



CEPLAS

Cluster of Excellence on Plant Sciences

Preparation

before November 11th, 2024

Dominik Brilhaus, [CEPLAS Data](#)



Checklist hands-on sessions

💡 Please prepare the following before the class:

- ✓ Register at DataPLANT
- ✓ Find your command line
- ✓ Install and configure Git on your computer
- ✓ Install ARC Commander on your computer
- ✓ Install ARCitect on your computer
- ✓ Install Swate on your computer
- ✓ (optional) Install VS Code

DataPLANT Registration

If you do not have a DataPLANT account, please register at the [DataPLANT website](#).

The screenshot shows a registration form titled "SIGN UP" with the sub-instruction "Get access to DataPLANT infrastructure and services". It features three input fields: "Email", "First name", and "Last name". Below these is a blue "Next" button with a right-pointing arrow. To the right of the form is a large blue text block explaining the benefits of registration, mentioning FAIR Digital Objects, seamless collaborations, and Git versioning.

SIGN UP

Get access to DataPLANT infrastructure and services

Email

First name

Last name

Next ▶

The infrastructure, tools, and workflows we offer support you in transforming your results into FAIR Digital Objects and enable seamless collaborations between you and your lab-members or even project partners from multiple labs. Thanks to the versioning feature of Git, every step is traceable at any time, preserving the provenance of each dataset. Do not hesitate and register using our Keycloak Single Sign-On solution.

Role and consortium

Please add your Project/consortium (e.g. CEPLAS) and choose the role Guest

The command line

Find the **command-line interface (CLI)** on your system.

- On Windows: Enter `powershell` into the explorer path
- On MacOS: Search `terminal` via spotlight (`⌘ + ⌂`) or navigate to `Applications` -> `Utilities` -> `Terminal`

 In our tutorials we sometimes use *terminal*, *command-line interface (CLI)* and *powershell* interchangeably.

Git Installation

Please install [Git](#) and [Git LFS](#) on your system

- 💡 Git LFS may already be installed with your Git installation (at least on Windows)
- 💡 For macOS we recommend to install via homebrew as described on the site above

Configuration of Git

Check the git user configuration on your system, by executing

```
git config --global --get-regexp user
```

This should prompt two lines

```
user.name <Your Name>
```

```
user.email <Your Email>
```

 Configuration needs to be done once after installation of git on your system.

Git configuration

Set the git user configuration on your system, by executing

1. Your name

```
git config --global user.name "Your Name"
```

2. Your email address

```
git config --global user.email "Your Email"
```

ARC Commander Installation

Please install the latest version of the ARC Commander and dependencies for your operating system according to the manual's [setup instructions](#).

Check if the ARC Commander is functional by displaying the ARC Commander version and help menu:

```
arc --version
```

Setup ▾

- [Installing Dependencies](#)
- [Configure Git](#)
- [Installing the ARC Commander](#)
 - [Windows](#)
 - [MacOS](#)
 - [Linux](#)
- [DataHUB Access](#)
- [Before we start](#)

ARCitect Installation

Please follow the instructions to install the latest version of ARCitect.

- [macOS](#)
- [Windows](#)

Swate Installation

Please follow [these instructions](#) to install the latest version of Swate.

Have a simple text editor ready

- Windows Notepad
- MacOS TextEdit

Recommended text editor with code highlighting, git support, terminal, etc: [Visual Studio Code](#)

Resources

DataPLANT (nfdi4plants)

Website: <https://nfdi4plants.org/>

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

DataHUB: <https://git.nfdi4plants.org>

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



CEPLAS

Cluster of Excellence on Plant Sciences

Good Data Management Practices

*part of M4468 – Plant developmental genetics, evolution
and biostatistics in the CEPLAS research program*

November 8th, 2023

Dominik Brilhaus, CEPLAS Data Science



Welcome

House-keeping

Pad: <https://pad.hhu.de/oI-NjeUtSHSMzk5huWRkJw>

Materials

Slides will be shared via DataPLANT knowledge base and the Sciebo folder

Tentative agenda

Day 1

| Time | Topics |
|---------------|--------------------------|
| 09:30 - 10:45 | Intro to RDM and ARC |
| 10:45 - 11:00 | <i>Short break</i> |
| 11:00 - 12:00 | ARC Hands-on |
| 12:30 - 13:30 | <i>Lunch</i> |
| 13:00 - 15:30 | Data storage and sharing |
| 15:30 - 16:00 | Wrap-up |

Day 2

| Time | Topics |
|---------------|------------------------------------|
| 09:30 - 10:30 | ARC Feedback session |
| 10:30 - 10:45 | <i>Short break</i> |
| 10:45 - 12:00 | ISA and Metadata |
| 12:30 - 13:30 | <i>Lunch</i> |
| 13:00 - 15:00 | Hands-on Swate |
| 15:00 - 15:30 | ARC ecosystem: Additional features |
| 15:30 - 16:00 | Wrap-up |

Goals

- Appreciate FAIR principles
- Tools and services for FAIR data management
- Effectively manage your own research data
- Communication and terminology

Why Research Data Management (RDM)?

- Increase transparency
- Make data accessible
- Save time (writing, reusing)
- Reduce the risk of data loss
- Optimize the costs
- Facilitate future reuse and sharing
- Improve citations

How is your data analysis going?

Can't understand the data

... and the data collector
does not answer my
emails or my phone calls

That is terrible and so
cruel !

Who is it, who collected the
data ?

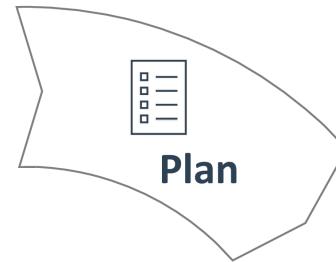
I did... 3 years ago



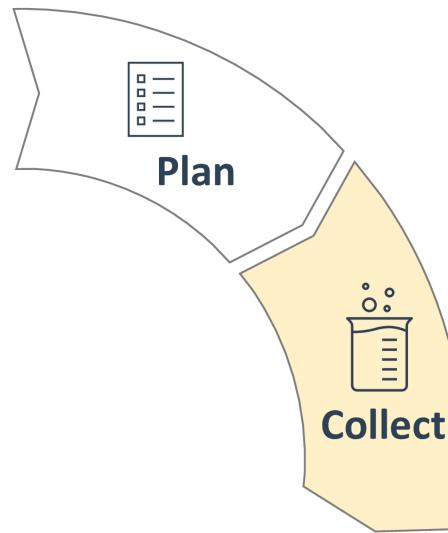
Your first collaborators
are your future selves,
be nice to them !

your future self, by Julien Colomb, CC-BY-NC, derived from .NORM Normal File Format, CC-BY-NC, by Randall Munroe

The Research Data Lifecycle



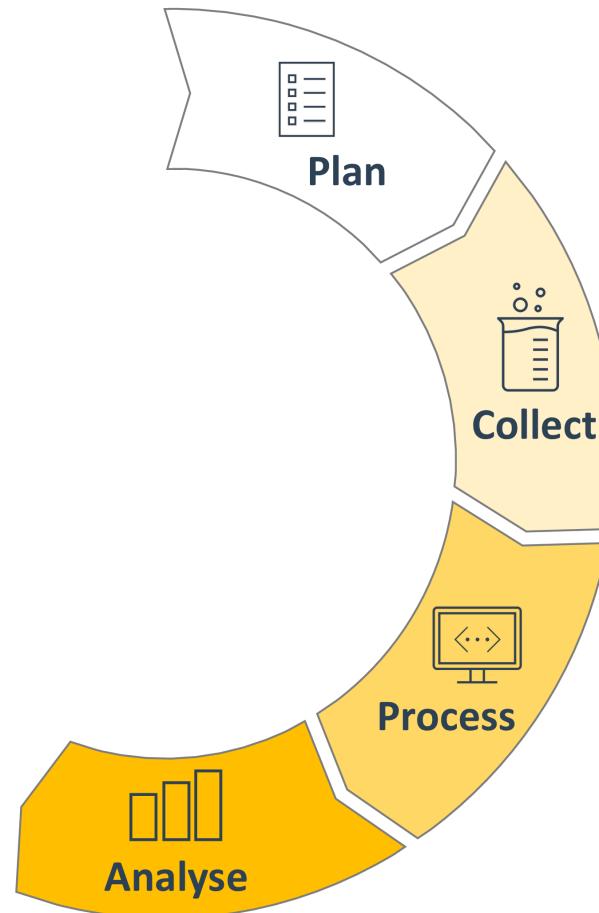
The Research Data Lifecycle



The Research Data Lifecycle



The Research Data Lifecycle



The Research Data Lifecycle



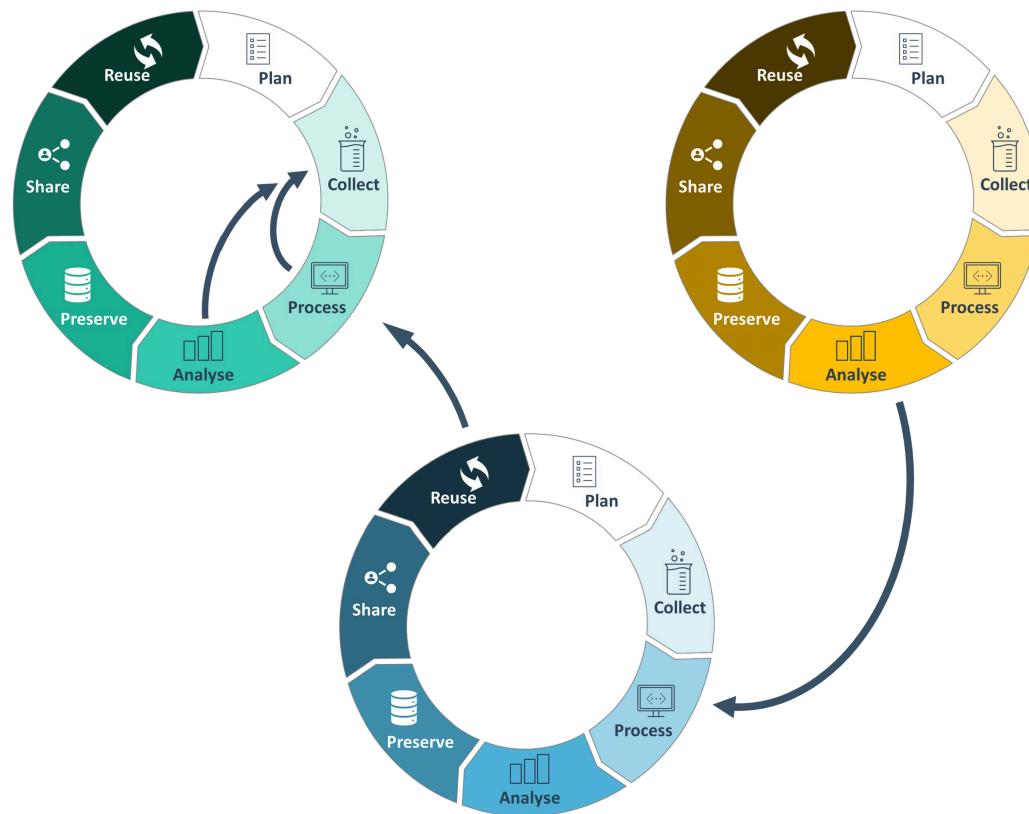
The Research Data Lifecycle



The Research Data Lifecycle



The Research Data Lifecycle *is mutable*



FAIR

- Findable
- Accessible
- Interoperable
- Reusable

<https://doi.org/10.1038/sdata.2016.18>

[nature](#) > [scientific data](#) > [comment](#) > [article](#)

[Open Access](#) | [Published: 15 March 2016](#)

The FAIR Guiding Principles for scientific data management and stewardship

[Mark D. Wilkinson](#), [Michel Dumontier](#), [IJsbrand Jan Aalbersberg](#), [Gabrielle Appleton](#), [Myles Axton](#), [Arie Baak](#), [Niklas Blomberg](#), [Jan-Willem Boiten](#), [Luiz Bonino da Silva Santos](#), [Philip E. Bourne](#), [Jildau Bouwman](#), [Anthony J. Brookes](#), [Tim Clark](#), [Mercè Crosas](#), [Ingrid Dillo](#), [Olivier Dumon](#), [Scott Edmunds](#), [Chris T. Evelo](#), [Richard Finkers](#), [Alejandra Gonzalez-Beltran](#), [Alasdair J.G. Gray](#), [Paul Groth](#), [Carole Goble](#), [Jeffrey S. Grethe](#), [Jaap Heringa](#), [Peter A.C. 't Hoen](#), [Rob Hooft](#), [Tobias Kuhn](#), [Ruben Kok](#), [Joost Kok](#), [Scott J. Lusher](#), [Maryann E. Martone](#), [Albert Mons](#), [Abel L. Packer](#), [Bengt Persson](#), [Philippe Rocca-Serra](#), [Marco Roos](#), [Rene van Schaik](#), [Susanna-Assunta Sansone](#), [Erik Schultes](#), [Thierry Sengstag](#), [Ted Slater](#), [George Strawn](#), [Morris A. Swertz](#), [Mark Thompson](#), [Johan van der Lei](#), [Erik van Mulligen](#), [Jan Velterop](#), [Andra Waagmeester](#), [Peter Wittenburg](#), [Katherine Wolstencroft](#), [Jun Zhao](#) & [Barend Mons](#) 

— Show fewer authors

[Scientific Data](#) 3, Article number: 160018 (2016) | [Cite this article](#)

The FAIR principles

| | | | |
|---|---|---|---|
| <p>Findable Accessible Interoperable Reusable</p> | <p>Easier collaboration & sharing</p> <pre>graph TD; A(()) --> B(()); A(()) --> C(()); A(()) --> D(()); B(()) --> E(()); C(()) --> E(()); D(()) --> E(());</pre> | <p>Increased findability and visibility</p> | <p>Reproducibility</p> |
| <p>Added-value to the research community</p> <p>nfdi NCBI EMBL-EBI</p> | <p>Compliance with funding policies</p> | <p>Receive due credit</p> <p>Reuse Citations</p> | <p>Saves time & workload</p> <p>FAIR Time wasted</p> |

Is your data FAIR?

Findable | Accessible | Interoperable | Reusable

- Where do you store your data?
- How do you annotate your data?
- How do you share your data?
- What tools do you use to analyse your data?
- How do you reuse other people's data?

Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

- F1. (Meta)data are assigned a globally unique and persistent identifier.
- F2. Data are described with rich metadata (defined by R1 below).
- F3. Metadata clearly and explicitly include the identifier of the data they describe.
- F4. (Meta)data are registered or indexed in a searchable resource.

Accessible

Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorisation.

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
 - A1.1 The protocol is open, free, and universally implementable
 - A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles.
- I3. (Meta)data include qualified references to other (meta)data.

Reusable

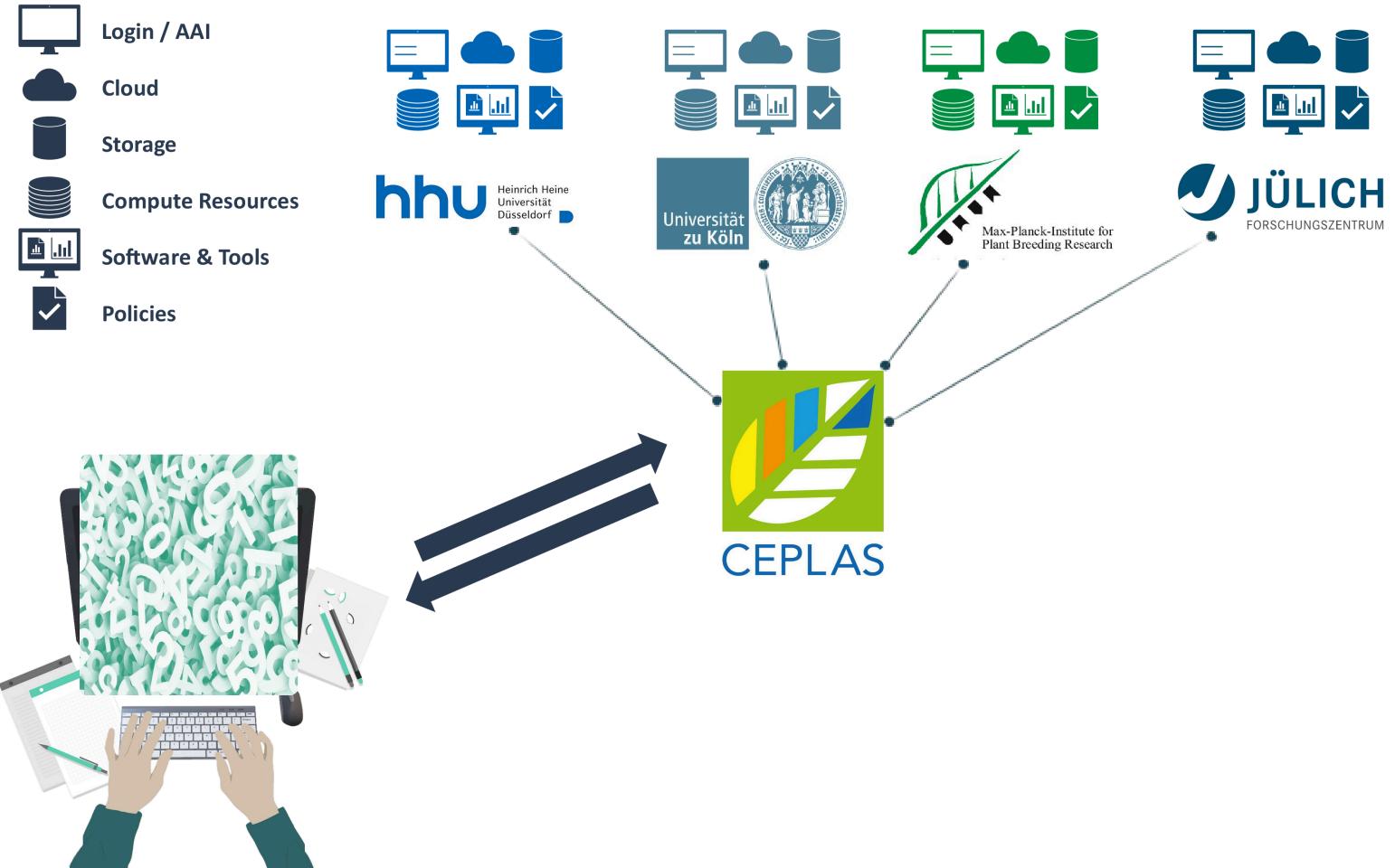
The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes
- R1.1. (Meta)data are released with a clear and accessible data usage license
- R1.2. (Meta)data are associated with detailed provenance
- R1.3. (Meta)data meet domain-relevant community standards

FAIR on multiple layers

The principles refer to three types of entities: **data** (or any digital object), **metadata** (information about that digital object), and **infrastructure**.

Scattered Data Silos



Scattered Data Silos

FAIR Data for everyone



Contributors

Slides presented here include contributions by

- name: Dominik Brilhaus
github: <https://github.com/brilator>
orcid: <https://orcid.org/0000-0001-9021-3197>



Understand your colleague's project



1. Go to the shared folder prepared before class
2. Try to identify one experiment that led to results (e.g. a figure in the thesis)
3. What are the samples (e.g. plants, bacteria)?
4. How were the samples prepared (~ materials)?
5. How was the experiment performed (~ methods)?
6. What is the raw data (~ results)?
7. How was the data analyzed (~ computational methods, statistics)?
8. Collect all of the above in a `README_<YourArbitraryParticipantID>.txt` in the same folder.

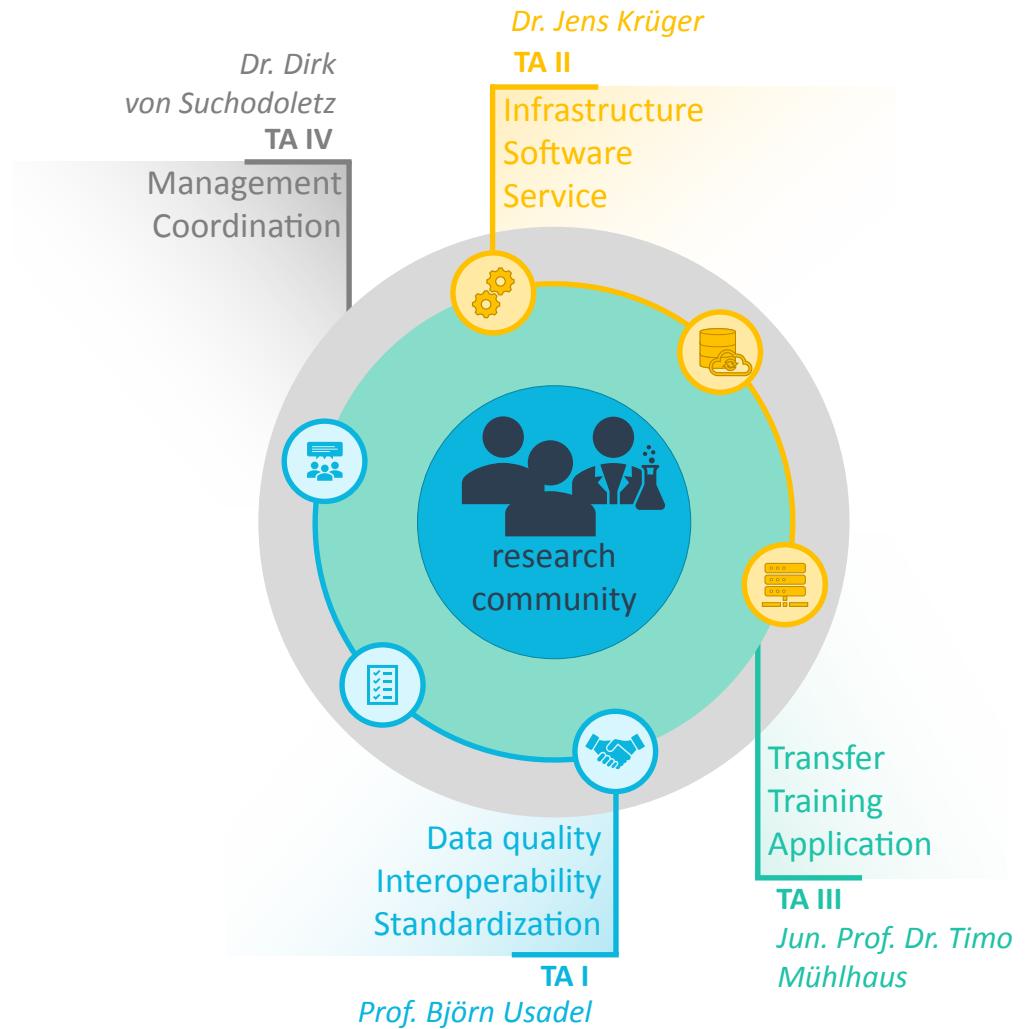
Assignment

| Participant | looks at project of |
|---------------|---------------------|
| Participant01 | Participant02 |
| Participant02 | Participant03 |
| ... | ... |
| Participant n | Participant n+1 |

Intro to DataPLANT and ARC

DataPLANT – The NFDI4Plants

- NFDI: "Nationale Forschungsdaten Infrastruktur" – www.nfdi.de
- Funded since end of 2020



Data Stewardship between DataPLANT and the community

Community



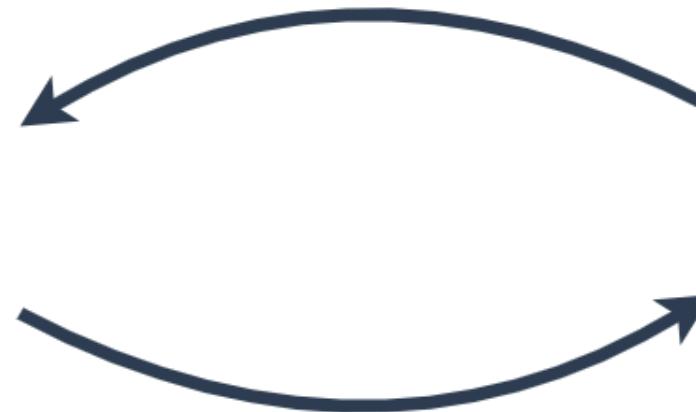
CEPLAS

Domain experts
User experience
Training

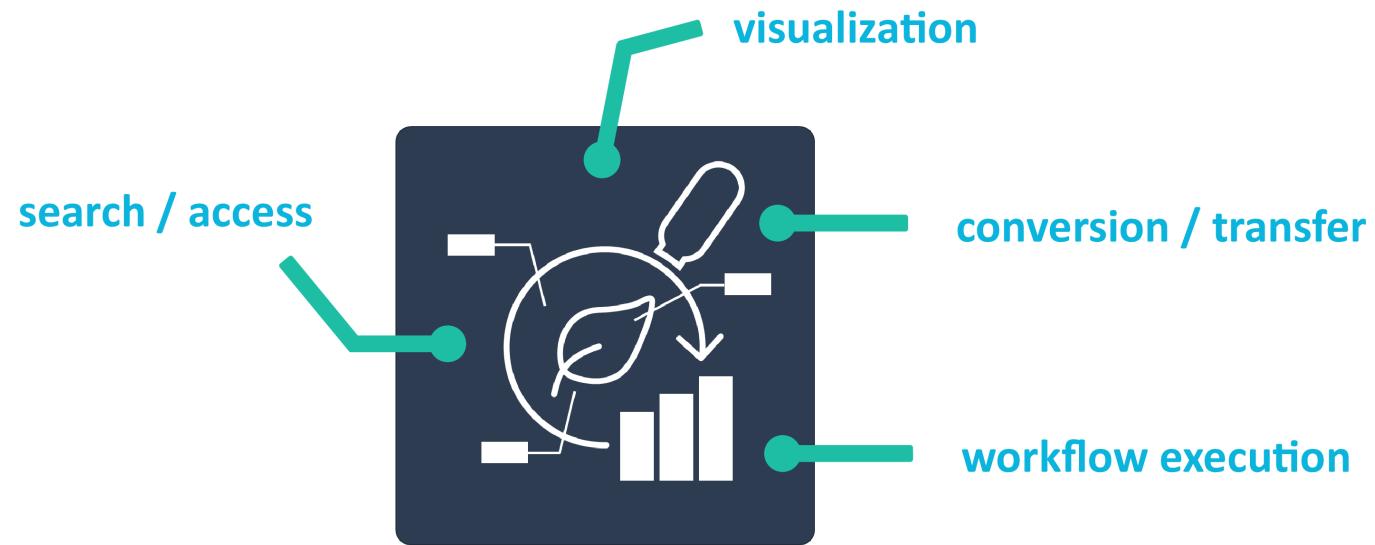
nfdi4plants



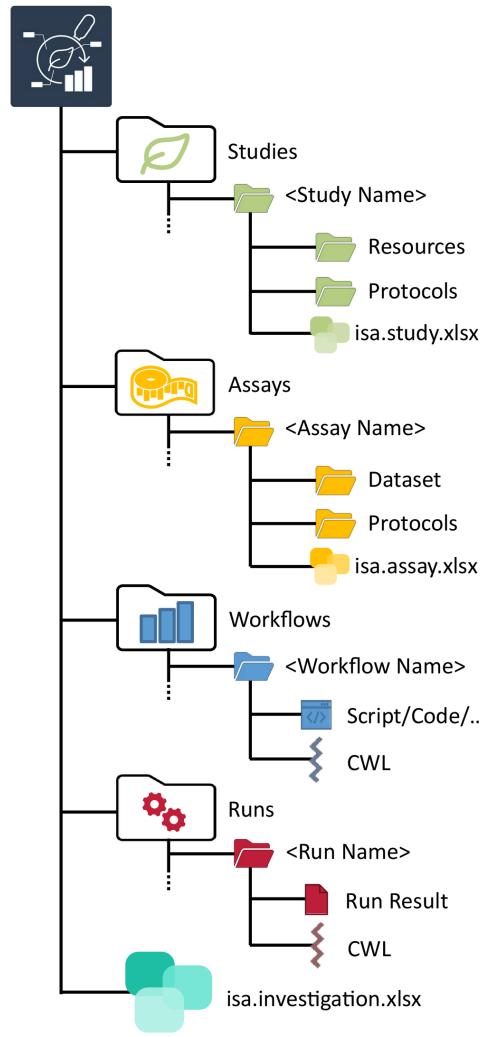
Service provider
Developers
Tech experts



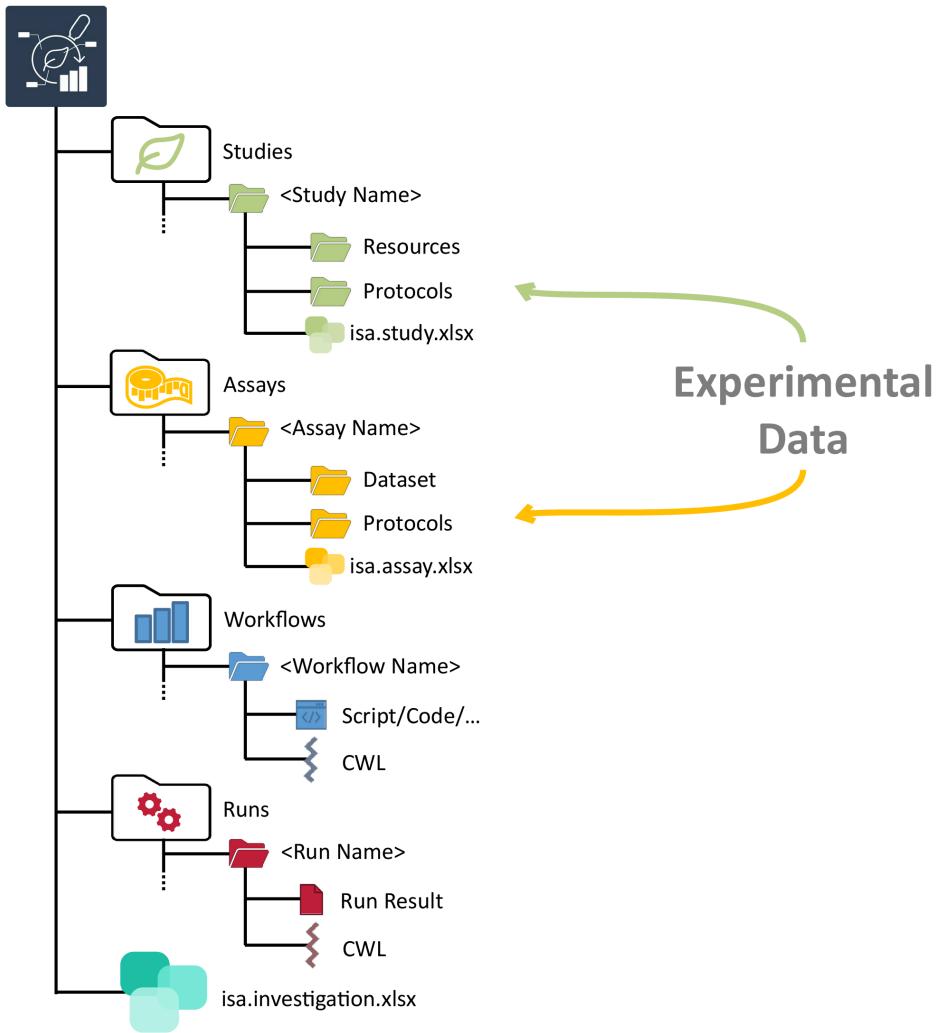
Annotated Research Context (ARC)



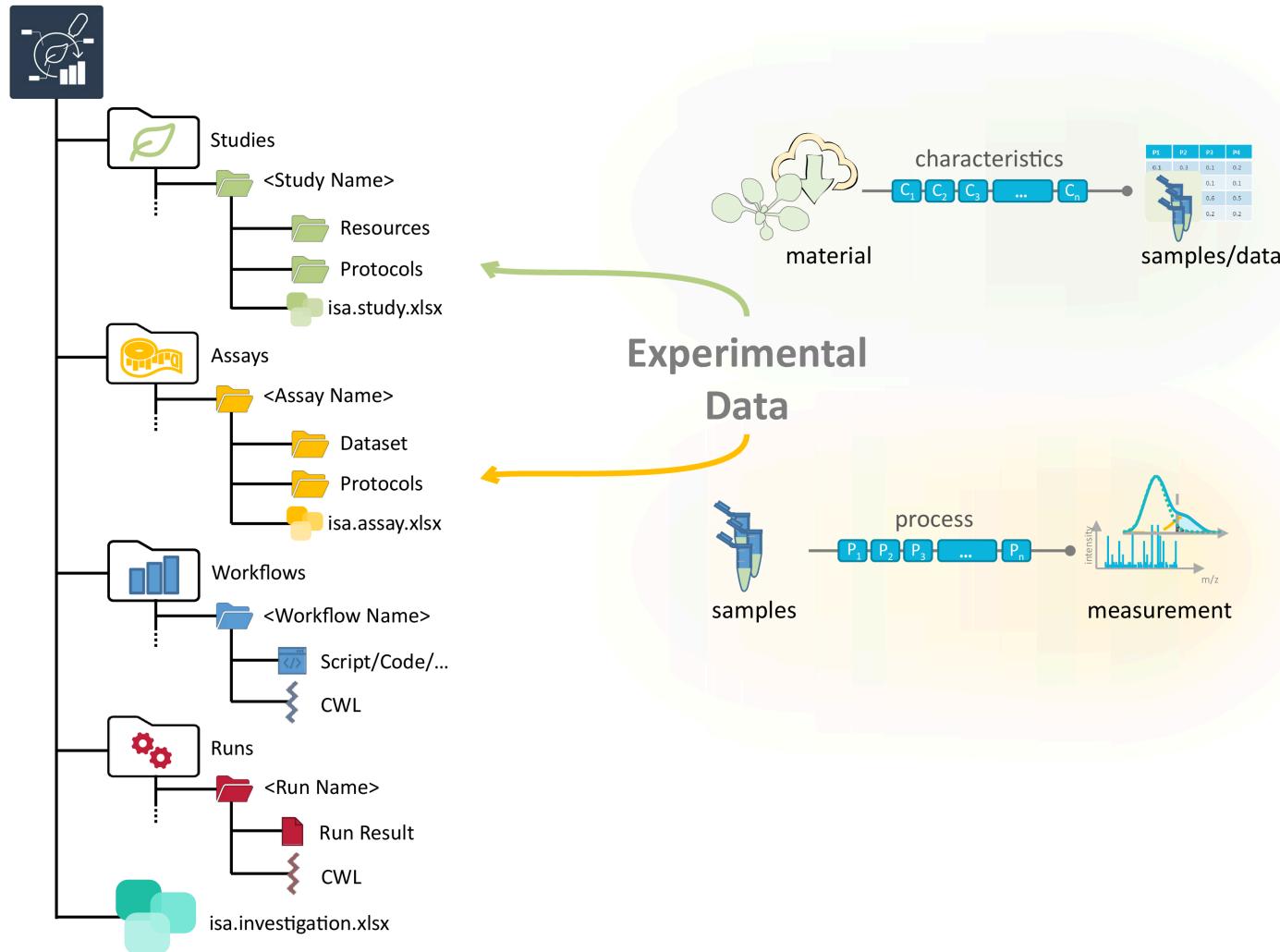
What does an ARC look like?



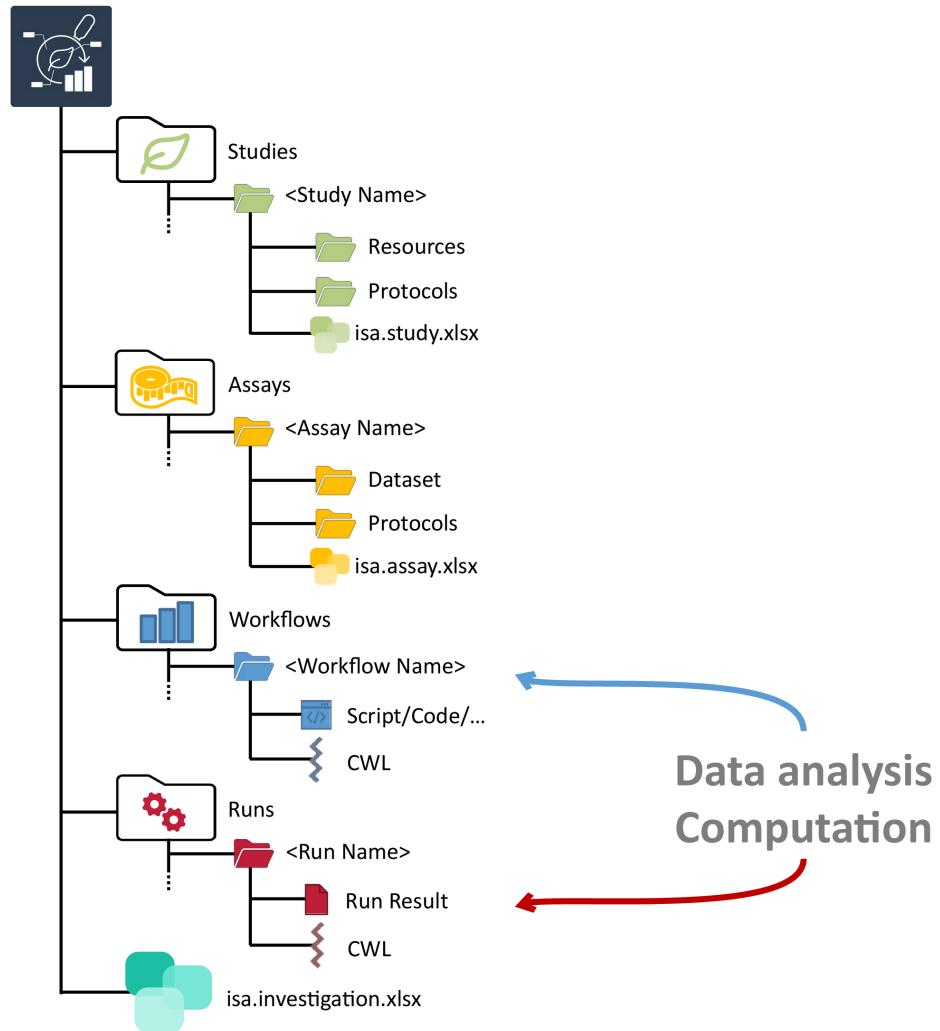
What does an ARC look like?



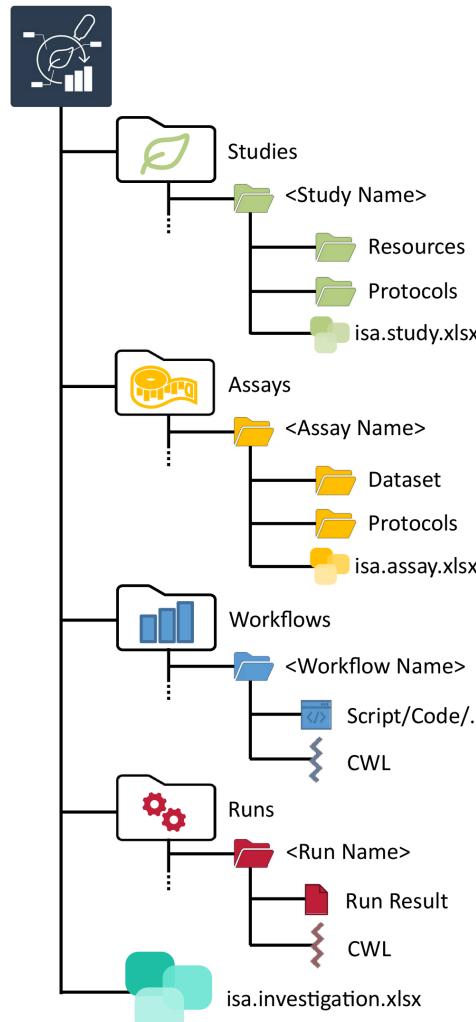
What does an ARC look like?



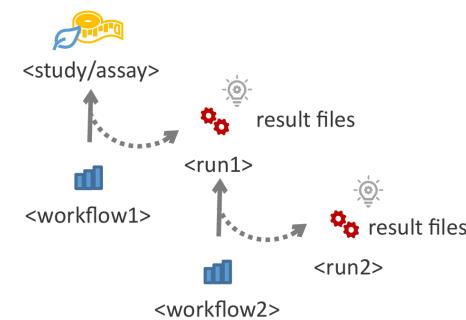
What does an ARC look like?



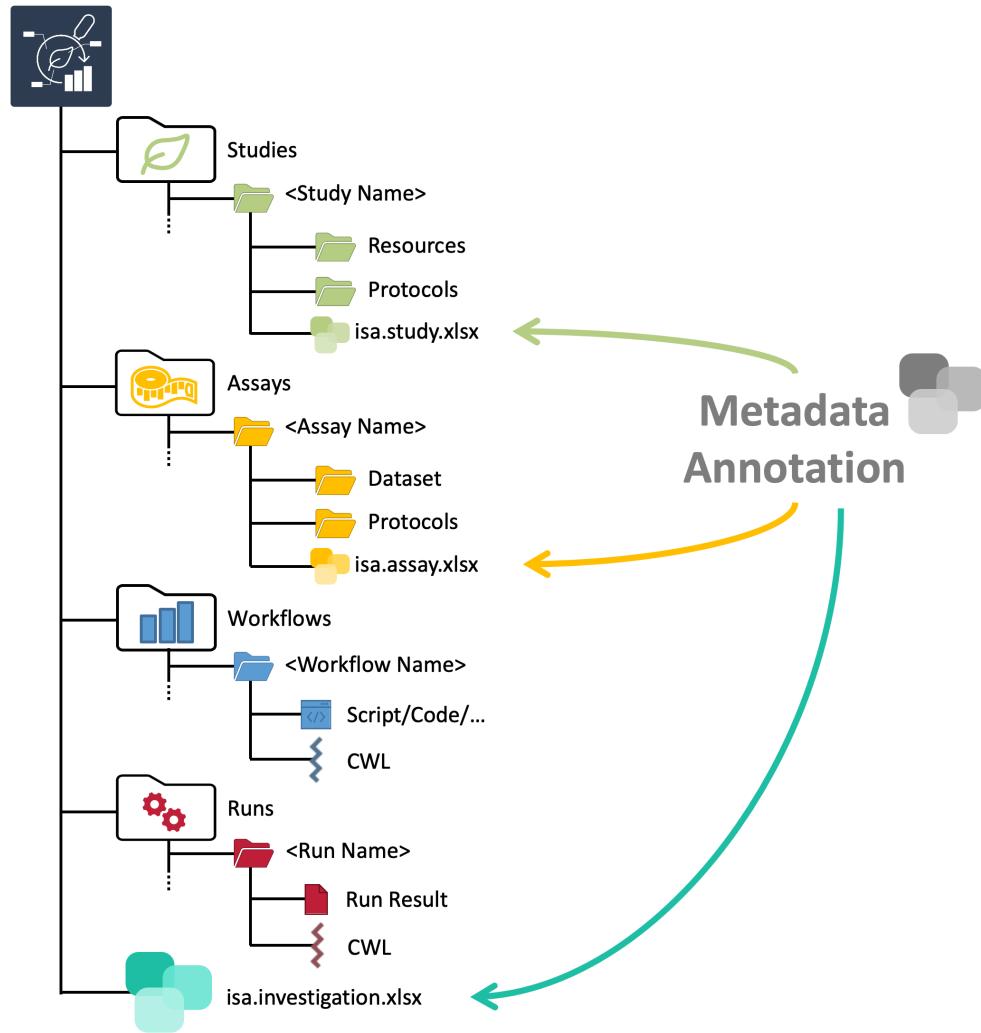
What does an ARC look like?



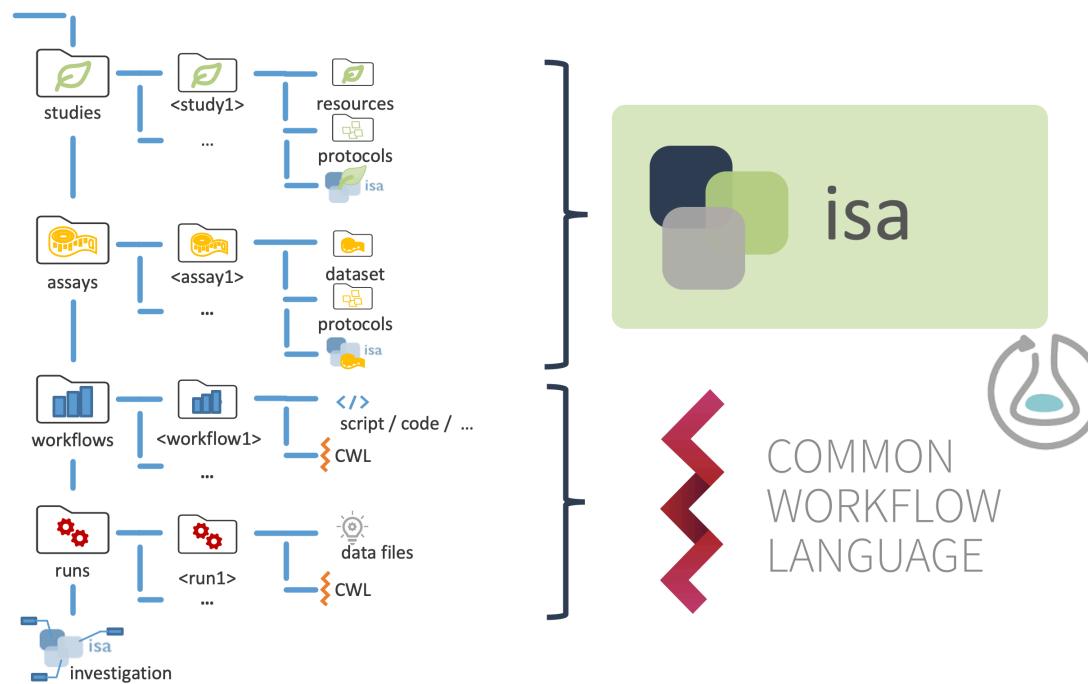
Data analysis
Computation



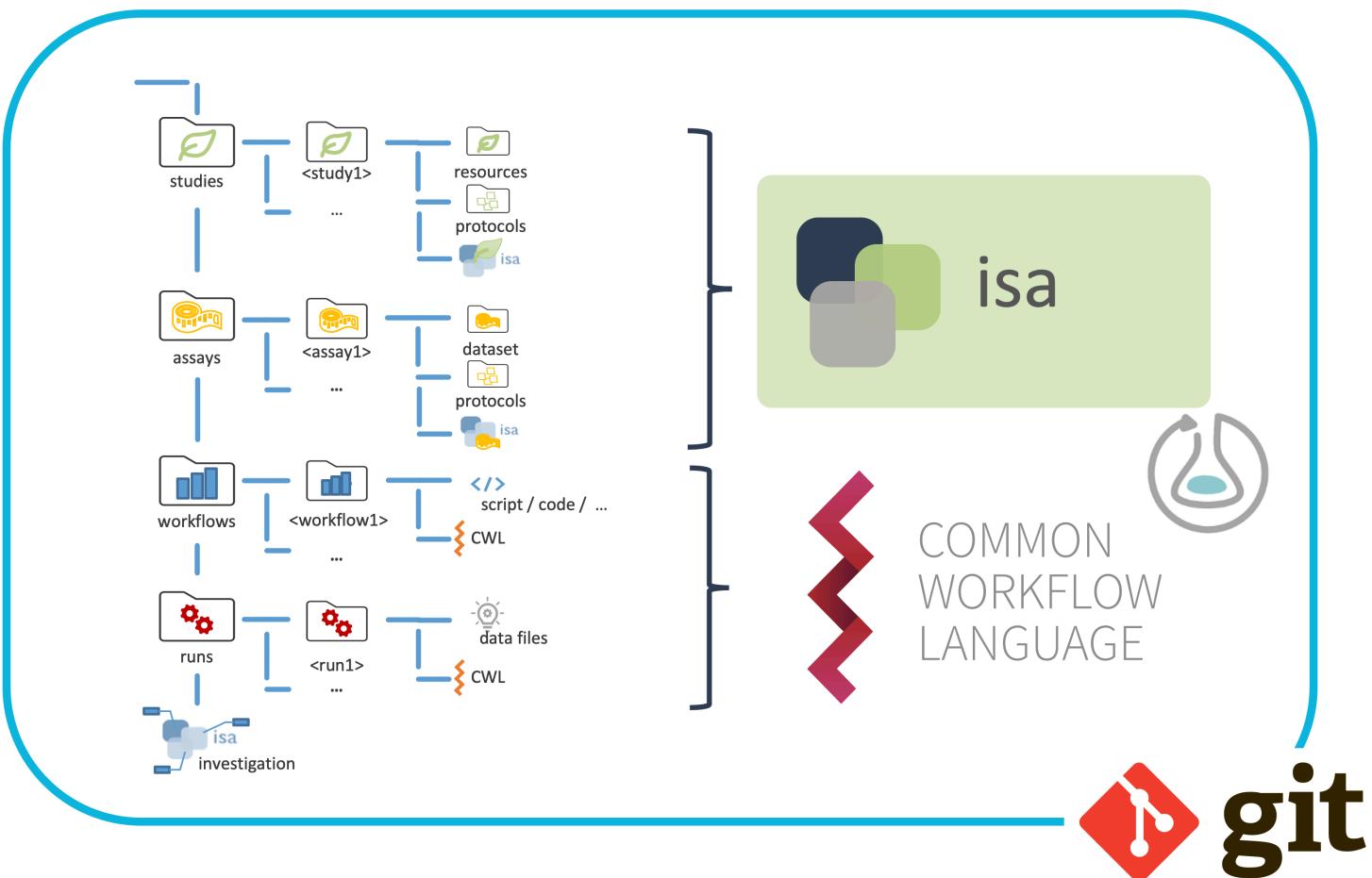
What does an ARC look like?



ARC builds on standards



ARC builds on standards + Git



Contributors

Slides presented here include contributions by

- name: Dominik Brilhaus
github: <https://github.com/brilator>
orcid: <https://orcid.org/0000-0001-9021-3197>
- name: Cristina Martins Rodrigues
github: <https://github.com/CMR248>
orcid: <https://orcid.org/0000-0002-4849-1537>
- name: Martin Kuhl
github: <https://github.com/Martin-Kuhl>
orcid: <https://orcid.org/0000-0002-8493-1077>

ARC Commander Hands-on

Registration

Everyone [signed-up](#) at the DataHUB?

Check your installation

Open a terminal and one after the other execute

```
git --version
```

```
git-lfs --version
```

```
arc --version
```

 If you see a warning at any of these, let us know.

Config

```
git config --global --get-regexp user
```

 If this does not display your user name and email, you need to [configure git](#).

Have a simple text editor ready

- Windows Notepad
- MacOS TextEdit

Recommended text editor with code highlighting, git support, terminal, etc: [Visual Studio Code](#)

Create a fresh folder for your ARCs

For this workshop, create a new folder somewhere on your machine where you want to store ARCs, e.g. in your documents folder:

- `C:\Users\<username>\Documents\workshop-arcs` (windows)
- `~/Documents/workshop-arcs` (mac)

 Ideally this folder is not "watched" by any cloud service (Sciebo, google drive, iCloud, etc.)

Hands-on with demo data

First steps towards your ARC using the **ARC Commander**

Download the demo data

```
git clone "https://demo-user:1_eznikmzxzARAbUxxnF@git.nfdi4plants.org/teaching/demo-arc_level0.git"
```

You just received your data

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

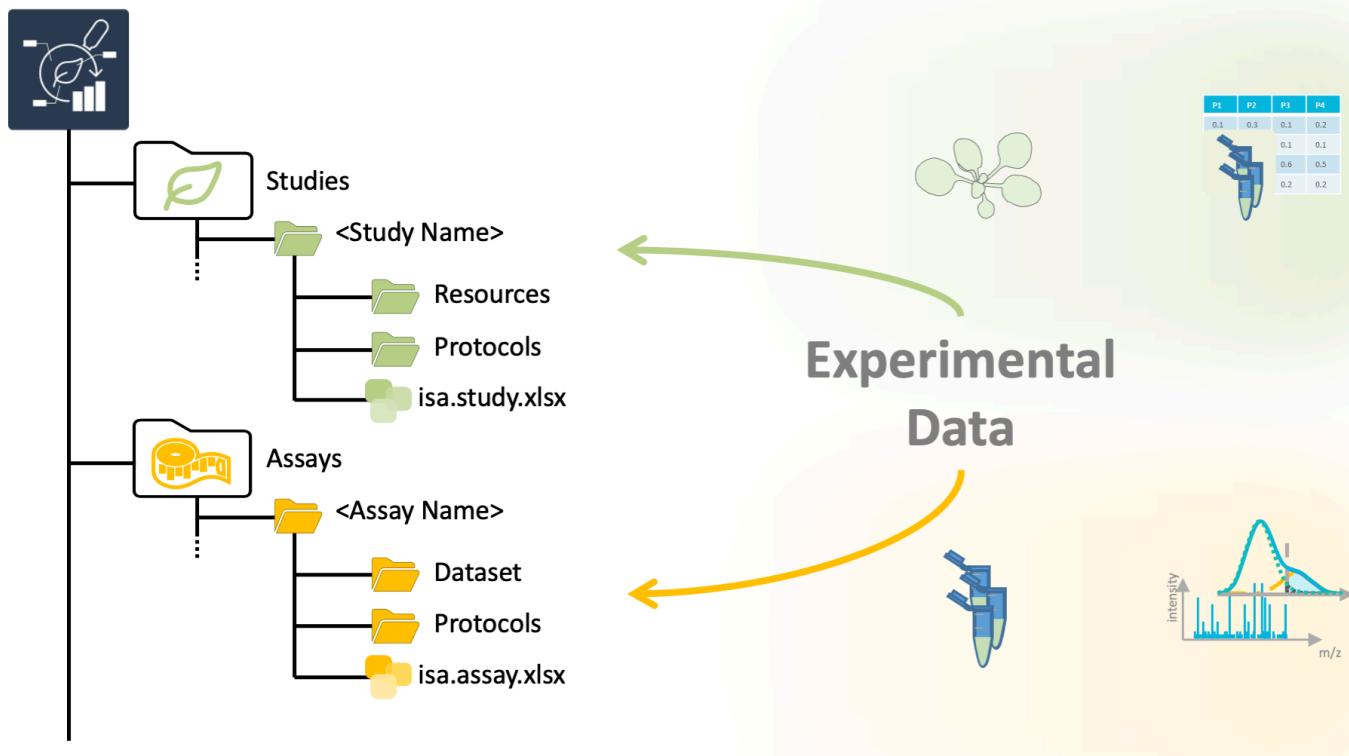
Goal

- Structure,
- Annotate, and
- Share your experimental data.



We'll talk about data annotation later

Structure your data



Your fresh ARC folder

1. Create a new folder, which you want to initialize as an ARC.
2. Open the command line inside the folder or navigate via command line to that folder.

For example:

```
mkdir -p ~/Documents/workshop-arcs/arc-demo  
cd ~/Documents/workshop-arcs/arc-demo
```

Initiate the ARC folder structure

```
arc init
```

Create an investigation

```
arc investigation create -i TalinumPhotosynthesis
```

Add a person

```
arc i person register --lastname Brilhaus --firstname Dominik --email brilhaus@hhu.de --affiliation CEPLAS
```

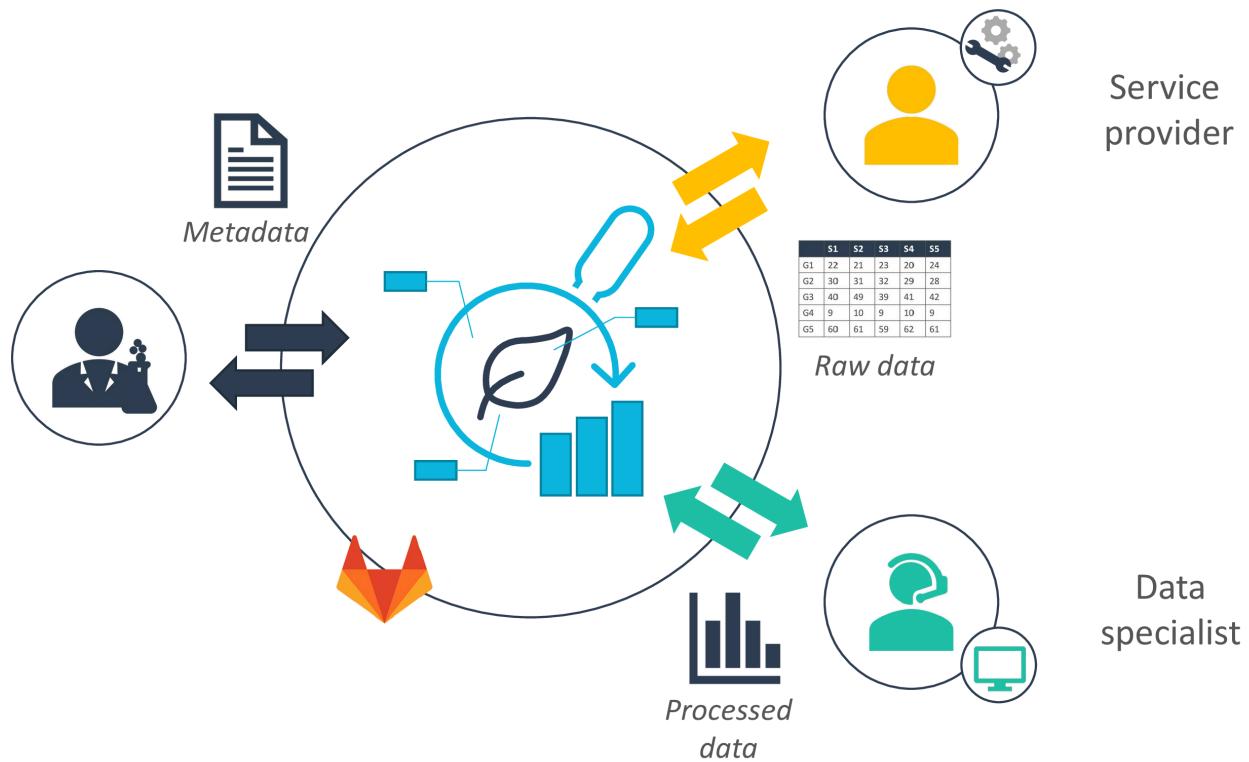
Add a study

```
arc study add -s talinum_drought
```

Add assays

```
arc assay add -s talinum_drought -a rnaseq  
arc assay add -s talinum_drought -a metabolomics
```

Collaborate and share

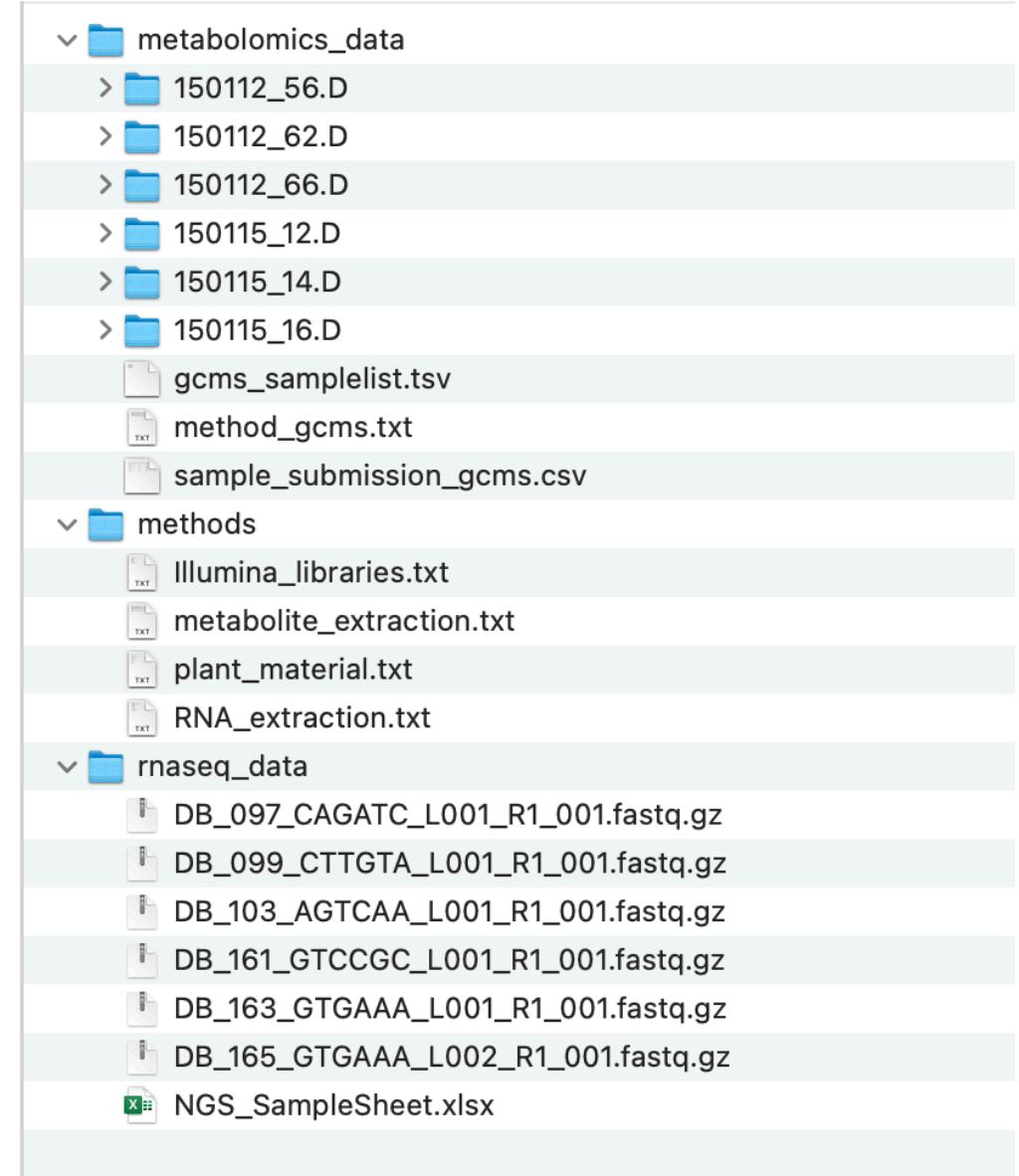


Upload your local ARC to the DataHUB

```
arc sync -f -r https://git.nfdi4plants.org/<username>/arc-demo
```

Sort the demo data into the ARC

Identify "raw dataset(s)" and "protocols" and move them to the proper subfolders in the ARC.



Sync your ARC to the DataHUB

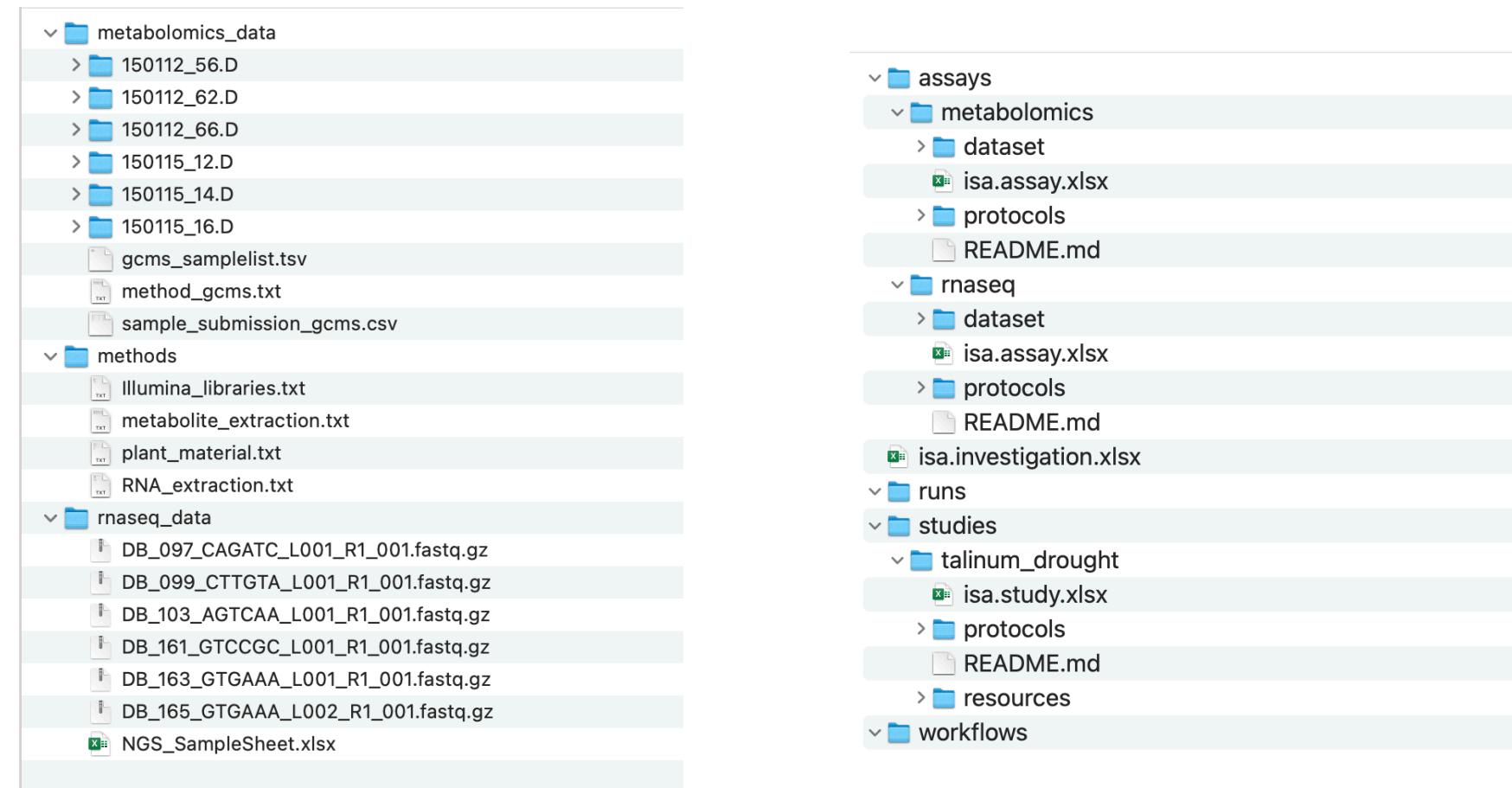
To save the changes, sync the ARC to the DataHUB including a message.

```
arc sync -m "sorted the demo data"
```

Check the ARC in the DataHUB

Navigate to <https://git.nfdi4plants.org/<username>/arc-demo> to visit your ARC in the DataHUB

Your ARC is ready



Contributors

Slides presented here include contributions by

- name: Dominik Brilhaus
github: <https://github.com/brilator>
orcid: <https://orcid.org/0000-0001-9021-3197>

ARCitect hands-on

ARCitect installation

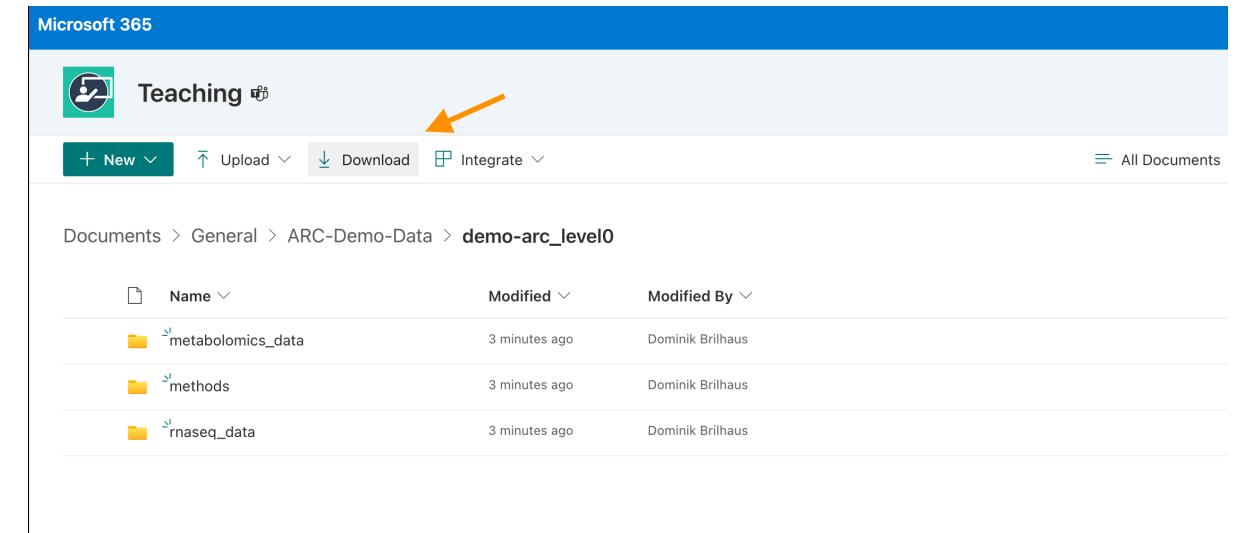
Please install version **v0.0.10** of the ARCitect:

<https://github.com/nfdi4plants/ARCitect/releases/tag/v0.0.10>

🔥 (released September 20th, 2023) 🔥

Download the demo data

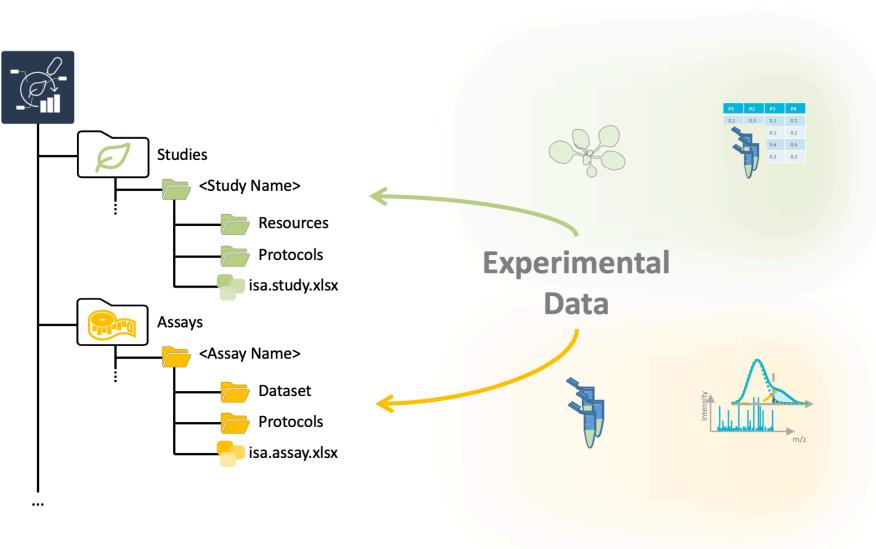
https://nfdi4plant.sharepoint.com/:f/s/Teaching/Eik7koJiMREgZ24kt07sIYBGxHmmZIS_Kzf7psk-5w-xg?e=u0sADd



The screenshot shows a Microsoft 365 SharePoint interface. At the top, there's a blue header bar with the text "Microsoft 365". Below it is a navigation bar with icons for "New", "Upload", "Download" (which is highlighted with an orange arrow), and "Integrate". To the right of the navigation bar is a link to "All Documents". The main content area shows a file structure under "Documents > General > ARC-Demo-Data > demo-arc_level0". There are three items listed: "metabolomics_data", "methods", and "rnaseq_data", all modified 3 minutes ago by Dominik Brilhaus.

| Name | Modified | Modified By |
|-------------------|---------------|------------------|
| metabolomics_data | 3 minutes ago | Dominik Brilhaus |
| methods | 3 minutes ago | Dominik Brilhaus |
| rnaseq_data | 3 minutes ago | Dominik Brilhaus |

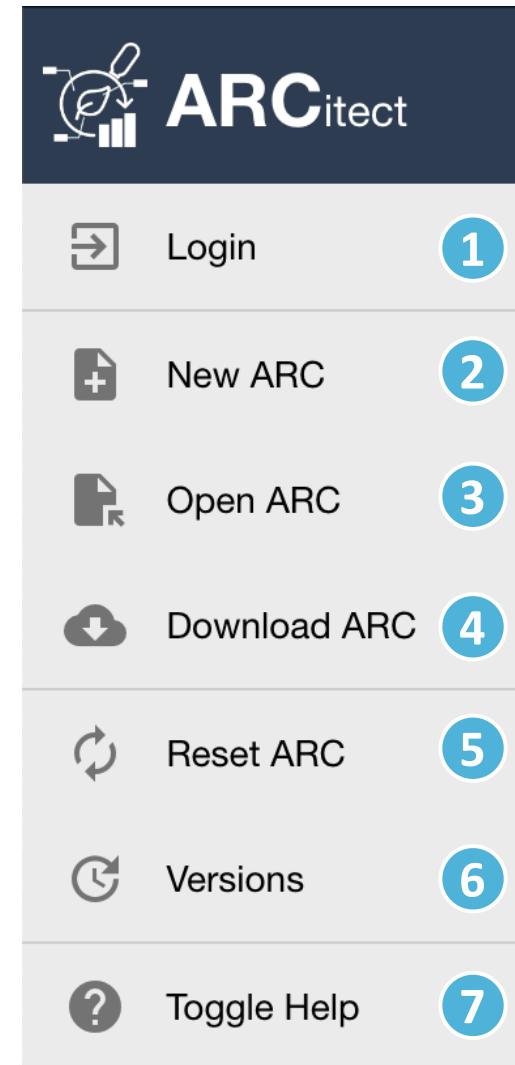
Sort Demo data in an ARC



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

Open ARCitect

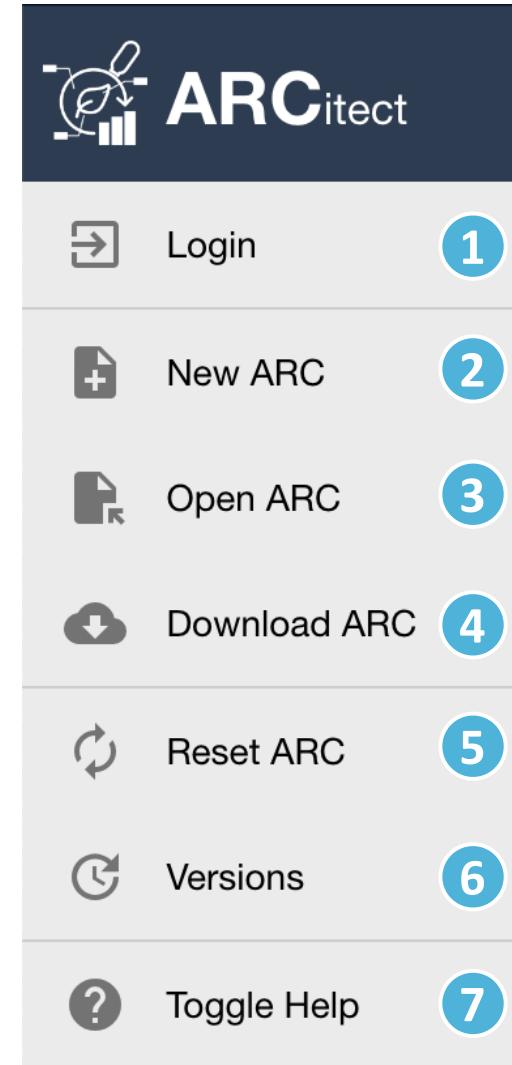
1. Login to DataHUB (1)



Initiate the ARC folder structure

1. Create a **New ARC** (2)
2. Select a location and name it

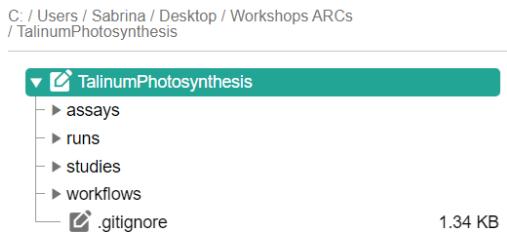
TalinumPhotosynthesis



Your ARC's name

- 💡 By default, your ARC's name will be used
 - for the ARC folder on your machine
 - to create your ARC in the DataHUB at
<https://git.nfdi4plants.org/<YourUserName>/<YourARC>>
(see next steps)
 - as the identifier for your investigation
 - 💡 Make sure that no ARC exists at
<https://git.nfdi4plants.org/<YourUserName>/<YourARC>> .
Otherwise you will sync to that ARC.
 - 💡 Don't use spaces in ARC's name
-  [TalinumPhotosynthesis](#)
 - ► assays
 - ► runs
 - ► studies
 - ► workflows

Add a description to your investigation



Identifier
TalinumPhotosynthesis

Title
Talinum Photosynthesis

Description
This is a very interesting investigation about life and photosynthesis

Add a contributor

Contacts

Your First Name Your Last Name

Your ORCID

6/10 ▾

First Name

Last Name

Mid Initials

ORCID

Search

Affiliation

Address

Email

Phone

Fax

Roles

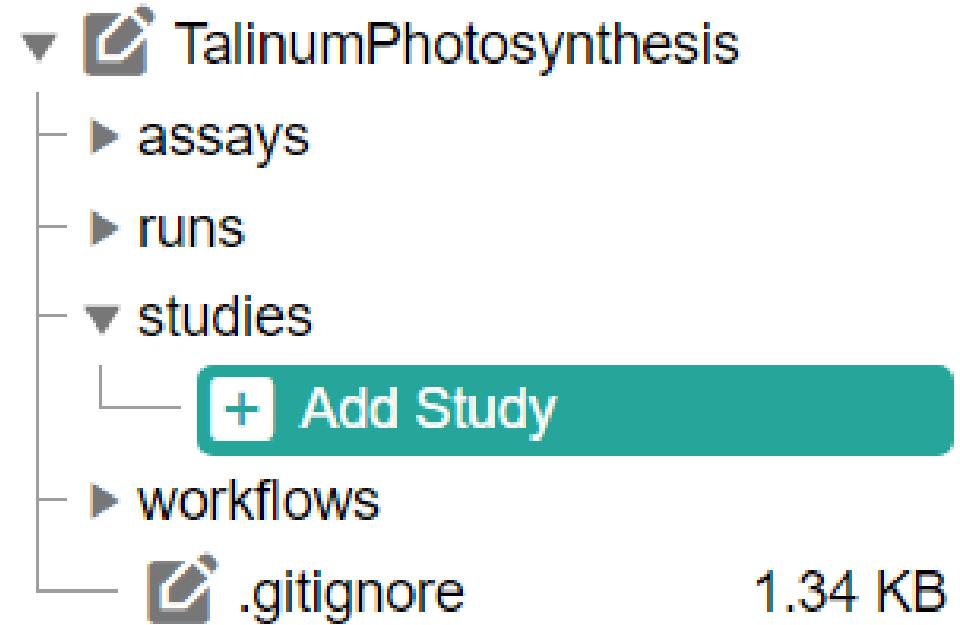
1. ✓

X+Delete

Add a study

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

Use **talinum_drought** as an identifier



Study panel

In the study panel you can add

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

Identifier

Description

Contacts

Publications

Submission Date

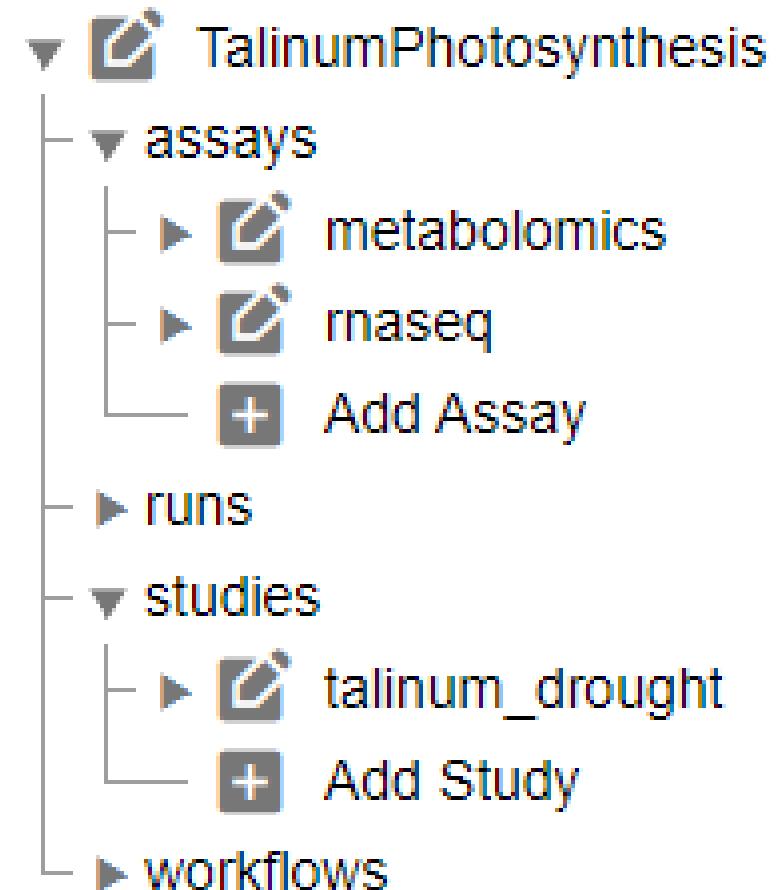
Public Release Date

Study Design Descriptors

Add an assay

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

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



Link your assay to a study

You can either

- link your new assay to an existing study in your ARC or
- create a new one

Link your assays to your
talinum_drought study

Add Assay

Add Assay
rnaseq

Study Identifiers
talinum_drought 

talinum_drought

 ADD ASSAY CANCEL

Add information about your assay

In the assay panel you can

1. link or unlink the assay to studies, and
2. define the assay's
 - measurement type
 - technology type, and
 - technology platform.
3. add data process information

| | | | |
|----------------------------|----------------------------------|----------------------|--|
| Identifier | rnaseq | | |
| Measurement Type | | | |
| Term Name | TSR | TAN | |
| <input type="text"/> | <input type="text"/> | <input type="text"/> | |
| Technology Type | | | |
| Term Name | TSR | TAN | |
| <input type="text"/> | <input type="text"/> | <input type="text"/> | |
| Technology Platform | | | |
| Term Name | TSR | TAN | |
| <input type="text"/> | <input type="text"/> | <input type="text"/> | |
| Performers | <input type="button" value="+"/> | | |
| Comments | <input type="button" value="+"/> | | |

Add protocols

You can either

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

Create or Import Protocol

Protocol Name



NEW PROTOCOL



IMPORT PROTOCOL

CANCEL

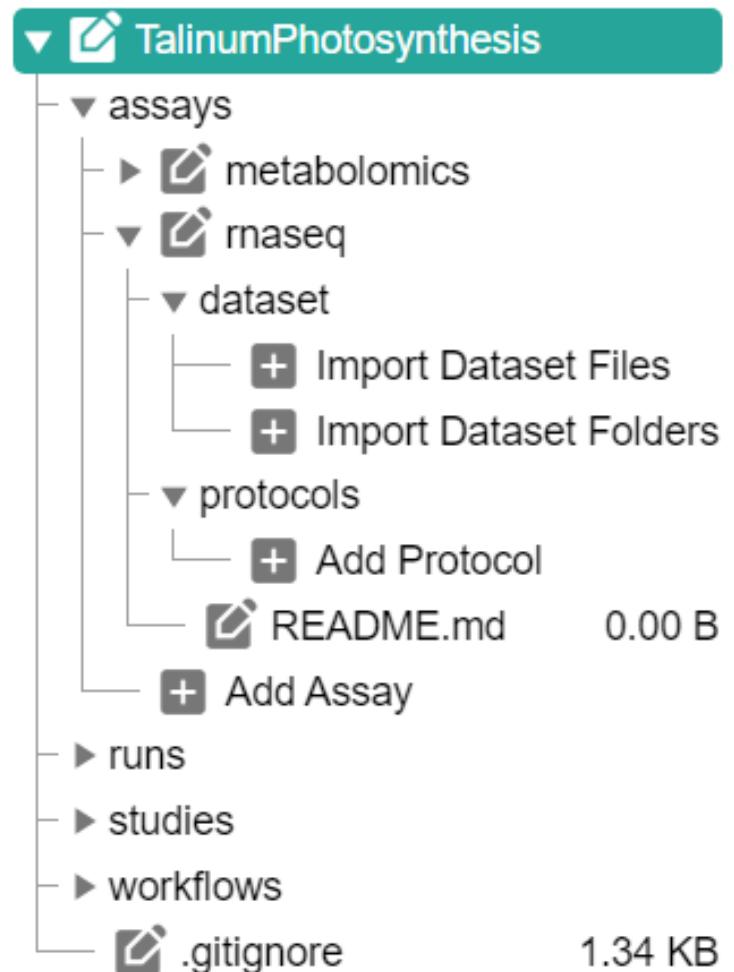
Add protocols and datasets

In the file tree you can

- **add a dataset** and
- **protocols** associated to that dataset.

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

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



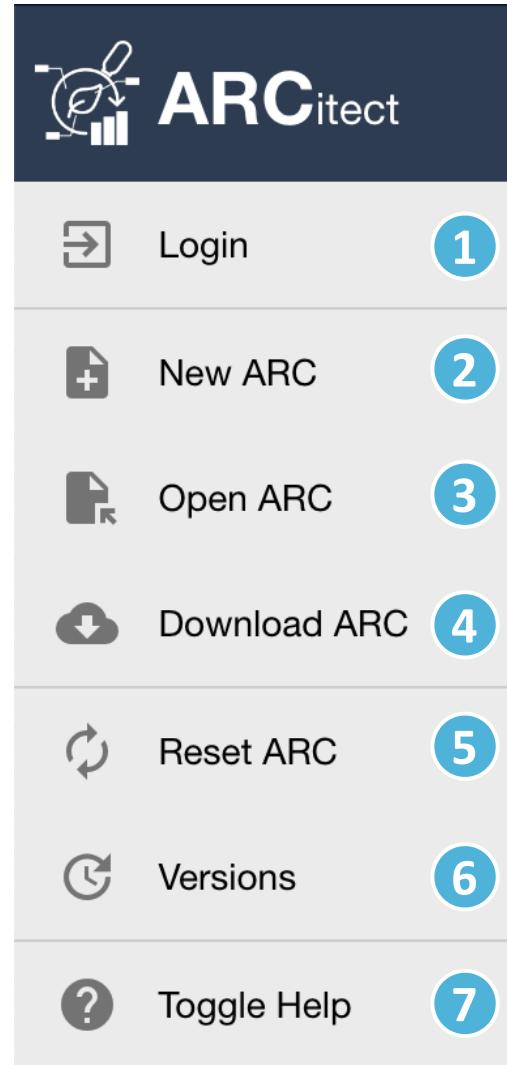
Sort Demo Data to your ARC

- 💡 protocols can directly be imported via ARCitect
- 💡 to add multiple datasets folders, they have to be added manually via file browser

Login to the DataHUB

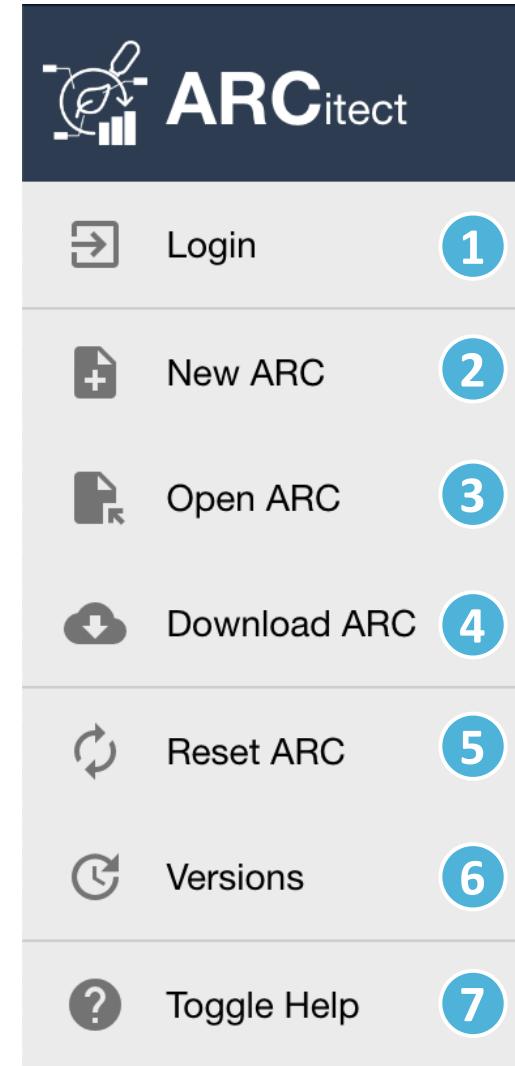
Click **Login** (1) in the sidebar to login to the DataHUB.

 This automatically opens your browser at the DataHUB (<https://git.nfdi4plants.org>) and asks you to login, if you are not already logged in.



Upload your local ARC to the DataHUB

From the sidebar, navigate to **Versions** (6)



Versions

The versions panel allows you to

- store the local changes to your ARC in form of "commits",
- sync the changes to the DataHUB, and
- check the history of your ARC

The screenshot shows the 'Update' section of the DataHUB interface. It includes fields for 'Full Name' (Demo User), 'eMail' (demo@nfdi4plants.org), and 'Remote' (https://git.nfdi4plants.org/demouser/Demo-ARC.git). A large text area for 'Commit Message' is present with a placeholder 'A short description of the made changes'. Below it is a 'Changes' list with items like '.arc/', 'assays/', 'isa.investigation.xlsx', 'runs/', 'studies/', and 'workflows/'. At the bottom are buttons for 'REFRESH', 'COMMIT', 'UPLOAD', and 'DOWNLOAD'.

Update
Commit changes and upload ARC

Full Name
Demo User

eMail
demo@nfdi4plants.org

Remote
https://git.nfdi4plants.org/demouser/Demo-ARC.git

Commit Message

A short description of the made changes

Changes

- + .arc/
- + assays/
- + isa.investigation.xlsx
- + runs/
- + studies/
- + workflows/

REFRESH COMMIT UPLOAD DOWNLOAD

History
Inspect ARC history

Connection to the DataHUB

If you are logged in, the versions panel shows

- your DataHUB's *Full Name* and *eMail*
- the URL of the current ARC in the DataHUB

<https://git.nfdi4plants.org/<YourUserName>/<YourARC>>

The screenshot shows the 'Update' section of the DataHUB interface. It includes fields for 'Full Name' (Demo User), 'eMail' (demo@nfdi4plants.org), and 'Remote' (https://git.nfdi4plants.org/demouser/Demo-ARC.git). A large 'Commit Message' field is present with a placeholder 'A short description of the made changes'. Below it is a 'Changes' list with items like '.arc/', 'assays/', 'isa.investigation.xlsx', 'runs/', 'studies/', and 'workflows/'. At the bottom are buttons for 'REFRESH', 'COMMIT', 'UPLOAD', and 'DOWNLOAD'.

Update
Commit changes and upload ARC

Full Name
Demo User

eMail
demo@nfdi4plants.org

Remote
https://git.nfdi4plants.org/demouser/Demo-ARC.git

Commit Message

A short description of the made changes

Changes

- + .arc/
- + assays/
- + isa.investigation.xlsx
- + runs/
- + studies/
- + workflows/

REFRESH COMMIT UPLOAD DOWNLOAD

History
Inspect ARC history

Check if your ARC is successfully uploaded

1. [sign in](#) to the DataHUB
2. Check your projects

Your ARC is ready

 Initiated an ARC

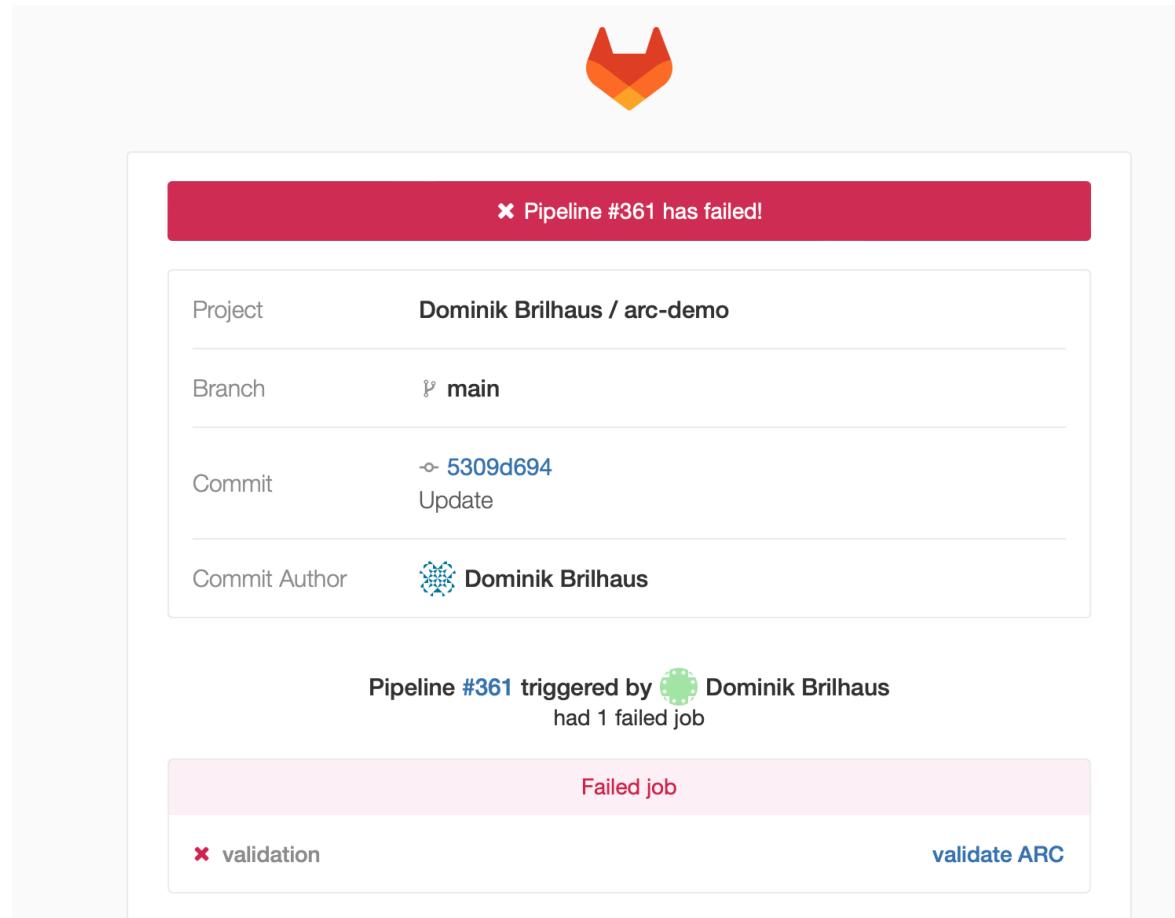
 Structured and ...

 ... annotated experimental data

 Shared with collaborators



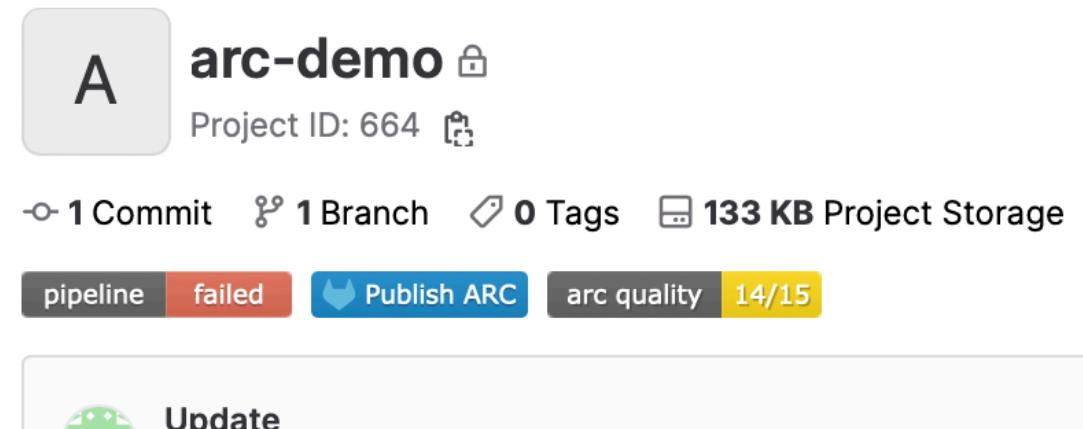
Received two emails from "GitLab" about a failed pipeline?



Pipeline Failed

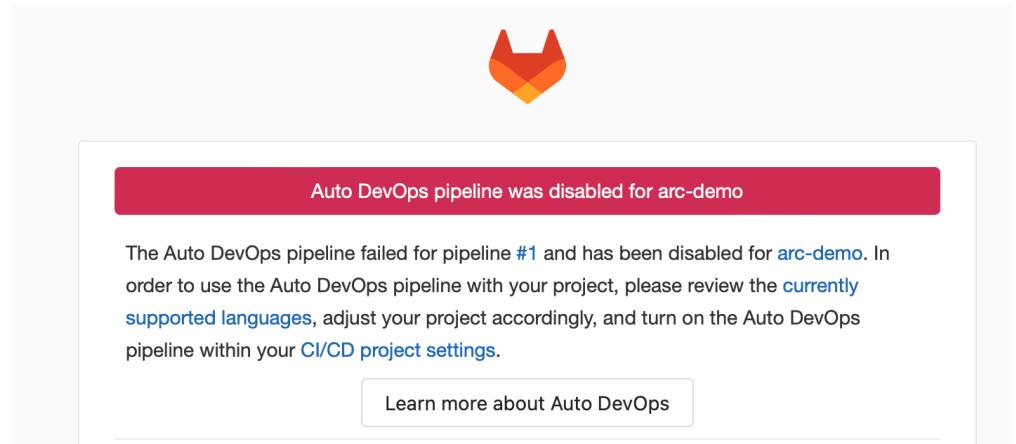
- a "continuous quality control" (CQC) pipeline validates your ARC
- This fails if one of the following metadata items is missing:

```
Investigation Identifier
Investigation Title
Investigation Description
Investigation Person Last Name
Investigation Person First Name
Investigation Person Email
Investigation Person Affiliation
```



Pipeline Failed

If the pipeline has failed once, it is disabled by default



The screenshot shows a GitHub Auto DevOps pipeline status page. At the top, there's a large orange and yellow icon resembling a cat's head. Below it, a red banner displays the message "Auto DevOps pipeline was disabled for arc-demo". The main content area contains text explaining the failure: "The Auto DevOps pipeline failed for pipeline #1 and has been disabled for [arc-demo](#). In order to use the Auto DevOps pipeline with your project, please review the [currently supported languages](#), adjust your project accordingly, and turn on the Auto DevOps pipeline within your [CI/CD project settings](#)." A button at the bottom right says "Learn more about Auto DevOps".

Reactivate the CQC pipeline

To reactivate it and let the DataHUB validate your ARC again:

1. navigate to CI/CD setting `<arc-url>/-/settings/ci_cd`
2. expand "Auto DevOps"
3. check box "Default to Auto DevOps pipeline"
4. Save changes

The screenshot shows the GitLab CI/CD settings interface. On the left, a sidebar lists various project settings: Security & Compliance, Deployments, Packages and registries, Infrastructure, Monitor, Analytics, Wiki, Snippets, Settings (selected), General, Integrations, Webhooks, Access Tokens, Repository, Merge requests, CI/CD (selected), Packages and registries, Monitor, and Usage Quotas. The main content area is titled "Auto DevOps" with a "Collapse" button. It includes a sub-section "How do I get started?" with a checked checkbox for "Default to Auto DevOps pipeline" (instance enabled). Below this, there's a note about adding a Kubernetes cluster integration or creating an AUTO_DEVOPS_PLATFORM_TARGET CI variable. A "Deployment strategy" section contains three radio buttons: "Continuous deployment to production" (selected), "Continuous deployment to production using timed incremental rollout", and "Automatic deployment to staging, manual deployment to production". At the bottom, a "Save changes" button is visible.

Contributors

Slides presented here include contributions by

- name: Dominik Brilhaus
github: <https://github.com/brilator>
orcid: <https://orcid.org/0000-0001-9021-3197>
- name: Cristina Martins Rodrigues
github: <https://github.com/CMR248>
orcid: <https://orcid.org/0000-0002-4849-1537>
- name: Sabrina Zander
github: <https://github.com/SabrinaZander>
orcid: <https://orcid.org/0009-0000-4569-6126>

Data Storage and Versioning

Data stores

Local hard disks



Institute server



University server



Cloud services



Your data



labfolder

Electronic lab
notebooks



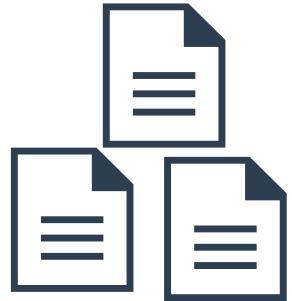
Git repositories

Backup vs. Archive

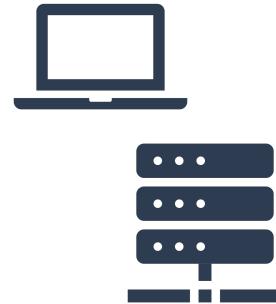
| | Backup | Archive |
|--------------|---------------------|-------------------------------|
| Storage type | Short-, mid-term | Long-term |
| Purpose | Disaster recovery | Long-term storage, compliance |
| Reason | Duplication | Migration |
| Usage | Work in progress | Cold, Unused data |
| Changes | Short-term updates | No updates |
| Trend | Cyclic, Replacement | Growing |
| Latency | Short/Costly | High/Cheaper |

3-2-1 backup rule

*3 copies
of data*



*2 storage
media*



*1 copy
off-site*



Version control and track changes

It's good practice to document:

- What was changed?
- Who is responsible?
- When did it happen?
- Why the changes?

Types of Version Control

- by file name (_v1, _v2)
- cloud services
 - dropbox, icloud, gdrive
- distributed version control system
 - e.g. Git

Which files need to be "versioned"?



- paper manuscript (.docx)
- single-cell RNASeq reads (.fastq.gz)
- spread sheet with photometer measurements (.xlsx)
- calendar invitation (.ical)
- photo of SDS-PAGE (.jpeg)
- excel workbook with calculations (.xlsx)
- presentation for a conference (.pdf)
- data analysis script (.py)

Concept of Git and git-based platforms

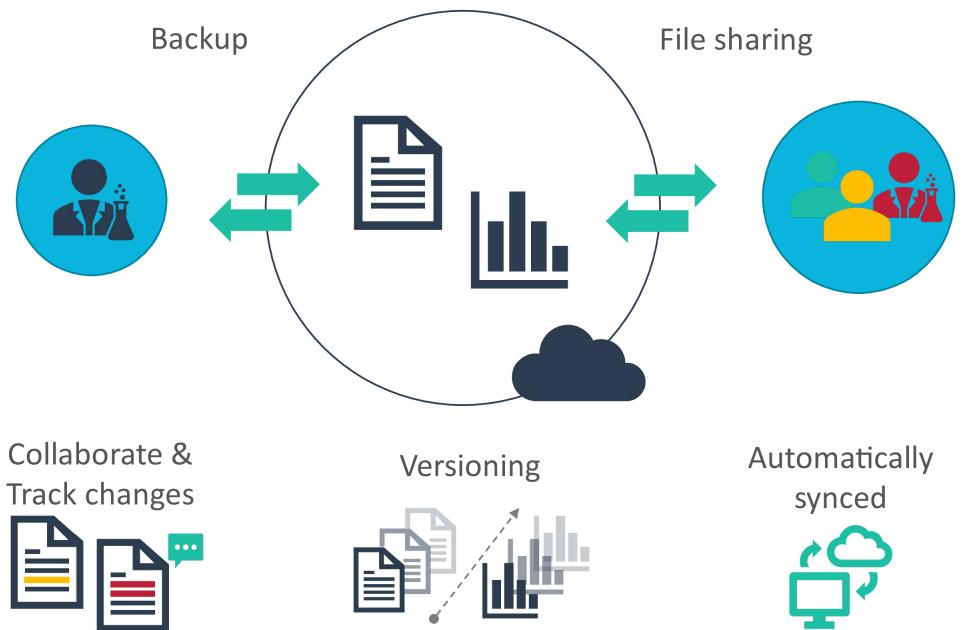
Cloud Services

- ✓ Documents
- ✓ Small data
- ✓ Presentations

X Code

X Data analytical projects

X Big (“raw”) data



Git and git platforms

- ~ Documents
- ✓ Small data
- ~ Presentations

- ✓✓ Code
- ✓✓ Data analytical projects
- ~ Big (“raw”) data

Why git? => Why code?

- Save time
- Avoid doing repetitive tasks “by hand”
- Reuse scripts, analyses, pipelines
- Reproduce results

A simple example: RNASeq project

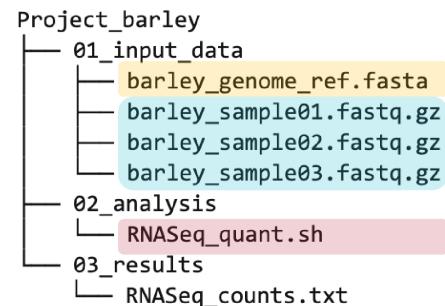
```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
└── 02_analysis
    └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

A simple example: RNASeq project

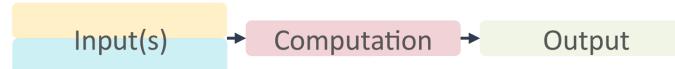
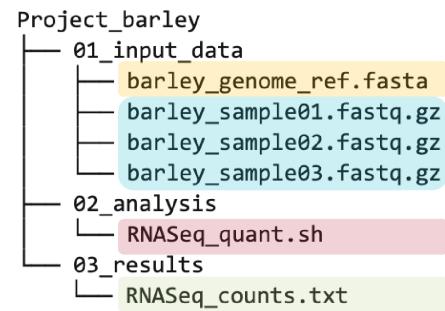
```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
└── 02_analysis
    └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

Input(s)

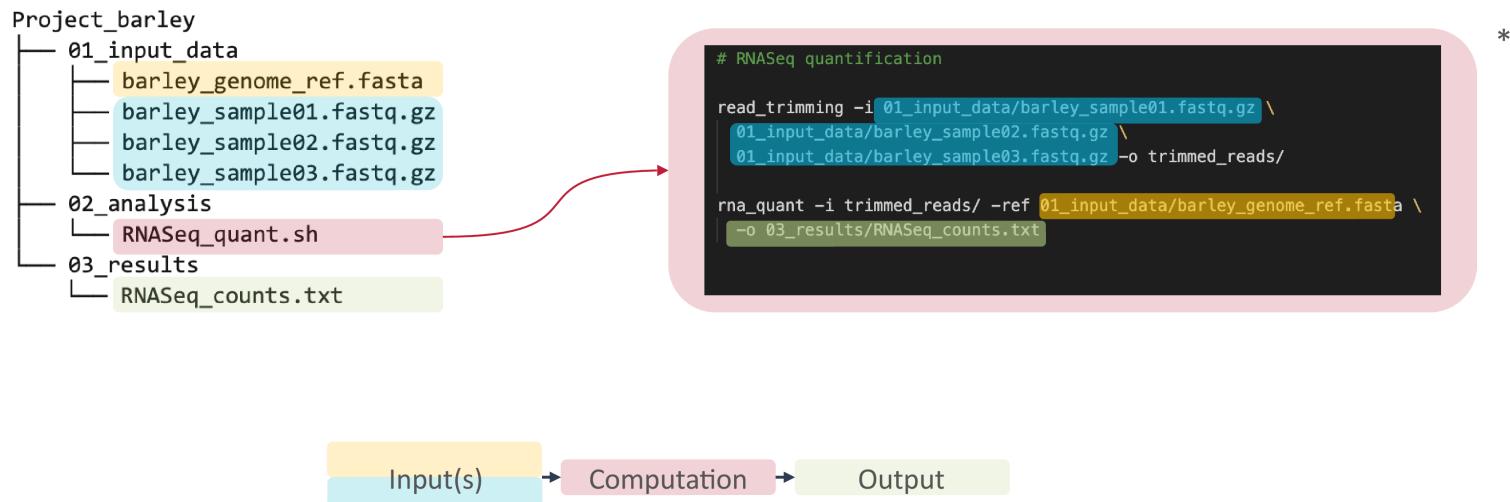
A simple example: RNASeq project



A simple example: RNASeq project



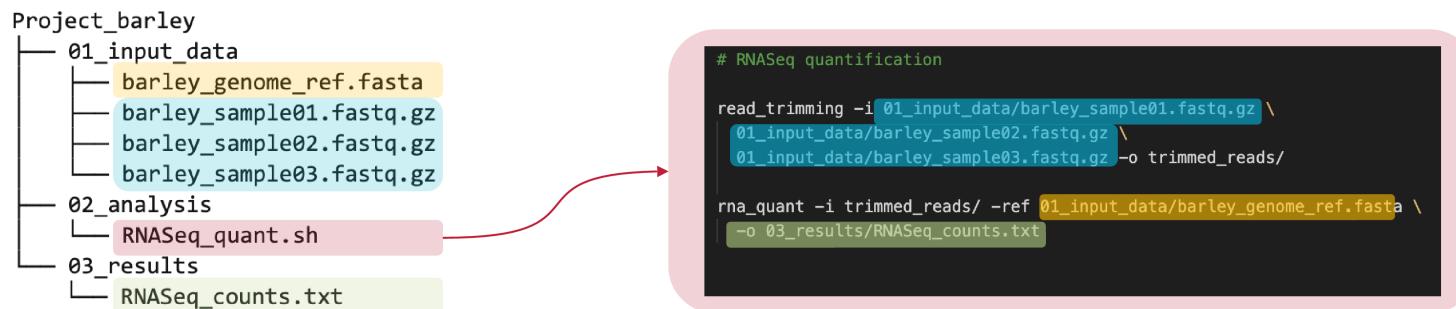
A simple example: RNASeq project



* Disclaimer: this is not a good example for reusable code

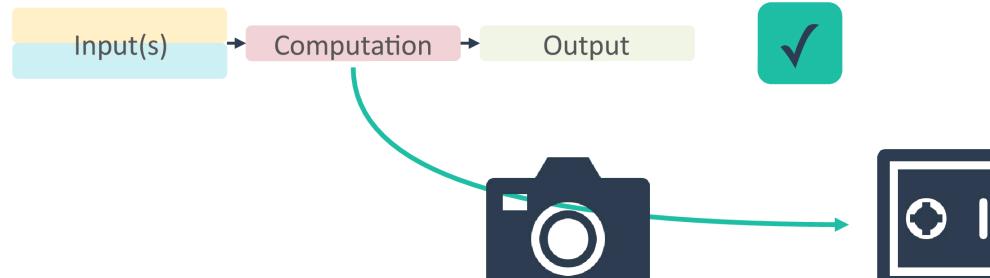
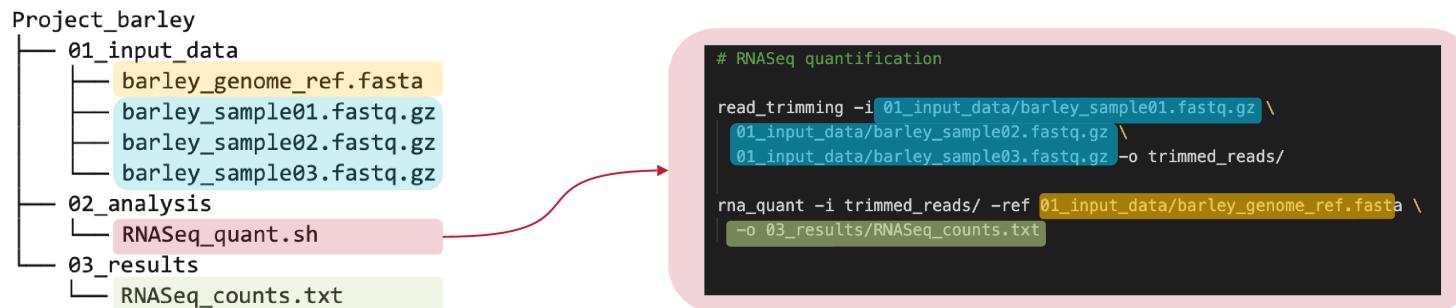
Take snapshots of your code work...

(... as long as it works)



Take snapshots of your code work...

(... as long as it works)



Scenario 1: More data

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
└── 02_analysis
    └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

```
# RNASeq quantification
read_trimming -i 01_input_data/barley_sample01.fastq.gz \
| 01_input_data/barley_sample02.fastq.gz \
| 01_input_data/barley_sample03.fastq.gz -o trimmed_reads/
rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
-o 03_results/RNASeq_counts.txt
```

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   ├── barley_sample03.fastq.gz
│   ├── barley_sample04.fastq.gz
│   ├── barley_sample05.fastq.gz
│   └── barley_sample06.fastq.gz
```

Scenario 1: More data

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
└── 02_analysis
    └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   ├── barley_sample03.fastq.gz
│   ├── barley_sample04.fastq.gz
│   ├── barley_sample05.fastq.gz
│   └── barley_sample06.fastq.gz
```

```
# RNASeq quantification

read_trimming -i 01_input_data/barley_sample01.fastq.gz \
01_input_data/barley_sample02.fastq.gz \
01_input_data/barley_sample03.fastq.gz -o trimmed_reads/

rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
-o 03_results/RNASeq_counts.txt
```

```
# RNASeq quantification

read_trimming -i 01_input_data/barley_sample01.fastq.gz \
01_input_data/barley_sample02.fastq.gz \
01_input_data/barley_sample03.fastq.gz \
01_input_data/barley_sample04.fastq.gz \
01_input_data/barley_sample05.fastq.gz \
01_input_data/barley_sample06.fastq.gz -o trimmed_reads/

rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
-o 03_results/RNASeq_counts.txt
```

Scenario 1: More data

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
└── 02_analysis
    └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   ├── barley_sample03.fastq.gz
│   ├── barley_sample04.fastq.gz
│   ├── barley_sample05.fastq.gz
│   └── barley_sample06.fastq.gz
└── 02_analysis
    ├── RNASeq_quant.sh
    ├── RNASeq_quant_first_samples.sh
    ├── RNASeq_quant_including_all_samples.sh
    ├── RNASeq_quant_including_all_samples_updated.sh
    └── RNASeq_quant_including_all_samples_updated_v2.sh
└── 03_results
    └── RNASeq_counts.txt
```

```
# RNASeq quantification
read_trimming -i 01_input_data/barley_sample01.fastq.gz \
01_input_data/barley_sample02.fastq.gz \
01_input_data/barley_sample03.fastq.gz -o trimmed_reads/
rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
-o 03_results/RNASeq_counts.txt
```

```
# RNASeq quantification
read_trimming -i 01_input_data/barley_sample01.fastq.gz \
01_input_data/barley_sample02.fastq.gz \
01_input_data/barley_sample03.fastq.gz \
01_input_data/barley_sample04.fastq.gz \
01_input_data/barley_sample05.fastq.gz \
01_input_data/barley_sample06.fastq.gz -o trimmed_reads/
rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
-o 03_results/RNASeq_counts.txt
```

Let git track changes and keep things clean

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   ├── barley_sample03.fastq.gz
│   ├── barley_sample04.fastq.gz
│   ├── barley_sample05.fastq.gz
│   └── barley_sample06.fastq.gz
└── 02_analysis
    └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

```
Project_barley > 02_analysis > $ RNASeq_quant.sh
1 # RNASeq quantification
2
3 read_trimming -i 01_input_data/barley_sample01.fastq.gz \
4 01_input_data/barley_sample02.fastq.gz \
5 01_input_data/barley_sample03.fastq.gz -o trimmed_reads/
6
7 rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
8 -o 03_results/RNASeq_counts.txt
9
10
11
```

“version 1”

```
1 # RNASeq quantification
2
3 read_trimming -i 01_input_data/barley_sample01.fastq.gz \
4 01_input_data/barley_sample02.fastq.gz \
5+ 01_input_data/barley_sample03.fastq.gz \
6+ 01_input_data/barley_sample04.fastq.gz \
7+ 01_input_data/barley_sample05.fastq.gz \
8+ 01_input_data/barley_sample06.fastq.gz -o trimmed_reads/
9
10 rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
11 -o 03_results/RNASeq_counts.txt
12
13
14
```

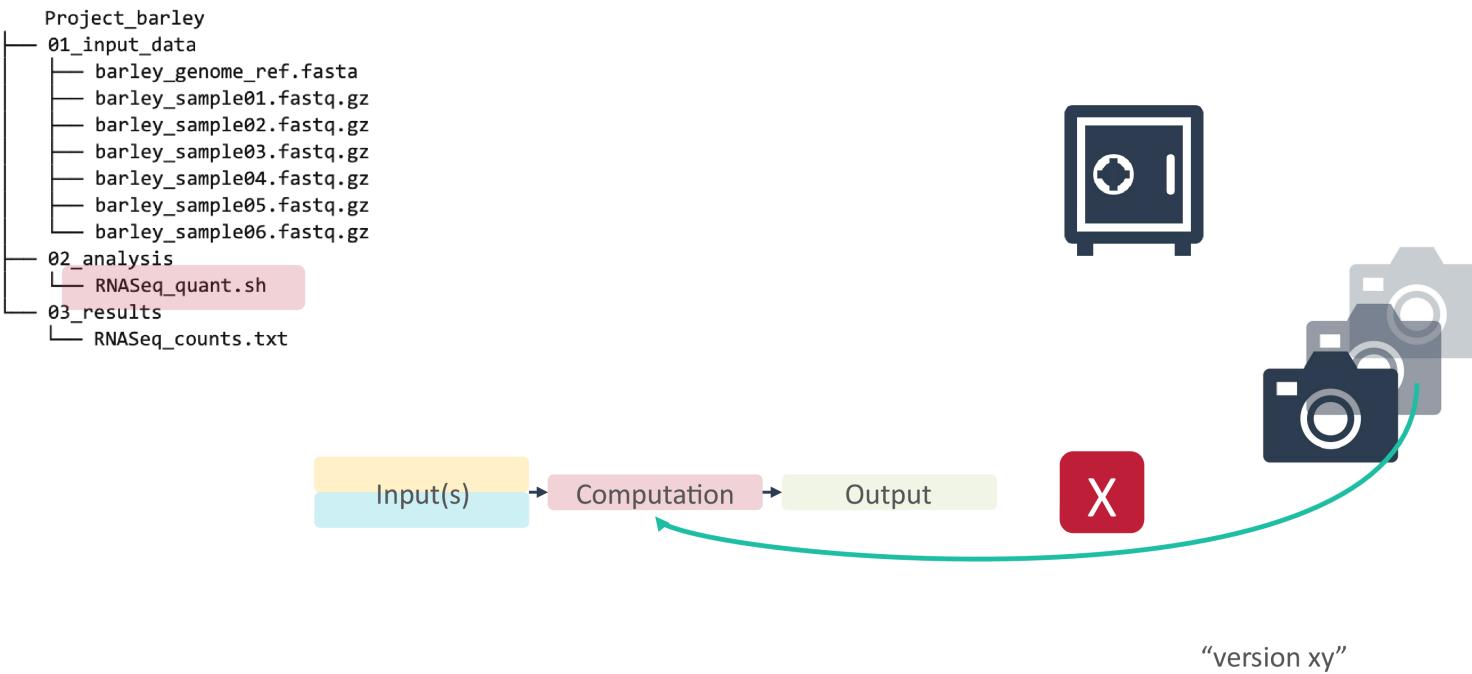
“version 2”

Scenario 2: Pipeline breaks

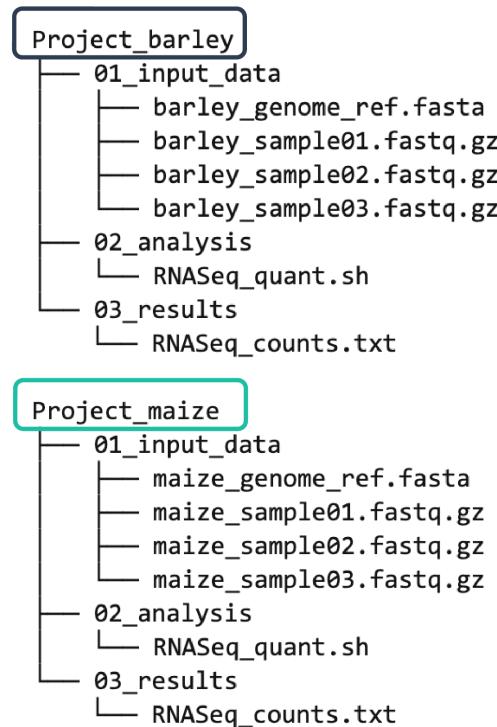
```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   ├── barley_sample03.fastq.gz
│   ├── barley_sample04.fastq.gz
│   ├── barley_sample05.fastq.gz
│   └── barley_sample06.fastq.gz
└── 02_analysis
    └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```



Revert to snapshot



Scenario 3: New project, same type of data and analysis



Scenario 3: New project, same type of data and analysis

```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
└── 02_analysis
    └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

```
Project_maize
├── 01_input_data
│   ├── maize_genome_ref.fasta
│   ├── maize_sample01.fastq.gz
│   ├── maize_sample02.fastq.gz
│   └── maize_sample03.fastq.gz
└── 02_analysis
    └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

```
# RNASeq quantification

read_trimming -i 01_input_data/barley_sample01.fastq.gz \
  01_input_data/barley_sample02.fastq.gz \
  01_input_data/barley_sample03.fastq.gz -o trimmed_reads/

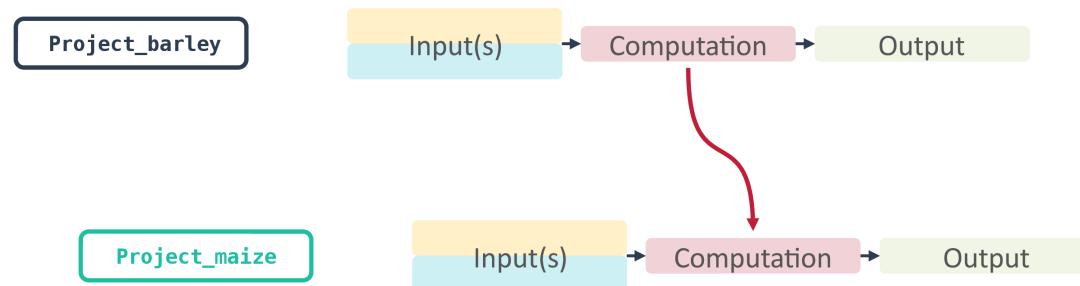
rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
-o 03_results/RNASeq_counts.txt
```

```
# RNASeq quantification

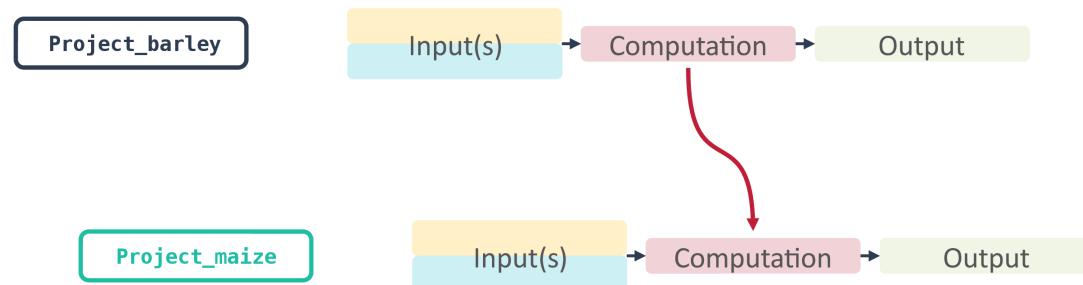
read_trimming -i 01_input_data/maize_sample01.fastq.gz \
  01_input_data/maize_sample02.fastq.gz 01_input_data/maize_sample03.fastq.gz \
-o trimmed_reads/

rna_quant -i trimmed_reads/ -ref 01_input_data/ maize_genome_ref.fasta \
-o 03_results/RNASeq_counts.txt
```

Re-use code



Re-use code



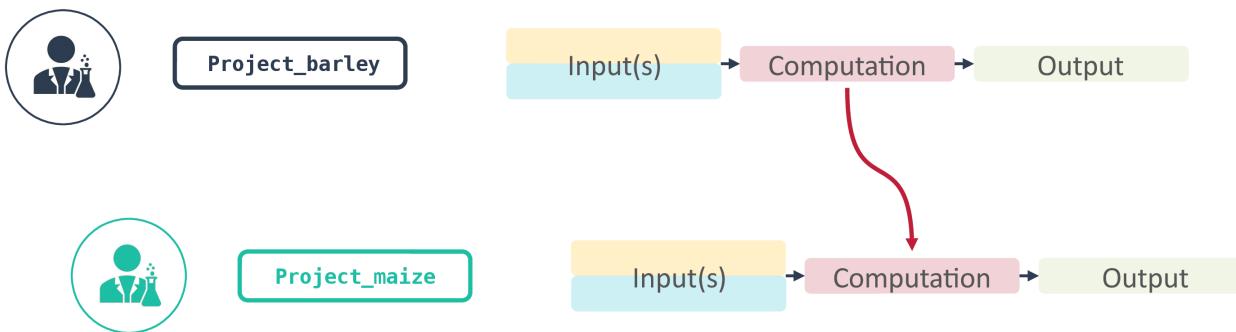
```
1 # RNASeq quantification
2
3 read_trimming -i 01_input_data/barley_sample01.fastq.gz \
4 01_input_data/barley_sample02.fastq.gz \
5 01_input_data/barley_sample03.fastq.gz \
6 -o trimmed_reads/
7
8 rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
9 -o 03_results/RNASeq_counts.txt
10
```

“version barley”

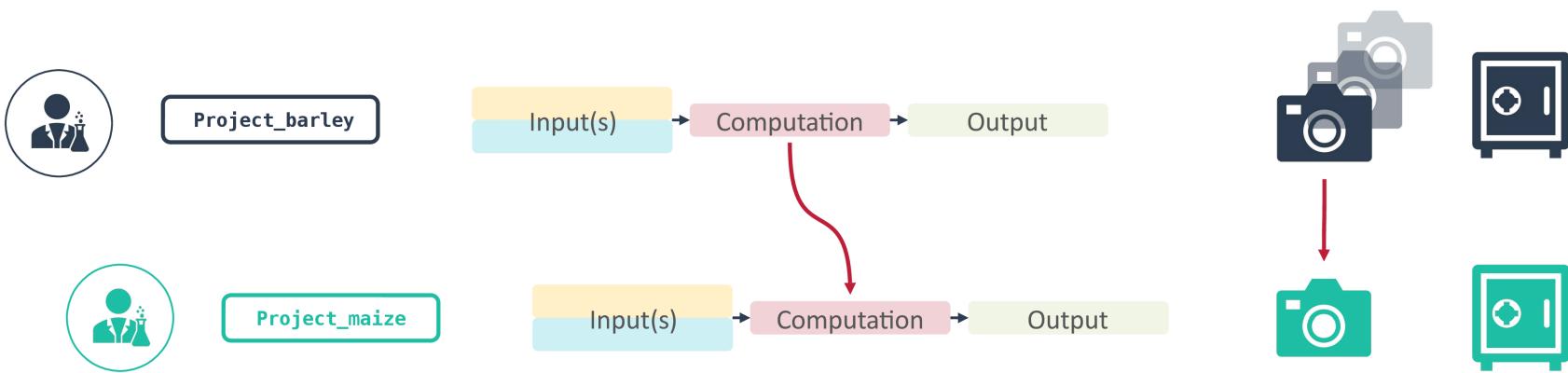
```
1 # RNASeq quantification
2
3+ read_trimming -i 01_input_data/maize_sample01.fastq.gz \
4+ 01_input_data/maize_sample02.fastq.gz 01_input_data/maize_sample03.fastq.gz
5 -o trimmed_reads/
6
7+ rna_quant -i trimmed_reads/ -ref 01_input_data/ maize_genome_ref.fasta -o 03_results/RNASeq_counts.txt
```

“version maize”

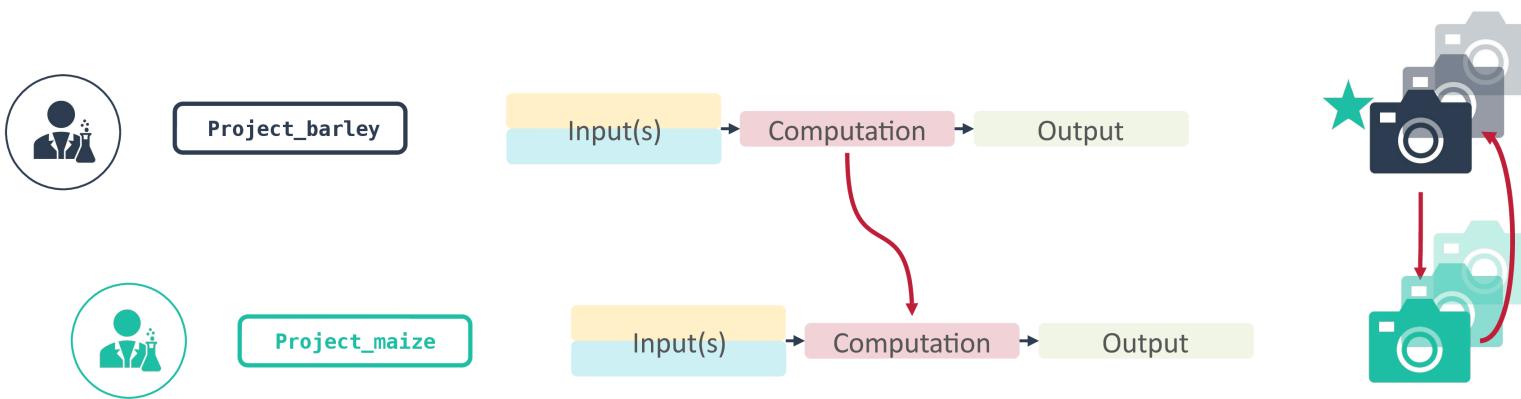
Re-use code – People have done this



Re-use code – People have done this



Re-use code – Link and contribute



Git: summary

- Version control system
- Git “repository” = a central data package (directory)
- Allows to track changes to any file in the repository
 - **What** was changed
 - **When** was it changed
 - **By whom** was it changed
 - **Why** was it changed?

GitHub and GitLab

- A well-documented cloud environment
- Active syncing
- Not automatically synced
- Non-automated version control
- You have the control what changes to track and what to sync
- Time machine to go back to older versions

GitHub and Gitlab team projects

Simplifies concurrent work & merging changes

- Online service to host our projects
- Share code with other developers
- Others can download our projects, work on and contribute to them
- They can upload their changes and merge them with the main project

Cloud vs. Git

| | Cloud services | Git / GitHub / GitLab |
|---------------|---|---|
| Track changes | | |
| Collaboration | | |
| Versioning | | |
| Syncing | | |
| Access | | |
| Data security | | |
| | Cloud services | Git / GitHub / GitLab |
| | | |
| | ✓ Documents ✓ Small data ✓ Presentations | ✓ Code ✓ Data analytical projects |
| | Automated | issue tracker, tracked contribution |
| | Automated | Well-documented (commit history) |
| | Oftentimes only within organization / institution | Active / controlled by user |
| | Private / commercial | Easily collaborate across institutions |
| | | GitLab: on-premise and custom solutions |

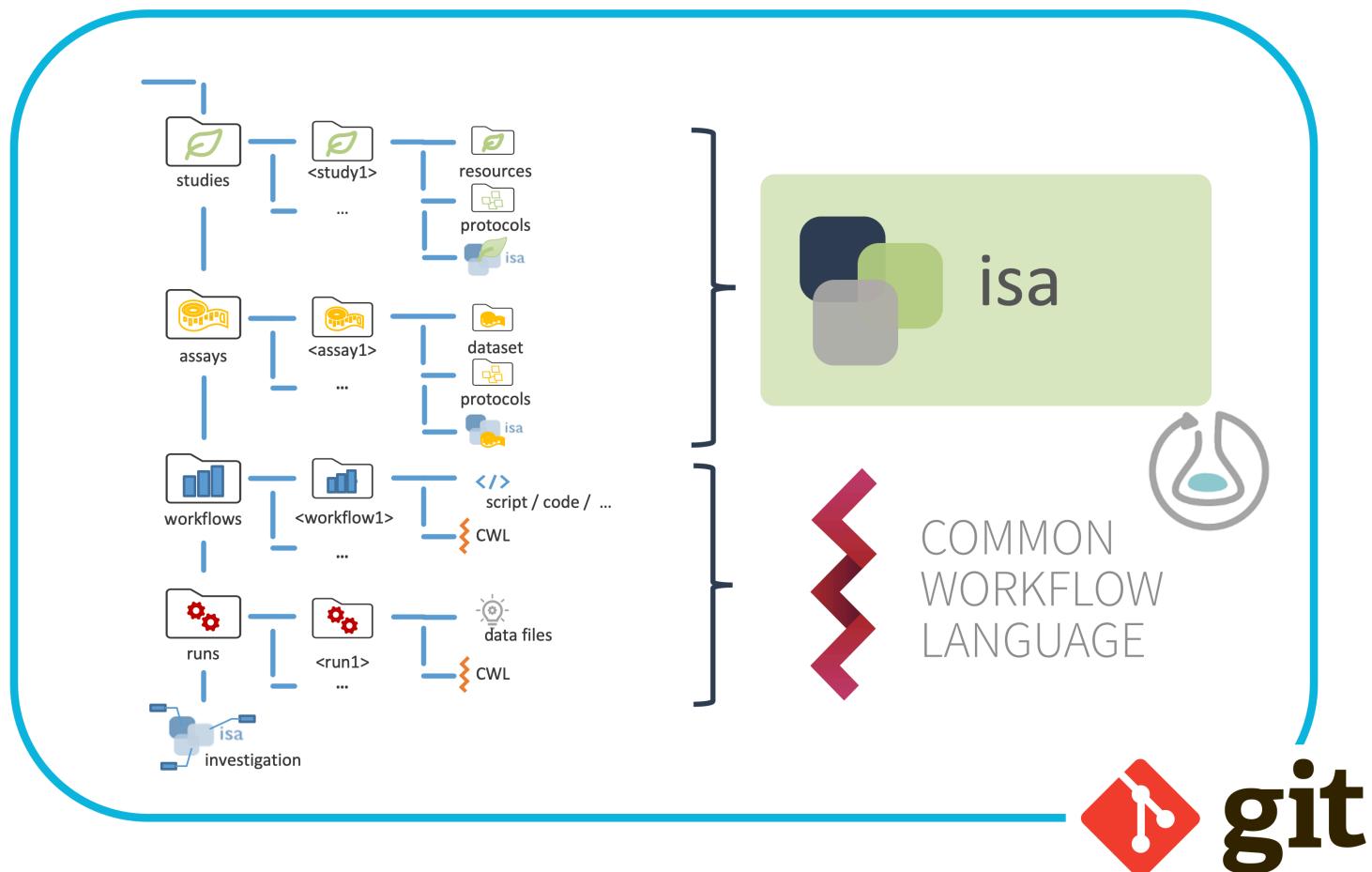
Contributors

Slides presented here include contributions by

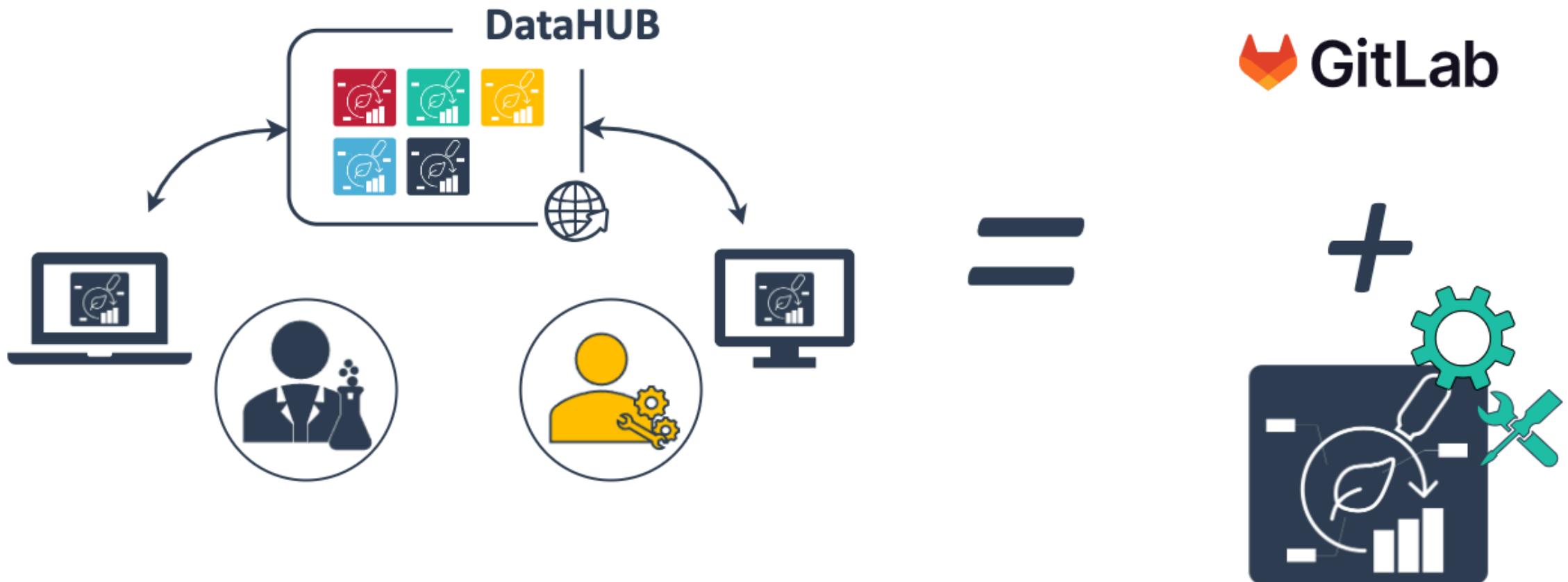
- name: Dominik Brilhaus
github: <https://github.com/brilator>
orcid: <https://orcid.org/0000-0001-9021-3197>
- name: Hajira Jabeen
github: <https://github.com/HajiraJabeen>
orcid: <https://orcid.org/0000-0003-1476-2121>

DataPLANT DataHUB

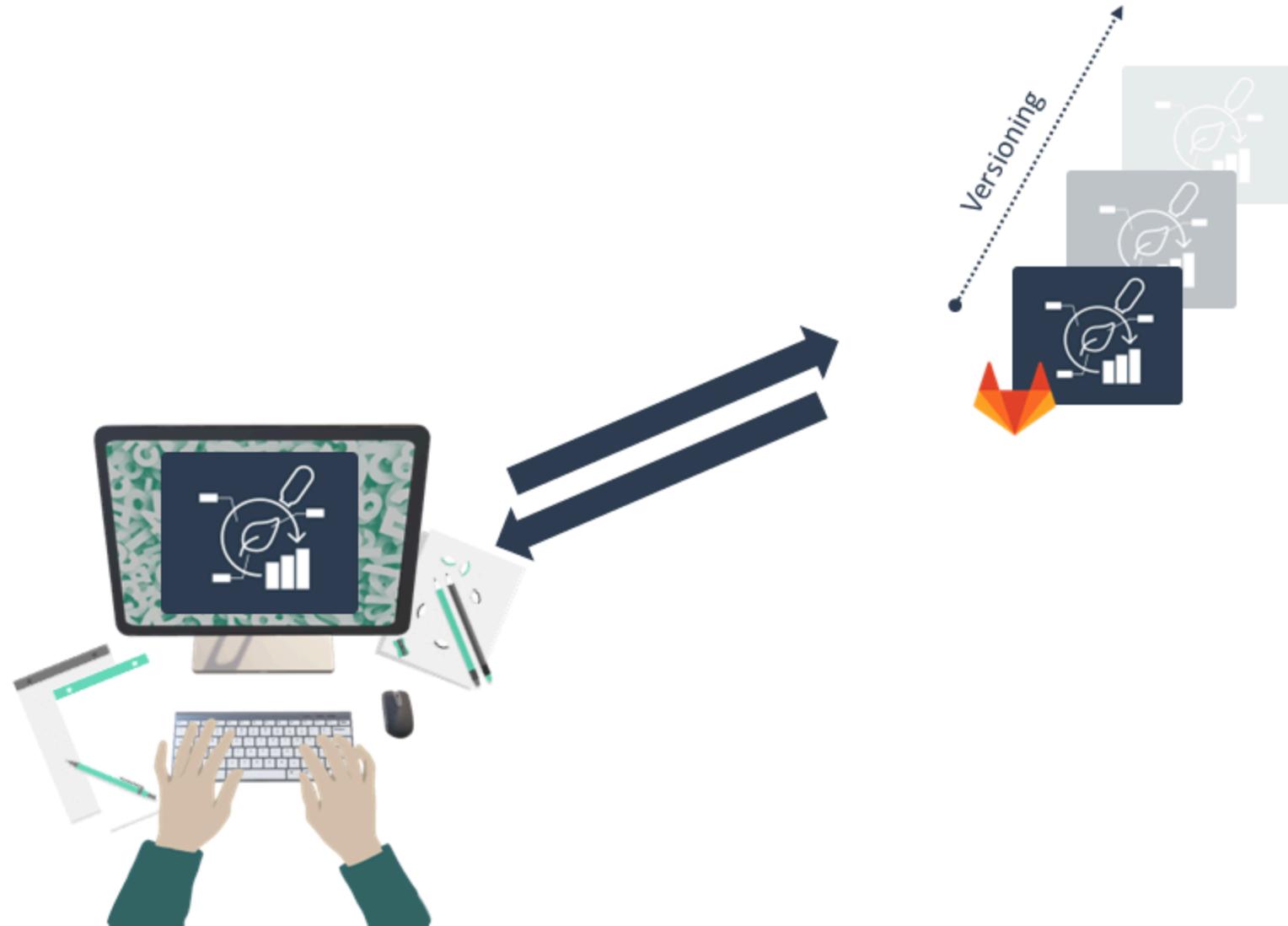
ARC builds on standards + Git



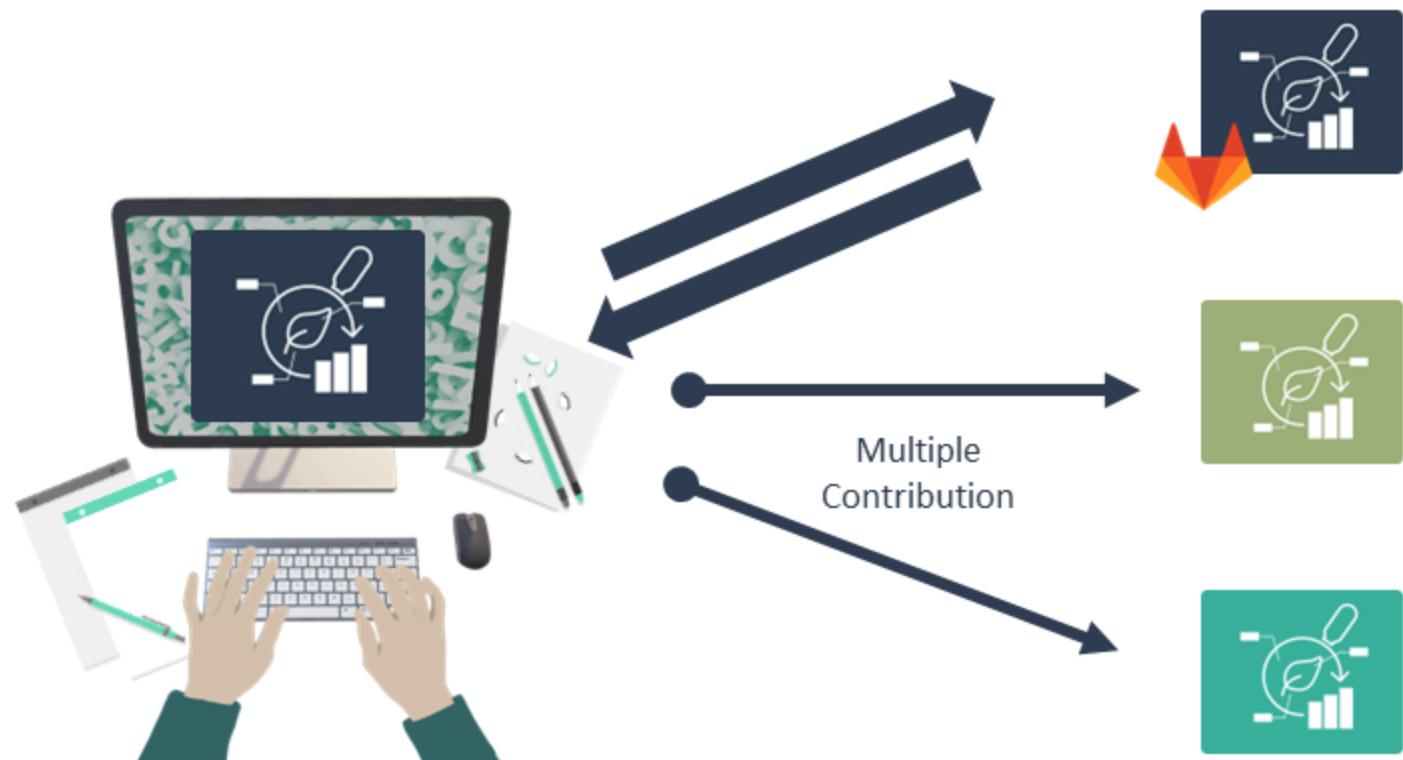
The DataPLANT DataHUB – a GitLab *Plus*



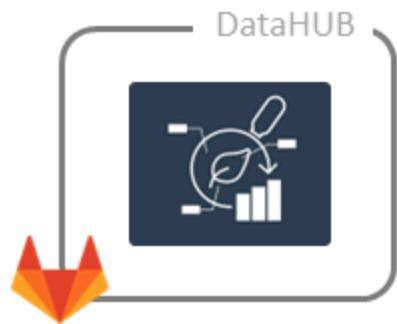






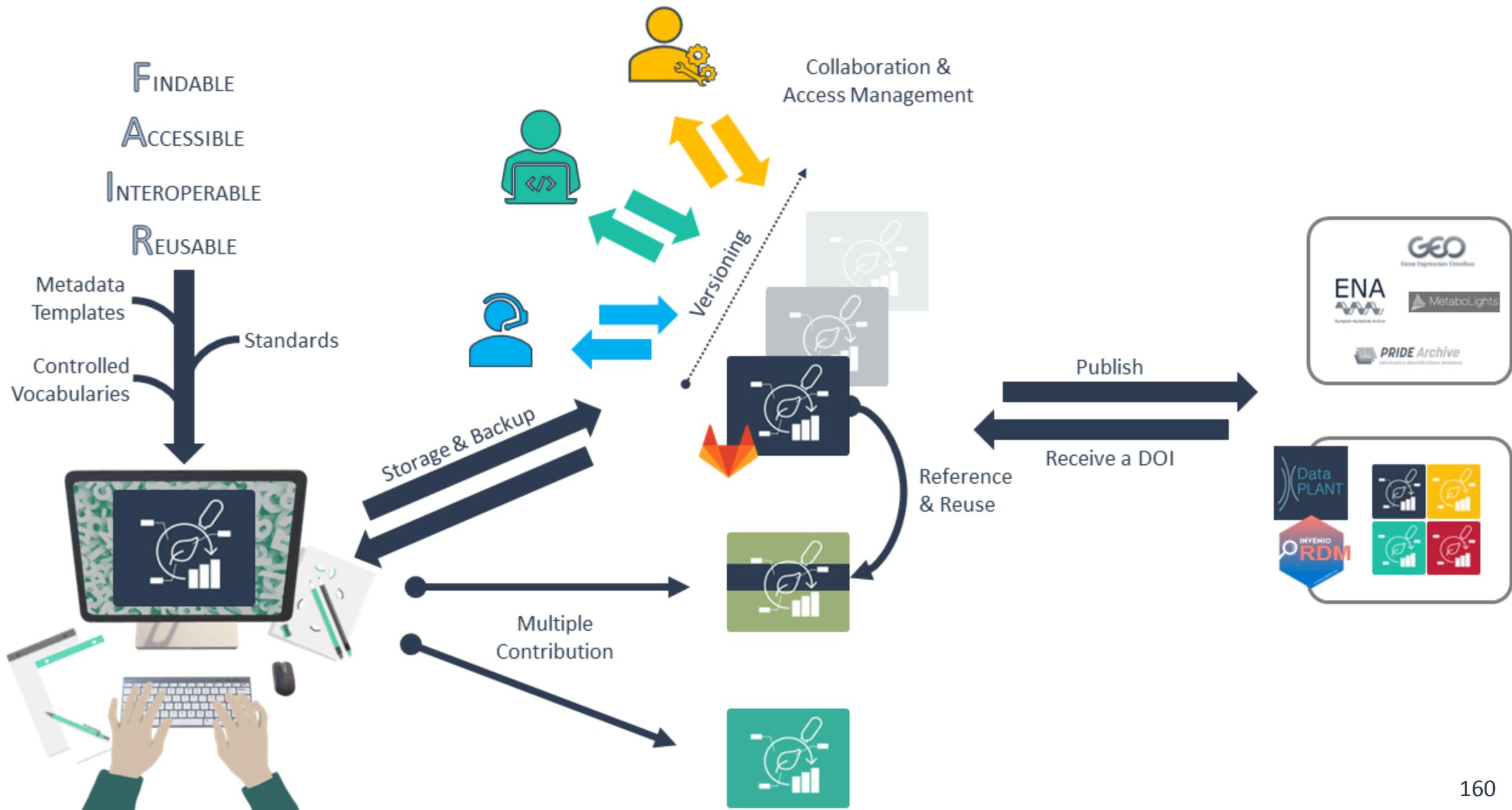




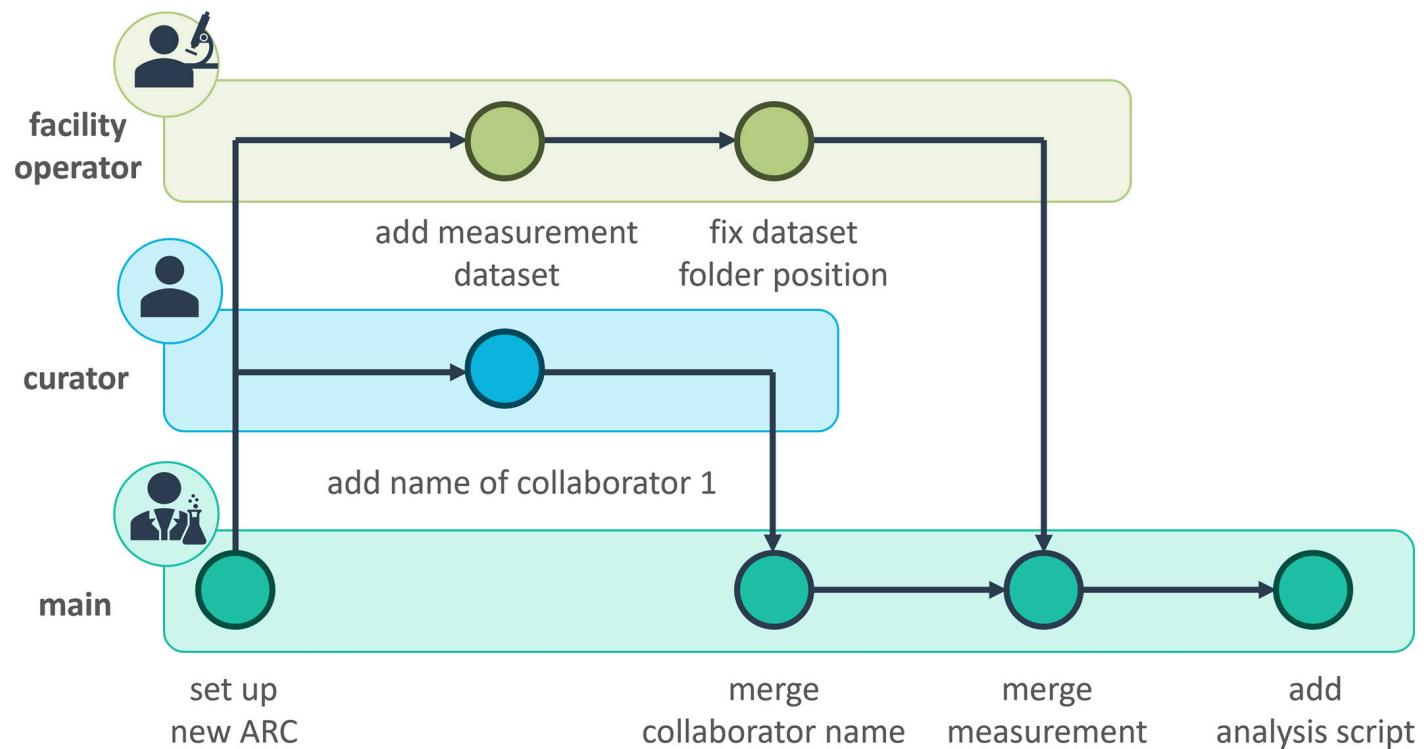


Publish
Receive a DOI

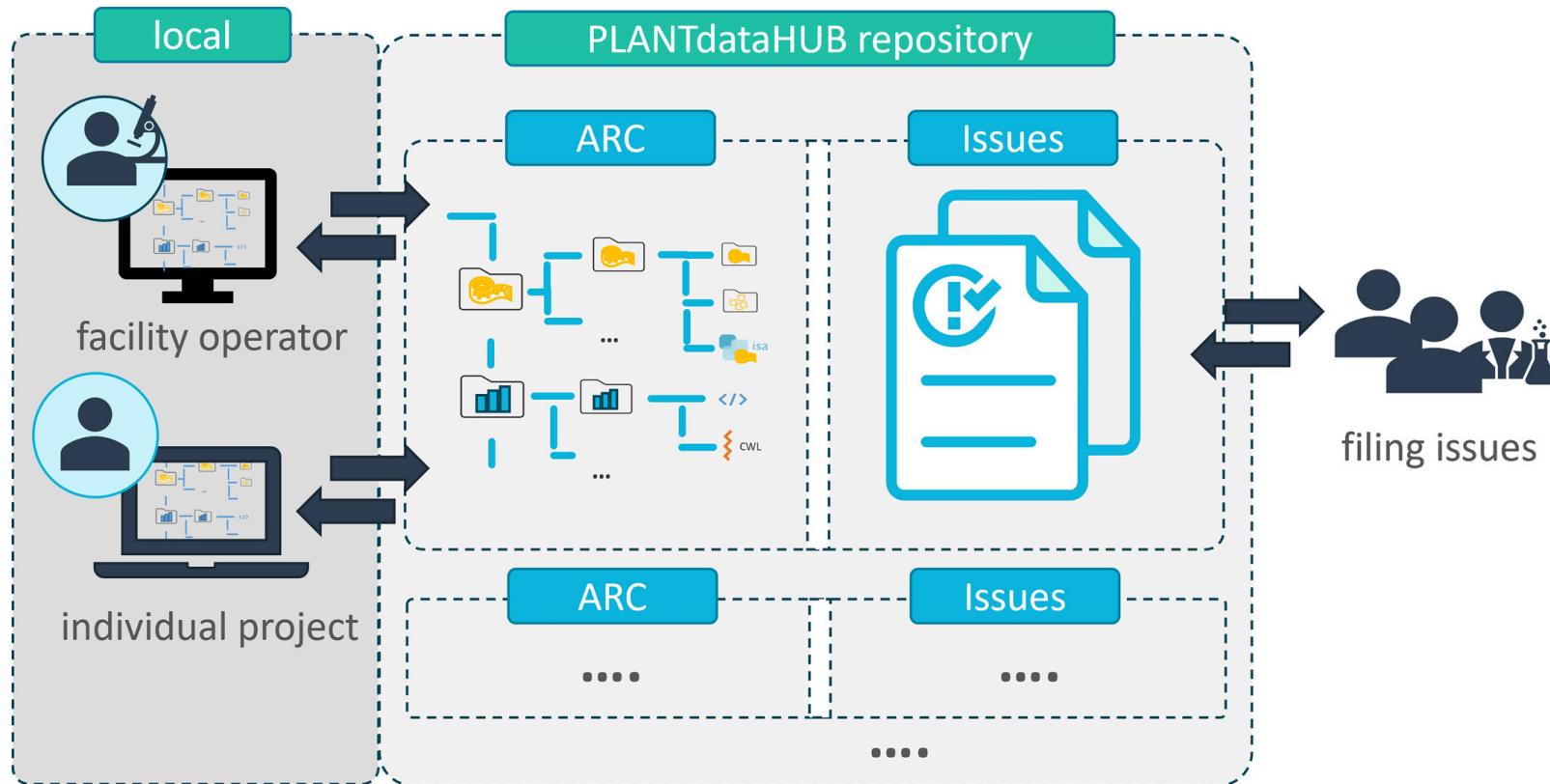




Mutable data life cycle



Project management



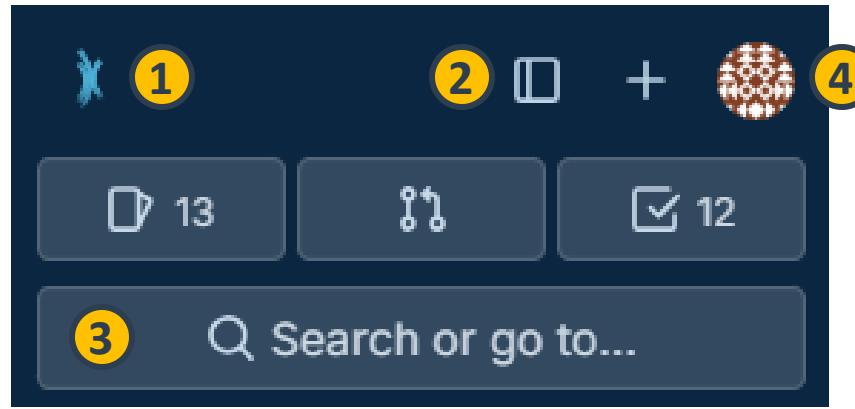
Contributors

Slides presented here include contributions by

- name: Dominik Brilhaus
github: <https://github.com/brilator>
orcid: <https://orcid.org/0000-0001-9021-3197>
- name: Cristina Martins Rodrigues
github: <https://github.com/CMR248>
orcid: <https://orcid.org/0000-0002-4849-1537>

DataHub Hands-On

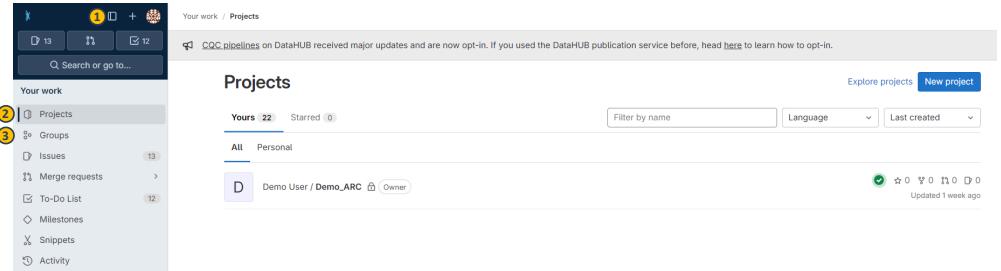
Navigation Bar



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

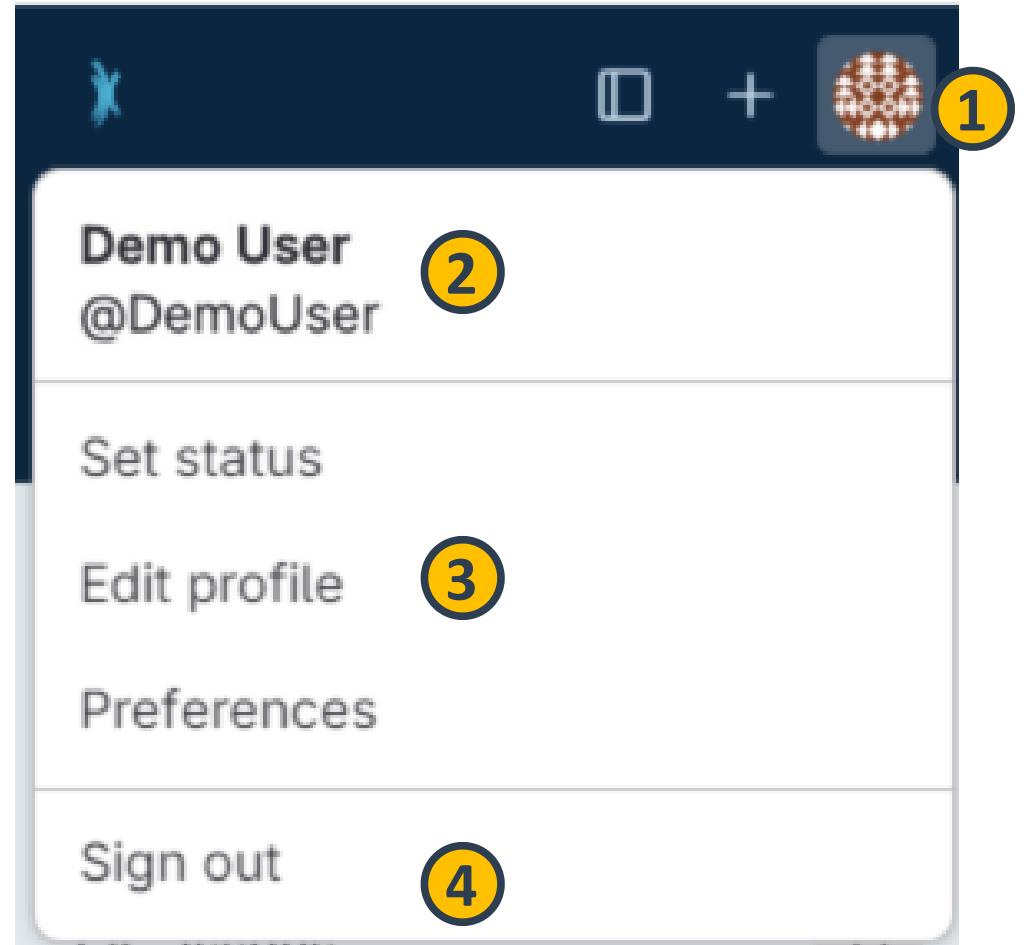
Hamburger Menu

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



Avatar Menu

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



Projects Panel

The screenshot shows the 'Projects' panel in DataHub. On the left is a sidebar with 'Your work' sections: Projects (selected), Groups, Issues (13), Merge requests, To-Do List (12), Milestones, Snippets, and Activity. At the bottom are DataPLANT and CC-BY logos. The main area has tabs 'Your work / Projects' and 'CQC pipelines...' (with a note about opt-in). It includes a search bar, filters for 'Yours' (22), 'Starred' (0), 'All' (selected), 'Personal', and 'Explore projects' (5) or 'New project' (6). A project card for 'Demo User / Demo_ARC' (Owner) is shown, with metrics: 0 stars, 0 forks, 0 issues, 0 pull requests, and updated 23 minutes ago. Callouts numbered 1 through 6 point to various UI elements.

- ① Choose a tab (1) to see only your ARCs, or explore other publicly available ARCs.
- ② Click on a project card (2) to view its details.
- ③ View project metrics (3).
- ④ View project activity (4).
- ⑤ Explore projects (5).
- ⑥ New project (6).

ARC Panel

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

The screenshot shows the ARC Panel interface for a project named "Demo_ARC".

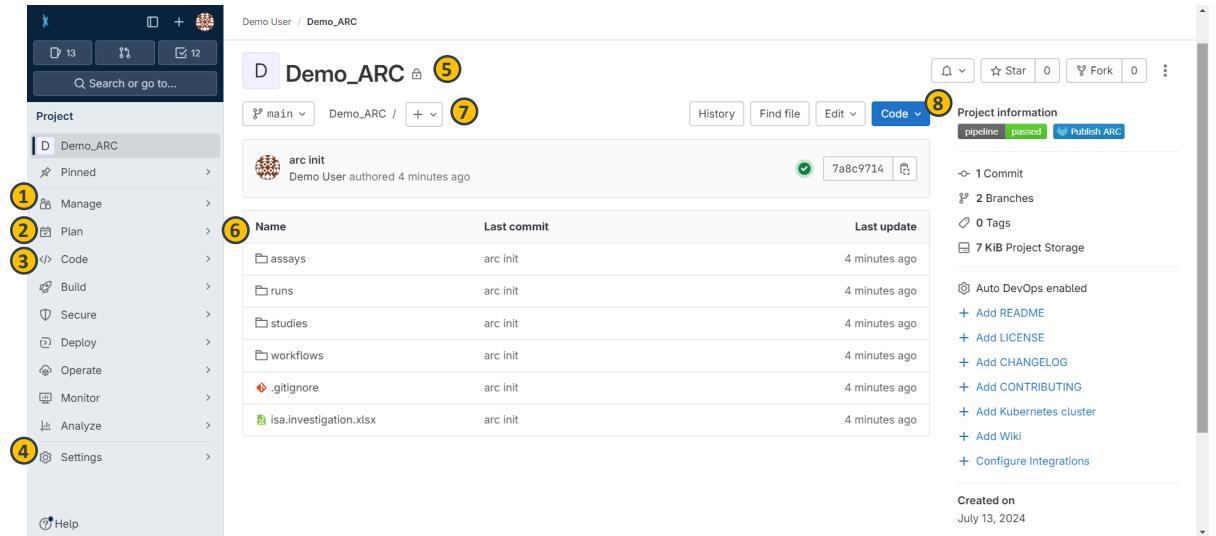
- Left Sidebar:** A navigation menu with the following items:
 - Manage (1)
 - Plan (2)
 - Code (3)
 - Build
 - Secure
 - Deploy
 - Operate
 - Monitor
 - Analyze
 - Settings (4)
 - Help
- Project Header:** Shows the project name "Demo_ARC" with a lock icon (5) and a commit count (7).
- Code Repository:** A table showing file commits:

| 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 |

(6)
- Code Tab:** A dropdown menu with options: History, Find file, Edit, Code (selected), and Project information.
- Project Information:** Displays pipeline status (passed), project stats (1 Commit, 2 Branches, 0 Tags, 7 KiB Project Storage), and integration links (Auto DevOps enabled, Add README, Add LICENSE, Add CHANGELOG, Add CONTRIBUTING, Add Kubernetes cluster, Add Wiki, Configure Integrations).
- Bottom Right:** Created on July 13, 2024.

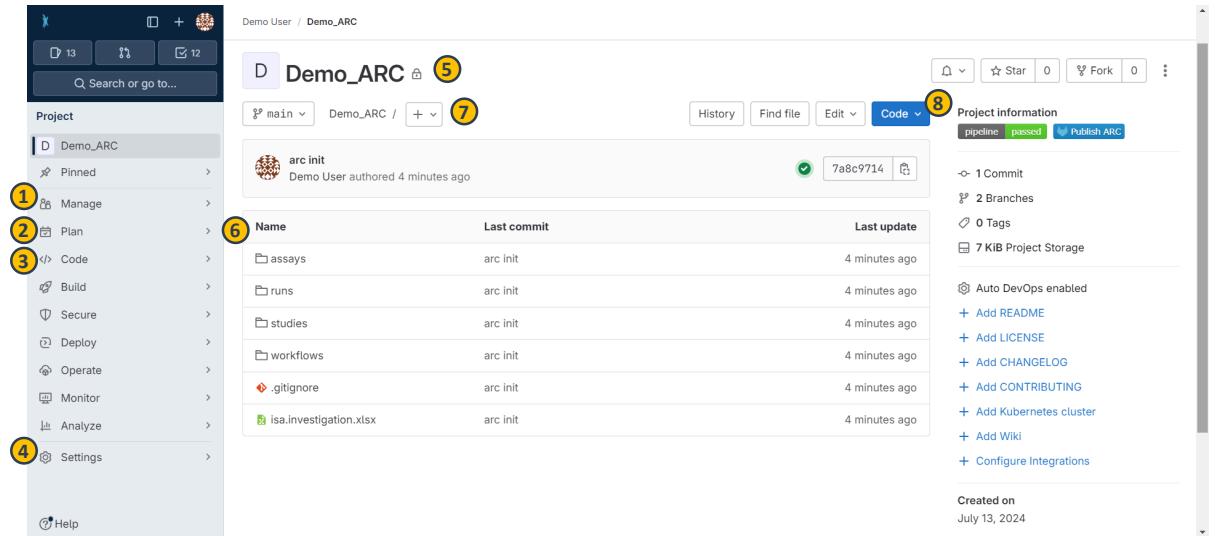
ARC Panel – sidebar

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



ARC Panel – main panel

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



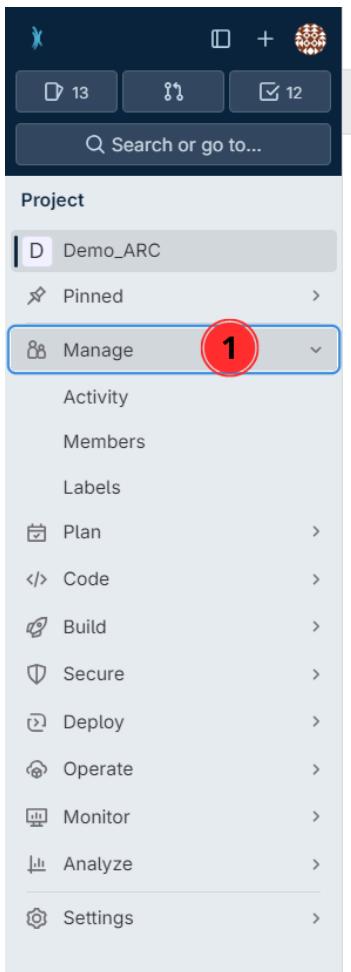
Collaborate and share



Invite collaborators

- Unless changed, your ARC is set to private by default.
- To collaborate, you can invite lab colleagues or project partners to your ARC by following the steps on the subsequent slides.
- To get started [sign in](#) to the DataHUB and open the ARC you want to share.

1. Click on Project Information in the left navigation panel



The screenshot shows the DataHUB project management interface. On the left, a sidebar lists various project management sections: Project, Pinned, Manage (highlighted with a red circle), Activity, Members, Labels, Plan, Code, Build, Secure, Deploy, Operate, Monitor, Analyze, and Settings. The main area displays the 'Demo_ARC' project details. At the top, there's a message about CQC pipelines. Below it, the project name 'Demo_ARC' is shown with a lock icon, and a breadcrumb trail indicates the current location: Demo_ARC / Demo_ARC. A 'Code' dropdown menu is open. To the right, there's a 'Project information' section with a green 'pipeline passed' status and a 'Publish ARC' button. The 'Manage' section also includes a 'Project storage' table with rows for assays, runs, studies, workflows, .gitignore, and isa.investigation.xlsx, all last updated 1 week ago.

| Name | Last commit | Last update |
|------------------------|-------------|-------------|
| assays | arc init | 1 week ago |
| runs | arc init | 1 week ago |
| studies | arc init | 1 week ago |
| workflows | arc init | 1 week ago |
| .gitignore | arc init | 1 week ago |
| isa.investigation.xlsx | arc init | 1 week ago |

Project information

pipeline passed | Publish ARC

-o 1 Commit

2 Branches

0 Tags

21 KiB Project Storage

Auto DevOps enabled

Wiki

+ Add README

+ Add LICENSE

+ Add CHANGELOG

+ Add CONTRIBUTING

+ Add Kubernetes cluster

+ Configure Integrations

Created on

July 13, 2024

2. Click on Members

The screenshot shows the DataHub interface for the project 'Demo_ARC'. The left sidebar has a 'Project' section with 'Demo_ARC' pinned, 'Manage' (circled 1), 'Activity', 'Members' (circled 2), 'Labels', 'Plan', 'Code', 'Build', 'Secure', 'Deploy', 'Operate', 'Monitor', 'Analyze', and 'Settings'. The main area is titled 'Project members' and says 'You can invite a new member to Demo_ARC or invite another group.' It shows one member: 'Demo User @DemoUser' (It's you). The member details include 'Source: Direct member by Demo User', 'Max role: Owner', and an 'Expiration' section with 'Expiration date: Sep 27, 2023', 'Jul 13, 2024', and 'Jul 21, 2024'. Buttons for 'Import from a project', 'Invite a group', and 'Invite members' are at the top right.

| Account | Source | Max role | Expiration | Activity |
|----------------------------|-------------------------------|----------|---|----------|
| Demo User @DemoUser | Direct member by Demo User | Owner | Expiration date Sep 27, 2023 ✓ Jul 13, 2024 ✗ Jul 21, 2024 | |

3. Click on Invite members

The screenshot shows the 'Members' section of the DataHub interface for the 'Demo_ARC' project. The left sidebar has 'Members' selected (2). The main area shows one member ('Demo User') with an 'Invite members' button (3) highlighted.

Demo User / Demo_ARC / Members

CQC_pipelines on DataHUB received major updates and are now opt-in. If you used the DataHUB publication service before, head [here](#) to learn how to opt-in.

Project members

You can invite a new member to Demo_ARC or invite another group.

Members 1

Filter members Account ▾

| Account | Source | Max role | Expiration | Activity |
|---|-------------------------------|----------|---|----------------------------------|
| Demo User @DemoUser It's you | Direct member by Demo User | Owner | Expiration date <input type="button" value="Sep 27, 2023"/> | ✓ Jul 13, 2024 ✗ Jul 21, 2024 |

4. Search for potential collaborators

Invite members

You're inviting members to the **Demo_ARC** project.

Username, name or email address 4

Select members or type email addresses

Select a role

Guest ▼

[Read more about role permissions](#)

5. Select a role

Invitations

X

ARC project.

4

5

Guest

Reporter

Developer

Maintainer

Owner

Guest

Read more about role permissions

Choosing the proper role

Guests

Have the least rights. They will not be able to see the content of your ARC (only the wiki page).

Reporters

Have **read access** to your ARC. This is recommended for people you ask for consultancy.

Developers

The choice for most people you want to invite to your ARC. Developers have **read and write access**, but cannot maintain the project on the DataHUB, e.g. inviting others.

Maintainers

Gives the person the same rights as you have (except of removing you from your own project). This is recommended for inviting PIs or group leaders allowing them to add their group members for data upload or analysis to the project as well.

A detailed list of all permissions for the individual roles can be found [here](#)

Congratulations!

You have just shared your ARC with a collaborator.



Version control

- Commit history

Project Management

- Issues

ARCs come with their own wiki space

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

The screenshot shows a user interface for a research collaboration platform. On the left is a sidebar with a dark header containing icons for search, refresh, and notifications (13, 11, 12). Below the header are buttons for 'Search or go to...', 'Project' (Demo_ARC), 'Pinned', 'Manage', 'Plan', 'Issues' (0), 'Issue boards', 'Milestones', 'Wiki' (selected), 'Code', 'Build', 'Secure', and 'Help'. The main content area has a header 'Demo User / Demo_ARC / Wiki / Home'. A message at the top right says 'QC pipelines on DataHUB received major updates and are now opt-in. If you used the DataHUB publication service before, head [here](#) to learn how to opt-in.' Below this is a 'Home' section with an 'Edit' button, a timestamp 'Last edited by Demo User just now', and a note 'This is the wiki to Demo_ARC. We will announce meeting schedules here.' To the right are sections for 'Pages' (2), 'Home', 'Meeting Schedule' (with items for Kick-off, Proposal discussion, and RNA-seq pipeline), and 'Ideas and drafts' (with a Golden Gate protocol item).

Contributors

Slides presented here include contributions by

- name: Dominik Brilhaus
github: <https://github.com/brilator>
orcid: <https://orcid.org/0000-0001-9021-3197>
- name: Cristina Martins Rodrigues
github: <https://github.com/CMR248>
orcid: <https://orcid.org/0000-0002-4849-1537>



Structure your project as an ARC



1. Follow the slides to build an ARC for your project
2. Go back to the shared folder prepared before class
3. Add data from your project to your ARC

Q&A and Wrap-up Day1

Preparation for next day

- Please try to prepare your own ARC
- Please install SWATe

Resources

DataPLANT (nfdi4plants)

Website: <https://nfdi4plants.org/>

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

DataHUB: <https://git.nfdi4plants.org>

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

HelpDesk: <https://helpdesk.nfdi4plants.org>

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

Overview of Institutional services at UoC and HHU

UoC

- C3RDM: <https://fdm.uni-koeln.de/en/home>
- Data storage and sharing: <https://rrzk.uni-koeln.de/daten-speichern-teilen>
- HPC: <https://rrzk.uni-koeln.de/hpc-projekte>
- service overview: <https://fdm.uni-koeln.de/en/rdm-services/service-catalogue>

HHU

- RDM Competence Center: <https://www.fdm.hhu.de>
- Support for research including HPC: <https://www.zim.hhu.de/servicekatalog/forschungsunterstuetzung>
- Processing & storing data: <https://www.zim.hhu.de/servicekatalog/rechnen-und-speichern>

Five-Finger-Feedback





CEPLAS

Cluster of Excellence on Plant Sciences

Good Data Management Practices

*part of M4468 – Plant developmental genetics, evolution
and biostatistics in the CEPLAS research program*

November 10th, 2023

Dominik Brilhaus, CEPLAS Data Science



House-keeping

Pad: <https://pad.hhu.de/oI-NjeUtSHSMzk5huWRkJw>

Tentative agenda

Day 2

| Time | Topics |
|---------------|-------------------------|
| 09:30 - 10:30 | Feedback session |
| 10:30 - 10:45 | <i>Short break</i> |
| 10:45 - 12:30 | ISA, Metadata and Swate |
| 12:30 - 13:30 | <i>Lunch</i> |
| 13:30 - 15:00 | Hands-on Swate |
| 15:00 - 15:30 | Data Publication |
| 15:30 - 16:00 | Wrap-up |

Points to discuss from and since day 1

Metadata and ISA

**What is
metadata?**

Viola's PhD Project

Exercise: Take 5 minutes to note down the metadata

Viola investigates the effect of the plant circadian clock on sugar metabolism in *W. mirabilis*. For her PhD project, which is part of an EU-funded consortium in Prof. Beetroot's lab, she acquires seeds from a South-African botanical society. Viola grows the plants under different light regimes, harvests leaves from a two-day time series experiment, extracts polar metabolites as well as RNA and submits the samples to nearby core facilities for metabolomics and transcriptomics measurements, respectively. After a few weeks of iterative consultation with the facilities' heads as well as technicians and computational biologists involved, Viola receives back a wealth of raw and processed data. From the data she produces figures and wraps everything up to publish the results in the Journal of Wonderful Plant Sciences.

Metadata everywhere

Viola investigates the effect of the plant circadian clock on sugar metabolism in *W. mirabilis*. For her PhD project, which is part of an EU-funded consortium in Prof. Beetroot's lab, she acquires seeds from a South-African botanical society. Viola grows the plants under different light regimes, harvests leaves from a two-day time series experiment, extracts polar metabolites as well as RNA and submits the samples to nearby core facilities for metabolomics and transcriptomics measurements, respectively. After a few weeks of iterative consultation with the facilities' heads as well as technicians and computational biologists involved, Viola receives back a wealth of raw and processed data. From the data she produces figures and wraps everything up to publish the results in the Journal of Wonderful Plant Sciences.

Project metadata

project design

- researcher
- institute and project
- biological context
- research question
- purpose of data collection
- ...

experimental processes

- origin and nature of the biological material
- lab protocols
- instrument model
- ...

data-analytical processes

- algorithms
- tools
- software versions and dependencies employed
- ...

Other types of metadata

bibliographic

- Title
- Publication date and title
- Description
- Author
- Contacts
- Keywords
- ...

legal or administrative

- data origin, ownership, provenance,
- licensing
- ethical aspects
- ...

technical

- expected data volume
- storage location
- file formats
- ...

Metadata from a FAIR perspective

Findable

- metadata names the content of the data
- basis for search engines
- makes it categorizable for people and machines

Interoperable

- metadata identifies software and file formats
- required conversions between file formats

Reusable

- obtain and reuse research data according to clear rules described in licenses

Accessible

- information about origin
- location of storage
- access rights

Metadata "Standards"

Examples from [Minimum Information for Biological and Biomedical Investigations \(MIBBI\)](#):

- MIAPPE | Minimum Information About a Plant Phenotyping Experiment
<https://www.miappe.org>
 - MIAME | Minimum Information About a Microarray Experiment
<https://www.fged.org/projects/miame/>
 - MIAPE | Minimum Information About a Proteomics Experiment
<https://www.psidev.info/miape>
 - MINSEQE | Minimum Information about a high-throughput SEQuencing Experiment
<https://www.fged.org/projects/minseqe>
-  Check out <https://fairsharing.org/> for more examples

Metadata standards ≈ Checklists

- Determine (minimal) required information
- Usually **do not** determine the format (i.e. shape or file type)

A small Interactive detour

-> favorite Movie

How does google "know"?!

Google X 🔍

Bilder Videos Cast Bedeutung Handlung Hinkebein Netflix Soundtrack Tanz Alle Filter ▾ | Suchfilter

Ungefähr 37.300.000 Ergebnisse (0,39 Sekunden)

Pulp Fiction FSK 16 1994 · 2 h 34 min : Übersicht Besetzung Film ansehen Rezensionen Trailer und Clips

Besetzung >



Quentin Tarantino
John Travolta
Samuel L. Jackson
Uma Thurman
Bruce Willis
Tim Roth

Jimmie Dimmick
Vincent Vega
Jules Winnfield
Mia Wallace
Butch Coolidge
Pumpkin

Wikipedia https://de.wikipedia.org/wiki/Pulp_Fiction :

Pulp Fiction

Pulp Fiction ist ein US-amerikanischer Gangsterfilm von und mit Quentin Tarantino aus dem Jahr 1994. Der Film wurde für sieben Oscars nominiert – darunter ...

[Maria de Medeiros](#) · [Peter Greene](#) · [Eric Stoltz](#) · [Paul Calderón](#)

Weitere Fragen

Was ist so besonders an Pulp Fiction? ▾

Was bedeutet der Titel Pulp Fiction? ▾

Warum ist Pulp Fiction ein Kultfilm? ▾

Film ansehen

DIENSTE BEARBEITEN

 Jetzt ansehen Premium-Abo  Angesehen  Möchte ich sehen

 Ab 2,99 €  Ansehen

 Ab 2,99 €  Ansehen

 Ab 3,99 €  Ansehen

[Alle Optionen zum Ansehen](#) ▾

Info

 Pulp Fiction | Official Trailer (HD) - John Tra...  1:39

8,9/10  4,8/5  4,5/5  

IMDb Amazon Wer streamt ...

Dieser Film gefiel 92 % der Nutzer   

Google-Nutzer

Schemas and machine-readability

Structured data and the internet

Schema.org

- create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, ...
- Structured data can be used to *mark up* all kinds of items from products to events to recipes
- Communicate with search engines (-> SEO, search engine optimization)
- Enhance findability from search engine results
- Provide context to an ambiguous webpage
- Metadata interoperability and standardization across all website using schema.org

Structured data and the internet: Schema.org

<https://schema.org/Person>

```
<script type="application/ld+json">
{
  "@context": "https://schema.org",
  "@type": "Person",
  "address": {
    "@type": "PostalAddress",
    "addressLocality": "Seattle",
    "addressRegion": "WA",
    "postalCode": "98052",
    "streetAddress": "20341 Whitworth Institute 405 N. Whitworth"
  },
  "colleague": [
    "http://www.xyz.edu/students/alicejones.html",
    "http://www.xyz.edu/students/bobsmith.html"
  ],
  "email": "mailto:jane-doe@xyz.edu",
  "image": "janedoe.jpg",
  "jobTitle": "Professor",
  "name": "Jane Doe",
  "telephone": "(425) 123-4567",
  "url": "http://www.janedoe.com"
}
</script>
```

JSON-LD

JSON-LD = JavaScript Object Notation for Linked Data

```
<script type="application/ld+json">
{
  "@context": "https://schema.org",
  "@type": "SportsTeam",
  "name": "San Francisco 49ers",
  "member": {
    "@type": "OrganizationRole",
    "member": {
      "@type": "Person",
      "name": "Joe Montana"
    },
    "startDate": "1979",
    "endDate": "1992",
    "roleName": "Quarterback"
  }
}
</script>
```

RDFa

RDFa = Resource Description Framework in Attributes

```
<div vocab="http://schema.org/" typeof="SportsTeam">
  <span property="name">San Francisco 49ers</span>
  <div property="member" typeof="OrganizationRole">
    <div property="member" typeof="http://schema.org/Person">
      <span property="name">Joe Montana</span>
    </div>
    <span property="startDate">1979</span>
    <span property="endDate">1992</span>
    <span property="roleName">Quarterback</span>
  </div>
</div>
```

Standards

Dublin Core

<https://www.dublincore.org/schemas/>

DataCite Schema

- Schema: <http://schema.datacite.org/meta/kernel-4.3/metadata.xsd>
- Full Example: <https://schema.datacite.org/meta/kernel-4.3/example/datacite-example-full-v4.xml>

DataCite Schema: Simple Example

```
...
<identifier identifierType="DOI">10.5072/D3P26Q35R-Test</identifier>
<creators>
  <creator>
    <creatorName nameType="Personal">Fosmire, Michael</creatorName>
    <givenName>Michael</givenName>
    <familyName>Fosmire</familyName>
  </creator>
  <creator>
    <creatorName nameType="Personal">Wertz, Ruth</creatorName>
    <givenName>Ruth</givenName>
    <familyName>Wertz</familyName>
  </creator>
  <creator>
    <creatorName nameType="Personal">Purzer, Senay</creatorName>
    <givenName>Senay</givenName>
    <familyName>Purzer</familyName>
  </creator>
</creators>
<titles>
  <title xml:lang="en">Critical Engineering Literacy Test (CELT)</title>
</titles>
<publisher xml:lang="en">Purdue University Research Repository (PURR)</publisher>
<publicationYear>2013</publicationYear>
<subjects>
  <subject xml:lang="en">Assessment</subject>
  <subject xml:lang="en">Information Literacy</subject>
  <subject xml:lang="en">Engineering</subject>
  <subject xml:lang="en">Undergraduate Students</subject>
  <subject xml:lang="en">CELT</subject>
  <subject xml:lang="en">Purdue University</subject>
</subjects>
<language>en</language>
<resourceType resourceTypeGeneral="Dataset">Dataset</resourceType>
...

```

Ontologies

Ontology

(Sometimes also referred to "semantic model")

An ontology combines features of

- a **dictionary**,
- a **taxonomy**, and
- a **thesaurus**

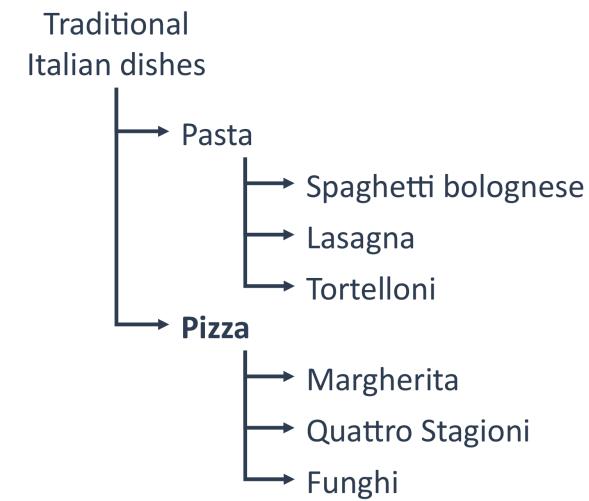
Dictionary

Alphabetically lists terms and their definitions

Pizza: *"a dish made typically of flattened bread dough spread with a savory mixture usually including tomatoes and cheese and often other toppings and baked"*

Taxonomy

Hierarchy or classification



Thesaurus

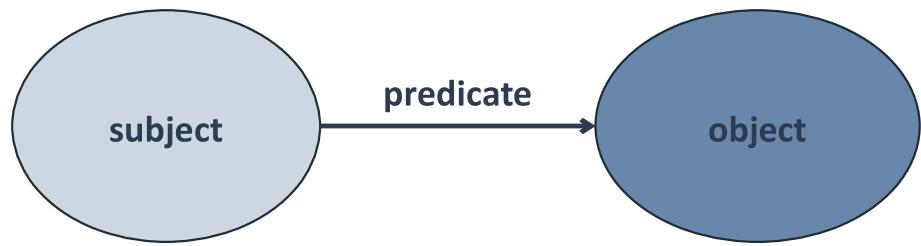
Dictionary of synonyms and relations

Pizza ≈ Lahmacun ≈ Focaccia ≈ Flammkuchen

Ontology

- Structures a set of **concepts** in a particular area and the relations between them in a **graph-like manner**
- Can be used in disambiguation, defining hierarchies, a standard to define terms
- Define a common vocabulary of concepts and their relationships to **model** a particular domain while making it **machine understandable**

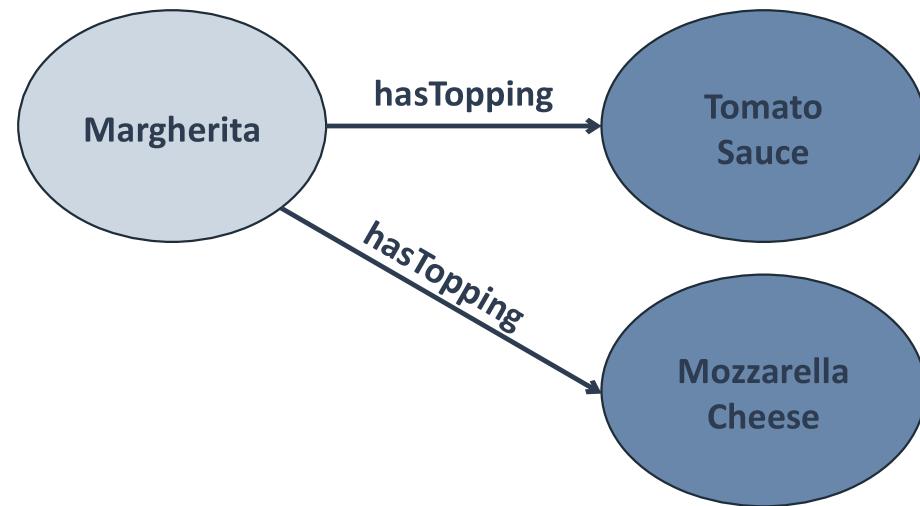
The semantic triple



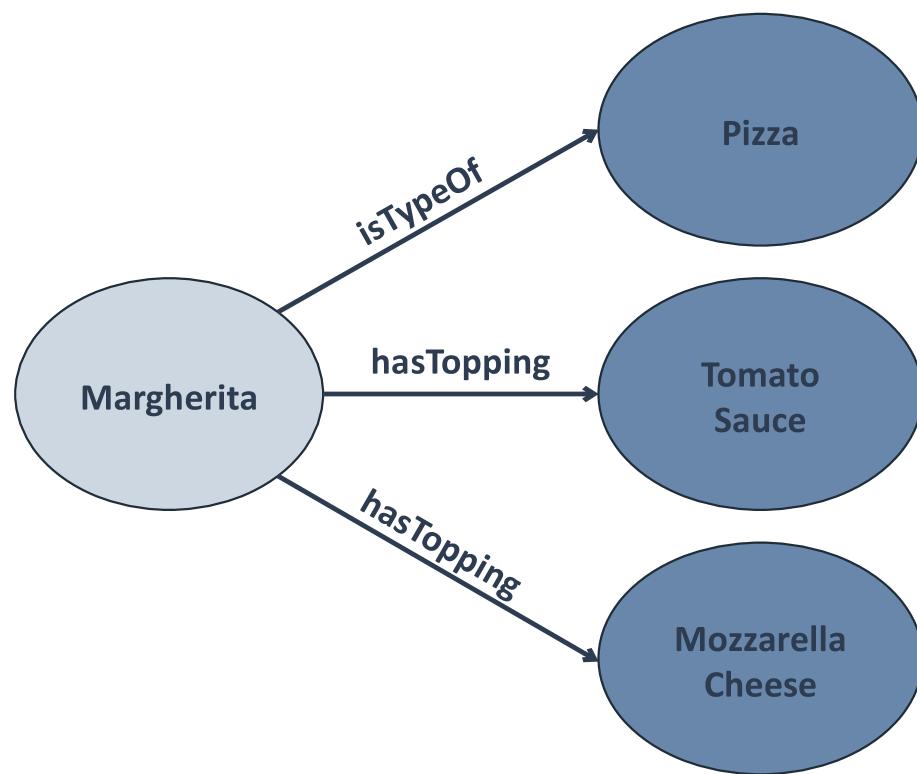
Modeling a pizza menu



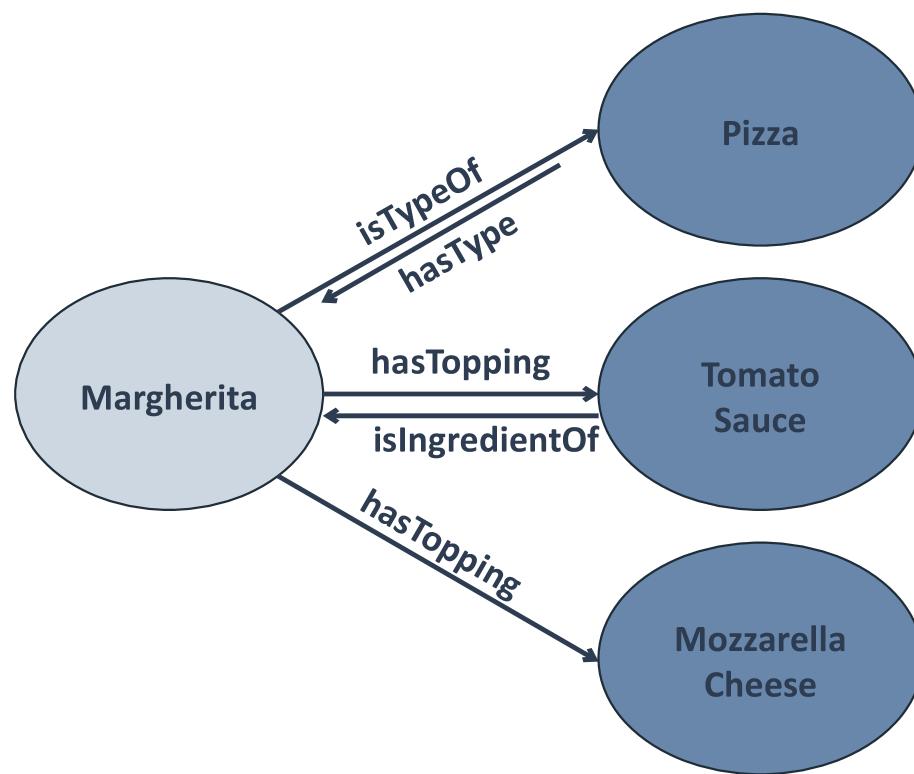
Modeling a pizza menu



Modeling a pizza menu

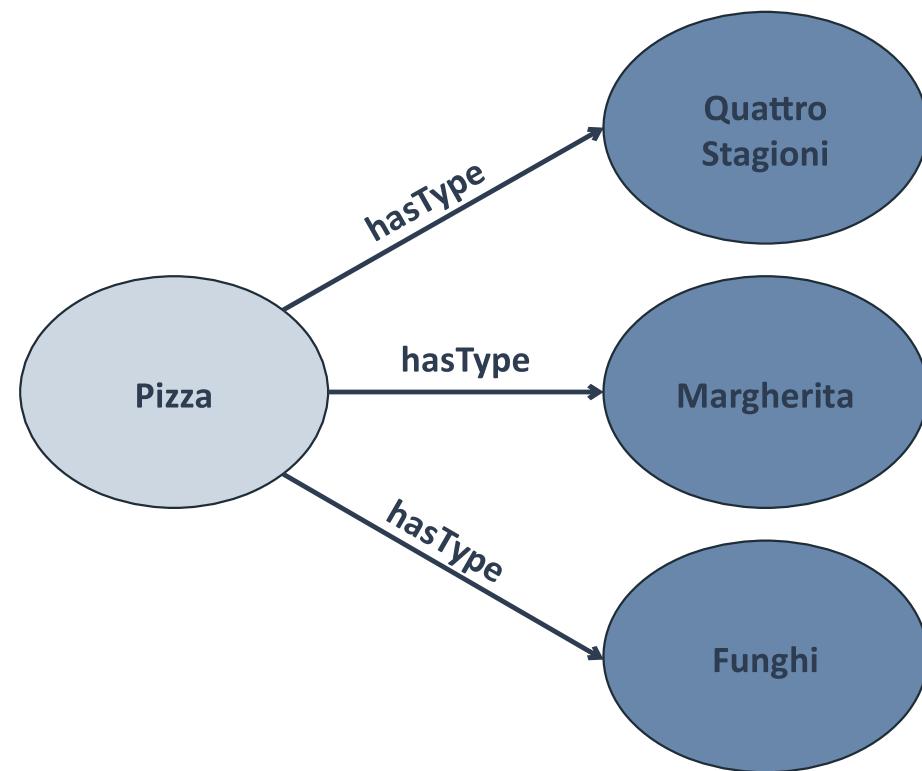


Predicates have two directions

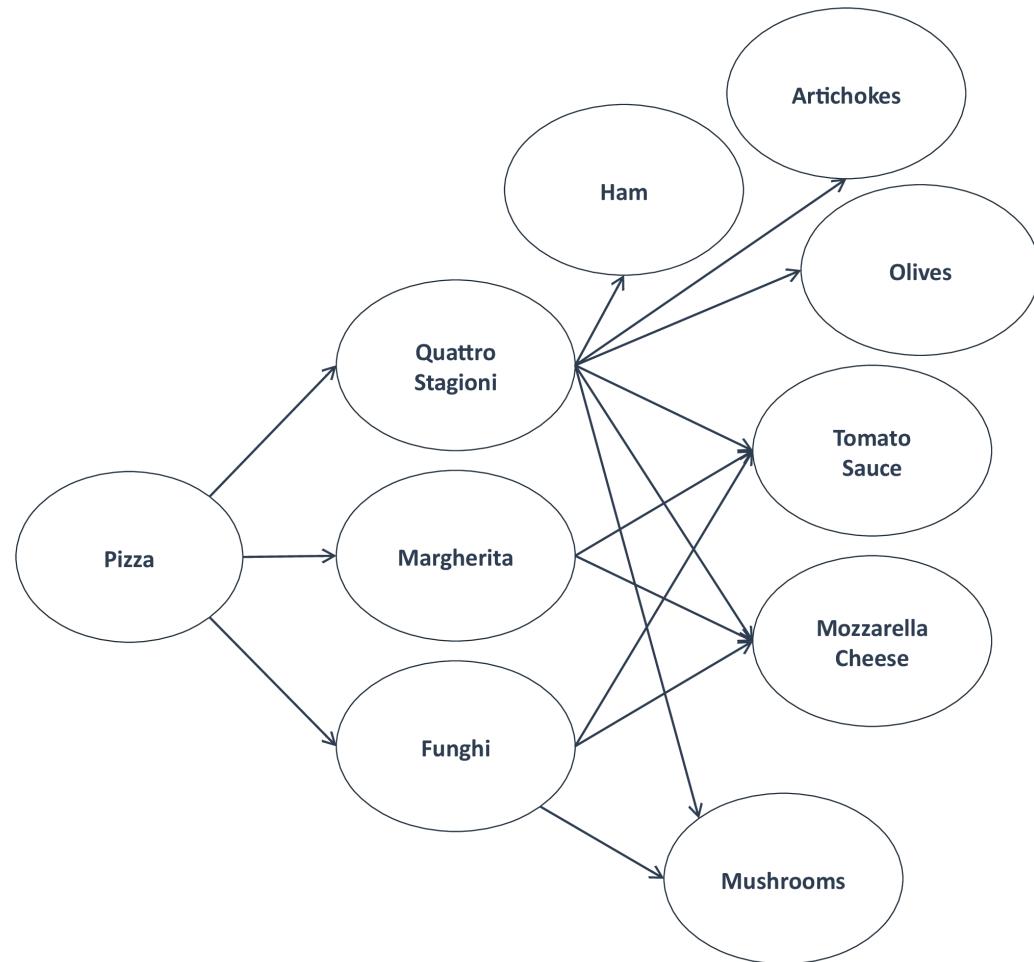


Looking at the menu from a different perspective

An object of one triplet can be the subject to another



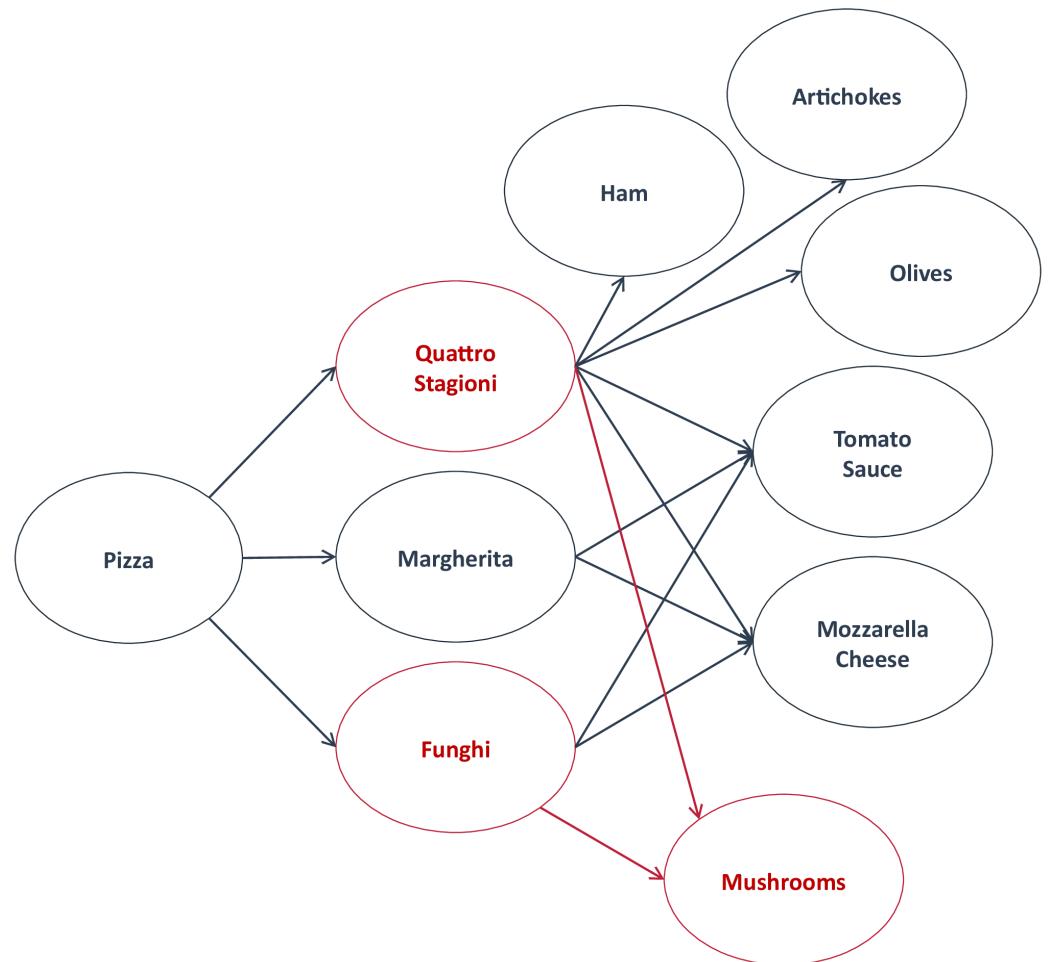
(Towards) a knowledge graph



Searching the menu

An ontology can be queried:

- *"name all pizzas with topping mushrooms"*



The Pizza Ontology

- Example from protege: <https://protege.stanford.edu/ontologies/pizza/pizza.owl>
- Visualize via WebVOWL <http://vowl.visualdataweb.org/webvowl.html>

Example ontologies

EDAM ontology

- Description: <http://edamontology.org/page>
- Browser: <https://edamontology.github.io/edam-browser>

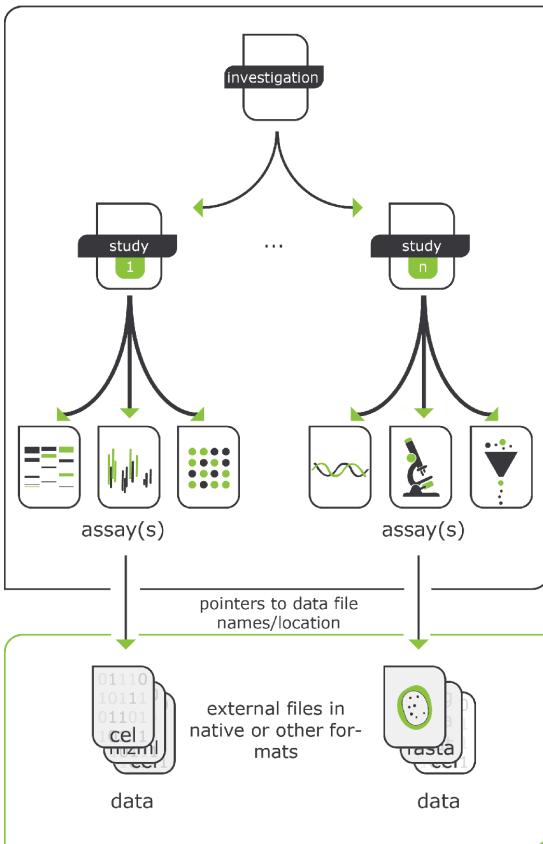
PECO ontology

- Human-readable: <https://www.ebi.ac.uk/ols/ontologies/peco>
- Raw (OWL): <http://purl.obolibrary.org/obo/peco.owl>

Explore more examples

- <https://www.ebi.ac.uk/ols/>
- <https://bioportal.bioontology.org>

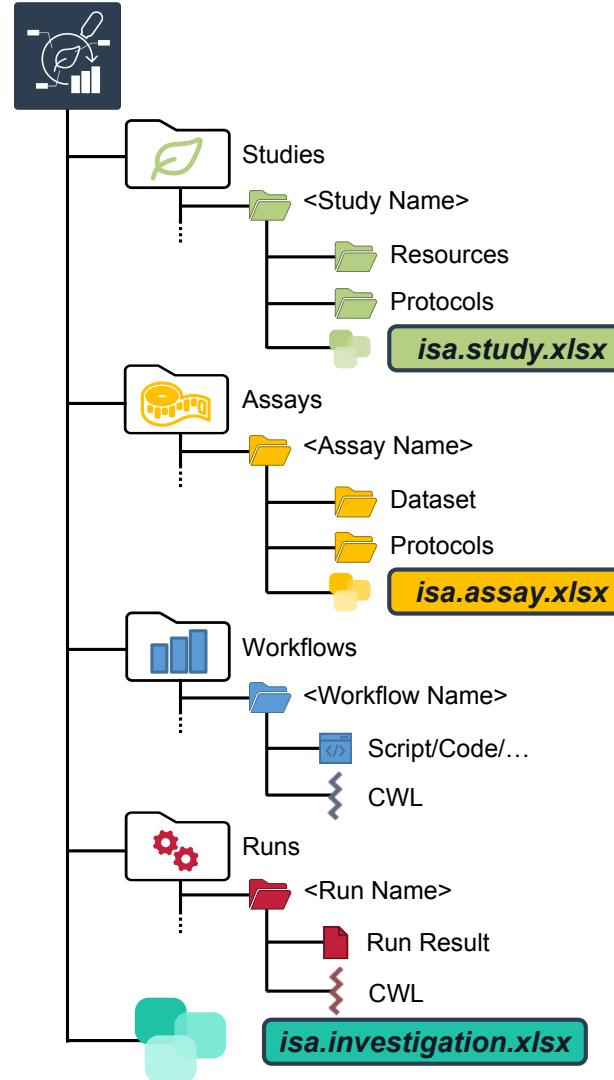
ARC builds on ISA



Investigation
Overall goals
Scientific context

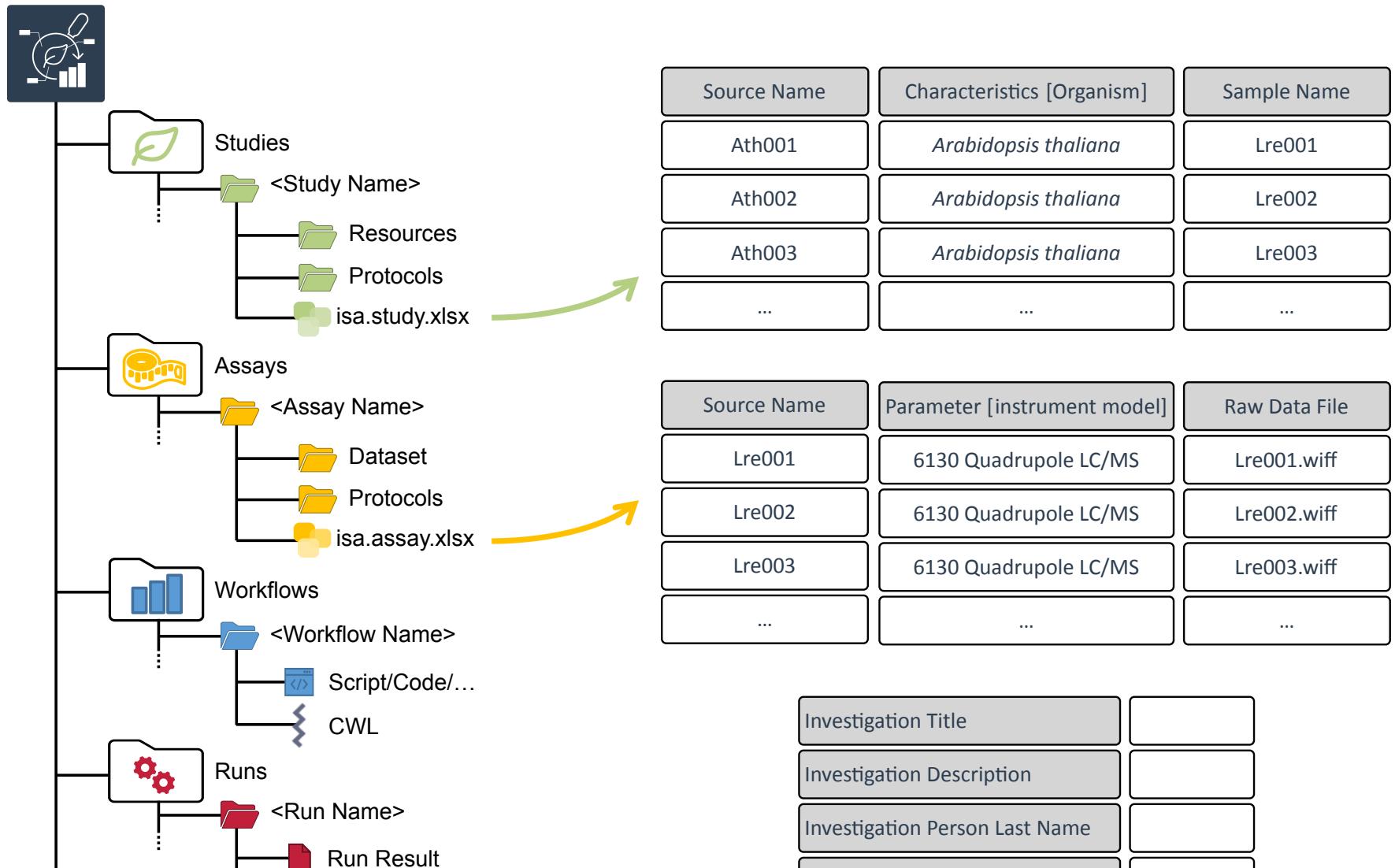
Study
Experimental design

Assay
Leading to (raw) data



ARC builds on ISA

Metadata Annotations



isa.<>.xlsx files within ARCs

isa.investigation.xlsx

| A | B | C | D | E | F | G | H | I | J |
|---|--|--------------------|-------------------------|-----------------|---------|-----------------------|------------|---|---|
| ONTOLOGY SOURCE REFERENCE | | | | | | | | | |
| 1 Term Source Name | OBI | BTO | NENT | LO | CHEBI | PATO | EFO | | |
| 2 | http://biportal.bioArrayExpress Experimental Factor Ontology | | | | | | | | |
| 3 Term Source Version | 47893 | v1.26 | v1.26 | v1.26 | v1.26 | v1.26 | v1.26 | | |
| 4 Term Source Description | Ontology for Biomed BRENDa tissue / NEWT UniProt Tax Unit Ontology Chemical Ent Phenotypic c ArrayExpress Experimental Factor Ontology | | | | | | | | |
| 5 INVESTIGATION | | | | | | | | | |
| 6 Investigation Identifier | BII-1 | | | | | | | | |
| 7 Investigation Title | Growth control of the eukaryote cell: a systems biology study in yeast | | | | | | | | |
| 8 Investigation Description | Background Cell growth underlies many key cellular and developmental processes, yet a limited number of studies have been carried out on cell growth. | | | | | | | | |
| 9 Investigation Start Date | 30.04.07 | | | | | | | | |
| 10 Investigation End Date | 10.05.09 | | | | | | | | |
| 11 Investigation Public Release Date | | | | | | | | | |
| 12 Comment[Entered With Configuration] | isaconfig default_v2013_02_13 | | | | | | | | |
| 13 Comment[Last Opened With Configuration] | | | | | | | | | |
| 14 INVESTIGATION PUBLICATIONS | | | | | | | | | |
| 15 Investigation PubMed ID | 17439666 | | | | | | | | |
| 16 Investigation Publication DOI | doi:10.1186/jbio54 | | | | | | | | |
| 17 Investigation Publication Author List | Castroillo J, Zeeb LA, Hoyle DC, Zhang N, Hayes A, Gardner DC, Cornell M, Petty J, Hakes L, Wardleworth L, Rash B, Brown M, Dunn WB, Broadhurst | | | | | | | | |
| 18 Investigation Publication Title | Growth control of the eukaryote cell: a systems biology study in yeast. | | | | | | | | |
| 19 Investigation Publication Status | published | | | | | | | | |
| 20 Investigation Publication Status Term Accession Number | | | | | | | | | |
| 21 Investigation Publication Status Term Source REF | | | | | | | | | |
| 22 Investigation Publication Status Term Source URI | | | | | | | | | |
| 23 Investigation Person Last Name | Stephen | Castroillo | Zeeb | | | | | | |
| 24 Investigation Person First Name | Oliver | Juan | Lao | | | | | | |
| 25 Investigation Person Mid Initials | G | | A | | | | | | |
| 26 Investigation Person Email | | | | | | | | | |
| 27 Investigation Person Phone | | | | | | | | | |
| 28 Investigation Person Fax | | | | | | | | | |
| 29 Investigation Person Address | Oxford Road, Manch Oxford Road, M2 Oxford Road, Manchester M13 9PT, UK | | | | | | | | |
| 30 Investigation Person Affiliation | Faculty of Life Sci: Faculty of Life Sci: Faculty of Life Sciences; Michael Smith Building, University of Manchester | | | | | | | | |
| 31 Investigation Person Roles | corresponding authc author | | | | | | | | |
| 32 Investigation Person Roles Term Accession Number | | | | | | | | | |
| 33 Investigation Person Roles Term Source REF | | | | | | | | | |
| 34 Comment[Investigation Person REF] | | | | | | | | | |
| 35 STUDY | | | | | | | | | |
| 36 Study Identifier | BII-S-1 | | | | | | | | |
| 37 Study Title | Study of the impact of changes in flux on the transcriptome, proteome, endometabolome and exometabolome of the yeast <i>Saccharomyces cerevisiae</i> . We wished to study the impact of growth rate on the total complement of mRNA molecules, proteins, and metabolites in <i>S. cerevisiae</i> . Independent | | | | | | | | |
| 38 Study Description | | | | | | | | | |
| 39 Comment[Study Grant Number] | | | | | | | | | |
| 40 Comment[Study Funding Agency] | | | | | | | | | |
| 41 Study Submission Date | 30.04.07 | | | | | | | | |
| 42 Study Public Release Date | 10.05.09 | | | | | | | | |
| 43 Study File | 5_BII-S-1.txt | | | | | | | | |
| 44 STUDY DESIGN DESCRIPTORS | | | | | | | | | |
| 45 Study Design Type | Intervention design | | | | | | | | |
| 46 Study Design Type Term Accession Number | http://purl.obolibrary.org/obo/OBI_0001115 | | | | | | | | |
| 47 Study Design Type Term Source REF | OBI | | | | | | | | |
| 48 STUDY PUBLICATIONS | | | | | | | | | |
| 49 Study PubMed ID | 17439666 | | | | | | | | |
| 50 Study Publication DOI | doi:10.1186/jbio54 | | | | | | | | |
| 51 Study Publication Author List | Castroillo J, Zeeb LA, Hoyle DC, Zhang N, Hayes A, Gardner DC, Cornell M, Petty J, Hakes L, Wardleworth L, Rash B, Brown M, Dunn WB, Broadhurst | | | | | | | | |
| 52 Study Publication Title | Growth control of the eukaryote cell: a systems biology study in yeast. | | | | | | | | |
| 53 Study Publication Status | published | | | | | | | | |
| 54 Study Publication Status Term Accession Number | | | | | | | | | |
| 55 Study Publication Status Term Source REF | | | | | | | | | |
| 56 STUDY FACTORS | | | | | | | | | |
| 57 Study Factor Name | limiting nutrient | rate | | | | | | | |
| 58 Study Factor Type | chemical compound | rate | | | | | | | |
| 59 Study Factor Type Term Accession Number | http://purl.obolibrary.org/obo/PATO_000161 | | | | | | | | |
| 60 Study Factor Type Term Source REF | PATO | | | | | | | | |
| 61 STUDY ASSAYS | | | | | | | | | |
| 62 Study Assay Measurement Type | protein expression | metabolite profile | transcription profiling | | | | | | |
| 63 Study Assay Measurement Type Term Accession Number | http://purl.obolibrary.org/obo/OBI_0400148 | | | | | | | | |
| 64 Study Assay Measurement Type Term Source REF | OBI | OBI | OBI | | | | | | |
| 65 Study Assay Technology Type | mass spectrometry | mass spectrometry | RNA microarray | | | | | | |
| 66 Study Assay Technology Type Term Accession Number | http://purl.obolibrary.org/obo/OBI_0400148 | | | | | | | | |
| 67 Study Assay Technology Type Term Source REF | OBI | OBI | OBI | | | | | | |
| 68 Study Assay Technology Platform | ITRAQ | LC-MS/MS | Affymetrix | | | | | | |
| 69 Study Assay File Name | a_proteome.txt | a_metabolome.txt | a_transcriptome.txt | | | | | | |
| 70 STUDY PROTOCOLS | | | | | | | | | |
| 71 Study Protocol Name | growth | mRNA extraction | protein extraction | biotin labeling | EuKEW54 | metabolite extraction | | | |
| 72 Study Protocol Type | growth | mRNA extraction | protein extraction | labeling | loading | hybridization | extraction | | |
| 73 Study Protocol Type Term Accession Number | | | | | | | | | |
| 74 Study Protocol Type Term Source REF | | | | | | | | | |
| 75 Study Protocol Description | 1. Biomass samples (1. Biomass samples (45 ml) were tak. This was done using Enzo! For each target, a hybridisation cocktail was made using the | | | | | | | | |
| 76 Study Protocol URI | | | | | | | | | |
| 77 Study Protocol | | | | | | | | | |
| 78 Study Protocol Type | | | | | | | | | |
| 79 Study Protocol Type Term Accession Number | | | | | | | | | |
| 80 Study Protocol Type Term Source REF | | | | | | | | | |
| 81 Study Protocol Description | | | | | | | | | |
| 82 Study Protocol URI | | | | | | | | | |

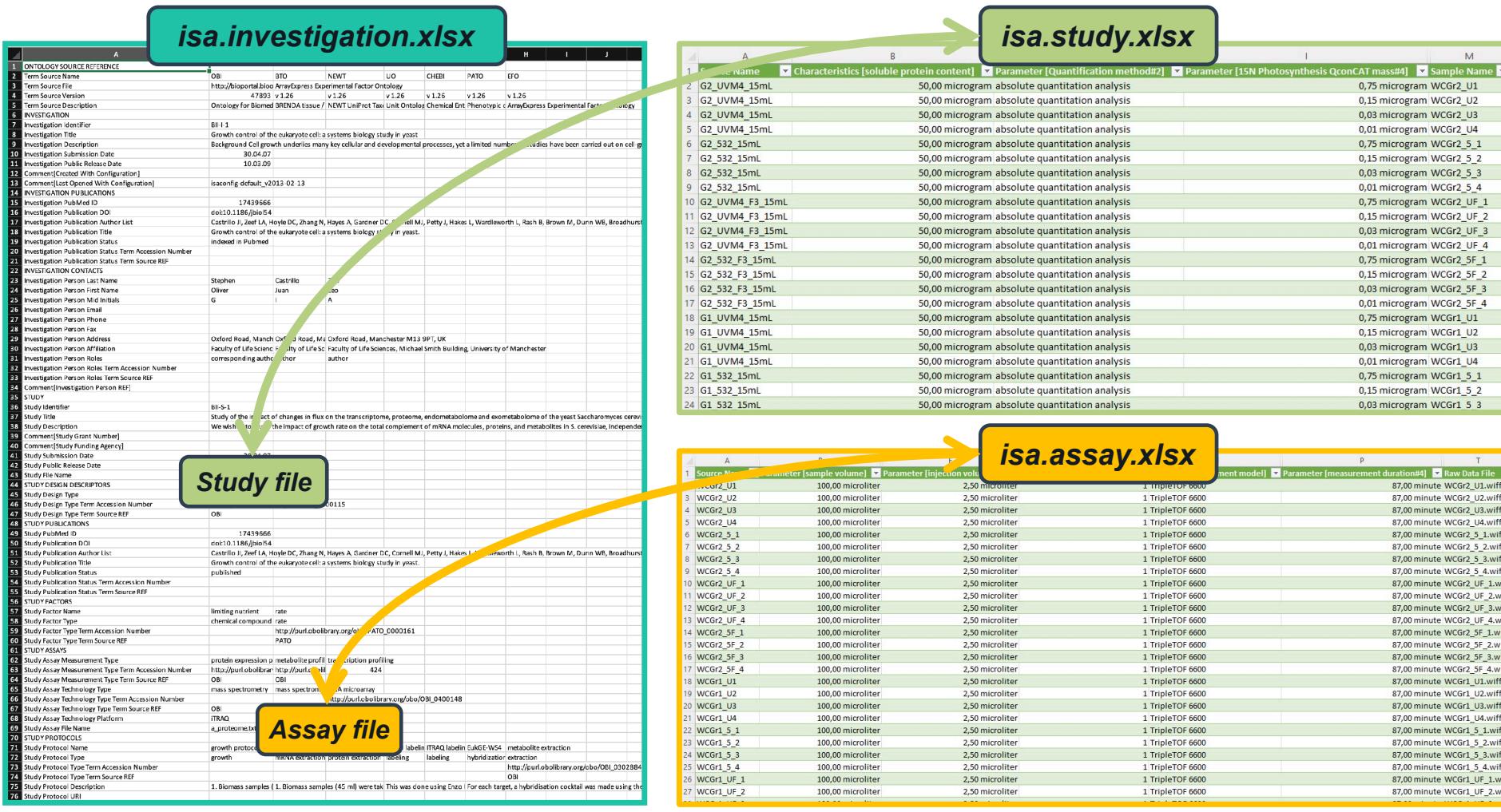
isa.study.xlsx

| A | B | C | D | E | F | G | H | I | J |
|--------------------|-------|-----------|--------------------------------|---|---|---|---|---|---|
| Source Name | | | | | | | | | |
| 1 G2_UVM4_15mL | | | | | | | | | |
| 2 G2_UVM4_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 3 G2_UVM4_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 4 G2_UVM4_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 5 G2_S32_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 6 G2_S32_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 7 G2_S32_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 8 G2_S32_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 9 G2_S32_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 10 G2_UVM4_F3_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 11 G2_UVM4_F3_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 12 G2_UVM4_F3_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 13 G2_UVM4_F3_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 14 G2_S32_F3_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 15 G2_S32_F3_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 16 G2_S32_F3_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 17 G2_S32_F3_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 18 G1_UVM4_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 19 G1_UVM4_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 20 G1_UVM4_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 21 G1_UVM4_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 22 G1_S32_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 23 G1_S32_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |
| 24 G1_S32_15mL | 50,00 | microgram | absolute quantitation analysis | | | | | | |

isa.assay.xlsx

| A | B | C | D | E | F | G | H | I | J |
|--------------------|--------|------------|------|------------|---|----------------|---|---|---|
| Source Name | | | | | | | | | |
| 1 WCGr1_U1 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 2 WCGr1_U2 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 3 WCGr1_U3 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 4 WCGr1_U4 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 5 WCGr2_U1 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 6 WCGr2_U2 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 7 WCGr2_U3 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 8 WCGr2_U4 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 9 WCGr2_U5 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 10 WCGr2_U6 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 11 WCGr2_U7 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 12 WCGr2_U8 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 13 WCGr2_U9 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 14 WCGr2_U10 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 15 WCGr2_S3_1 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 16 WCGr2_S3_2 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 17 WCGr2_S3_3 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 18 WCGr1_U1 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 19 WCGr1_U2 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 20 WCGr1_U3 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 21 WCGr1_U4 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 22 WCGr1_S3_1 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 23 WCGr1_S3_2 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 24 WCGr1_S3_3 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 25 WCGr1_U1 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 26 WCGr1_U2 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |
| 27 WCGr1_U2 | 100,00 | microliter | 2,50 | microliter | 1 | TripleTOF 6600 | | | |

Study and assay files are registered in the investigation file



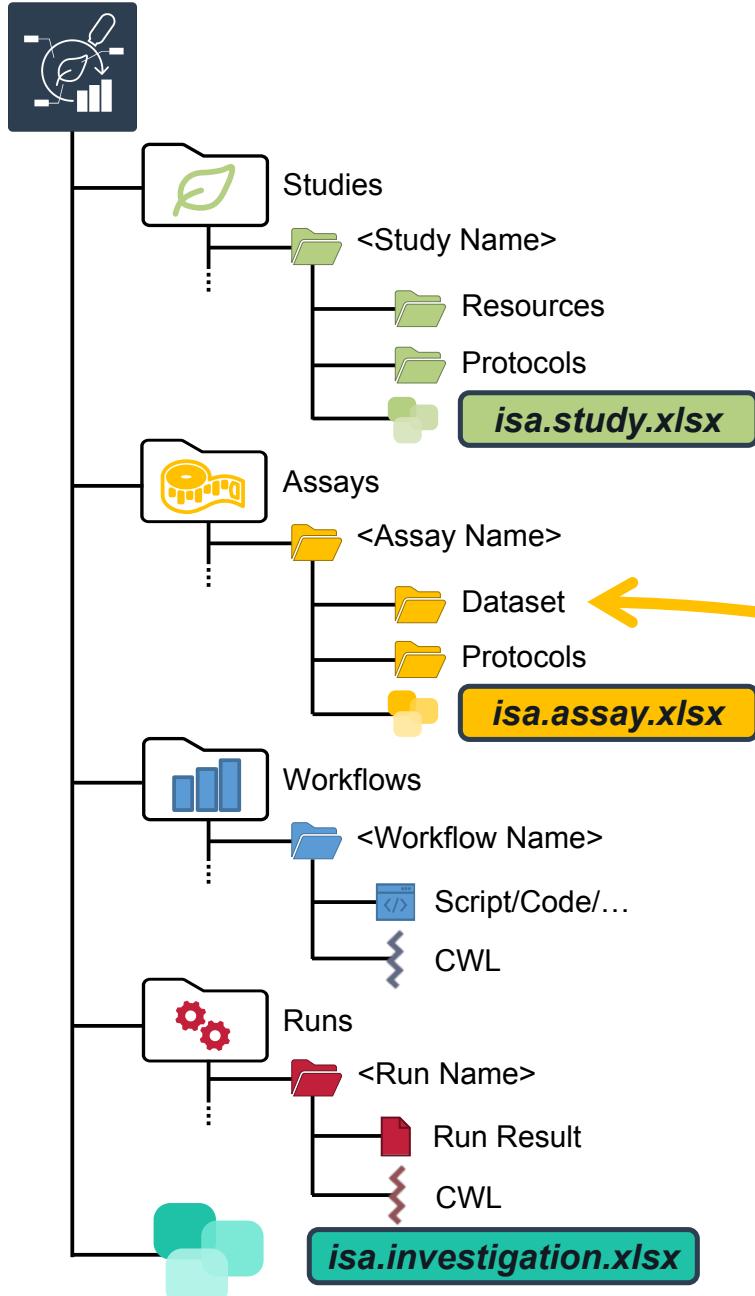
The output of a study or assay file can function as input for a new isa.assay.xlsx

Output building blocks:

- Sample Name
- Raw Data File
- Derived Data File

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|----|-----------------|---|-------------------------------------|---|-------------|---|---|---|---|---|---|---|---------------------------|
| 1 | Source Name | Characteristics [soluble protein content] | Parameter [Quantification method#2] | Parameter [15N Photosynthesis QconCAT mass#4] | Sample Name | | | | | | | | |
| 2 | G2_UVM4_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,75 microgram WCGr2_U1 |
| 3 | G2_UVM4_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,15 microgram WCGr2_U2 |
| 4 | G2_UVM4_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,03 microgram WCGr2_U3 |
| 5 | G2_UVM4_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,01 microgram WCGr2_U4 |
| 6 | G2_532_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,75 microgram WCGr2_5_1 |
| 7 | G2_532_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,15 microgram WCGr2_5_2 |
| 8 | G2_532_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,03 microgram WCGr2_5_3 |
| 9 | G2_532_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,01 microgram WCGr2_5_4 |
| 10 | G2_UVM4_F3_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,75 microgram WCGr2_UF_1 |
| 11 | G2_UVM4_F3_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,15 microgram WCGr2_UF_2 |
| 12 | G2_UVM4_F3_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,03 microgram WCGr2_UF_3 |
| 13 | G2_UVM4_F3_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,01 microgram WCGr2_UF_4 |
| 14 | G2_532_F3_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,75 microgram WCGr2_SF_1 |
| 15 | G2_532_F3_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,15 microgram WCGr2_SF_2 |
| 16 | G2_532_F3_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,03 microgram WCGr2_SF_3 |
| 17 | G2_532_F3_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,01 microgram WCGr2_SF_4 |
| 18 | G1_UVM4_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,75 microgram WCGr1_U1 |
| 19 | G1_UVM4_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,15 microgram WCGr1_U2 |
| 20 | G1_UVM4_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,03 microgram WCGr1_U3 |
| 21 | G1_UVM4_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,01 microgram WCGr1_U4 |
| 22 | G1_532_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,75 microgram WCGr1_5_1 |
| 23 | G1_532_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,15 microgram WCGr1_5_2 |
| 24 | G1_532_15mL | 50,00 microgram | absolute quantitation analysis | | | | | | | | | | 0,03 microgram WCGr1_5_3 |

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|----|-------------|---------------------------|---------------------------|---|----------------|---|---|---|---|---|---|---|------------------------------|
| 1 | Source Name | Parameter [sample volume] | Parameter [injection volu | | | | | | | | | | |
| 2 | WCGr2_U1 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr2_U1.wiff |
| 3 | WCGr2_U2 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr2_U2.wiff |
| 4 | WCGr2_U3 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr2_U3.wiff |
| 5 | WCGr2_U4 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr2_U4.wiff |
| 6 | WCGr2_5_1 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr2_5_1.wiff |
| 7 | WCGr2_5_2 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr2_5_2.wiff |
| 8 | WCGr2_5_3 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr2_5_3.wiff |
| 9 | WCGr2_5_4 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr2_5_4.wiff |
| 10 | WCGr2_UF_1 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr2_UF_1.wiff |
| 11 | WCGr2_UF_2 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr2_UF_2.wiff |
| 12 | WCGr2_UF_3 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr2_UF_3.wiff |
| 13 | WCGr2_UF_4 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr2_UF_4.wiff |
| 14 | WCGr2_SF_1 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr2_SF_1.wiff |
| 15 | WCGr2_SF_2 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr2_SF_2.wiff |
| 16 | WCGr2_SF_3 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr2_SF_3.wiff |
| 17 | WCGr2_SF_4 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr2_SF_4.wiff |
| 18 | WCGr1_U1 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr1_U1.wiff |
| 19 | WCGr1_U2 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr1_U2.wiff |
| 20 | WCGr1_U3 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr1_U3.wiff |
| 21 | WCGr1_U4 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr1_U4.wiff |
| 22 | WCGr1_5_1 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr1_5_1.wiff |
| 23 | WCGr1_5_2 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr1_5_2.wiff |
| 24 | WCGr1_5_3 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr1_5_3.wiff |
| 25 | WCGr1_5_4 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr1_5_4.wiff |
| 26 | WCGr1_UF_1 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr1_UF_1.wiff |
| 27 | WCGr1_UF_2 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | | 87,00 minute WCGr1_UF_2.wiff |



isa.study.xlsx

| A | B | C | D | E | F | G | H | I | J | K | L | M |
|----|-----------------|--|-------------------------------------|---|---------------------------|---|---|---|---|---|---|---|
| 1 | Source Name | Characteristics [soluble protein content] | Parameter [Quantification method#2] | Parameter [15N Photosynthesis QconCAT mass#4] | Sample Name | | | | | | | |
| 2 | G2_UVM4_15mL | 50,00 microgram absolute quantitation analysis | | | 0,75 microgram WCGr2_U1 | | | | | | | |
| 3 | G2_UVM4_15mL | 50,00 microgram absolute quantitation analysis | | | 0,15 microgram WCGr2_U2 | | | | | | | |
| 4 | G2_UVM4_15mL | 50,00 microgram absolute quantitation analysis | | | 0,03 microgram WCGr2_U3 | | | | | | | |
| 5 | G2_UVM4_15mL | 50,00 microgram absolute quantitation analysis | | | 0,01 microgram WCGr2_U4 | | | | | | | |
| 6 | G2_532_15mL | 50,00 microgram absolute quantitation analysis | | | 0,75 microgram WCGr2_5_1 | | | | | | | |
| 7 | G2_532_15mL | 50,00 microgram absolute quantitation analysis | | | 0,15 microgram WCGr2_5_2 | | | | | | | |
| 8 | G2_532_15mL | 50,00 microgram absolute quantitation analysis | | | 0,03 microgram WCGr2_5_3 | | | | | | | |
| 9 | G2_532_15mL | 50,00 microgram absolute quantitation analysis | | | 0,01 microgram WCGr2_5_4 | | | | | | | |
| 10 | G2_UVM4_F3_15mL | 50,00 microgram absolute quantitation analysis | | | 0,75 microgram WCGr2_UF_1 | | | | | | | |
| 11 | G2_UVM4_F3_15mL | 50,00 microgram absolute quantitation analysis | | | 0,15 microgram WCGr2_UF_2 | | | | | | | |
| 12 | G2_UVM4_F3_15mL | 50,00 microgram absolute quantitation analysis | | | 0,03 microgram WCGr2_UF_3 | | | | | | | |
| 13 | G2_UVM4_F3_15mL | 50,00 microgram absolute quantitation analysis | | | 0,01 microgram WCGr2_UF_4 | | | | | | | |
| 14 | G2_532_F3_15mL | 50,00 microgram absolute quantitation analysis | | | 0,75 microgram WCGr2_5F_1 | | | | | | | |
| 15 | G2_532_F3_15mL | 50,00 microgram absolute quantitation analysis | | | 0,15 microgram WCGr2_5F_2 | | | | | | | |
| 16 | G2_532_F3_15mL | 50,00 microgram absolute quantitation analysis | | | 0,03 microgram WCGr2_5F_3 | | | | | | | |
| 17 | G2_532_F3_15mL | 50,00 microgram absolute quantitation analysis | | | 0,01 microgram WCGr2_5F_4 | | | | | | | |
| 18 | G1_UVM4_15mL | 50,00 microgram absolute quantitation analysis | | | 0,75 microgram WCGr1_U1 | | | | | | | |
| 19 | G1_UVM4_15mL | 50,00 microgram absolute quantitation analysis | | | 0,15 microgram WCGr1_U2 | | | | | | | |
| 20 | G1_UVM4_15mL | 50,00 microgram absolute quantitation analysis | | | 0,03 microgram WCGr1_U3 | | | | | | | |
| 21 | G1_UVM4_15mL | 50,00 microgram absolute quantitation analysis | | | 0,01 microgram WCGr1_U4 | | | | | | | |
| 22 | G1_532_15mL | 50,00 microgram absolute quantitation analysis | | | 0,75 microgram WCGr1_5_1 | | | | | | | |
| 23 | G1_532_15mL | 50,00 microgram absolute quantitation analysis | | | 0,15 microgram WCGr1_5_2 | | | | | | | |
| 24 | G1_532_15mL | 50,00 microgram absolute quantitation analysis | | | 0,03 microgram WCGr1_5_3 | | | | | | | |

isa.assay.xlsx

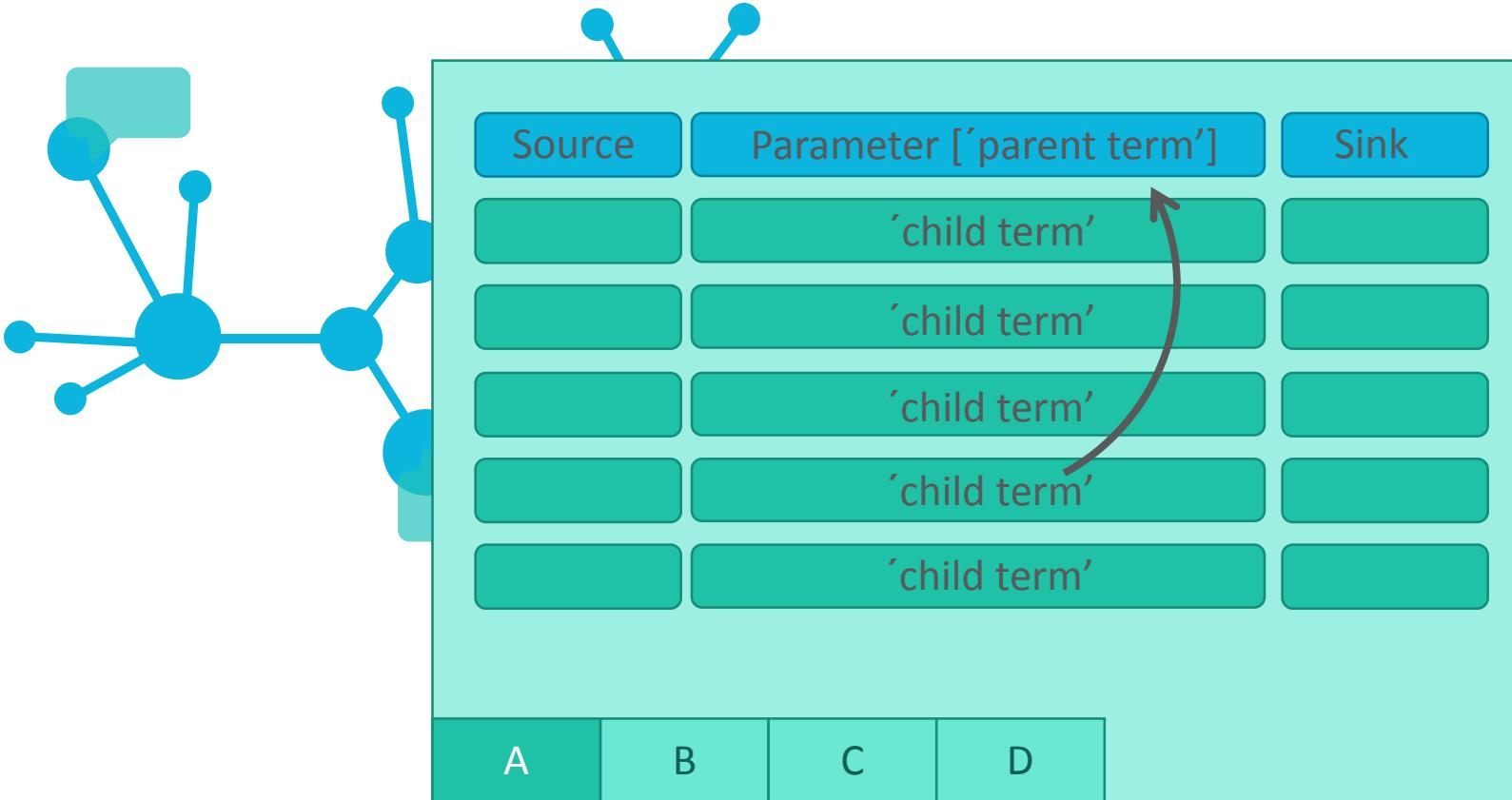
| A | B | C | D | E | F | G | H | I | J | K | L | M |
|----|-------------|---------------------------|---------------------------|---|----------------|---|---|---|---|---|---|---|
| 1 | Source Name | Parameter [sample volume] | Parameter [injection vol. | | | | | | | | | |
| 2 | WCGr2_U1 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 3 | WCGr2_U2 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 4 | WCGr2_U3 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 5 | WCGr2_U4 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 6 | WCGr2_5_1 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 7 | WCGr2_5_2 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 8 | WCGr2_5_3 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 9 | WCGr2_5_4 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 10 | WCGr2_UF_1 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 11 | WCGr2_UF_2 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 12 | WCGr2_UF_3 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 13 | WCGr2_UF_4 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 14 | WCGr2_5F_1 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 15 | WCGr2_5F_2 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 16 | WCGr2_5F_3 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 17 | WCGr2_5F_4 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 18 | WCGr1_U1 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 19 | WCGr1_U2 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 20 | WCGr1_U3 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 21 | WCGr1_U4 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 22 | WCGr1_5_1 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 23 | WCGr1_5_2 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 24 | WCGr1_5_3 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 25 | WCGr1_5_4 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 26 | WCGr1_UF_1 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |
| 27 | WCGr1_UF_2 | 100,00 microliter | 2,50 microliter | 1 | TripleTOF 6600 | | | | | | | |

Raw data

The screenshot shows a file browser with a folder containing WIFF files. The files are named after the runs listed in the assay table, such as WCGr2_5_1.wiff, WCGr2_5_2.wiff, etc. A yellow box highlights this folder, and an arrow points from the 'Raw data' label in the assay table to this folder.

Swate

Annotation by flattening the knowledge graph



- Low-friction metadata annotation
- Familiar spreadsheet, row/column-based environment

Annotation principle

| Sample | Parameter [instrument model] | Data |
|--------|------------------------------|------|
| | 'TripleTOF4600' | |
| A | B | C |
| D | | |

- Low-friction metadata annotation
- Familiar spreadsheet, row/column-based environment

Adding new building blocks (columns)

The screenshot shows a Microsoft Excel spreadsheet titled "isa.study (1).xlsx" with a table of experimental data. The table has columns for Source Name, Protocol Type, Characteristic [sample label], Factor [temperature], Parameter [Instrument model], Component [Software], and Sample Name. A callout bubble points to the "Parameter" column with the text "New Parameter".

The Swate add-in is open on the right side of the screen, displaying a "Building Blocks" interface. It lists various instrument models and instruments, each with a dropdown arrow. The interface includes a search bar and a note about parameter columns.

| Source Name | Protocol Type | Characteristic [sample label] | Factor [temperature] | Parameter [Instrument model] | Component [Software] | Sample Name |
|-----------------|--------------------------|-------------------------------|----------------------|------------------------------|----------------------|-------------|
| G2_UVMA_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr2_U1 |
| G2_UVMA_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr2_U2 |
| G2_UVMA_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr2_U3 |
| G2_UVMA_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr2_U4 |
| G2_532_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr2_5_1 |
| G2_532_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr2_5_2 |
| G2_532_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr2_5_3 |
| G2_532_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr2_5_4 |
| G2_UVMA_F3_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr2_UF_1 |
| G2_UVMA_F3_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr2_UF_2 |
| G2_UVMA_F3_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr2_UF_3 |
| G2_UVMA_F3_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr2_UF_4 |
| G2_532_F3_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr2_SF_1 |
| G2_532_F3_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr2_SF_2 |
| G2_532_F3_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr2_SF_3 |
| G2_532_F3_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr2_SF_4 |
| G1_UVMA_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr1_U1 |
| G1_UVMA_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr1_U2 |
| G1_UVMA_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr1_U3 |
| G1_UVMA_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr1_U4 |
| G1_532_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr1_5_1 |
| G1_532_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr1_5_2 |
| G1_532_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr1_5_3 |
| G1_532_15mL | data extraction protocol | 15N | 30,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr1_5_4 |
| G1_UVMA_F7_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr1_UF_1 |
| G1_UVMA_F7_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr1_UF_2 |
| G1_UVMA_F7_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr1_UF_3 |
| G1_UVMA_F7_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr1_UF_4 |
| G1_532_F10_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr1_SF_1 |
| G1_532_F10_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr1_SF_2 |
| G1_532_F10_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr1_SF_3 |
| G1_532_F10_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr1_SF_4 |
| G3_UVMA_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr3_U1 |
| G3_UVMA_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr3_U2 |
| G3_UVMA_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr3_U3 |
| G3_UVMA_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr3_U4 |
| G3_532_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr3_5_1 |
| G3_532_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr3_5_2 |
| G3_532_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr3_5_3 |
| G3_532_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr3_5_4 |
| G3_UVMA_F1_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr3_UF_1 |
| G3_UVMA_F1_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr3_UF_2 |
| G3_UVMA_F1_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr3_UF_3 |
| G3_UVMA_F1_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr3_UF_4 |
| G3_532_F2_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr3_SF_1 |
| G3_532_F2_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr3_SF_2 |
| G3_532_F2_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr3_SF_3 |
| G3_532_F2_15mL | data extraction protocol | 15N | 4,00 degree Celsius | 6130 Quadrupole LC/MS | Analyst | WCGr3_SF_4 |

Annotation Building Block types

- Source Name (Input)
- Protocol Columns
 - Protocol Type, Protocol Ref
- Characteristic
- Parameter
- Factor
- Component
- Output Columns
 - Sample Name, Raw Data File, Derived Data File

The screenshot shows a Microsoft Excel spreadsheet titled 'isa.study (1).xlsx' with a single sheet named 'Sheet1'. The data consists of approximately 50 rows of experimental runs, each with columns for Source Name, Protocol Type, Characteristic, Factor, Component, and Sample Name. Several annotations are overlaid on the data:

- New Parameter**: A callout points to the 'Protocol Type/Protocol Ref' column.
- Characteristic**: A callout points to the 'Characteristic [sample label]' column.
- Component**: A callout points to the 'Component [instrument model]' column.
- Factor**: A callout points to the 'Factor [temperature]' column.
- Sample Name/ Raw Data File Derived Data File**: A callout points to the 'Sample Name' column.

To the right of the spreadsheet, the 'Swate' application interface is visible, showing a sidebar with 'Building Blocks' and a list of annotations like 'Parameter', 'Instrument Model', and 'Instrument'.

Let's take a detour on [Annotation Principles | slides](#)

Ontology term search

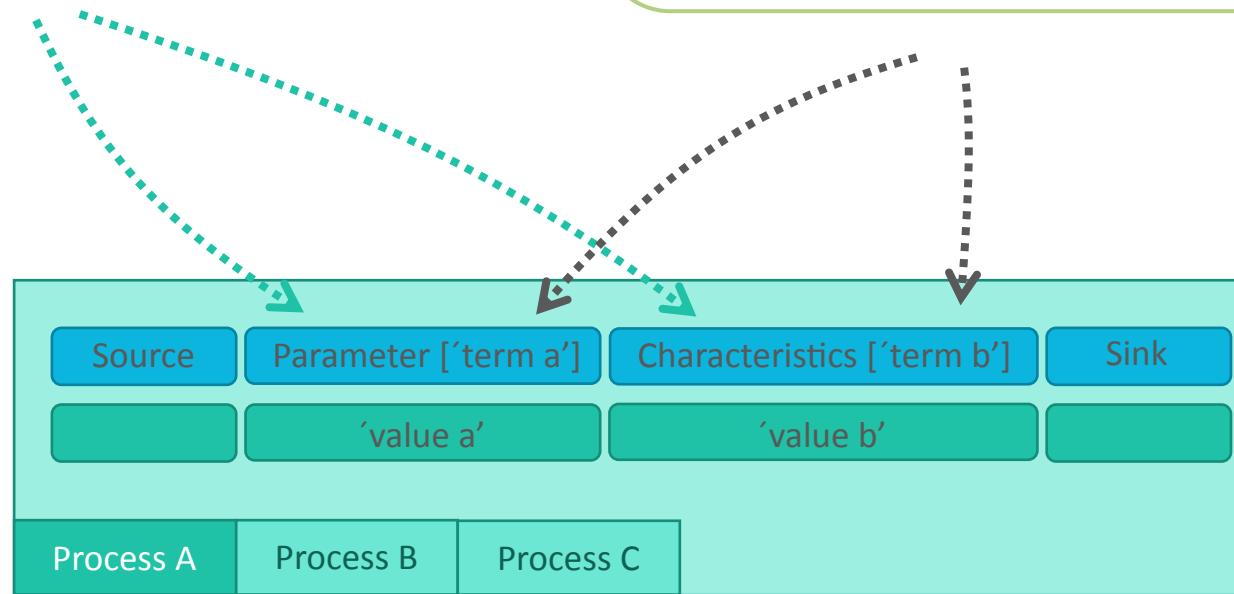
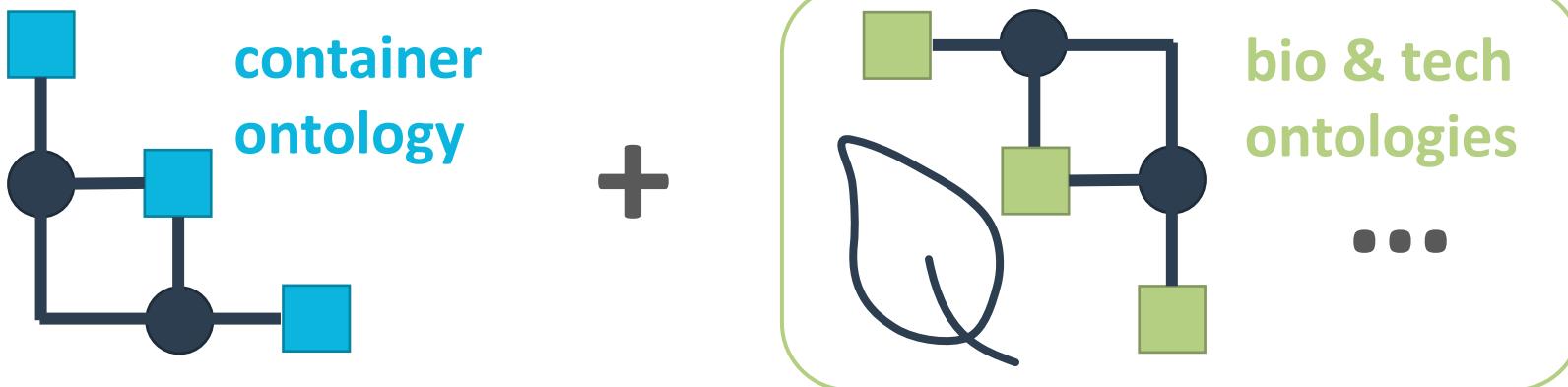
The screenshot shows a Microsoft Excel spreadsheet titled "Sheet1" with data in columns A through AB. The data consists of rows 1 through 52, each containing information such as Source Name, Protocol Type, Characteristic [sample label], Factor [temperature], Parameter [Instrument model], Component [software], and Sample Name. The "Sample Name" column contains entries like "WCGR2_U1", "WCGR2_U2", etc. The "Parameter [Instrument model]" column contains entries like "Analyst", "WCGR2_SF_1", etc.

To the right of the spreadsheet, a "Swate" window is open. The title bar says "SWATE". The main area is titled "Ontology term search" and contains the text "Search for an ontology term to fill into the selected field(s)". Below this is a search input field with the value "instrument n 6130". Underneath the input field, it says "6130 Quadrupole MS:1000470 LC/MS". There is also a link "Advanced Search!" and a note "Cant find the Term you are looking for? Try Advanced Search!". At the bottom of the window, there is a link "Still can't find what you need? Get in contact with us!" and the text "Swate Release Version 0.6.2".

The Excel ribbon at the top includes tabs for File, Home, Insert, Draw, Page Layout, Formulas, Data, Review, View, Help, and Table Design. The "Data" tab is selected, showing various tools like Get Data, Queries & Connections, Sort & Filter, Data Tools, Forecast, Outline, Group, Ungroup, Subtotal, Core, Experts, Swate, and Swate.

Fill your table with ontology terms

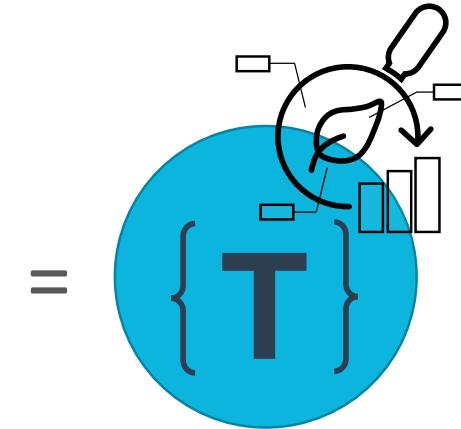
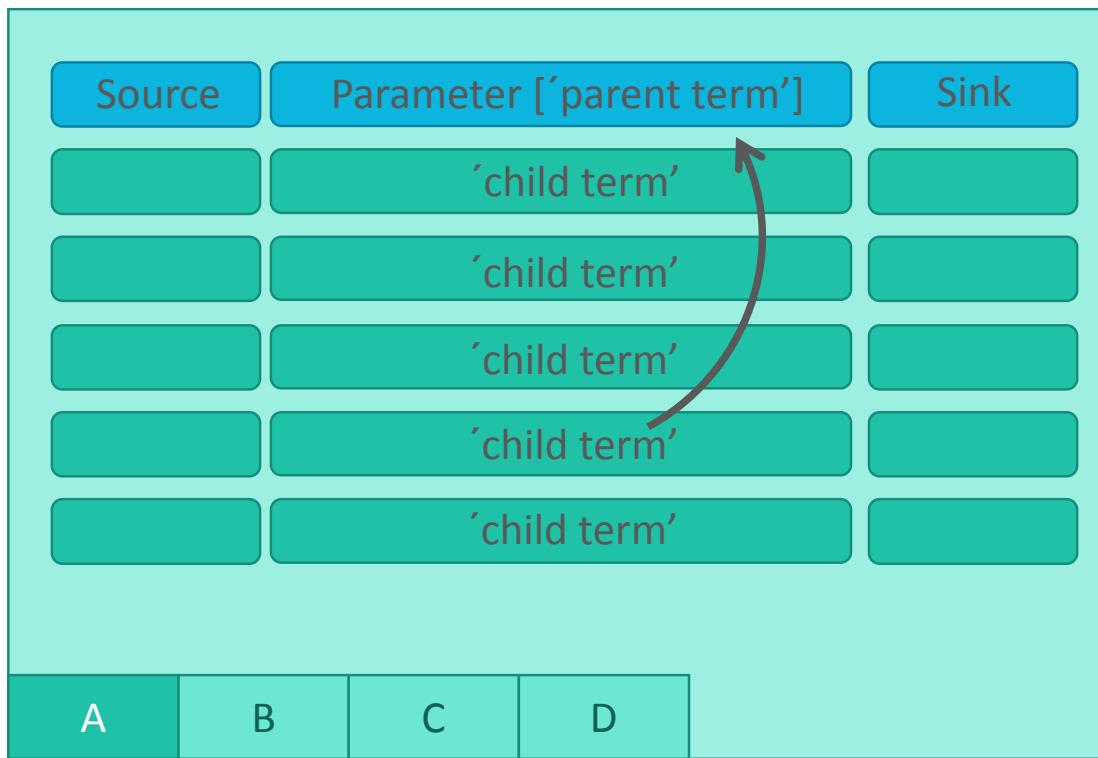
Hierarchical combination of ontologies



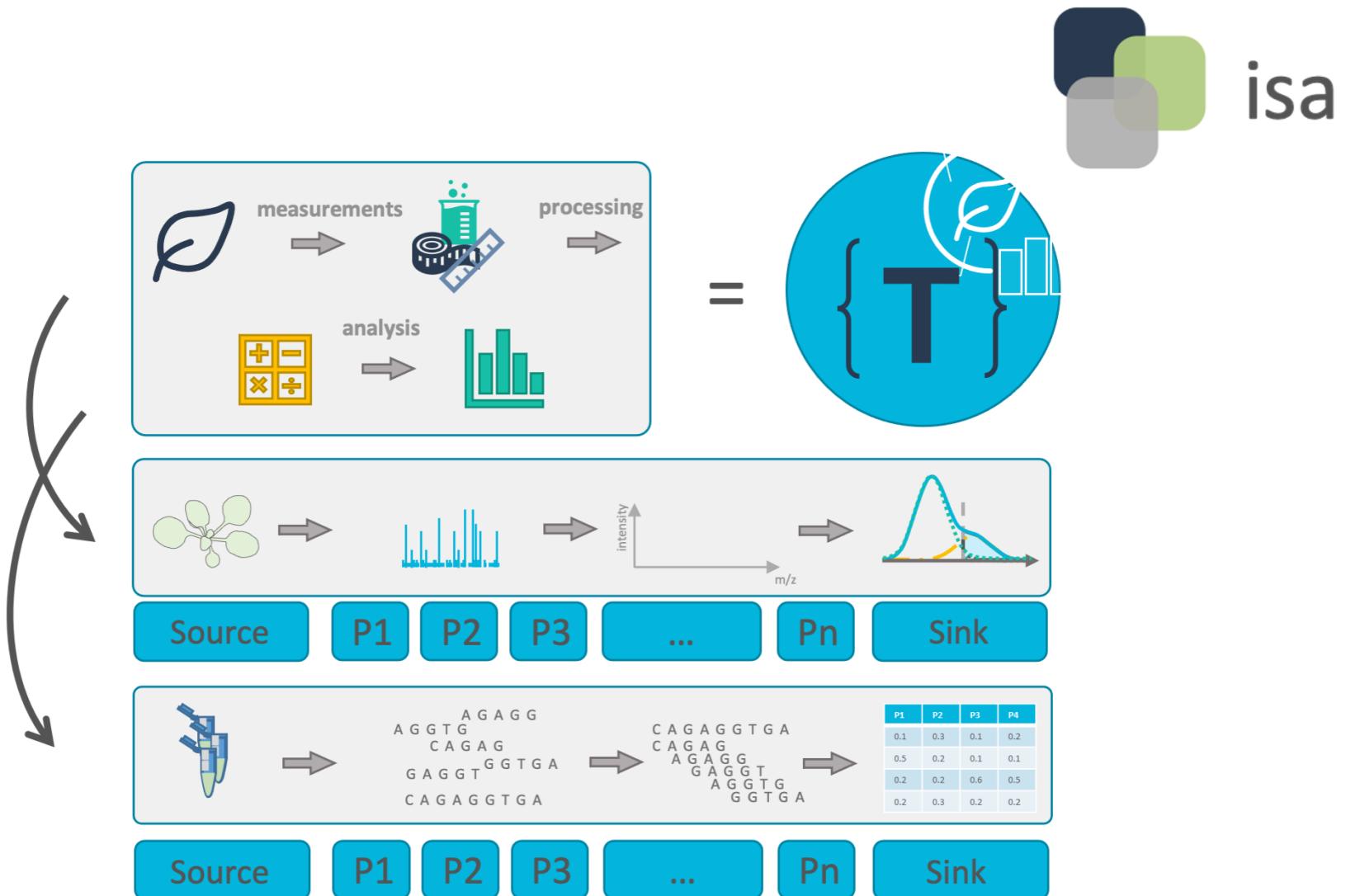
[isa.study.xlsx or isa.assay.xlsx](#)

Swate templates

Checklists and Templates



Realization of lab-specific metadata templates



Directly import templates via Swate

- DataPLANT curated
- Community templates

The screenshot shows the Swate software interface. At the top, there's a green header bar with the word "Swate" and various icons. Below the header is a toolbar with buttons for creating new templates, searching, and other functions. The main area is titled "Templates → Template Search" and contains a sub-instruction: "Search the database for a protocol template you want to use." It includes two search fields: "Search by protocol name" and "Search for tags". The main content is a table listing protocol templates. The columns are "Protocol Name", "Protocol Version", and "Uses". Each row shows a template name, its status (curated or community), its version, the number of uses, and a dropdown arrow. A vertical teal bar on the left side of the table indicates it's scrollable. At the bottom right of the table, it says "Swate Release Version 0.6.2".

| Protocol Name | Protocol Version | Uses |
|-------------------------------------|------------------|------|
| Plant growth | 1.1.13 | 0 |
| RNA extraction | 1.1.6 | 0 |
| Protein extraction | 1.1.6 | 0 |
| Metabolite Extraction | 1.1.8 | 0 |
| DNA extraction | 1.1.6 | 0 |
| Imaging extraction | 1.0.2 | 0 |
| RNA-Seq Assay | 1.1.7 | 0 |
| Proteomics MassSpec Assay | 1.1.6 | 0 |
| Metabolomics MassSpec Assay | 1.1.8 | 1 |
| Genomics Assay | 1.1.6 | 0 |
| Imaging assay | 1.0.2 | 0 |
| RNA-Seq Computational Analysis | 1.1.7 | 0 |
| Proteomics Computational Analyses | 1.1.6 | 0 |
| Metabolomics Computational Analysis | 1.1.8 | 0 |
| Genome assembly | 1.1.6 | 0 |
| Imaging computation | 1.0.2 | 0 |
| MAdLand Fragmentanalyzer | 1.0.0 | 0 |

Contributors

Slides presented here include contributions by

- name: Dominik Brilhaus
github: <https://github.com/brilator>
orcid: <https://orcid.org/0000-0001-9021-3197>
- name: Martin Kuhl
github: <https://github.com/Martin-Kuhl>
orcid: <https://orcid.org/0000-0002-8493-1077>

Swate hands-on

Goals

- Get familiar with ISA metadata and Swate
- Annotate data in your ARC

Check Swate installation

 Make sure [Swate is installed](#):

1. Open Excel (online or Desktop)
2. Go to the [Insert](#) tab: Click the arrow next to "My Add-ins". There you should be able to select Swate.
3. Go to the [Data](#) tab: you should see the Swate (Core) add-in.

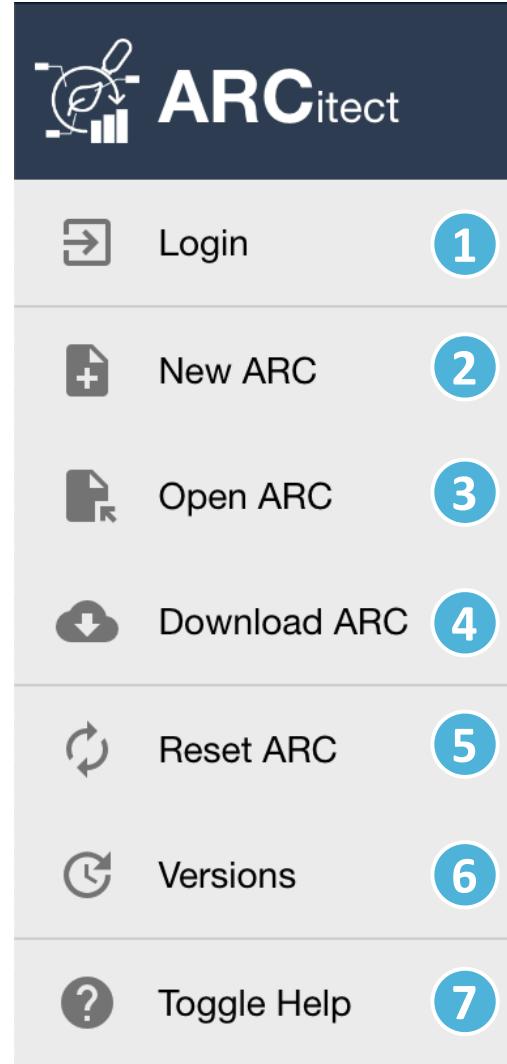
 Alternatively, you can use [Swate standalone](#)

(⚠️ this is however *work in progress* and likely to change)

Have a simple text editor ready

- Windows Notepad
- MacOS TextEdit

Recommended text editor with code highlighting, git support, terminal, etc: [Visual Studio Code](#)



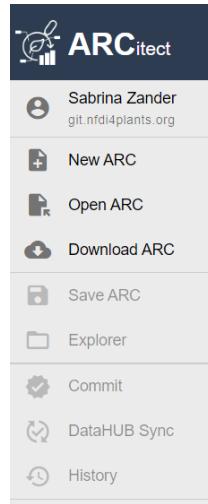
Download the demo data

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

Download the demo data

4. Search for **Talinum-CAM-Photosynthesis**

5. Click the download button, select a location and open the ARC.



Open ARC

A screenshot of the 'Download ARC' interface. The title bar says 'Download ARC' and 'Download ARCs from the nfdi4plants DataHUB'. The search bar contains 'Talinum-CAM-Photosynthesis' with a search icon. Below it, a dropdown shows 'Host git.nfdi4plants.org' and a refresh icon. A list item 'Talinum-CAM-Photosynthesis [2023-10-11T09:24:10.208Z] Teaching' is shown with a green circular icon containing a white letter 'T', a search icon, and a download icon.

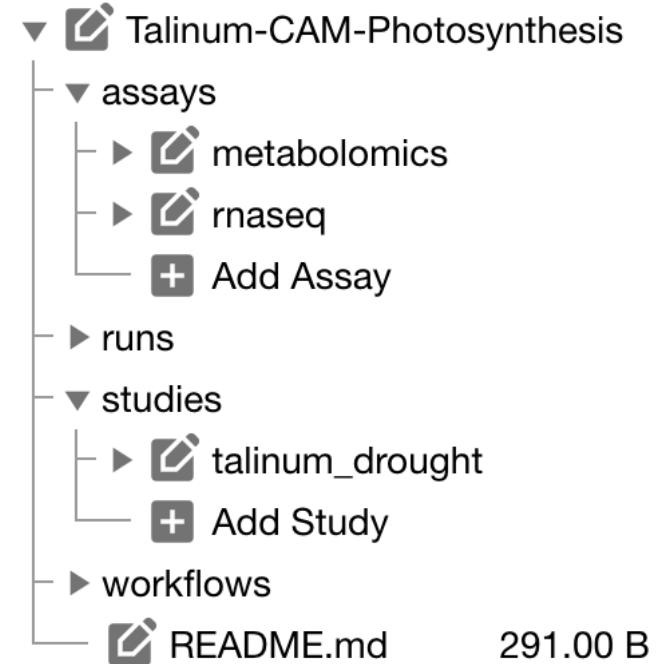
This is basically the ARC we created last session.

Where we left off last time

 Initiated an ARC

 Structured and ...

 Shared with collaborators



Today we want to

 ... annotate the experimental data

Swate hands-on with demo data

Swate Overview

The screenshot shows a Microsoft Excel spreadsheet titled "toolTalk.xlsx" with a single sheet named "Tabelle1". The data consists of 37 rows and 5 columns. The columns are labeled "Source Name", "Characteristics [sample label]", "Factor [temperature unit]", "Data File Name", and "Instrument". The "Source Name" column contains various sample names like "Heat_15A_OD_R1" through "Highlight_180A_OD_R2". The "Characteristics" column shows values such as "15N", "32.00 degree Celsius", and "4.00 degree Celsius". The "Factor" column shows values like "15N", "32.00 degree Celsius", and "4.00 degree Celsius". The "Data File Name" column lists corresponding .wiff files. The "Instrument" column is initially empty but has a tooltip indicating it is a "new parameter".

Annotations from the Swate add-in are overlaid on the spreadsheet:

- A callout points to the "Instrument" column header with the text "new parameter".
- A callout points to the "Instrument" column with the text "otation building block selection".
- A callout points to the "Instrument" column with the text "Add annotation building blocks (columns) to the annotation table."
- A callout points to the "Instrument" column with the text "Parameter instrument".
- A callout points to the "Instrument" column with the text "This Pa".
- A callout points to the "Instrument" column with the text "Use advanced search in".
- A callout points to the "Instrument" column with the text "Add/Update unit re".
- A callout points to the "Instrument" column with the text "Adds a unit to the co assigned, the new unit".
- A callout points to the "Instrument" column with the text "Can't find the Term you are looking for? Use Advanced Search".
- A callout points to the "Instrument" column with the text "You can also request a term by opening an Issue".
- A callout points to the "Instrument" column with the text "More about Parameter: Use parameters to annotate your experimental workflow. You can group parameters to create a protocol. You can find more information on our website".

Let's annotate the plant samples first

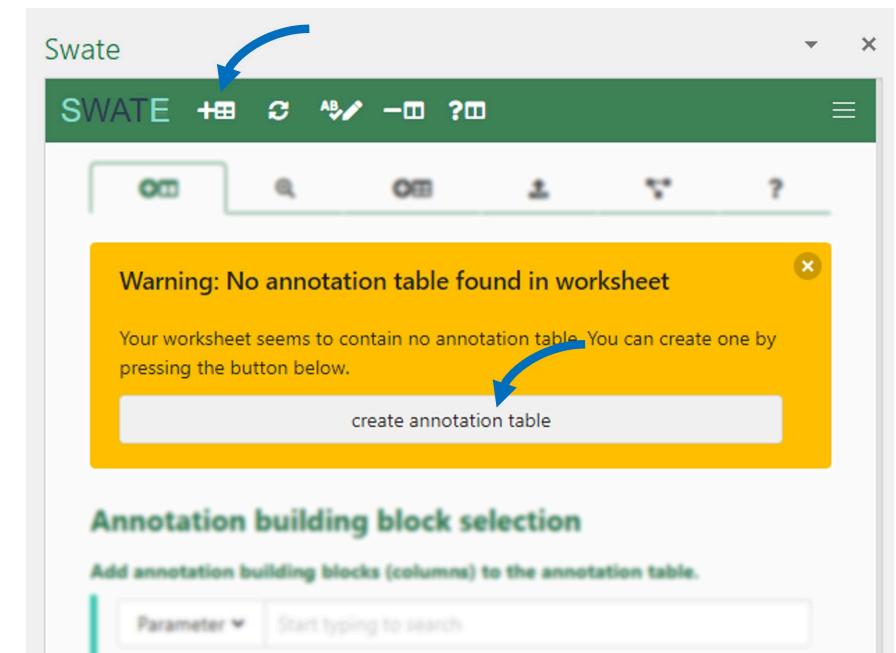
1. Navigate to the demo ARC.
2. Open the lab notes `studies/talinum_drought/protocols/plant_material.txt` in a text editor.
3. Open the empty `studies/talinum_drought/isa.study.xlsx` workbook in Excel.

Create an annotation table

Create a Swate annotation table via the
create annotation table button in the yellow pop-up box
OR click the Create Annotation Table quick access
button.

💡 Each table is by default created with one input (Source Name) and one output (Sample Name) column

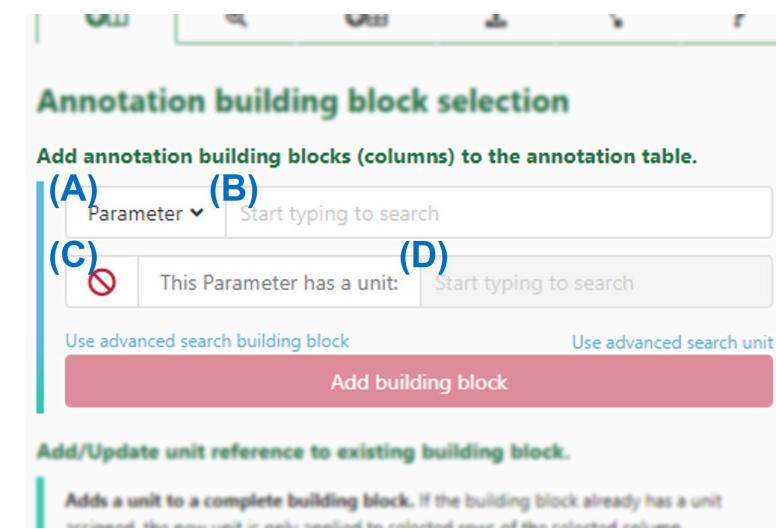
💡 Only one annotation table can be added per Excel sheet



Add a building block

1. Navigate to the *Building Blocks* tab via the navbar. Here you can add *Building Blocks* to the table.
2. Instead of *Parameter* select *Characteristic* from the drop-down menu (A)
3. Search for **organism** in the search bar (B). This search looks for suitable *Terms* in our *Ontology* database.
4. Select the Term with the id **OBI:0100026** and,
5. Click **Add building block**.

 This adds three columns to your table, one visible and two hidden.



Annotation building block selection

Add annotation building blocks (columns) to the annotation table.

(A) Parameter (B) Start typing to search

(C)  This Parameter has a unit: (D) Start typing to search

Use advanced search building block Use advanced search unit

Add building block

Add/Update unit reference to existing building block.

Adds a unit to a complete building block. If the building block already has a unit

Insert values to annotate your data

1. Navigate to the *Terms* tab in the Navbar
2. In the annotation table, select any number of cells below **Characteristic**
[organism]

3. Click into the search field in Swate.

|  You should see **organism** showing in a field in front of the search field
 The search will now yield results related to **organism**

4. In the search field, search for "Talinum fruticosum"

5. Select the first hit and click **Fill selected cells with this term**

Add a building block with a unit

1. In the *Building Blocks* tab, select *Parameter*, search for `light intensity exposure` and select the term with id `PEC0:0007224`.
2. Check the box for *This Parameter has a unit* and search for `microeinsteins per square meter per second` in the adjacent search bar.
3. Select `U0:0000160`.
4. Click `Add building block`.



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

Insert unit-values to annotate your data

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

 You can see the numbers being complemented with the chosen unit, e.g. 425.00 microeinsteins per square meter per second

Showing ontology reference columns

Hold **Ctrl** and click the *Autoformat Table* quick access button to adjust column widths and un-hide all hidden columns.

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

 This feature is currently not supported on MacOS

Update ontology reference columns

Click the **Update Ontology Terms** quick access buttons.

 This updates all reference columns according to the main column. In this case the reference columns for **Parameter [light intensity exposure]** are updated with the id and source ontology of the **microeinsteин per square meter per second** unit.

Your ISA table is growing

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

| ISA Table Overview | | | | |
|--------------------|---------------------|---------------------------|--|----------------------|
| | Input [Source Name] | Characteristic [organism] | Parameter [light intensity exposure] | Output [Sample Name] |
| 1 | | Talinum fruticosum | ✓ 425 microeinsteins per square meter per second | |
| 2 | | Talinum fruticosum | ✓ 425 microeinsteins per square meter per second | |
| 3 | | Talinum fruticosum | ✓ 425 microeinsteins per square meter per second | |
| 4 | | Talinum fruticosum | ✓ 425 microeinsteins per square meter per second | |
| 5 | | Talinum fruticosum | ✓ 425 microeinsteins per square meter per second | |
| 6 | | Talinum fruticosum | ✓ 425 microeinsteins per square meter per second | |

1 +

Hiding ontology reference columns

Click the  quick access button without holding  to hide all reference columns.

Exercise



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

Let's annotate the RNA Seq data

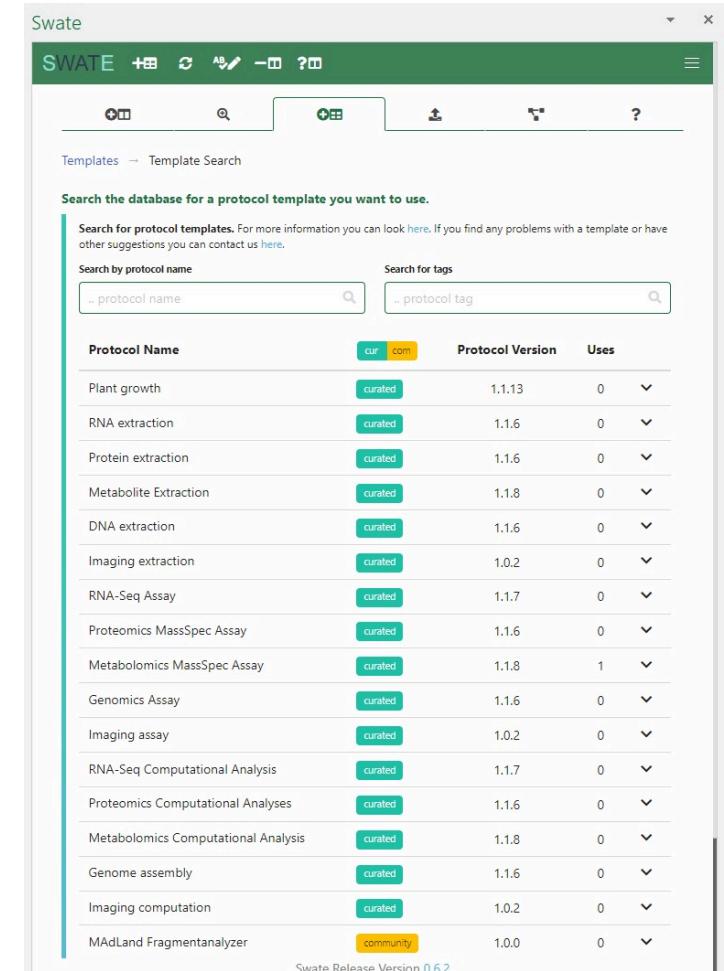
1. Navigate to the demo ARC.
2. Open the lab notes `assays/rnaseq/protocols/RNA_extraction.txt` in a text editor.
3. Open the empty `assays/rnaseq/isa.assay.xlsx` workbook in Excel.

Use a template

1. Navigate to *Templates* in the Navbar and click *Browse database* in the first function block.

 Here you can find community created workflow annotation templates

1. Search for **RNA extraction** and click **select**
 - You will see a preview of all building blocks which are part of this template.
2. Click **Add template** to add all Building Blocks from the template to your table, which do not exist yet.



The screenshot shows the 'Template Search' page in the Swate software. The top navigation bar includes icons for file operations (New, Open, Save, Print, etc.) and help. Below the bar, there are search fields for 'Search by protocol name' and 'Search for tags'. A main table lists various protocol templates:

| Protocol Name | cur | com | Protocol Version | Uses |
|-------------------------------------|-----------|-----|------------------|------|
| Plant growth | curated | | 1.1.13 | 0 |
| RNA extraction | curated | | 1.1.6 | 0 |
| Protein extraction | curated | | 1.1.6 | 0 |
| Metabolite Extraction | curated | | 1.1.8 | 0 |
| DNA extraction | curated | | 1.1.6 | 0 |
| Imaging extraction | curated | | 1.0.2 | 0 |
| RNA-Seq Assay | curated | | 1.1.7 | 0 |
| Proteomics MassSpec Assay | curated | | 1.1.6 | 0 |
| Metabolomics MassSpec Assay | curated | | 1.1.8 | 1 |
| Genomics Assay | curated | | 1.1.6 | 0 |
| Imaging assay | curated | | 1.0.2 | 0 |
| RNA-Seq Computational Analysis | curated | | 1.1.7 | 0 |
| Proteomics Computational Analyses | curated | | 1.1.6 | 0 |
| Metabolomics Computational Analysis | curated | | 1.1.8 | 0 |
| Genome assembly | curated | | 1.1.6 | 0 |
| Imaging computation | curated | | 1.0.2 | 0 |
| MADLand Fragmentanalyzer | community | | 1.0.0 | 0 |

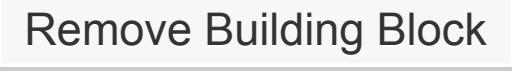
Swate Release Version 0.6.2

Adding / Updating unit references

Sometimes you need to add or update the unit of an existing building block.

1. Select any number of rows of the **Parameter [biosource amount]** building block to mark it for the next steps.
2. Open the *Building Blocks* tab
3. In the bottom panel "Add/Update unit reference to existing building block", search for the unit "milligram". Select the unit term and click **Update unit for cells**.
 If you already had values in the main column they will be updated automatically.
4. Click the *Update Ontology Terms* quick access button, to update the reference columns.

Remove building blocks

If there are any Building Blocks which do not fit your experiment you can use the  quick access button to remove it including all related (hidden) reference columns.

 Due to the hidden reference columns, we recommend not to delete table columns via usual Excel functions.

New process, new worksheet

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

Exercise



Try to fill the two sheets with the protocol details:

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

Your ISA table is ready 

Go ahead, adjust the Building Blocks you want to use to describe your experiment as you see fit.

Insert values using Swate Term search and add input and output.

A small detour on "Excel Tables"

Swate uses Excel's "table" feature to annotate workflows. Each table represents one *process* from input (e.g. plant leaf material) to output (e.g. leaf extract).

Example workflows with three *processes* each:

- Plant growth → sampling → extraction
- Measured data files → statistical analysis → result files

 Excel tables allow to group data that belongs together inside one sheet. This is not to be confused with a (work)sheet or workbook.

```
workbook          (e.g. "isa.assay.xlsx")
  └── worksheet    (e.g. "plant_growth")
    └── table      (e.g. "annotationTable")
```

Known issues with ARCitect and Swate (Nov 2023)

1. Annotation within ARCitect is not yet available.
2. Swate and ARCitect handle isa.study.xlsx / isa.assay.xlsx files differently.

Contributors

Slides presented here include contributions by

- name: Dominik Brilhaus
github: <https://github.com/brilator>
orcid: <https://orcid.org/0000-0001-9021-3197>
- name: Kevin Frey
github: <https://github.com/Freymaurer>
orcid: <https://orcid.org/0000-0002-8493-1077>
- name: Martin Kuhl
github: <https://github.com/Martin-Kuhl>
orcid: <https://orcid.org/0000-0002-8493-1077>
- name: Sabrina Zander



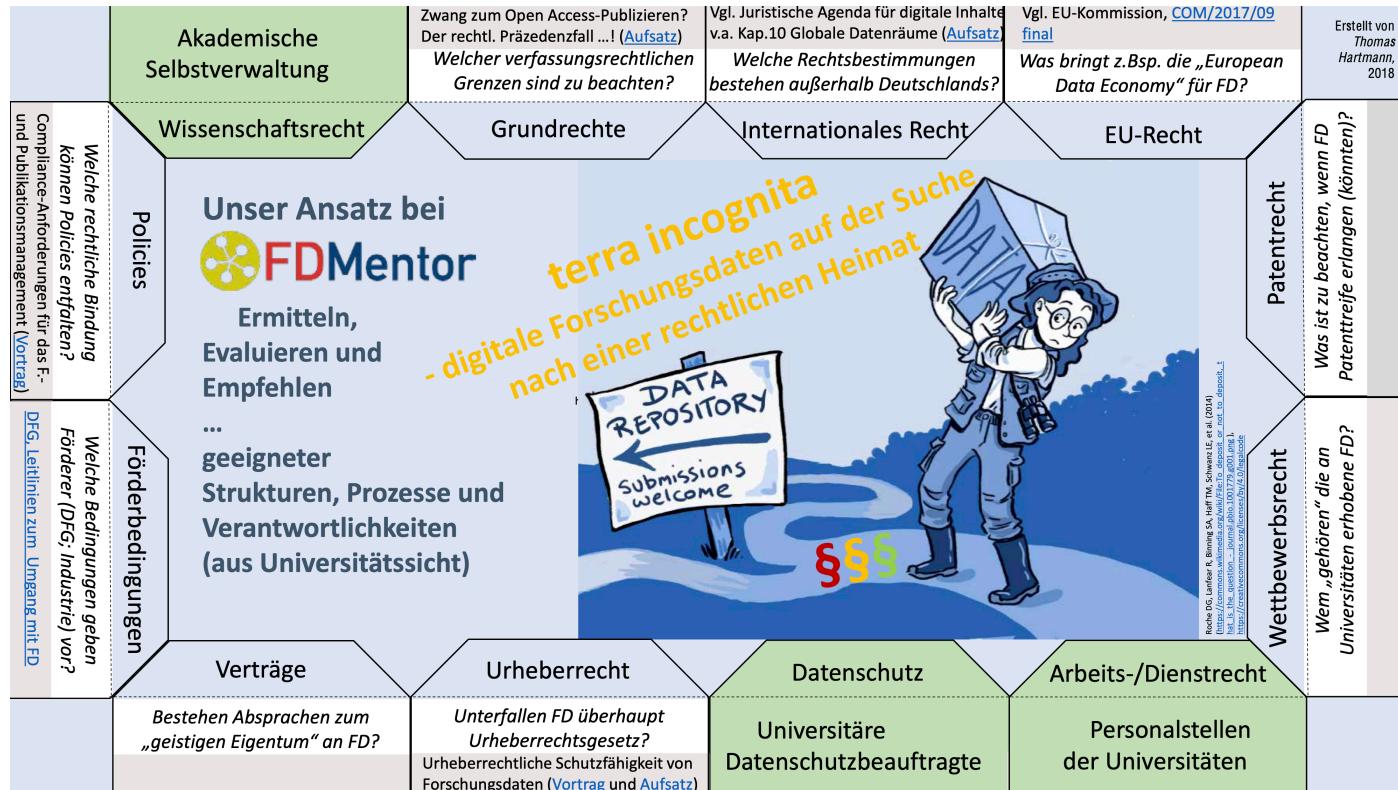
Annotate your ARC with metadata



1. Go to your ARC
2. Try to annotate studies and assays with proper metadata using Swate

Legal aspects of RDM

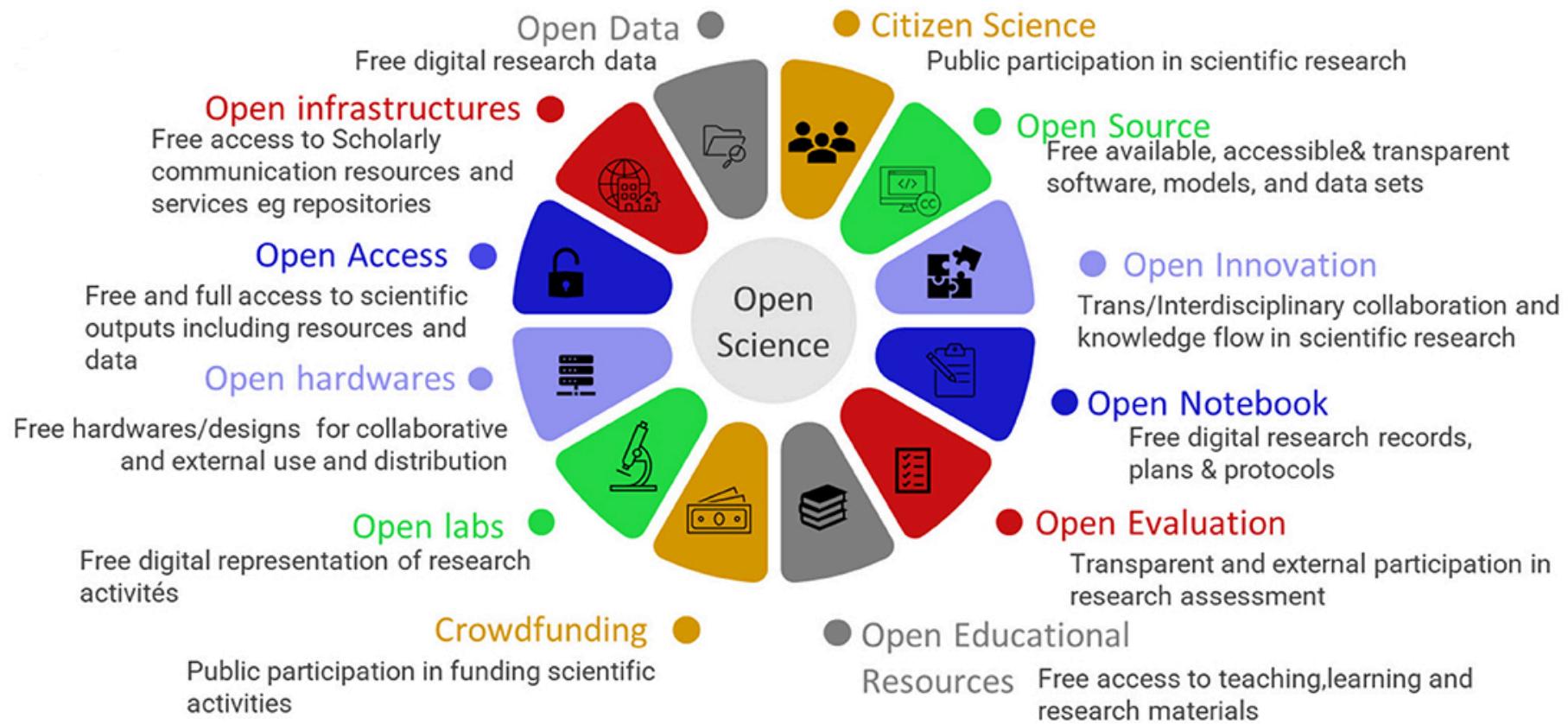
Different laws touched by RDM



Open Access (OA) categories

- Gold: Published in an open-access journal that is indexed by the DOAJ.
- Green: Toll-access on the publisher page, but there is a free copy in an OA repository.
- Hybrid: Free under an open license in a toll-access journal.
- Bronze: Free to read on the publisher page, but without a clearly identifiable license.
- Closed: All other articles, including those shared only on an Academic Social Network or in Sci-Hub.

Open Science is more than Open Access



Creative commons

Check out: <https://creativecommons.org/about/cclicenses/>



Attribution
CC BY



Attribution – ShareAlike
CC BY-SA



Attribution – NoDerivs
CC BY-ND



Attribution – NonCommercial
CC BY-NC



Attribution – NonCommercial – ShareAlike
CC BY-NC-SA



Attribution – NonCommercial – NoDerivs
CC BY-NC-ND

Data protection

GDPR: General Data Protection Regulation

DS-GVO (german): Datenschutz-Grundverordnung

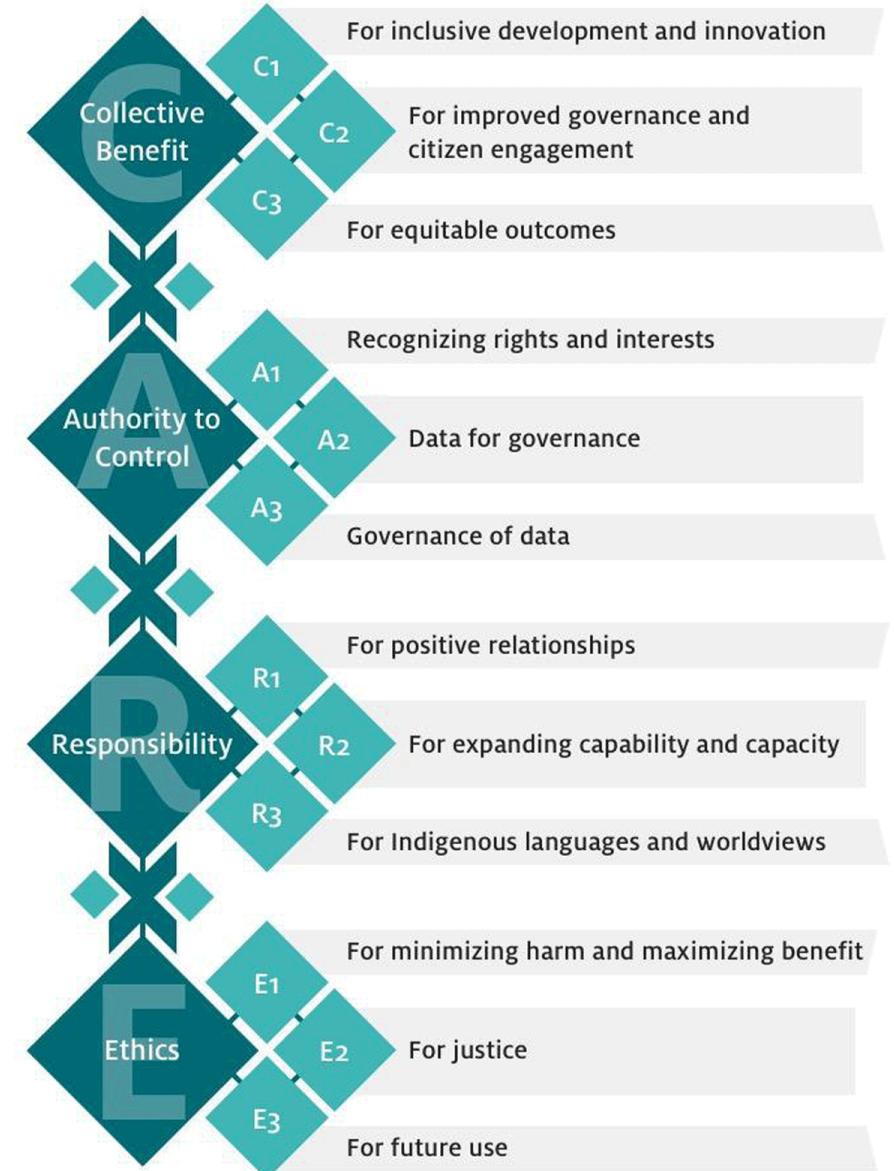
Use of biological materials

- MTA: material transfer agreement
- Nagoya Protocol: <https://www.cbd.int/abs/about/>
- DSIs: Digital sequence information

FAIR and CARE



CARE principles



Research Data policies

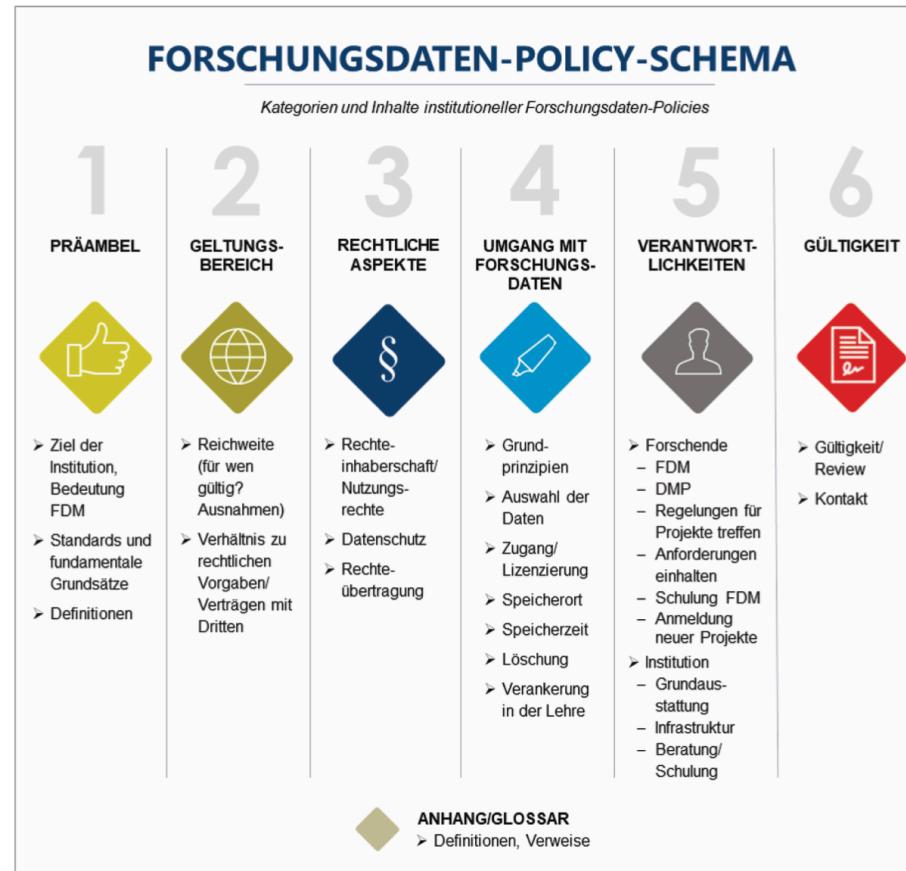


Abbildung 2: FD-Policy-Schema: Die sechs Kategorien einer FD-Policy mit ihren inhaltlichen Bestandteilen

CEPLAS relevant data handling guidelines & policies

- Deutsche Forschungsgemeinschaft (2015): DFG Guidelines on the Handling of Research Data
- Amtliche Mitteilungen der Universität zu Köln AM 07/2018: Leitlinie zum Umgang mit Forschungsdaten
- Amtliche Bekanntmachung der Heinrich-Heine-Universität Nr. 43/2022: Forschungsdaten-Richtlinie
- Leitlinie zum Umgang mit Forschungsdaten im Forschungszentrum Jülich 05/2019
- Senat der Max-Planck-Gesellschaft (2009): Regeln zur Sicherung guter wissenschaftlicher Praxis

The Data Management Plan (DMP)

- Covers the full research data lifecycle
- Frequently updated as your project develops
- Required to different extents by funding agencies (e.g. DFG, Horizon Europe, BMBF, BMEL, ...)

DMP tools

- Data Stewardship Wizard <https://ds-wizard.org/>
- RDMO <https://rdmorganiser.github.io/> (e.g. <https://rdmo.hhu.de>)
- Dataplan: <https://dmpg.nfdi4plants.org>

Check out the [Elixir RDMkit](#) for more

Contributors

Slides presented here include contributions by

- name: Dominik Brilhaus
github: <https://github.com/brilator>
orcid: <https://orcid.org/0000-0001-9021-3197>
- name: Hajira Jabeen
github: <https://github.com/HajiraJabeen>
orcid: <https://orcid.org/0000-0003-1476-2121>



Share your ARC



1. Follow the next slides to learn how to share your ARC with a colleague
2. Go to your colleague's ARC and try to understand it



Understand your colleague's ARC



1. Go to your colleague's ARC
2. Try to identify one experiment that led to results (e.g. a figure in the thesis)
3. What are the samples (e.g. plants, bacteria)?
4. How were the samples prepared (~ materials)?
5. How was the experiment performed (~ methods)?
6. What is the raw data (~ results)?
7. How was the data analyzed (~ computational methods, statistics)?
8. Collect all of the above in a `README_<YourArbitraryParticipantID>.md` in the same folder.

Assignment

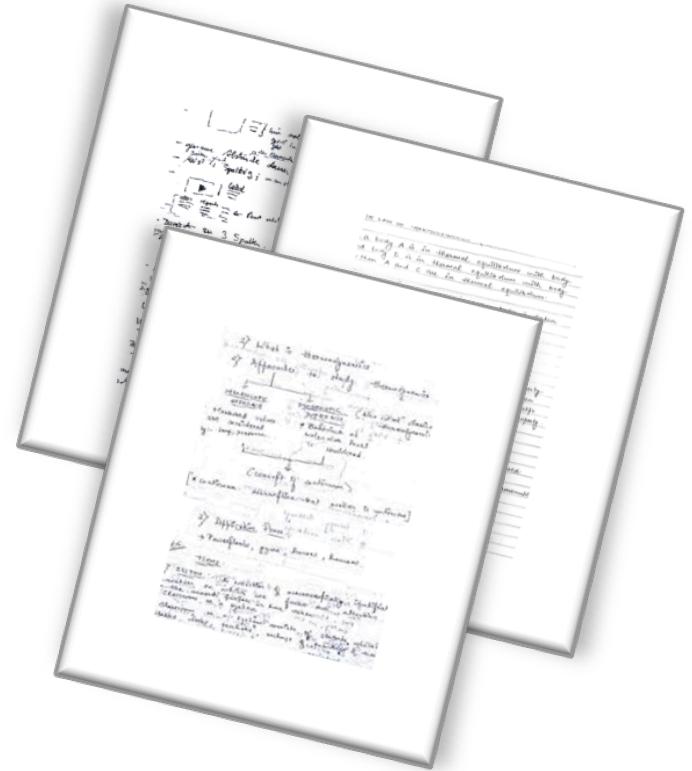
| Participant | looks at ARC of |
|---------------|-----------------|
| Participant02 | Participant01 |
| Participant03 | Participant02 |
| ... | ... |
| Participant n | Participant n-1 |

Electronic Lab Notebooks (ELNs)

ELN ≈ Digital Protocol Editors

- Documenting daily lab routine
- Lab methods & protocols
- Lab inventory (biologicals, chemicals, instruments)
- Local sharing & collaboration
- Backup (locally)

 ELNs help to digitalize research documentation



Are ELNs FAIR?

| FAIR indicator* | ELN |
|---|-----|
| Findable | |
| F1. (Meta)data are assigned a globally unique and persistent identifier. | |
| F2. Data are described with rich metadata (defined by R1 below). | |
| F3. Metadata clearly and explicitly include the identifier of the data they describe. | |
| F4. (Meta)data are registered or indexed in a searchable resource. | |
| Accessible | |
| A1. (Meta)data are retrievable by their identifier using a standardised communications protocol | |
| A1.1 The protocol is open, free, and universally implementable | |

*FAIR classified by <https://www.go-fair.org/fair-principles/>

Findable

| FAIR indicator* | elabFTW |
|---|---|
| F1. (Meta)data are assigned a globally unique and persistent identifier. |  /  |
| F2. Data are described with rich metadata (defined by R1 below). |  |
| F3. Metadata clearly and explicitly include the identifier of the data they describe. |  /  |
| F4. (Meta)data are registered or indexed in a searchable resource. |  |

Accessible

| FAIR indicator* | elabFTW |
|---|--|
| A1. (Meta)data are retrievable by their identifier using a standardised communications protocol | ● |
| A1.1 The protocol is open, free, and universally implementable | ● / ● |
| A1.2 The protocol allows for an authentication and authorisation procedure, where necessary | ? |
| A2. Metadata are accessible, even when the data are no longer available | ? |

Interoperable

| FAIR indicator* | elabFTW |
|--|---------|
| I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. | ● |
| I2. (Meta)data use vocabularies that follow FAIR principles. | ● |
| I3. (Meta)data include qualified references to other (meta)data. | ● |

Reusable

| FAIR indicator* | elabFTW |
|--|---------|
| R1. (Meta)data are richly described with a plurality of accurate and relevant attributes | ● |
| R1.1. (Meta)data are released with a clear and accessible data usage license | ● |
| R1.2. (Meta)data are associated with detailed provenance | ● |
| R1.3. (Meta)data meet domain-relevant community standards | ● |

Contributors

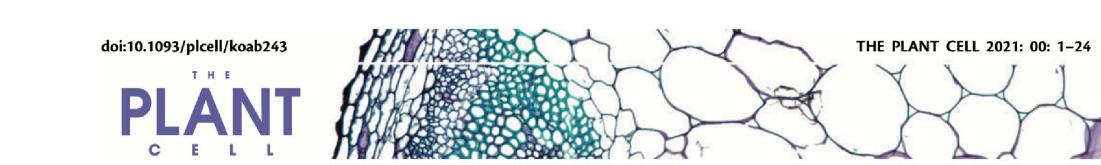
Slides presented here include contributions by

- name: Dominik Brilhaus
github: <https://github.com/brilator>
orcid: <https://orcid.org/0000-0001-9021-3197>

Data publication and repositories

Persistent Identifiers (PIIDs)

Spot the PIDs



Research Article

Interactions between SQUAMOSA and SHORT VEGETATIVE PHASE MADS-box proteins regulate meristem transitions during wheat spike development

Kun Li ^{1,2,†} Juan M. Debernardi ^{1,2,*†,‡} Chengxia Li ^{1,2} Huiqiong Lin ^{1,2} Chaozhong Zhang ¹,
Judy Jernstedt ¹, Maria von Korff ^{3,4} Jinshun Zhong ³ and Jorge Dubcovsky ^{1,2,*†}

1 Department of Plant Sciences, University of California, Davis, California 95616, USA

2 Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA

3 Institute for Plant Genetics, Heinrich Heine University, Düsseldorf 40225, Germany

4 Cluster of Excellence on Plant Sciences "SMART Plants for Tomorrow's Needs", Heinrich Heine University, Düsseldorf 40225, Germany

*Author for correspondence: jmdebernardi@ucdavis.edu (J.M.D), jdubcovsky@ucdavis.edu (J.D.)

†These authors contributed equally (K.L and J.M.D.)

‡Senior authors

C.L., J.M.D., and J.D. designed the research. K.L. performed most of the experimental work. J.M.D., C.L., H.L., and C.Z. performed research. J.J. contributed the SEM images. M.V.K. and J.Z. contributed *in situ* hybridizations. C.L., H.L., J.M.D., K.L., and J.D. analyzed the data. C.L., J.M.D., K.L., H.L., and J.D. wrote the article.

The authors responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (<https://academic.oup.com/plcell>) are: Jorge Dubcovsky (jdubcovsky@ucdavis.edu) and Juan Manuel Debernardi (jmdebernardi@ucdavis.edu).

Abstract

Inflorescence architecture is an important determinant of crop productivity. The number of spikelets produced by the wheat inflorescence meristem (IM) before its transition to a terminal spikelet (TS) influences the maximum number of grains per spike. Wheat MADS-box genes VERNALIZATION 1 (VRN1) and FRUITFULL 2 (FUL2) (in the SQUAMOSA-clade) are essential to promote the transition from IM to TS and for spikelet development. Here we show that SQUAMOSA genes contribute to

Downloaded from <https://academic.oup.com/plcell/advance-article/doi/10.1093/plcell/koab243/6415951>

Globally unique, stable, persistent identifiers (PIDs)

- Long-term findability
- Make data, digital objects, people,
... uniquely identifiable
- Diminish “dead links”
- Cope with name changes



Open
Researcher and Contributor ID
<https://orcid.org/>



Digital
Object Identifier
<https://www.doi.org>



Research
Resource
Identifiers
<https://www.rrids.org>



ePIC consortium
<https://www.pidconsortium.net>



Research
Organization Registry
<https://ror.org>



Global
Research Identifier Database
<https://grid.ac>

Properties of a PID

Ideally, PIDs are

- Stable and permanent
- Location-independent
- Globally unique and valid
- Addressable (citable)
- Clickable (resolvable)

Additional resources

- <https://www.doi.org>
- <https://www.orcid.org>
- <https://pidservices.org/>
- <https://datacite.org>
- <https://www.project-freya.eu/en>

Institutional publication guidelines

HHU Düsseldorf recommends use of ORCID and other PIDs

Publikationsrichtlinie der Heinrich-Heine-Universität Düsseldorf vom 09.11.2023:

<https://www.hhu.de/die-hhu/kontakt-und-services/zentrale-und-amtliche-bekanntmachungen/nr-34-2023>

Domain-specific data repositories

Good

- Assign PIDs / DOIs
- Long-term accessible
- Data type specific
- Apply metadata standards
- Usually recommended / required by journals
- Mostly accepted by the community

Intermediate

- User-friendliness
- Different metadata schema
- Complex and versatile submission routines

Domain-specific data repositories

| Repository | Description | Biological data domain |
|-----------------------|---|----------------------------------|
| EBI-ENA | European Nucleotide Archive | genome / transcriptome sequences |
| EBI-ArrayExpress | Archive of Functional Genomics Data | transcriptome |
| EBI-MetaboLights | Database of Metabolomics | metabolome |
| EBI-PRIDE | PRoteomics IDEntifications Database | proteome |
| EBI-Biolimage Archive | Stores and distributes biological images | imaging, microscopy |
| e!DAL-PGP | Plant Genomics & Phenomics Research Data Repository | phenome |

Choosing a data repository

Domain-specific >> Generic >> Institutional

Find repositories at:

- <https://www.re3data.org>
- <https://fairsharing.org>

Generic data repositories

Good

- Allow publication of any kind of data Assign PIDs / DOIs
- Long-term accessible
- Very simple to use



<https://zenodo.org>



<https://datadryad.org/>

Intermediate

- Only generic / high-level metadata schema
- Limited reusability



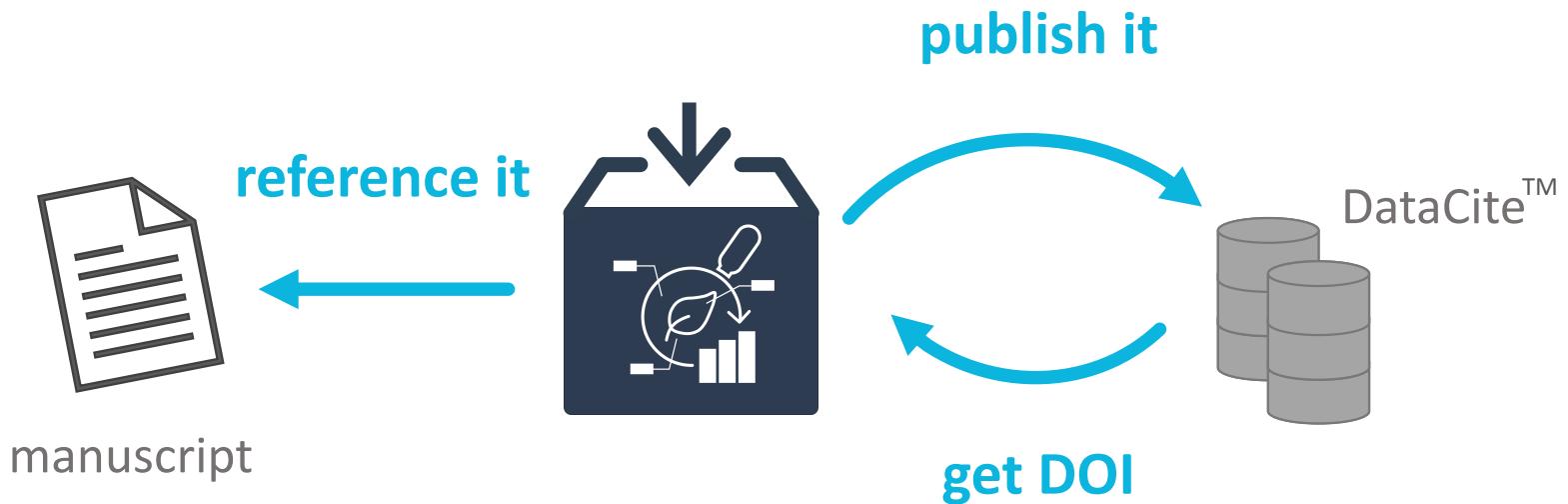
<https://figshare.com>

From ARC to repositories



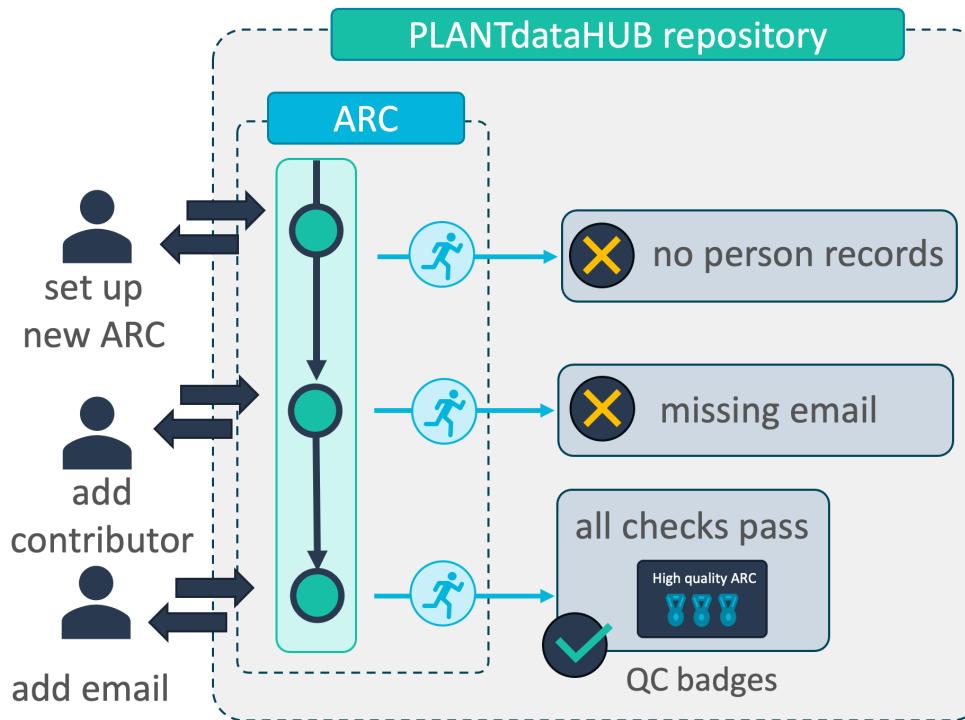
Moving from paper to data publications

Publish your ARC, get a DOI



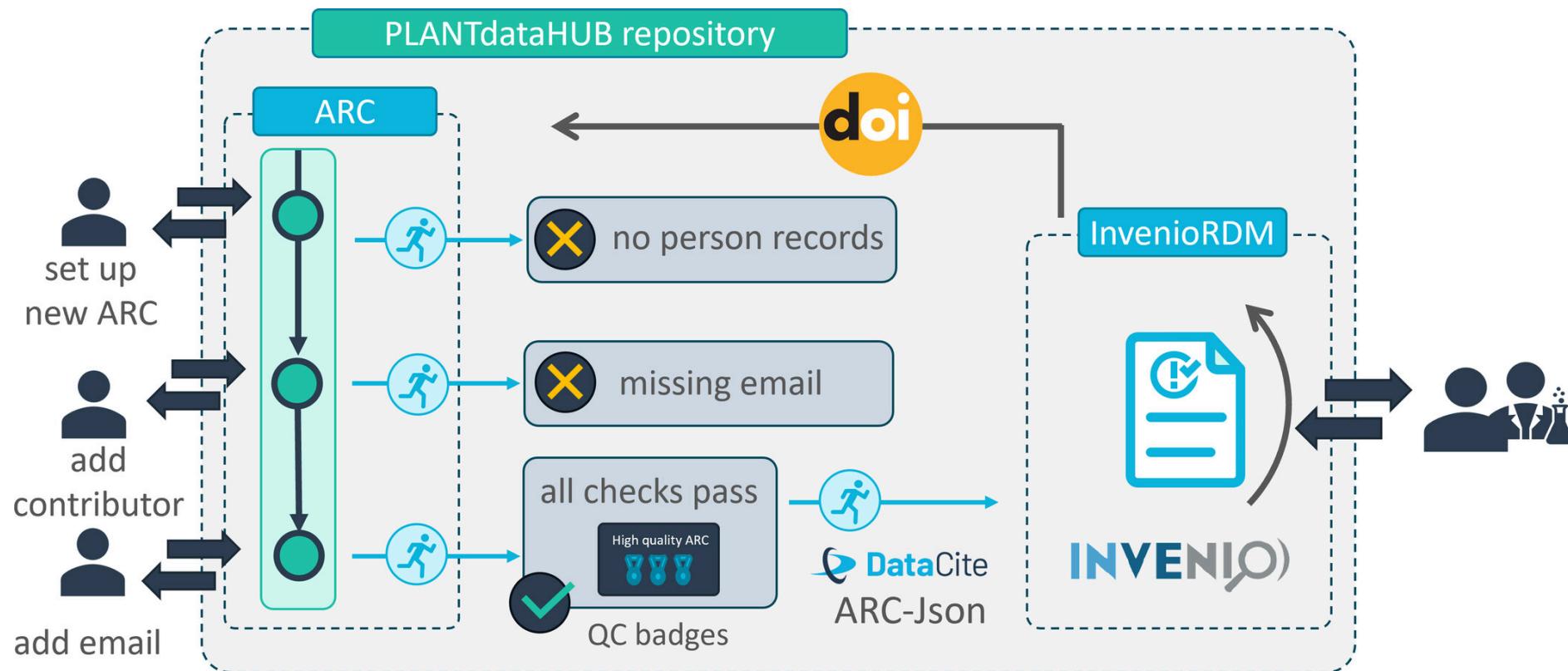
Publish your ARC with a few clicks

Validation



adapted from 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>

Validate & publish



Receive a DOI

Published September 7, 2023 | Version v1

The screenshot shows a dataset page with the following interface elements:

- Top right:** Dataset, Open buttons.
- Header:** Edit, New version, Share buttons.
- Versions:** Version v1 (10.60534/9e5jx-75d83), Sep 7, 2023.
- Details:** DOI (10.60534/9e5jx-75d83), Resource type (Dataset), Publisher (DataPLANT).
- Export:** JSON, Export button.

1

Citation

Style

APA

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

Description

hosted on: <https://git.nfdi4plants.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

Contributors

Slides presented here include contributions by

- name: Dominik Brilhaus
github: <https://github.com/brilator>
orcid: <https://orcid.org/0000-0001-9021-3197>

Code Notebooks and IDEs

- Visual Studio Code:
<https://code.visualstudio.com/>
- RMarkdown:
<http://rmarkdown.rstudio.com>
- Jupyter Notebooks: <https://jupyter.org/>
- ● Interactive (good start for non-coders)
- ● Document and comment code
- ● Often offer integrated version control (e.g. git plugin)
- ● Executable code + "result preview"
- ● Reproducibility (package / library dependencies)

Computational workflow languages

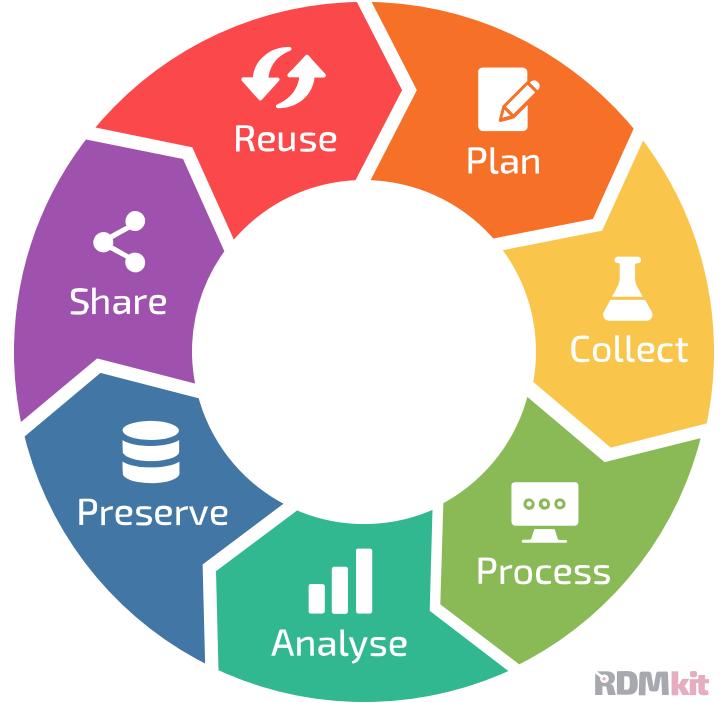
- Describe analysis workflows and tools <https://www.commonwl.org>
- Make them portable and scalable <https://www.nextflow.io>
- Across a variety of environments
(software and hardware) <https://nf-co.re/>
<https://snakemake.github.io>

<https://galaxyproject.eu/>

Contributors

Slides presented here include contributions by

- name: Dominik Brilhaus
github: <https://github.com/brilator>
orcid: <https://orcid.org/0000-0001-9021-3197>



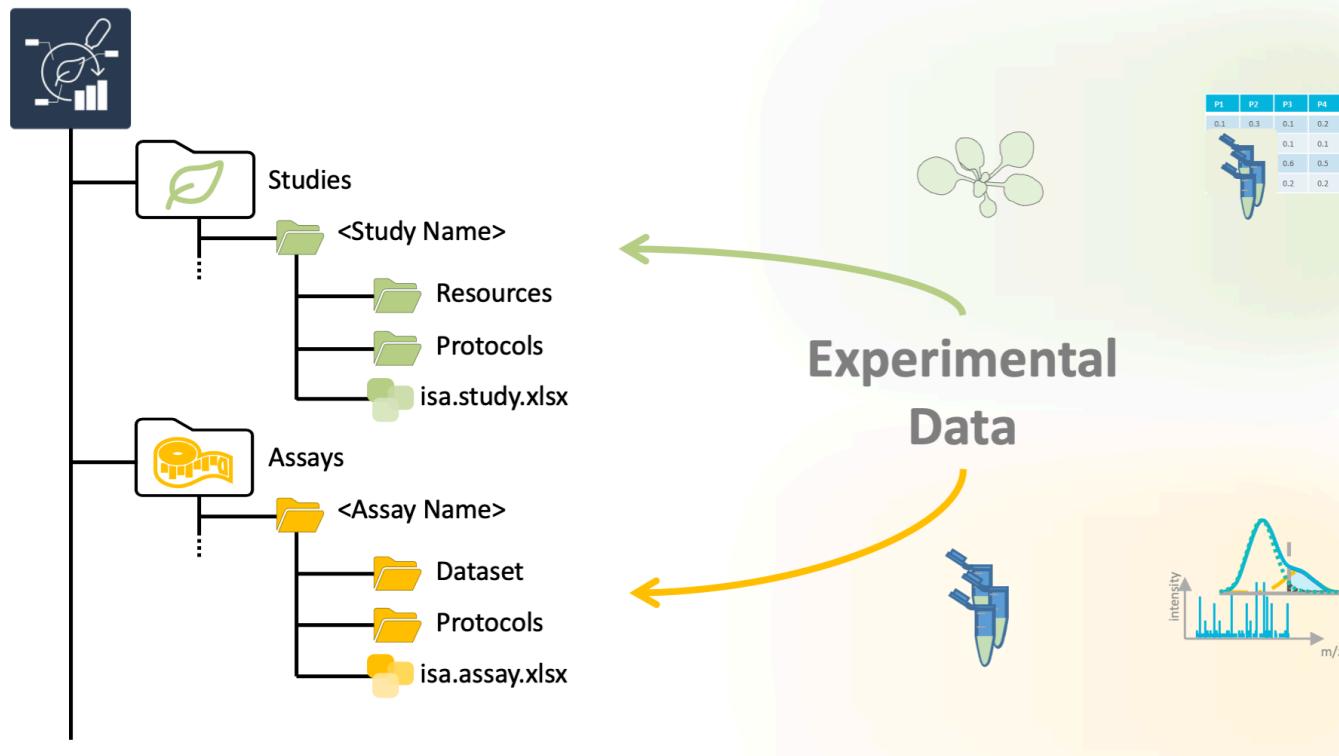
ARC Ecosystem Demo

"A FAIR RDM journey along a (mutable) data life cycle"

Dominik Brilhaus

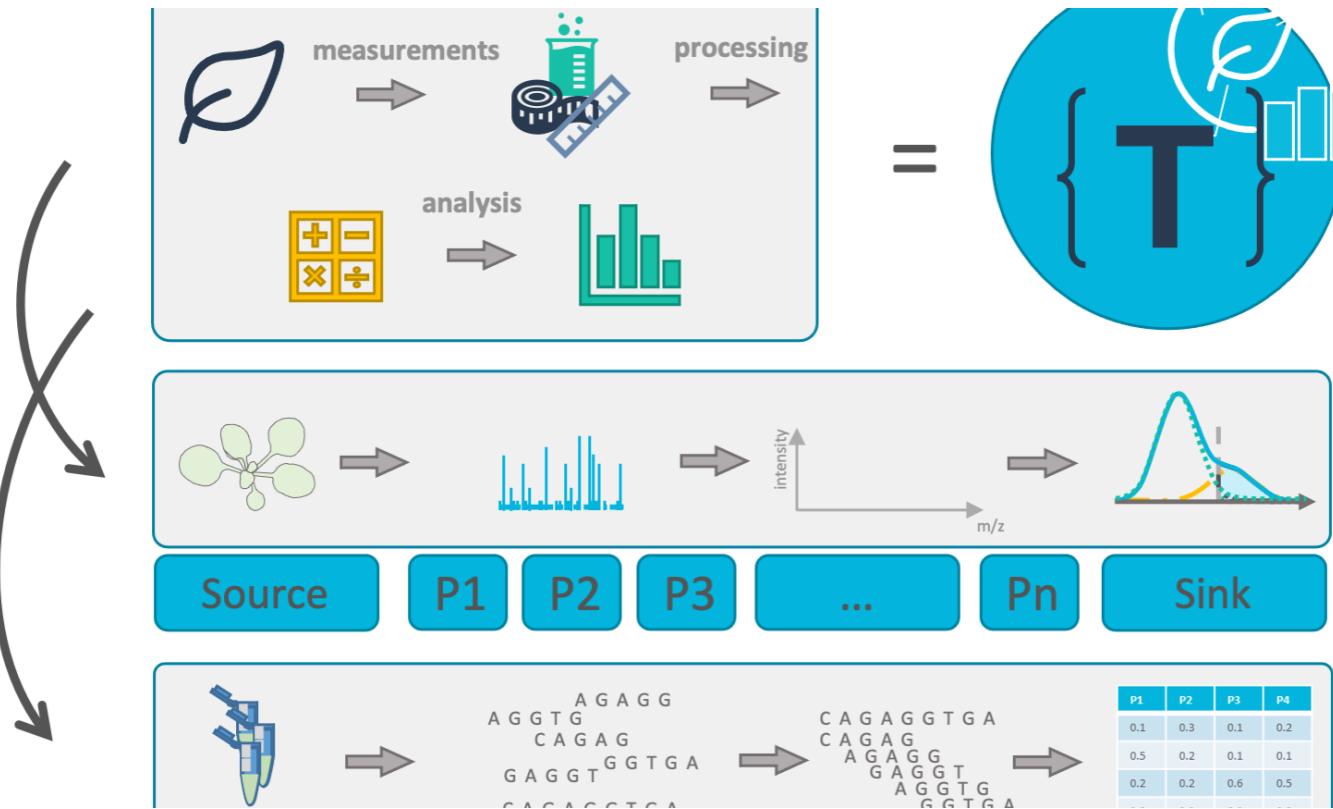


Collect



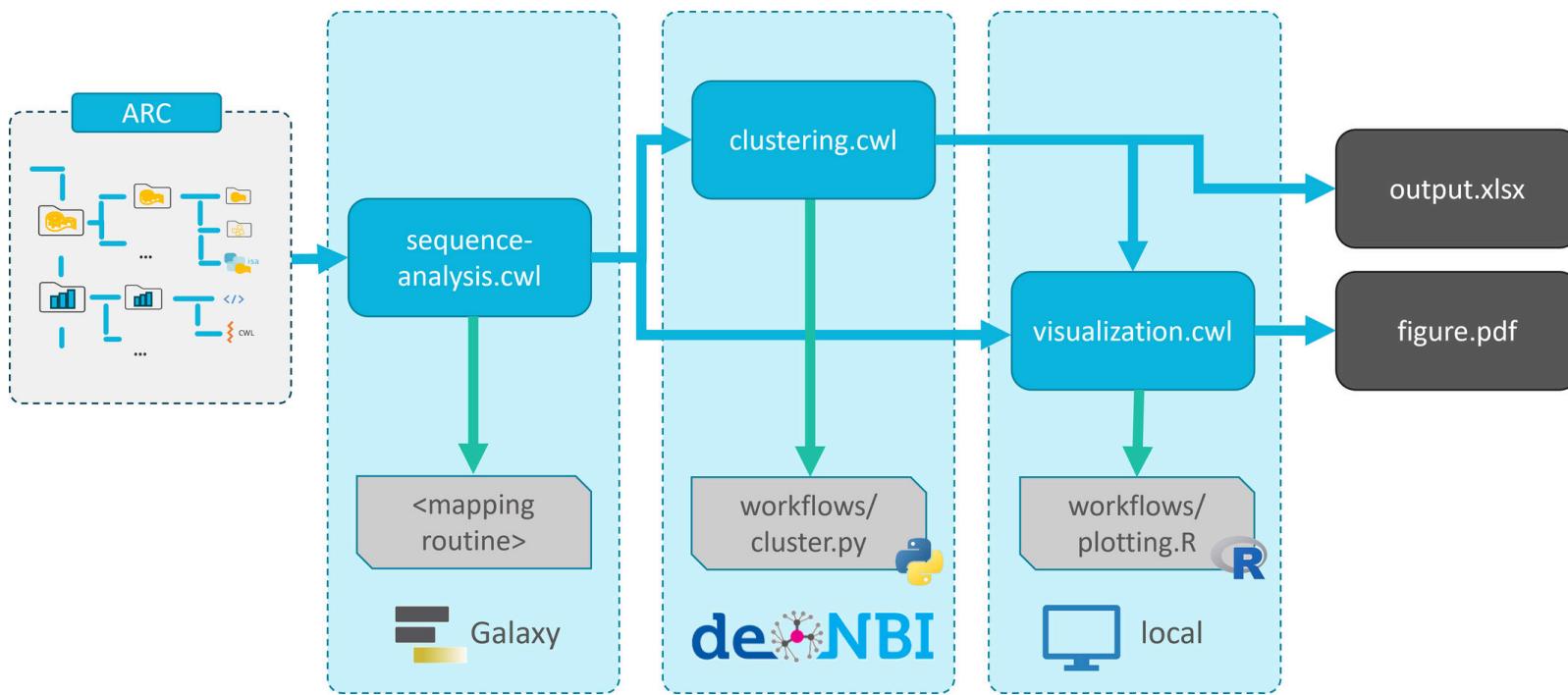


Process (e.g. annotate)



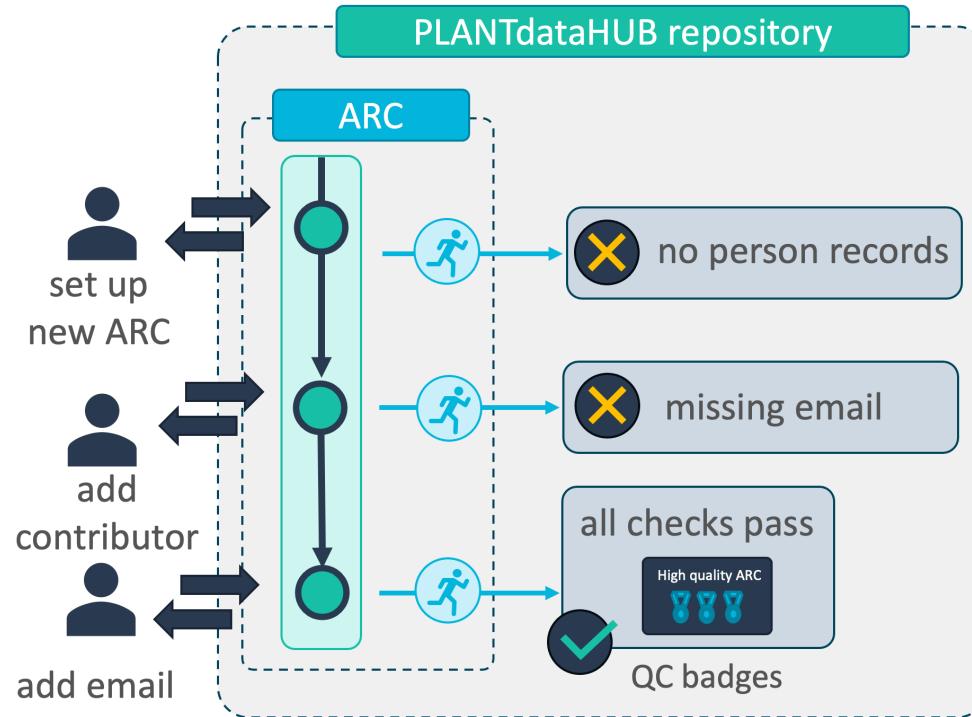


Analyse





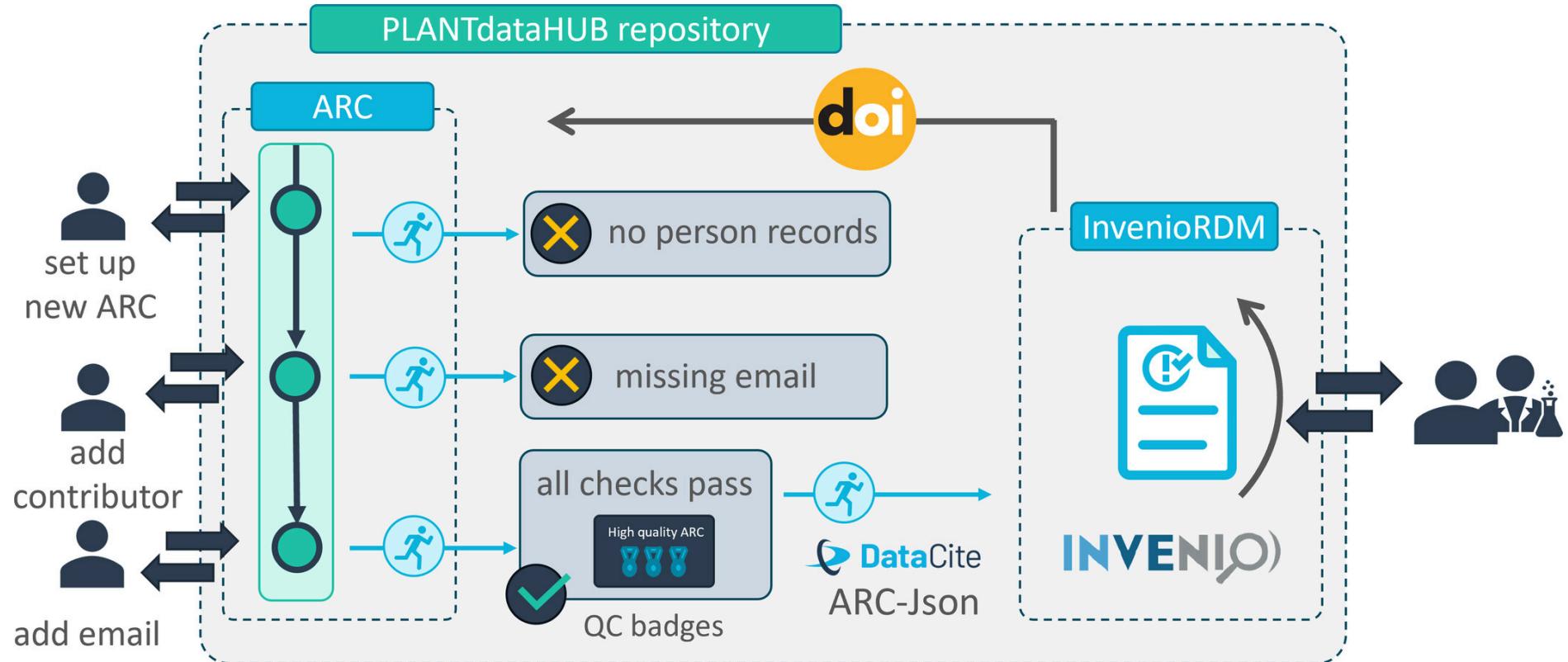
Preserve



adapted from 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>

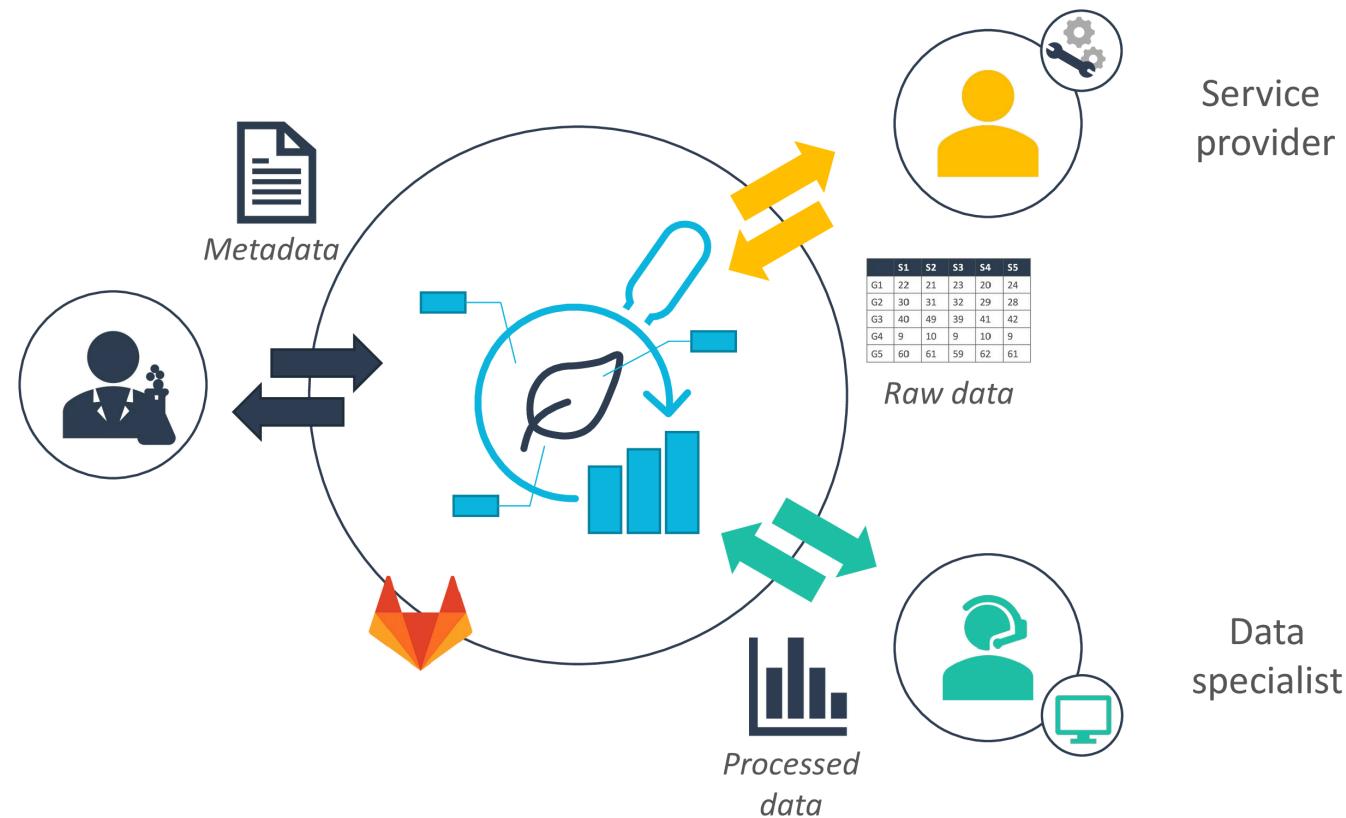


Preserve and publish



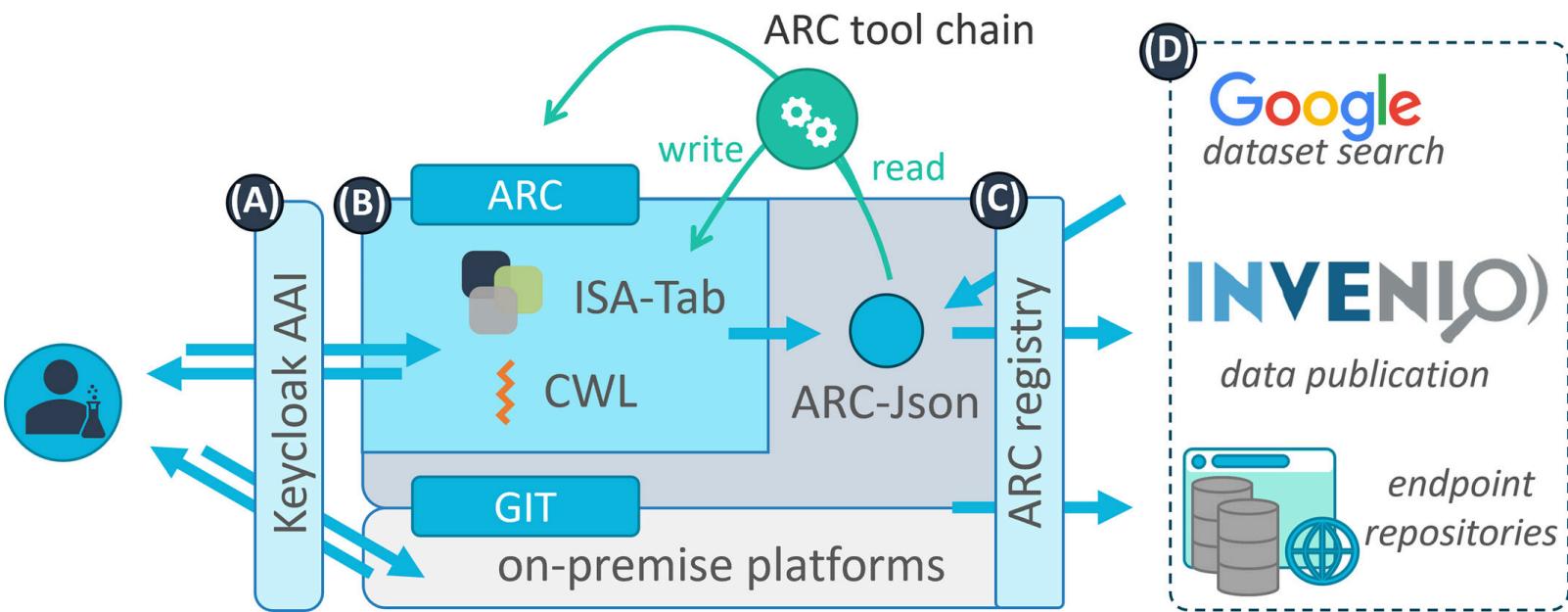


Share and collaborate

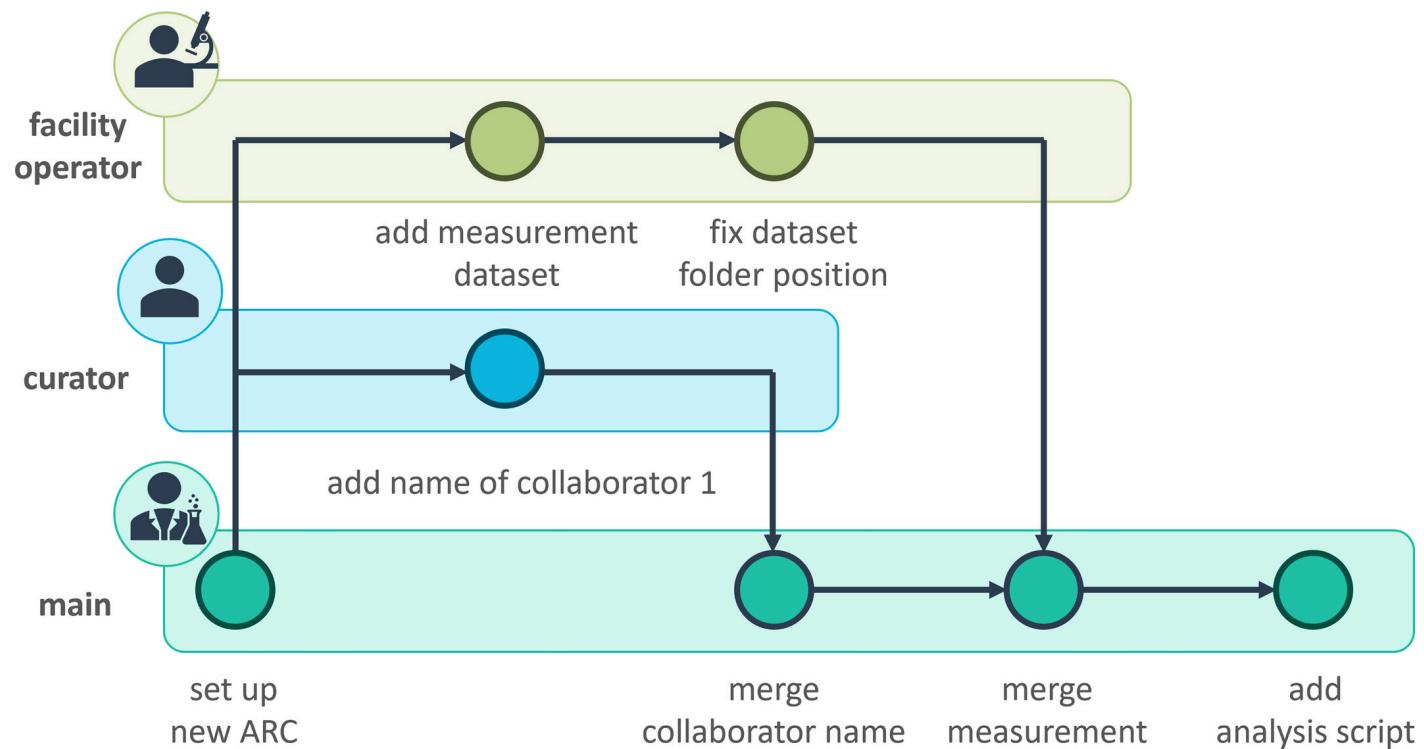




Reuse

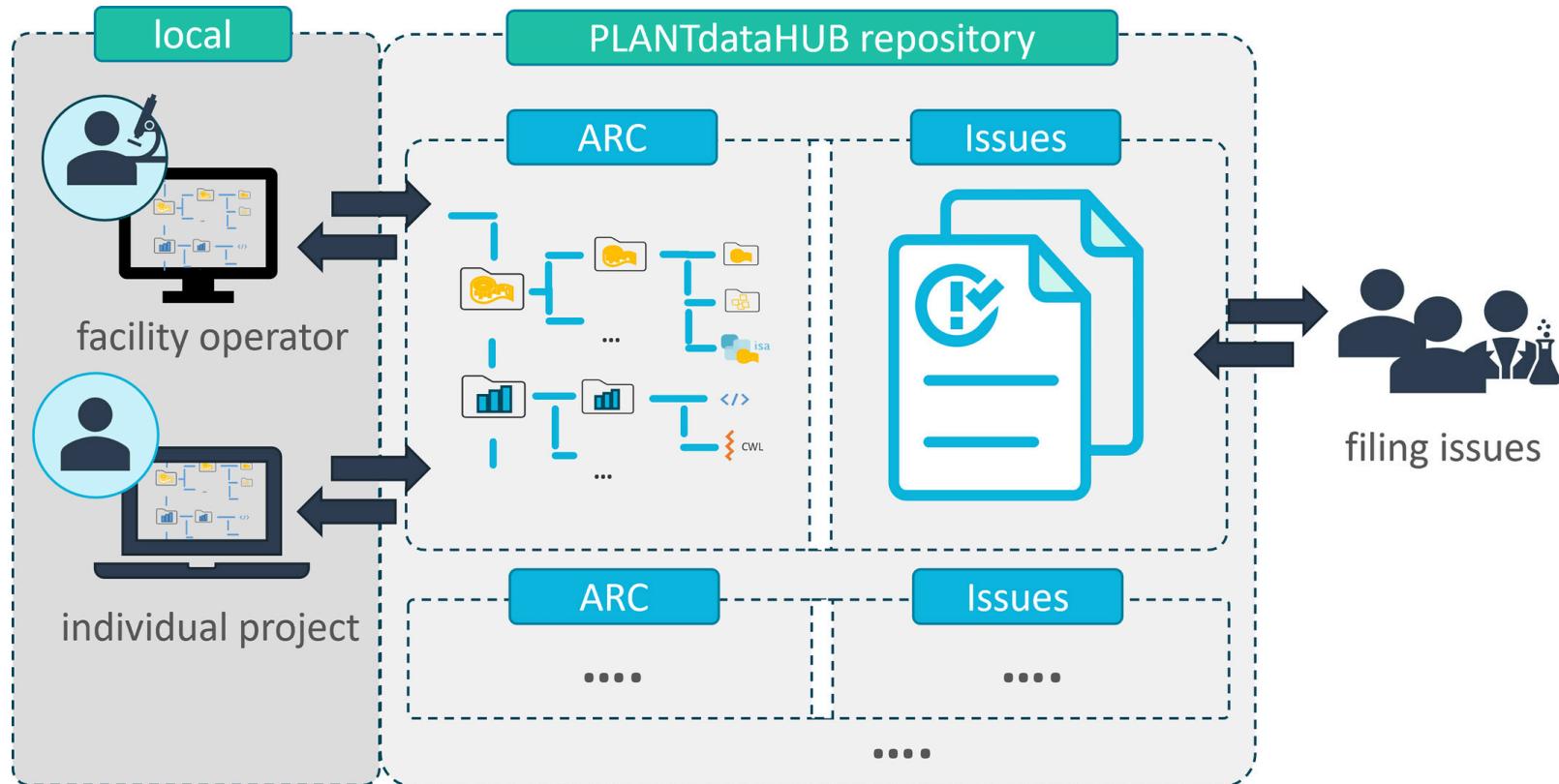


Mutable data life cycle





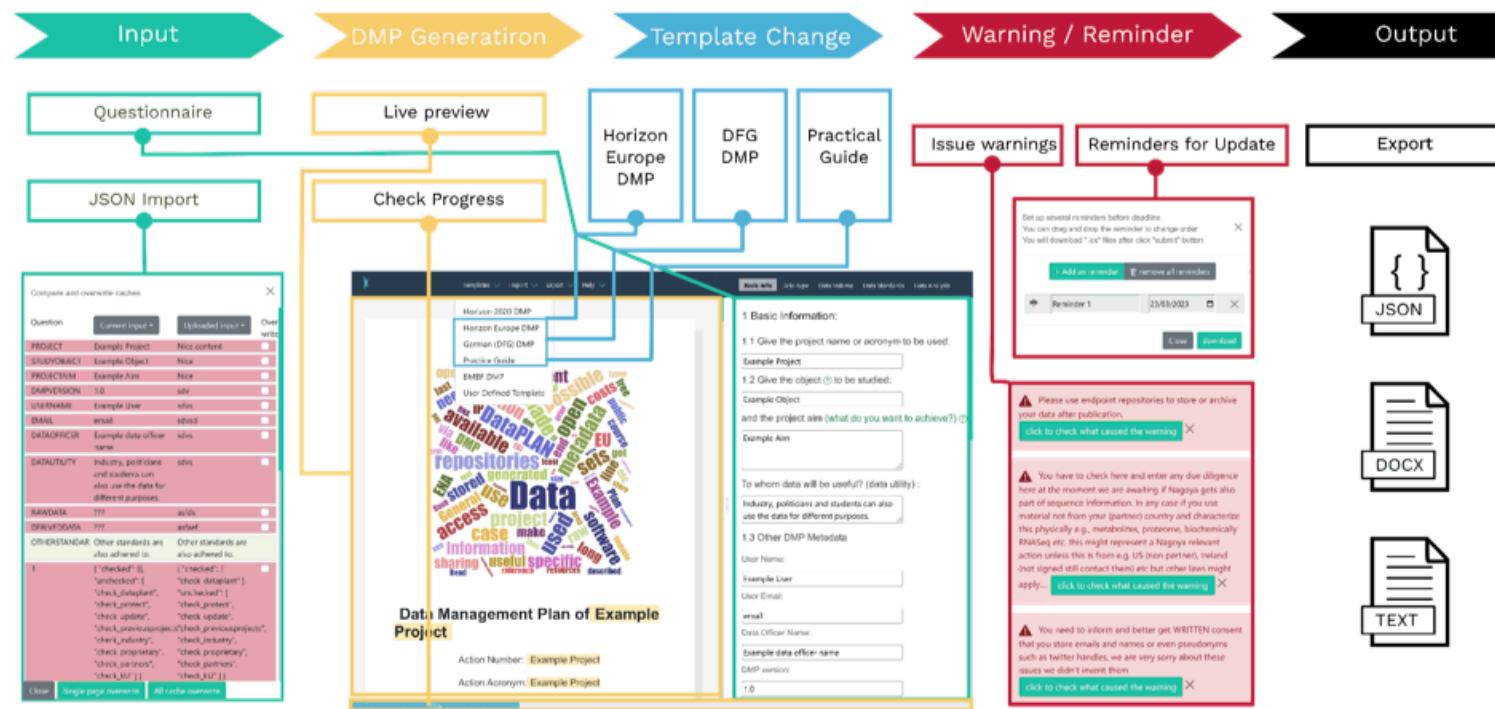
Plan (ARC scale)





Plan (proposal scale)

<https://dmpg.nfdi4plants.org>



Zhou et al. (2023), DataPLAN: a web-based data management plan generator for the plant sciences, bioRxiv 2023.07.07.548147; doi: <https://doi.org/10.1101/2023.07.07.548147>

Contributors

Slides presented here include contributions by

- name: Dominik Brilhaus
github: <https://github.com/brilator>
orcid: <https://orcid.org/0000-0001-9021-3197>

