

# Block 0 – Preparation

before September 21st, 2023



Sabrina Zander  
[MibiNet](#)



Dominik Brilhaus  
[CEPLAS Data Science](#)

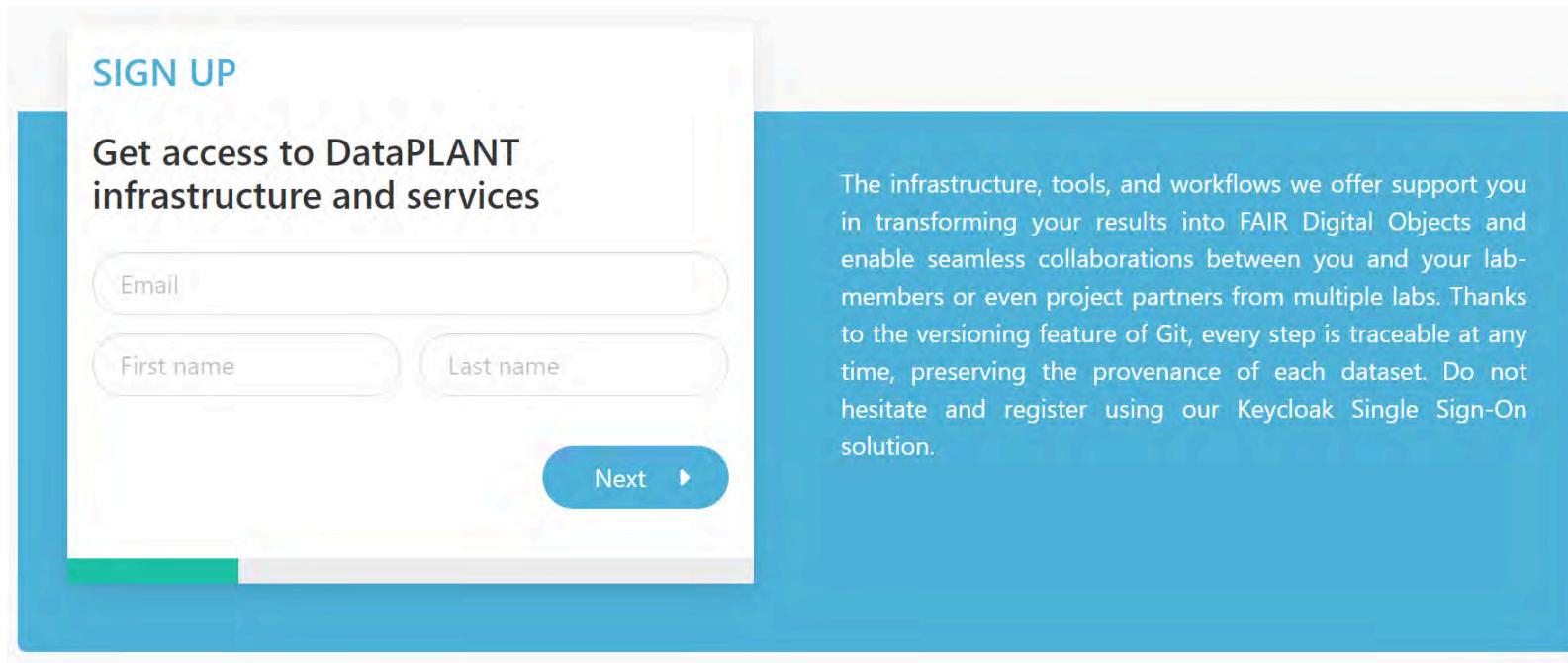
# Checklist hands-on sessions

💡 Please prepare the following before the workshop:

- Register at DataPLANT
- Install ARCitect on your computer
- Install Swate on your computer
- Bring your own data
- Find your command line
- Install ARC Commander and dependencies on your computer
- Install VS Code

# DataPLANT Registration

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



## Role and consortium

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

The screenshot shows a 'SIGN UP' form with 'Affiliation details' fields for 'Project/consortium' and 'Research interests'. Below these is a dropdown menu titled 'Choose your Role in DataPLANT' containing 'DataSteward', 'Developer', 'Member', and 'Guest', with 'Guest' selected.

SIGN UP

Affiliation details

Project/consortium

Research interests. Multiple interests need to be separated with a comma.

✓ Choose your Role in DataPLANT

- DataSteward
- Developer
- Member
- Guest

# 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.

# Hands-on: Bring your own data

In the hands-on session, we would like to start creating an ARC together.

To do so, please bring some data!

This can be data of your current research project or an already published manuscript with supplemental data. Anything that you feel familiar with.

# Recommended for trouble-shooting

 We will likely not use the tools on the next few slides. However, as of now (September 2023), it's probably better to have them ready for trouble-shooting and to show some inner workings of the ARC.

# 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.

# 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](#)

## 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>



# Contributors

Slides presented here include contributions by

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

# Block 1 – Welcome and Intro

September 21st, 2023



Sabrina Zander  
[MibiNet](#)



Dominik Brilhaus  
[CEPLAS Data Science](#)

# Welcome

# House-keeping

Pad: <https://pad.hhu.de/Aem023liTWKyfPysU0H8Gw>

# Your motivation

- how to **organise** and handle the data collected from my experiments
- grasp important concepts on **research data management**
- make data **available to the others** in the work group
- store the data in a sorted way to allow **reproducibilty** and **not loose any important data**
- learn how to **store** my data correctly
- help others to **find** important data
- familiarize myself with systems and organizing data in an easy but also **time efficient way**
- data has to be **accessible and usable** for other members of the project
- generate data that can be used by the **other projects**
- know how **data (including metadata)** should be structured and integrated uniformly in the ARC
- learn to build **ARCs** in order to have the data produced during the PhD in an organised manner
- ...

# Tentative agenda

## Day 1

Time	Topics
09:30 - 10:45	Welcome and intro to RDM
10:45 - 11:00	<i>Short break</i>
11:00 - 12:00	Intro to DataPLANT and ARC
12:00 - 13:00	<i>Lunch</i>
13:00 - 15:30	ARC Demo and ARC Hands-on

## 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
12:00 - 13:00	<i>Lunch</i>
13:00 - 15:00	Hands-on Swate
15:00 - 15:30	Wrap-up

# Introduce yourselves

- UoC / HHU
- CEPLAS / MibiNet
- Used code / programming language before
- Has an ORCID

# Goals

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



In this workshop we focus more on **how** and less on **why**

# 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 ?

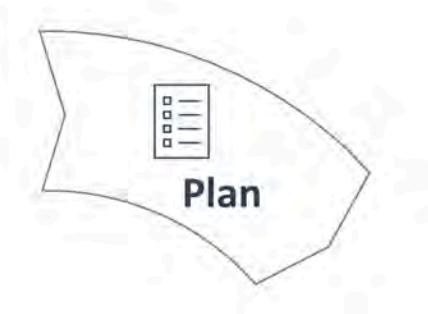
I did... 3 years ago



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



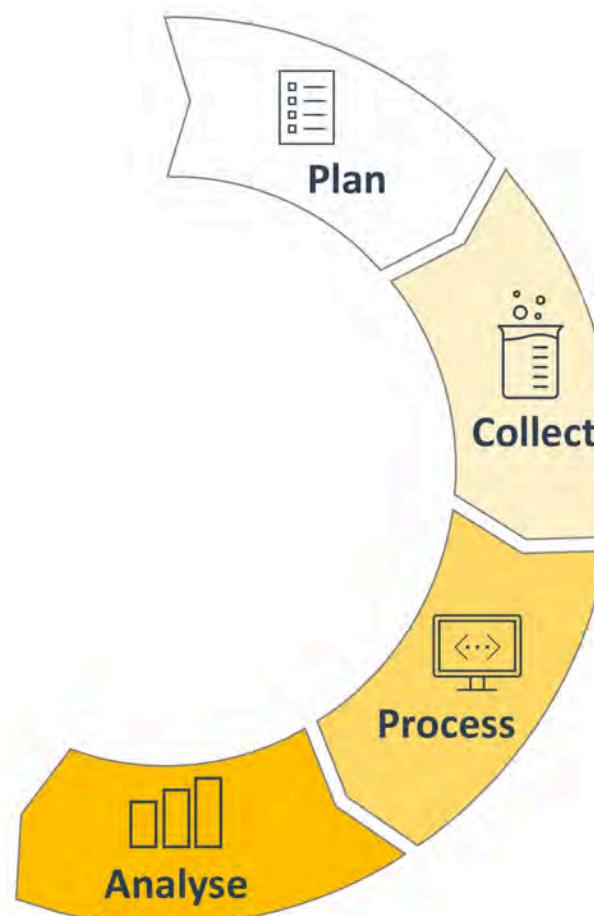
# The Research Data Lifecycle



# The Research Data Lifecycle



# The Research Data Lifecycle



# The Research Data Lifecycle



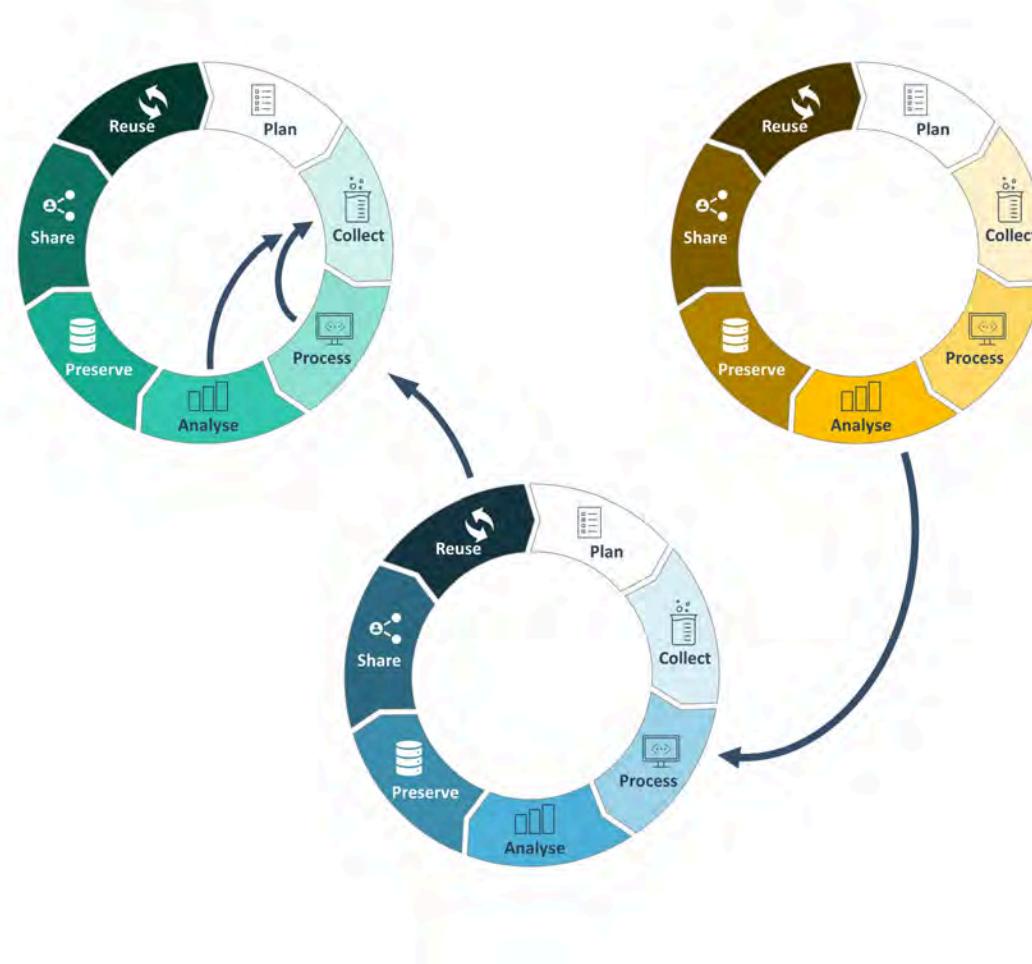
# The Research Data Lifecycle



# The Research Data Lifecycle



# The Research Data Lifecycle *is mutable*



**Have you ever heard about the  
FAIR principles?**

# 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



Findable



Accessible

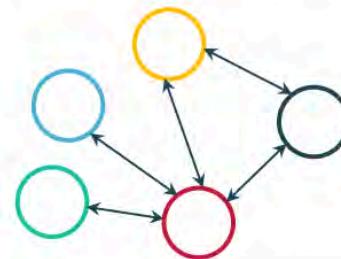


Interoperable



Reusable

Easier collaboration & sharing



Increased findability and visibility



Reproducibility



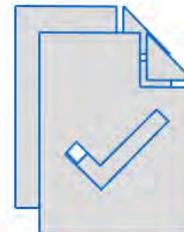
Added-value to the research community



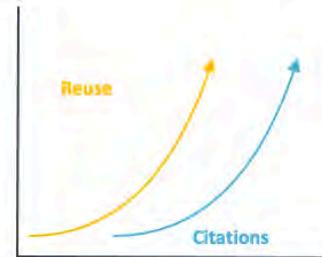
EMBL-EBI



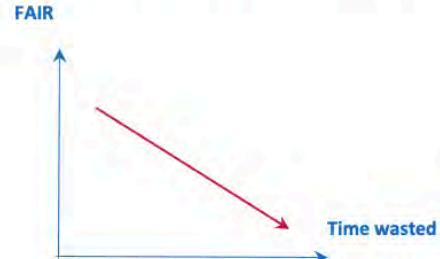
Compliance with funding policies



Receive due credit



Saves time & workload



# Is your data FAIR?

Findable | Accessible | Interoperable | Reusable

- Where do you store 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

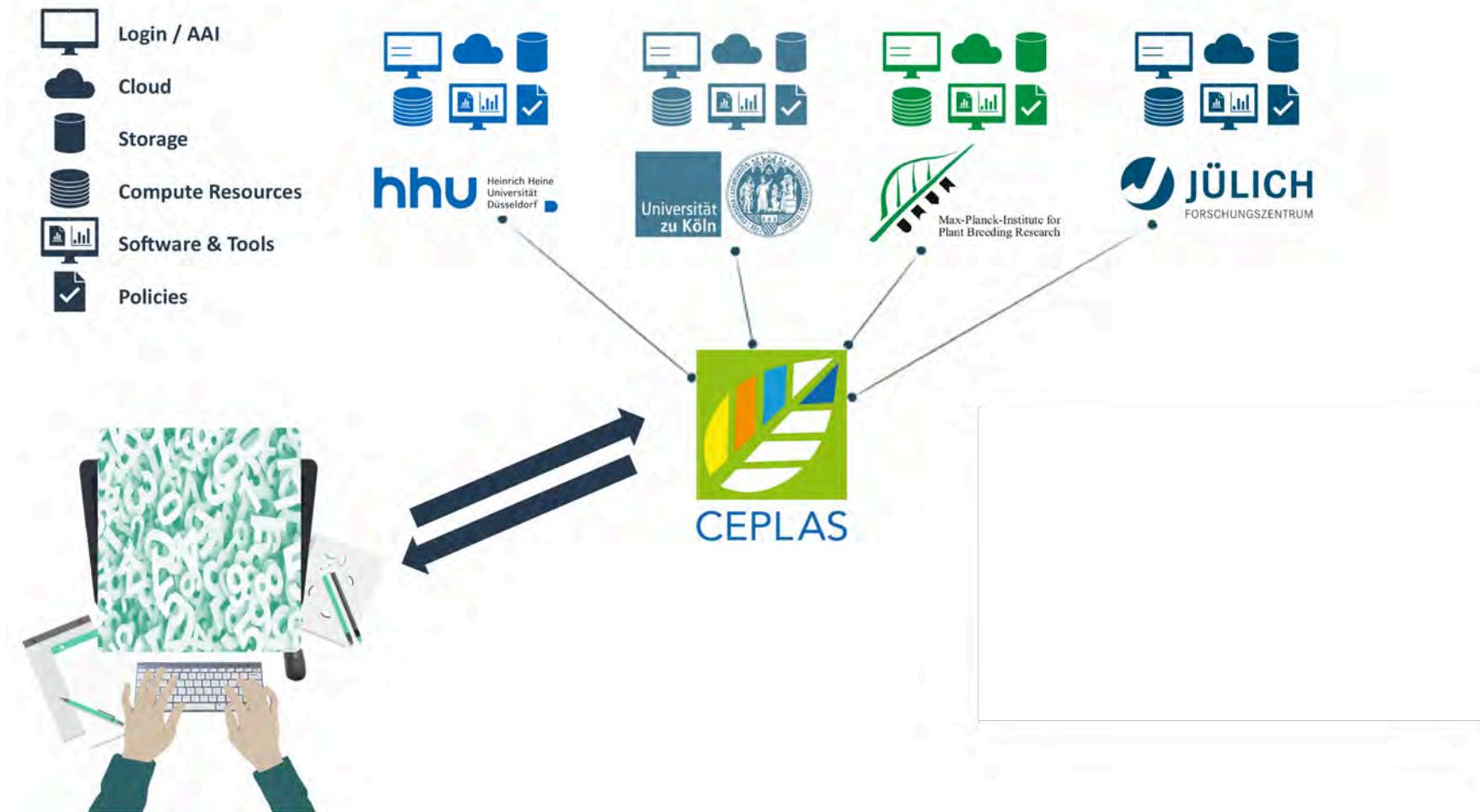
The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

- 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

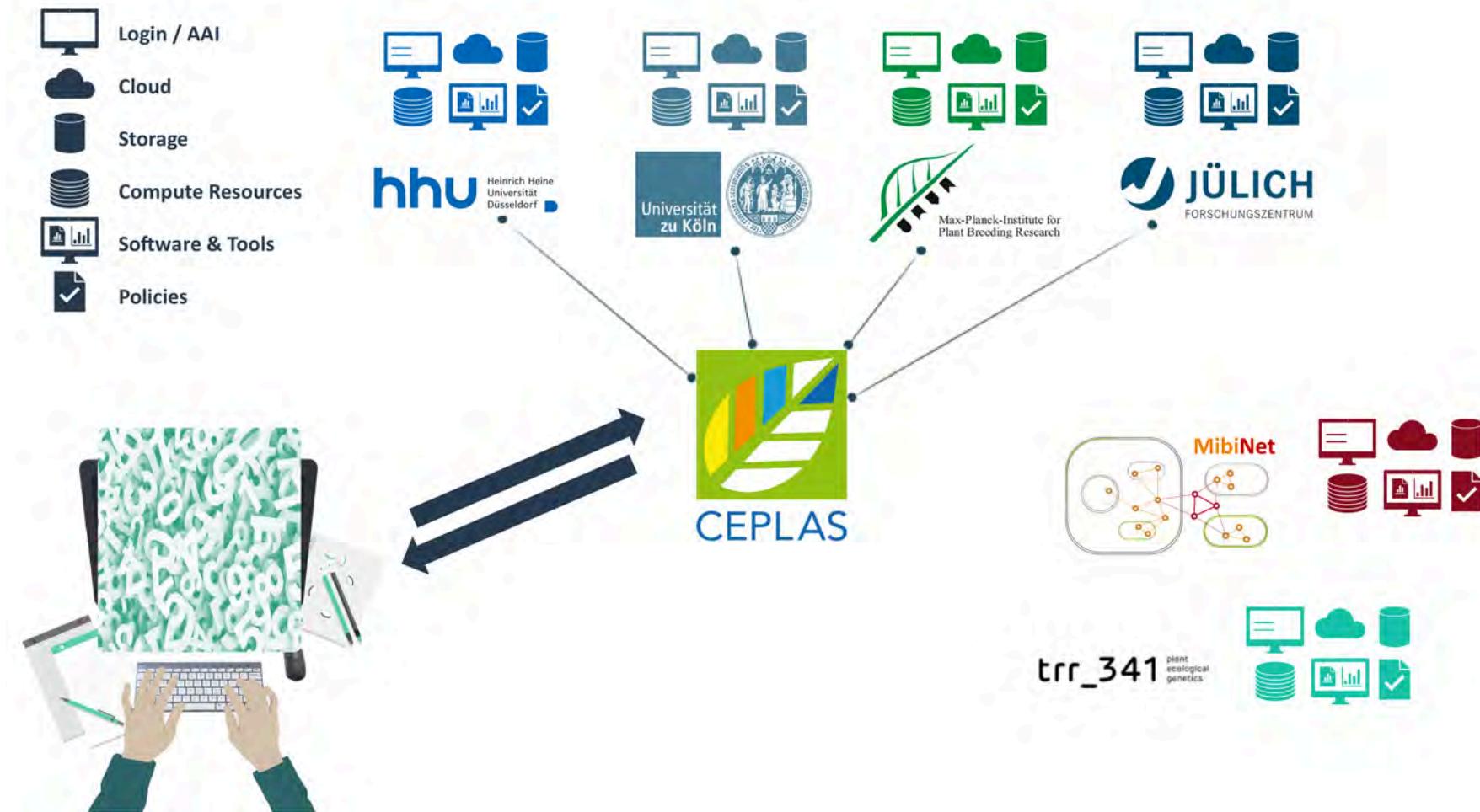
# FAIR on multiple layers

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



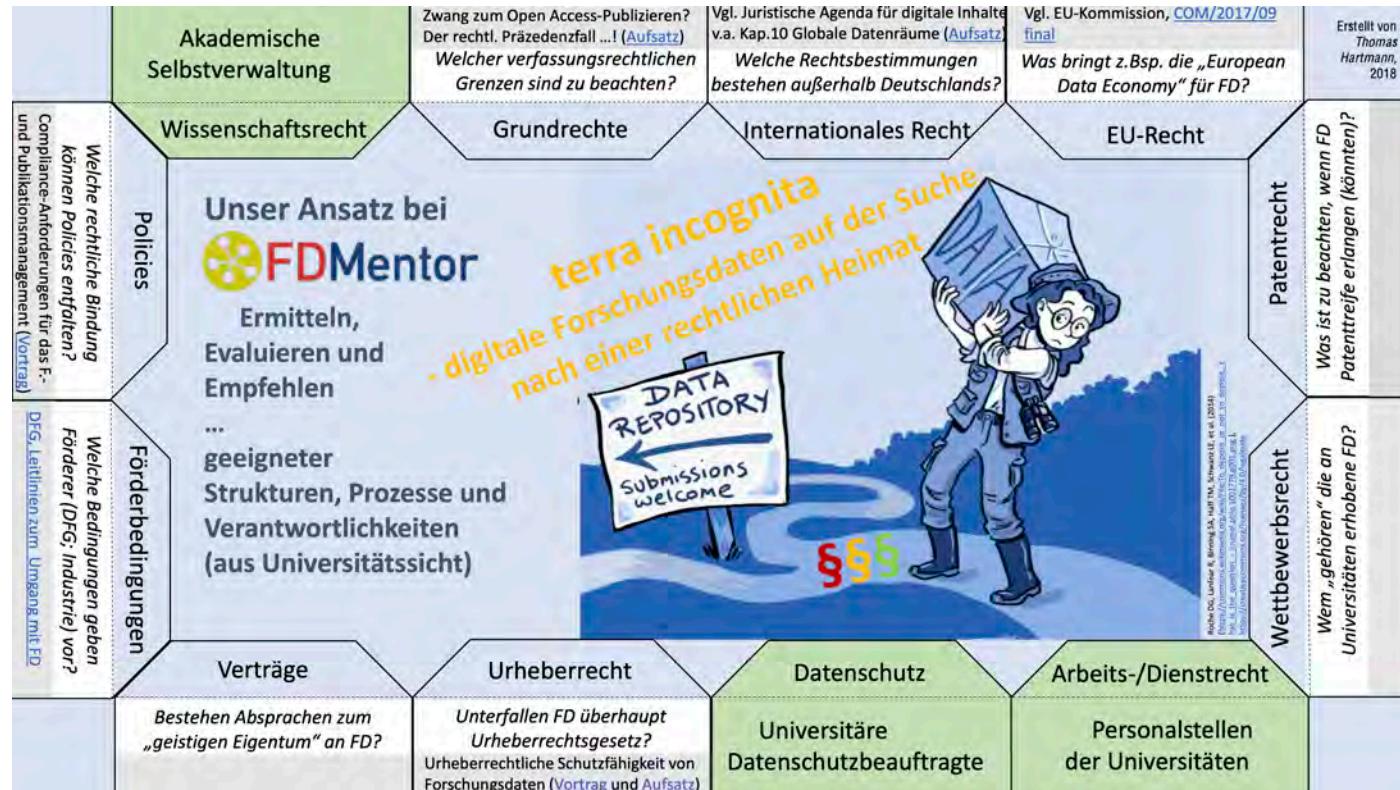
# RDM fundamentals

Dominik Brilhaus

Sept 20th, 2023

# Legal aspects of RDM

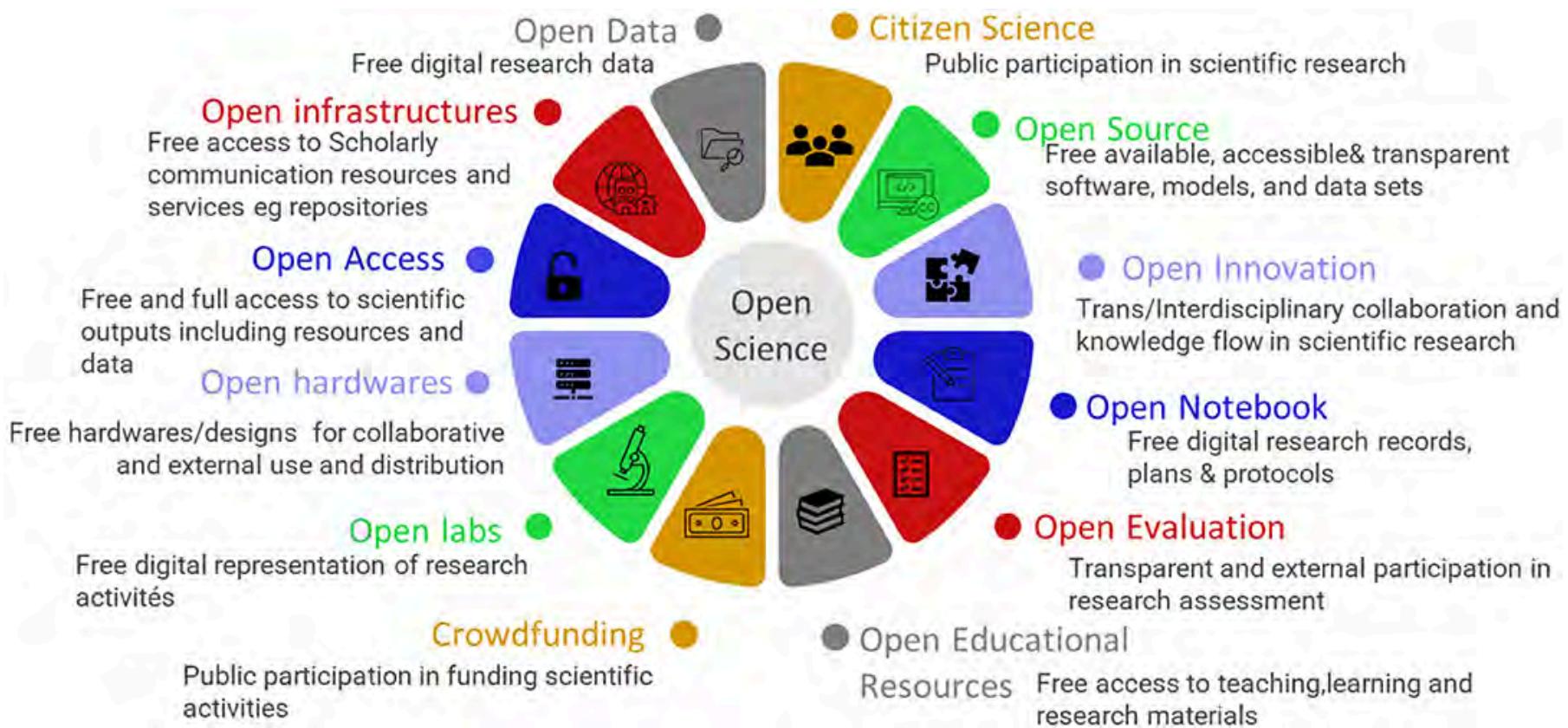
# Different laws touched by RDM



# Open Access (OA) categories

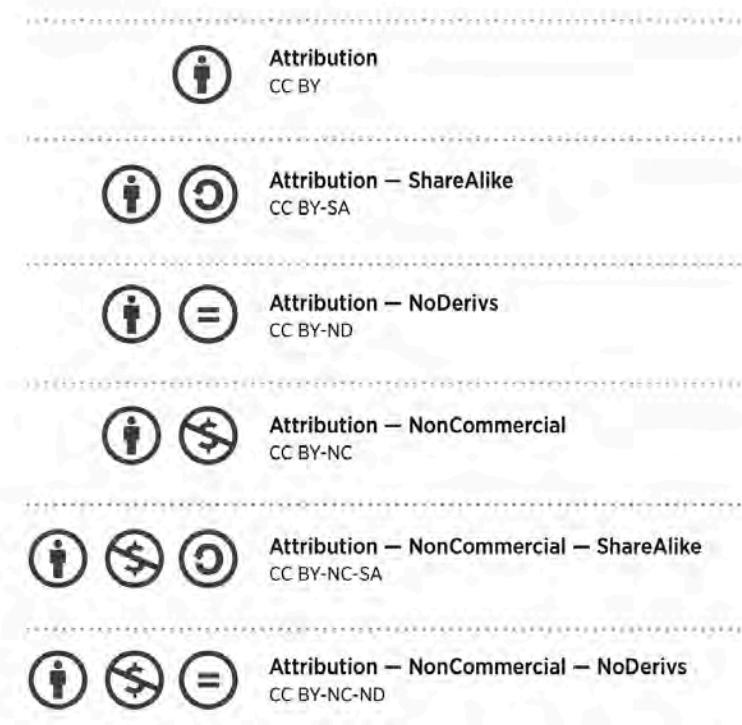
- Gold: Published in an open-access journal that is indexed by the [DOAJ](#).
- Green: Toll-access on the publisher page, but there is a free copy in an OA repository.
- Hybrid: Free under an open license in a toll-access journal.
- Bronze: Free to read on the publisher page, but without a clearly identifiable 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/>



# Data protection

GDPR: General Data Protection Regulation

DS-GVO (german): Datenschutz-Grundverordnung

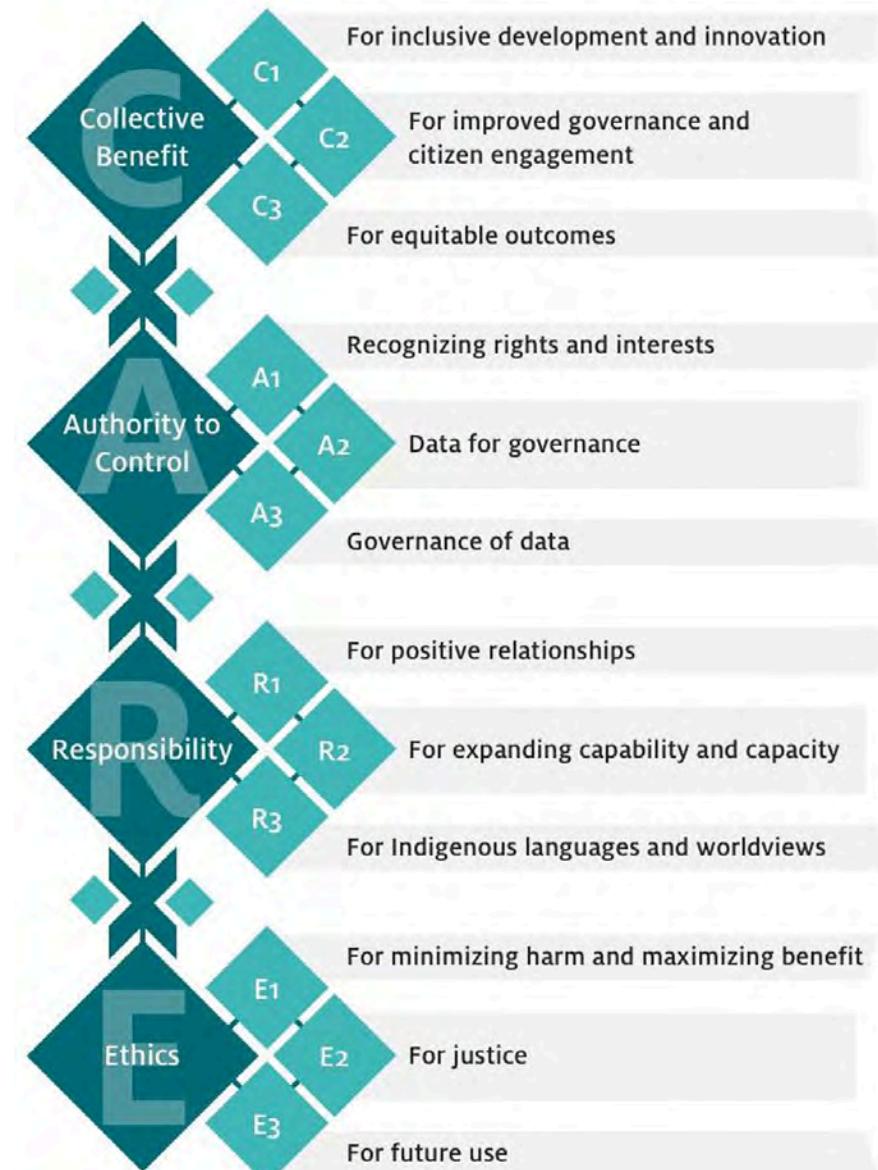
# Use of biological materials

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

# FAIR and CARE



# CARE principles



<https://datascience.codata.org/articles/10.5334/dsj-2020-043/>

# 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

# Public data repositories

# 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>

# 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

# 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

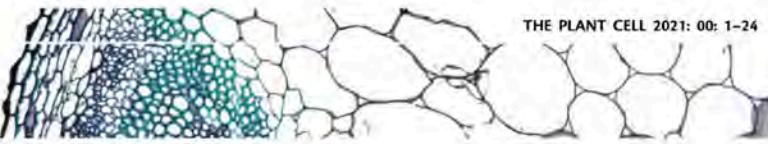


<https://figshare.com>

# Persistent Identifiers (PIPs)

# Spot the PIDs

doi:10.1093/plcell/koab243 THE PLANT CELL 2021; 00: 1–24



**Research Article**

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

Kun Li ,<sup>1,2,†</sup> Juan M. Debernardi ,<sup>1,2,\*†,†</sup> Chengxia Li ,<sup>1,2</sup> Huiqiong Lin ,<sup>1,2</sup> Chaozhong Zhang ,<sup>1</sup> Judy Jernstedt ,<sup>1</sup> Maria von Korff ,<sup>3,4</sup> Jinshun Zhong ,<sup>3</sup> and Jorge Dubcovsky ,<sup>1,2,\*†</sup>

<sup>1</sup> Department of Plant Sciences, University of California, Davis, California 95616, USA  
<sup>2</sup> Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA  
<sup>3</sup> Institute for Plant Genetics, Heinrich Heine University, Düsseldorf 40225, Germany  
<sup>4</sup> 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>

# Data stores



# Backup vs. Archive

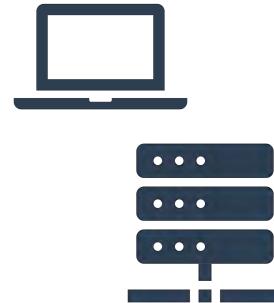
	Backup	Archive
Storage type	Short-, mid-term	Long-term
Purpose	Disaster recovery	Long-term storage, compliance
Reason	Duplication	Migration
Usage	Work in progress	Cold, Unused data
Changes	Short-term updates	No updates
Trend	Cyclic, Replacement	Growing
Latency	Short/Costly	High/Cheaper

# 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

# Data Sharing

# Cloud Services

- ✓ Documents
- ✓ Small data
- ✓ Presentations

X Code

X Data analytical projects

X Big (“raw”) data



# 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>



# Contributors

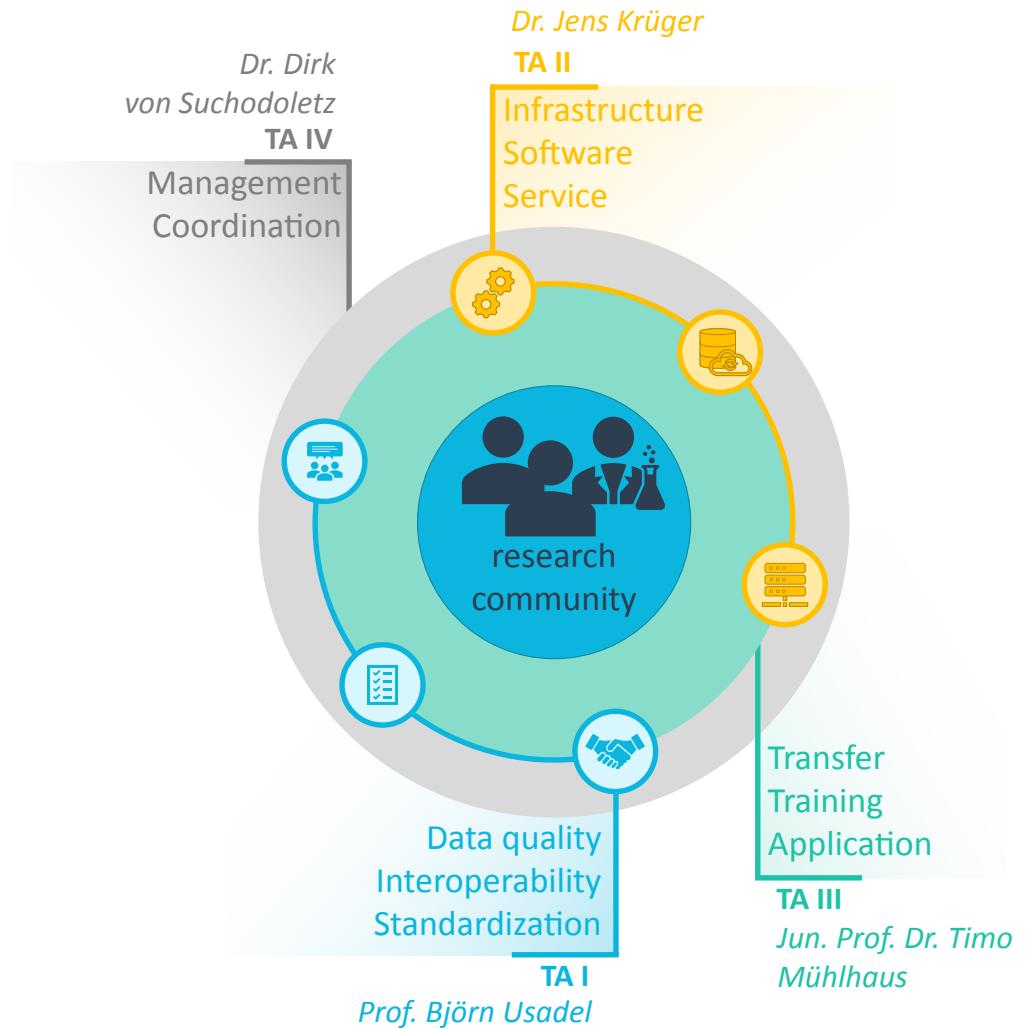
Slides presented here include contributions by

- 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>

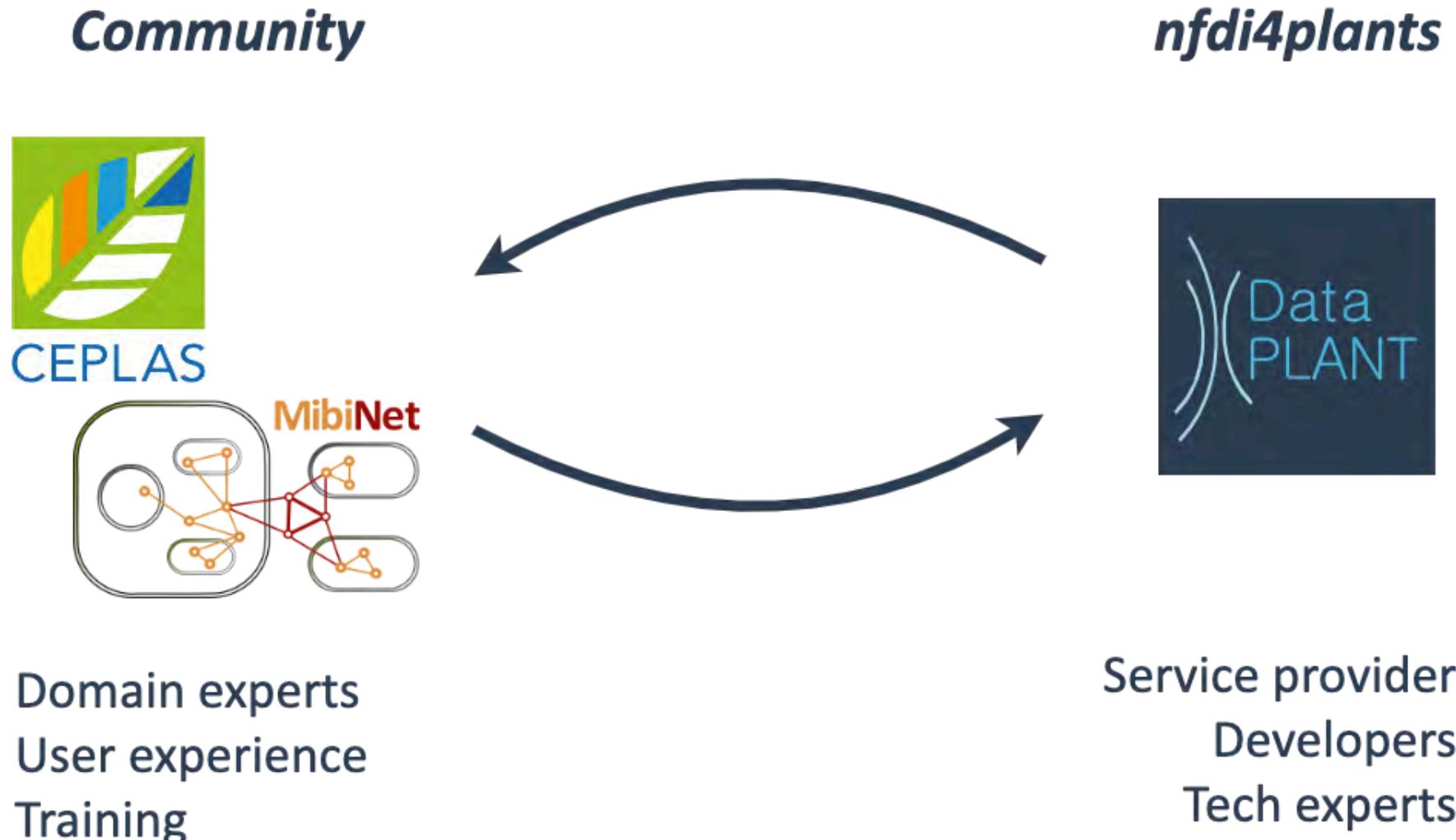
# Block 2 – Intro to DataPLANT and ARC

# DataPLANT – The NFDI4Plants

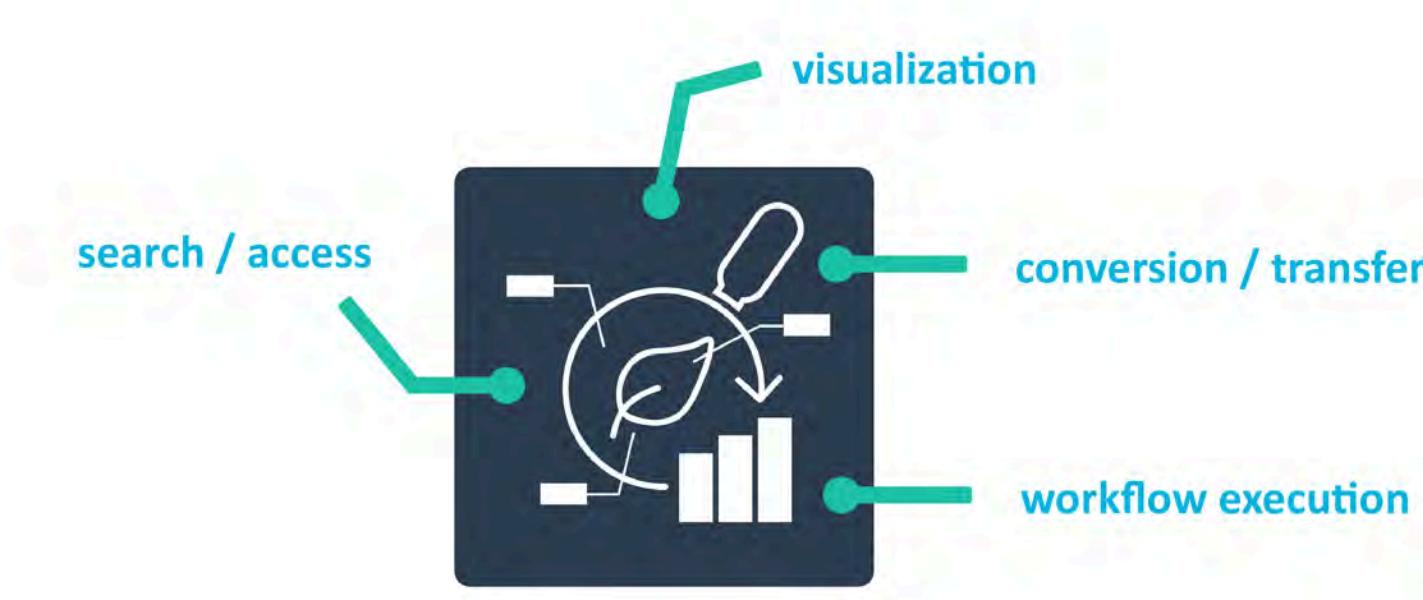
- NFDI: "Nationale Forschungsdaten Infrastruktur" – [www.nfdi.de](http://www.nfdi.de)
- Funded since end of 2020



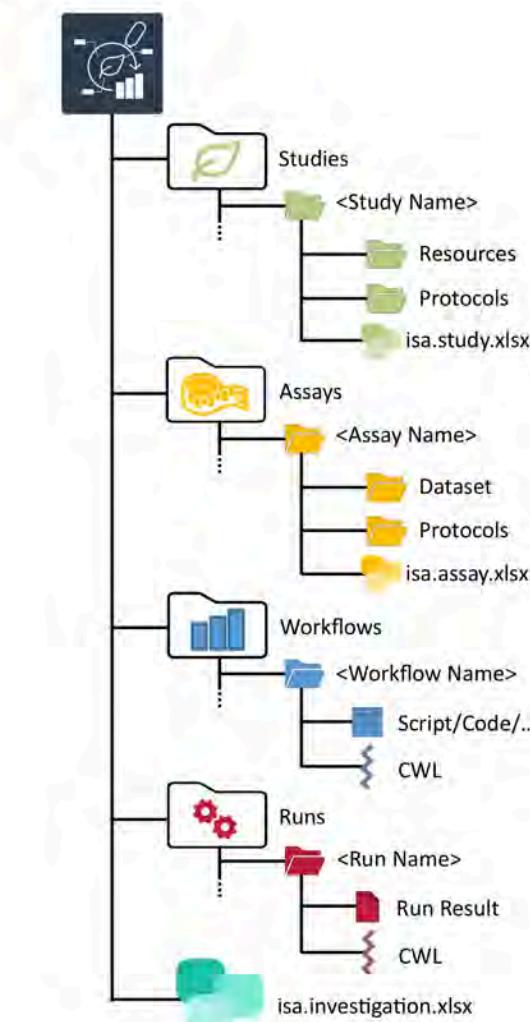
# Data Stewardship between DataPLANT and the community



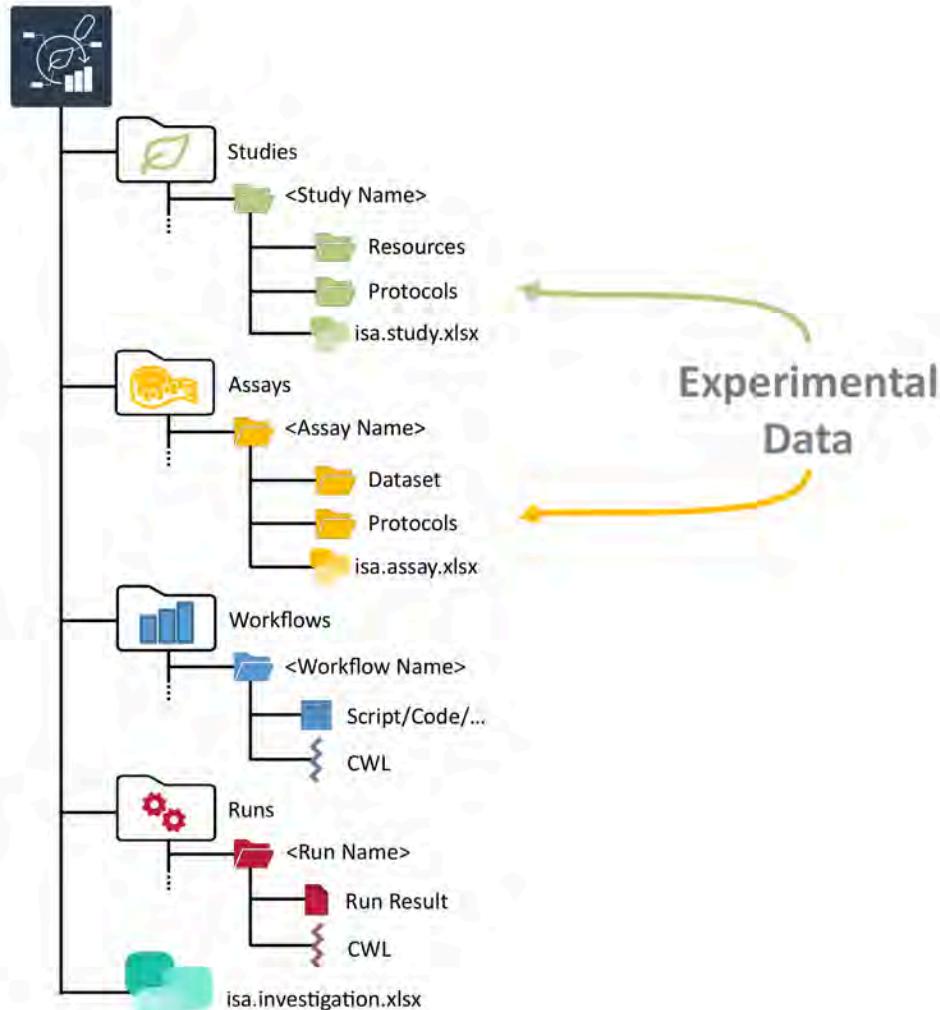
# Annotated Research Context (ARC)



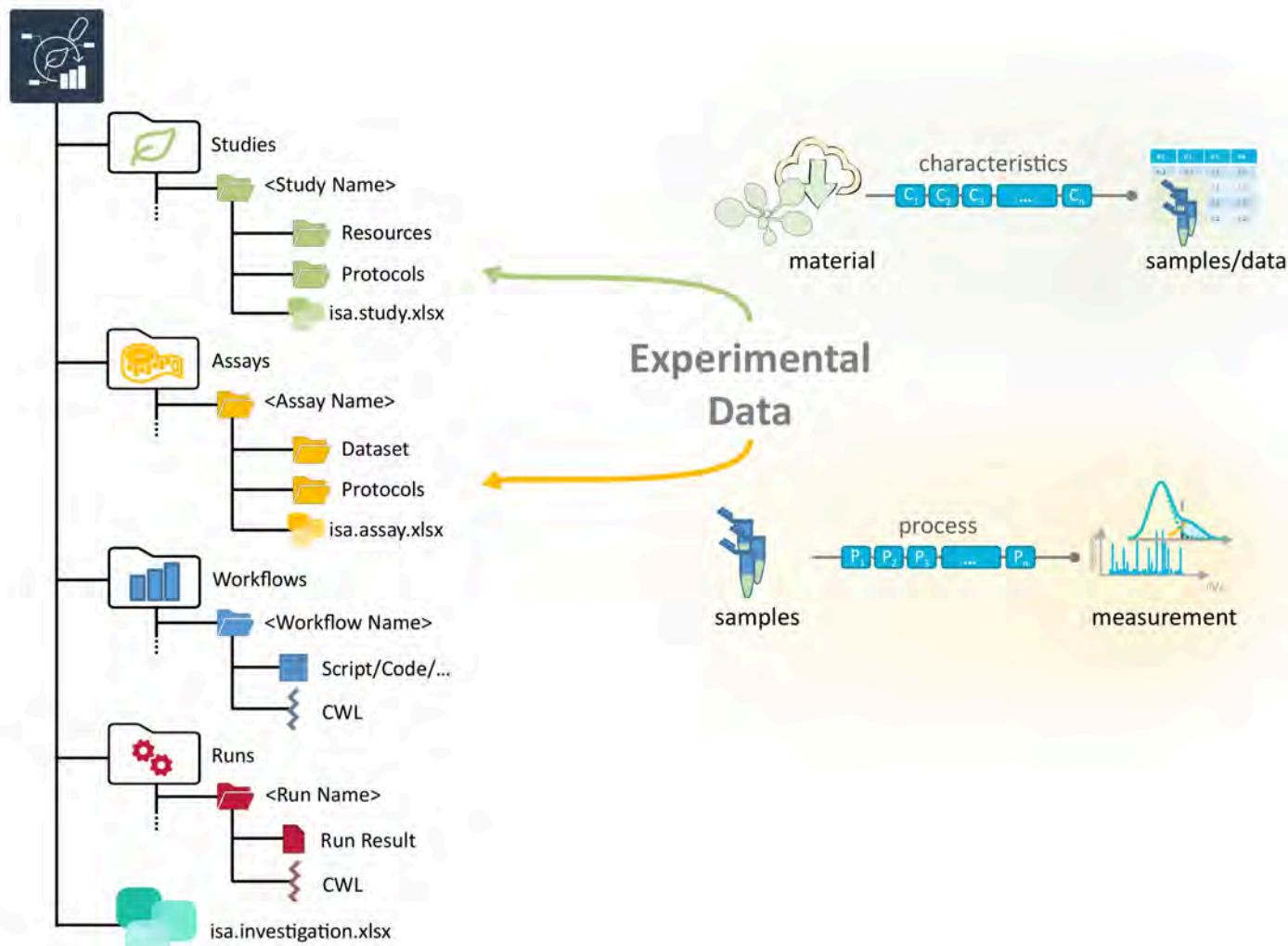
# What does an ARC look like?



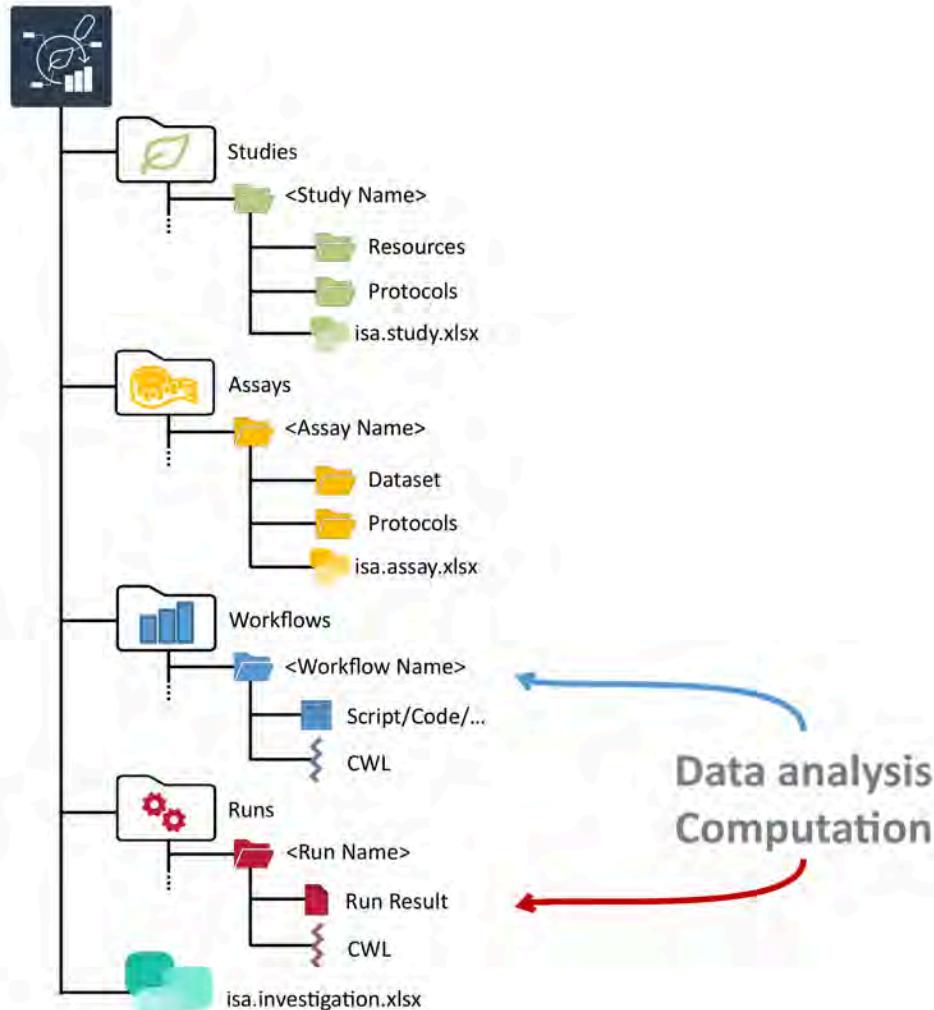
# What does an ARC look like?



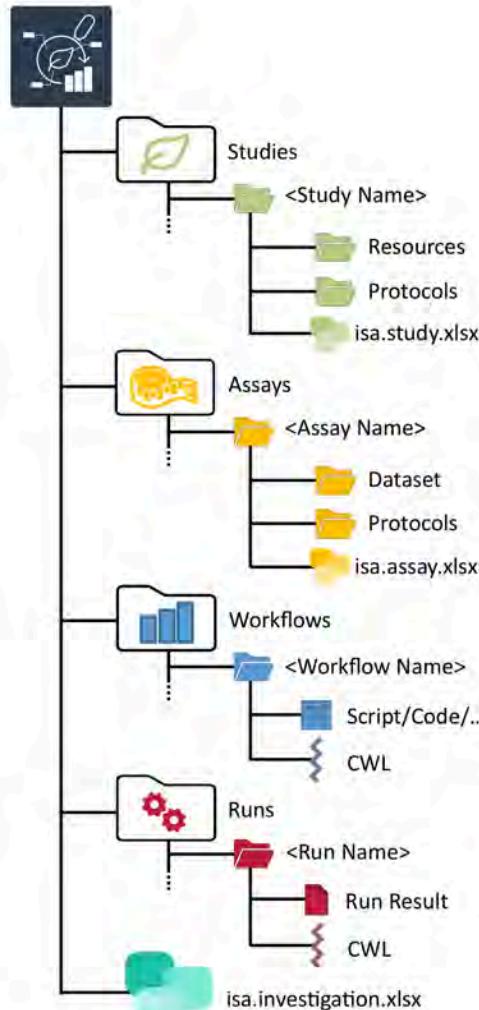
# What does an ARC look like?



# What does an ARC look like?



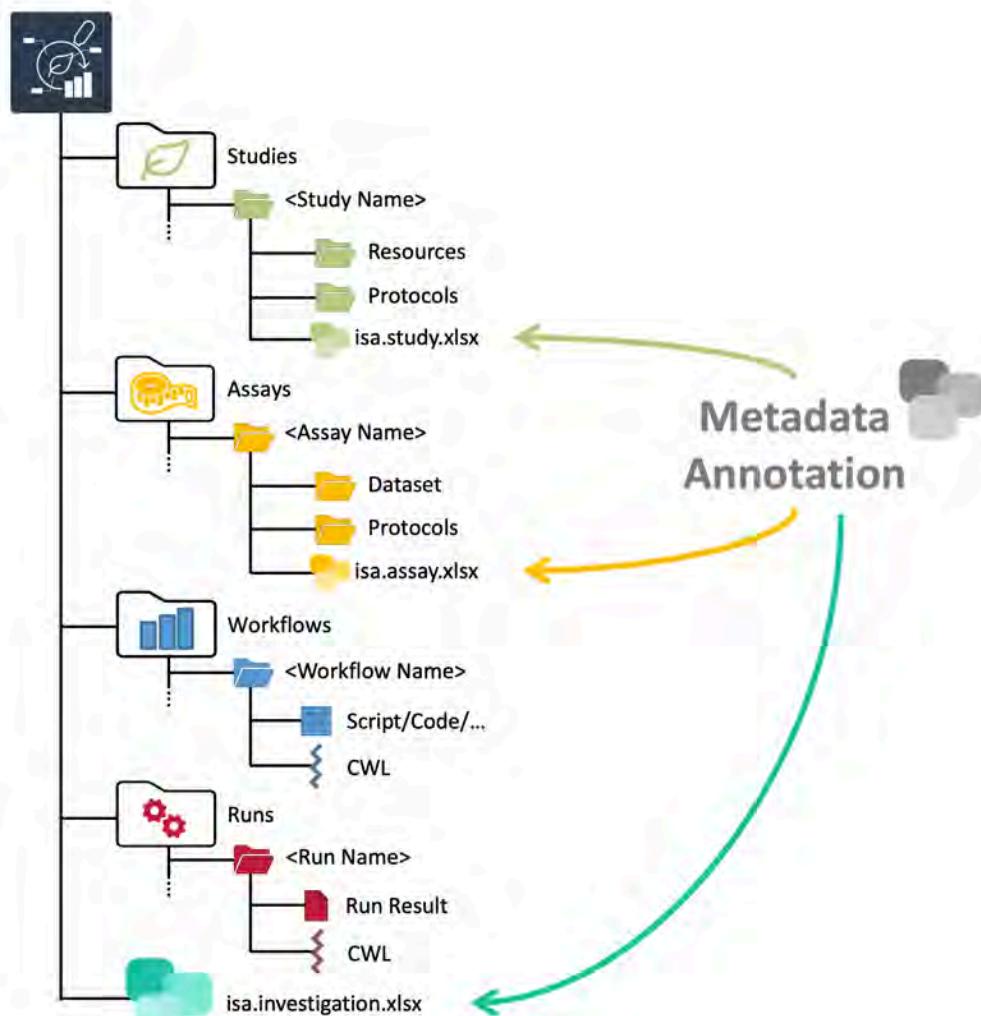
# What does an ARC look like?



Data analysis  
Computation



# What does an ARC look like?





FINDABLE

ACCESSIBLE

INTEROPERABLE

REUSABLE

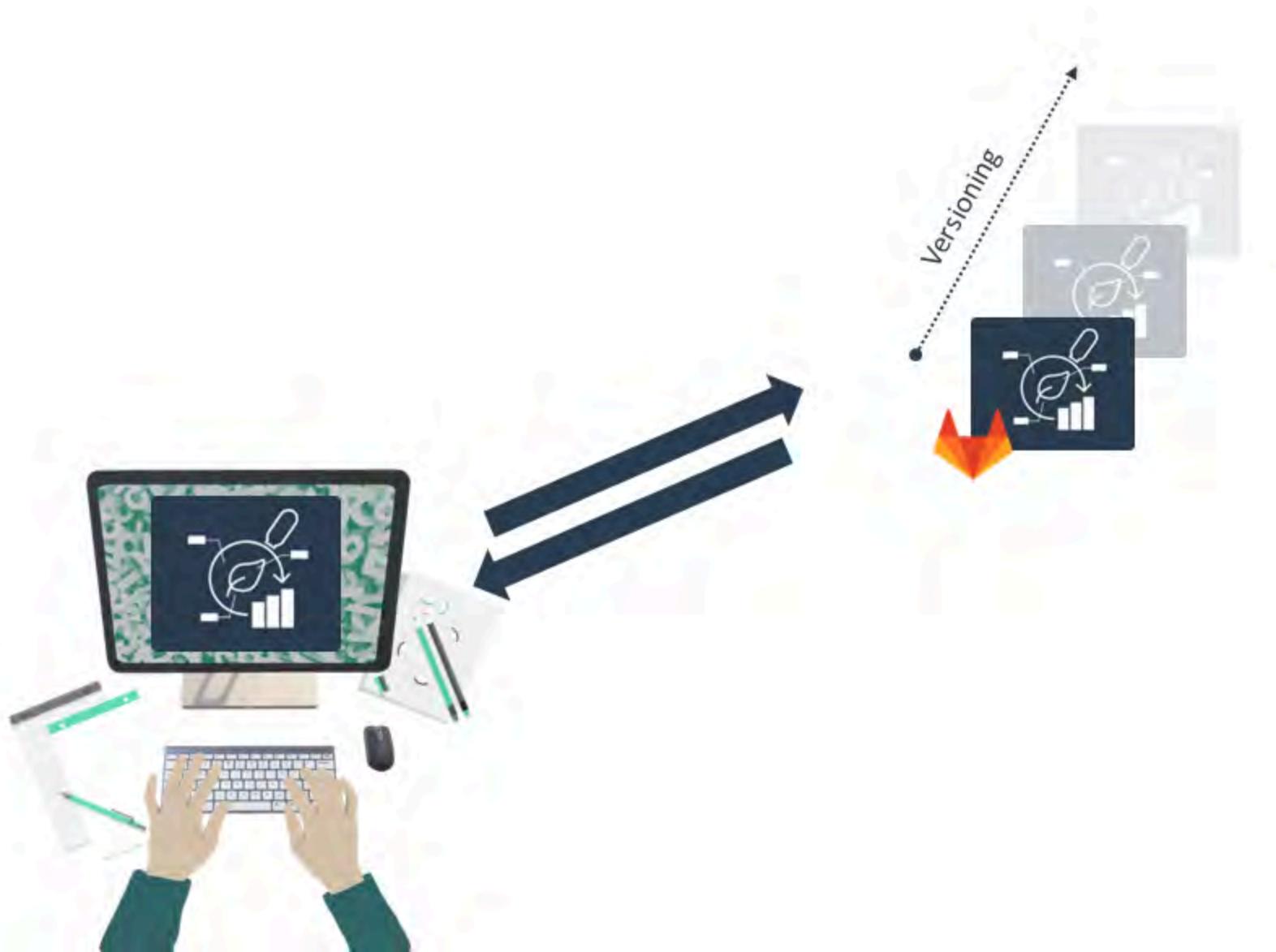
Metadata  
Templates

Controlled  
Vocabularies

Standards



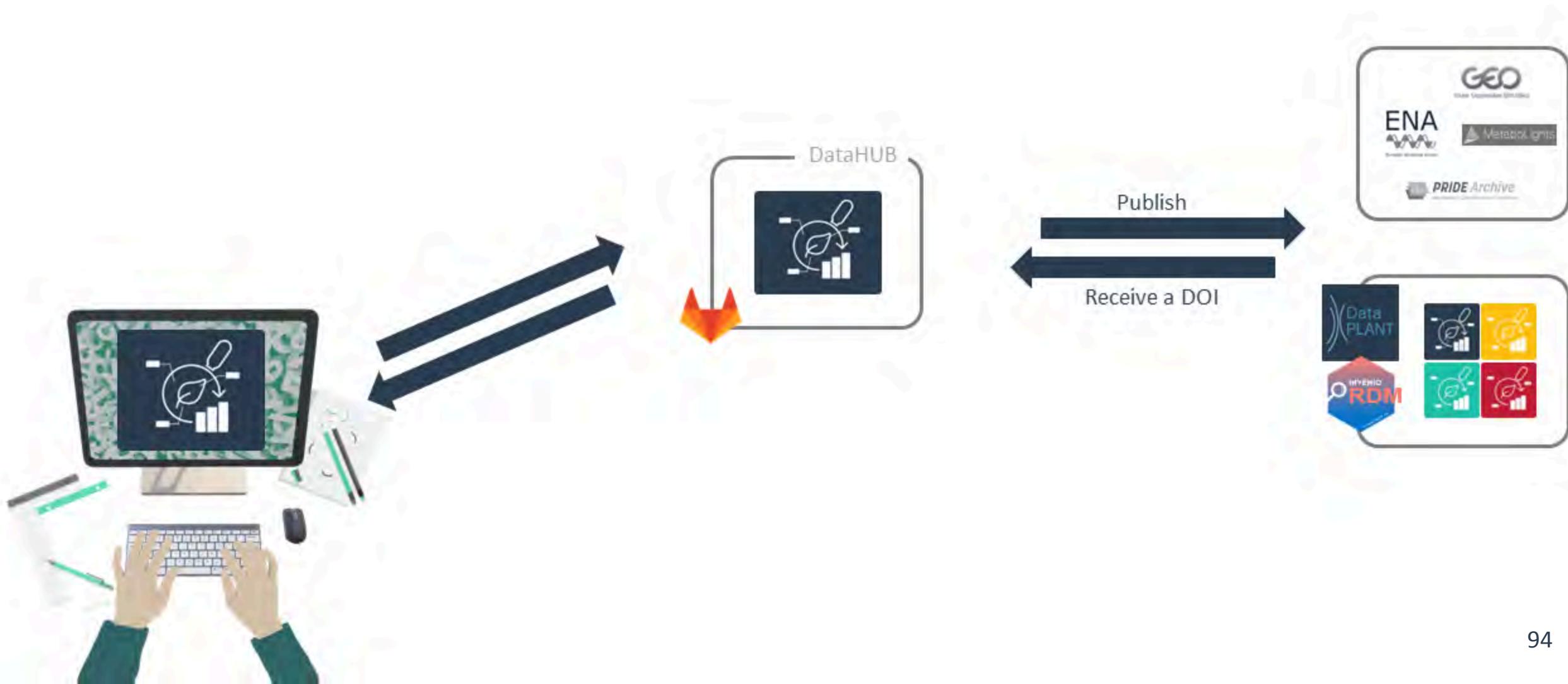


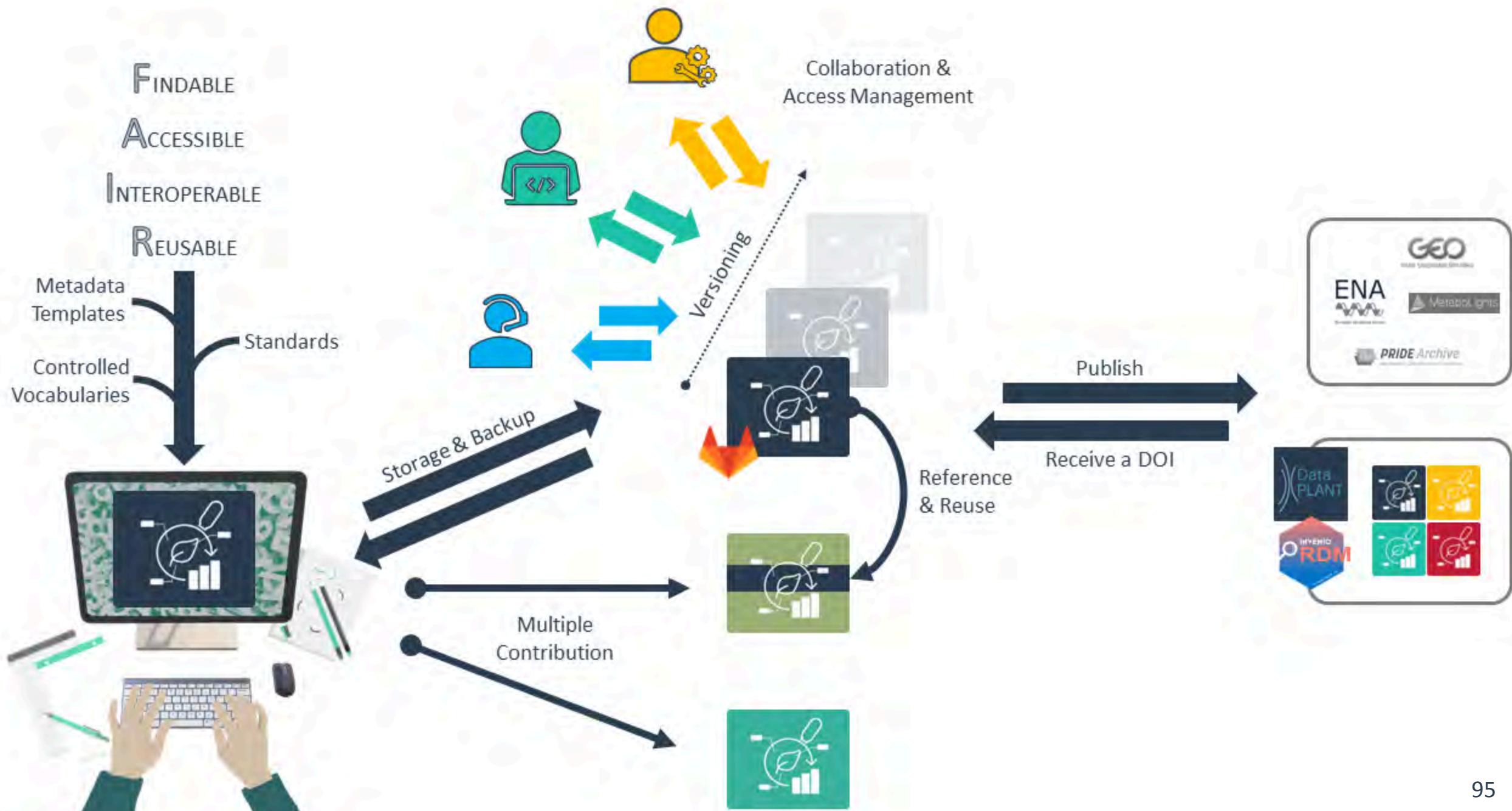










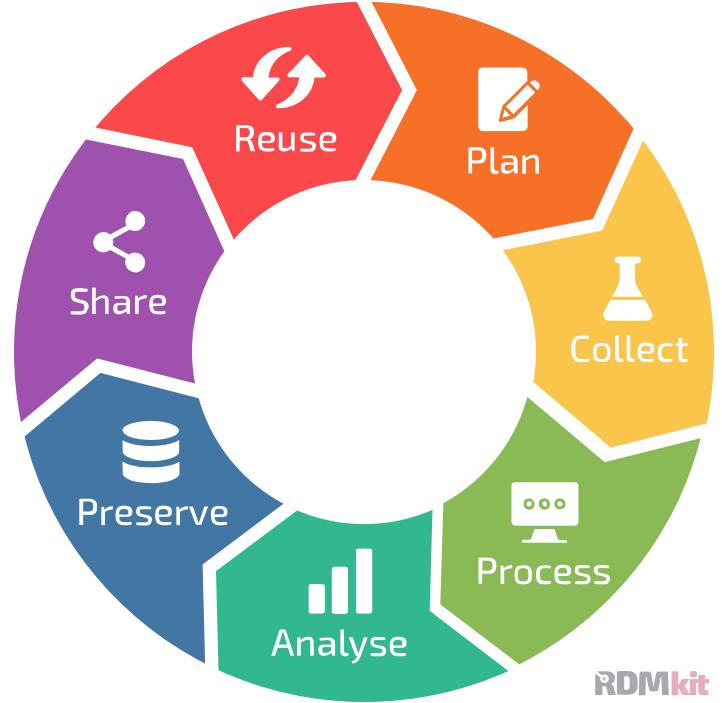




# Contributors

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>



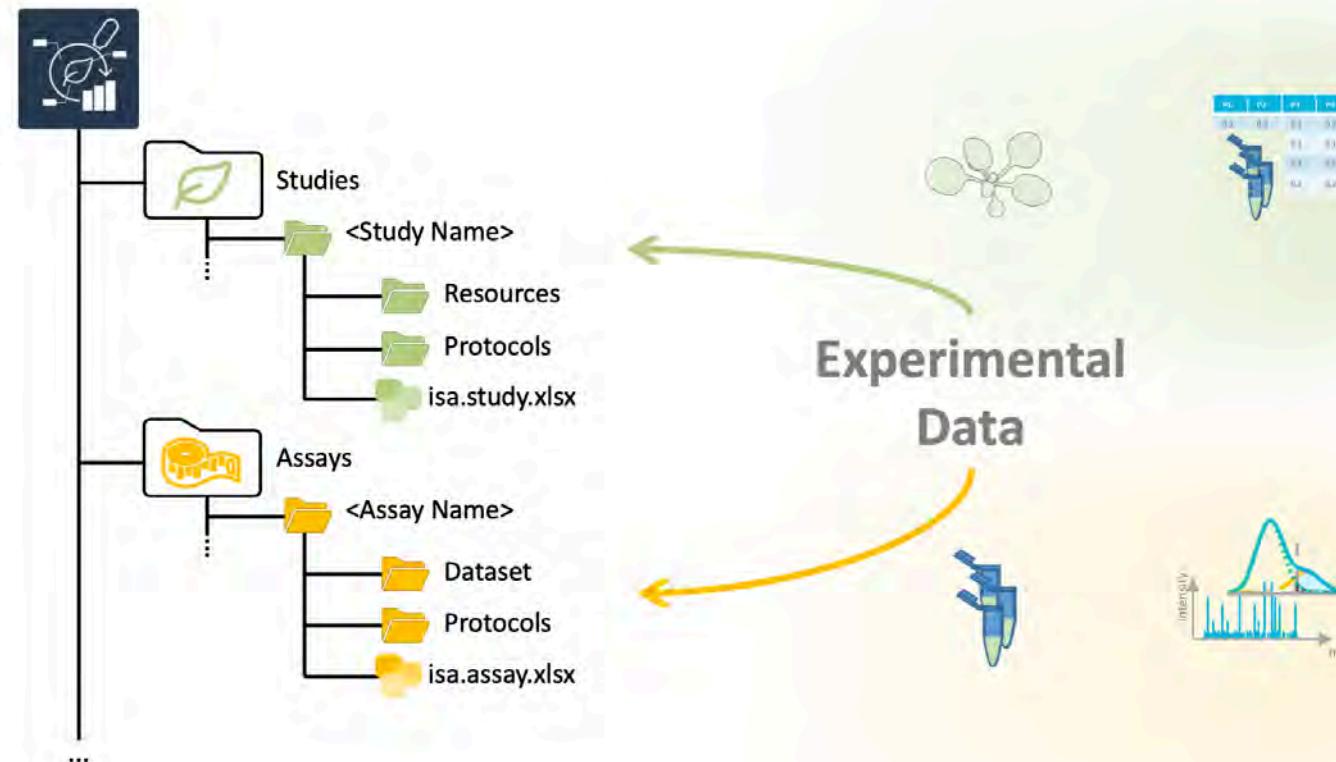
# ARC Ecosystem Demo

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

Dominik Brilhaus

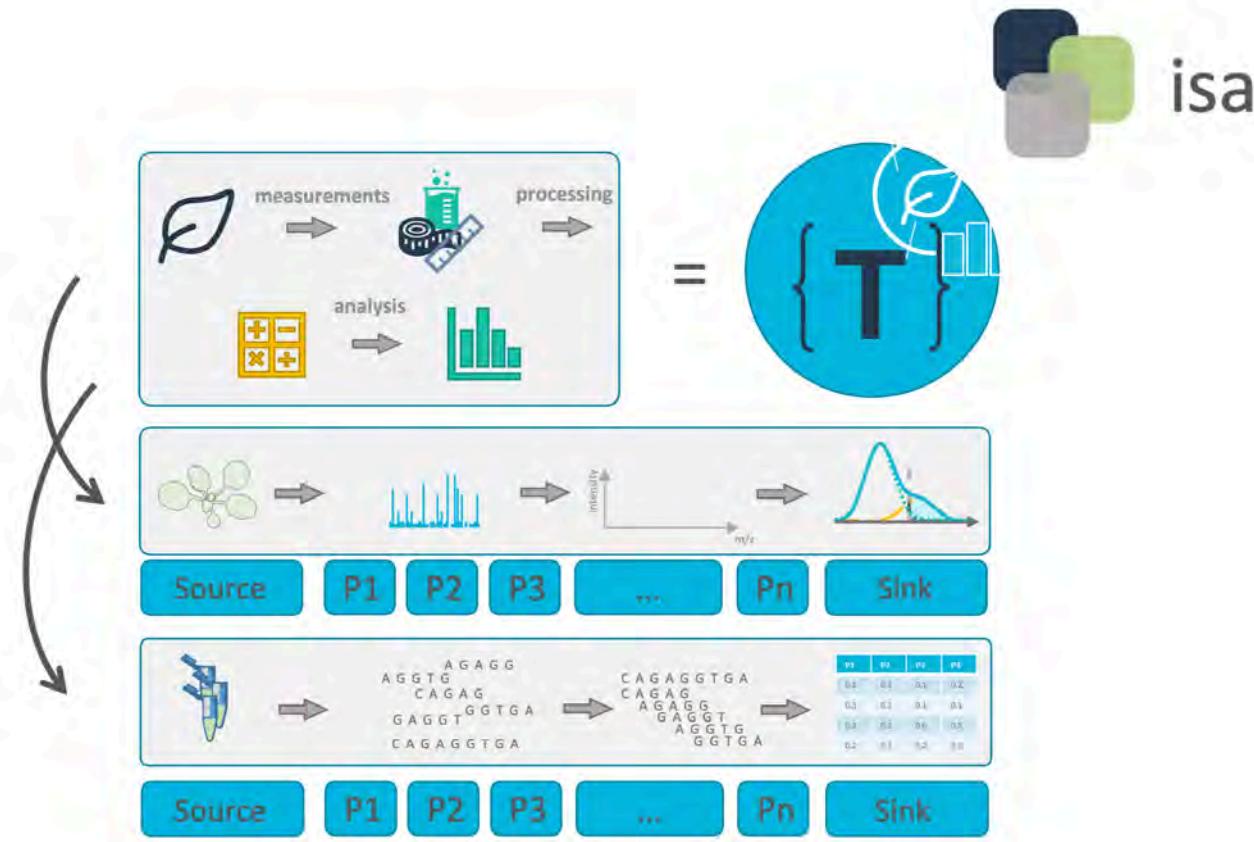


# Collect



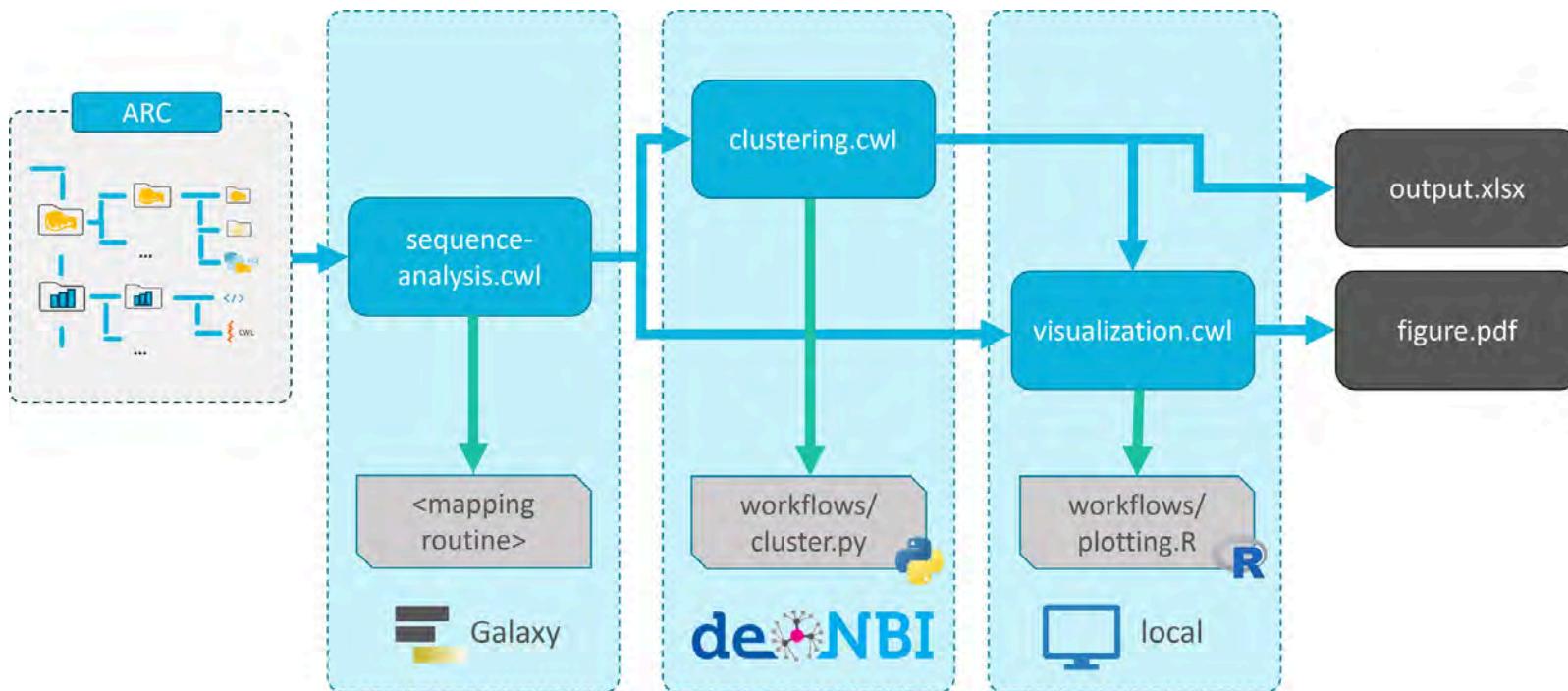


# Process (e.g. annotate)



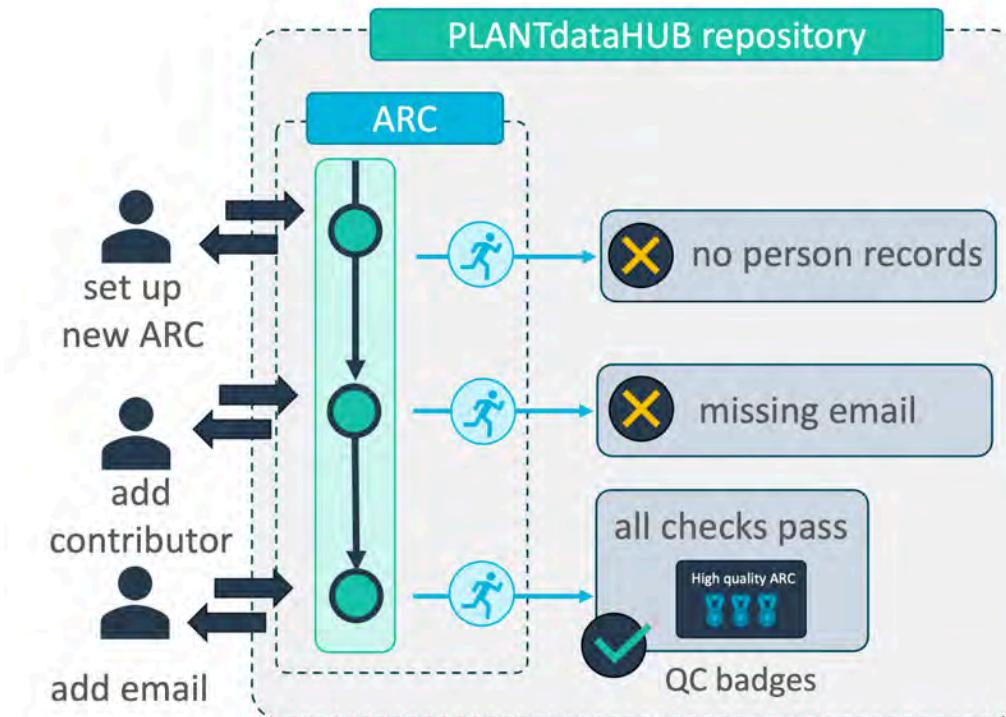


# Analyse





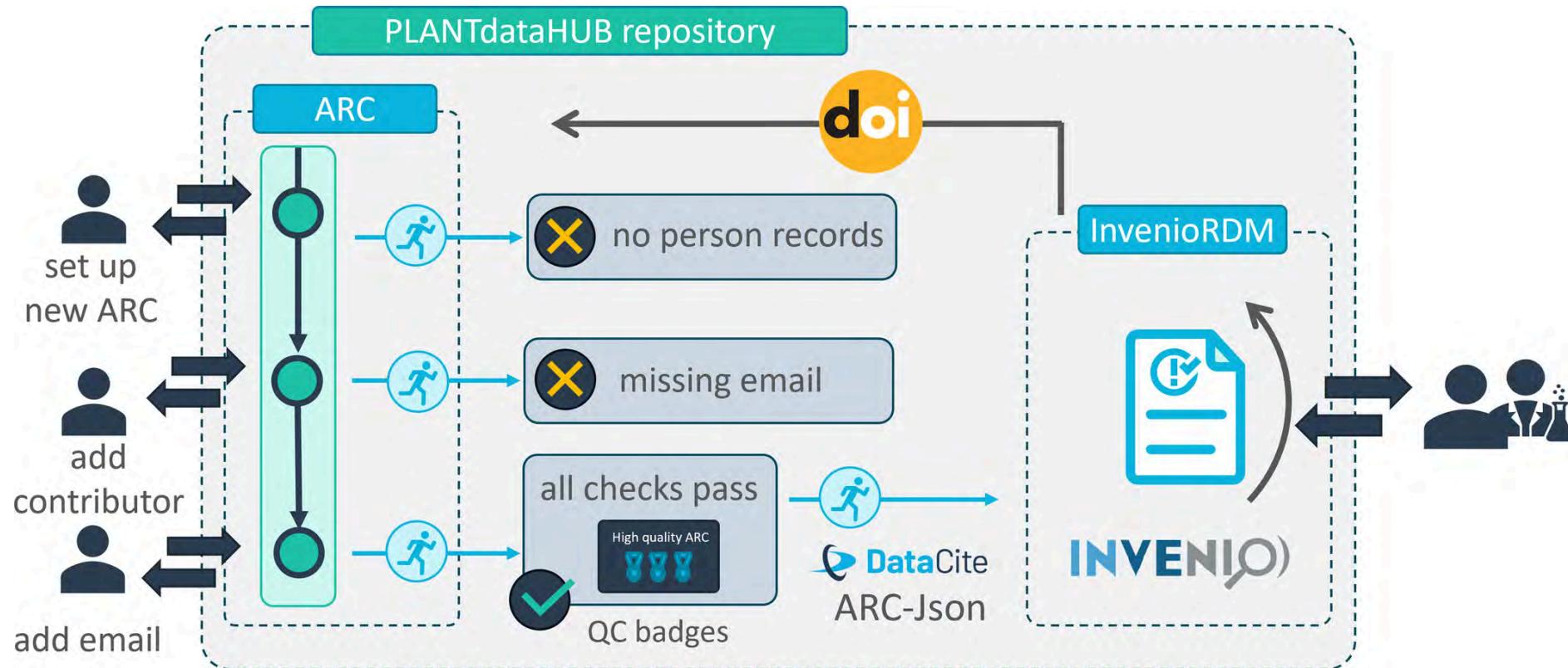
# Preserve



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



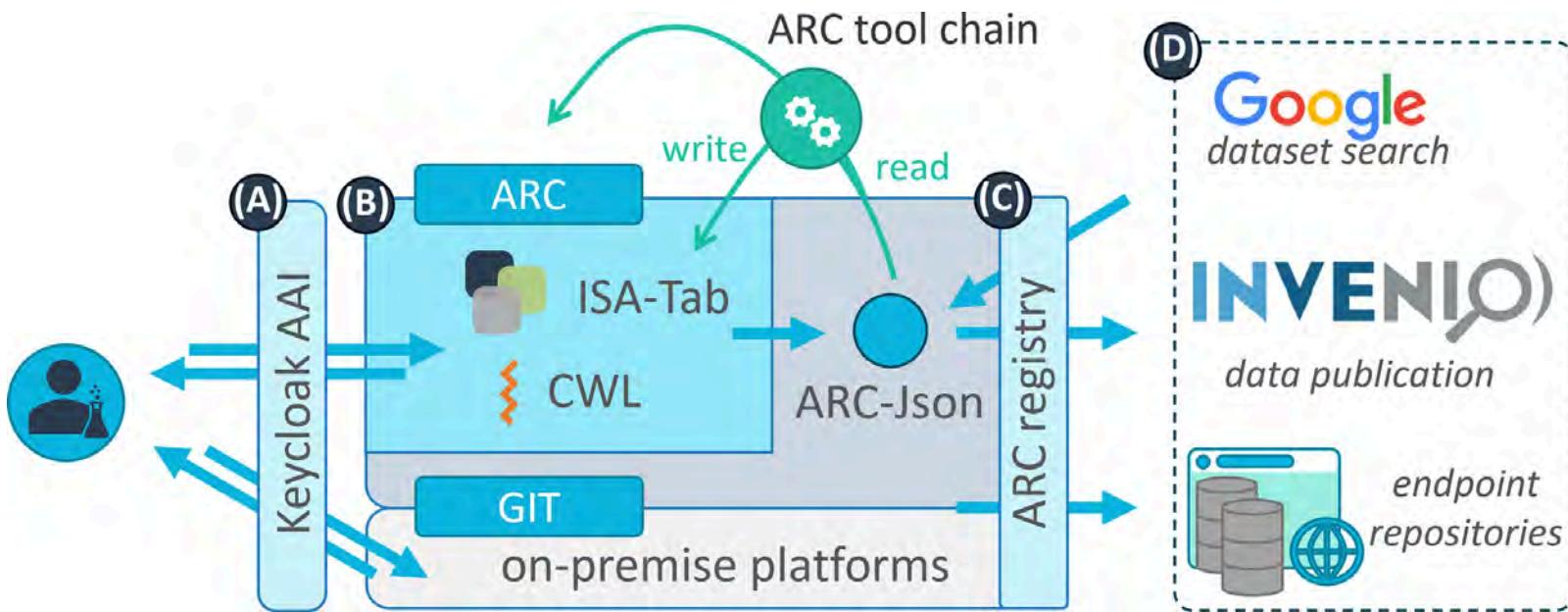


# Share and collaborate

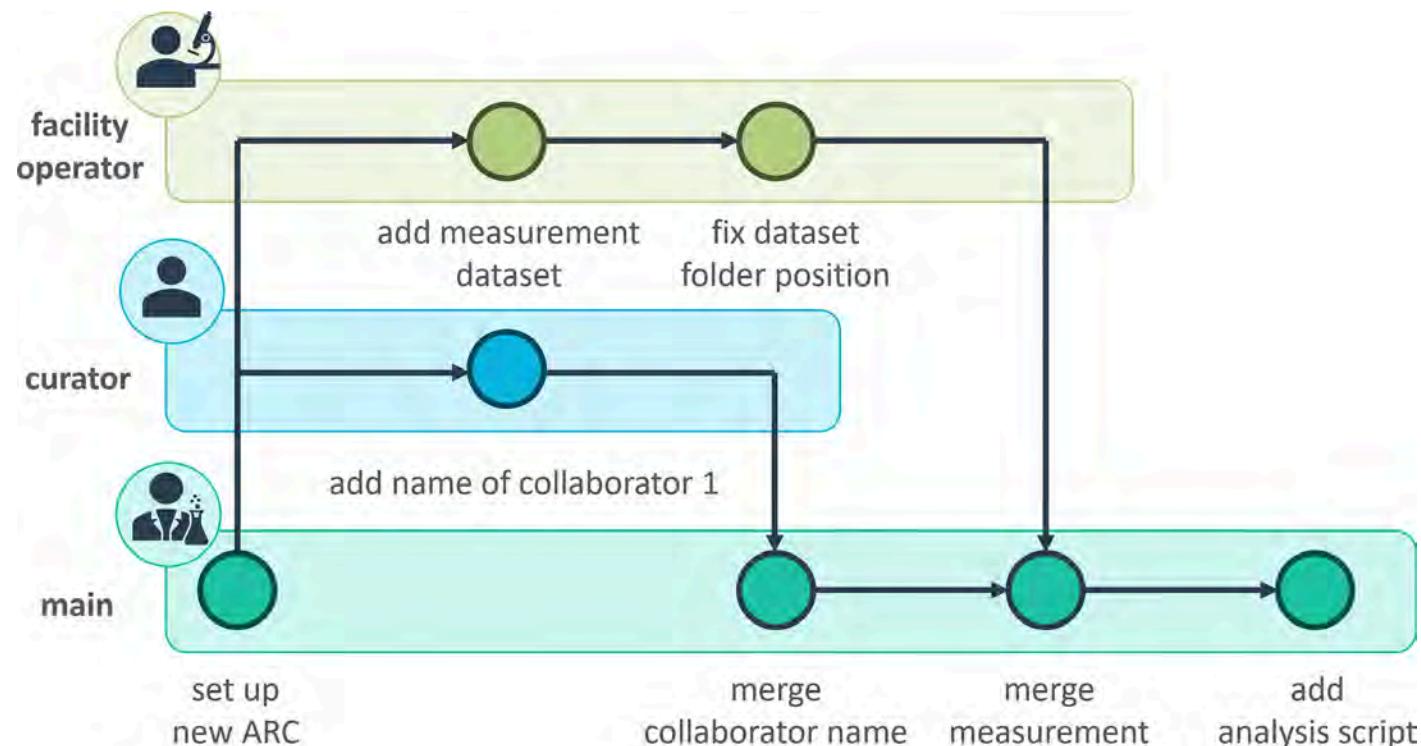




# Reuse

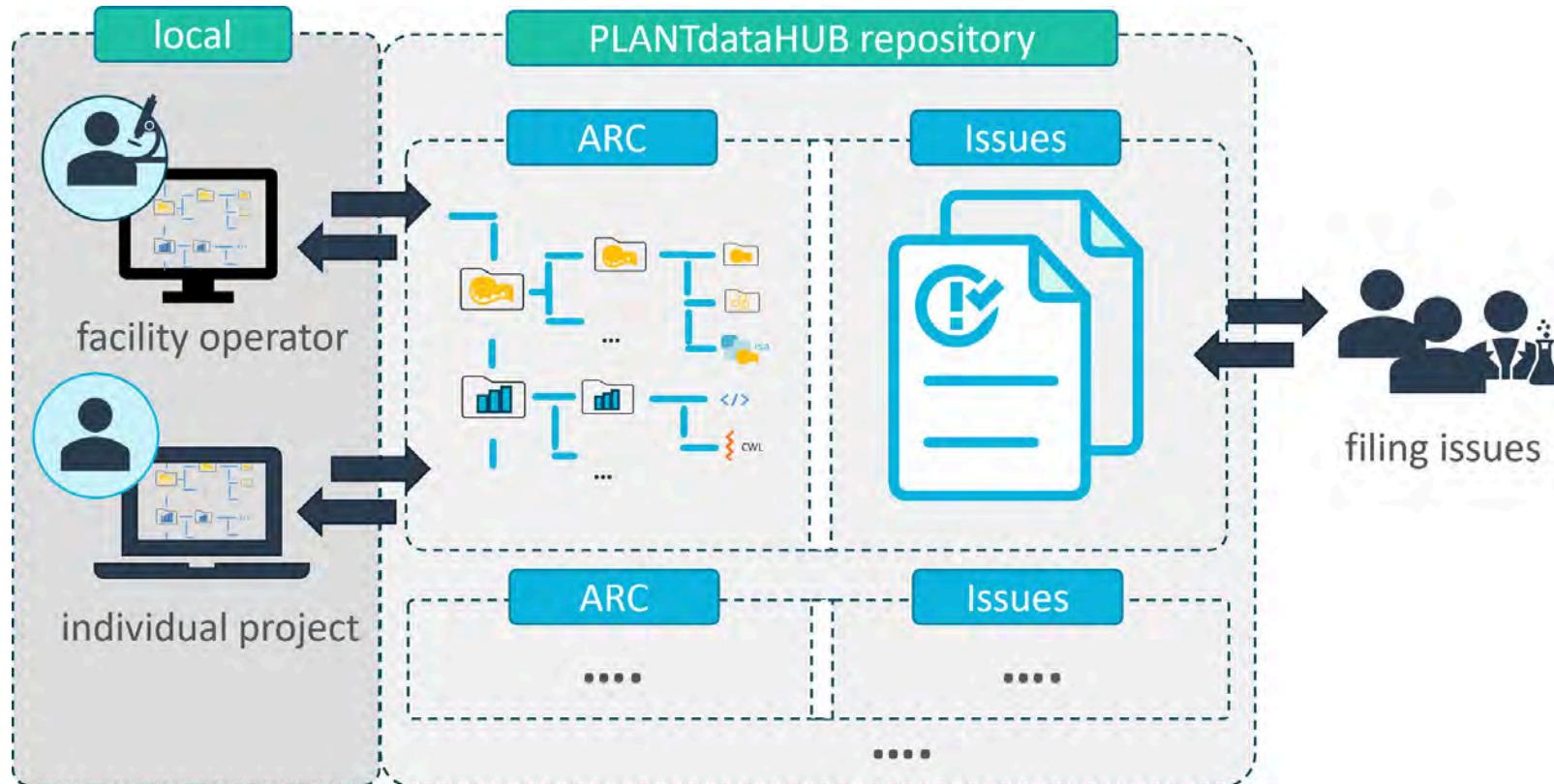


# Mutable data life cycle





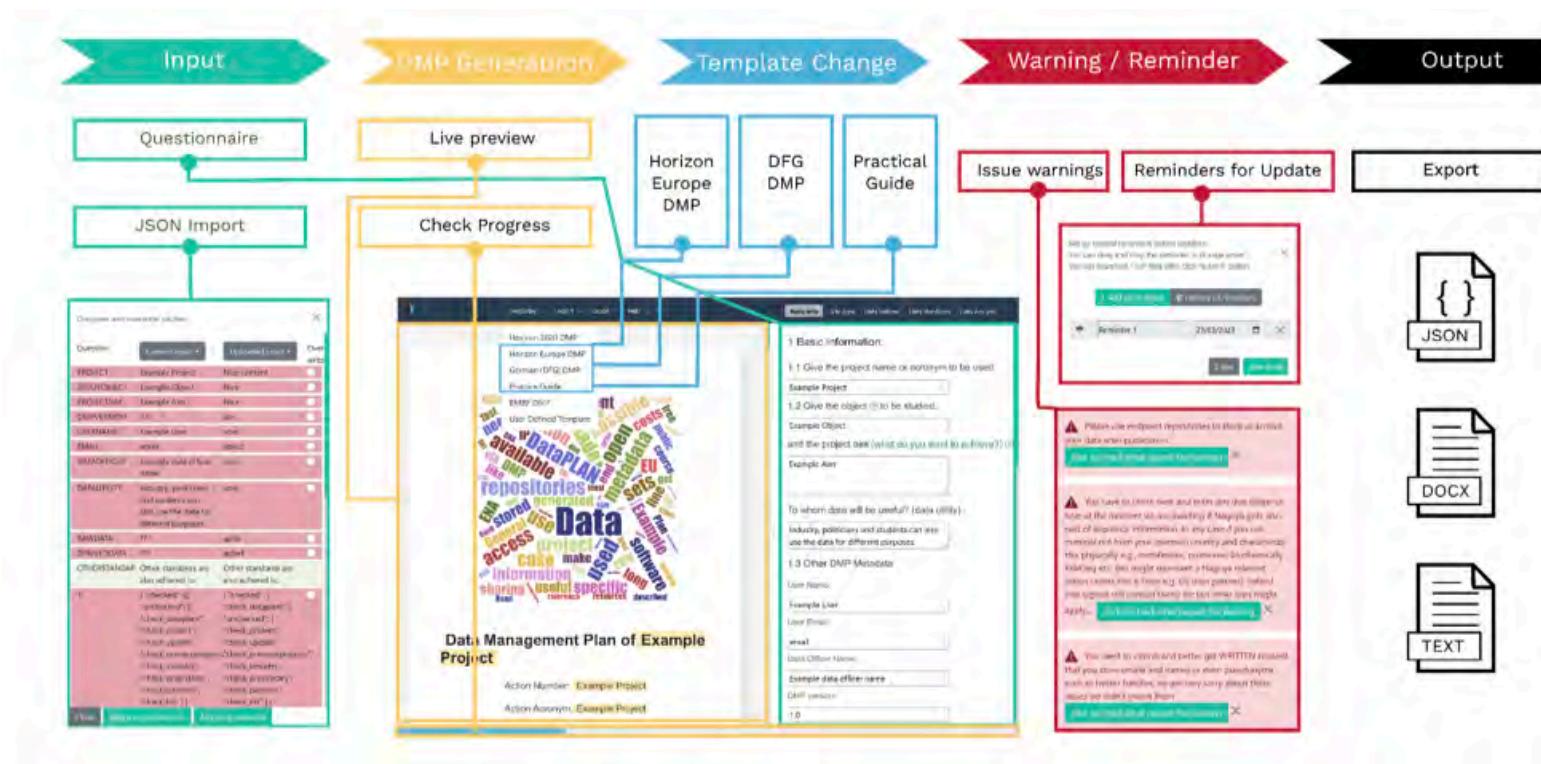
# Plan (ARC scale)





# 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>

# Check-in and ARC Commander Hands-on

Dominik Brilhaus – CEPLAS Data Science

# 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.)

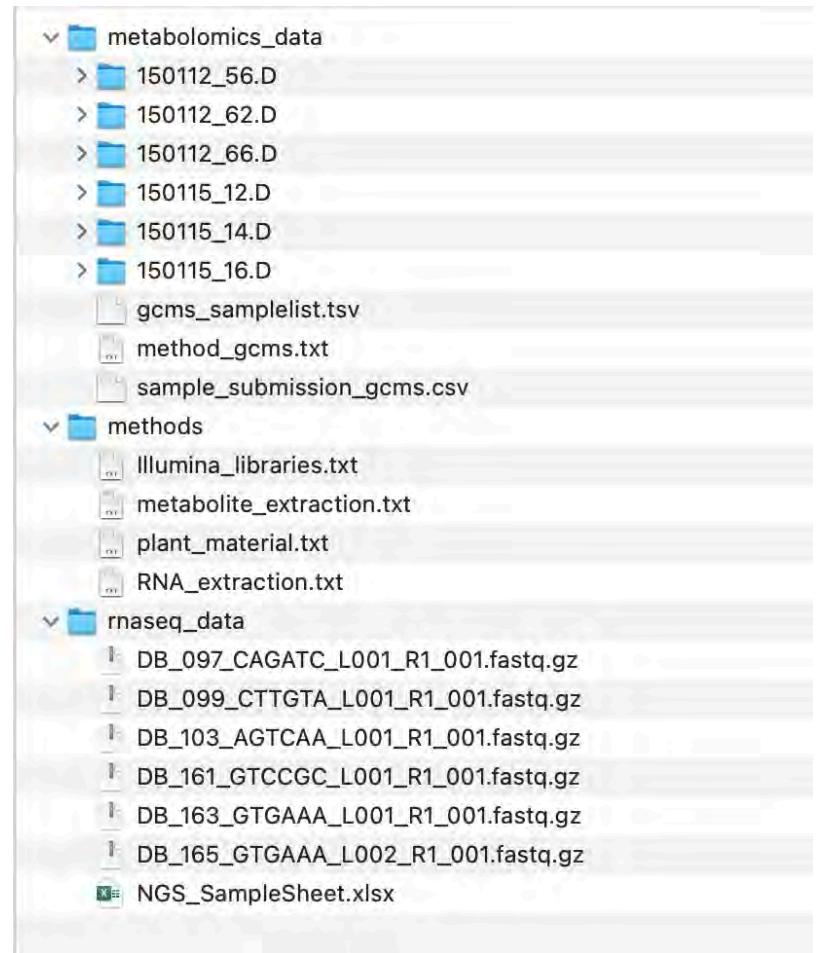
# Hands-on with demo data

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



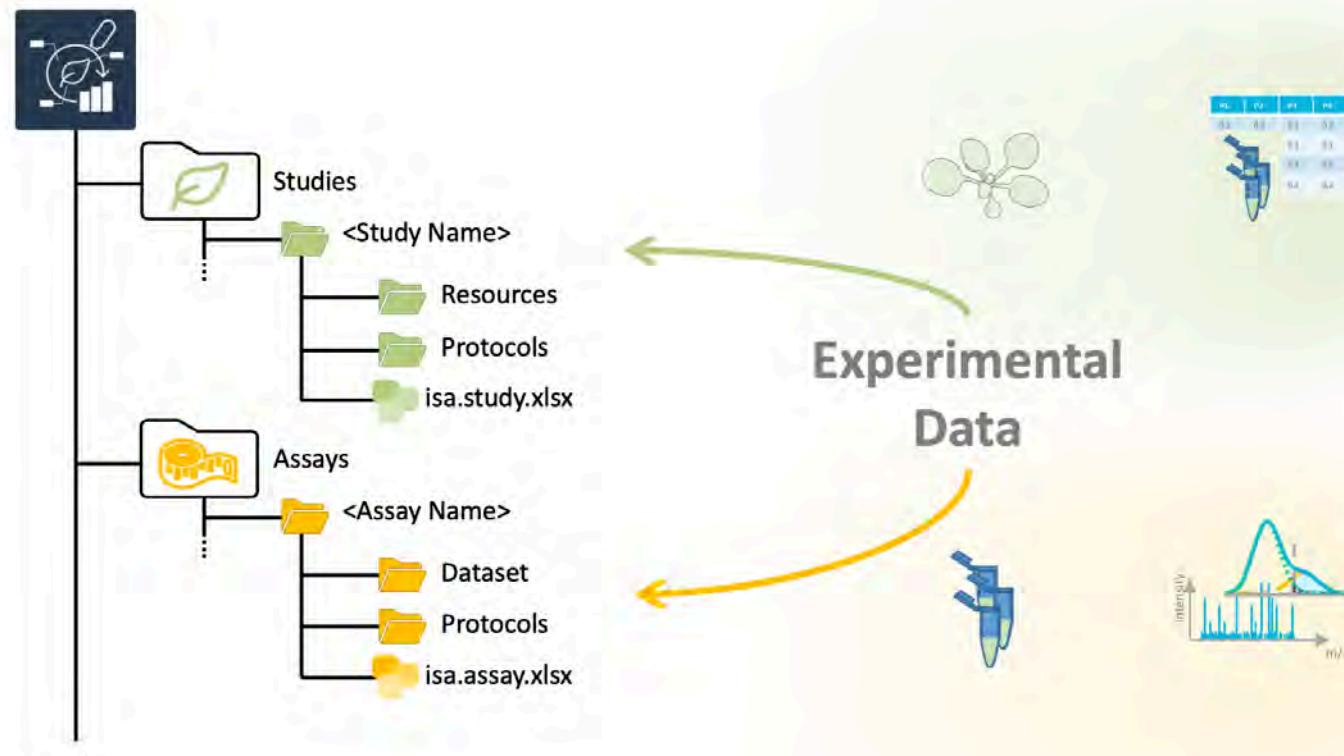
# Goal

- 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 --title TalinumPhotosynthesis --description "This is a very interesting investigation about life and photosynthesis"
```

## Add (at least one) person

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



For each person added, the minimum information is  
lastname | firstname | email | affiliation

## 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
```

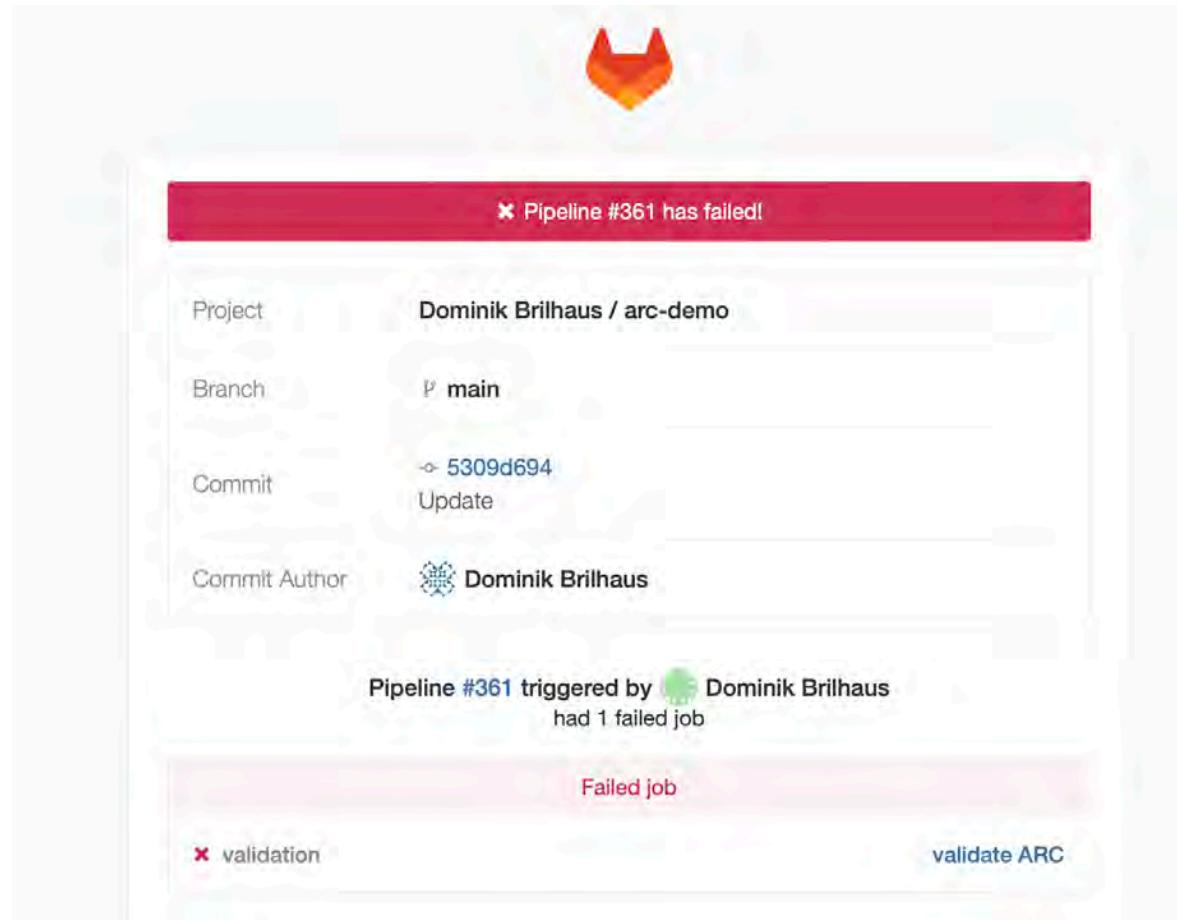
# Collaborate and share



# Upload your local ARC to the DataHUB

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

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

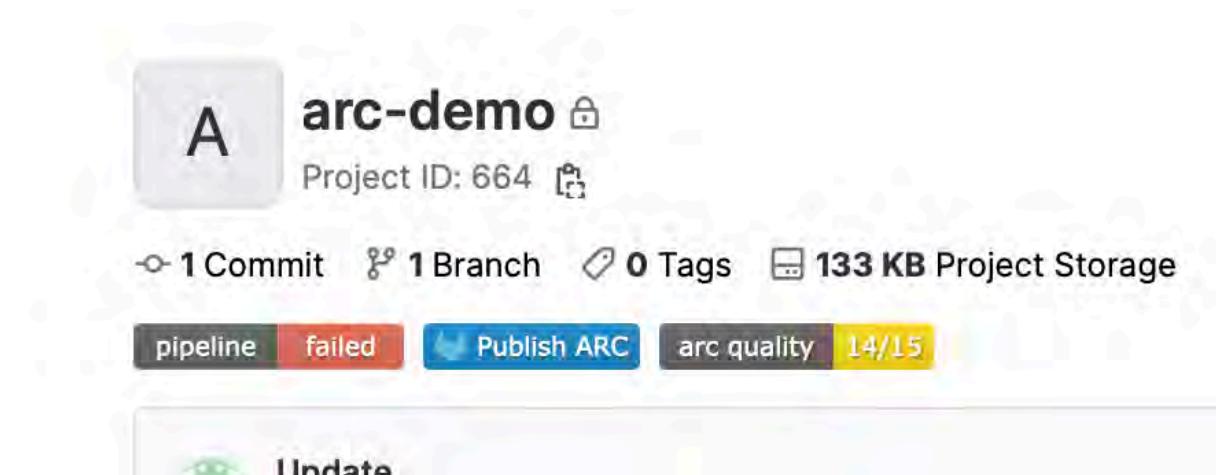


🔥 Don't worry 😊

# Pipeline Failed

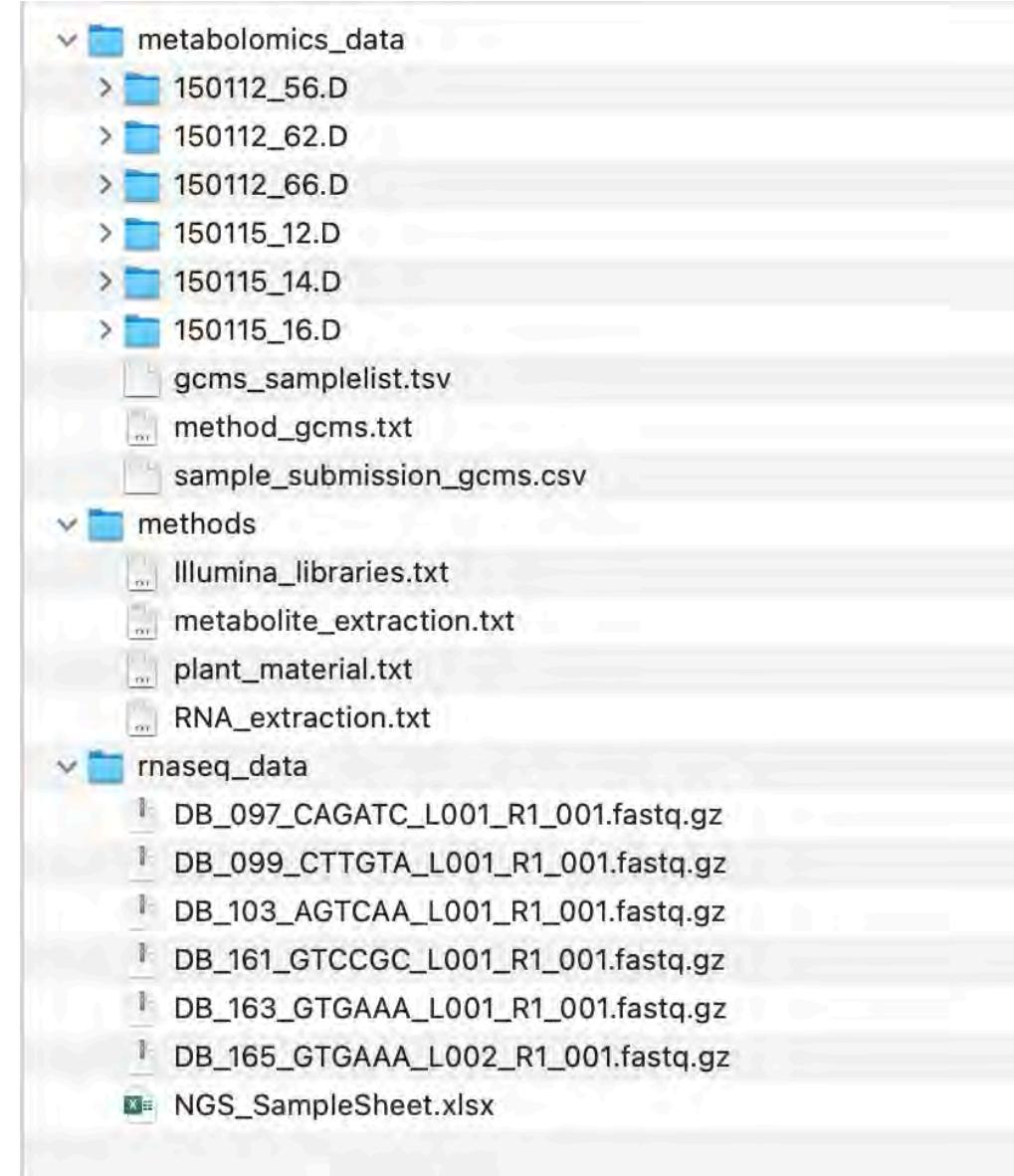
- a "continuous quality control" (CQC) pipeline validates your ARC
- This fails if one of the following metadata items is missing:

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



# 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

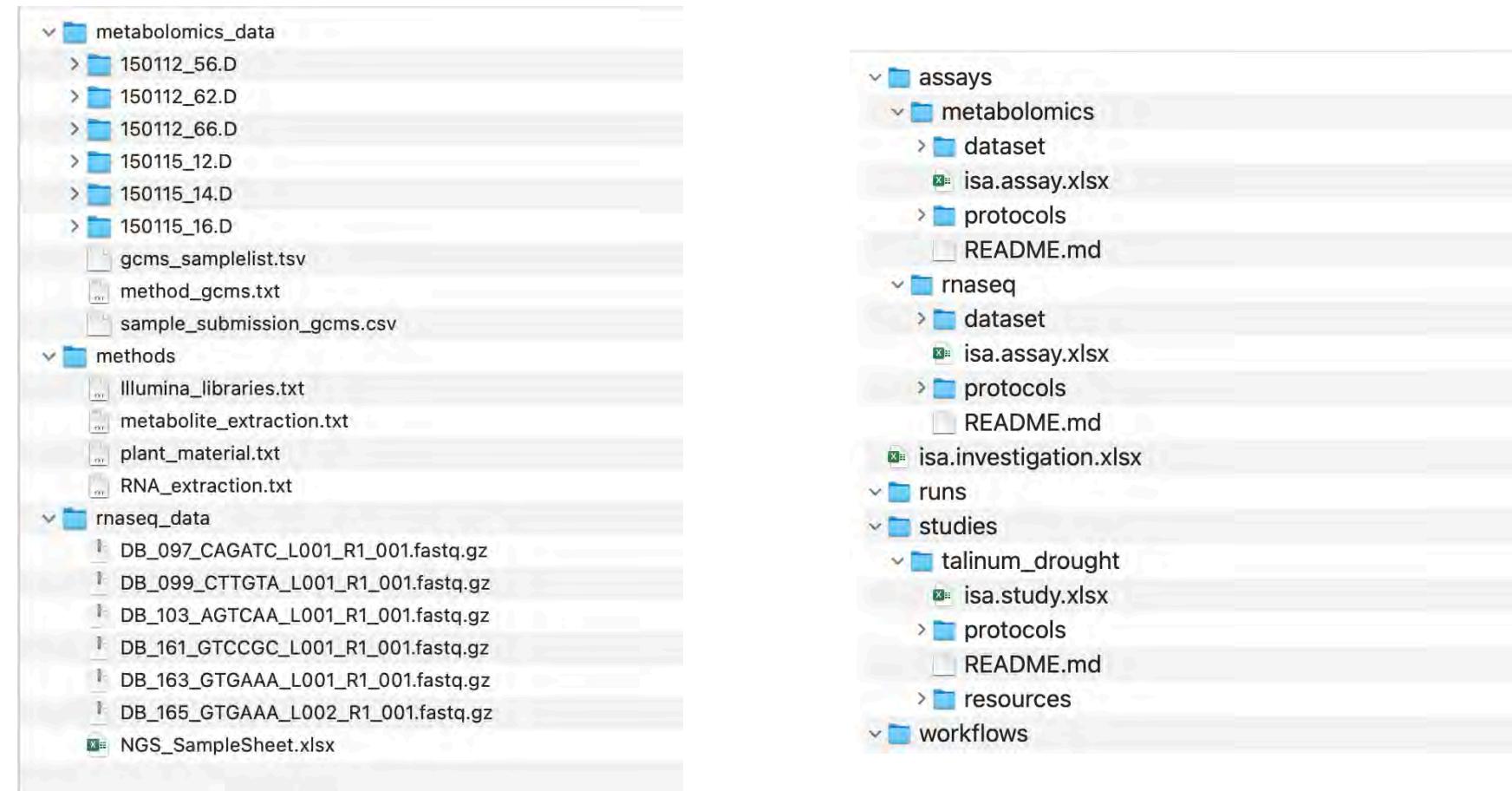
To save the changes, sync the ARC to the DataHUB including a message.

```
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>

## **Block 3 – ARCitect and hands-on**

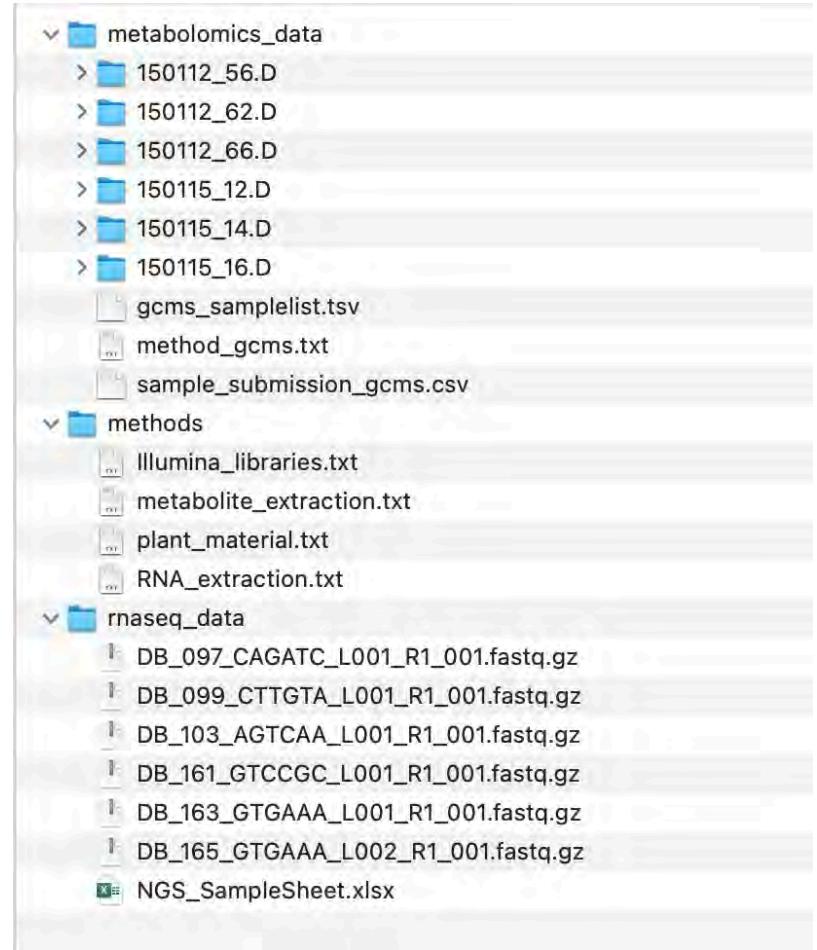
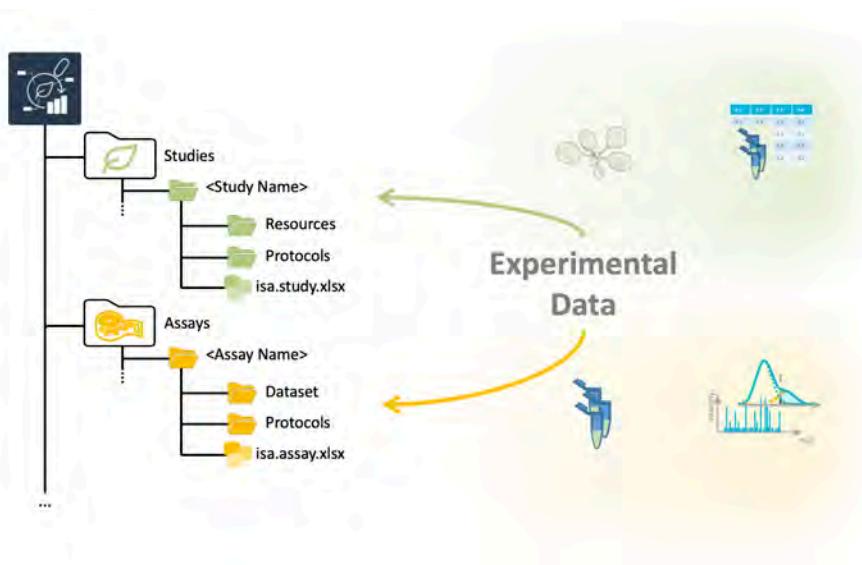
# ARCitect installation

Please install the latest version of the ARCitect: <https://github.com/nfdi4plants/ARCitect> 🔥 (released September 20th, 2023) 🔥

# Download the demo data

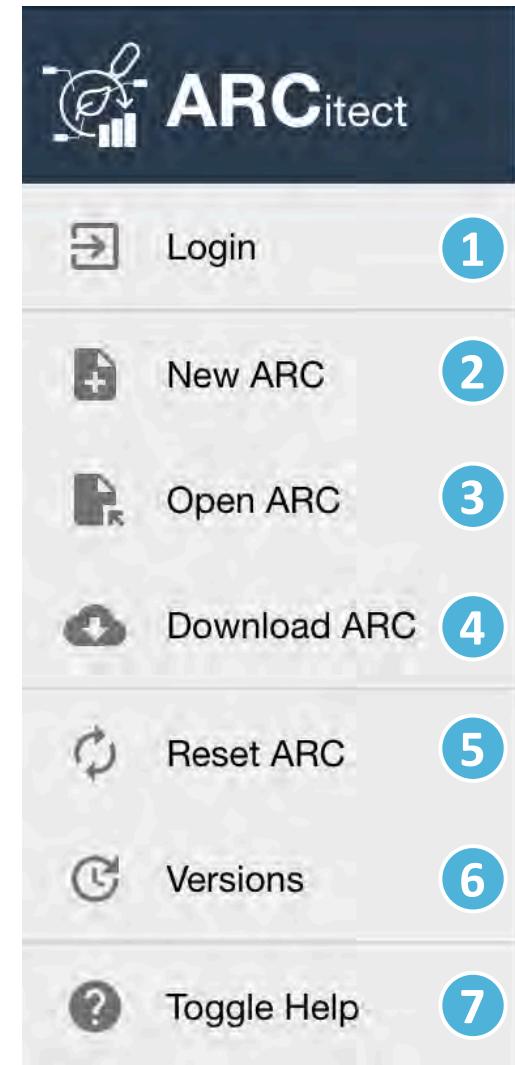
```
git clone "https://demo-user:1_eznikmzxzARAbUxxnF@git.nfdi4plants.org/teaching/demo-arc_level0.git"
```

# Sort Demo data in an ARC



# Open ARCitect

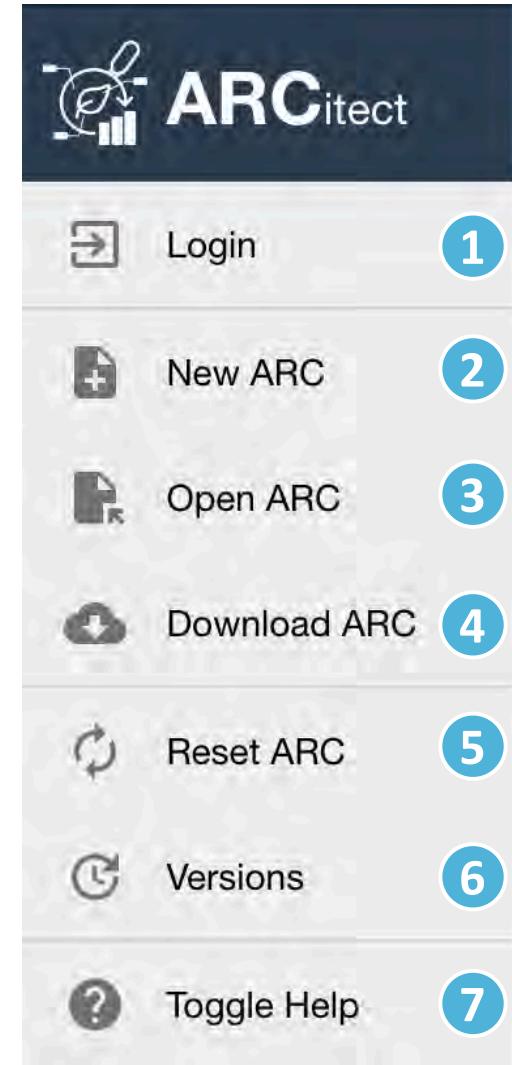
1. Login to DataHUB (1)



# Initiate the ARC folder structure

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

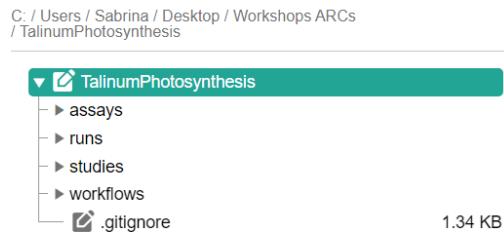
**TalinumPhotosynthesis**



# 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



A screenshot of an investigation creation form. The form includes fields for Identifier (TalinumPhotosynthesis), Title (Talinum Photosynthesis), and Description (This is a very interesting investigation about life and photosynthesis).

# Add (at least one) contributor

**Contacts**

Your First Name Your Last Name  
Your ORCID 6/10 ▾

<b>First Name</b>	<b>Last Name</b>
Your First Name	Your Last Name

<b>Mid Initials</b>	<b>ORCID</b>
	Your ORCID <span style="float: right;">Search</span>

<b>Affiliation</b>	<b>Address</b>
Your Affiliation	

<b>Email</b>	<b>Phone</b>	<b>Fax</b>
yourEmailAdress@uni.de		

**Roles**

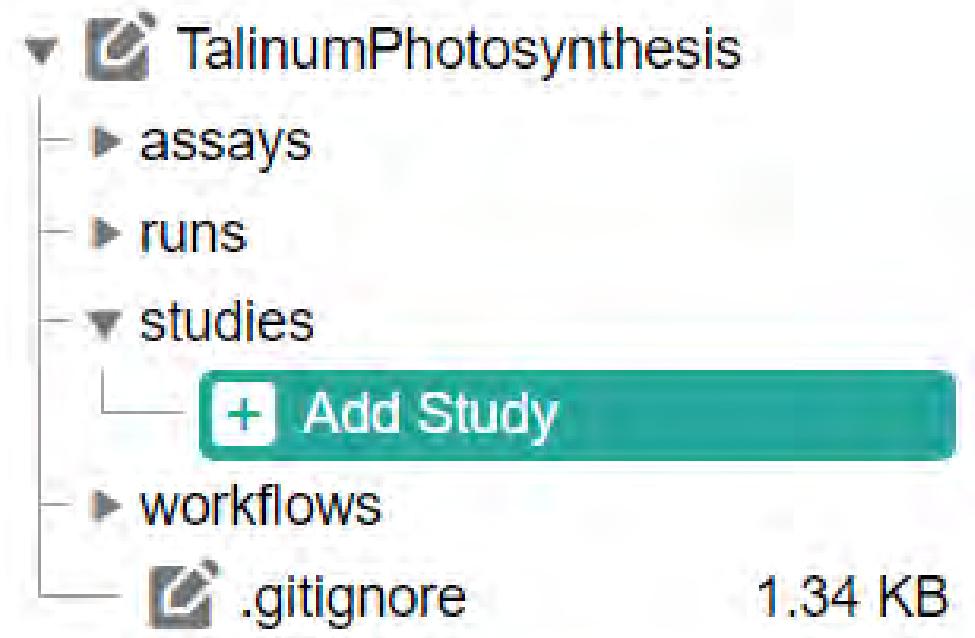
1. Author ✓ NCIT NCIT:C42781 X

+ 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 ReleaseDate

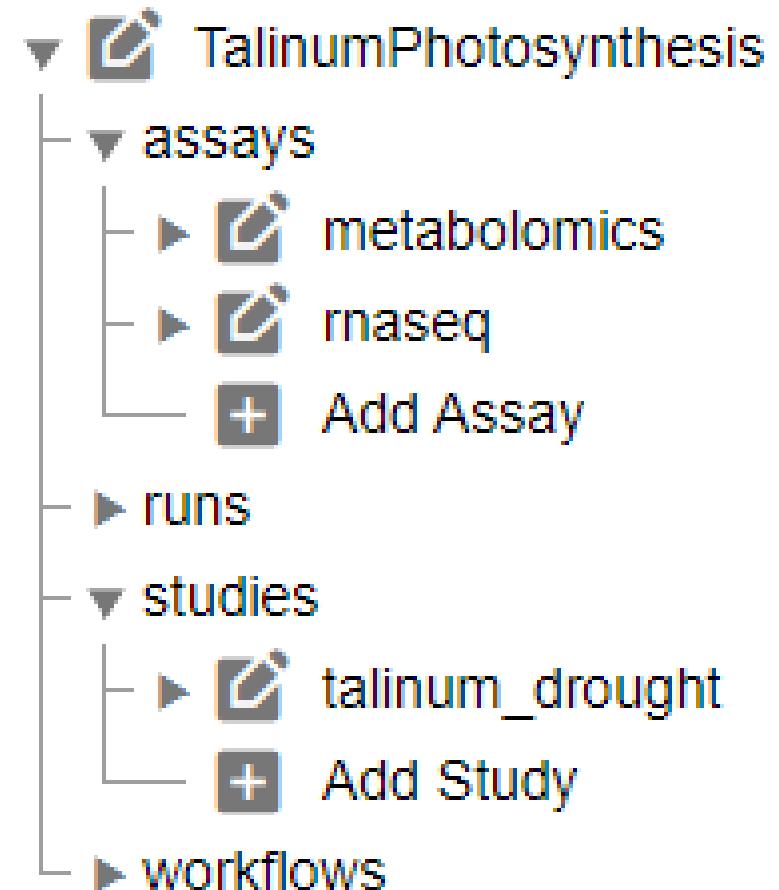
 

Study Design Descriptors

# Add an assay

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

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

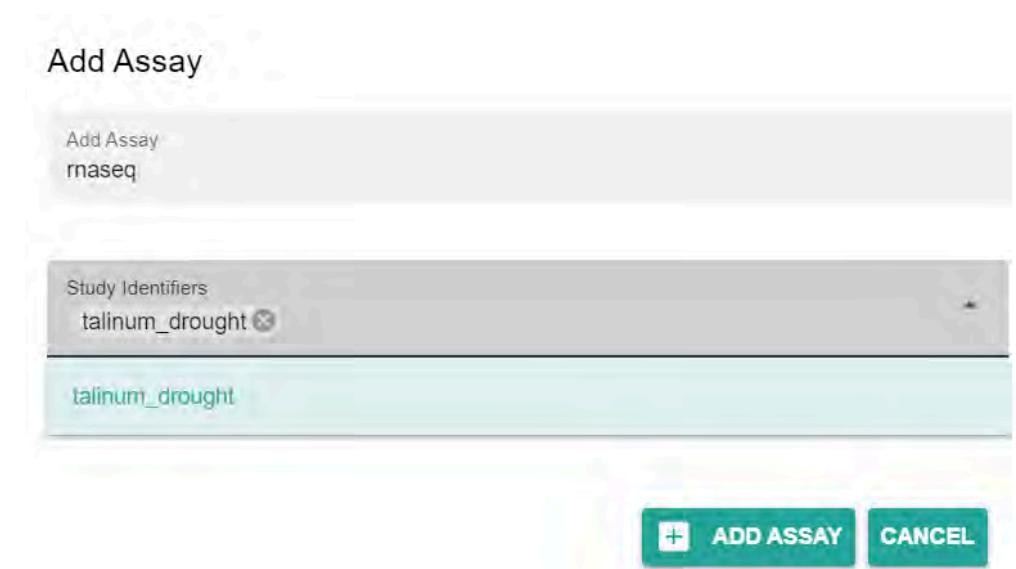


# Link your assay to a study

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 information about your assay

In the assay panel you can

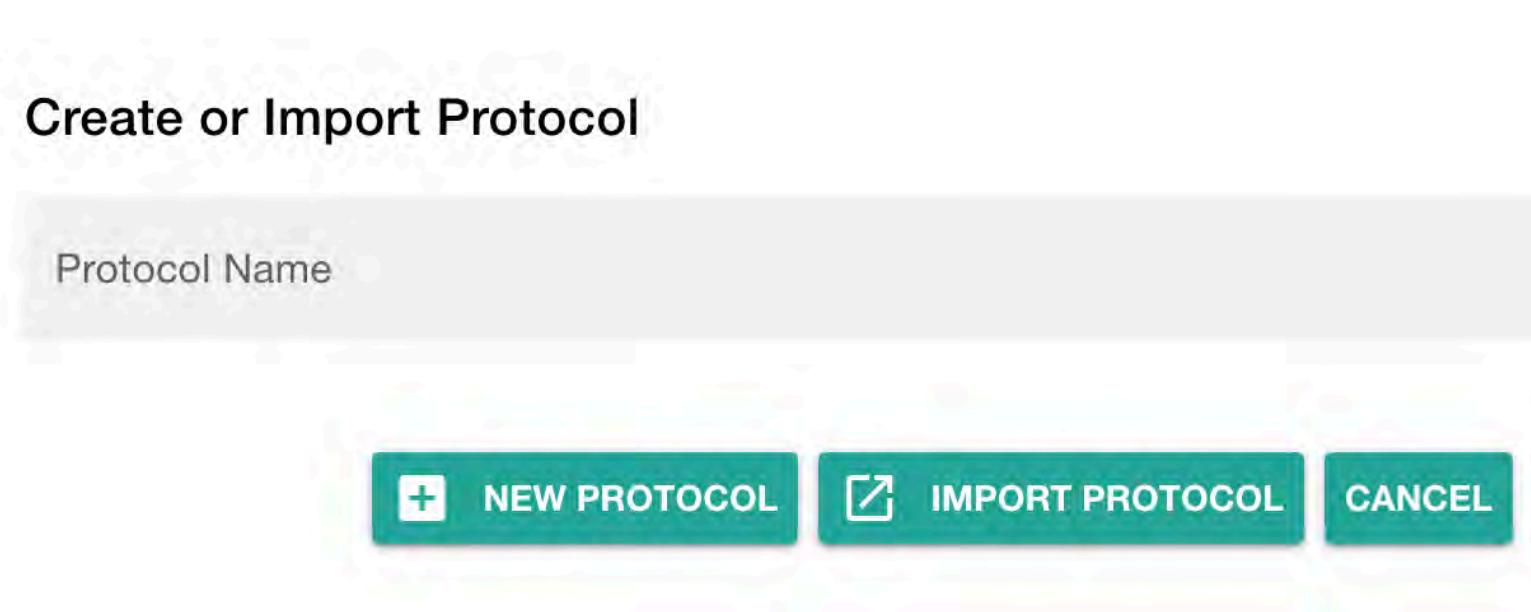
1. link or unlink the assay to studies, and
2. define the assay's
  - measurement type
  - technology type, and
  - technology platform.
3. add data process information

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



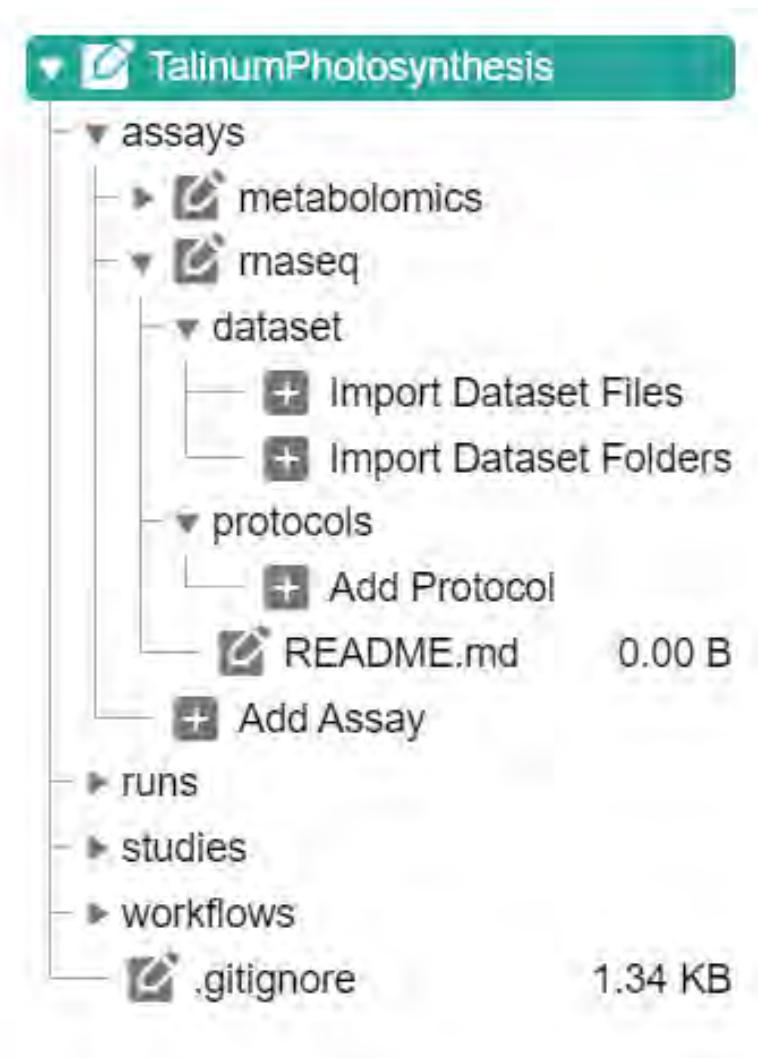
# Add protocols and datasets

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.



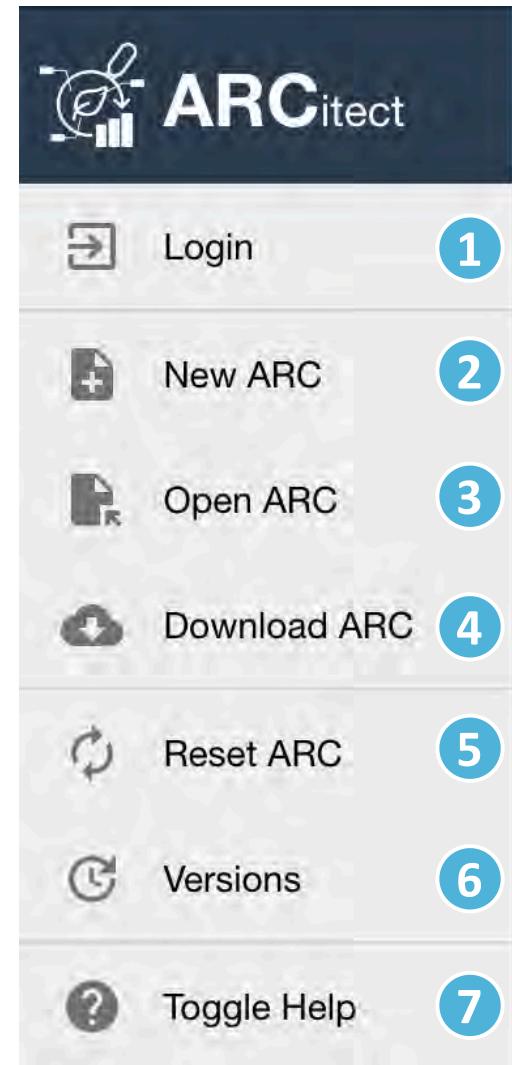
# Sort Demo Data to your ARC

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

# Login to the DataHUB

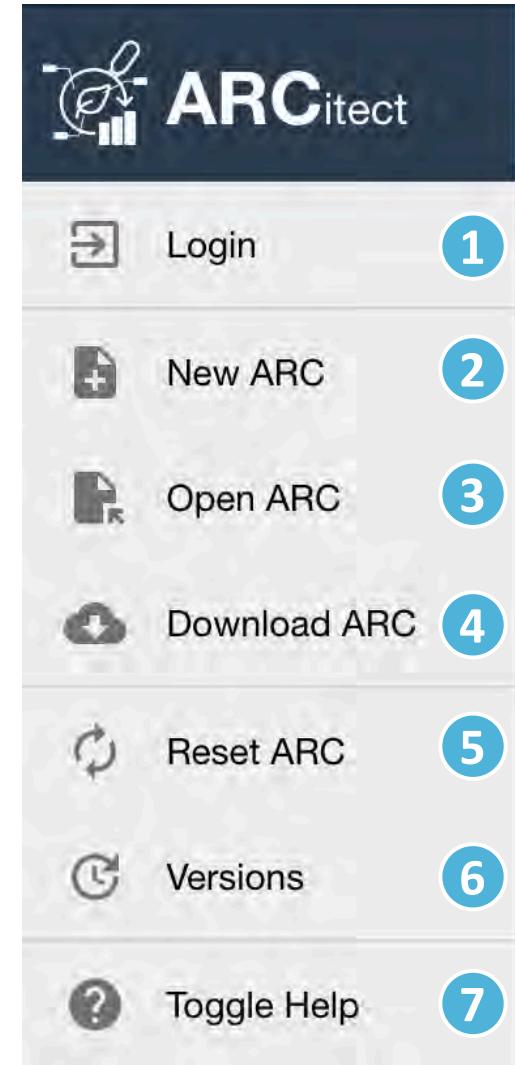
Click **Login** (1) in the sidebar to login to the DataHUB.

 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

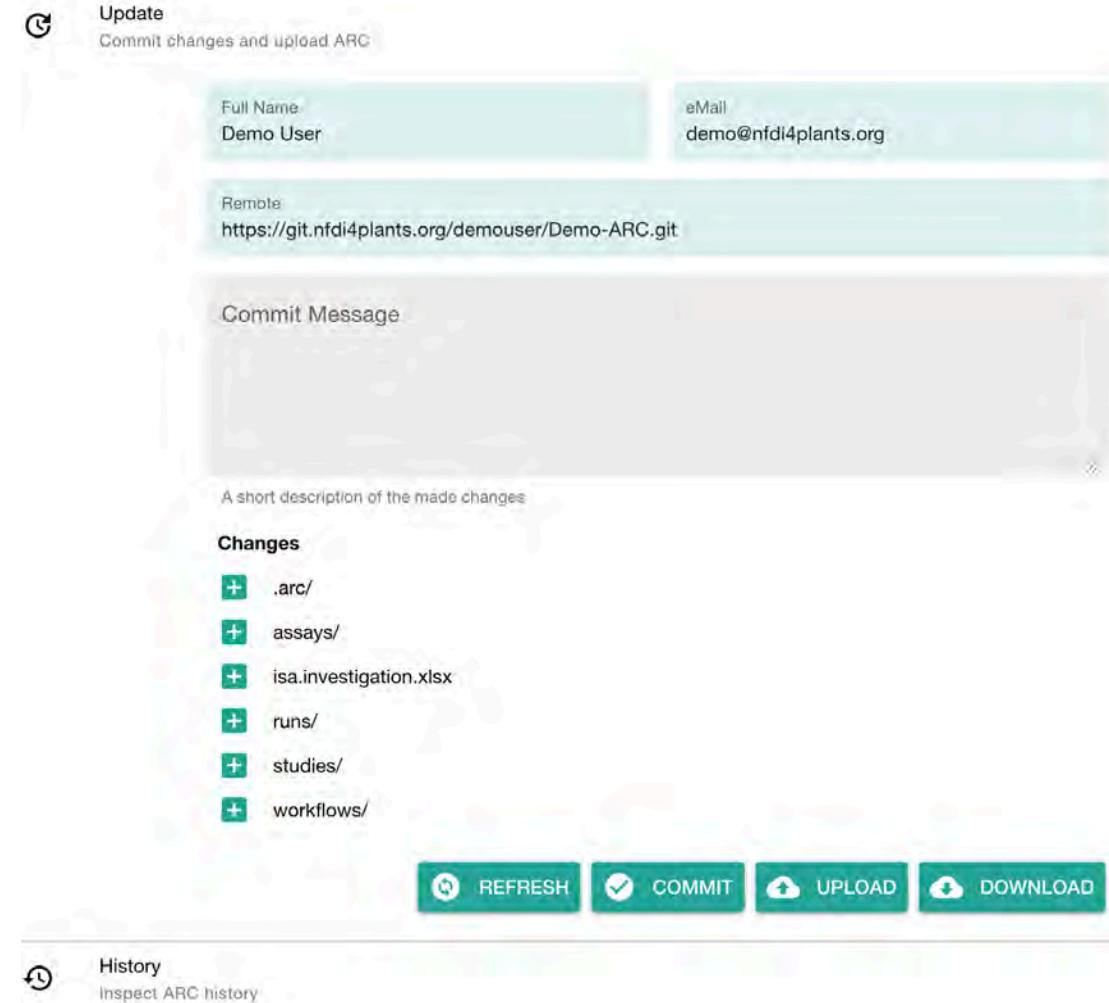
The screenshot shows the 'Update' section of the DataHUB interface. It includes fields for 'Full Name' (Demo User), 'eMail' (demo@nfdi4plants.org), and 'Remote' (https://git.nfdi4plants.org/demouser/Demo-ARC.git). A 'Commit Message' field is present with placeholder text: 'A short description of the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

# 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

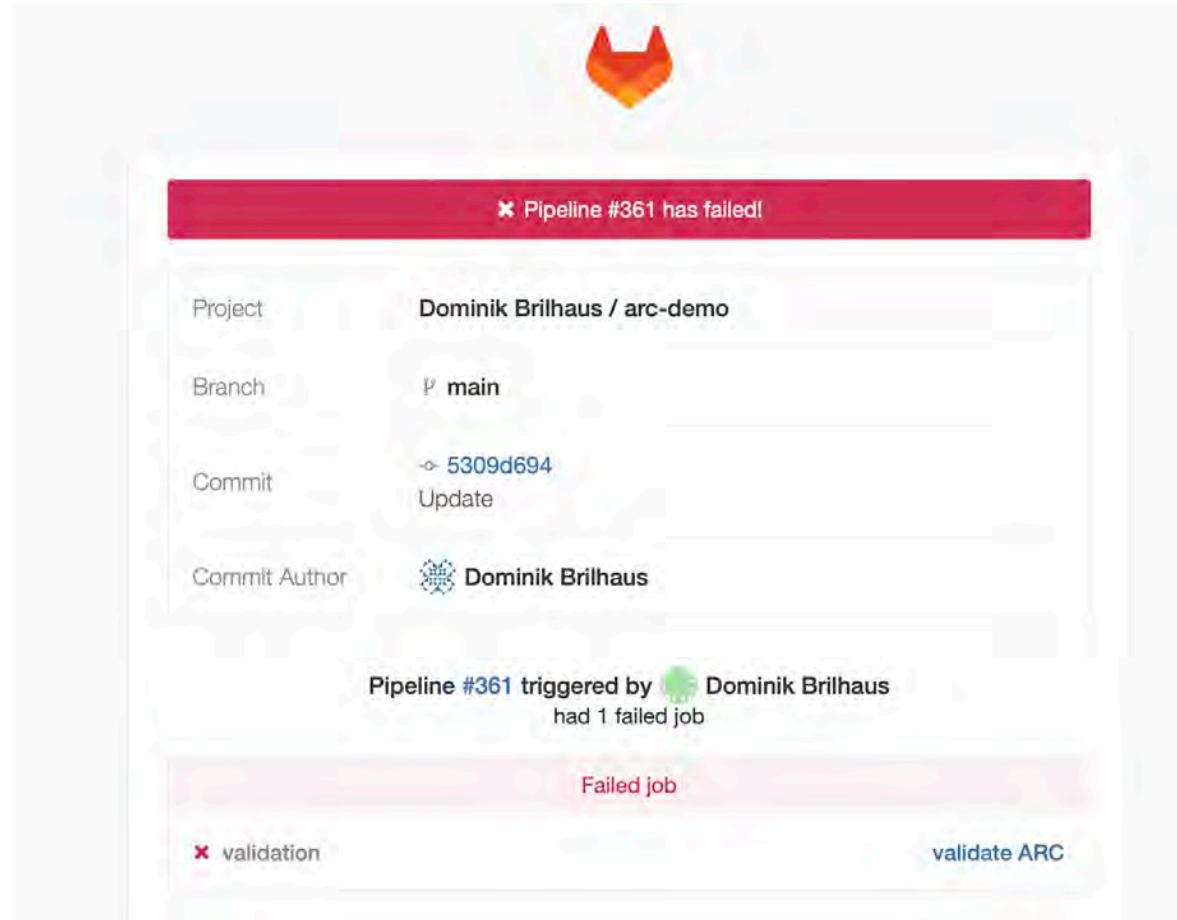
<https://git.nfdi4plants.org/<YourUserName>/<YourARC>>



# Check if your ARC is successfully uploaded

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

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

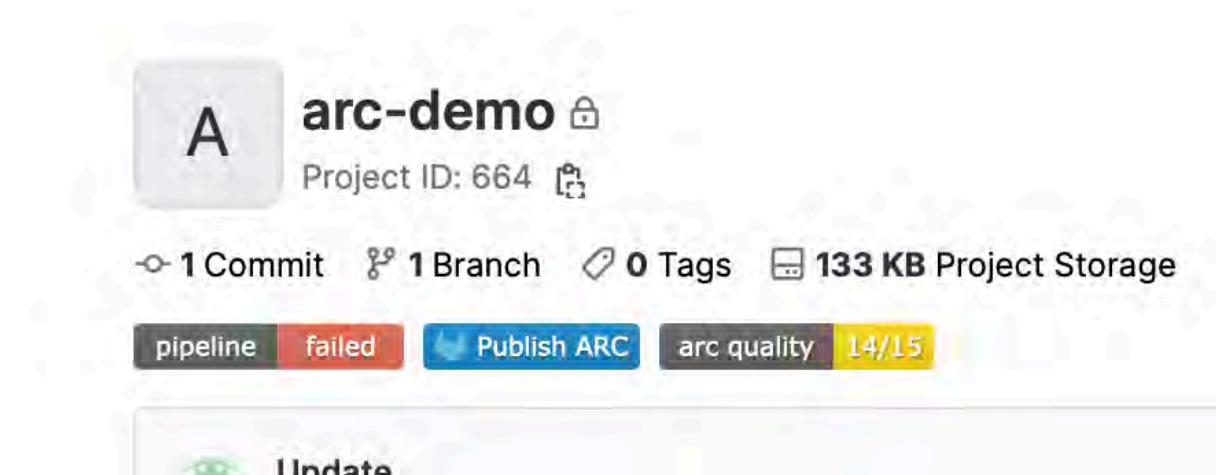


🔥 Don't worry 😊

# Pipeline Failed

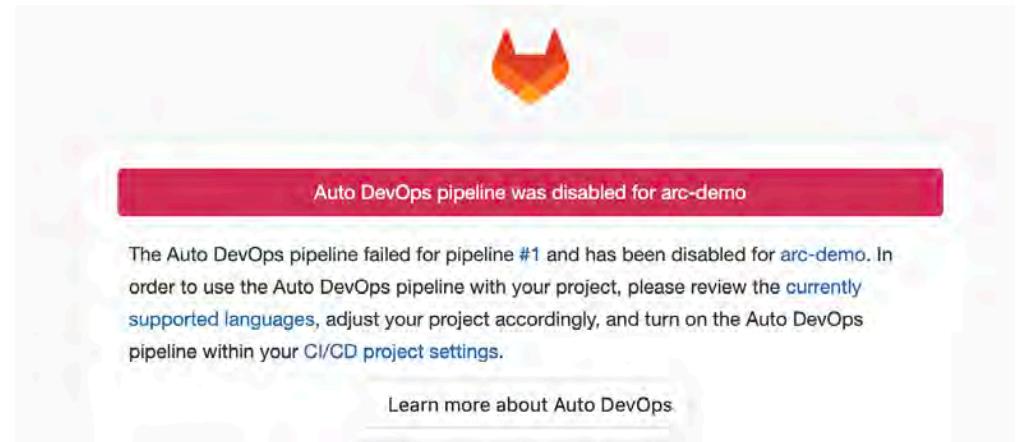
- a "continuous quality control" (CQC) pipeline validates your ARC
- This fails if one of the following metadata items is missing:

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



# 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 'CI/CD' settings page in GitLab. On the left, there is a sidebar with various project management and monitoring tools like Security & Compliance, Deployments, Packages and registries, Infrastructure, Monitor, Analytics, Wiki, Snippets, Settings, General, Integrations, Webhooks, Access Tokens, Repository, Merge requests, and CI/CD. The 'CI/CD' option is currently selected. The main content area is titled 'Auto DevOps' with the sub-instruction 'Automate building, testing, and deploying your applications based on your continuous integration and delivery configuration.' Below this, there is a note 'How do I get started?' and a checked checkbox for 'Default to Auto DevOps pipeline' with the status 'instance enabled'. A note says 'The Auto DevOps pipeline runs if no alternative CI configuration file is found.' There is also a link to 'Learn more'. A yellow callout box suggests adding a 'Kubernetes cluster integration' or creating an environment variable. Under 'Deployment strategy', there are 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, there is a 'Save changes' button.

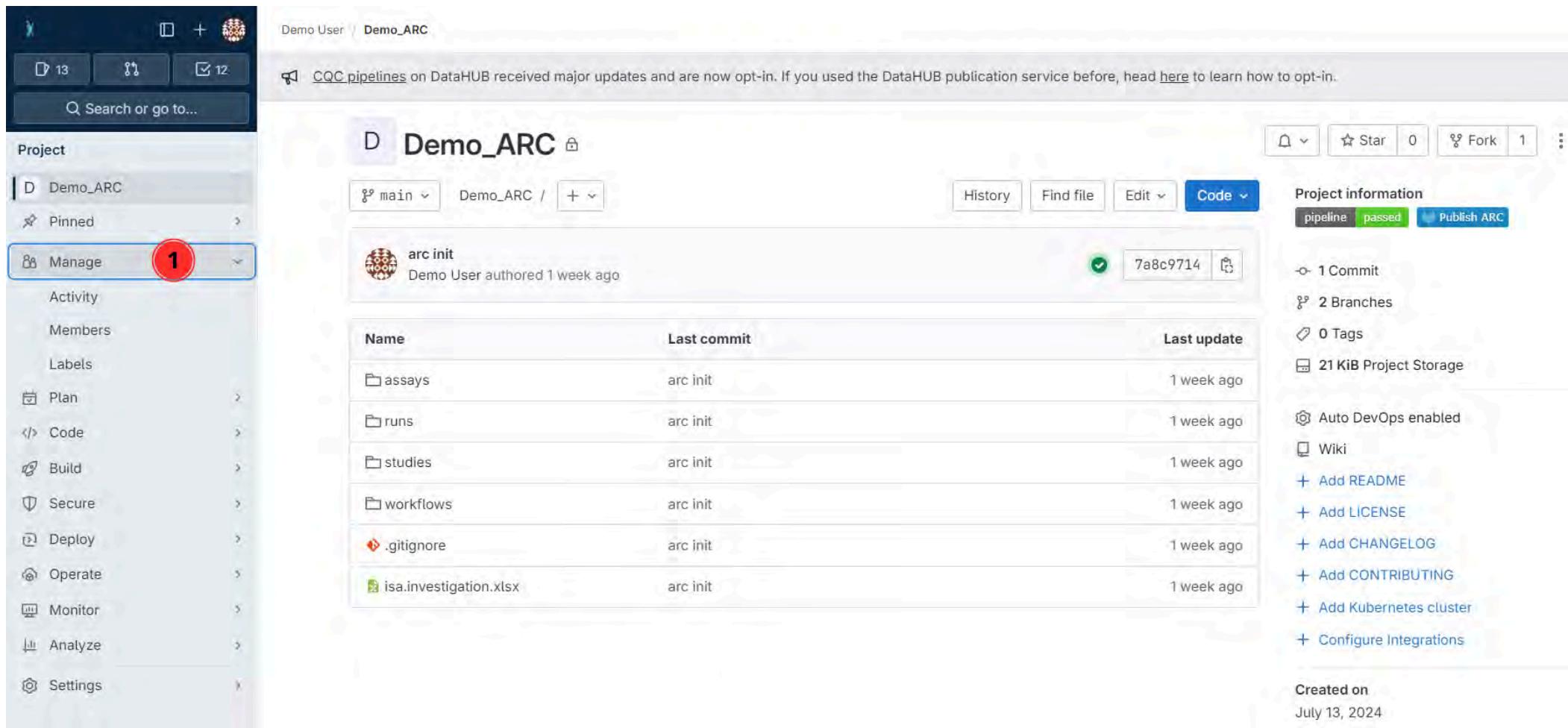
# Collaborate and share



# 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 interface for the 'Demo\_ARC' project. The left sidebar has a 'Manage' button highlighted with a red circle containing the number 1. The main area displays a list of files and their commit history. On the right, there is a 'Project information' sidebar with various details and action buttons.

**Project Information:**

- pipeline: passed
- Publish ARC
- 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

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

## 2. Click on Members

The screenshot shows the DataHUB interface with the following details:

- Project Sidebar:** On the left, there is a sidebar with various project management options: Demo ARC (selected), Pinned, Manage (circled with red number 1), Activity, Members (circled with red number 2), Labels, Plan, Code, Build, Secure, Deploy, Operate, Monitor, Analyze, and Settings.
- Header:** The header shows the path: Demo User / Demo\_ARC / Members. A message about CQC pipelines opt-in is displayed.
- Project members:** The main content area is titled "Project members". It says "You can invite a new member to Demo\_ARC or invite another group." and shows a table of members.
- Table Headers:** The table has columns: Account, Source, Max role, Expiration, and Activity.
- Table Data:** One member is listed:

Account	Source	Max role	Expiration	Activity
Demo User @DemoUser	Direct member by Demo User	Owner	Expiration date 8+ Sep 27, 2023 ✓ Jul 13, 2024 ✗ Jul 21, 2024	
- Buttons:** At the top right, there are buttons for "Import from a project", "Invite a group", and "Invite members".

### 3. Click on Invite members

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 <span style="background-color: #00AEEF; color: white; padding: 2px 5px;">It's you</span>	Direct member by Demo User	Owner	Expiration date <input type="button" value="Sep 27, 2023"/>	8+ Sep 27, 2023 ✓ Jul 13, 2024 ✗ Jul 21, 2024

Import from a project

#### 4. Search for potential collaborators

**Invite members** X

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

The screenshot shows a user interface for selecting a role. On the left, a sidebar lists five roles: Guest, Reporter, Developer, Maintainer, and Owner. The 'Guest' role is currently selected, indicated by a blue highlight and a checked checkbox icon. To the right of the sidebar, there is a large, semi-transparent overlay window. Inside this window, the word 'ARC project.' is visible. Overlaid on the 'Developer' role in the sidebar is a red circle containing the number '4'. Overlaid on the 'Owner' role is another red circle containing the number '5'. At the bottom of the sidebar, there is a dropdown menu also set to 'Guest', which also has a red circle with the number '5' overlaid on it. At the very bottom of the screen, there is a link labeled 'Read more about role permissions'.

Guest

Reporter

Developer

Maintainer

Owner

ARC project.

4

5

Guest

5

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.

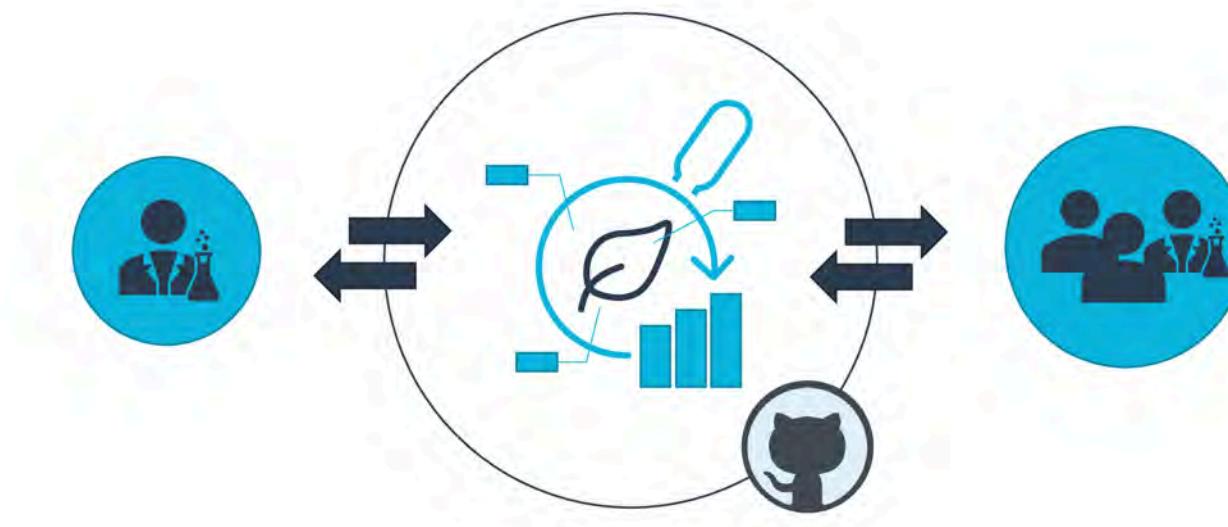
## Maintainers

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.



# Your ARC is ready

 Initiated an ARC

 Structured and ...

 ... annotated experimental data

 Shared with collaborators





# 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>

# Block 5 – Metadata and ISA

September 28th, 2023



Sabrina Zander  
[MibiNet](#)



Dominik Brilhaus  
[CEPLAS Data Science](#)

**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"?!

The screenshot shows a Google search results page for the query "pulp fiction". The search bar at the top contains the query. Below it, a navigation bar offers filters like Bilder, Videos, Cast, Bedeutung, Handlung, Hinkebein, Netflix, Soundtrack, and Tanz. The main search results include a summary card for the movie "Pulp Fiction" (1994), which has a rating of FSK 16, 2 hours 34 minutes, and 37,300,000 results. The card features a thumbnail, the title, and a brief synopsis. It also includes sections for the cast (Besetzung) and viewing options (Film ansehen). The cast section lists Quentin Tarantino, John Travolta, Samuel L. Jackson, Uma Thurman, Bruce Willis, and Tim Roth. The viewing options section lists services like Prime Video, YouTube, Google Play, and Apple TV, each with a price of 2,99 € or 3,99 €. Below the card, there's a link to the Wikipedia page for Pulp Fiction. A sidebar on the left provides additional information under "Weitere Fragen" (Further Questions) about the film's uniqueness, title meaning, and cult status.

Google

pulp fiction

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

W 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

YouTube Ab 2,99 € Ansehen

Google Play Filme & Serien Ab 2,99 € Ansehen

Apple TV 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>
...

```

# Ontology

(Sometimes also referred to "semantic model")

An ontology combines features of

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

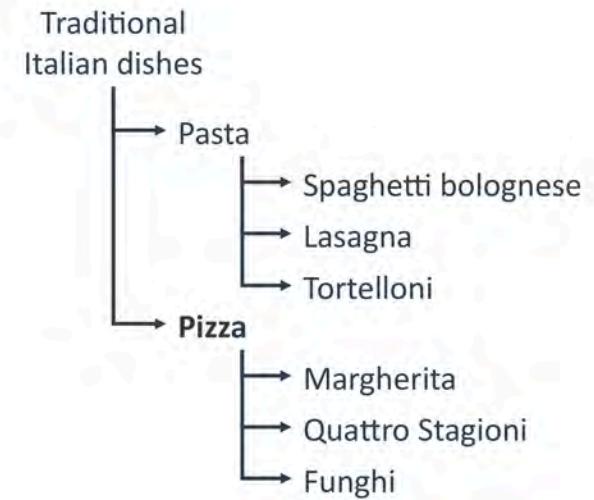
# Dictionary

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

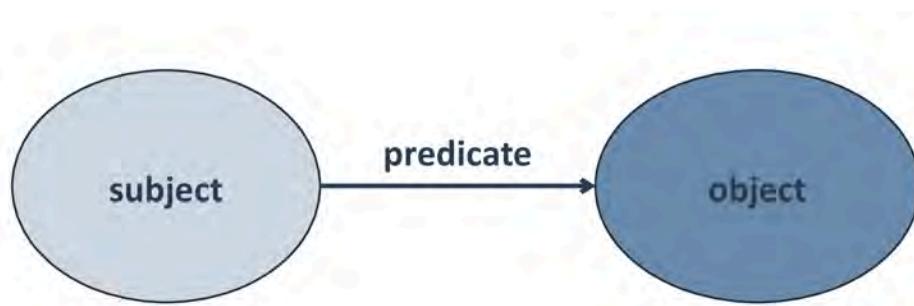
Dictionary of synonyms and relations

**Pizza** ≈ Lahmacun ≈ Focaccia ≈ Flammkuchen

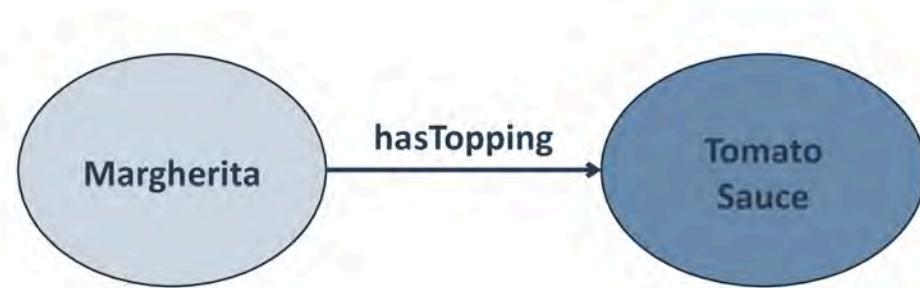
# Ontology

- 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**

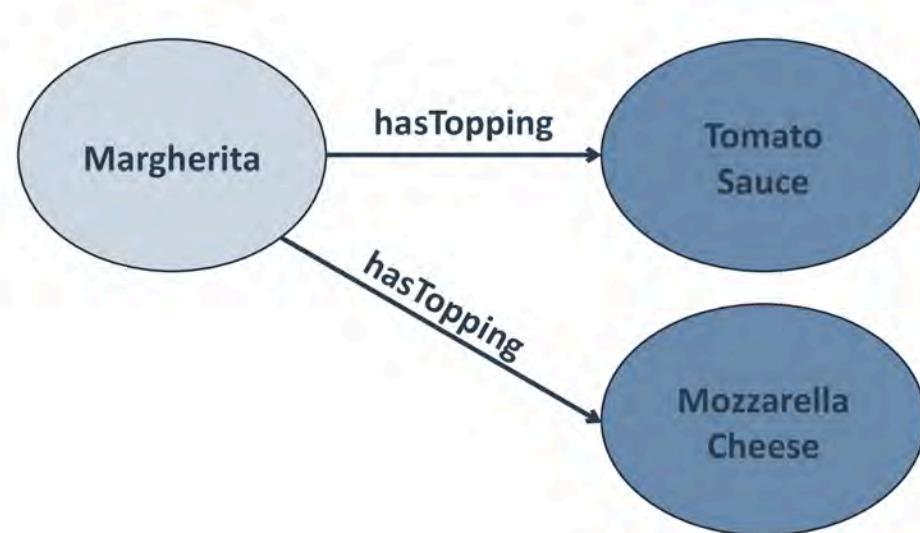
# The semantic triple



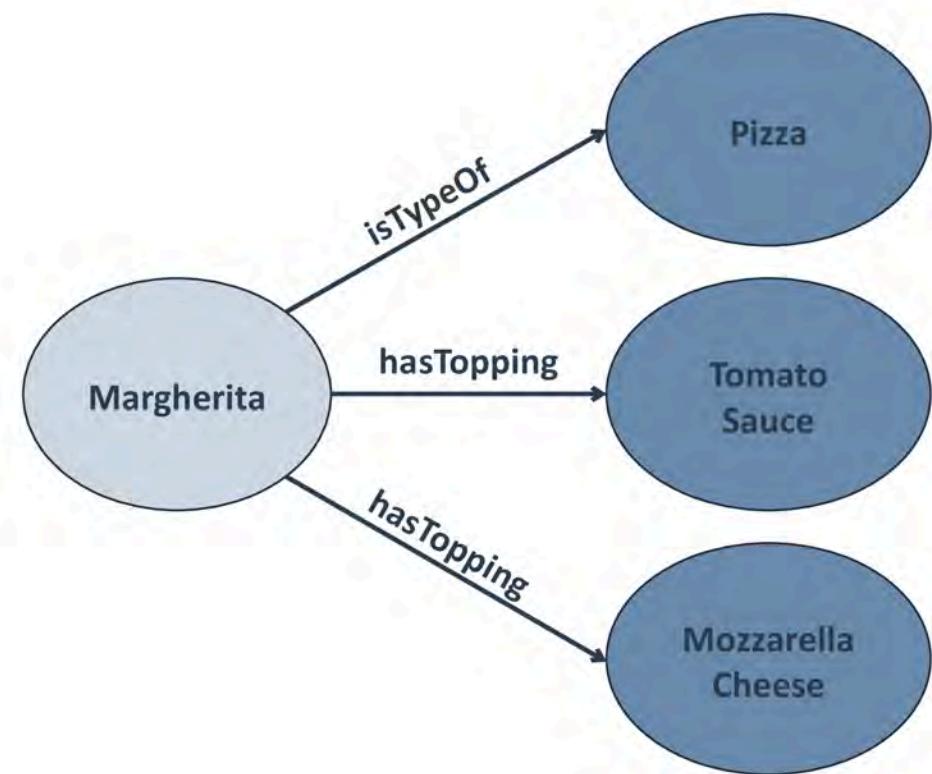
# Modeling a pizza menu



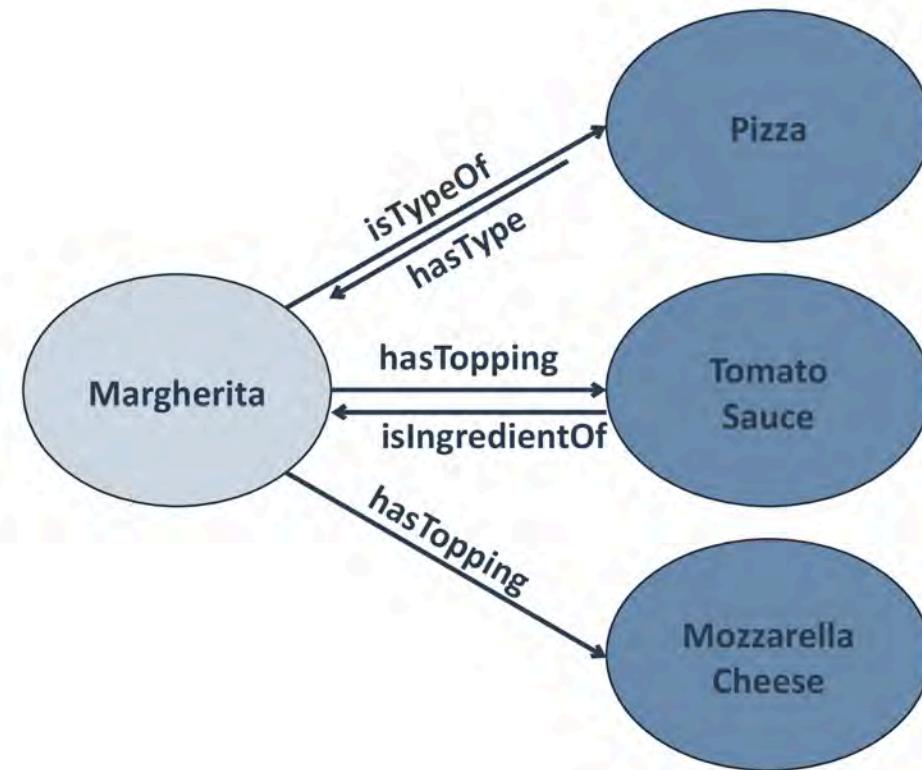
# Modeling a pizza menu



# Modeling a pizza menu

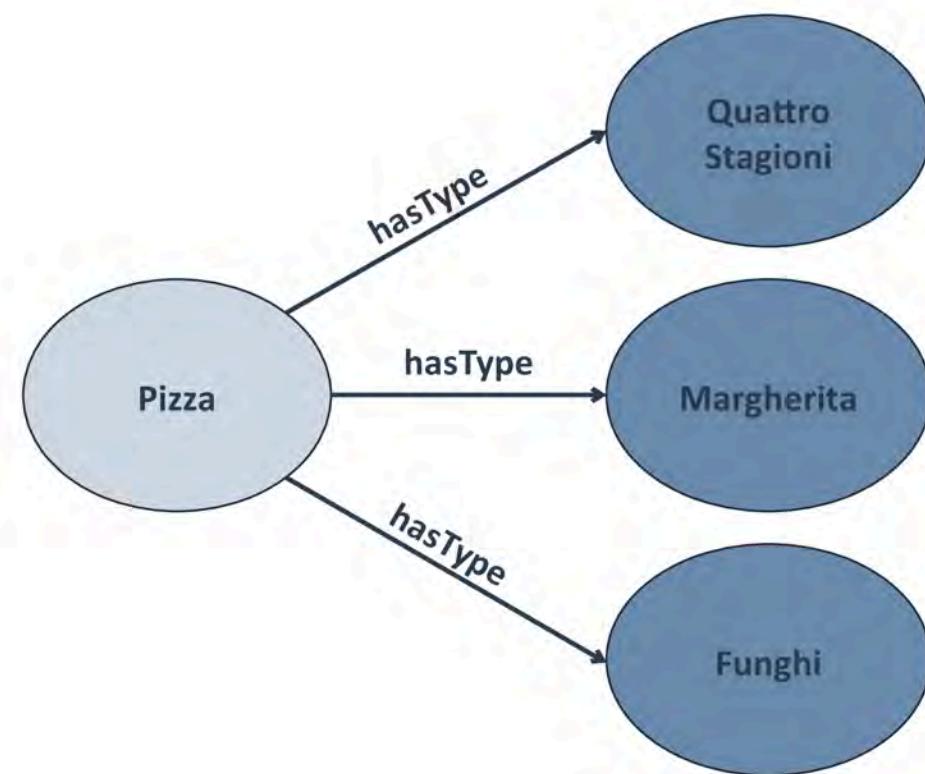


# Predicates have two directions

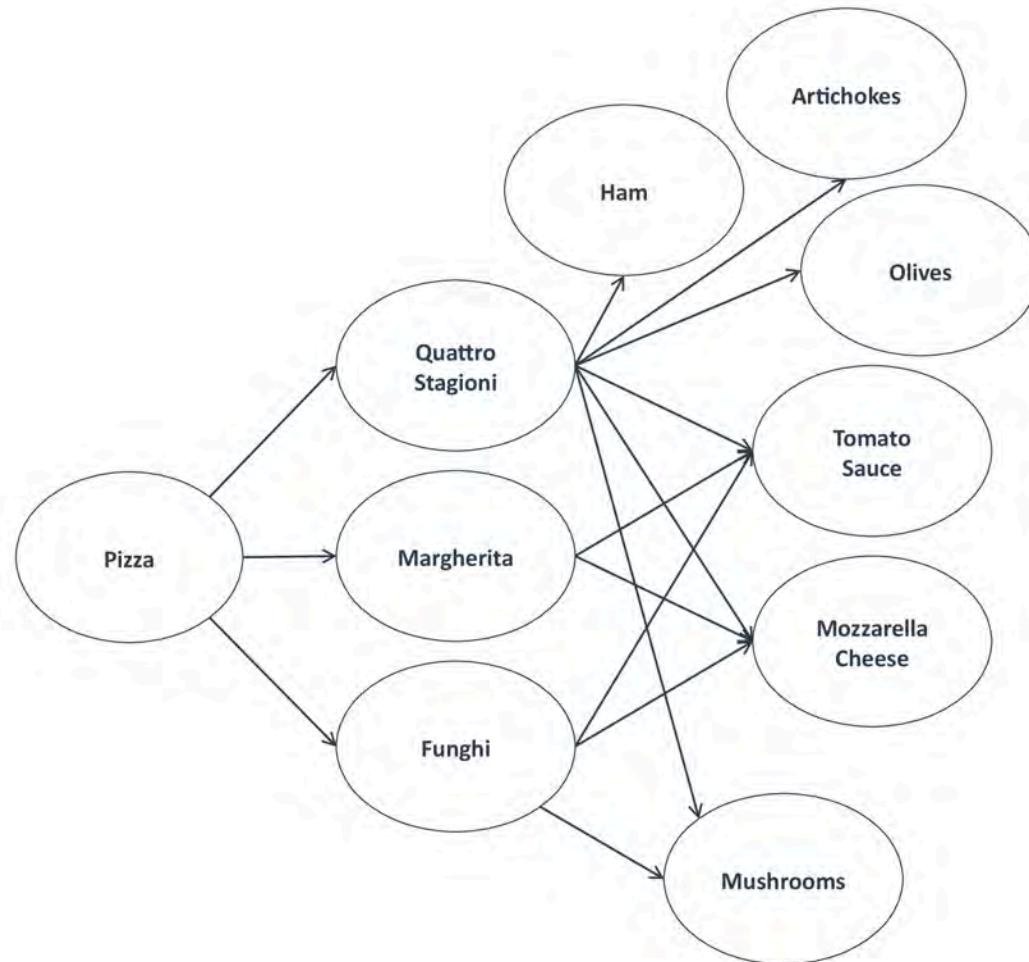


# Looking at the menu from a different perspective

An object of one triplet can be the subject to another



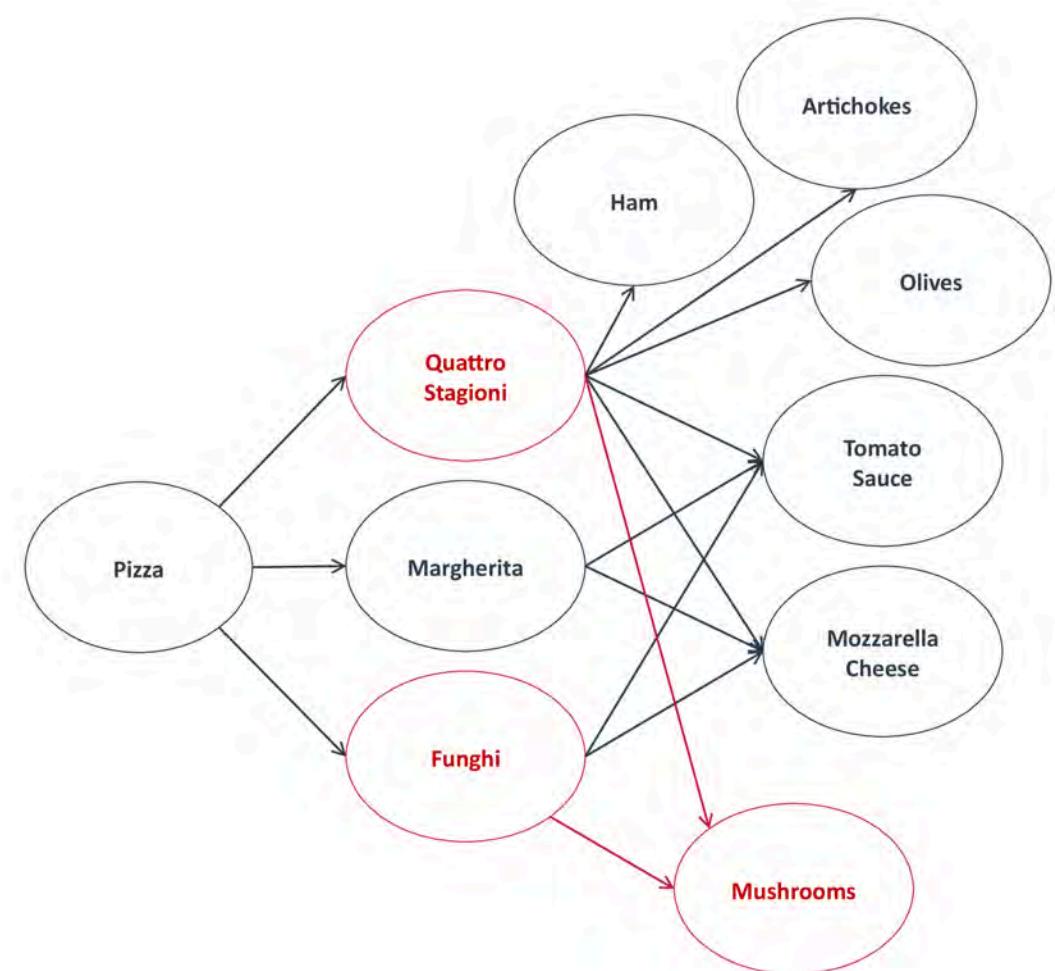
# (Towards) a knowledge graph



# 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>

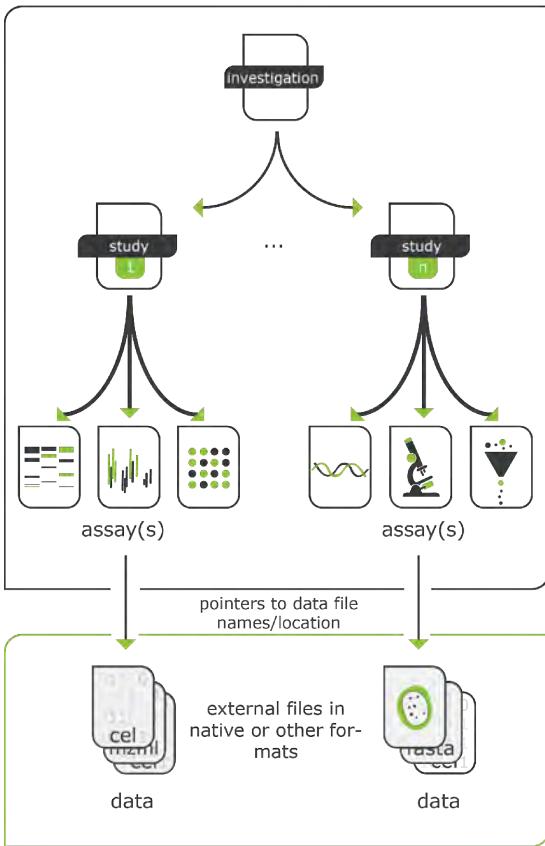
## PECO ontology

- 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>

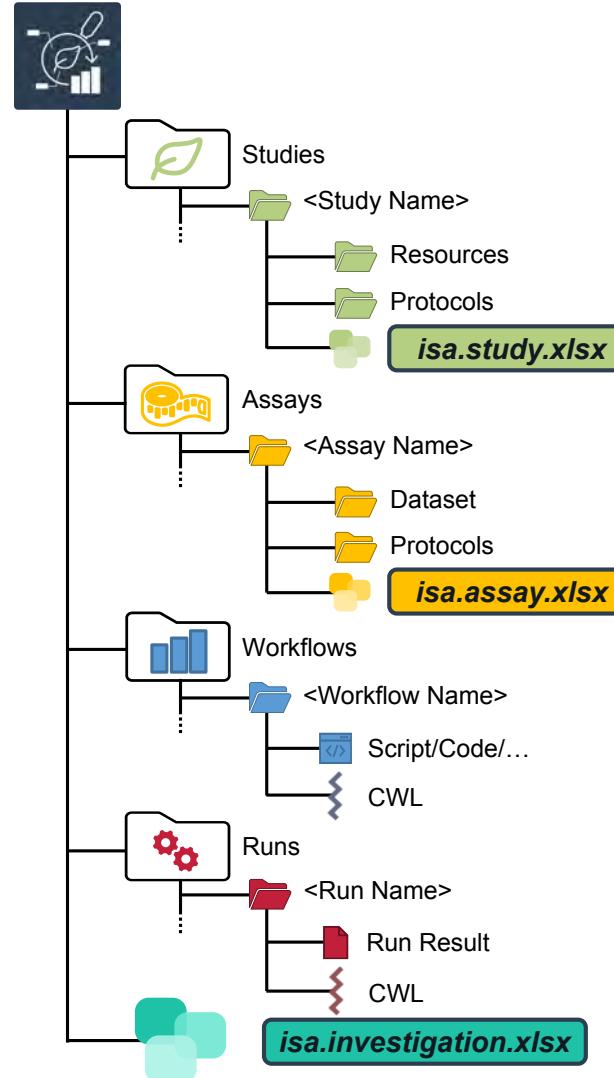
# ARC builds on ISA



**Investigation**  
Overall goals  
Scientific context

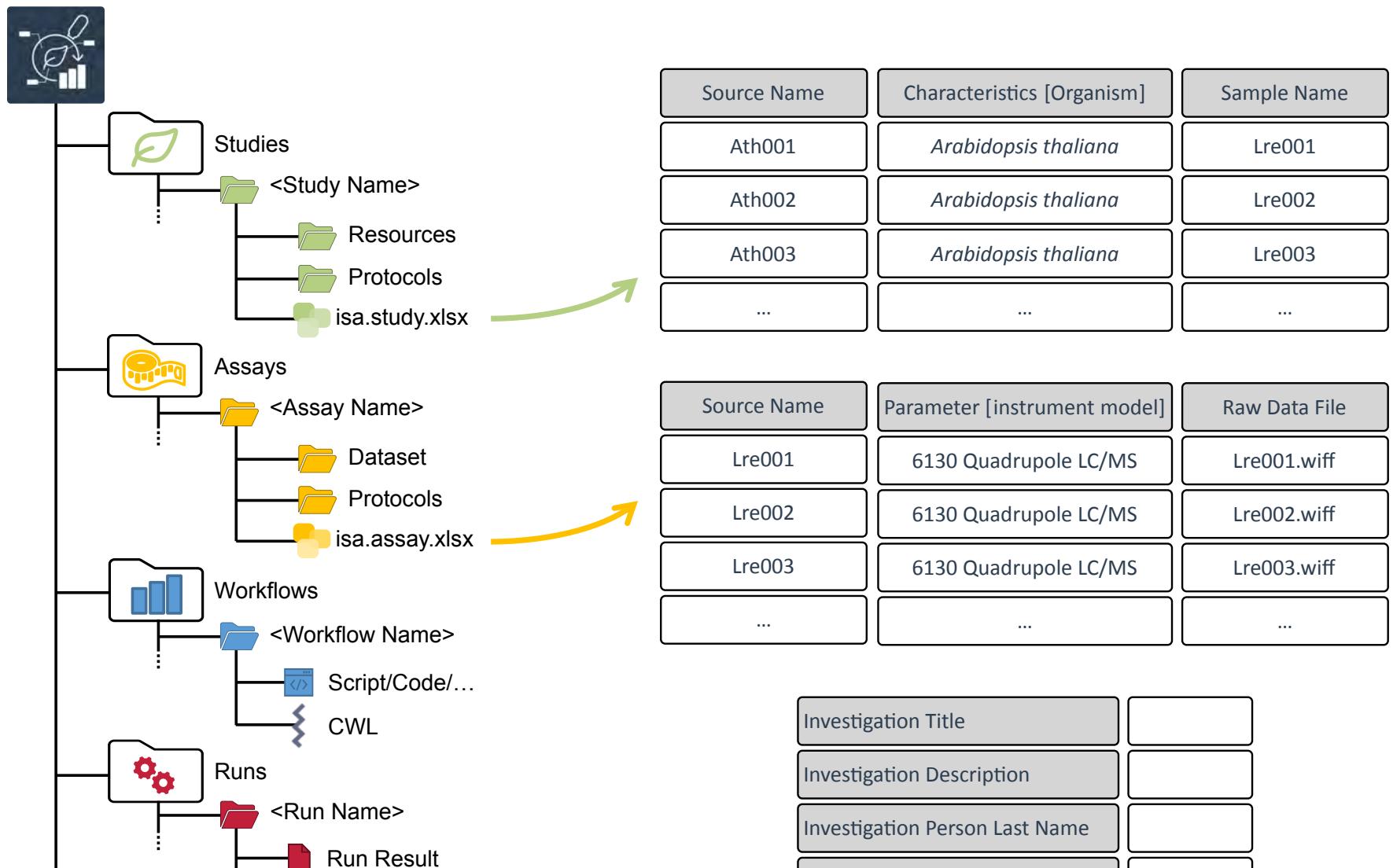
**Study**  
Experimental design

**Assay**  
Leading to (raw) data



# ARC builds on ISA

Metadata Annotations



# isa.<>.xlsx files within ARCs

*isa.investigation.xlsx*

DATACLOUD SOURCE REFERENCE	OB	STO	NENT	I_O	CIBIO	PATD	EFO
Term Source File	<a href="http://bioportal.bioontology.org/ontologies/Experimental Factor Ontology">http://bioportal.bioontology.org/ontologies/Experimental Factor Ontology</a>						
Term Source Version	47803_v126	v1.26	v1.26	v1.26	v1.26	v1.26	v1.26
Term Source Description	Ontology for Biomed BRINDA Issue / NWTF UniProt Tax Link Ontology Chemical Entity Phenotypic Array Exchange Diamerical factor Ontology						
INVESTIGATION							
Investigation Identifier	BII_1						
Investigation Title	Growth control of the yeast cell: a systems biology study in yeast						
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 control. This study aims to address this gap by using a systems biology approach to study cell growth control in yeast.						
Investigation Submission Date	30.04.07						
Investigation Public Release Date	19.03.09						
Investigation Publication Status	Published						
Investigation Publication Status Term Accession Number							
Investigation Publication Status Term Source REF							
INVESTIGATION PUBLICATIONS							
Investigation PubMed ID	17439666						
Investigation Publication DO	<a href="https://pubmed.ncbi.nlm.nih.gov/17439666/">https://pubmed.ncbi.nlm.nih.gov/17439666/</a>						
Investigation Publication Author List	Castroli J, Zeeb JA, Hoyle DC, Zhang N, Hayes A, Gardner DC, Cornell MJ, Petty J, Hayes L, Wettieworth L, Rash B, Brown JV, Dunn WB, Broadhurst C, Smith A, Smith M, Oliver S, Oliver SG						
Investigation Publication Title	Growth control of the yeast cell: a systems biology study in yeast						
Investigation Publication Status	published						
Investigation Publication Status Term Accession Number							
Investigation Publication Status Term Source REF							
STUDY							
Study Identifier	BII_1						
Study Title	Study of the impact of changes in flux on the transcriptome, proteome, endometabolome and exometabolome of the yeast Saccharomyces cerevisiae						
Study Description	We wished to study the impact of growth rate on the total component of mRNA molecules, proteins, and metabolites in S. cerevisiae. Independent						
Comment[Study Grant Number]							
Comment[Study Funding Agency]							
Study Submission Date	30.04.07						
Study Release Date	10.05.09						
Study Identifier	BII_1						
STUDY DESIGN DESCRIPTIONS							
Study Design Type	intervention design						
Study Design Type Term Accession Number	<a href="http://purl.bioontology.org/obo/ATO_0000015">http://purl.bioontology.org/obo/ATO_0000015</a>						
Study Design Type Term Source REF	OB						
STUDY PUBLICATIONS							
Study PubMed ID	17439666						
Study Publication DOI	<a href="https://doi.org/10.1186/gb-2008-9-10-s1">https://doi.org/10.1186/gb-2008-9-10-s1</a>						
Study Publication Author List	Castroli J, Zeeb JA, Hoyle DC, Zhang N, Hayes A, Gardner DC, Cornell MJ, Petty J, Hayes L, Wettieworth L, Rash B, Brown JV, Dunn WB, Broadhurst C, Smith A, Smith M, Oliver S, Oliver SG						
Study Publication Title	Growth control of the yeast cell: a systems biology study in yeast						
Study Publication Status	published						
Study Publication Status Term Accession Number							
Study Publication Status Term Source REF							
STUDY FACTORS							
Study Factor Name	Arabinose						
Study Factor Type	chemical compound						
Study Factor Type Term Accession Number	<a href="http://purl.bioontology.org/obo/ATO_0000161">http://purl.bioontology.org/obo/ATO_0000161</a>						
Study Factor Type Term Source REF	PATO						
STUDY ASSAYS							
Study Assay Measurement Type	protein expression profile; transcription profiling						
Study Assay Measurement Type Term Accession Number	<a href="http://purl.bioontology.org/obo/ATC_040014#">http://purl.bioontology.org/obo/ATC_040014#</a>						
Study Assay Measurement Type Term Source REF	OB	CBI	CBI				
Study Assay Technique Type	mass spectrometry						
Study Assay Technique Type Term Accession Number	<a href="http://purl.bioontology.org/obo/ATC_040014#">http://purl.bioontology.org/obo/ATC_040014#</a>						
Study Assay Technique Type Term Source REF	OB	CBI	CBI				
Study Assay Technique Platform	ITRAQ	LC-MS/MS	Affymetrix				
Study Assay File Name	<a href="#">#processsing</a>	<a href="#">#metabolite</a>	<a href="#">#transcript</a>				
STUDY PROTOCOLS							
Study Protocol Name	Growth protocol						
Study Protocol Type	mRNA extraction protein extraction; protein labeling						
Study Protocol Type Term Accession Number	<a href="http://purl.bioontology.org/obo/ATC_0402894">http://purl.bioontology.org/obo/ATC_0402894</a>						
Study Protocol Type Term Source File	OB						
Study Protocol Description	1. Biomass samples (1. Biomass samples (45 ml) were taken. This was done using Dmso. For each target, a hybridization cocktail was made using the						
Study Protocol JID							

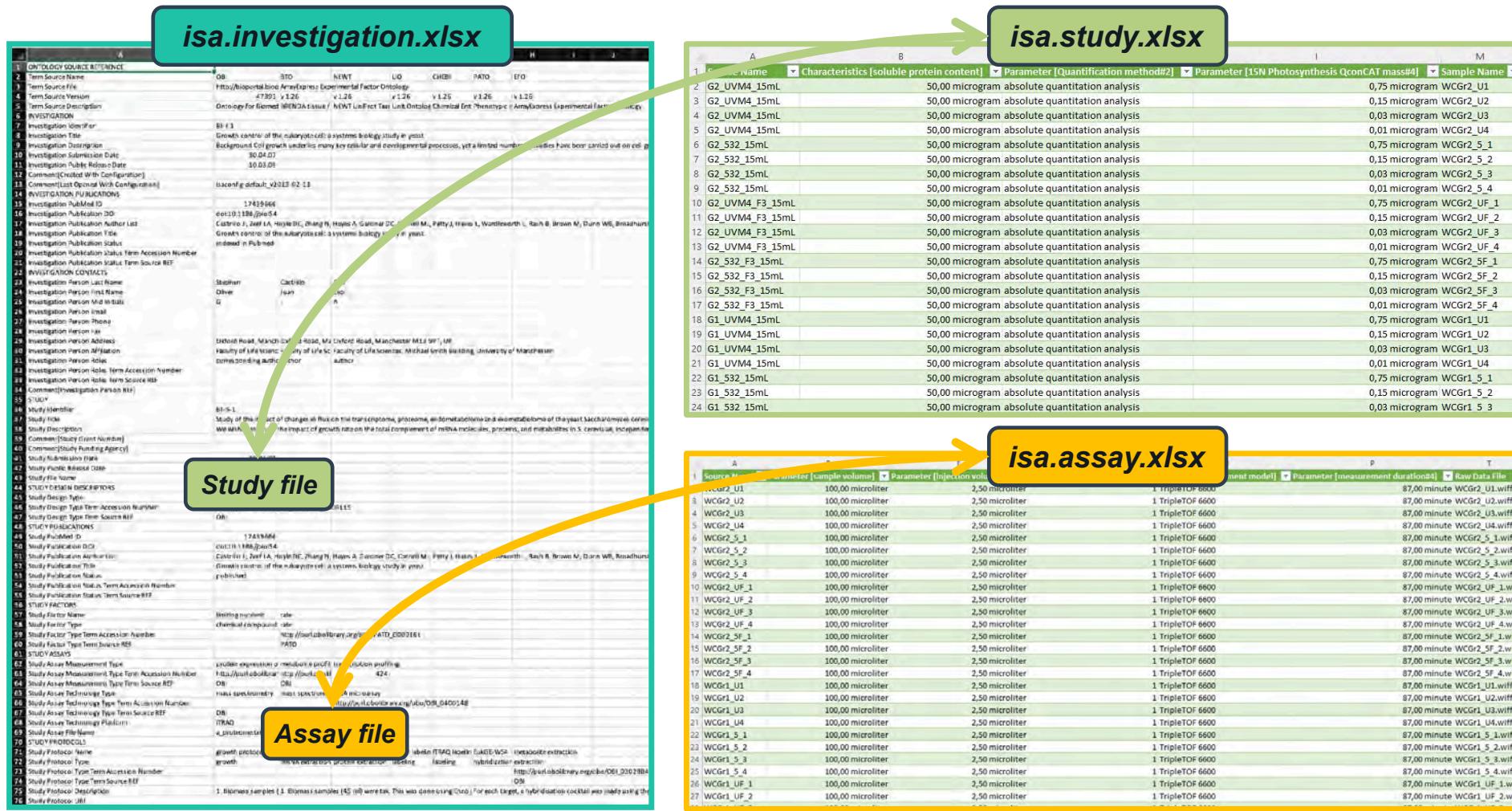
*isa.study.xlsx*

A	B	C	D	F
Source Name	Characteristics [soluble protein content]	Parameter [Quantification method]	Parameter [15N Photosynthesis QconCAT mass#4]	Sample Name
G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WGCr2_U1
G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WGCr2_U2
G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WGCr2_U3
G2_S32_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WGCr2_U4
G2_S32_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WGCr2_5_1
G2_S32_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WGCr2_5_2
G2_S32_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WGCr2_5_3
G2_S32_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WGCr2_5_4
G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WGCr2_UF_1
G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WGCr2_UF_2
G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WGCr2_UF_3
G2_S32_F3_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WGCr2_UF_4
G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WGCr1_U1
G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WGCr1_U2
G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WGCr1_U3
G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,01 microgram	WGCr1_U4
G1_S32_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WGCr1_5_1
G1_S32_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	WGCr1_5_2
G1_S32_15mL	50,00 microgram	absolute quantitation analysis	0,03 microgram	WGCr1_5_3

*isa.assay.xlsx*

A	B	C	D	E	F	G	H	I	J	K	L	M
Source Name	Parameter [sample volume]	Parameter [injection vol]	Parameter [measurement model]	Parameter [measurement duration]	Raw Data File							
WGCr2_U1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr2_U1.wiff							
WGCr2_U2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr2_U2.wiff							
WGCr2_U3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr2_U3.wiff							
WGCr2_U4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr2_U4.wiff							
WGCr2_5_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr2_5_1.wiff							
WGCr2_5_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr2_5_2.wiff							
WGCr2_5_3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr2_5_3.wiff							
WGCr2_5_4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr2_5_4.wiff							
WGCr2_UF_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr2_UF_1.wiff							
WGCr2_UF_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr2_UF_2.wiff							
WGCr2_UF_3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr2_UF_3.wiff							
WGCr2_UF_4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr2_UF_4.wiff							
WGCr2_SF_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr2_SF_1.wiff							
WGCr2_SF_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr2_SF_2.wiff							
WGCr2_SF_3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr2_SF_3.wiff							
WGCr2_SF_4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr2_SF_4.wiff							
WGCr1_U1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr1_U1.wiff							
WGCr1_U2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr1_U2.wiff							
WGCr1_U3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr1_U3.wiff							
WGCr1_U4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr1_U4.wiff							
WGCr1_S3_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr1_S3_1.wiff							
WGCr1_S3_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr1_S3_2.wiff							
WGCr1_S3_3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr1_S3_3.wiff							
WGCr1_S3_4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr1_S3_4.wiff							
WGCr1_UF_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr1_UF_1.wiff							
WGCr1_UF_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WGCr1_UF_2.wiff							

# 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

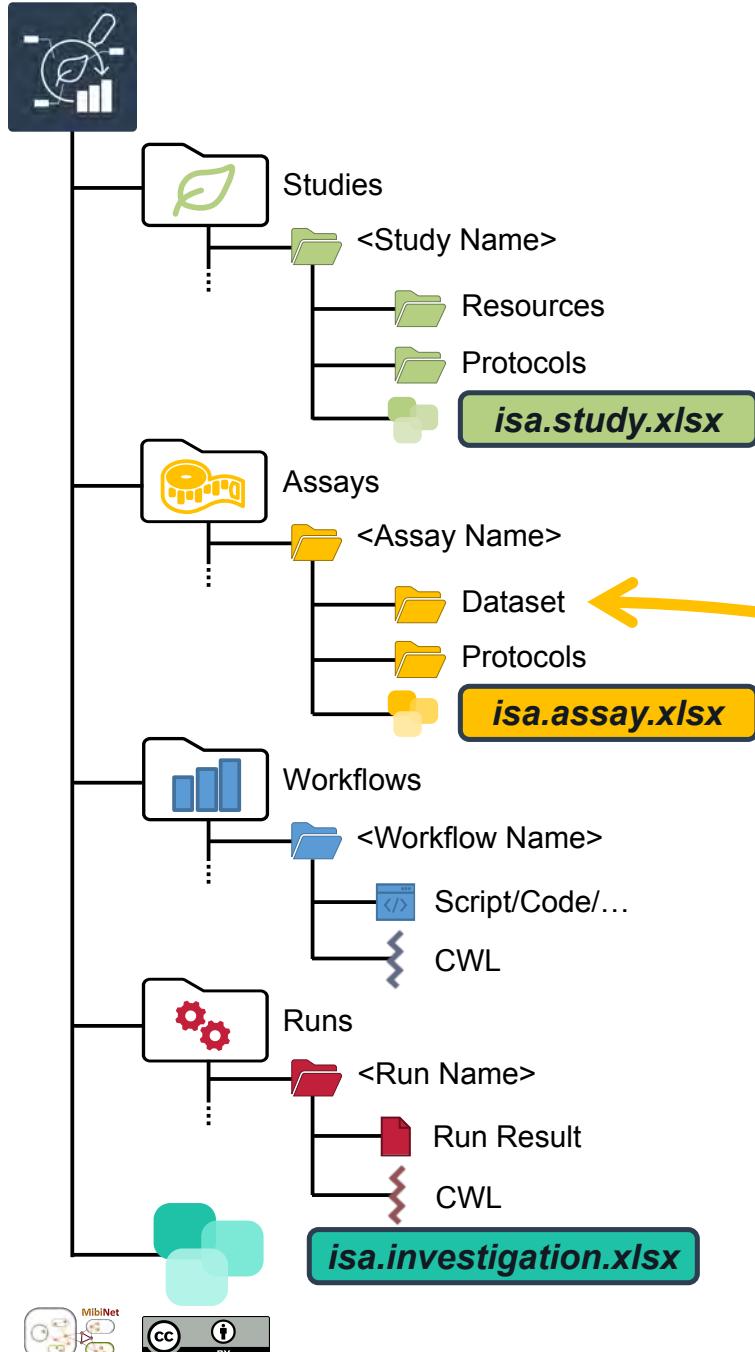
A	B	C	D	E	F	G	H	I	J	K	L	M
Source Name	Characteristics [soluble protein content]	Parameter [Quantification method#2]	Parameter [15N Photosynthesis QconCAT mass#4]	Sample Name								
G2_UVM4_15mL	50,00 microgram absolute quantitation analysis											0,75 microgram WCGr2_U1
G2_UVM4_15mL	50,00 microgram absolute quantitation analysis											0,15 microgram WCGr2_U2
G2_UVM4_15mL	50,00 microgram absolute quantitation analysis											0,03 microgram WCGr2_U3
G2_UVM4_15mL	50,00 microgram absolute quantitation analysis											0,01 microgram WCGr2_U4
G2_532_15mL	50,00 microgram absolute quantitation analysis											0,75 microgram WCGr2_5_1
G2_532_15mL	50,00 microgram absolute quantitation analysis											0,15 microgram WCGr2_5_2
G2_532_15mL	50,00 microgram absolute quantitation analysis											0,03 microgram WCGr2_5_3
G2_532_15mL	50,00 microgram absolute quantitation analysis											0,01 microgram WCGr2_5_4
G2_UVM4_F3_15mL	50,00 microgram absolute quantitation analysis											0,75 microgram WCGr2_UF_1
G2_UVM4_F3_15mL	50,00 microgram absolute quantitation analysis											0,15 microgram WCGr2_UF_2
G2_UVM4_F3_15mL	50,00 microgram absolute quantitation analysis											0,03 microgram WCGr2_UF_3
G2_UVM4_F3_15mL	50,00 microgram absolute quantitation analysis											0,01 microgram WCGr2_UF_4
G2_532_F3_15mL	50,00 microgram absolute quantitation analysis											0,75 microgram WCGr2_5F_1
G2_532_F3_15mL	50,00 microgram absolute quantitation analysis											0,15 microgram WCGr2_5F_2
G2_532_F3_15mL	50,00 microgram absolute quantitation analysis											0,03 microgram WCGr2_5F_3
G2_532_F3_15mL	50,00 microgram absolute quantitation analysis											0,01 microgram WCGr2_5F_4
G1_UVM4_15mL	50,00 microgram absolute quantitation analysis											0,75 microgram WCGr1_U1
G1_UVM4_15mL	50,00 microgram absolute quantitation analysis											0,15 microgram WCGr1_U2
G1_UVM4_15mL	50,00 microgram absolute quantitation analysis											0,03 microgram WCGr1_U3
G1_UVM4_15mL	50,00 microgram absolute quantitation analysis											0,01 microgram WCGr1_U4
G1_532_15mL	50,00 microgram absolute quantitation analysis											0,75 microgram WCGr1_5_1
G1_532_15mL	50,00 microgram absolute quantitation analysis											0,15 microgram WCGr1_5_2
G1_532_15mL	50,00 microgram absolute quantitation analysis											0,03 microgram WCGr1_5_3

isa.study.xlsx

Samples

A	B	C	D	E	F	G	H	I	J	K	L	M
Source Name	Parameter [sample volume]	Parameter [injection volu										
WCGr2_U1	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr2_U1.wiff
WCGr2_U2	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr2_U2.wiff
WCGr2_U3	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr2_U3.wiff
WCGr2_U4	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr2_U4.wiff
WCGr2_5_1	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr2_5_1.wiff
WCGr2_5_2	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr2_5_2.wiff
WCGr2_5_3	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr2_5_3.wiff
WCGr2_5_4	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr2_5_4.wiff
WCGr2_UF_1	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr2_UF_1.wiff
WCGr2_UF_2	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr2_UF_2.wiff
WCGr2_UF_3	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr2_UF_3.wiff
WCGr2_UF_4	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr2_UF_4.wiff
WCGr2_SF_1	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr2_SF_1.wiff
WCGr2_SF_2	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr2_SF_2.wiff
WCGr2_SF_3	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr2_SF_3.wiff
WCGr2_SF_4	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr2_SF_4.wiff
WCGr1_U1	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr1_U1.wiff
WCGr1_U2	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr1_U2.wiff
WCGr1_U3	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr1_U3.wiff
WCGr1_U4	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr1_U4.wiff
WCGr1_5_1	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr1_5_1.wiff
WCGr1_5_2	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr1_5_2.wiff
WCGr1_5_3	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr1_5_3.wiff
WCGr1_5_4	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr1_5_4.wiff
WCGr1_UF_1	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr1_UF_1.wiff
WCGr1_UF_2	100,00 microliter	2,50 microliter	1	TripleTOF 6600								87,00 minute WCGr1_UF_2.wiff

isa.assay.xlsx



**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							

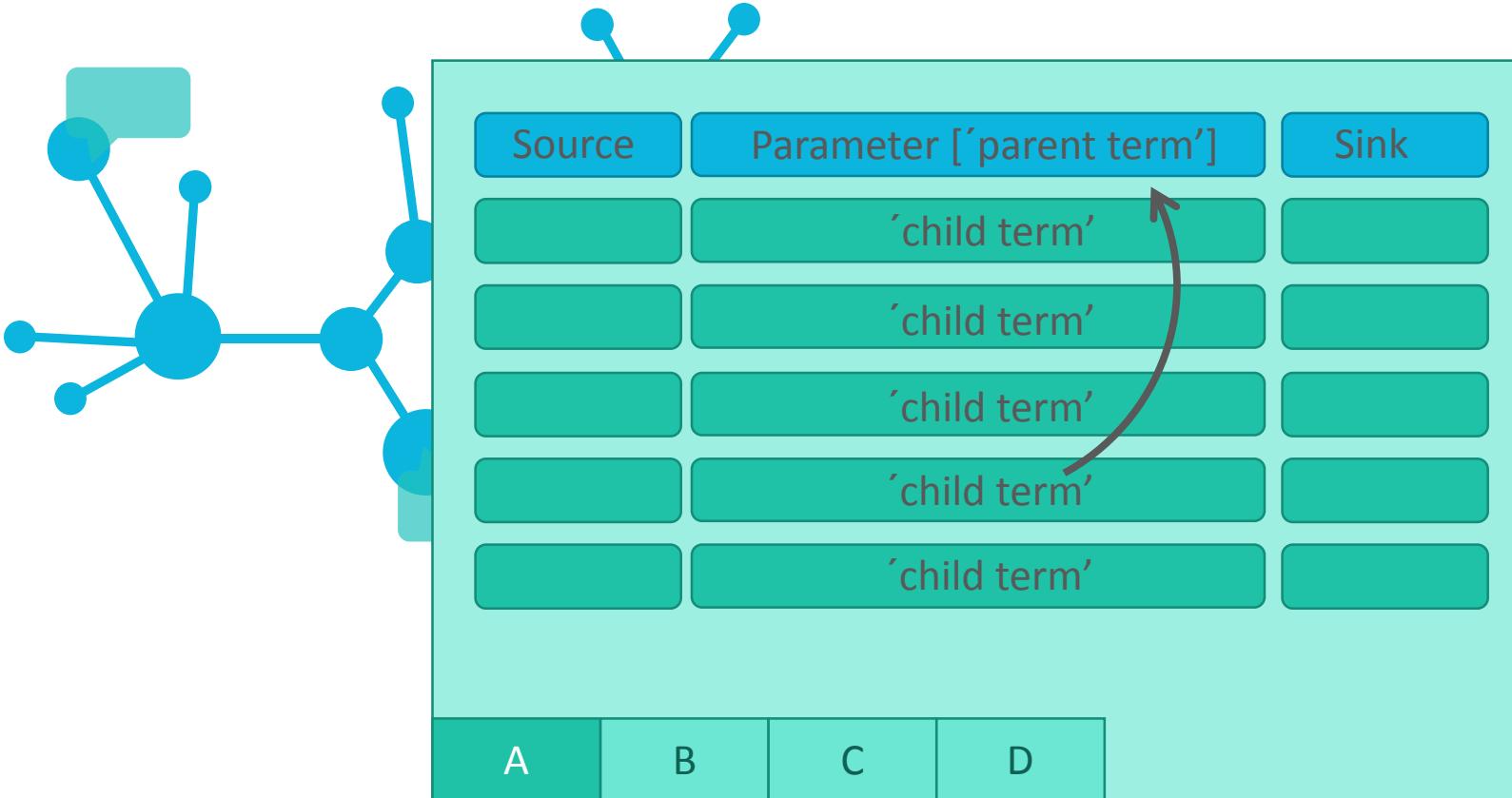
**isa.investigation.xlsx**

A	B	C	D	E	F	G	H	I	J	K	L	M
1	Source Name	Parameter [sample volume]	Parameter [injection vol]									
2	WCGr2_5_1.wiff			WIFF File								
3	WCGr2_5_2.wiff			WIFF File								
4	WCGr2_5_3.wiff			WIFF File								
5	WCGr2_5_4.wiff			WIFF File								
6	WCGr1_U1.wiff			WIFF File								
7	WCGr1_U2.wiff			WIFF File								
8	WCGr1_U3.wiff			WIFF File								
9	WCGr1_U4.wiff			WIFF File								
10	WCGr1_5_1.wiff			WIFF File								
11	WCGr1_5_2.wiff			WIFF File								
12	WCGr1_5_3.wiff			WIFF File								
13	WCGr1_5_4.wiff			WIFF File								
14	WCGr1_UF_1.wiff			WIFF File								
15	WCGr1_UF_2.wiff			WIFF File								
16	WCGr1_UF_3.wiff			WIFF File								
17	WCGr1_UF_4.wiff			WIFF File								

**Raw data**

# 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 the Microsoft Excel ribbon at the top with various tabs like File, Home, Insert, Draw, Page Layout, Formulas, Data, Review, View, Help, and Table Design. Below the ribbon is a toolbar with icons for Get & Transform Data, Queries & Connections, Data Types, Sort & Filter, Data Tools, Forecast, Group, Ungroup, Subtotal, Outline, and Swate. The main area displays a table titled "Sheet1" with columns A through AB. Column A contains "Source Name" and column AB contains "Sample Name". A callout bubble labeled "New Parameter" points to a modal window titled "SWATE". The modal has a tab bar with "Building Blocks" selected. It contains a section for "Add annotation building blocks (columns) to the annotation table." This section lists several entries under "Parameter instrument mod": "instrument model" (MS100003), "Instrument Model" (NCITC1774), "instrument" (MS100004), and "instrument" (EFO000054). There is also a section for "Agilent instrument model" with one entry (MS100049). A note at the bottom says, "Can't find the term you are looking for? Try..." followed by a link to a website.

# 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-like spreadsheet application titled "Swate" with the ribbon menu. The main window displays a table with columns labeled: Source Name, Protocol Ref, Characteristic (sample label), Factor (temperature), Annotated Instrument model, Component (software), and Sample Name. A "Building Blocks" sidebar on the right lists categories like Parameter, instrument model, and instrument. Several annotations are overlaid on the table:

- A callout box labeled "Characteristic" points to the "Characteristic (sample label)" column.
- A callout box labeled "Protocol Type/Protocol REF" points to the "Protocol Ref" column.
- A callout box labeled "Factor" points to the "Factor (temperature)" column.
- A callout box labeled "Component" points to the "Component (software)" column.
- A callout box labeled "Sample Name/Raw Data File Derived Data File" points to the "Sample Name" column.
- A callout box labeled "New Parameter" points to the "Parameter" category in the sidebar.

The sidebar also includes a search bar and a note about parameter columns describing experimental steps.

Source Name	Protocol Ref	Characteristic (sample label)	Factor (temperature)	Annotated Instrument model	Component (software)	Sample Name
G1_UVMA_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U1
G2_UVMA_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U2
G3_UVMA_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U3
G4_UVMA_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U4
G5_UVMA_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_1
G6_UVMA_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_2
G7_UVMA_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_3
G8_UVMA_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_4
G9_UVMA_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U1
G10_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U2
G11_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U3
G12_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U4
G13_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_1
G14_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_2
G15_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_3
G16_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_4
G17_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U1
G18_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U2
G19_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U3
G20_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U4
G21_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_1
G22_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_2
G23_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_3
G24_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_4
G25_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U1
G26_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U2
G27_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U3
G28_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U4
G29_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_1
G30_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_2
G31_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_3
G32_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_4
G33_UVMA_F1_15ml		30.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U1
G34_UVMA_15ml		4.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U2
G35_UVMA_15ml		4.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U3
G36_UVMA_15ml		4.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U4
G37_UVMA_15ml		4.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_1
G38_UVMA_15ml		4.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_2
G39_UVMA_15ml		4.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_3
G40_UVMA_15ml		4.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_4
G41_UVMA_15ml		4.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U1
G42_UVMA_F1_15ml		4.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U2
G43_UVMA_F1_15ml		4.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U3
G44_UVMA_F1_15ml		4.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U4
G45_UVMA_F1_15ml		4.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_1
G46_UVMA_F1_15ml		4.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_2
G47_UVMA_F1_15ml		4.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_3
G48_UVMA_F1_15ml		4.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_S2_4
G49_UVMA_F1_15ml		4.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U1
G50_UVMA_F1_15ml		4.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U2
G51_UVMA_F1_15ml		4.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U3
G52_UVMA_F1_15ml		4.00 degree Celsius 6130 Quadrupole LC/MS			Analyst	WCGRI_U4

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 numbered 1 to 52, each containing several pieces of information: Source Name, Protocol Type, Characteristic [sample label], Factor [temperature], Parameter [instrument model], Component [software], and Sample Name. The "Parameter [instrument model]" column contains values like "G2\_UV4M\_15mL", "G2\_UV4M\_F3\_15mL", "G1\_UV4M\_15mL", etc., which are highlighted in green. The "Component [software]" column contains values like "Analyst", "WCGr2\_U1", "WCGr2\_U2", etc. The "Sample Name" column contains values like "WCGr2\_U1", "WCGr2\_U2", "WCGr2\_U3", etc.

To the right of the spreadsheet, a "Swate" window is open. The title bar says "SWATE". The main area has tabs for "Ontology term search" and "Instrument search". The "Ontology term search" tab is active, showing a search bar with the text "instrument n 6130" and a dropdown menu showing "6130 Quadrupole LC/MS". Below the search bar, there is a message: "Can't 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!".

# Fill your table with ontology terms

The screenshot shows a Microsoft Excel spreadsheet titled "Sheet1" with data in columns A through AB. The data includes fields such as Source Name, Protocol Type, Characteristic [sample label], Factor [temperature], Parameter [instrument model], Component [software], and Sample Name. Many rows are highlighted in green, indicating they are selected for ontology term filling. An "ontology term search" dialog box from the "Swate" add-in is overlaid on the spreadsheet. The search bar contains the text "6130 Quadrupole LC/MS". Below the search bar are buttons for "Fill selected cells with this term" and "Use related term directed search". The "Swate" ribbon tab is also visible at the top of the Excel window.

File Home Insert Draw Page Layout Formulas Data Review View Help Table Design

Get & Transform Data Queries & Connections Data Types Sort & Filter Data Tools Forecast Outline

Queries & Connections Refresh All Properties Stocks Geography Advanced Text to Columns Flash Fill Remove Duplicates Data Validation Consolidate Relationships Manage Data Model What-If Analysis Forecast Sheet Group Ungroup Subtotal Core Experts Swate Swate

L2 fx 6130 Quadrupole LC/MS

A	B	E	H	L	P	T	U	Y	AB
1	Source Name	Protocol Type	Characteristic [sample label]	Factor [temperature]	Parameter [instrument model]	Component [software]	Sample Name		
2	G2_UVMA_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr2_U1		
3	G2_UVMA_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr2_U2		
4	G2_UVMA_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr2_U3		
5	G2_UVMA_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr2_U4		
6	G2_S32_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr2_S1		
7	G2_S32_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr2_S2		
8	G2_S32_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr2_S3		
9	G2_S32_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr2_S4		
10	G2_UVMA_F3_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr2_UF1		
11	G2_UVMA_F3_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr2_UF2		
12	G2_UVMA_F3_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr2_UF3		
13	G2_UVMA_F3_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr2_UF4		
14	G2_S32_F3_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr2_SF1		
15	G2_S32_F3_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr2_SF2		
16	G2_S32_F3_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr2_SF3		
17	G2_S32_F3_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr2_SF4		
18	G1_UVMA_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr1_U1		
19	G1_UVMA_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr1_U2		
20	G1_UVMA_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr1_U3		
21	G1_UVMA_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr1_U4		
22	G1_S32_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr1_S1		
23	G1_S32_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr1_S2		
24	G1_S32_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr1_S3		
25	G1_S32_15mL	data extraction protocol	15N	30,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr1_S4		
26	G1_UVMA_F7_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr1_UF1		
27	G1_UVMA_F7_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr1_UF2		
28	G1_UVMA_F7_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr1_UF3		
29	G1_UVMA_F7_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr1_UF4		
30	G1_S32_F10_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr1_SF1		
31	G1_S32_F10_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr1_SF2		
32	G1_S32_F10_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr1_SF3		
33	G1_S32_F10_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr1_SF4		
34	G3_UVMA_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr3_U1		
35	G3_UVMA_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr3_U2		
36	G3_UVMA_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr3_U3		
37	G3_UVMA_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr3_U4		
38	G3_S32_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr3_S1		
39	G3_S32_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr3_S2		
40	G3_S32_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr3_S3		
41	G3_S32_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr3_S4		
42	G3_UVMA_F1_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr3_UF1		
43	G3_UVMA_F1_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr3_UF2		
44	G3_UVMA_F1_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr3_UF3		
45	G3_UVMA_F1_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr3_UF4		
46	G3_S32_F2_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr3_SF1		
47	G3_S32_F2_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr3_SF2		
48	G3_S32_F2_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr3_SF3		
49	G3_S32_F2_15mL	data extraction protocol	15N	4,00 degree Celsius	6130 Quadrupole LC/MS	Analyst	WCGr3_SF4		
50									
51									
52									

Sheet1

Count: 48 Display Settings

Swate

Ontology term search

Search for an ontology term to fill into the selected field(s)

6130 Quadrupole LC/MS

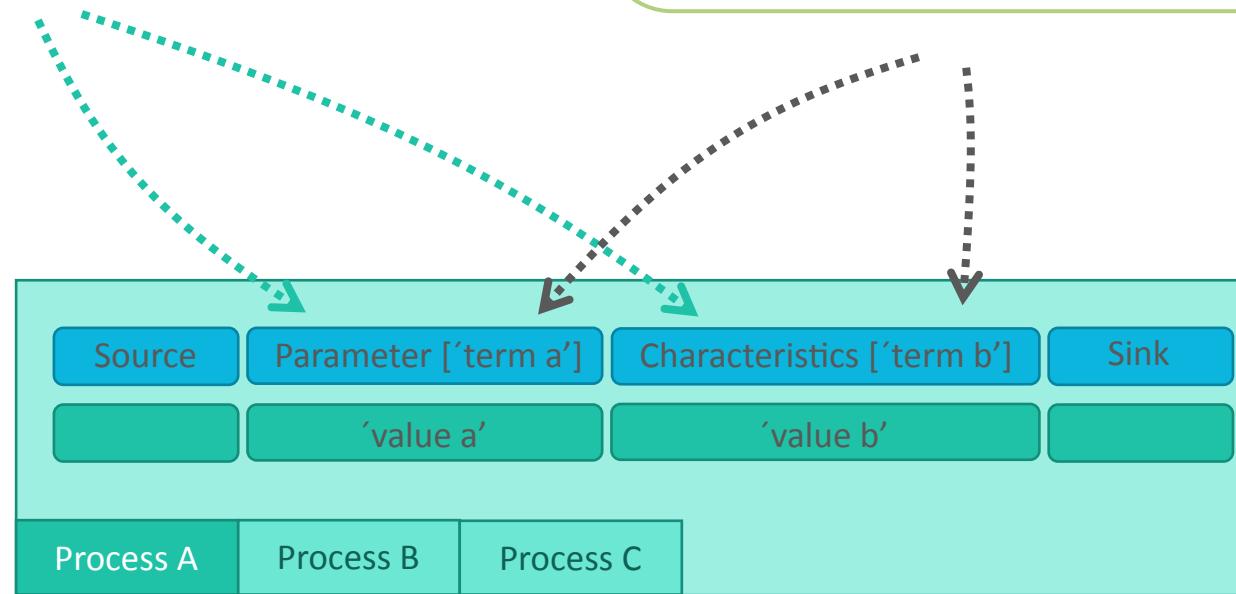
Use related term directed search.

Fill selected cells with this term

Swate Release Version 0.6.2

221

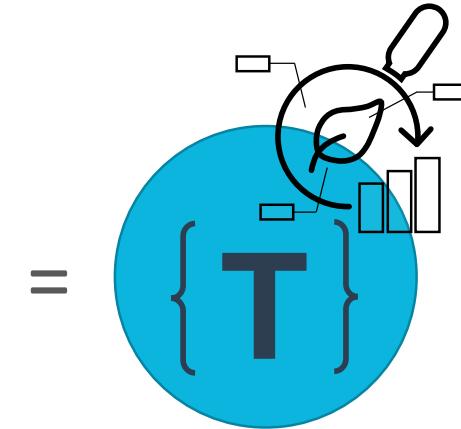
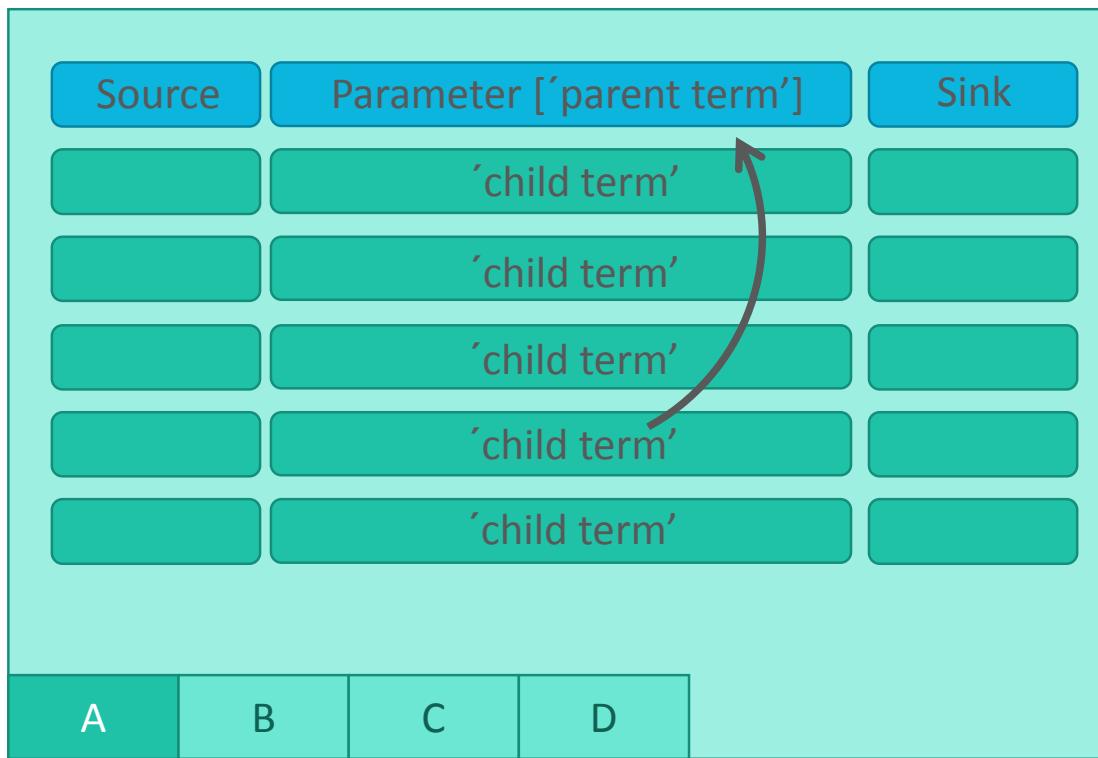
# Hierarchical combination of ontologies



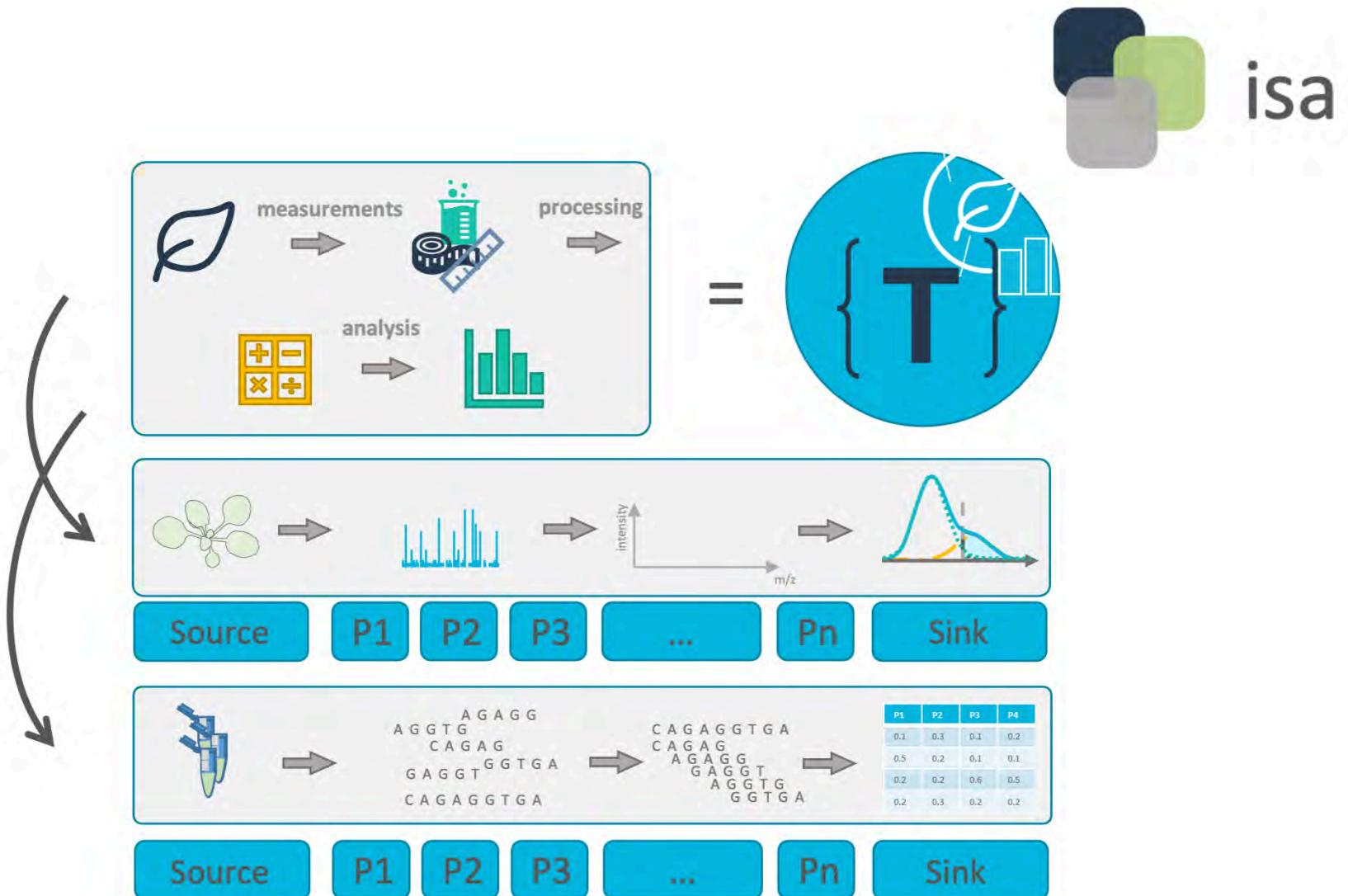
[isa.study.xlsx or isa.assay.xlsx](#)

# 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 toolbar with various icons. Below it is a navigation bar with 'Templates' and 'Template Search'. A search bar is present above a table. The table has columns for 'Protocol Name', 'Protocol Version', and 'Uses'. Each row contains a 'cur' or 'com' button, the version number, and a '0' with a dropdown arrow indicating zero uses. The table lists several biological and computational protocols.

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



# 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>

# Block 6 – Swate hands-on

September 28th, 2023



Sabrina Zander  
[MibiNet](#)



Dominik Brilhaus  
[CEPLAS Data Science](#)

# 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
- MacOSTextEdit

Recommended text editors with code highlighting:

- Visual Studio Code <https://code.visualstudio.com/>
- BBEdit <https://www.barebones.com/products/bbedit/>
- Sublime <https://www.sublimetext.com/>

## Download the demo data

```
git clone "https://demo-user:5ehDYeHcqP2MqVXsNNPu@git.nfdi4plants.org/teaching/demo-arc_level1.git"
```

# Where we left off last time

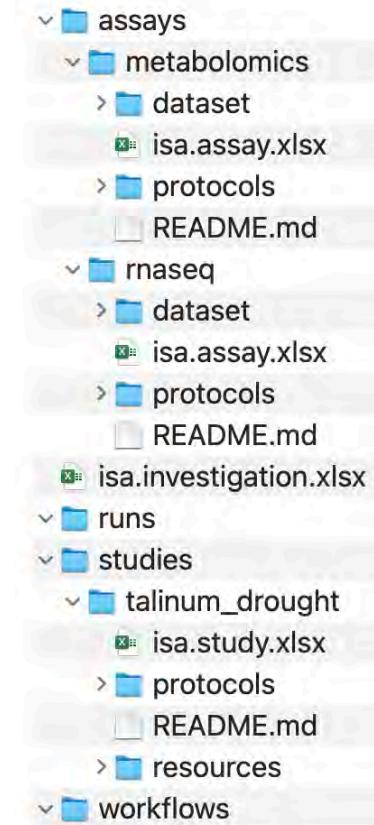
👤 Initiated an ARC

📁 Structured and ...

🌐 Shared with collaborators

Today we want to

S+ ... annotate the experimental data



# Swate hands-on with demo data

# Swate Overview

The screenshot shows a Microsoft Excel spreadsheet titled "tabelle1.xlsx" with data in columns A through K. The data includes rows for various samples (e.g., Heat\_15A\_OD\_R1, Heat\_15A\_OD\_R2, Heat\_180A\_OD\_R1, etc.) with their corresponding characteristics (e.g., 15N, 32.00 degree Celsius), factors (e.g., Heat\_15A\_OD\_R1.wiff, Heat\_15A\_OD\_R2.wiff), and data file names (e.g., Heat\_15A\_OD\_R1.wiff, Heat\_15A\_OD\_R2.wiff). The first row is labeled "source".

To the right of the Excel window, the "Swate" application is open. It displays a search bar and a list of annotations under "otation building block selection". The annotations include:

- Parameter instrument MS-1000468
- instrument model MS-100031
- instrument vendor MS-1001269
- medical instrument ENVO0001032
- Mascot:Instrument MS-1001655

Below the annotations, there is a search bar and a link to "Advanced Search". A note says "Can't find the Term you are looking for? Use Advanced Search".

At the bottom of the Swate window, it says "More about Parameter:" followed by a brief description: "Use parameters to annotate your experimental workflow. You can group parameters to create a protocol. You can find more information on our website." The version is listed as "Swate Release Version 0.4.7".

Source	Characteristics (sample label)	Factor [temperature unit]	Data File Name
1 Source Name			
2 Heat_15A_OD_R1	15N	32.00 degree Celsius	Heat_15A_OD_R1.wiff
3 Heat_15A_OD_R2	15N	32.00 degree Celsius	Heat_15A_OD_R2.wiff
4 Heat_180A_OD_R1	15N	32.00 degree Celsius	Heat_180A_OD_R1.wiff
5 Heat_180A_OD_R2	15N	32.00 degree Celsius	Heat_180A_OD_R2.wiff
6 Heat_2880A_OD_R1	15N	32.00 degree Celsius	Heat_2880A_OD_R1.wiff
7 Heat_2880A_OD_R2	15N	32.00 degree Celsius	Heat_2880A_OD_R2.wiff
8 Heat_5760A_OD_R1	15N	32.00 degree Celsius	Heat_5760A_OD_R1.wiff
9 Heat_5760A_OD_R2	15N	32.00 degree Celsius	Heat_5760A_OD_R2.wiff
10 Heat_5760A_15D_R1	15N	32.00 degree Celsius	Heat_5760A_15D_R1.wiff
11 Heat_5760A_15D_R2	15N	32.00 degree Celsius	Heat_5760A_15D_R2.wiff
12 Heat_5760A_180D_R1	15N	32.00 degree Celsius	Heat_5760A_180D_R1.wiff
13 Heat_5760A_180D_R2	15N	32.00 degree Celsius	Heat_5760A_180D_R2.wiff
14 Heat_5760A_2880D_R1	15N	32.00 degree Celsius	Heat_5760A_2880D_R1.wiff
15 Heat_5760A_2880D_R2	15N	32.00 degree Celsius	Heat_5760A_2880D_R2.wiff
16 Heat_5760A_5760D_R1	15N	32.00 degree Celsius	Heat_5760A_5760D_R1.wiff
17 Heat_5760A_5760D_R2	15N	32.00 degree Celsius	Heat_5760A_5760D_R2.wiff
18 Cold_15A_OD_R1	15N	4.00 degree Celsius	Cold_15A_OD_R1.wiff
19 Cold_15A_OD_R2	15N	4.00 degree Celsius	Cold_15A_OD_R2.wiff
20 Cold_180A_OD_R1	15N	4.00 degree Celsius	Cold_180A_OD_R1.wiff
21 Cold_180A_OD_R2	15N	4.00 degree Celsius	Cold_180A_OD_R2.wiff
22 Cold_2880A_OD_R1	15N	4.00 degree Celsius	Cold_2880A_OD_R1.wiff
23 Cold_2880A_OD_R2	15N	4.00 degree Celsius	Cold_2880A_OD_R2.wiff
24 Cold_5760A_OD_R1	15N	4.00 degree Celsius	Cold_5760A_OD_R1.wiff
25 Cold_5760A_OD_R2	15N	4.00 degree Celsius	Cold_5760A_OD_R2.wiff
26 Cold_5760A_15D_R1	15N	4.00 degree Celsius	Cold_5760A_15D_R1.wiff
27 Cold_5760A_15D_R2	15N	4.00 degree Celsius	Cold_5760A_15D_R2.wiff
28 Cold_5760A_180D_R1	15N	4.00 degree Celsius	Cold_5760A_180D_R1.wiff
29 Cold_5760A_180D_R2	15N	4.00 degree Celsius	Cold_5760A_180D_R2.wiff
30 Cold_5760A_2880D_R1	15N	4.00 degree Celsius	Cold_5760A_2880D_R1.wiff
31 Cold_5760A_2880D_R2	15N	4.00 degree Celsius	Cold_5760A_2880D_R2.wiff
32 Cold_5760A_5760D_R1	15N	4.00 degree Celsius	Cold_5760A_5760D_R1.wiff
33 Cold_5760A_5760D_R2	15N	4.00 degree Celsius	Cold_5760A_5760D_R2.wiff
34 Highlight_15A_OD_R1	15N	22.00 degree Celsius	Highlight_15A_OD_R1.wiff
35 Highlight_15A_OD_R2	15N	22.00 degree Celsius	Highlight_15A_OD_R2.wiff
36 Highlight_180A_OD_R1	15N	22.00 degree Celsius	Highlight_180A_OD_R1.wiff
37 Highlight_180A_OD_R2	15N	22.00 degree Celsius	Highlight_180A_OD_R2.wiff

## Let's annotate the plant samples first

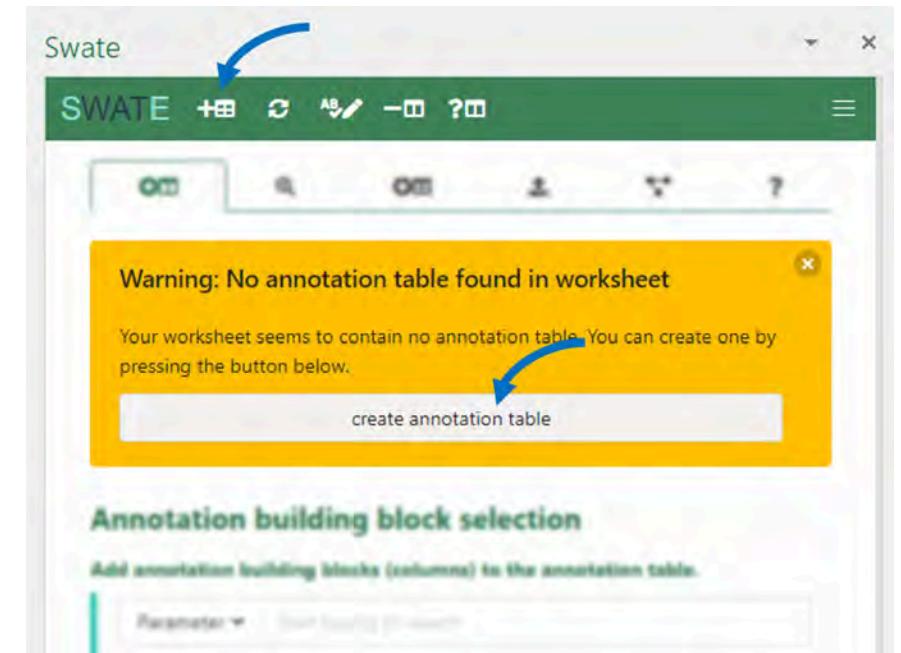
1. Navigate to the demo ARC.
2. Open the lab notes `studies/talinum_drought/protocols/plant_material.txt` in a text editor.
3. Open the empty `studies/talinum_drought/isa.study.xlsx` workbook in Excel.

# 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

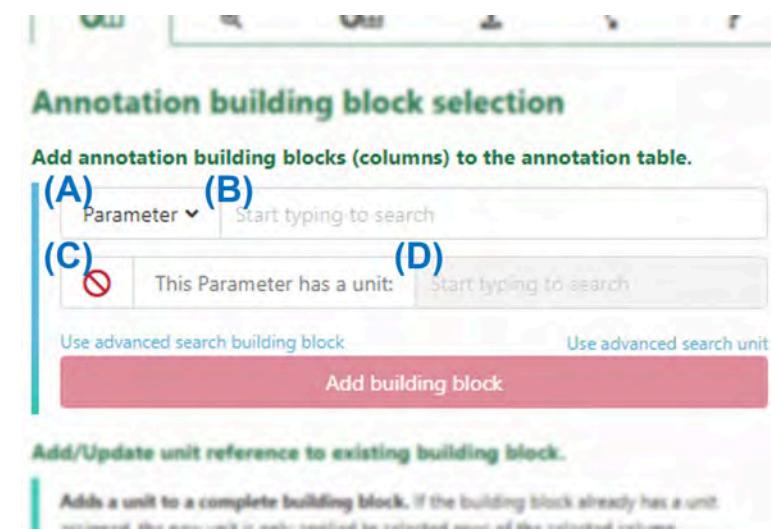
💡 Only one annotation table can be added per Excel sheet



# 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.



The screenshot shows the 'Annotation building block selection' interface. At the top, there's a header with various icons. Below it, a green bar says 'Annotation building block selection'. A sub-header says 'Add annotation building blocks (columns) to the annotation table.' There are two main search fields: one for 'Parameter' (labeled A) and one for 'Characteristic' (labeled C). The 'Parameter' field has a placeholder 'Start typing to search'. The 'Characteristic' field has a placeholder 'This Parameter has a unit: Start typing to search'. Below these fields are two buttons: 'Use advanced search building block' and 'Use advanced search unit'. At the bottom is a large red button labeled 'Add building block'.

# 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 `microeinsteин 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 microeinsteин 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.

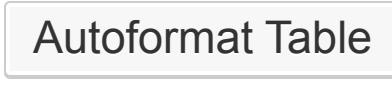
# Your ISA table is growing

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

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

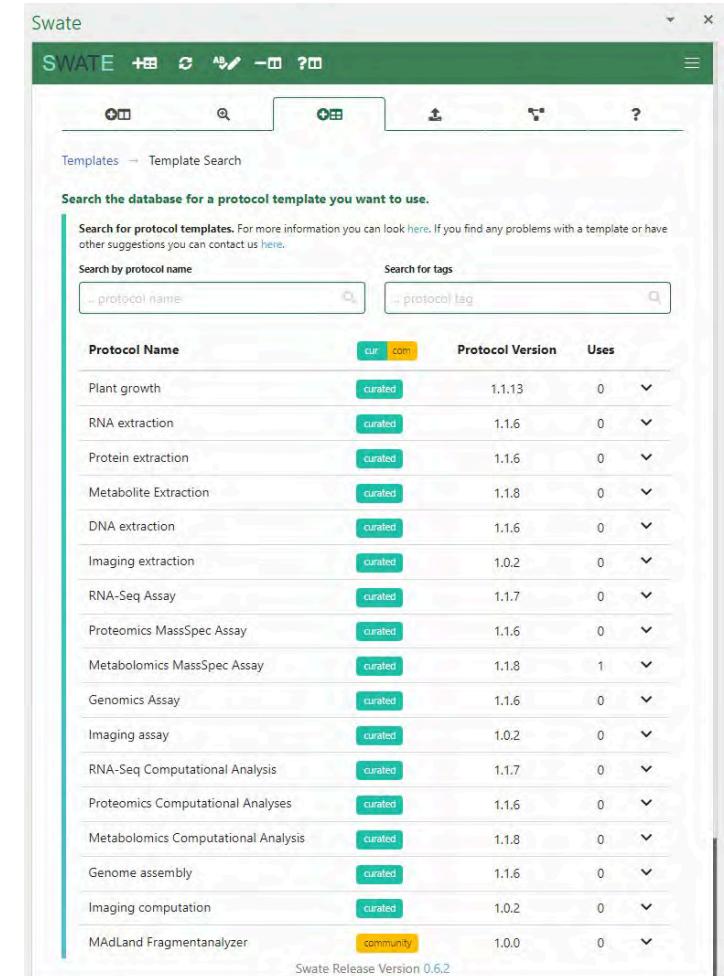
1. Navigate to the demo ARC.
2. Open the lab notes `assays/rnaseq/protocols/RNA_extraction.txt` in a text editor.
3. Open the empty `assays/rnaseq/isa.assay.xlsx` workbook in Excel.

# 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.

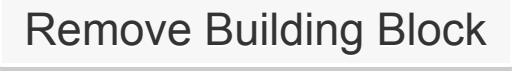


# Adding / Updating unit references

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")
```

# Annotation with ARCitect

# Process Information

 Assay  
General Meta Data of the Assay 

---

 Data  
Process Information 

 APPLY TEMPLATE

 UPDATE

 RESET



# 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

