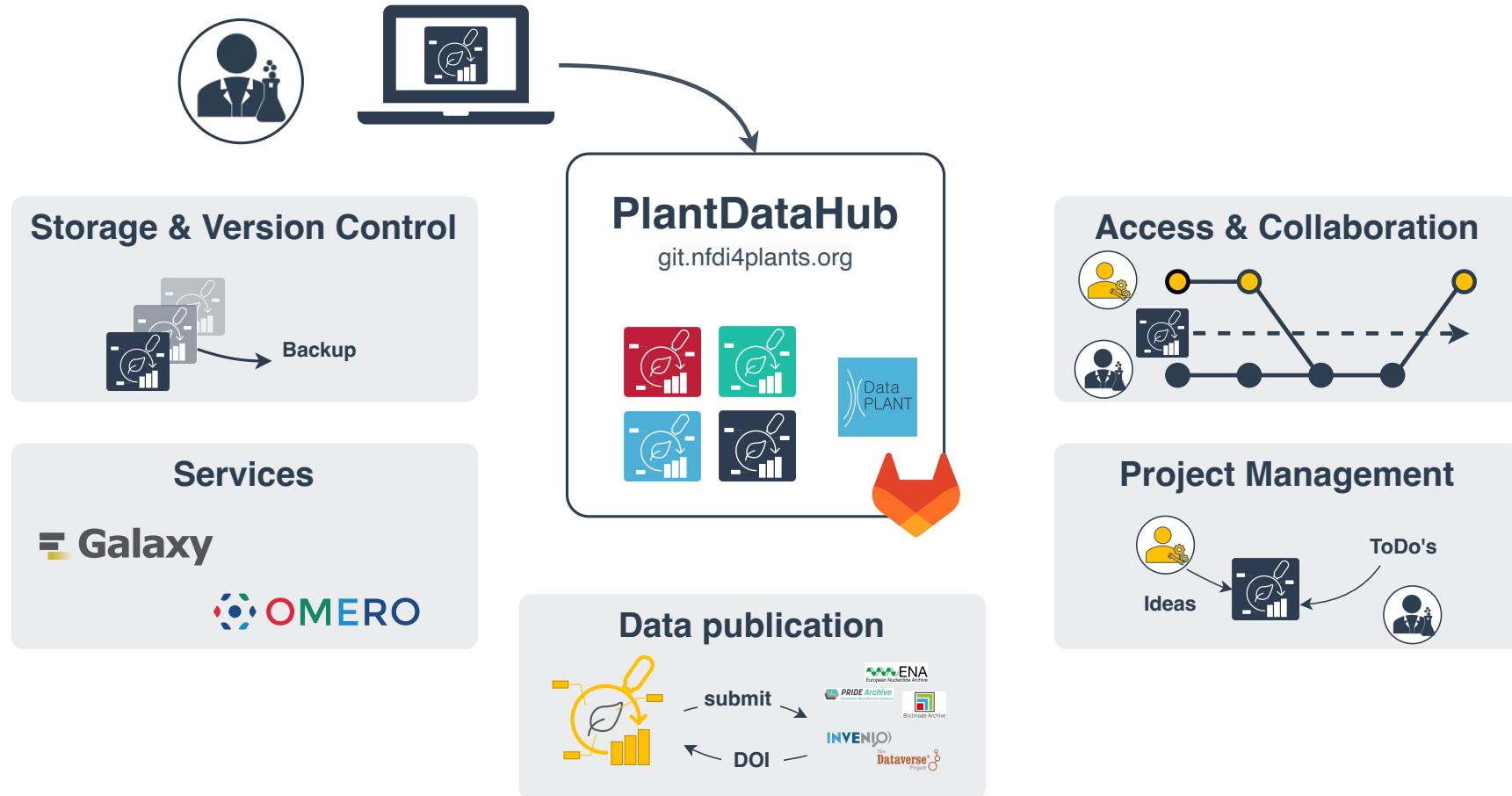
CWL

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

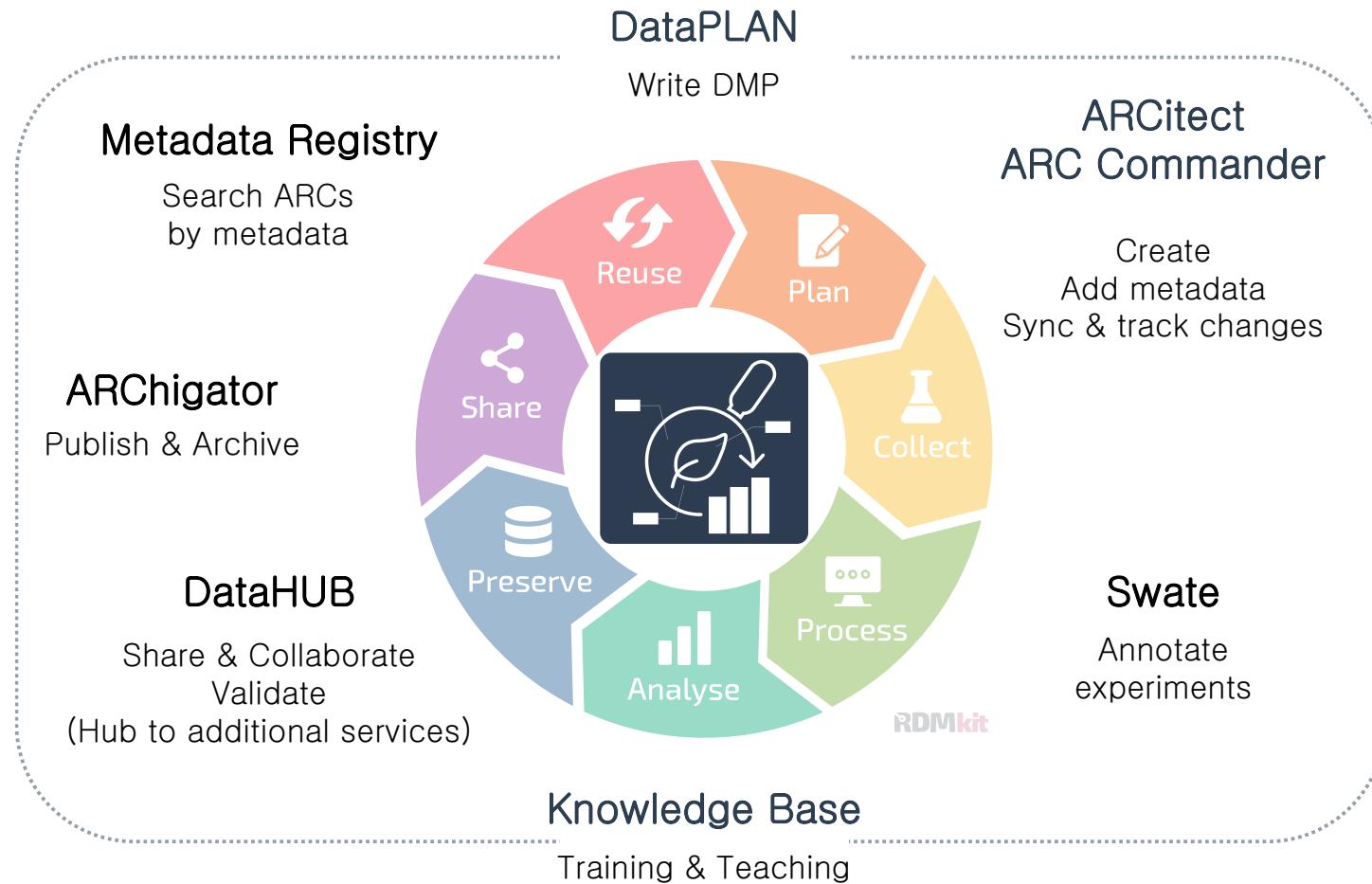
Git

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

ARC and DataHUB as entry point



The ARC ecosystem



Acknowledgements



Team Kaiserslautern

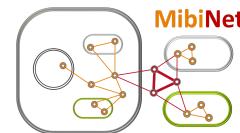
- Timo Mühlhaus
- Lukas Weil
- Kevin Frey
- Kevin Schneider
- Jonas Lukasczyk

Team Freiburg

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



- Björn Usadel
- Vittorio Tracanna
- Yaser Alashloo



- Sabrina Zander

Goals for today

- Get an idea of the ARC concept
- Create and upload a simple ARC
- Explore
- ...

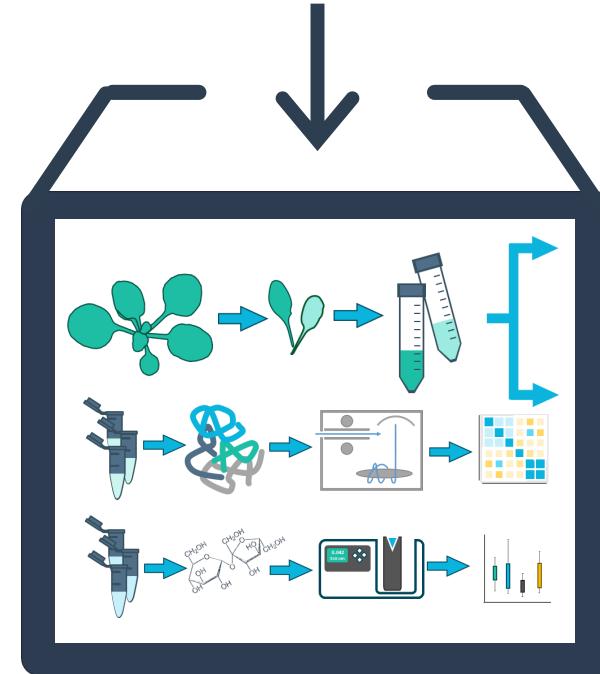
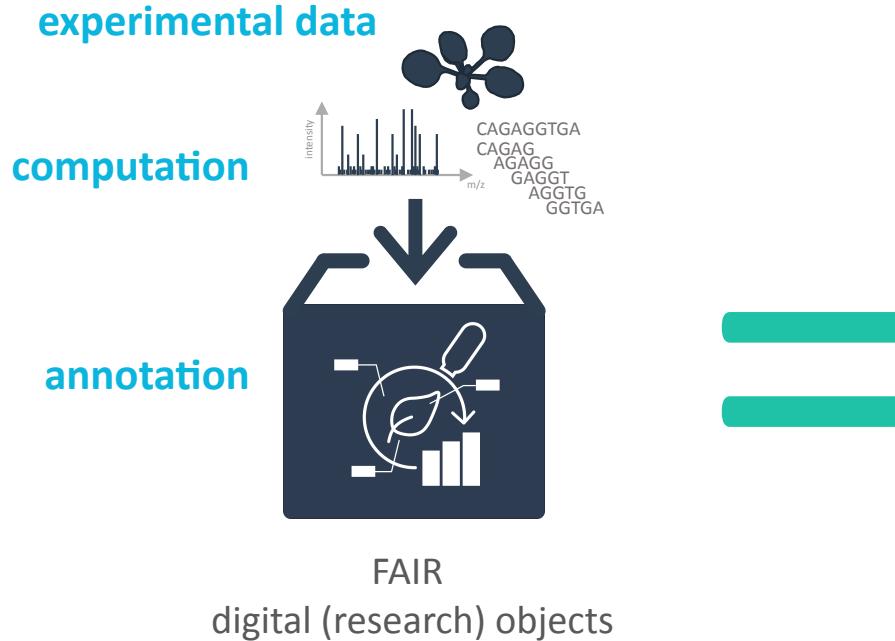
Hands-on part 1: Setup and ARCitect

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

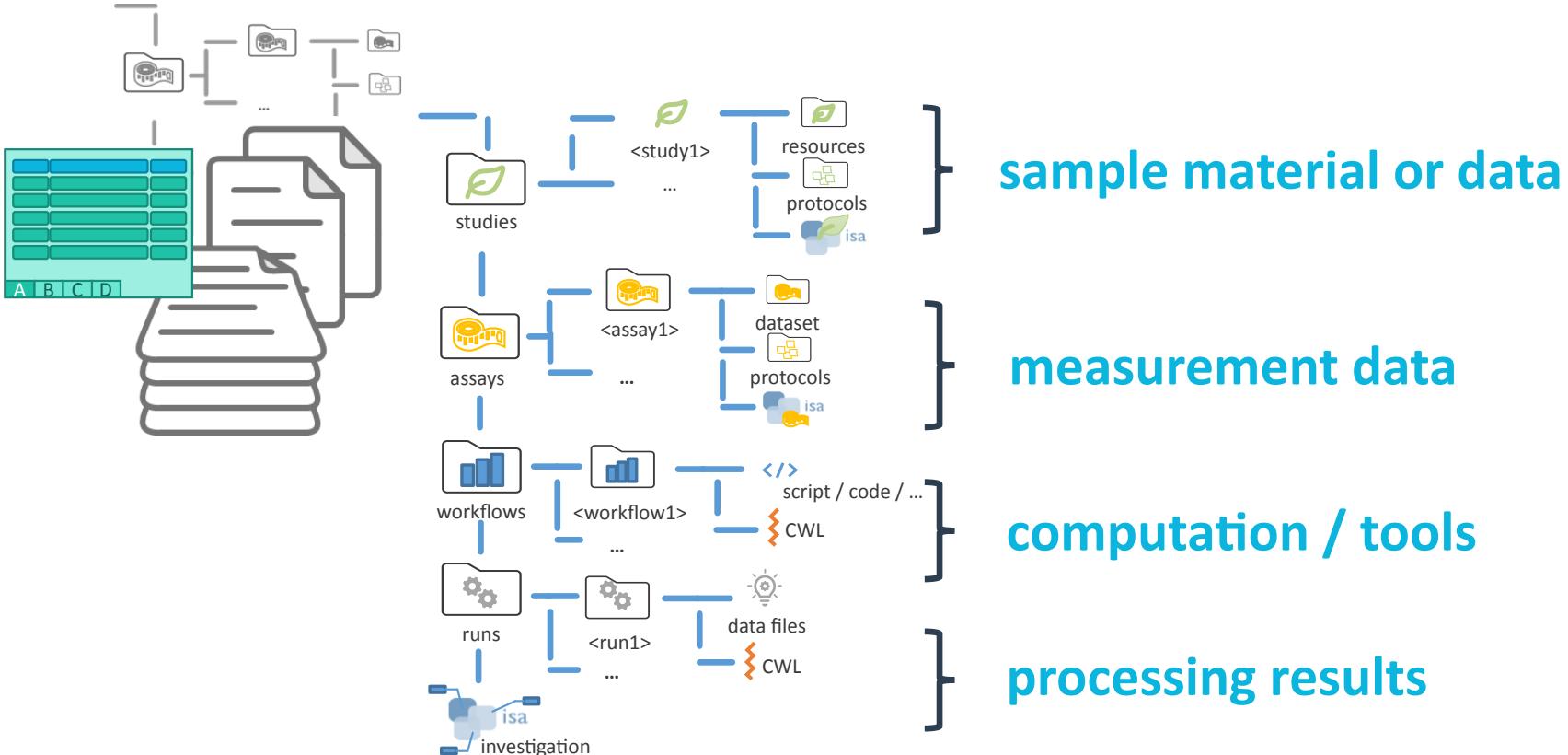


Until step Add a study

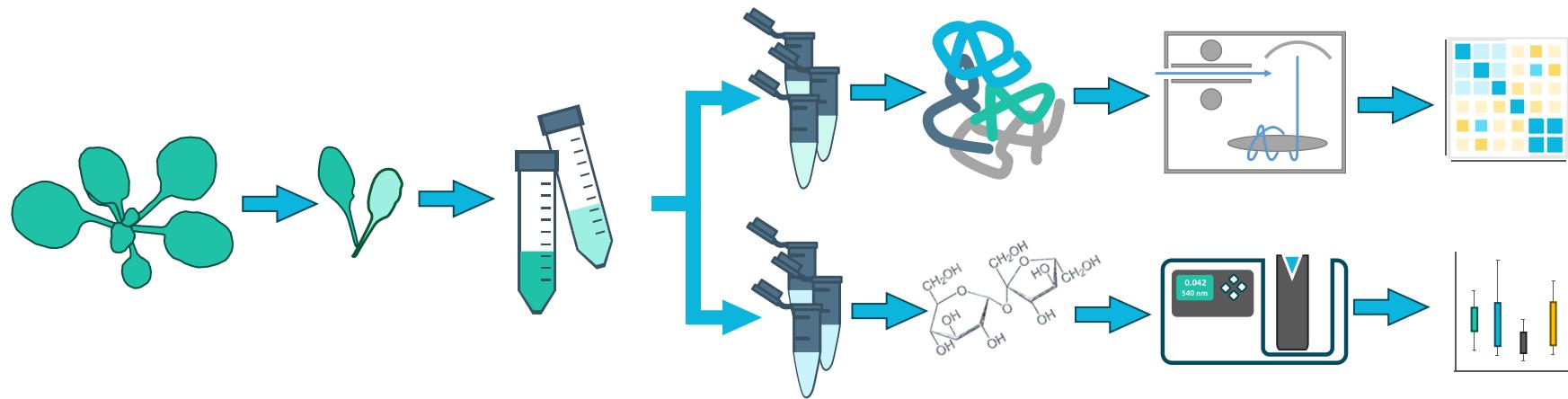
ARC: Annotated research context



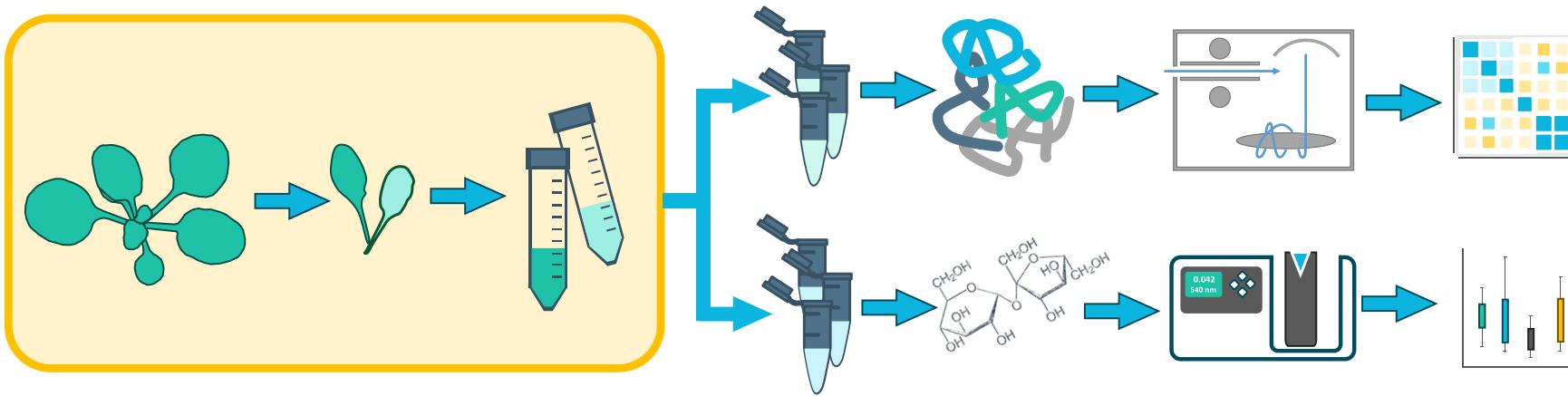
The ARC scaffold structure



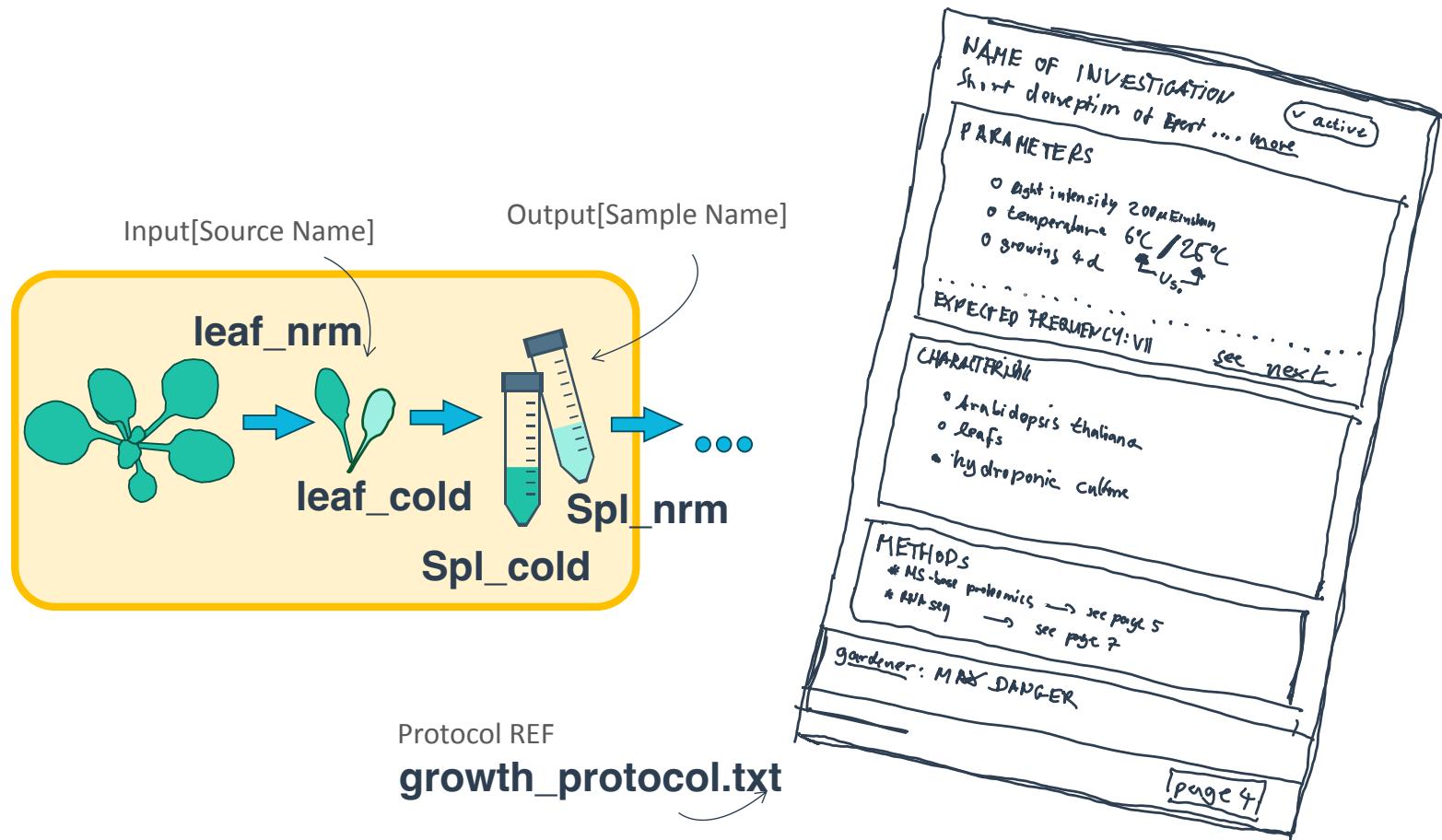
A small prototypic project



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		

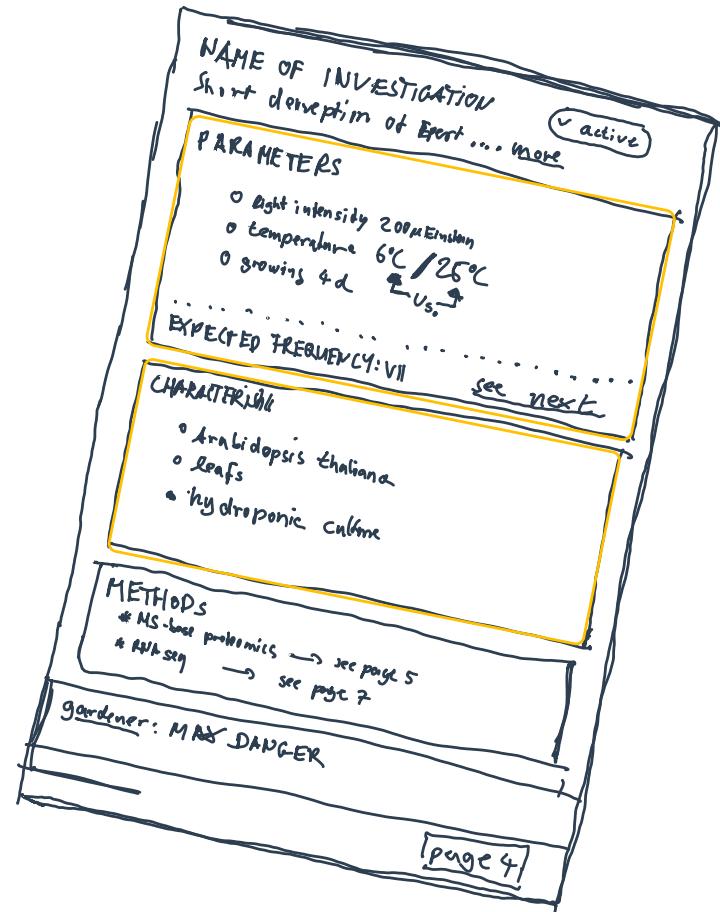
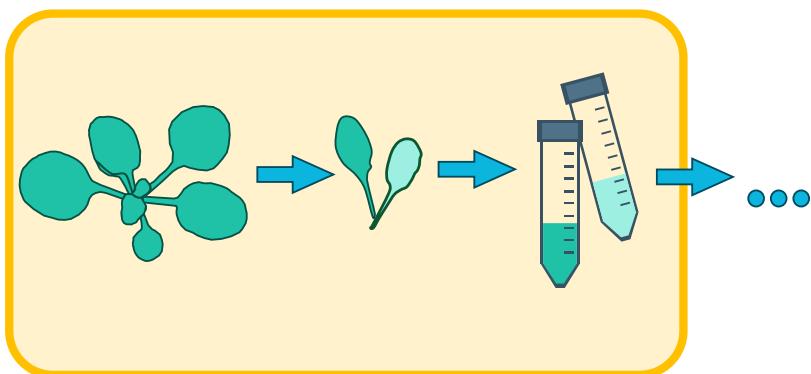
Referencing a protocol

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

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

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

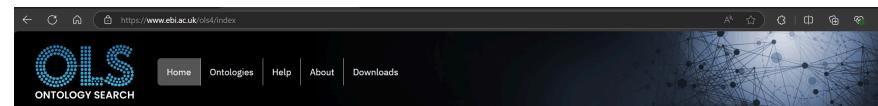
temperature PATO:0000146 

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 +



Welcome to the EMBL-EBI Ontology Lookup Service

temperature 

Exact match Include obsolete terms Include imported terms

Examples: diabetes, GO:0098743 

Looking for a particular ontology? 

Data Content
Updated 4 Oct 2024 Fri 17:50 +02:00
• 265 ontologies
• 8,595,500 classes
• 45,471 properties
• 697,188 individuals

 About OLS
The Ontology Lookup Service (OLS) is a repository for biomedical ontologies that aims to provide a single point of access to the latest ontology versions. You can browse the ontologies through the website as well as programmatically via the OLS API. OLS is developed and maintained by the Samples, Phenotypes and Ontologies Team (SPOT) at EMBL-EBI.

 Related Tools
In addition to OLS the SPOT team also provides the Oxo and ZOOMA services. Oxo provides cross-ontology mappings between terms from different ontologies. ZOOMA is a service to assist in mapping data to ontologies in OLS.

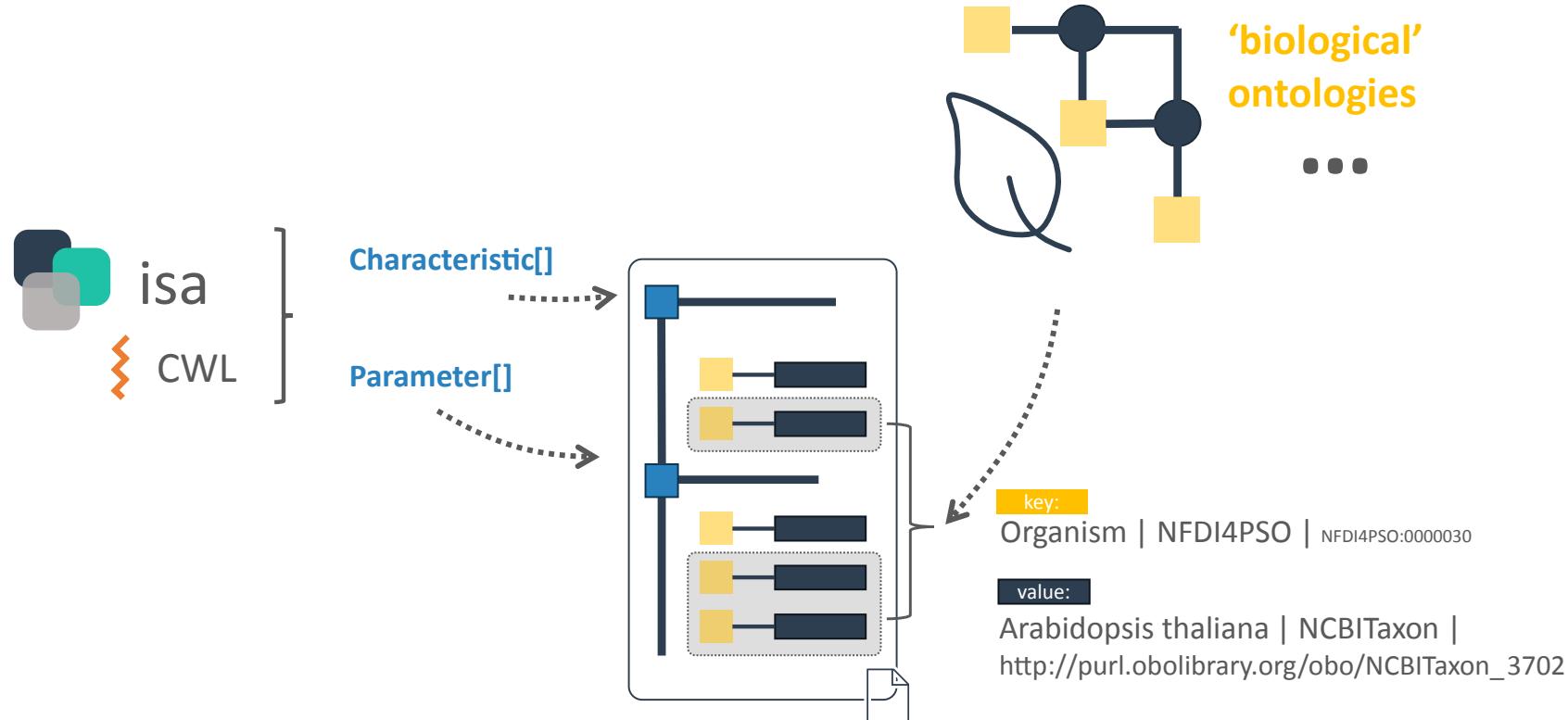
 Report an Issue
For feedback, enquiries or suggestion about OLS or to request a new ontology please use our GitHub issue tracker. For announcements relating to OLS, such as new releases and new features sign up to the OLS announce mailing list.

FOLLOW US 

© EMBL-EBI 2023 Licensing

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 μ Einstein
-   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

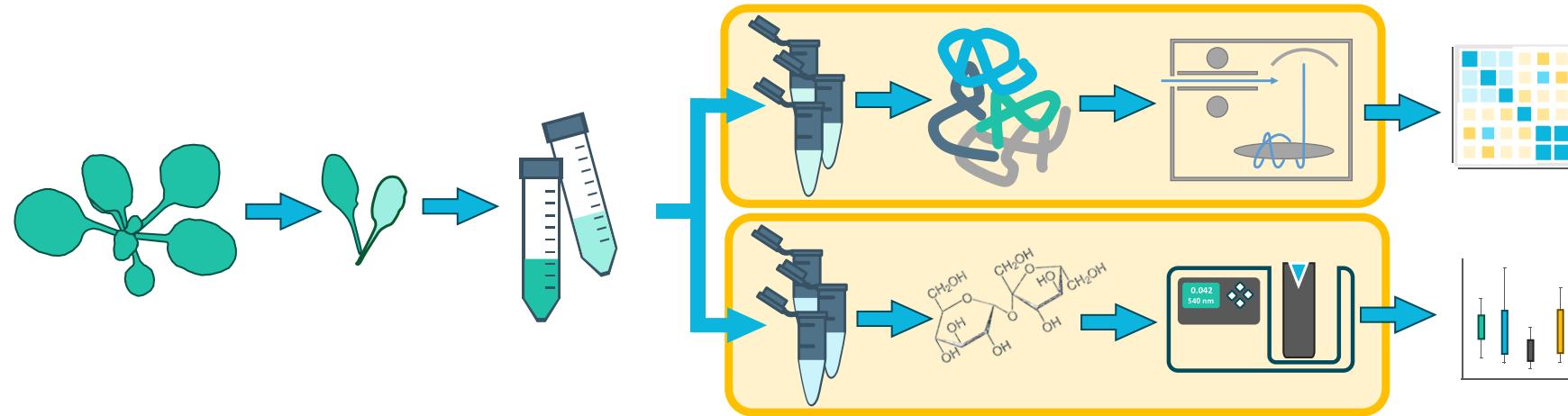
Follow the [Start Here guide](#) in the DataPLANT knowledge base.



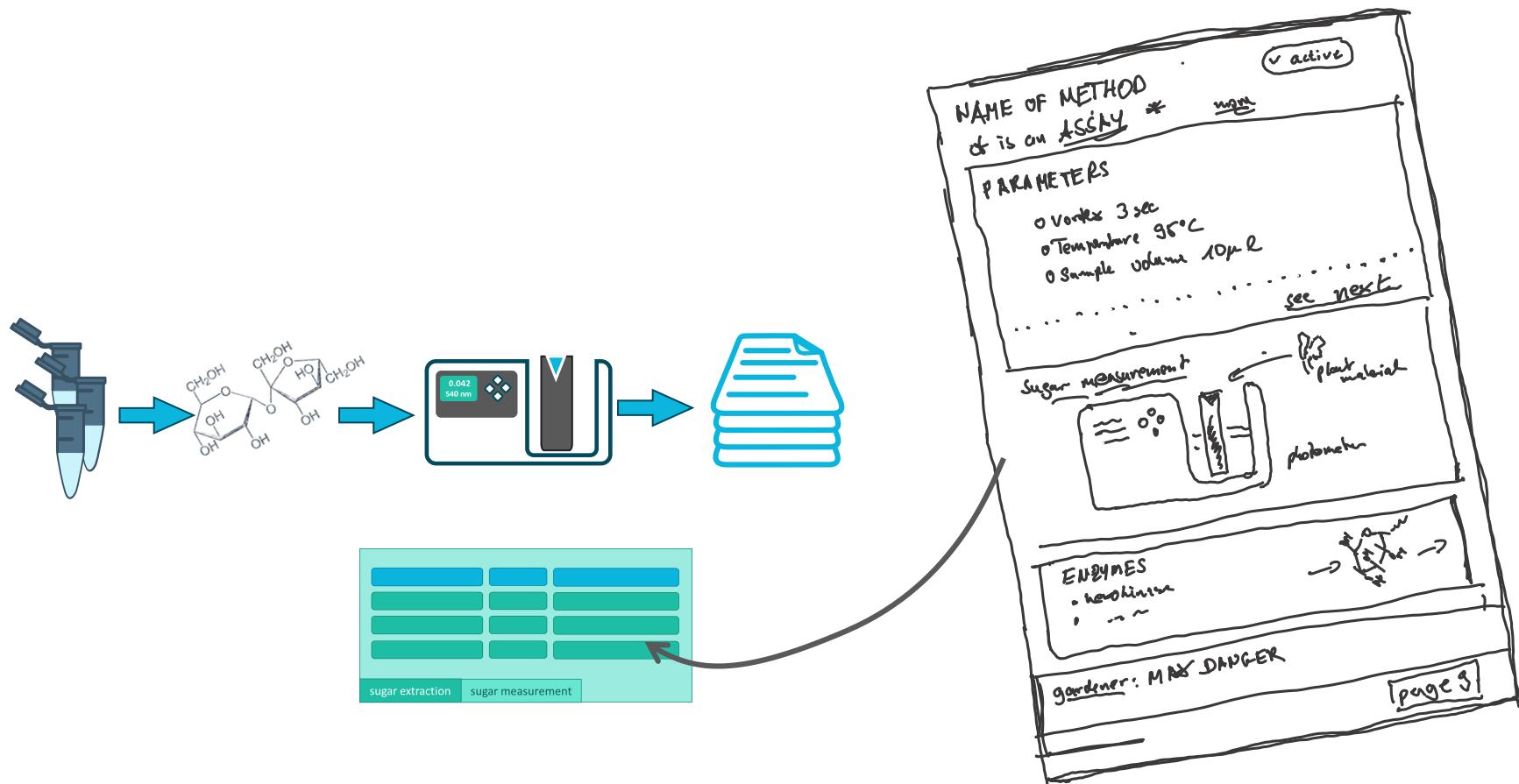
Until step Add a study

Hands-on part 2: ARCitect (and Swate)

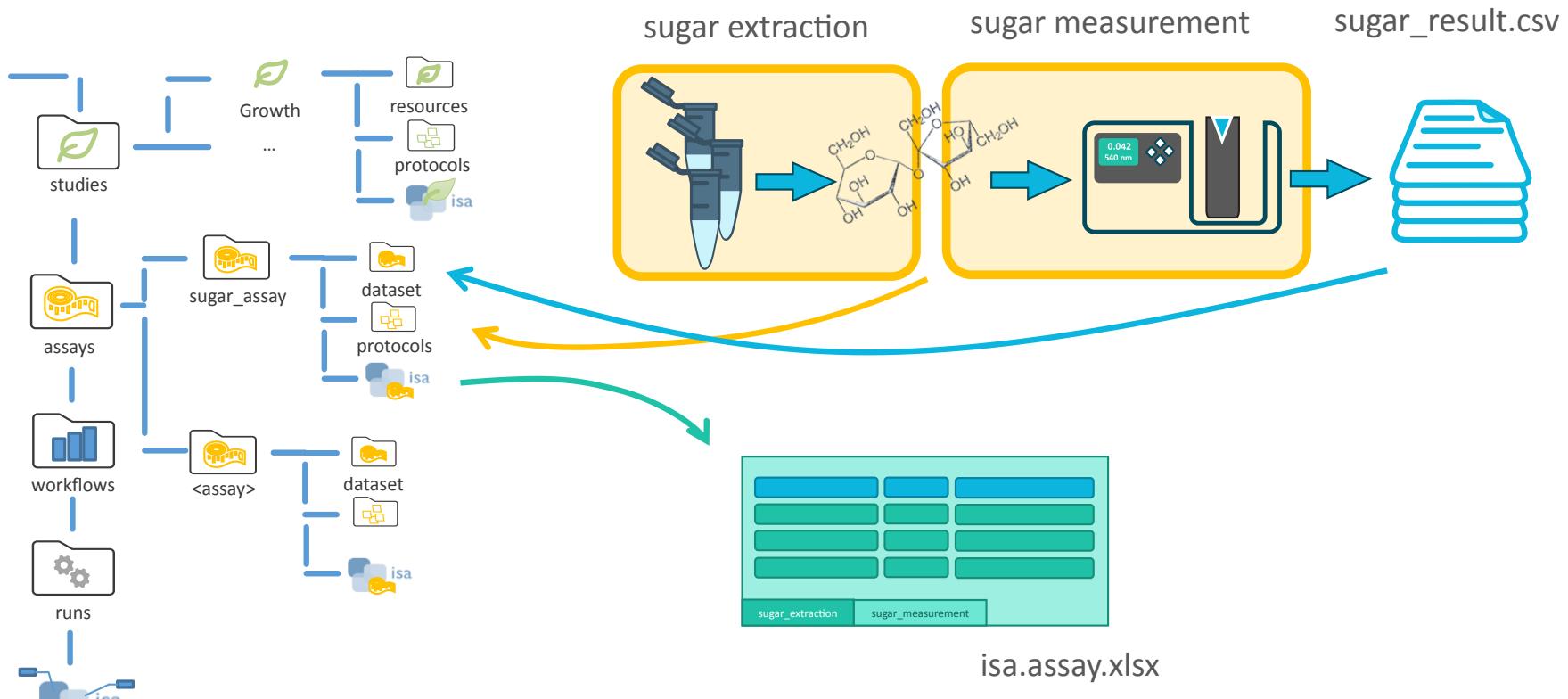
Identifying assays



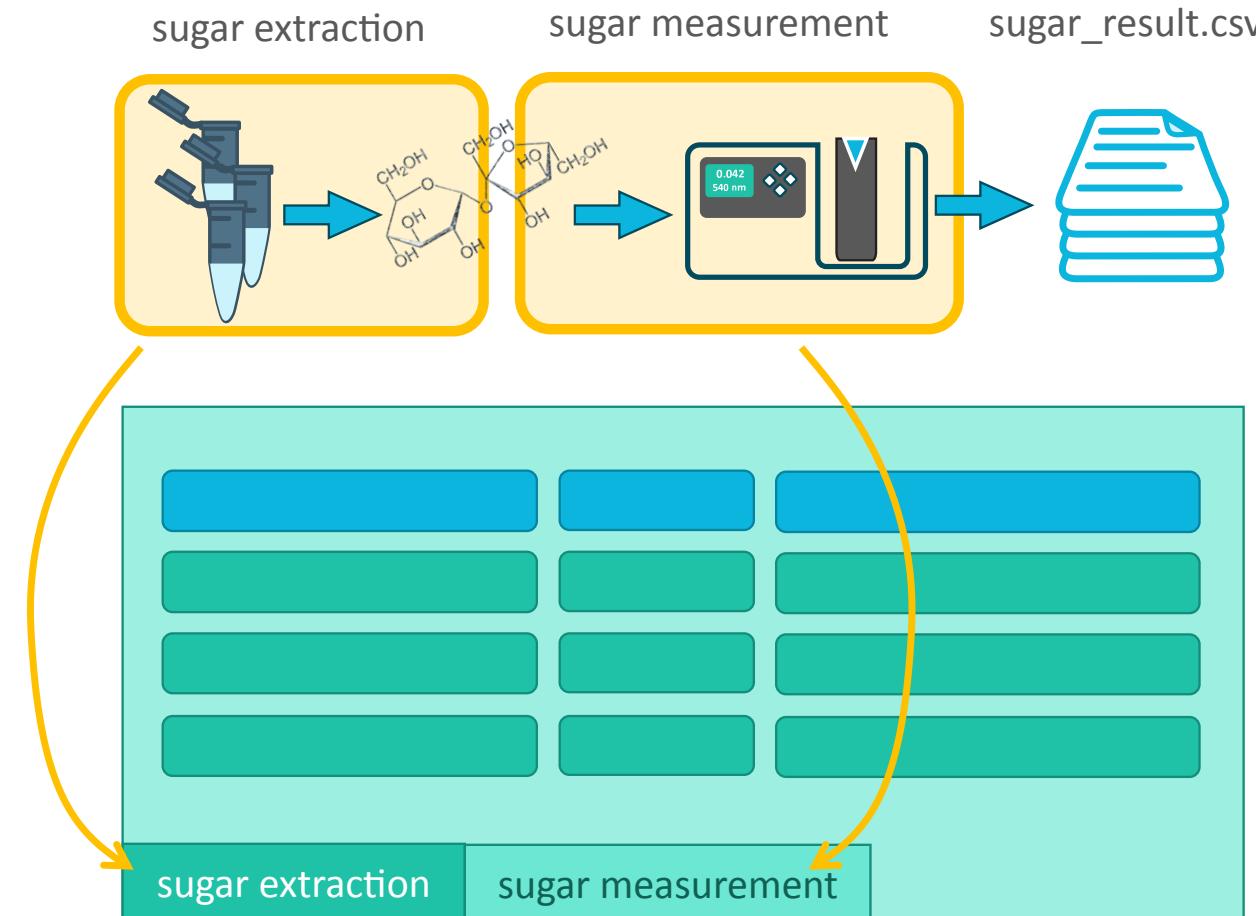
Assay for sugar measurement



Separating different assay elements



Isolating the lab processes in an assay



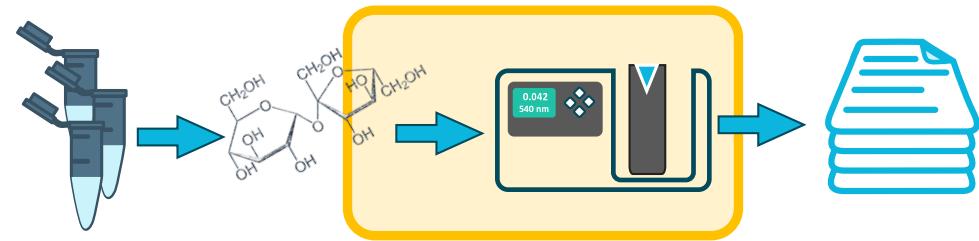
Parameteterization: sugar extraction

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



Parameterization: sugar measurement

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



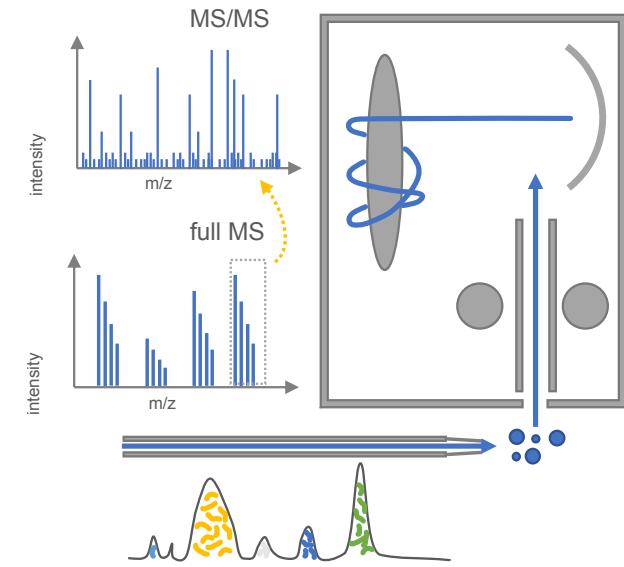
Save time using standard methods and SOPs

Parameter []

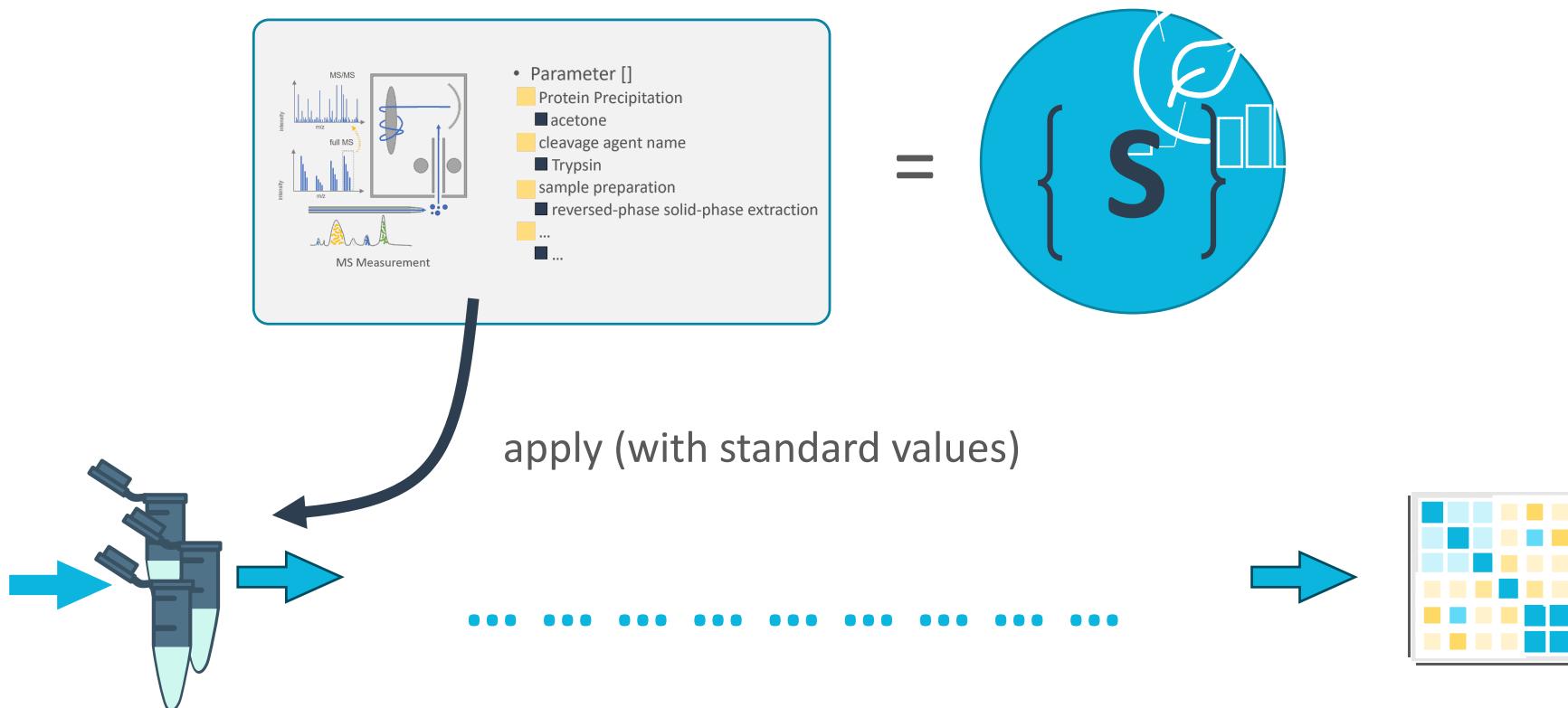
- Protein Precipitation
 - acetone
- cleavage agent name
 - Trypsin
- sample preparation
 - reversed-phase solid-phase extraction
- ...

Component []

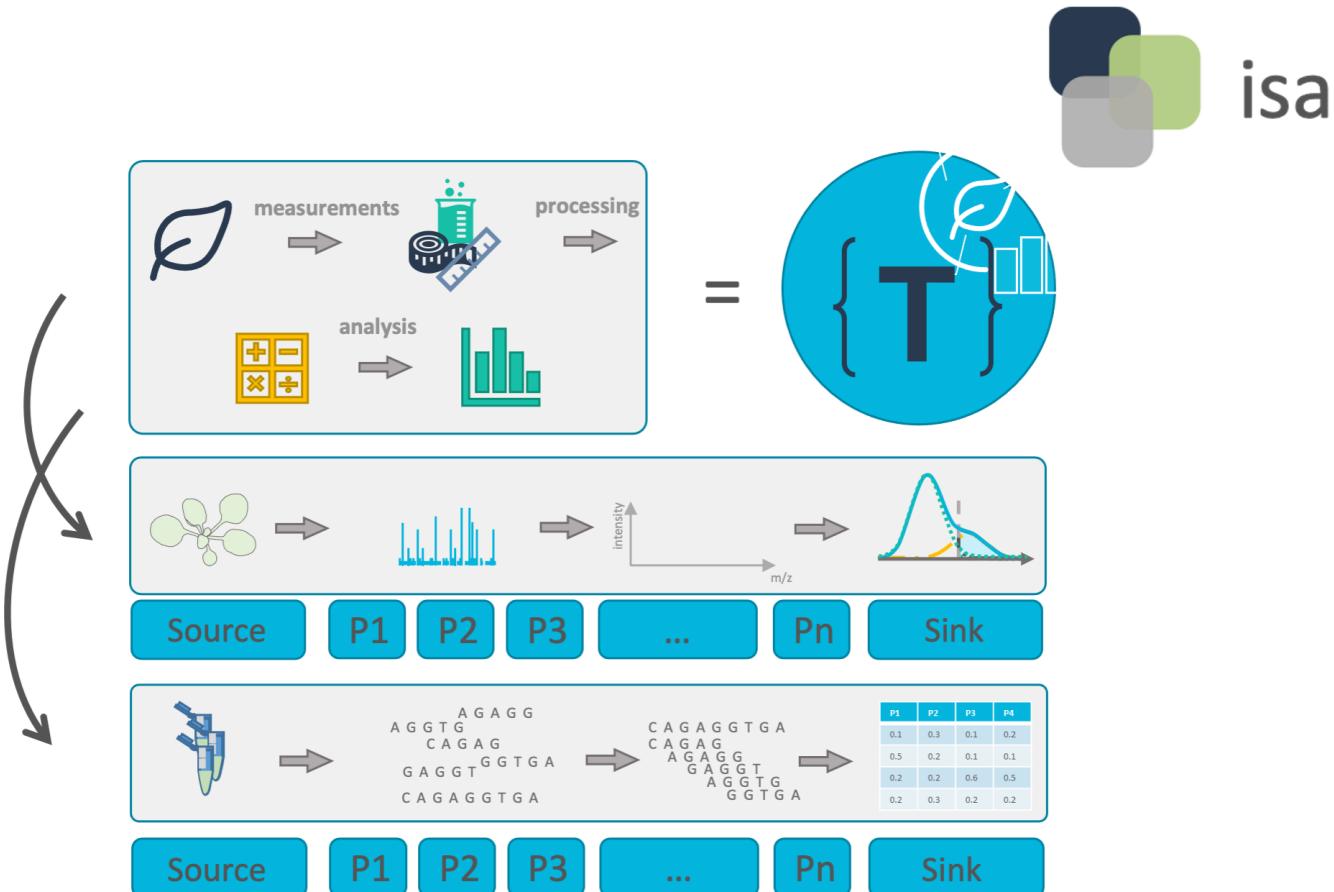
- chromatography instrument model
 - nanoElute2
- chromatography column model
 - PepSep C18 1.9u, 25cm x 75um ...



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.



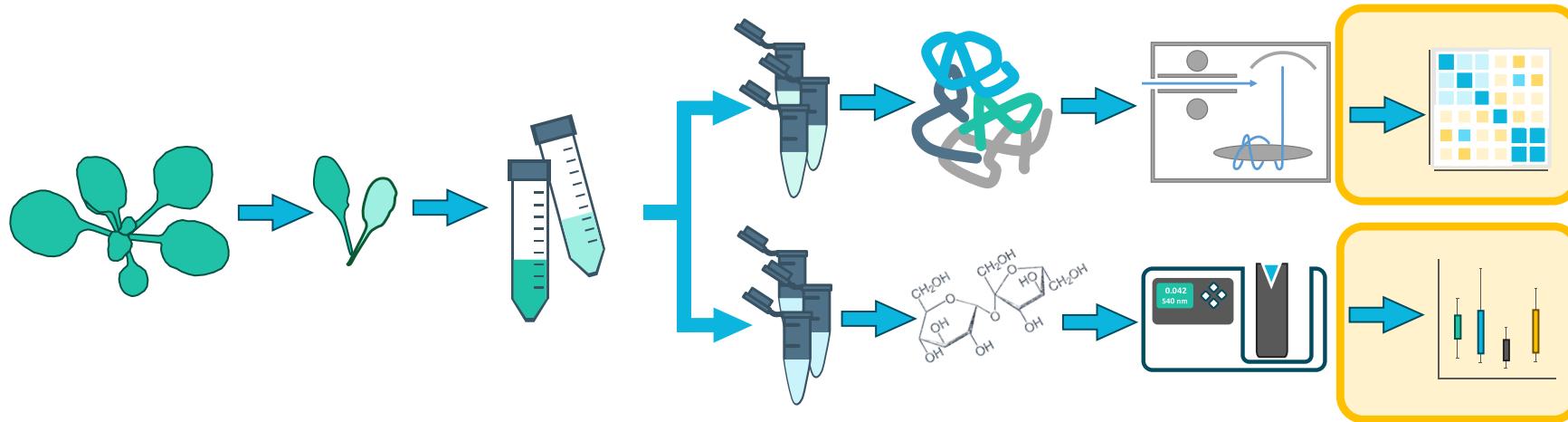
Until step Add assay data

Recap this Spring School

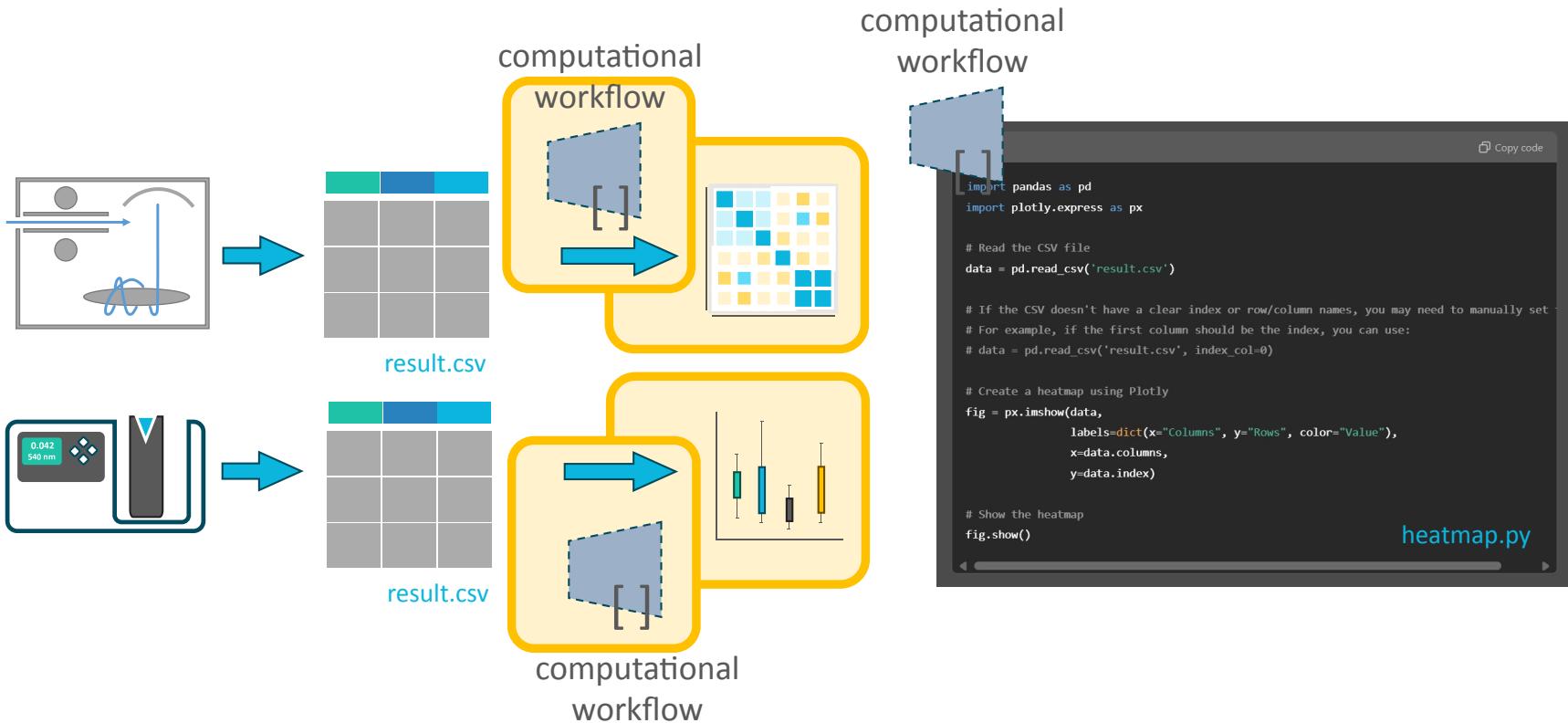
- What (types of) data did we learn about, e.g. in class, at IPK campus?
- Would you find a suitable place for the data in the ARC?
- Do you miss anything?

Hands-on part 3: Data

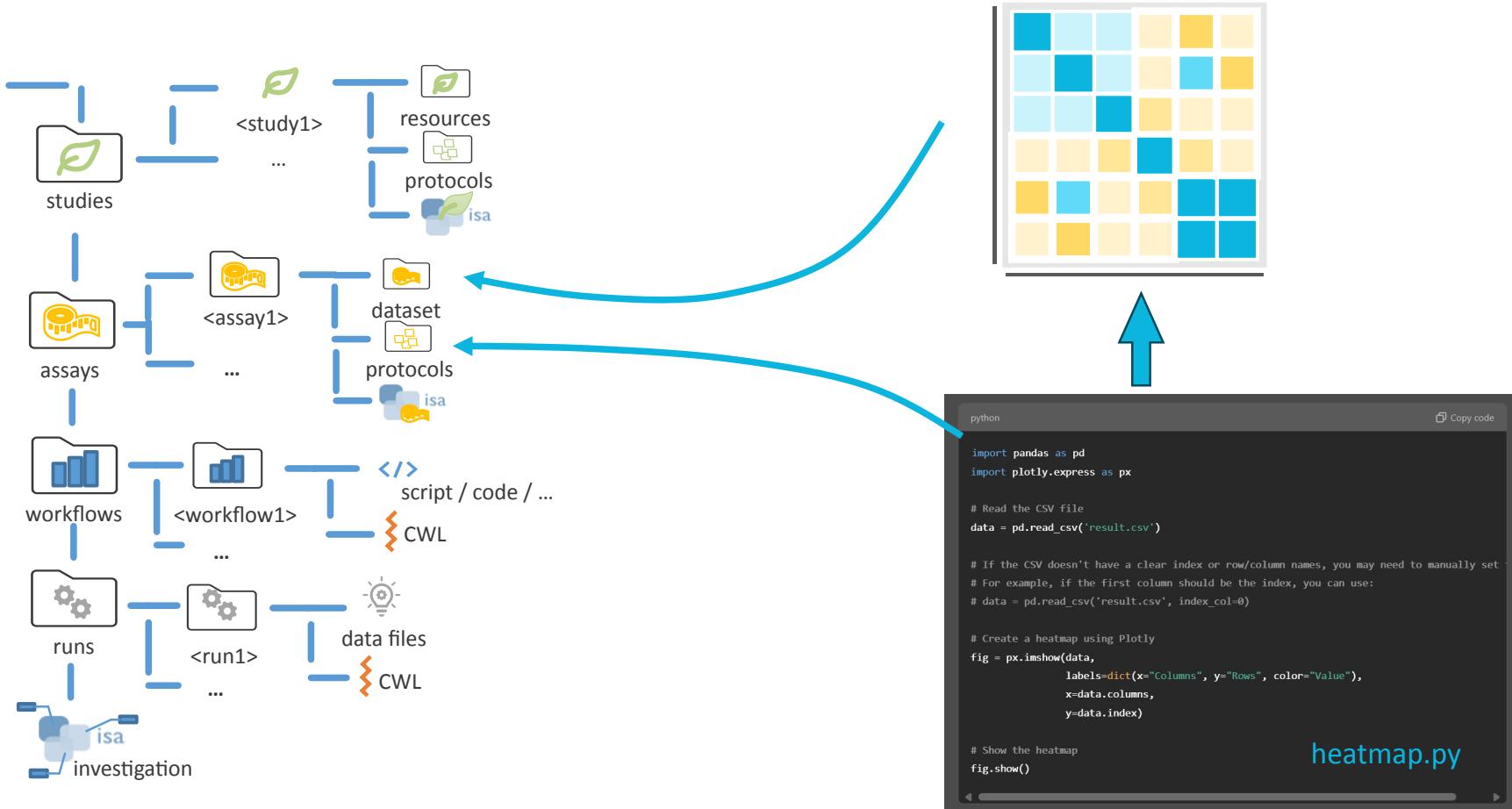
Options to annotate the data analysis



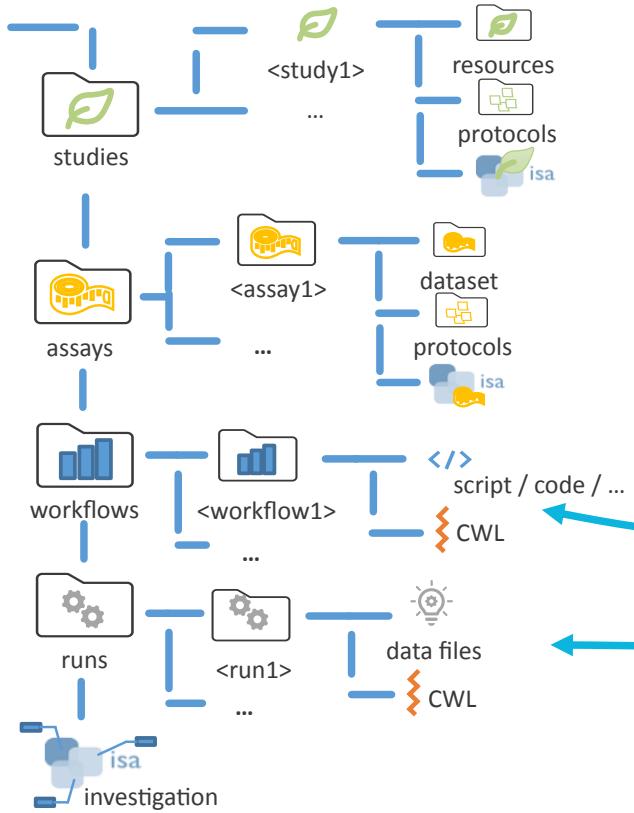
A computational workflow is like a protocol



Option I: Create a virtual assay



Option II: Create a workflow and run



A screenshot of a code editor showing a Python script named **heatmap.py**:

```
python
Copy code

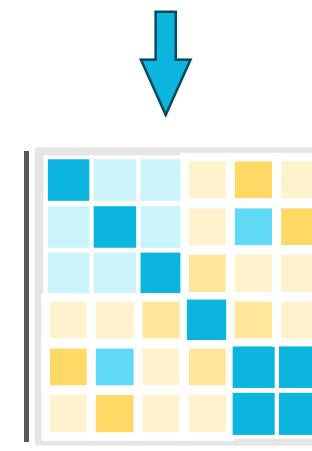
import pandas as pd
import plotly.express as px

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

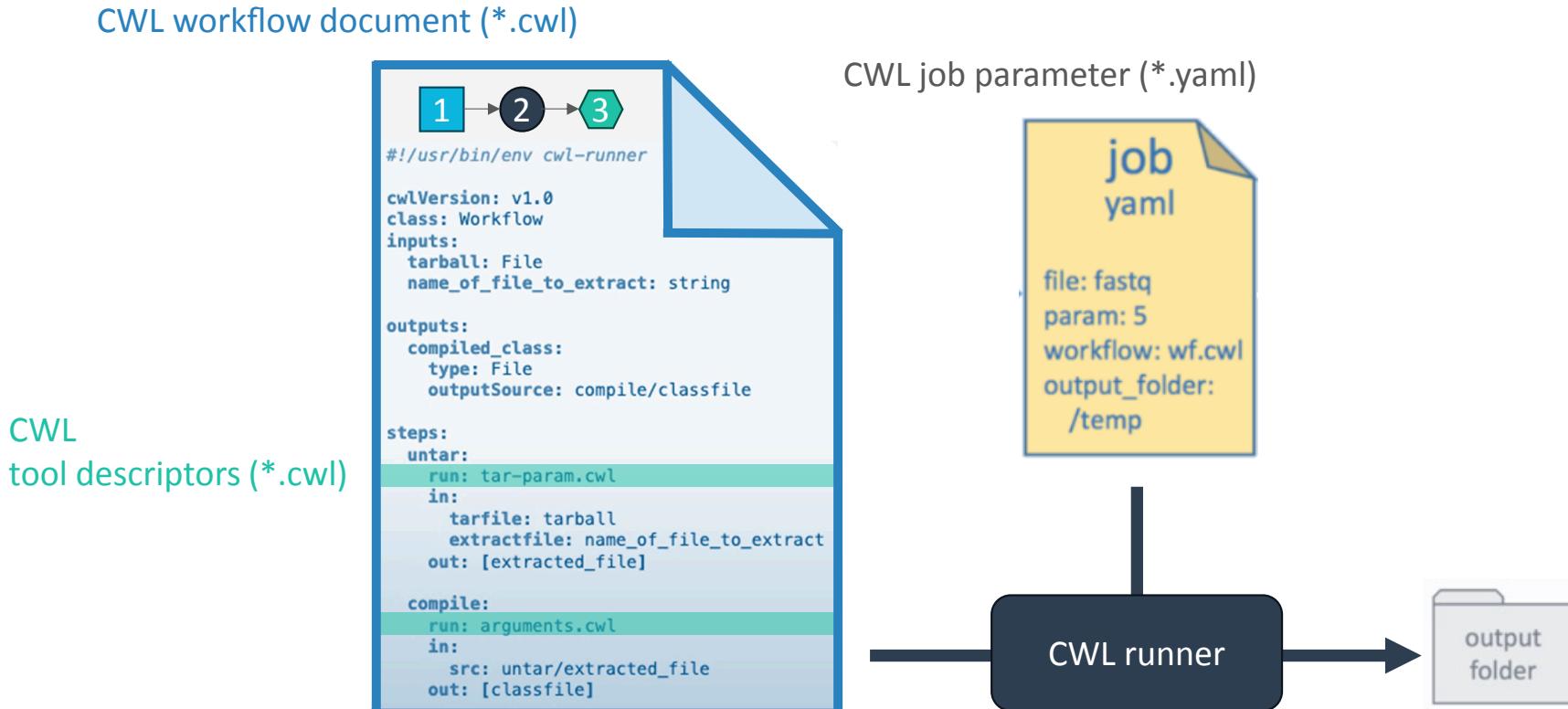
# If the CSV doesn't have a clear index or row/column names, you may need to manually set
# For example, if the first column should be the index, you can use:
# data = pd.read_csv('result.csv', index_col=0)

# 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()
```



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

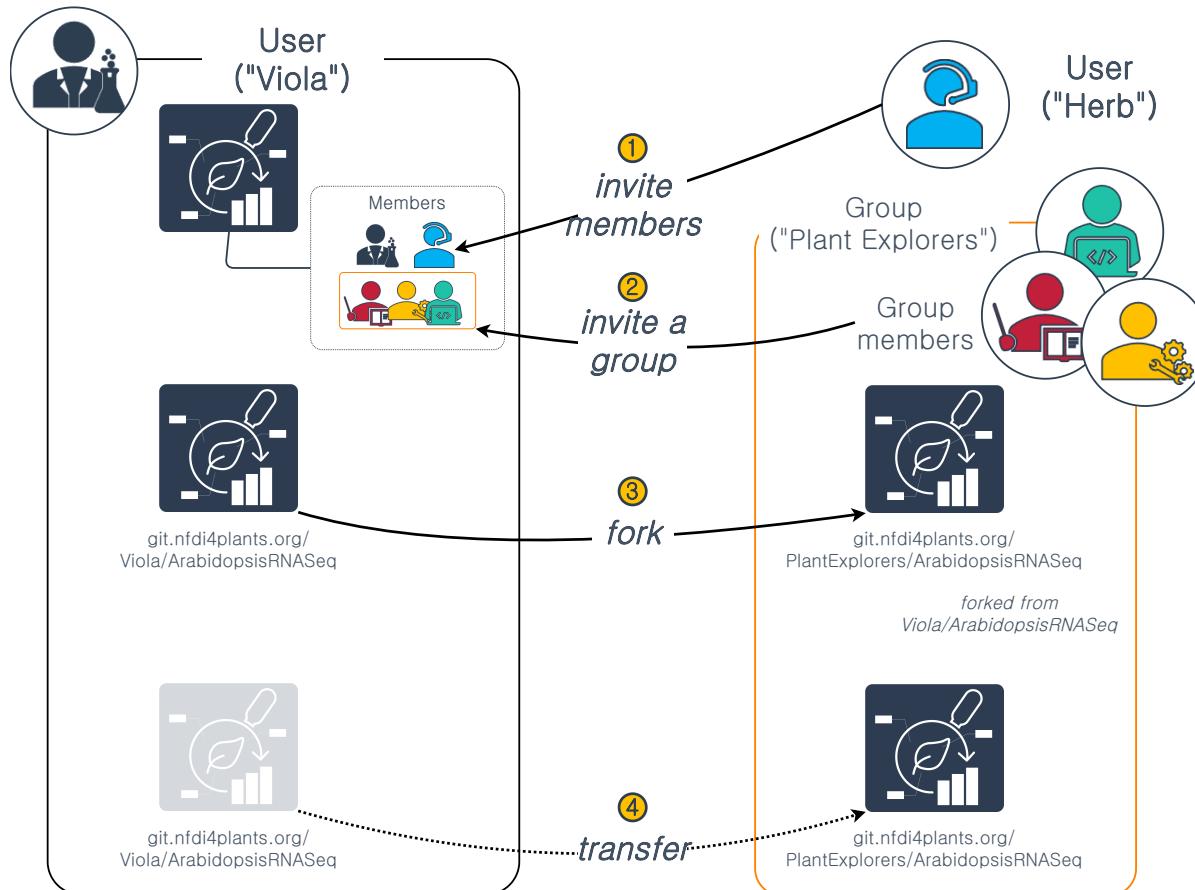
Project = ARC

- 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

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 members' input field is present. 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

The screenshot shows the 'ARCIct' interface for the 'Samuilov-2018-BOU-PSP' project. The left sidebar includes 'Login', 'New ARC', 'Open ARC', 'Download ARC', 'Save ARC', 'Explorer', 'Commit', 'DataHUB Sync', and 'History'. The right panel displays the project details and investigation contacts.

Identifier: Samuilov-2018-BOU-PSP

Title: Samuilov-2018-BOU-PSP

Description: (empty)

Contacts:

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

ARCIct: 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 for a project named "Demo_ARC". The interface includes a sidebar with navigation links (1-4), a main header with search and repository information (5), a breadcrumb trail (6), and a commit history table (7). The right side displays project information (8).

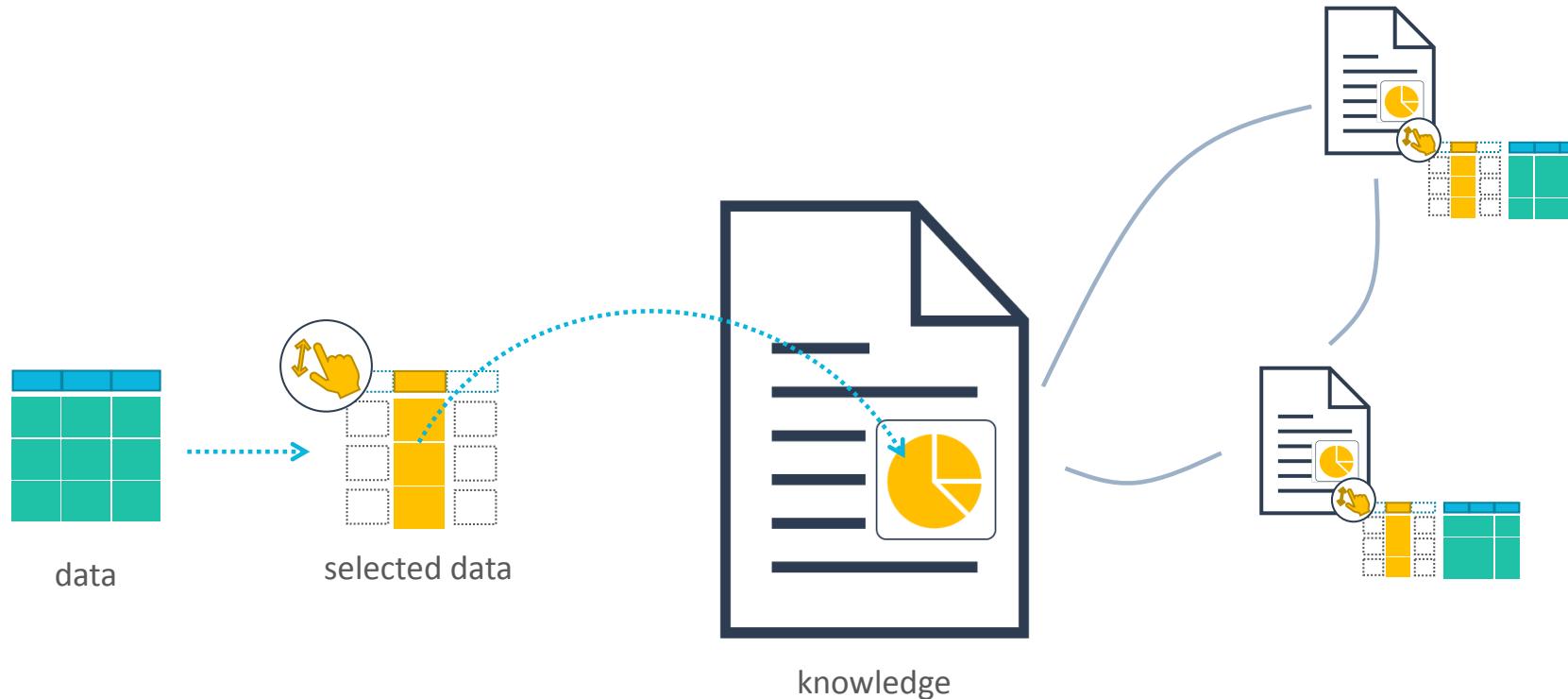
Project Information:

- pipeline: passed
- Star: 0
- Fork: 0
- Commits: 1 Commit
- Branches: 2 Branches
- Tags: 0 Tags
- Storage: 7 KIB Project Storage
- Auto DevOps enabled
- Add README
- Add LICENSE
- Add CHANGELOG
- Add CONTRIBUTING
- Add Kubernetes cluster
- Add Wiki
- Configure Integrations

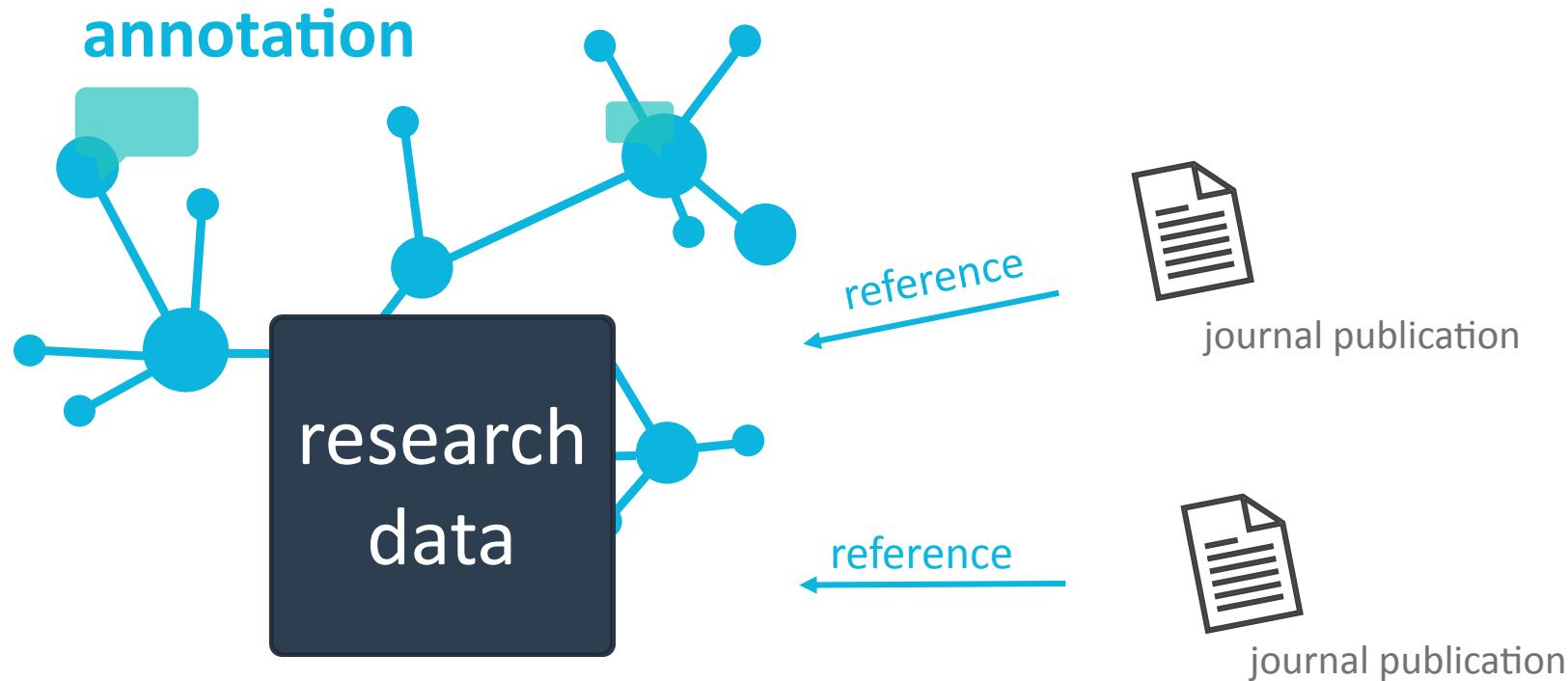
Created on: July 13, 2024

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

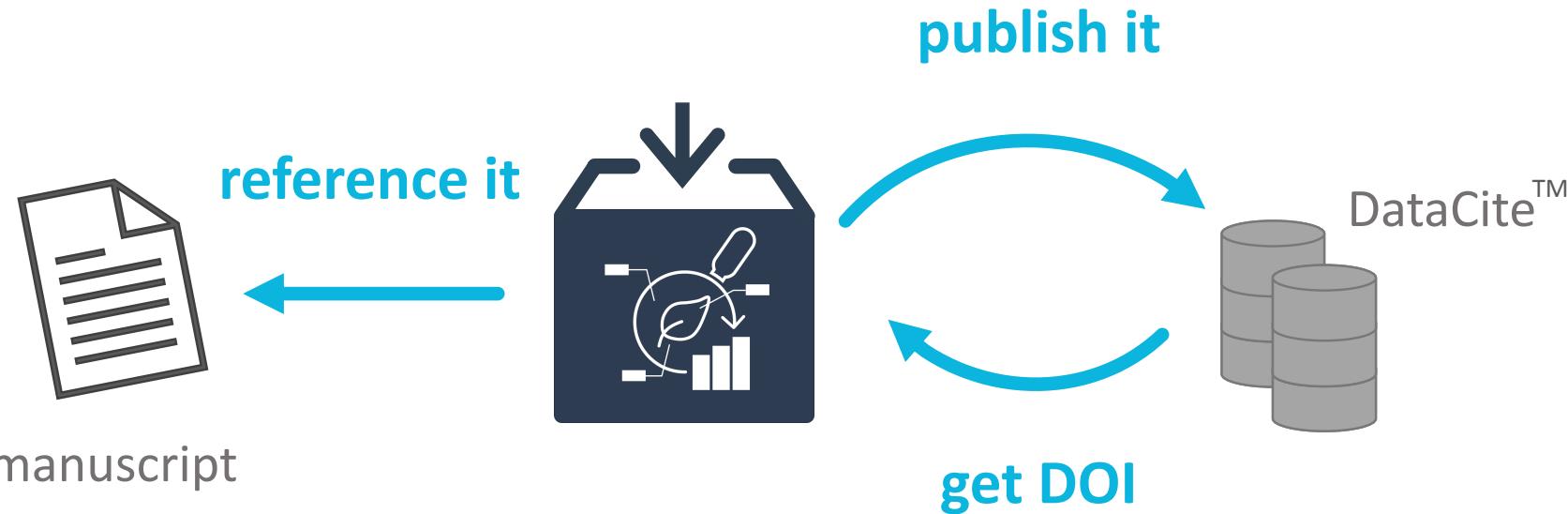
Hands-on part 5: ARC data publication



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

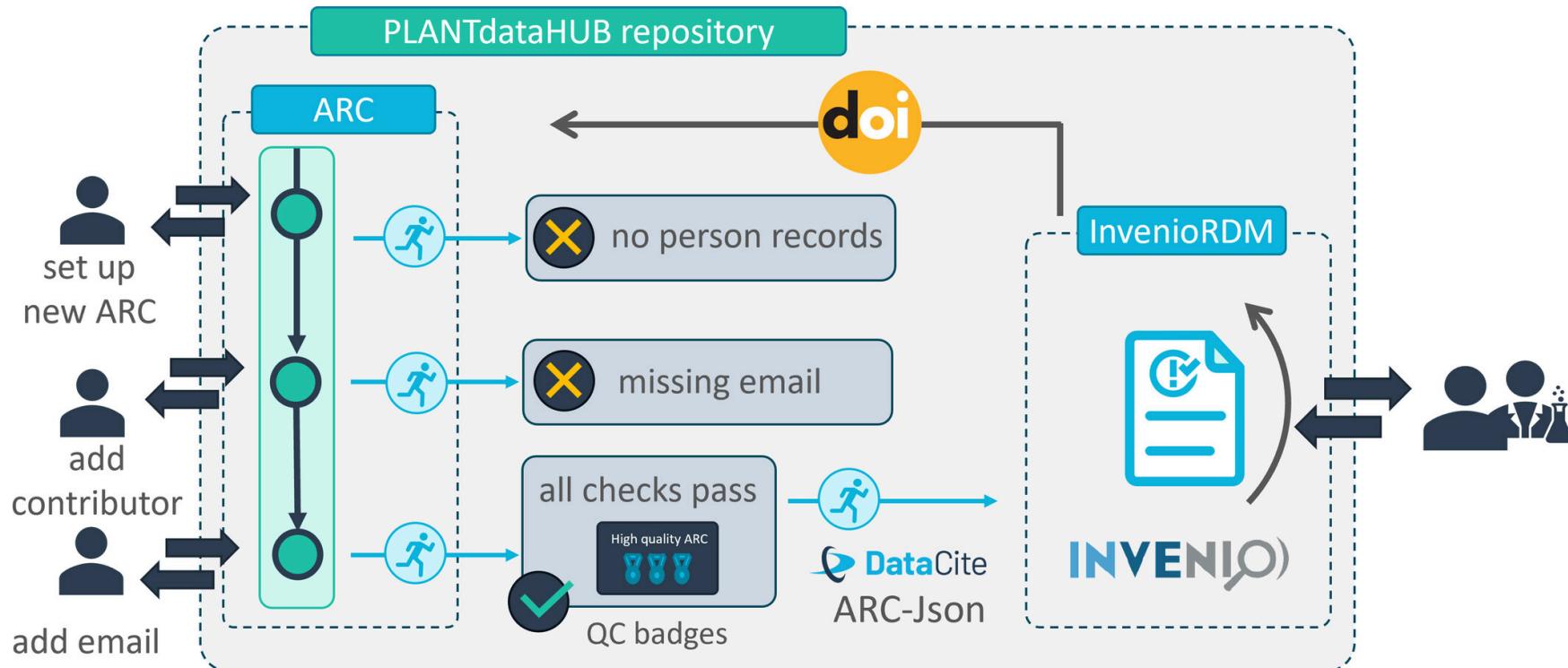
53 Commits 1 Branch 0 Tags 293.9 GB Project Storage

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, Ningning¹; Mattoon, Erin¹; McHargue, Will¹ ; Venn, Benedikt² ; Zimmer, David² ; Pecani, Kresti³; Jeong, Jooyeon¹; Anderson, Cheyenne⁴; Chen, Chen⁴; Berry, Jeffrey¹; Xia, Ming¹; Tzeng, Shin-Cheng¹ ; Becker, Eric¹; Pazouki, Leila¹; Evans, Bradley¹; Cross, Fred³; Cheng, Jianlin⁴; Czymmek, Kirk¹ ; Schröder, Michael⁵ ; Mühlhaus, Timo² ; Zhang, Ru¹ 

Show affiliations

Style APA 

1

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>

2

Details

DOI

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

Resource type

Dataset

Publisher

DataPLANT

Export

JSON

Export

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

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/>
- User Support Mailing List: [Click here to subscribe](#)

Open Source Development

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

Acknowledgements



Team Kaiserslautern

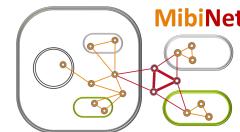
- Timo Mühlhaus
- Lukas Weil
- Kevin Frey
- Kevin Schneider
- Jonas Lukasczyk

Team Freiburg

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



- Björn Usadel
- Vittorio Tracanna
- Yaser Alashloo



- Sabrina Zander