Block 0 – Preparation

before September 21st, 2023

Sabrina Zander

MibiNet

Dominik Brilhaus

CEPLAS Data Science

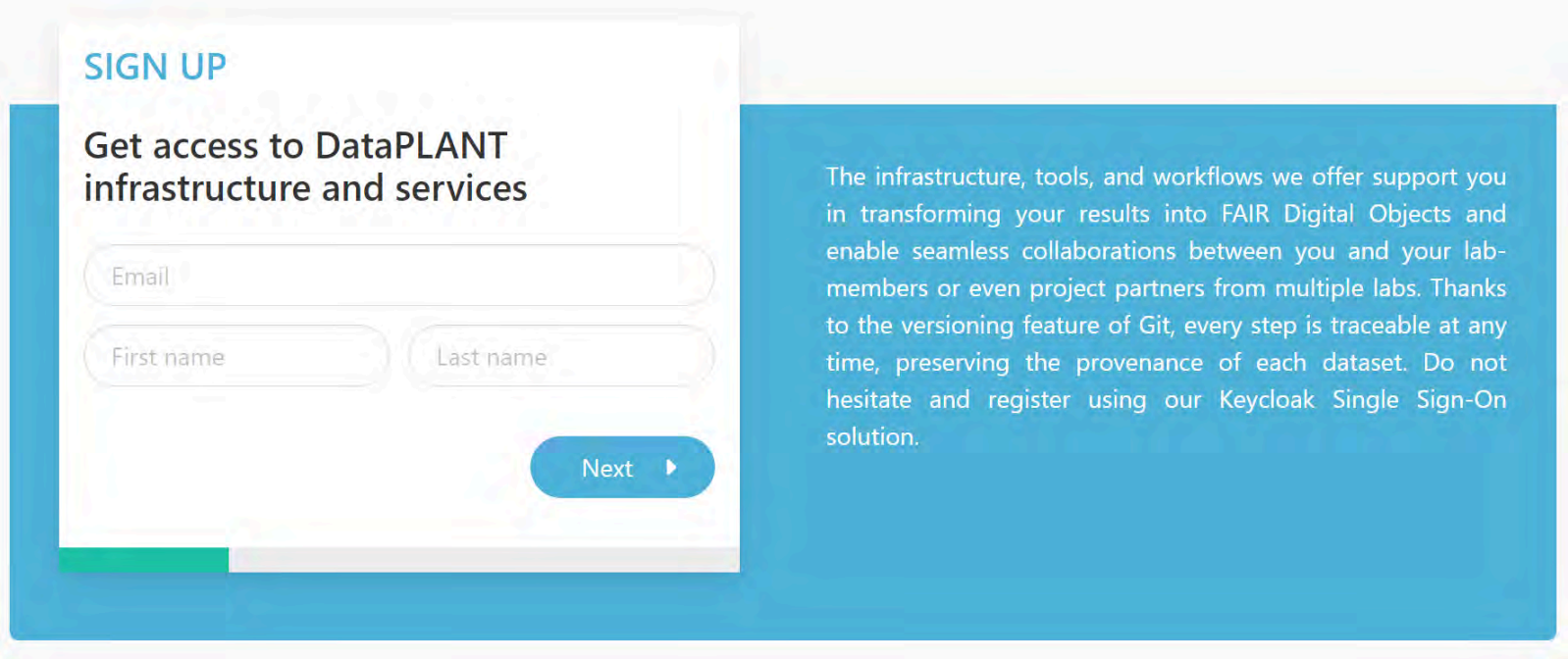
Checklist hands-on sessions

💡 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

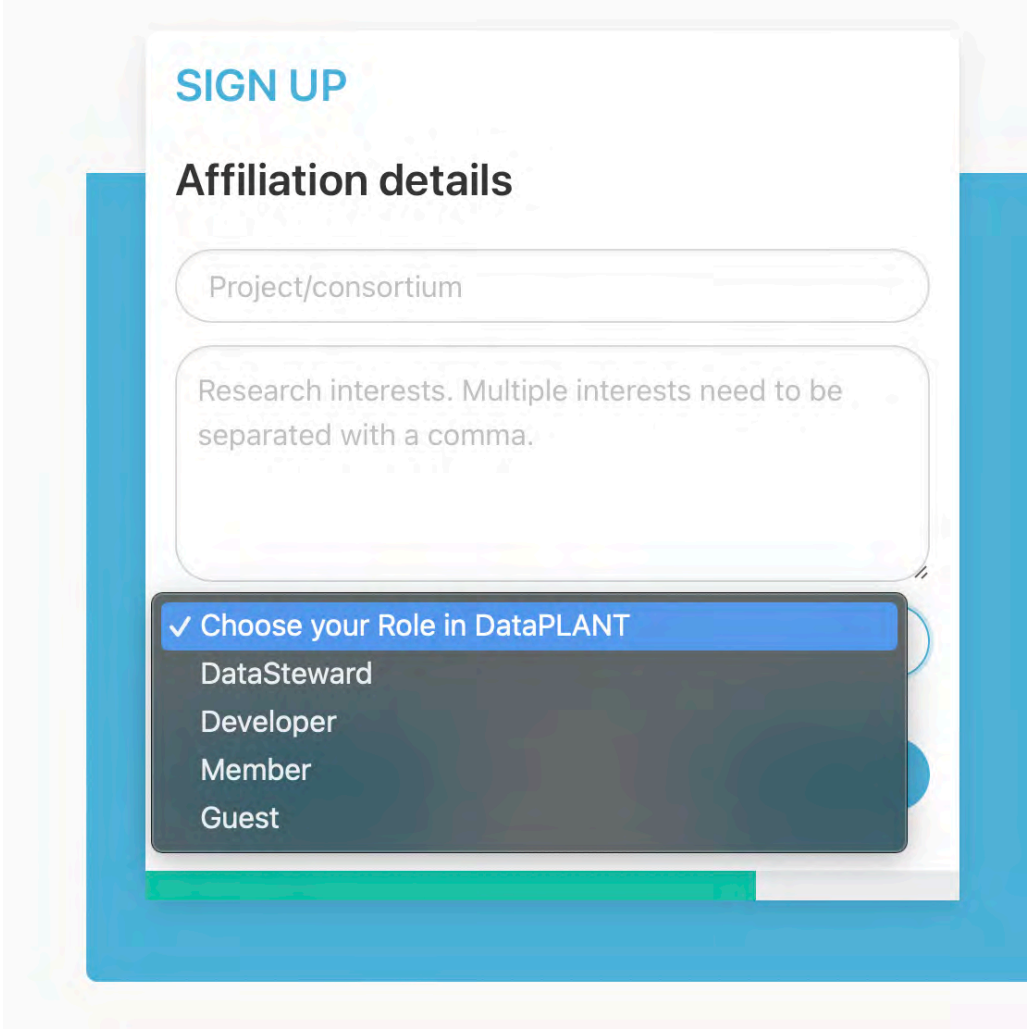
DataPLANT Registration

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`



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
```

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	*Short break*
11:00 - 12:00	Intro to DataPLANT and ARC
12:00 - 13:00	*Lunch*
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	*Short break*
10:45 - 12:00	ISA and Metadata
12:00 - 13:00	*Lunch*
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

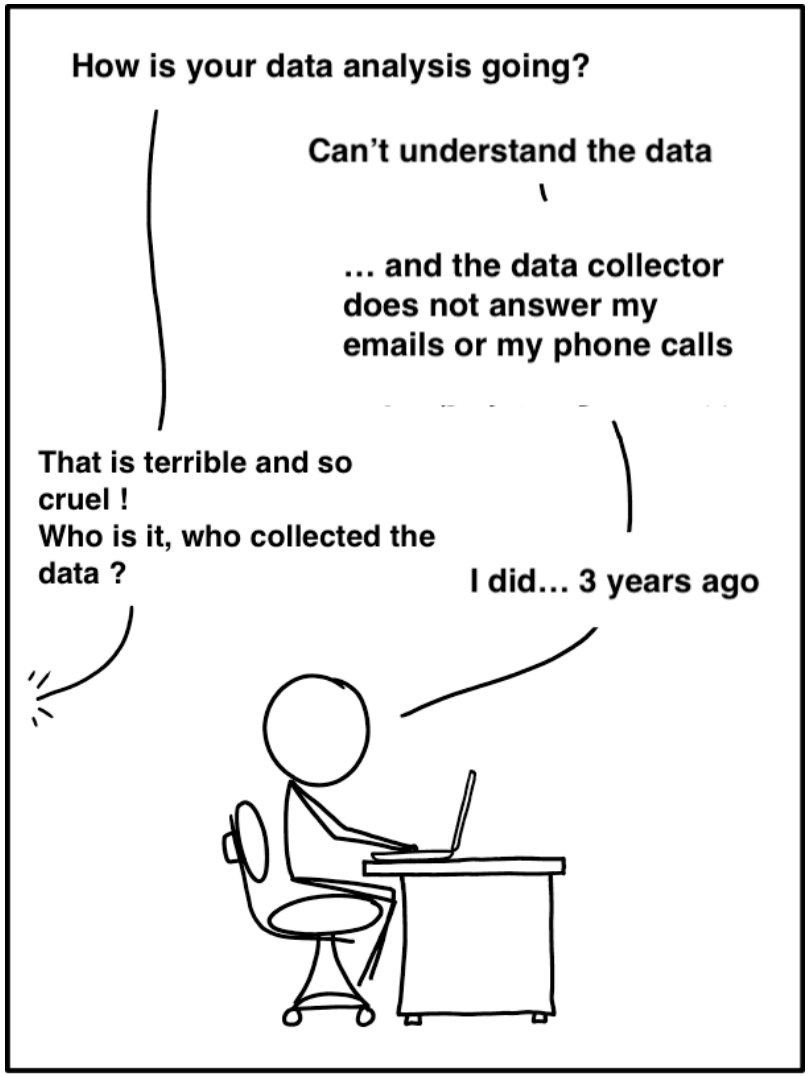
Goals

- 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)?

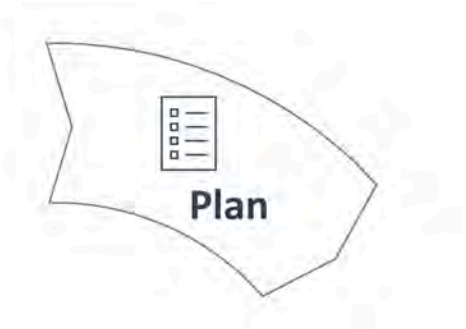
- Increase transparency
- Make data accessible
- Save time (writing, reusing)
- Reduce the risk of data loss
- Optimize the costs
- Facilitate future reuse and sharing
- Improve citations



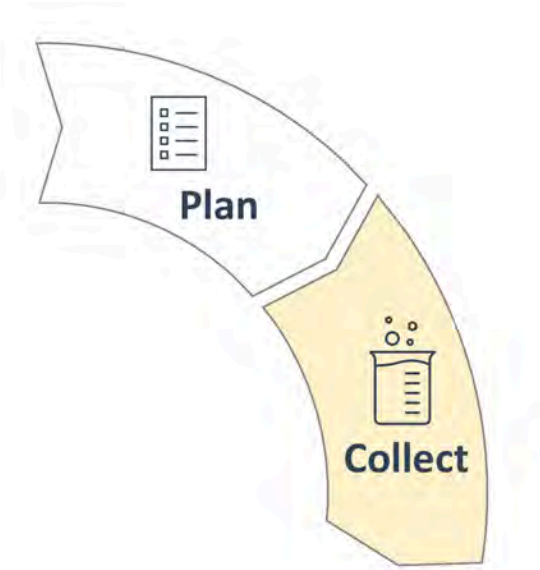
Your first collaborators are your future selves, be nice to them !

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

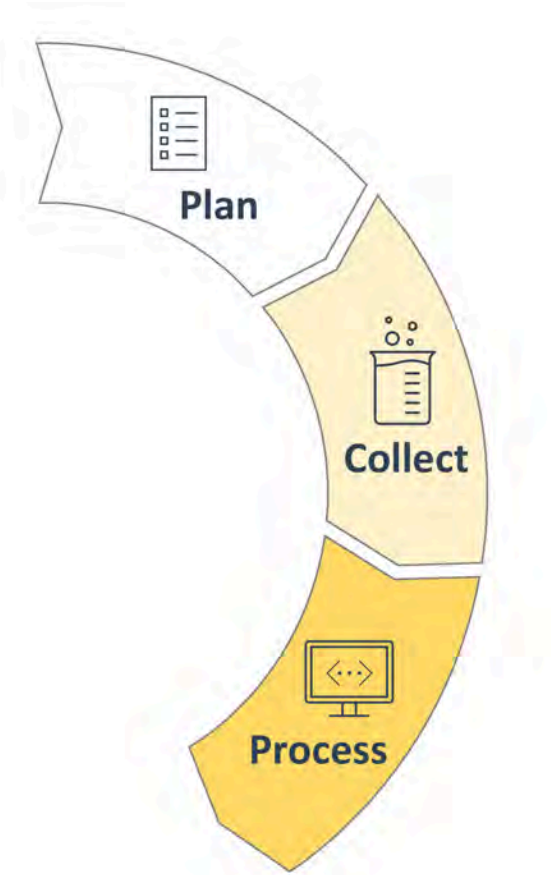
The Research Data Lifecycle



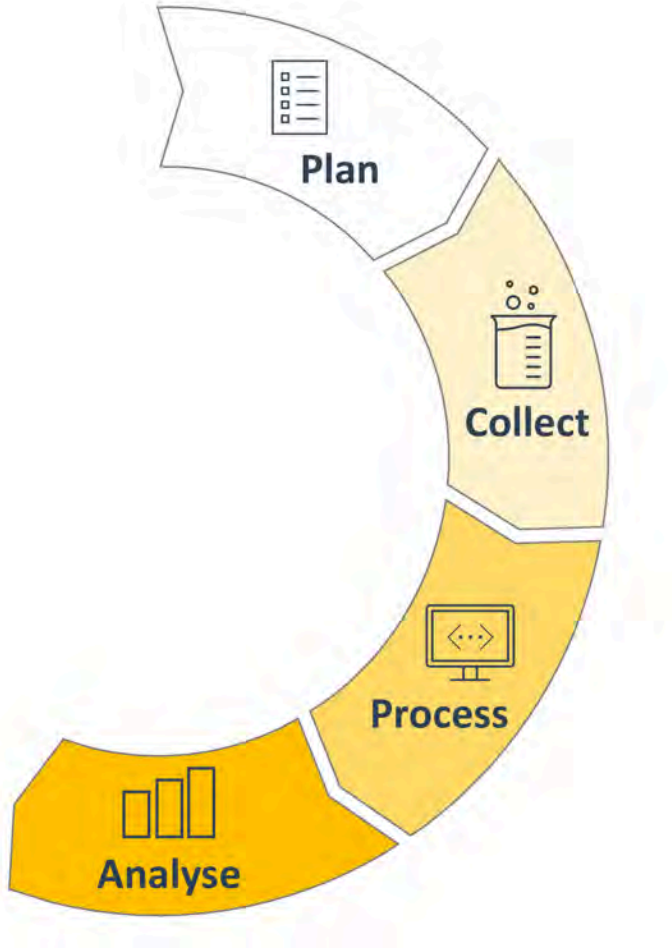
The Research Data Lifecycle



The Research Data Lifecycle



The Research Data Lifecycle

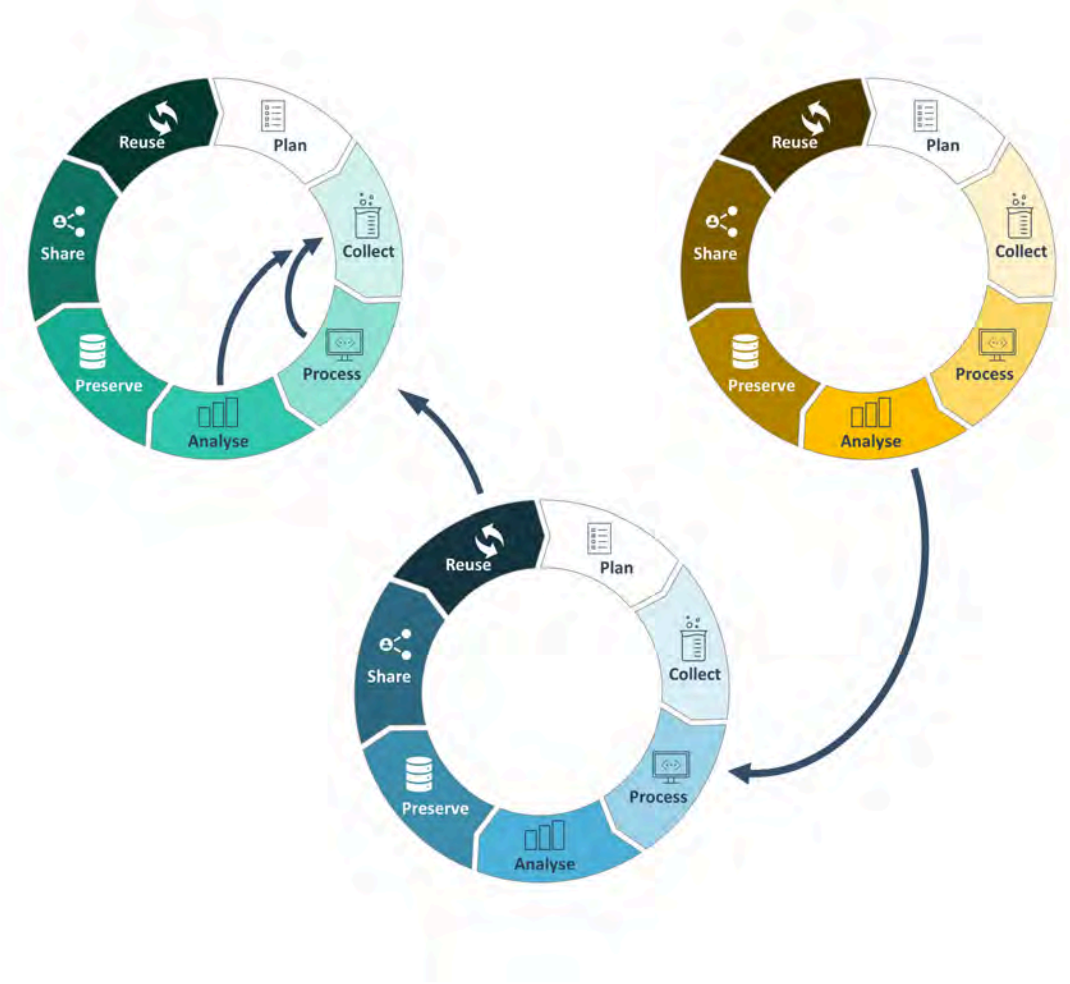


The Research Data Lifecycle

The Research Data Lifecycle

The Research Data Lifecycle

The Research Data Lifecycle *is mutable*

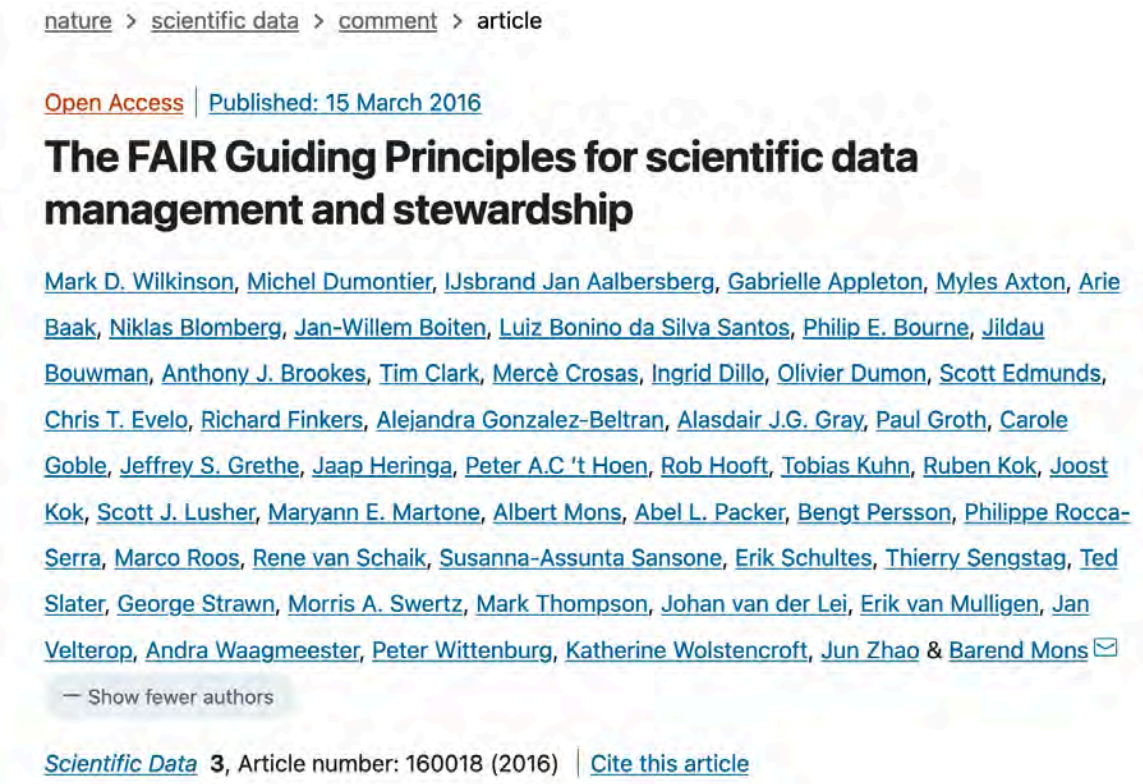


Have you ever heard about the FAIR principles?

FAIR

- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**eusable

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

Is your data FAIR?

Findable | **A**ccessible | **I**nteroperable | **R**eusable

- 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.

https://www.go-fair.org/fair-principles/

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

https://www.go-fair.org/fair-principles/

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.

https://www.go-fair.org/fair-principles/

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

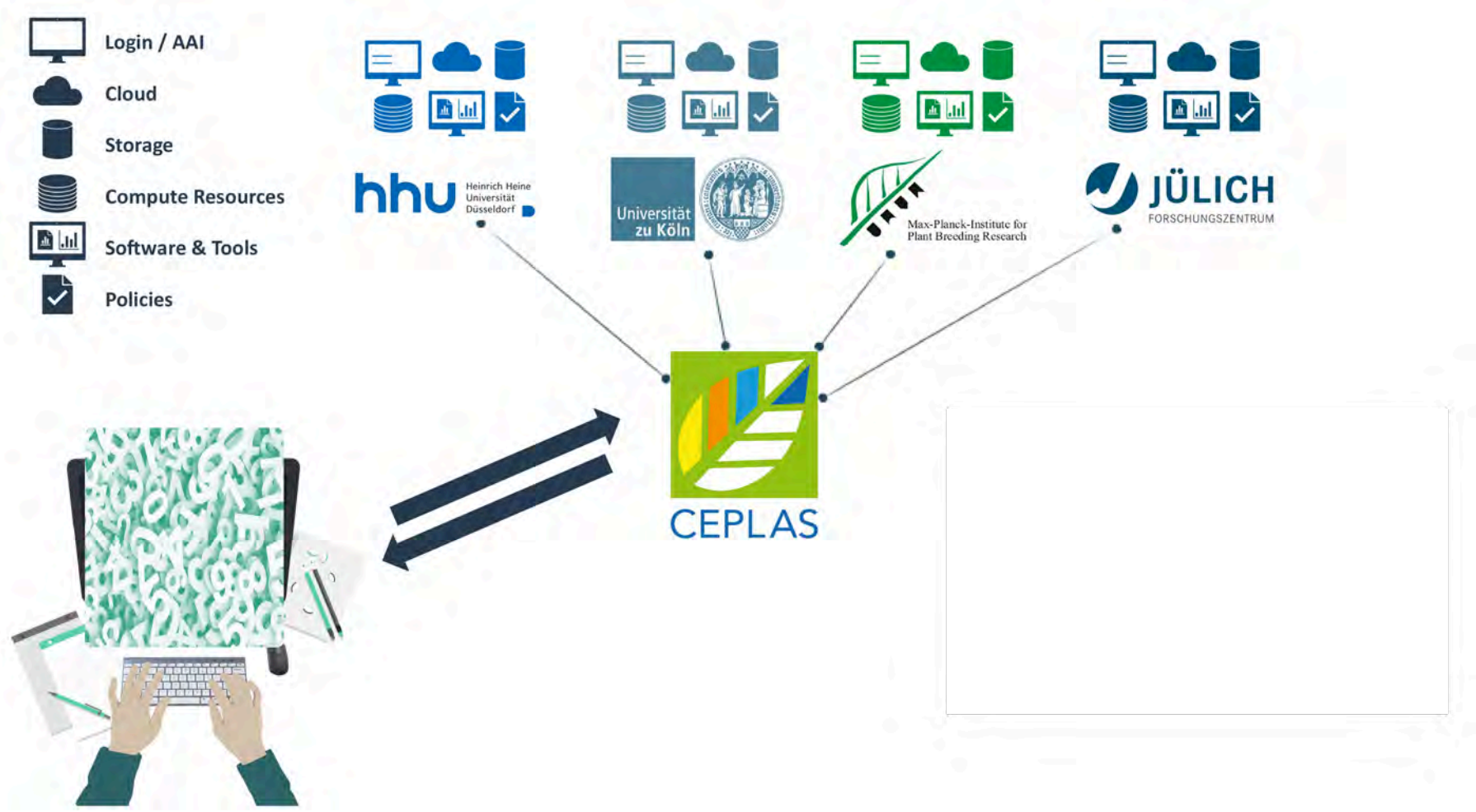
https://www.go-fair.org/fair-principles/

FAIR on multiple layers

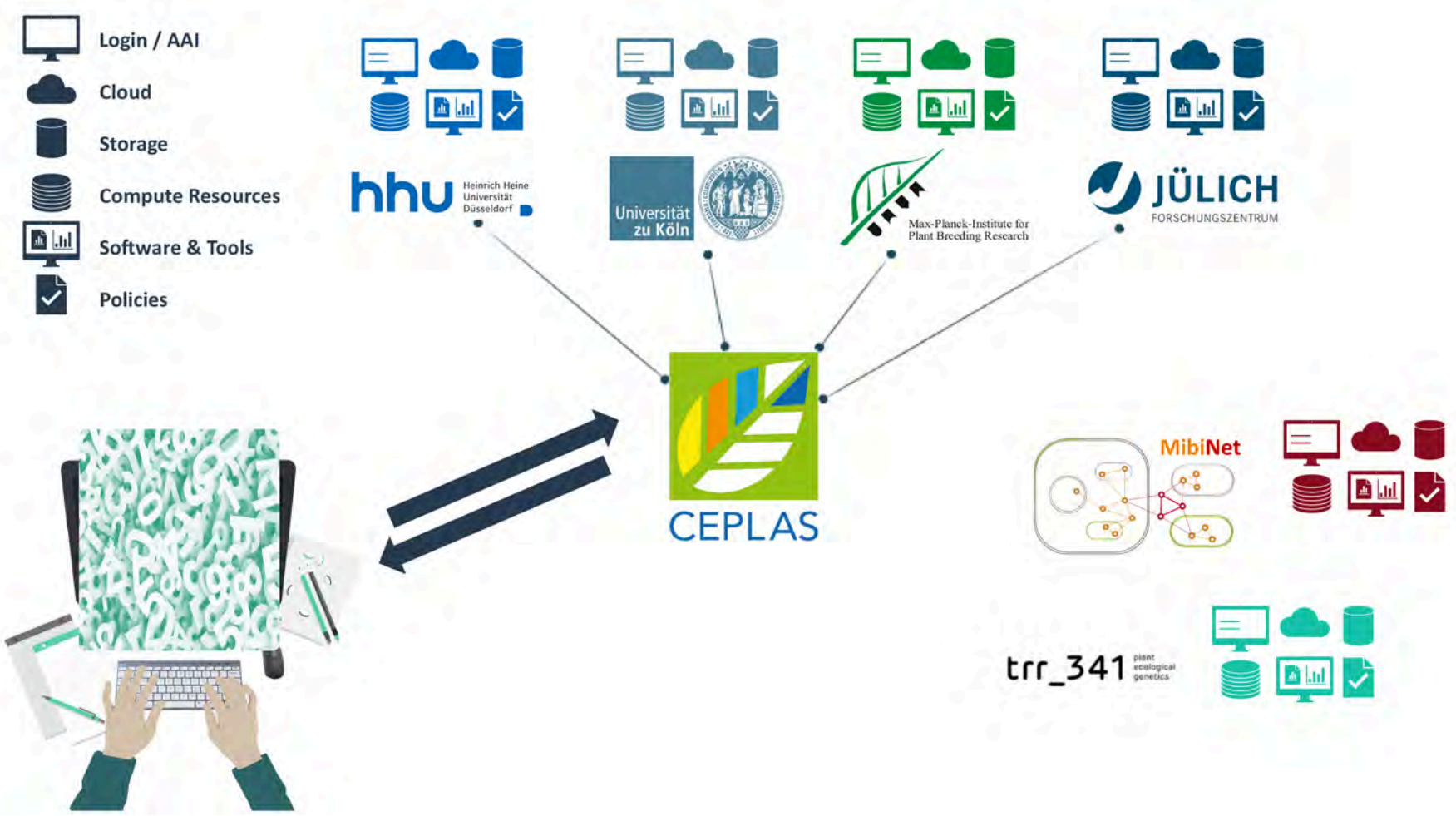
> The principles refer to three types of entities: **data** (or any digital object), **metadata** (information about that digital object), and **infrastructure**.

https://www.go-fair.org/fair-principles/

Scattered Data Silos



Scattered Data Silos



FAIR Data for everyone

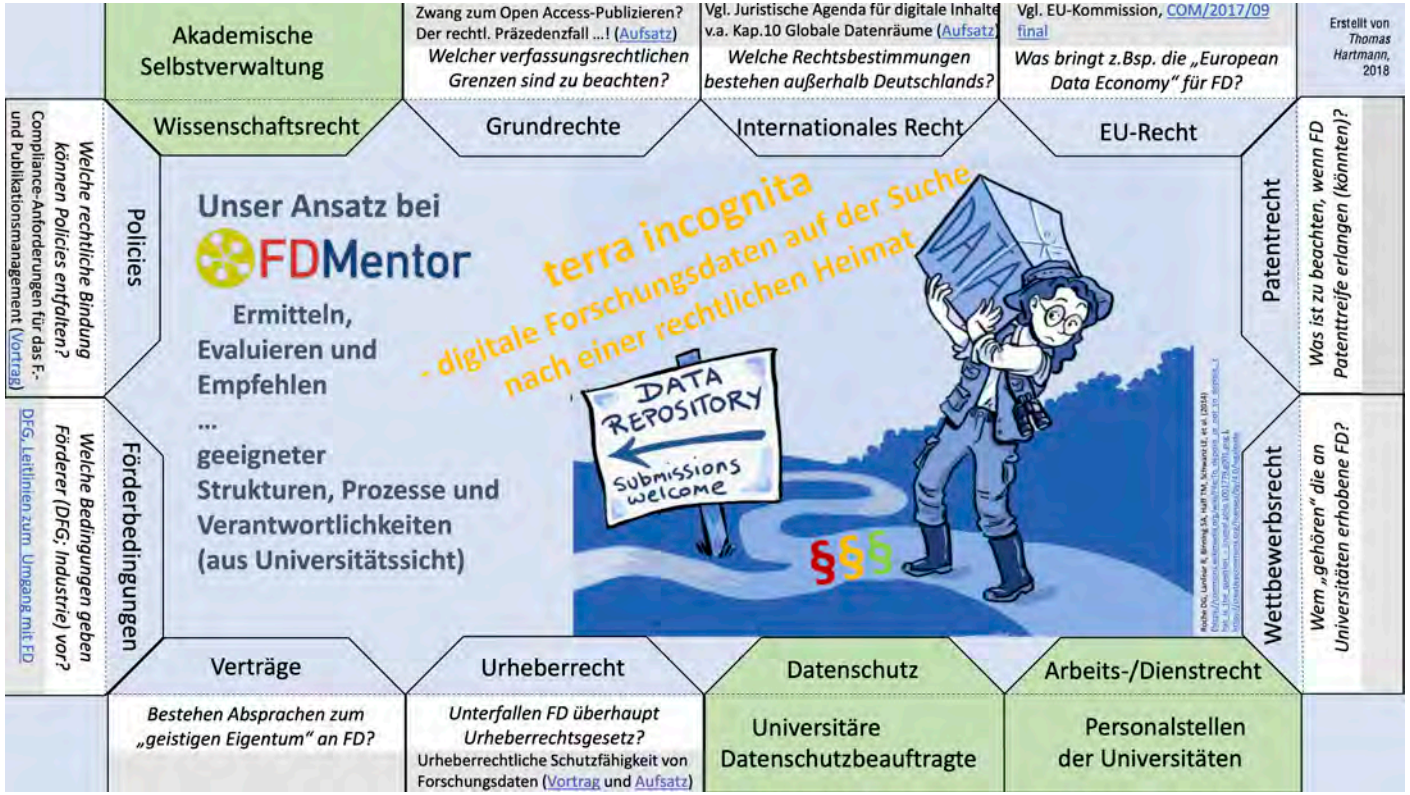
RDM fundamentals

Dominik Brilhaus

Sept 20th, 2023

Legal aspects of RDM

Different laws touched by RDM

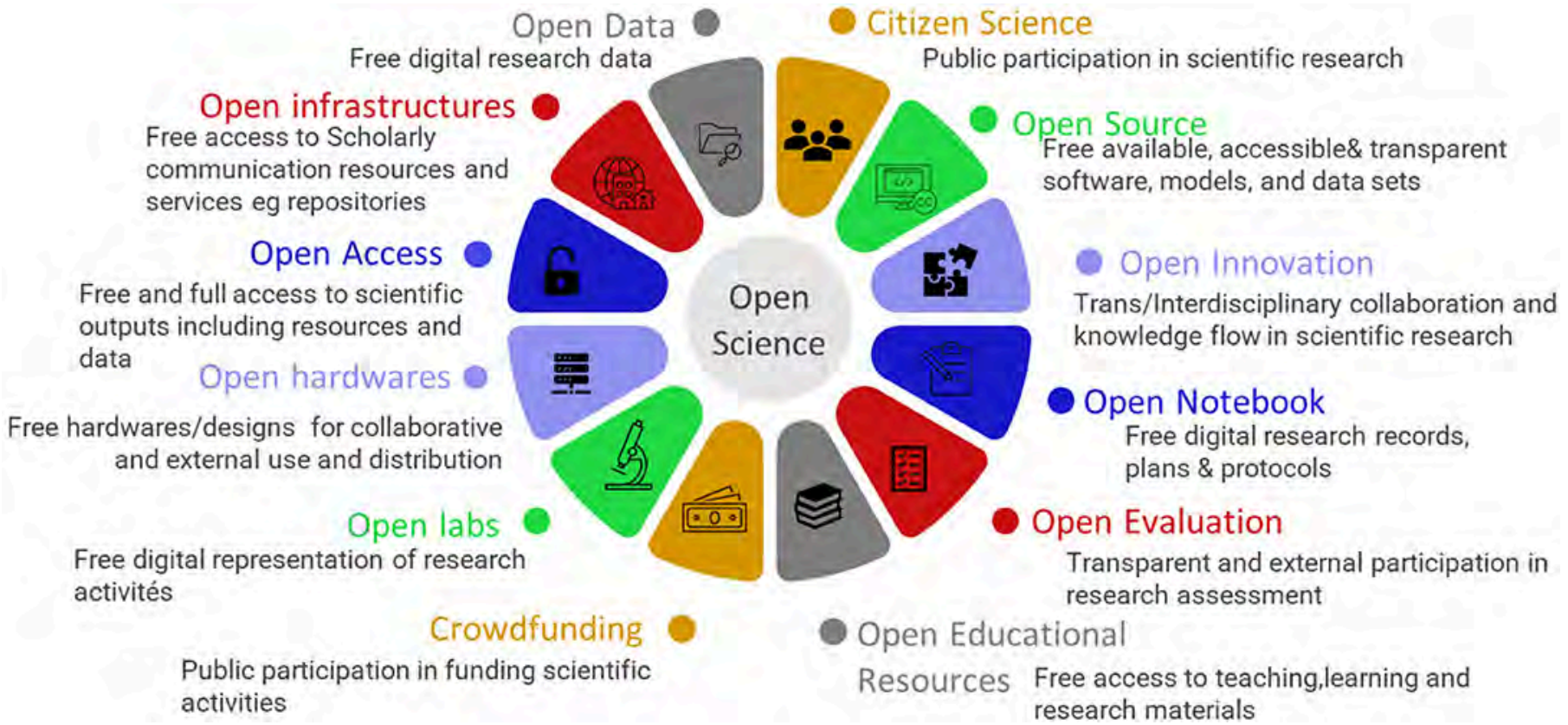


Hartmann, Thomas. (2019). Rechtsfragen: Institutioneller Rahmen und Handlungsoptionen für universitäres FDM. Zenodo. https://doi.org/10.5281/zenodo.2654306

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.

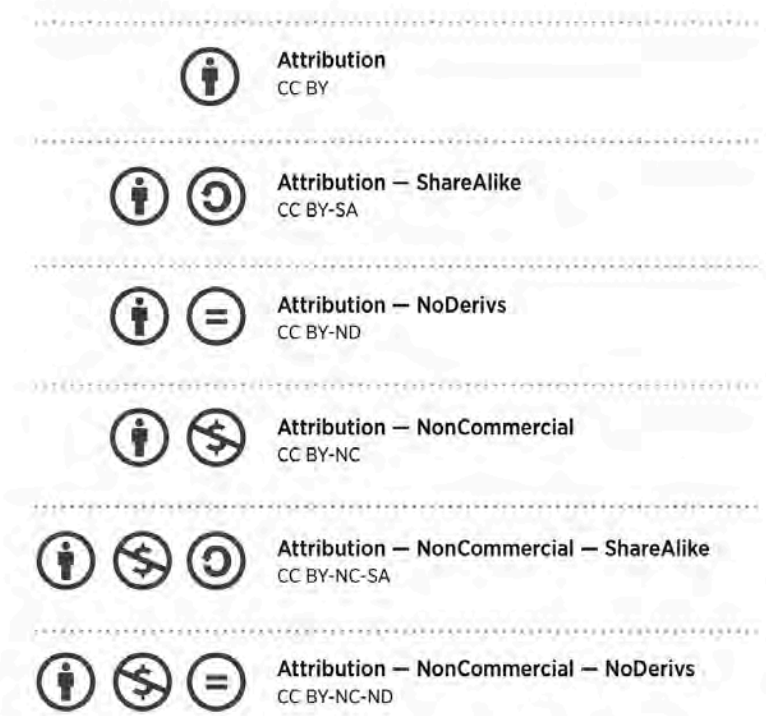
Open Science is more than Open Access



Okafor et al. (2022) https://doi.org/10.3389/frma.2022.855198, Figure 1

Creative commons

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



adapted from https://wiki.creativecommons.org/images/0/01/6licenses-folded.pdf

Data protection

GDPR: General Data Protection Regulation

DS-GVO (german): Datenschutz-Grundverordnung

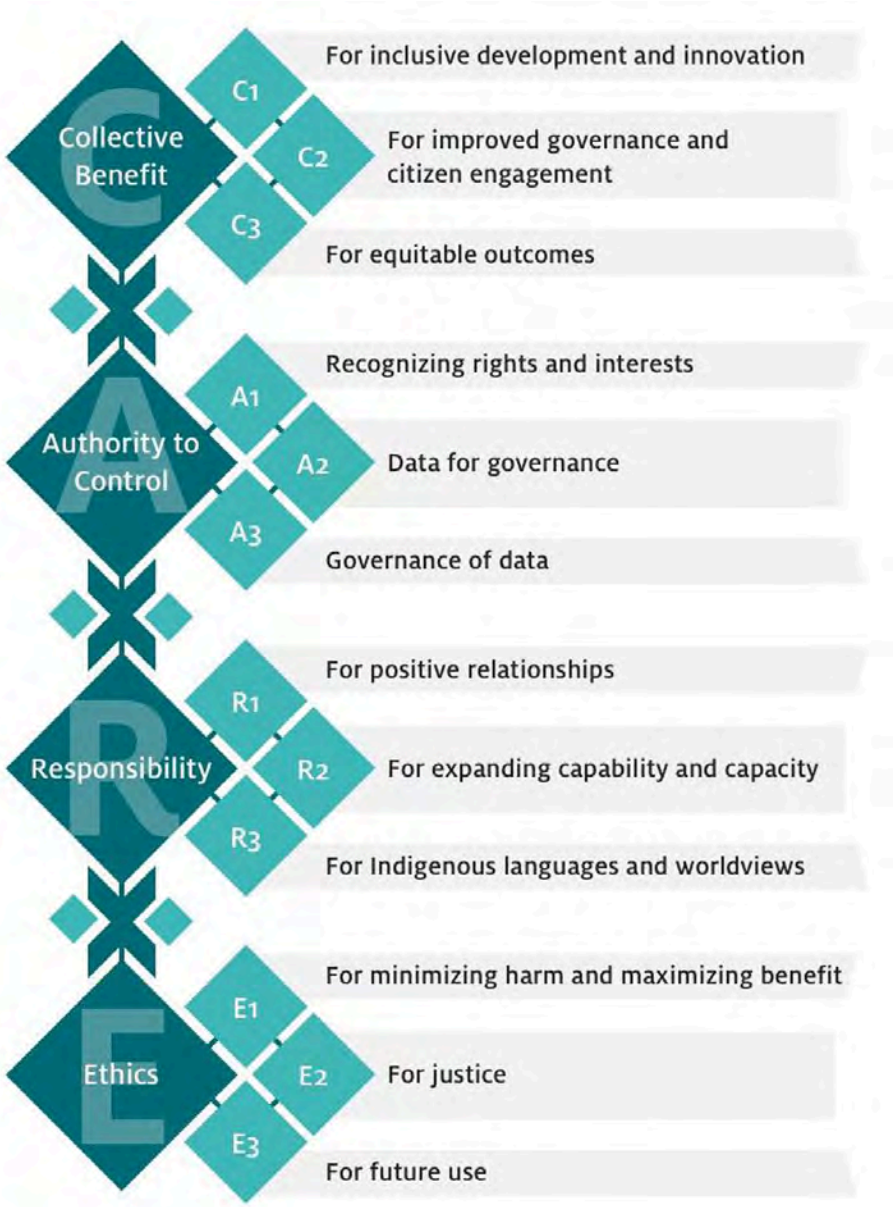
Use of biological materials

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

FAIR and CARE

https://www.gida-global.org/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

Hiemenz, Bea & Kuberek, Monika (2018) http://dx.doi.org/10.14279/depositonce-7521

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-BioImage 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

Intermediate

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

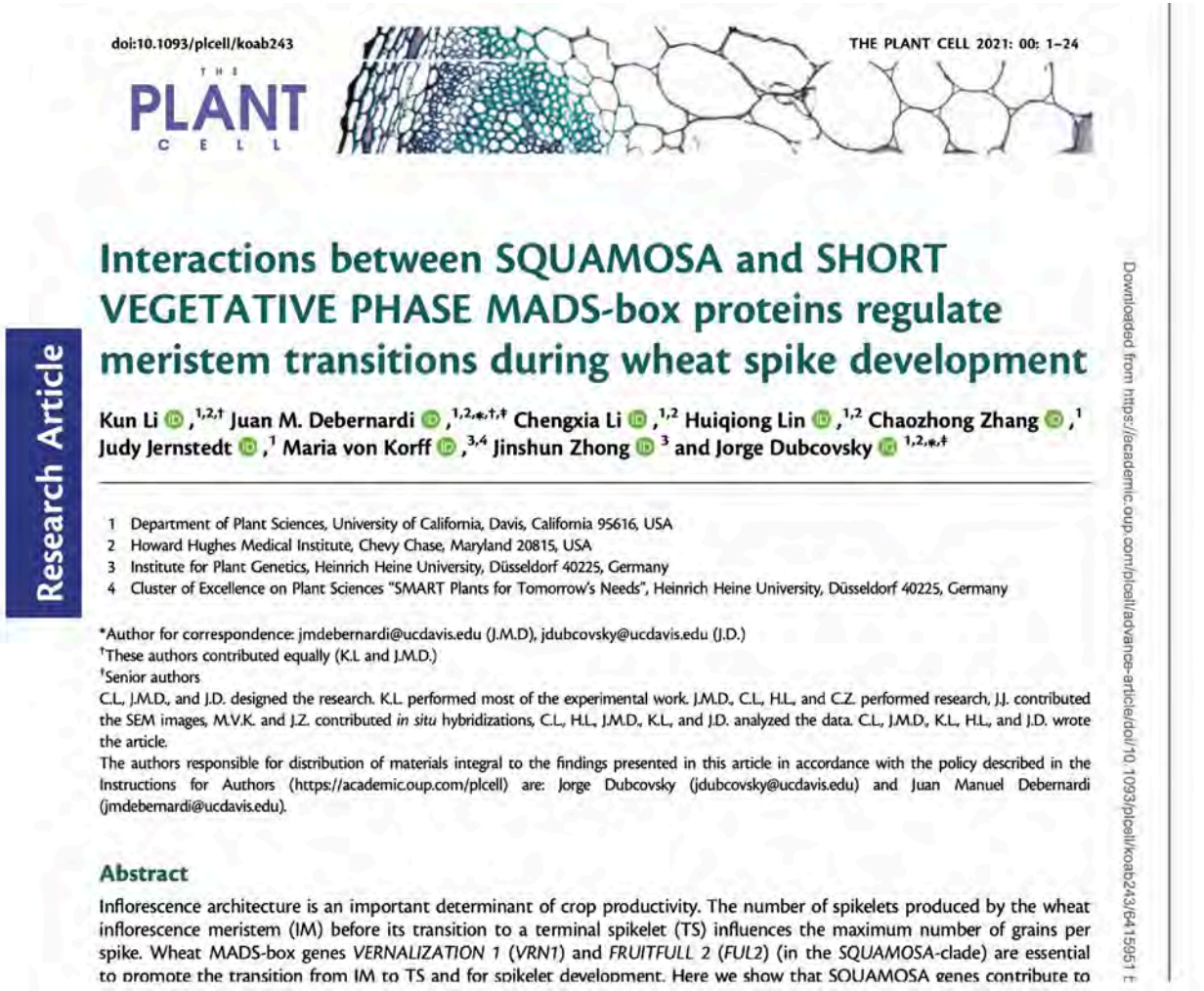
https://zenodo.org

https://datadryad.org/

https://figshare.com

Peristent Identifiers (PIDs)

Spot the PIDs



doi:10.1093/plcell/koab243

THE PLANT CELL

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

1 Department of Plant Sciences, University of California, Davis, California 95616, USA
2 Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA
3 Institute for Plant Genetics, Heinrich Heine University, Düsseldorf 40225, Germany
4 Cluster of Excellence on Plant Sciences "SMART Plants for Tomorrow's Needs", Heinrich Heine University, Düsseldorf 40225, Germany

*Author for correspondence: jmdebernardi@ucdavis.edu (J.M.D), jdubcovsky@ucdavis.edu (J.D.)
†These authors contributed equally (K.L and J.M.D.)
‡Senior authors

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

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

Abstract

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

https://doi.org/10.1093/plcell/koab243

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)

Adapted from https://www.ebi.ac.uk/rdf/documentation/good_practice_uri/

Additional resources

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

Data stores

Institute server

University server

Local hard disks

Cloud services

sciebo

Chat

Your data

Mail

Wiki, Project management

eLabFTW

labfolder

Electronic lab notebooks

Git repositories

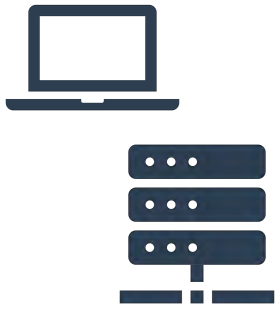
Backup vs. Archive

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

3-2-1 backup rule

3 copies of data

2 storage media



1 copy off-site

Version control and track changes

It’s good practice to document:

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

Types of Version Control

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

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

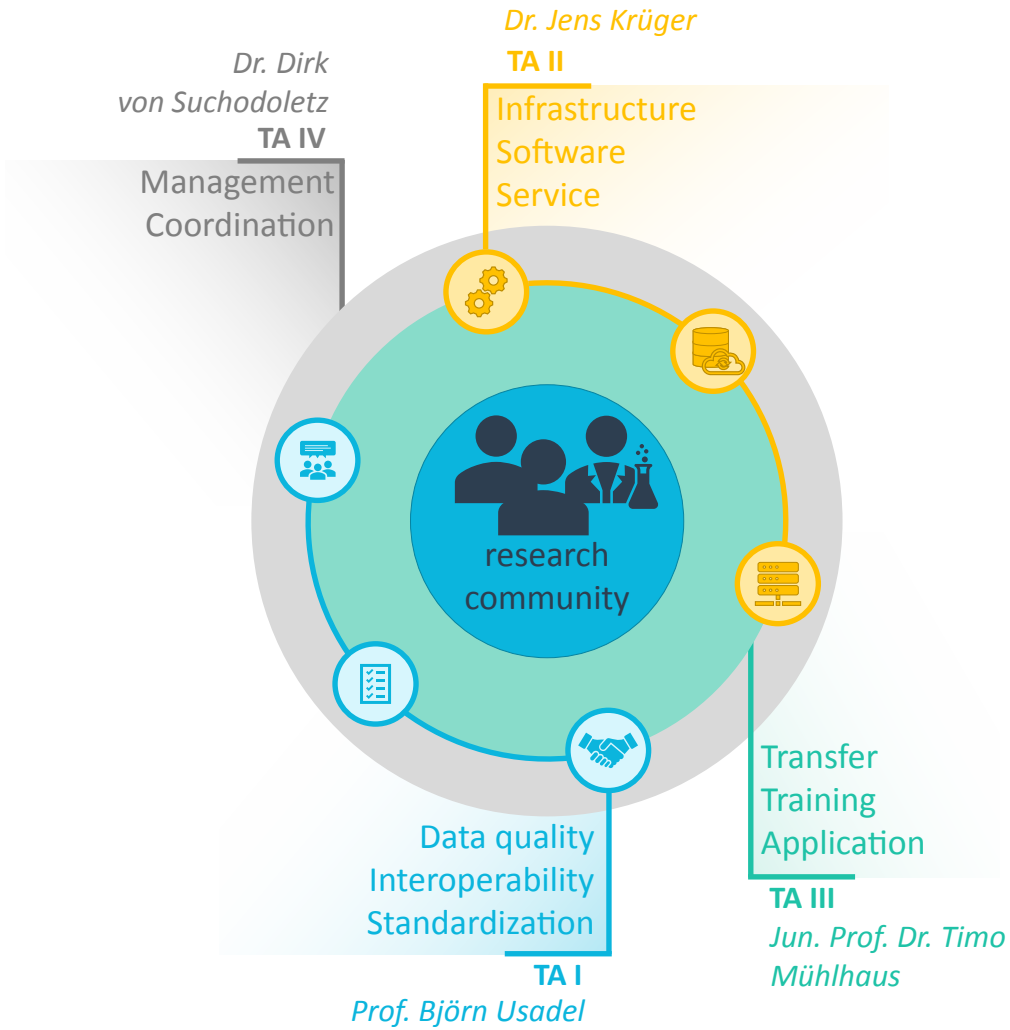
Slides presented here include contributions by

- 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

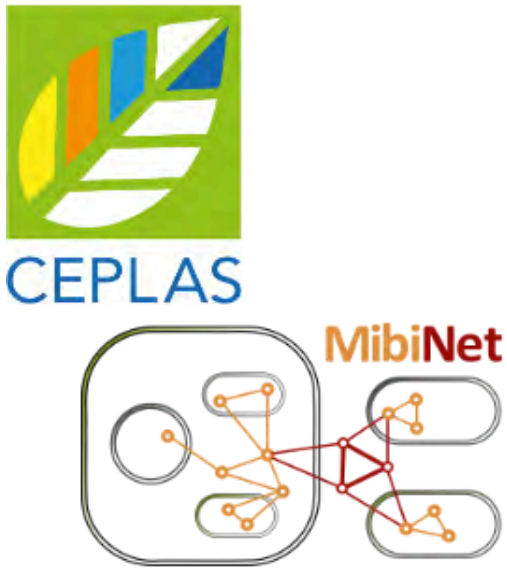
- NFDI: "Nationale Forschungsdaten Infrastruktur" – www.nfdi.de
- Funded since end of 2020

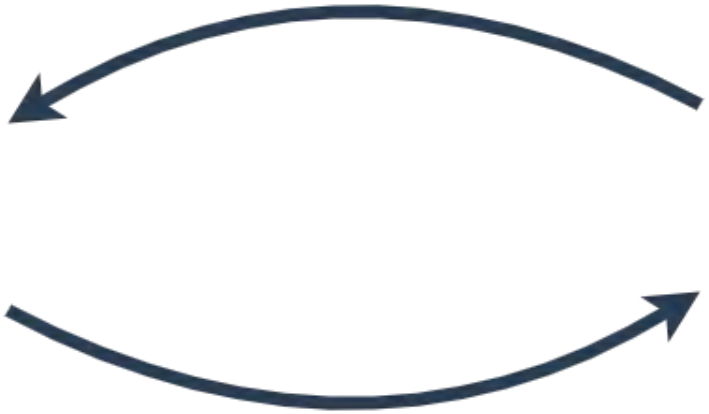


Data Stewardship between DataPLANT and the community

Community

nfdi4plants

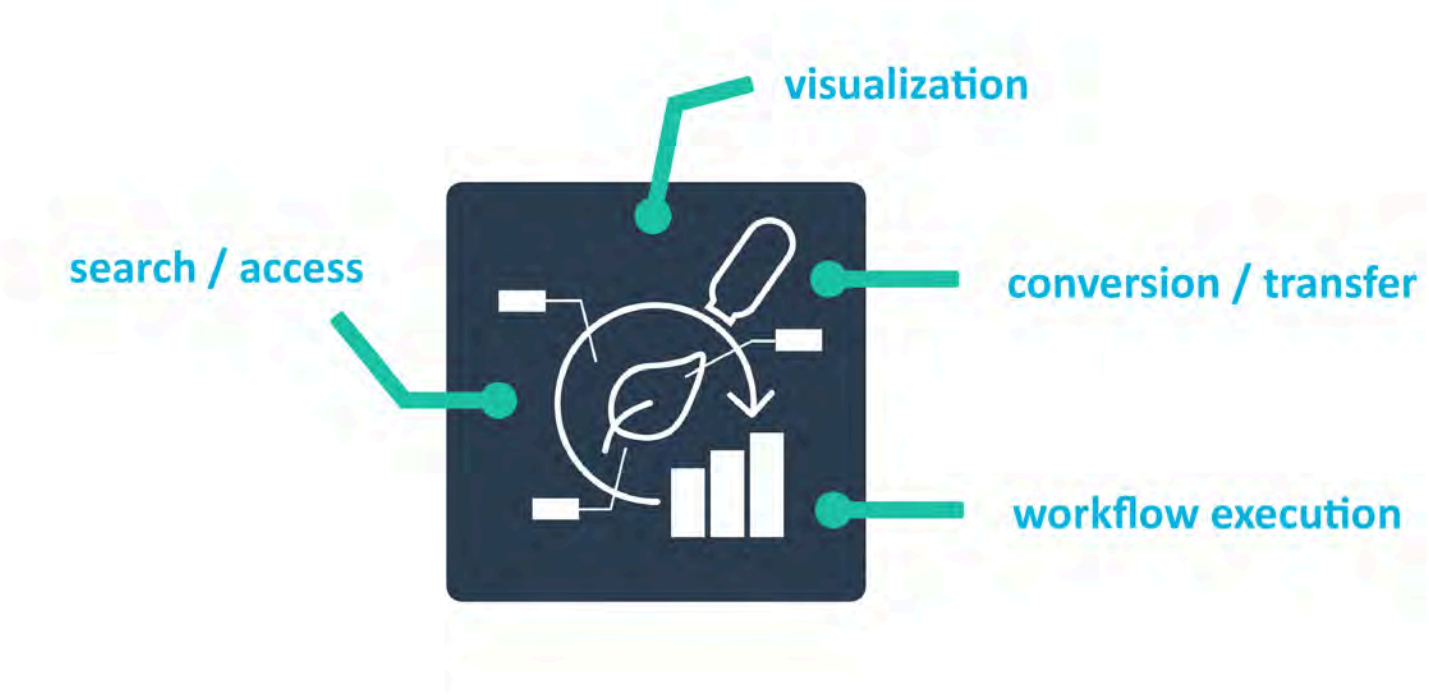




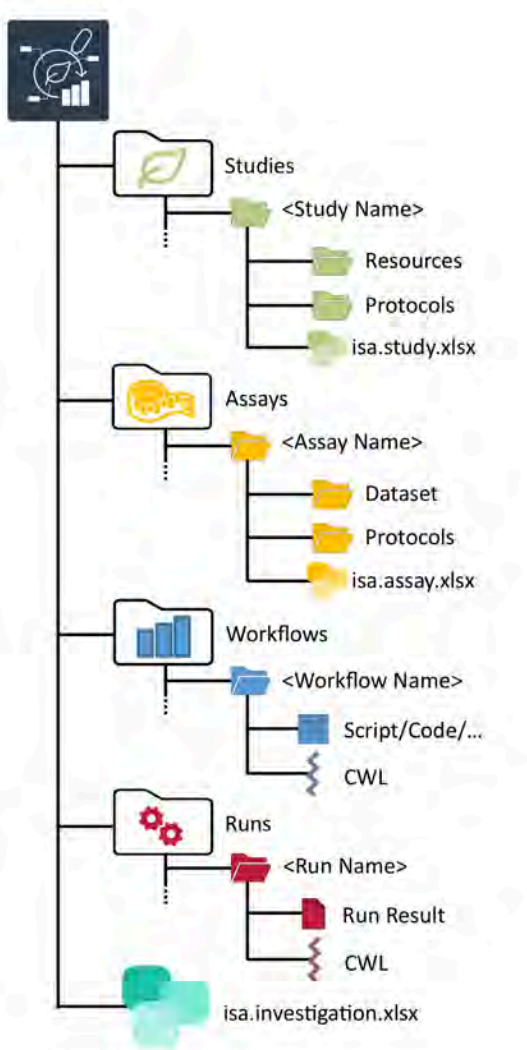
Domain experts
User experience
Training

Service provider
Developers
Tech experts

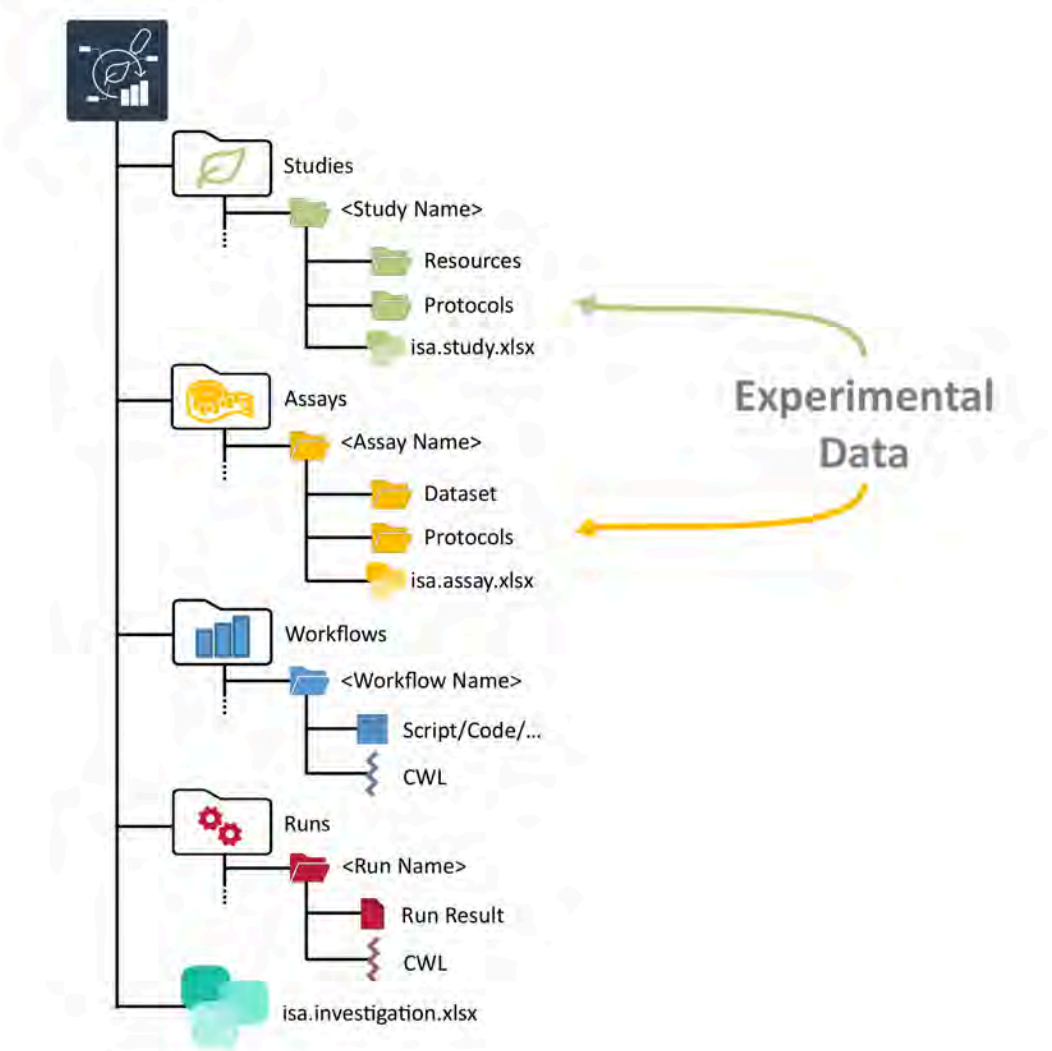
Annotated Research Context (ARC)



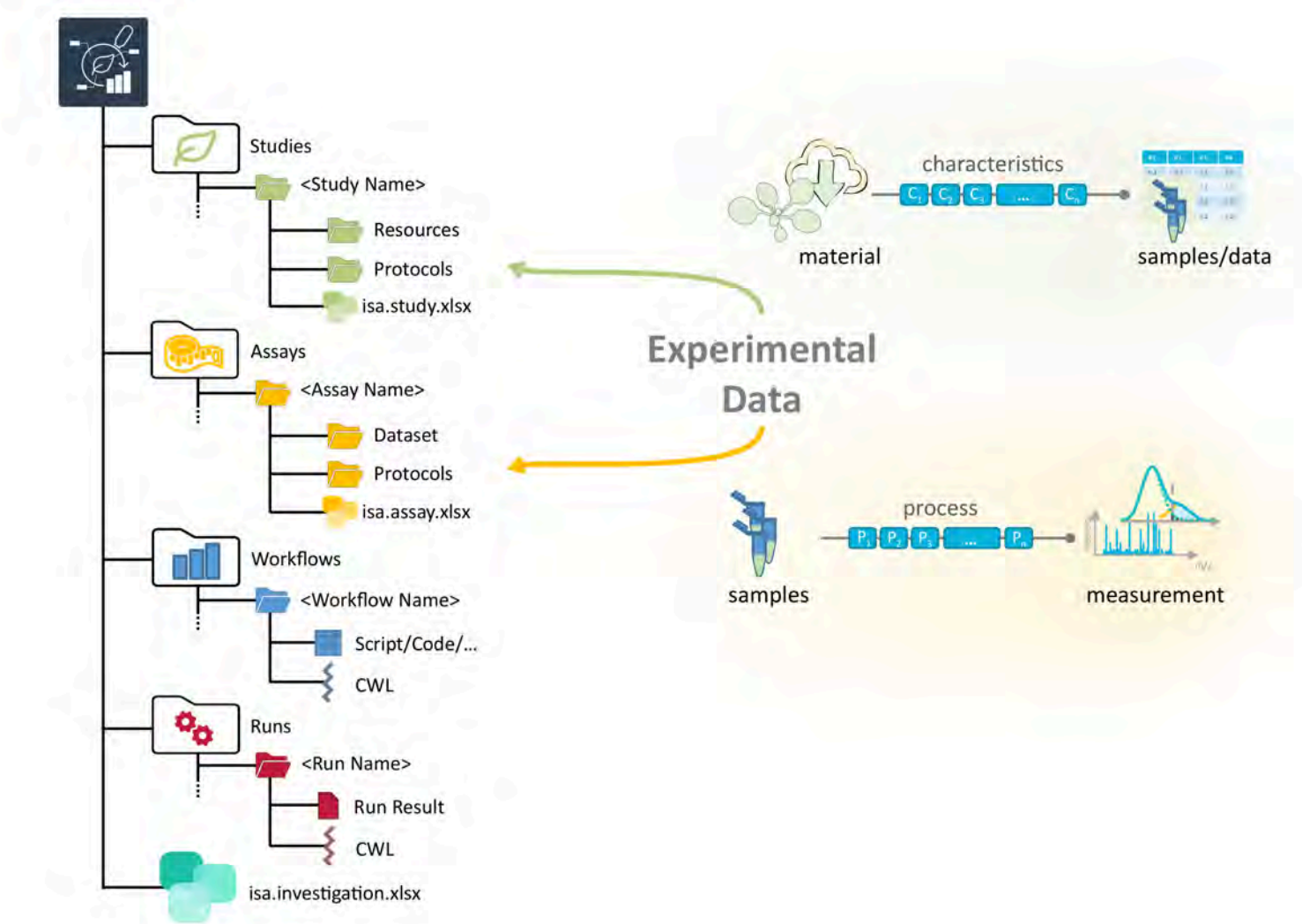
What does an ARC look like?



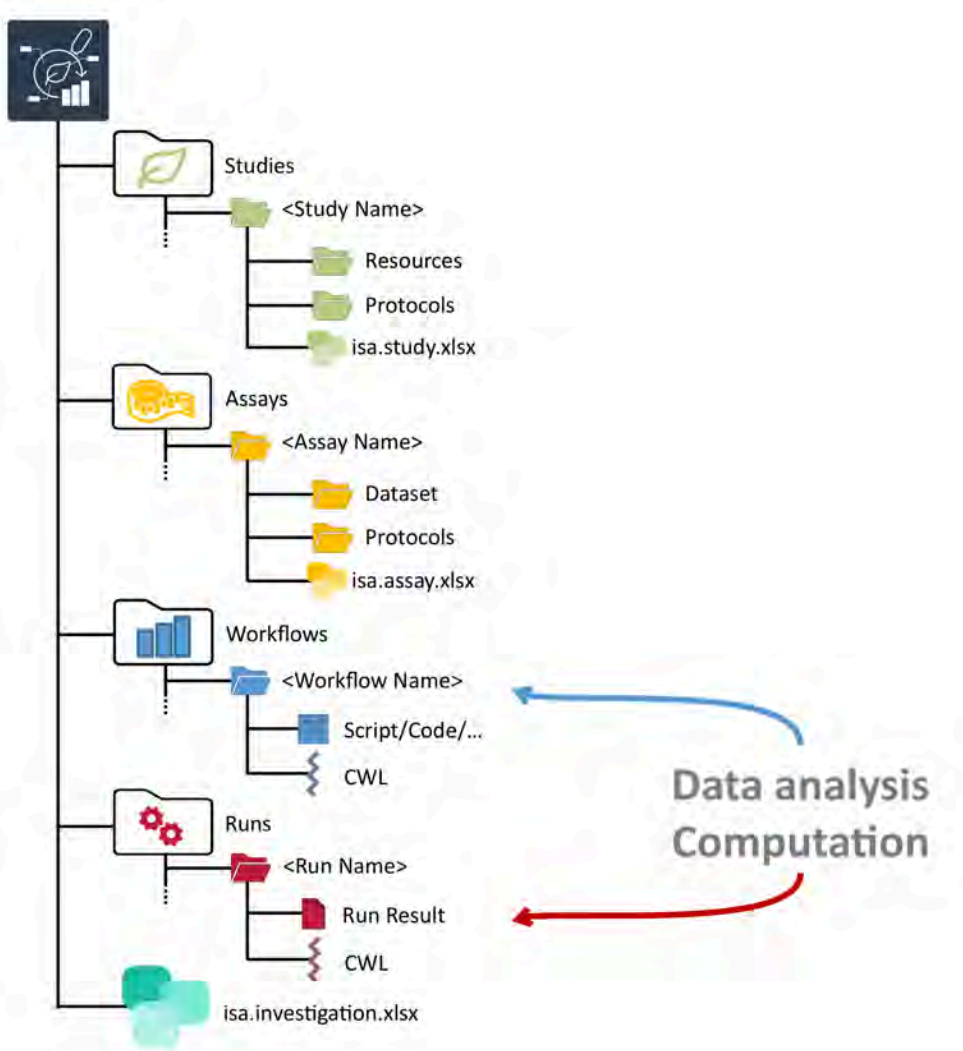
What does an ARC look like?



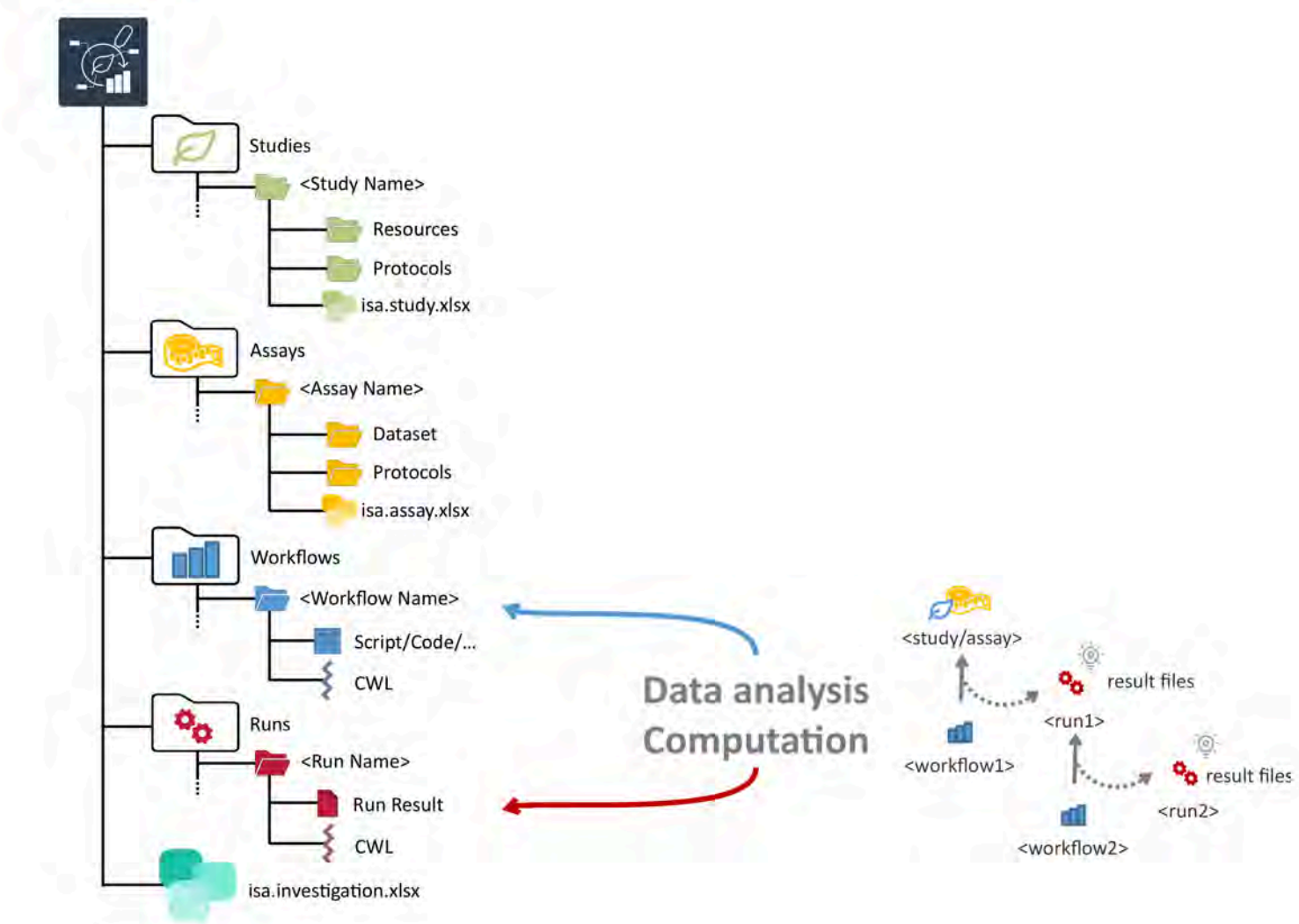
What does an ARC look like?



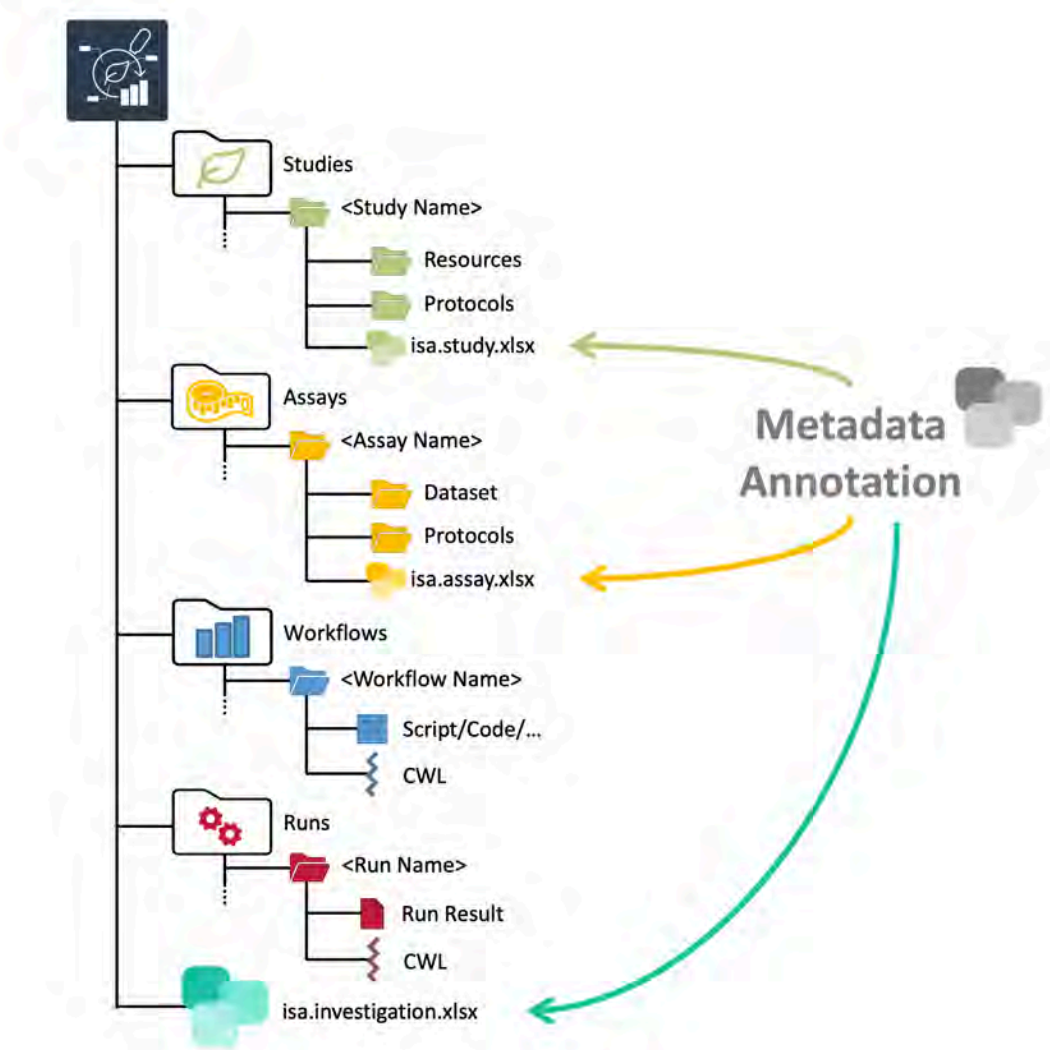
What does an ARC look like?



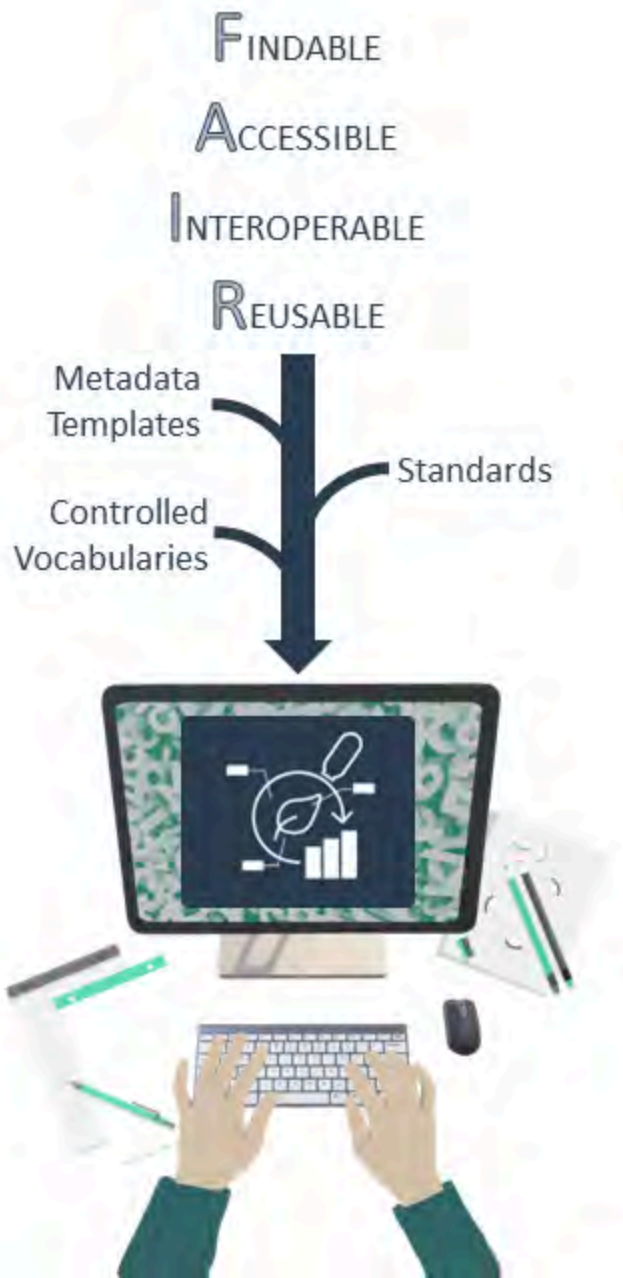
What does an ARC look like?



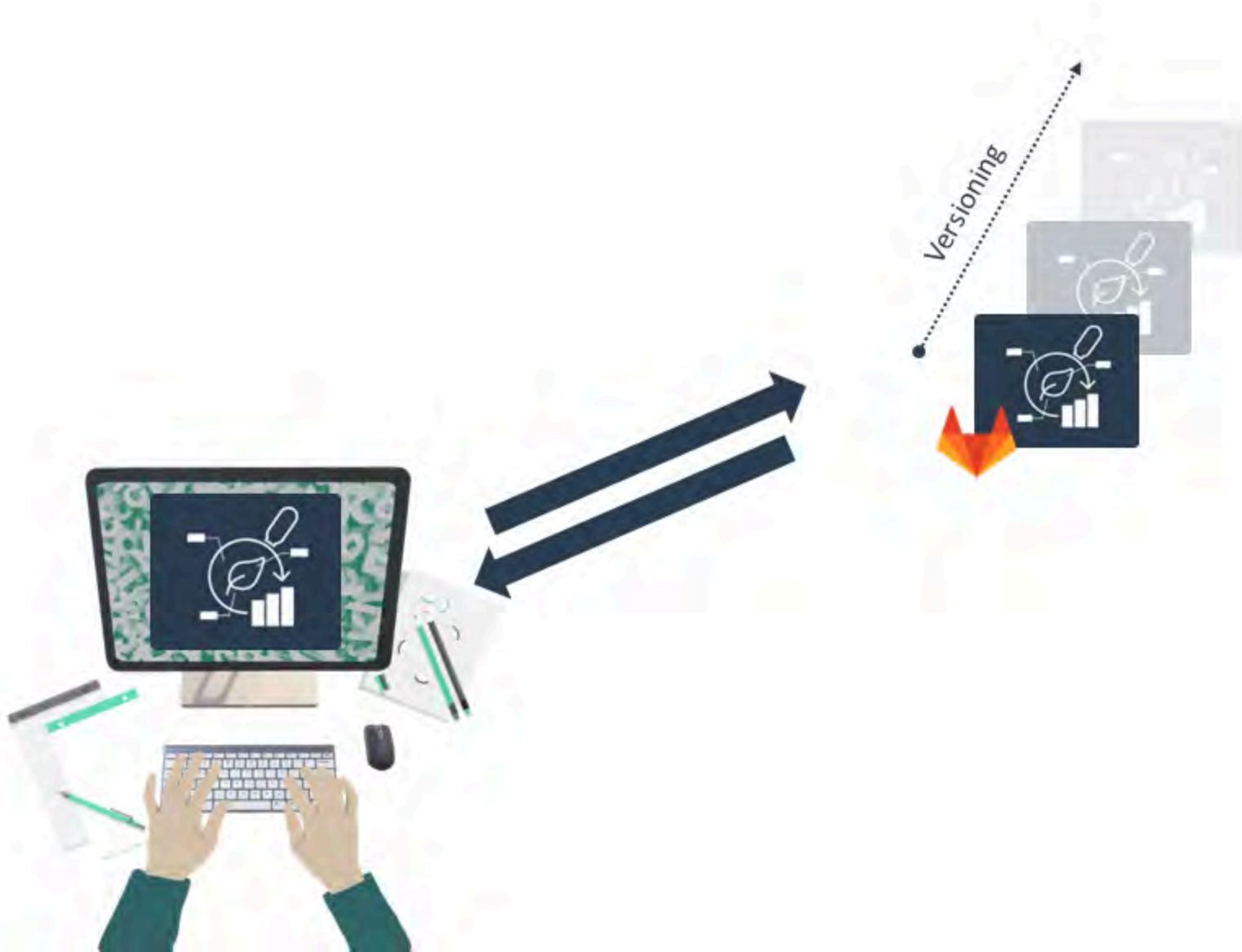
What does an ARC look like?




Repository
Collaboration
Standard
Provenance
Metadata
Computation
Ontology

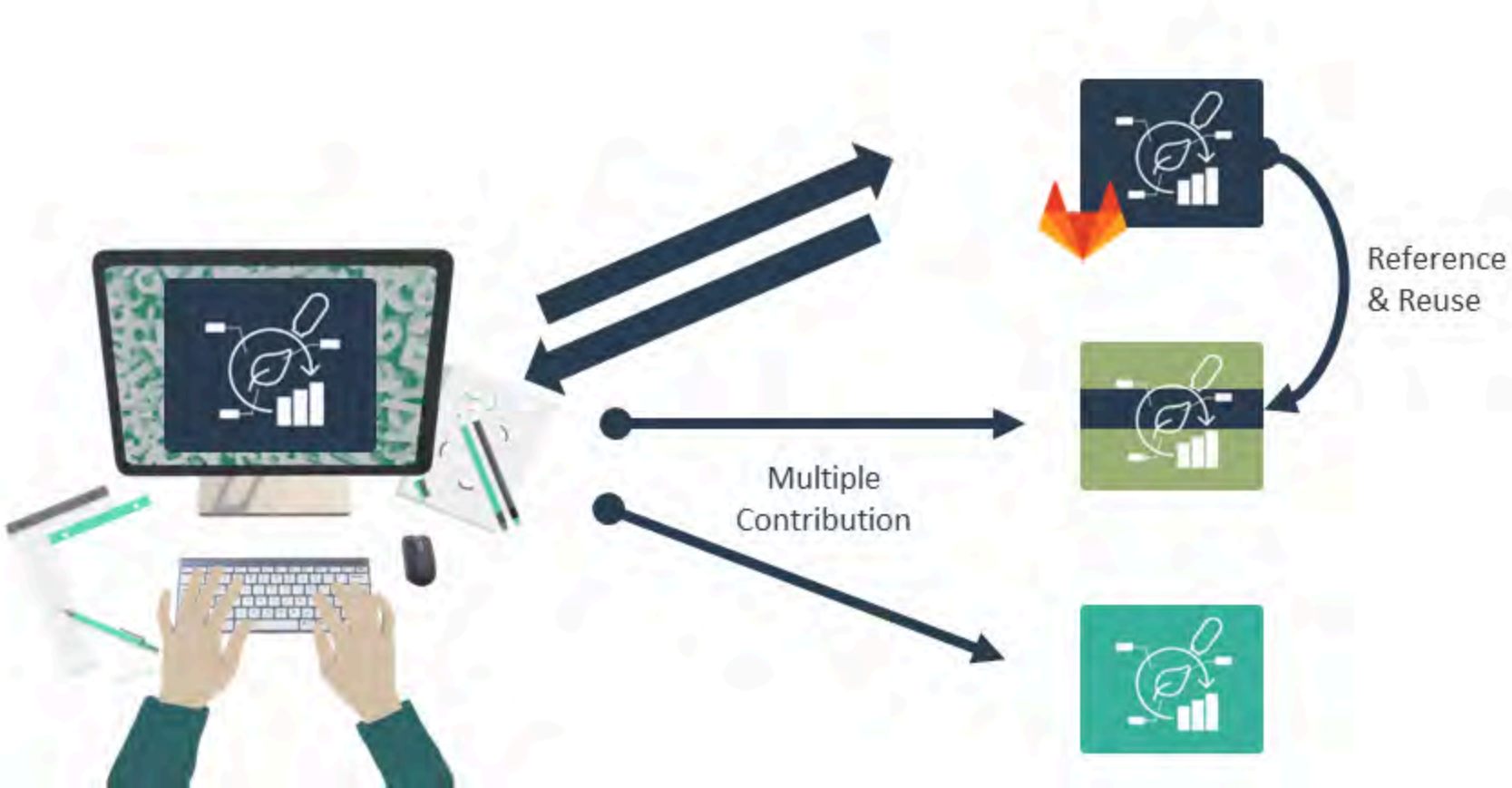

FINDABLE
ACCESSIBLE
INTEROPERABLE
REUSABLE
Metadata Templates
Controlled Vocabularies
Standards

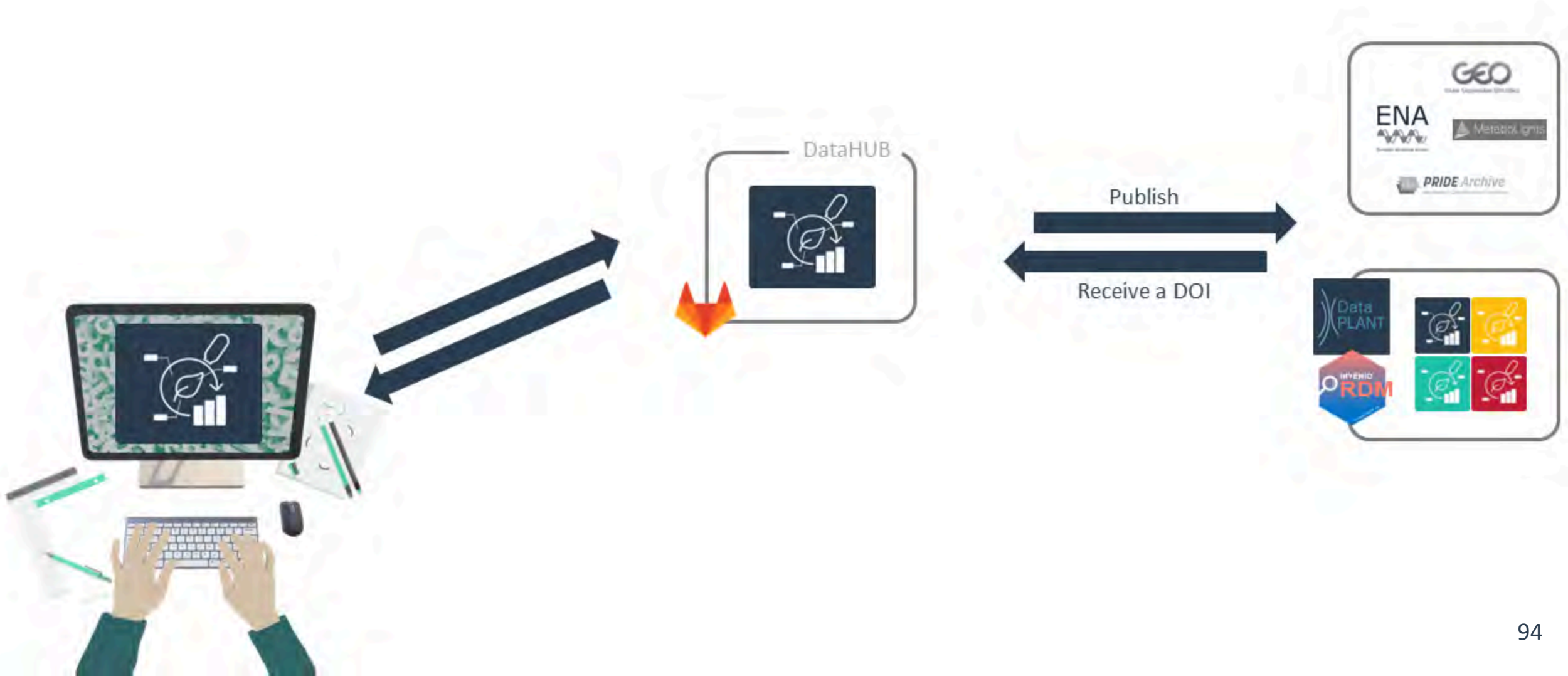

DataHUB
Storage & Backup

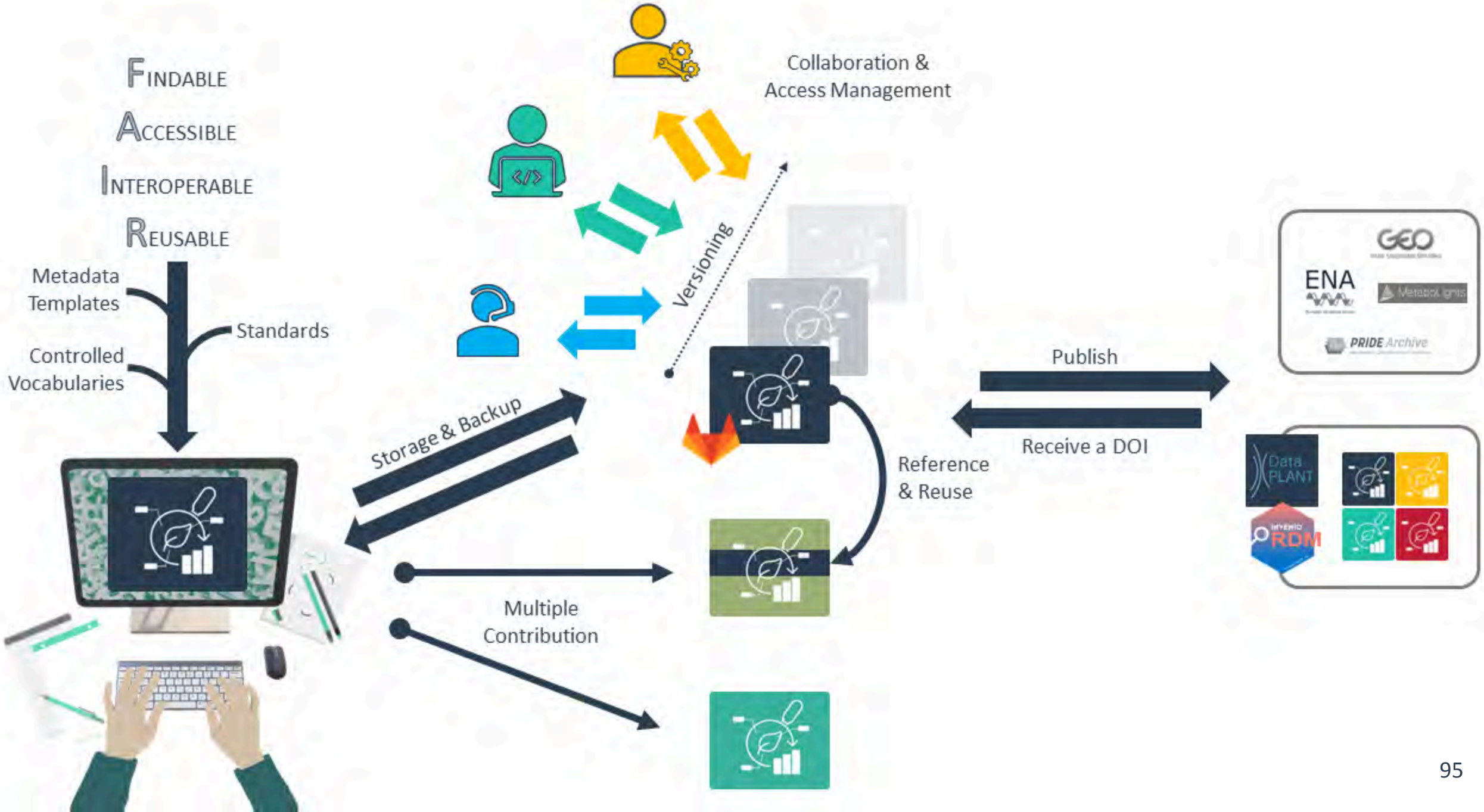

Versioning


Collaboration &
Access Management


Multiple
Contribution


Reference
& Reuse
Multiple
Contribution

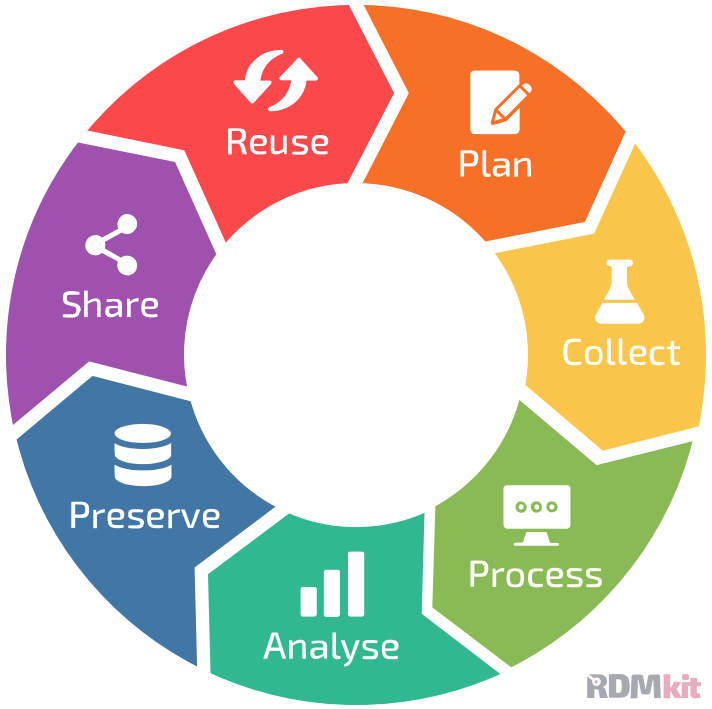

DataHUB
Publish
Receive a DOI
GEO
ENA
MetaboLights
PRIDE Archive
Data PLANT
INVENIO RDM


FINDABLE
ACCESSIBLE
INTEROPERABLE
REUSABLE
Metadata Templates
Controlled Vocabularies
Standards
Collaboration & Access Management
Versioning
Storage & Backup
Multiple Contribution
Reference & Reuse
Publish
Receive a DOI
GEO
ENA
MetaboLights
PRIDE Archive
Data PLANT
INVENIO RDM

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

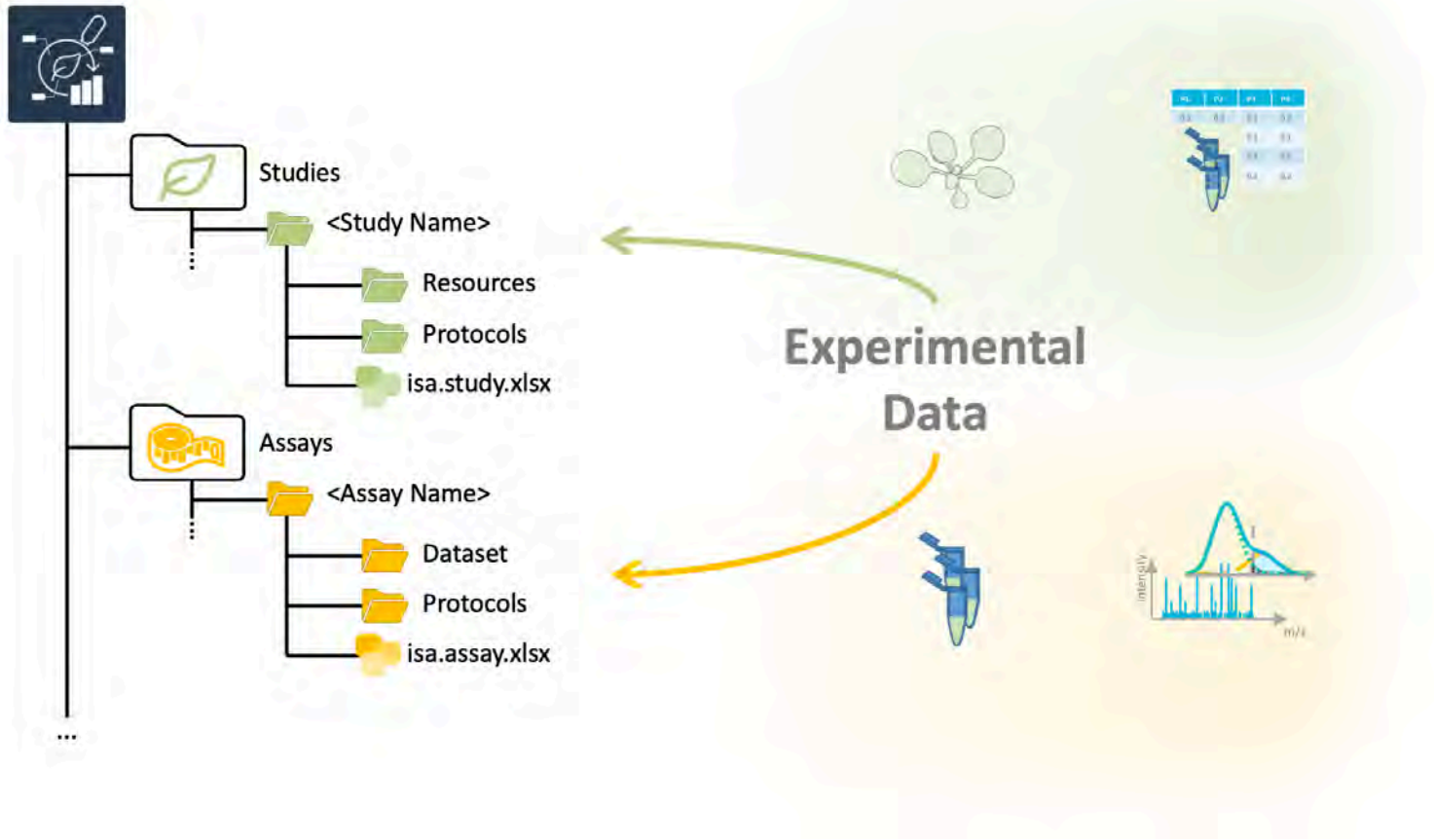


ARC Ecosystem Demo

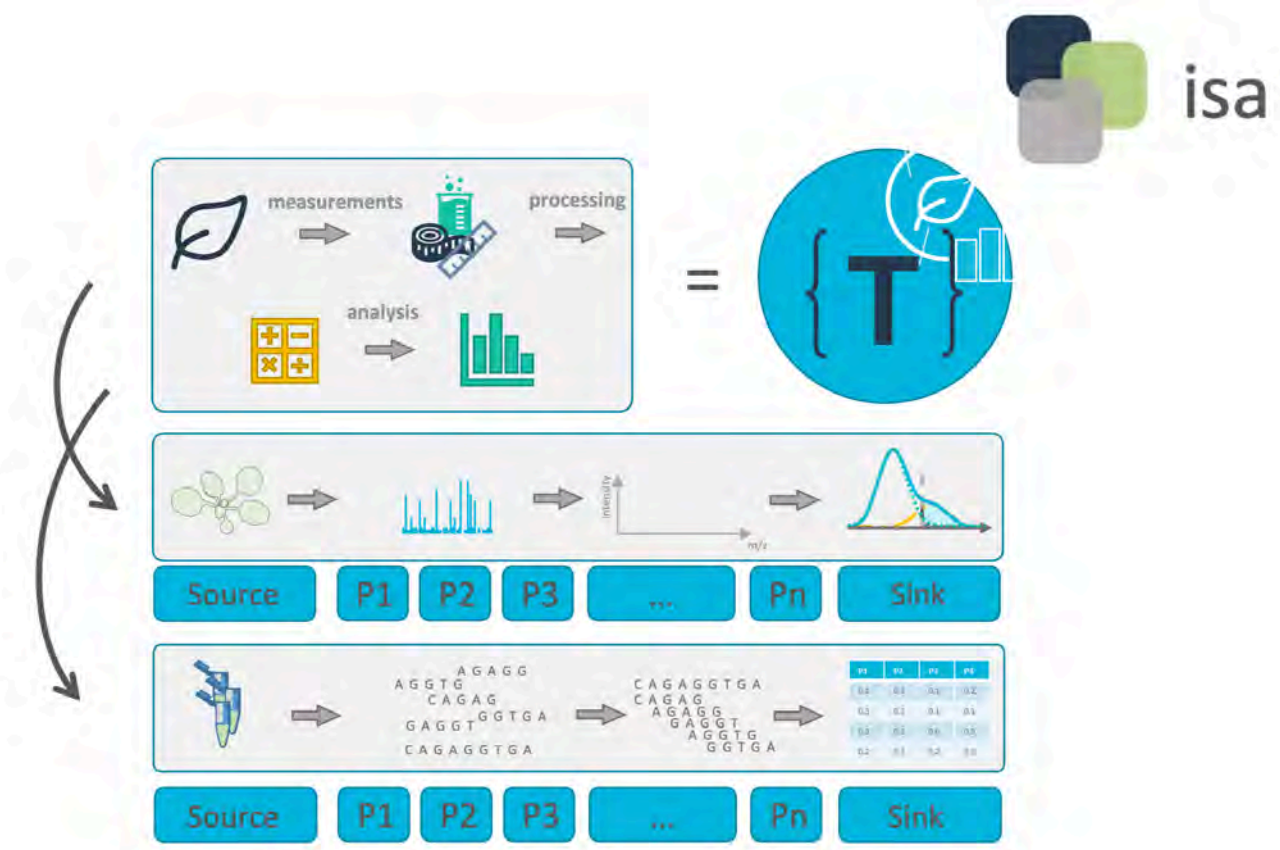
"A FAIR RDM journey along a (mutable) data life cycle"

Dominik Brilhaus

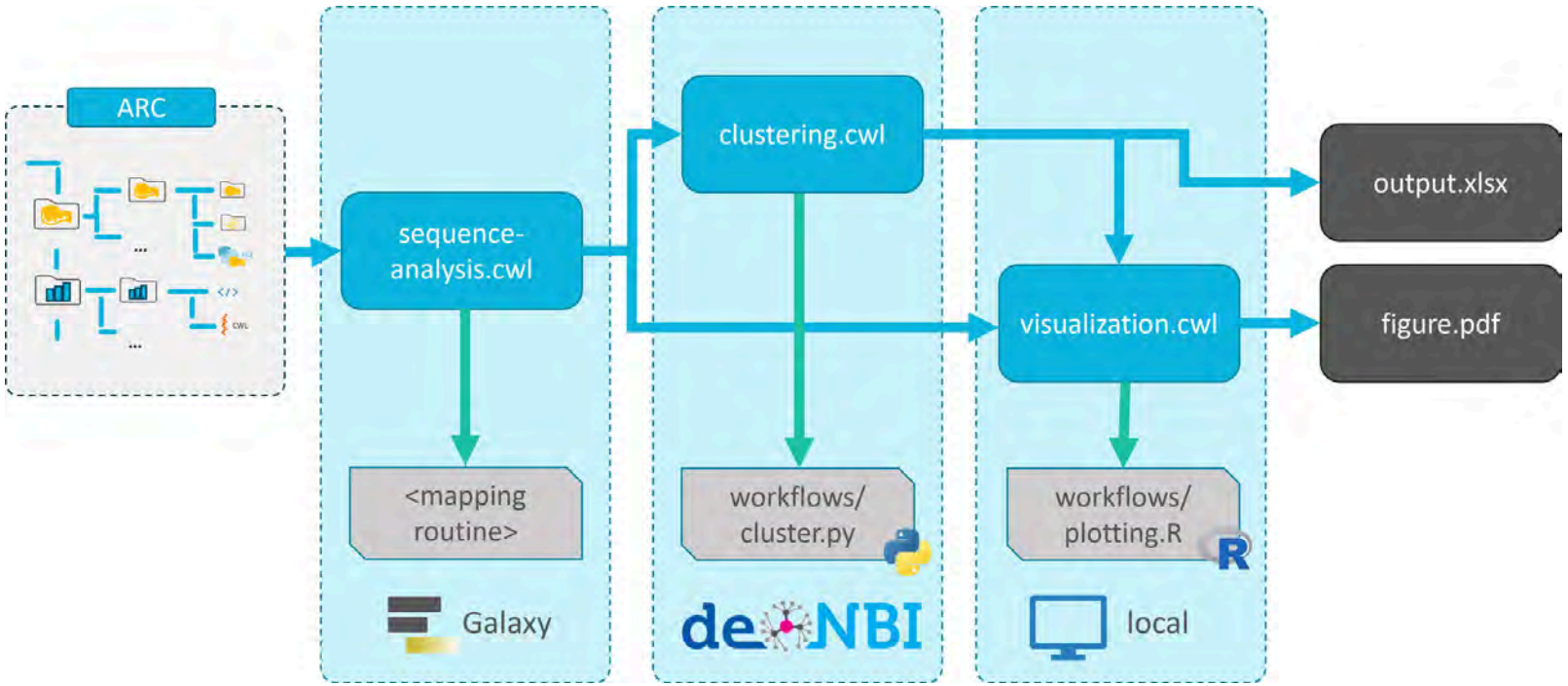
Collect


Studies
<Study Name>
Resources
Protocols
isa.study.xlsx
Assays
<Assay Name>
Dataset
Protocols
isa.assay.xlsx
...
Experimental Data

Process (e.g. annotate)

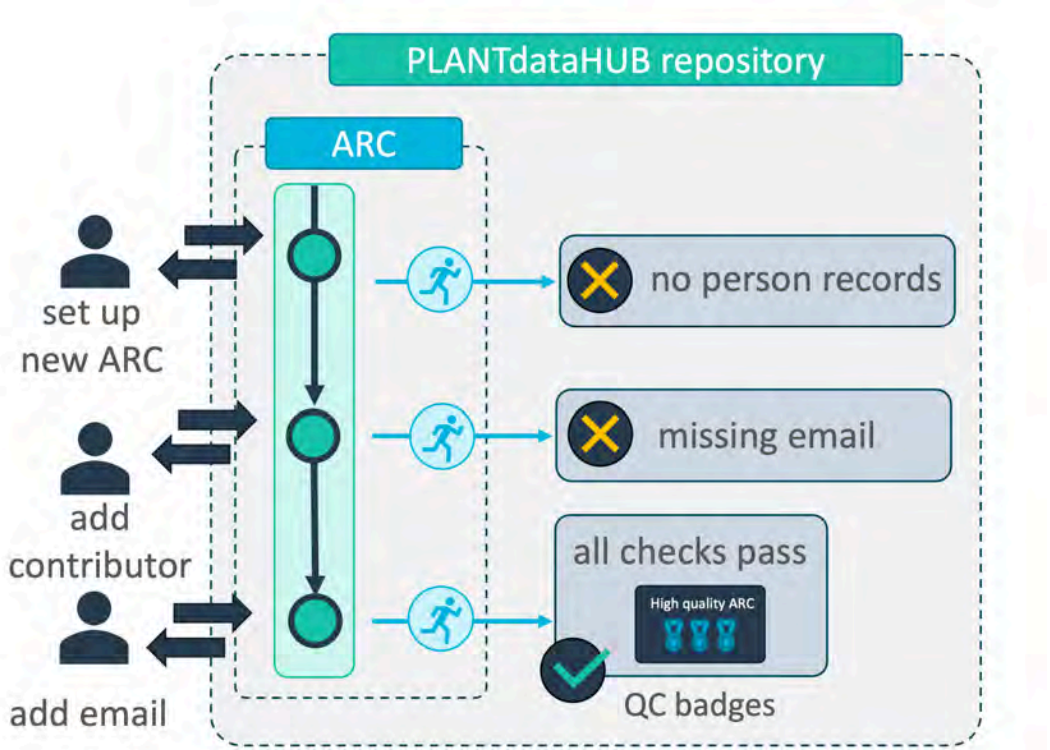


Analyse



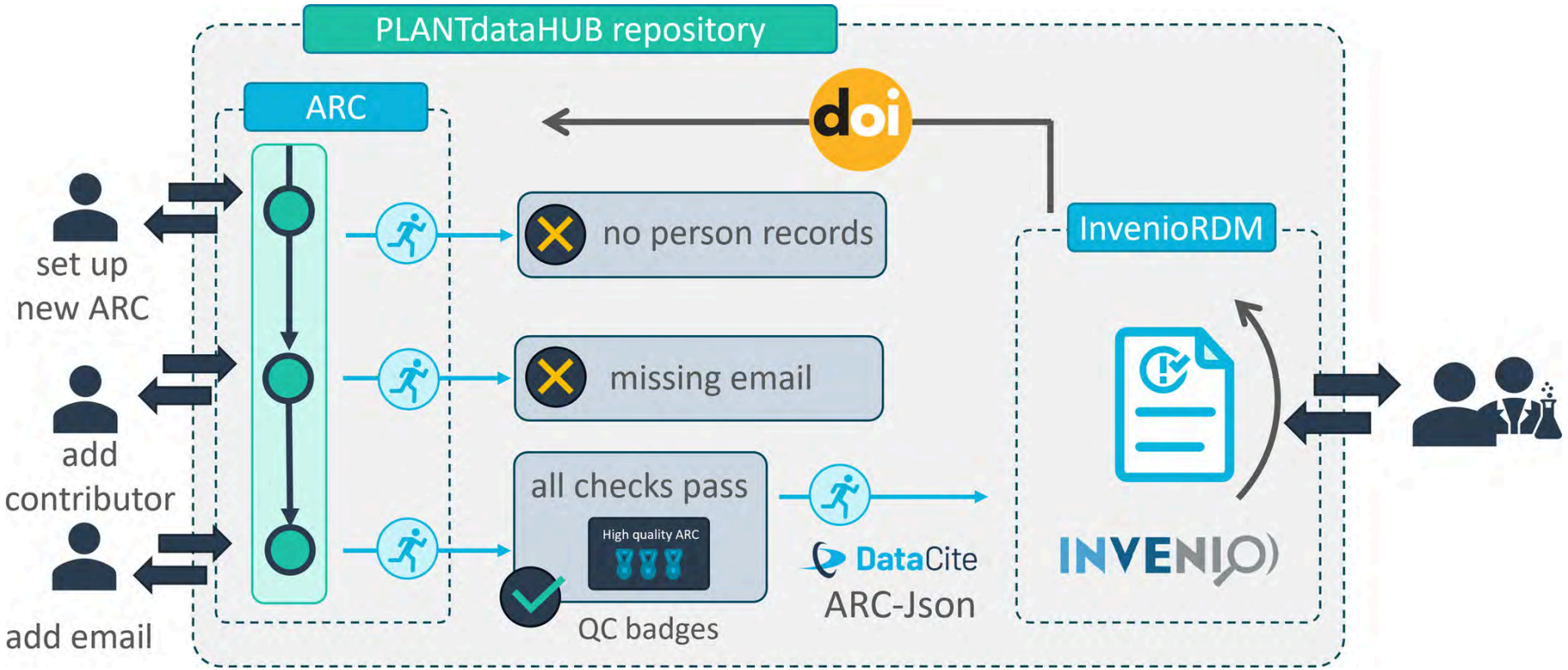
Weil, H.L., Schneider, K., et al. (2023), PLANTdataHUB: a collaborative platform for continuous FAIR data sharing in plant research. Plant J. https://doi.org/10.1111/tpj.16474

Preserve



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

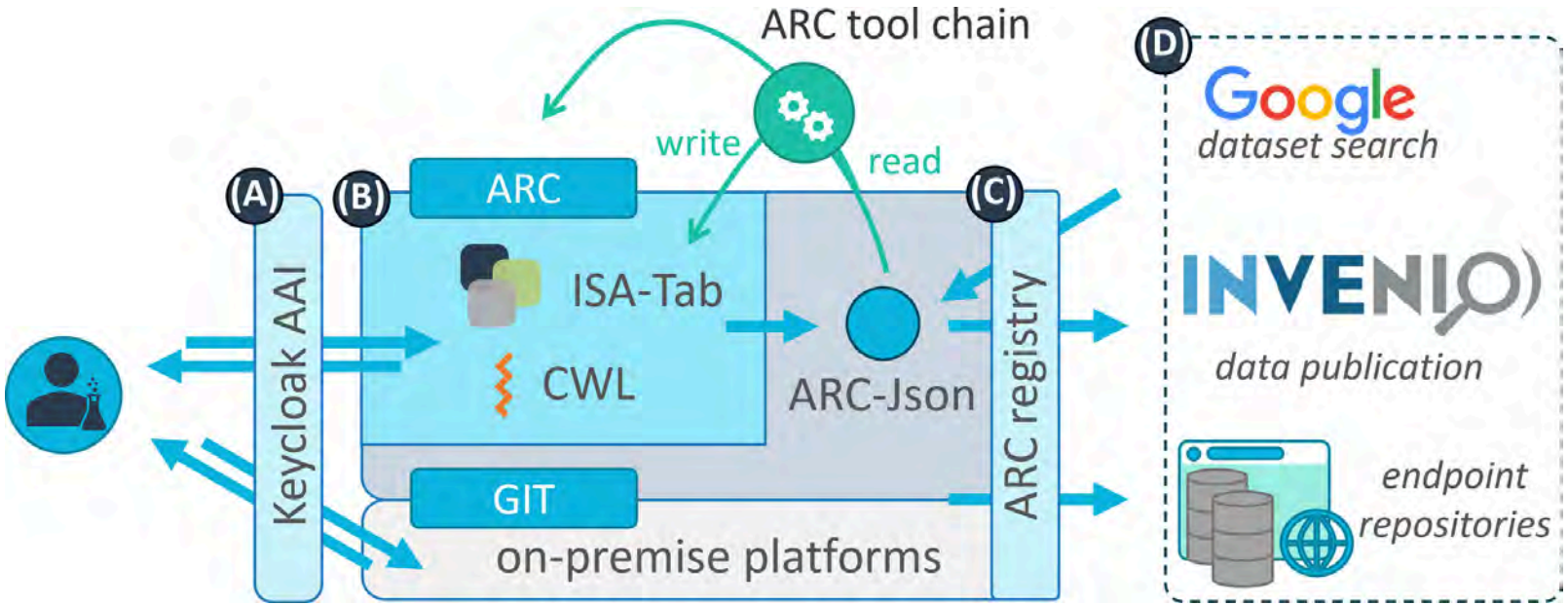
Preserve and publish



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

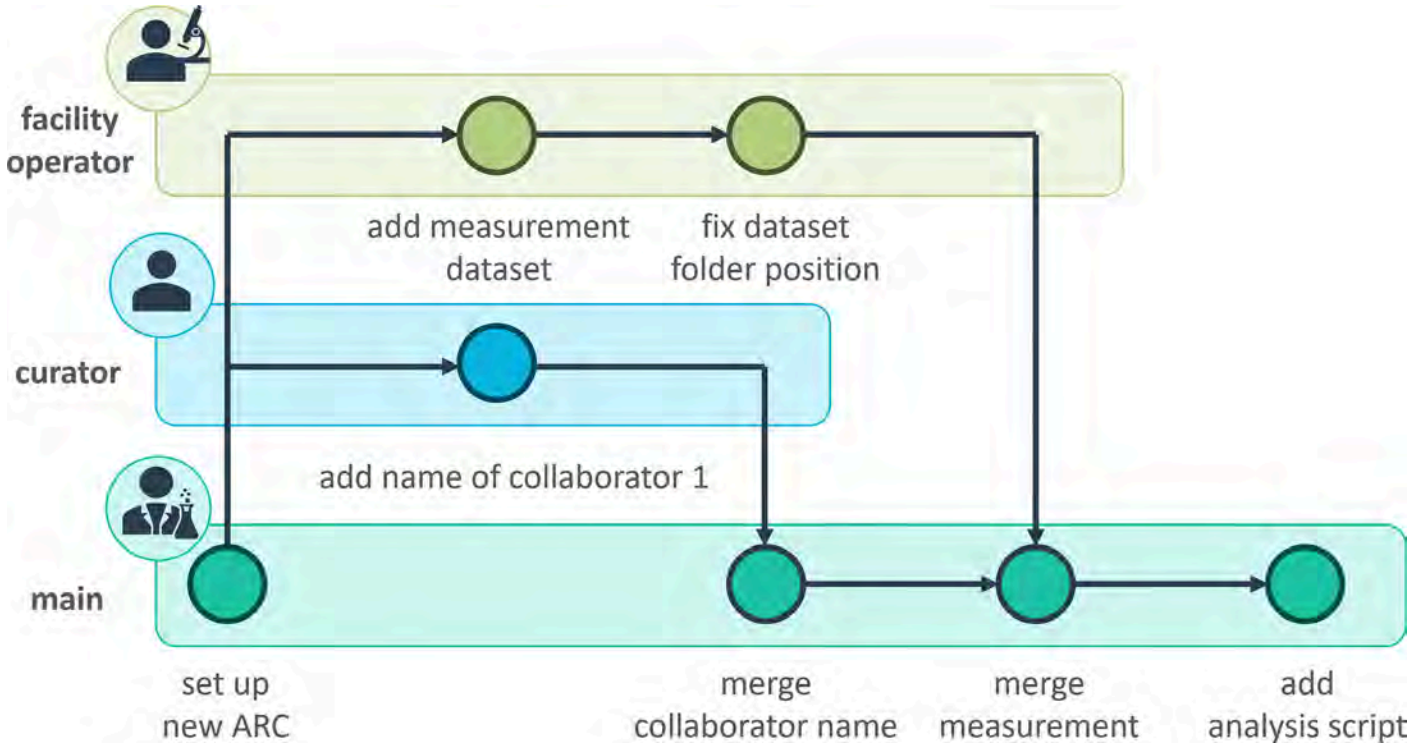
Share and collaborate

Reuse



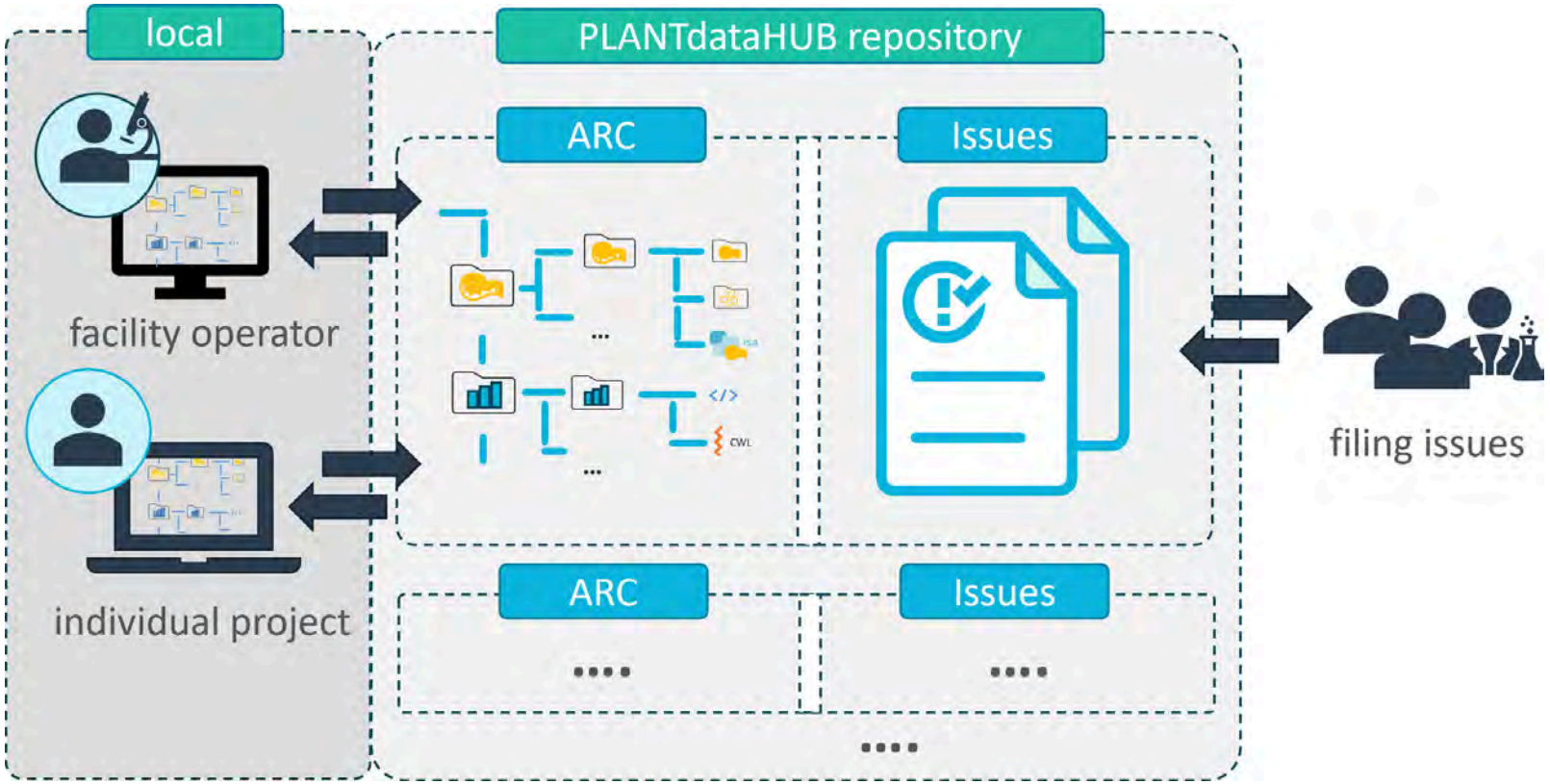
Weil, H.L., Schneider, K., et al. (2023), PLANTdataHUB: a collaborative platform for continuous FAIR data sharing in plant research. Plant J. https://doi.org/10.1111/tpj.16474

Mutable data life cycle



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

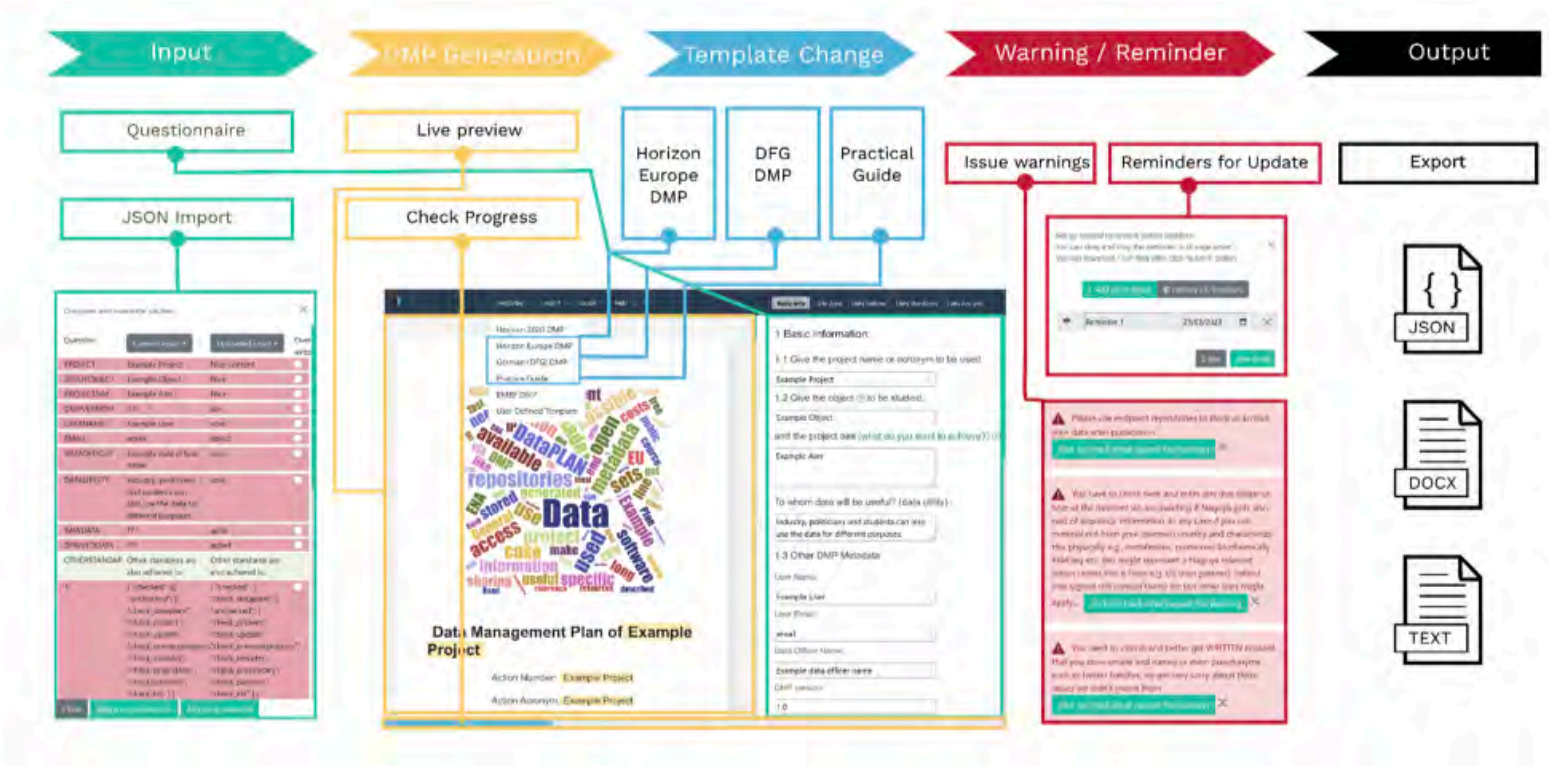
Plan (ARC scale)



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

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.)

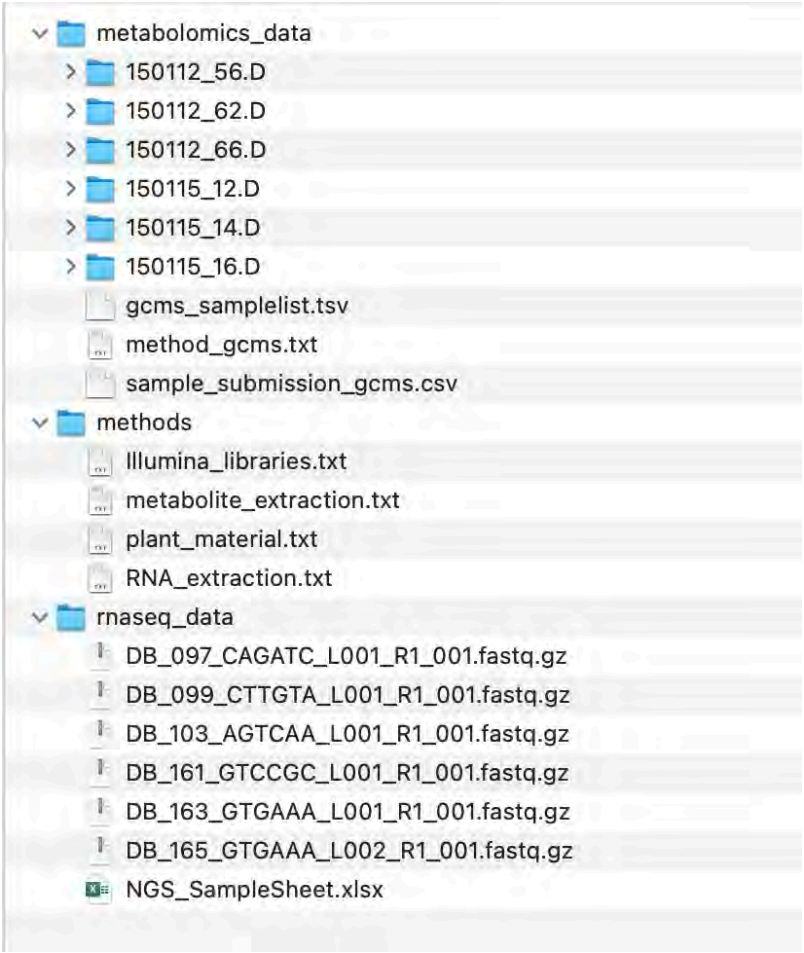
Hands-on with demo data

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

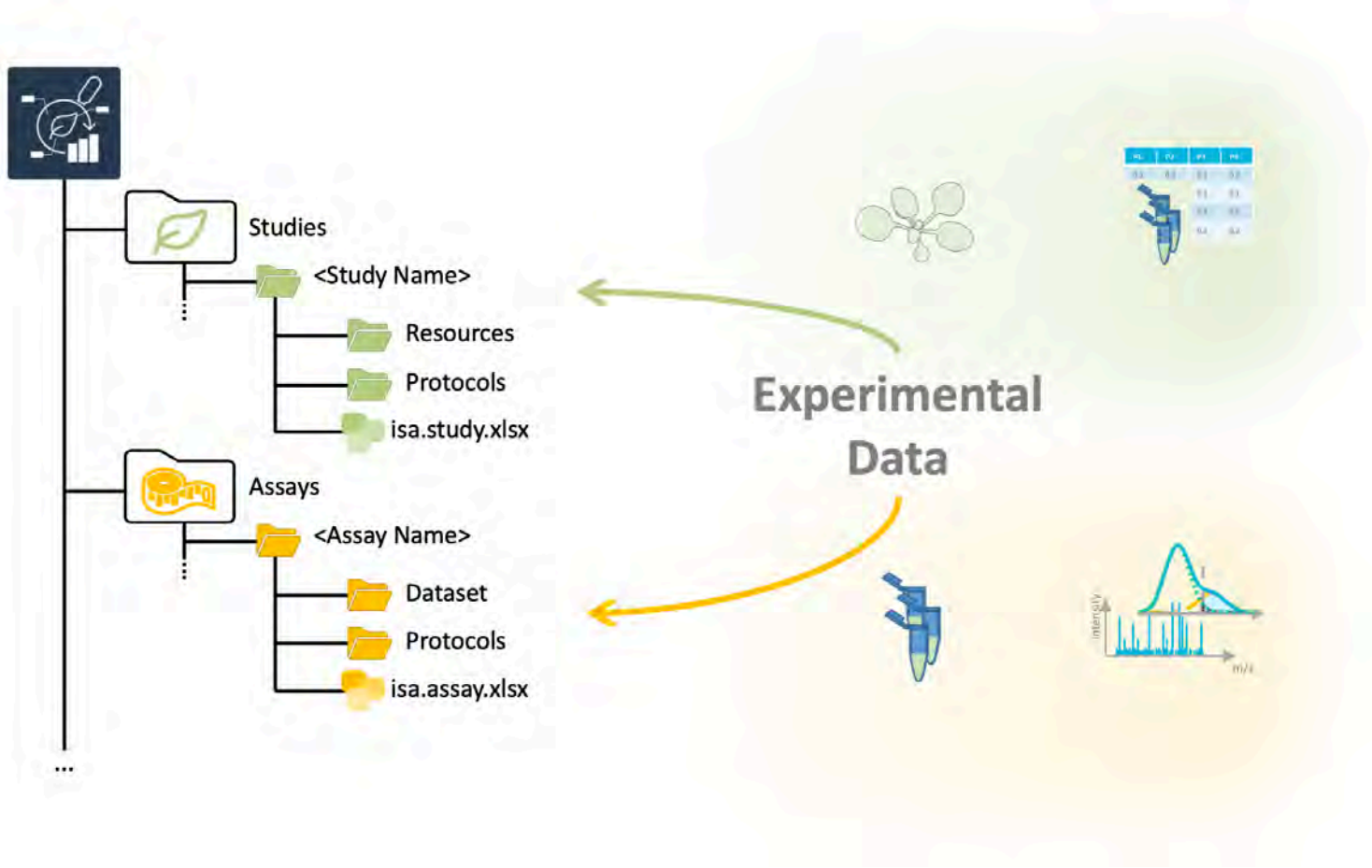


Goal

- 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

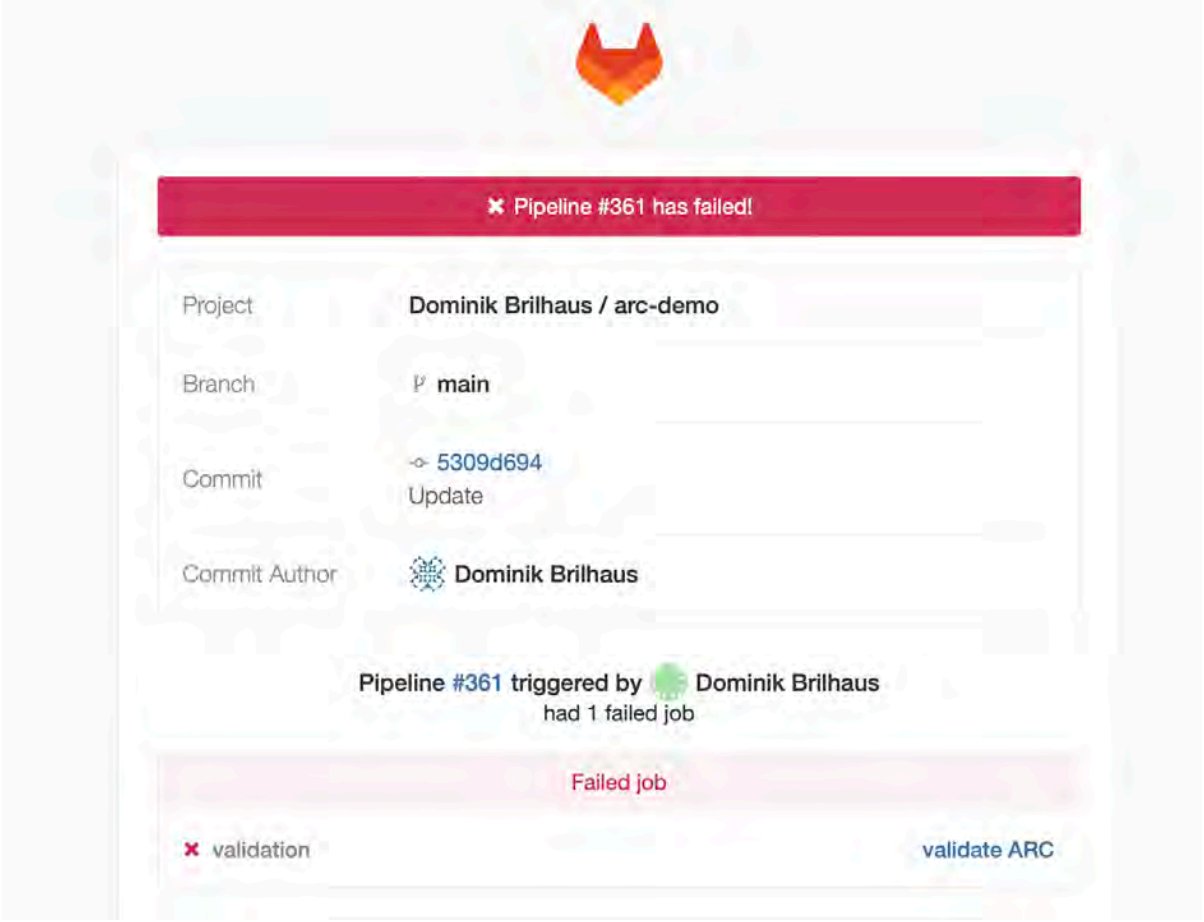
```
arc assay add -s talinum_drought -a rnaseq
arc assay add -s talinum_drought -a metabolomics
```

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?

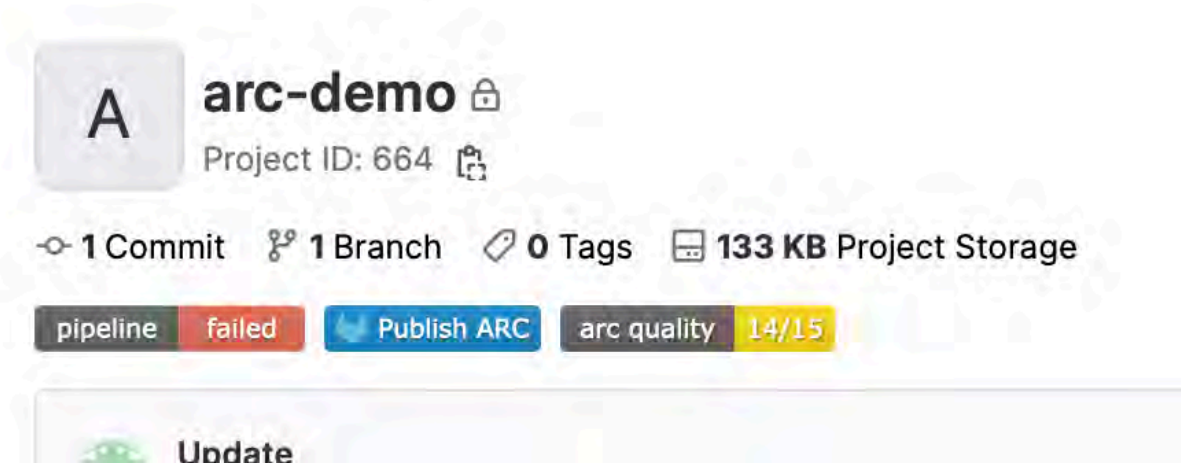


🔥 Don't worry 😄

Pipeline Failed

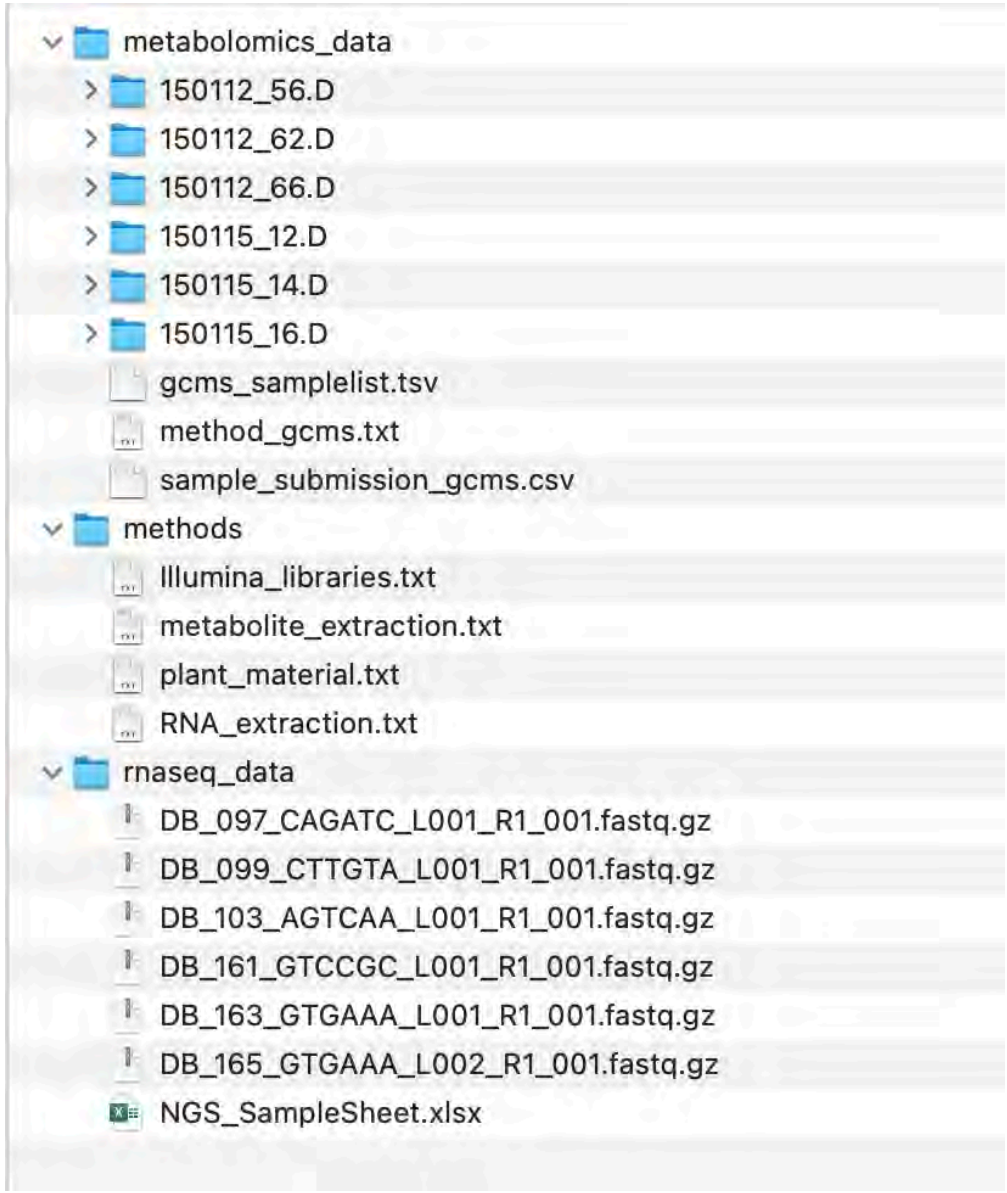
- a "continuous quality control" (CQC) pipeline validates your ARC
- This fails if one of the following metadata items is missing:

```
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

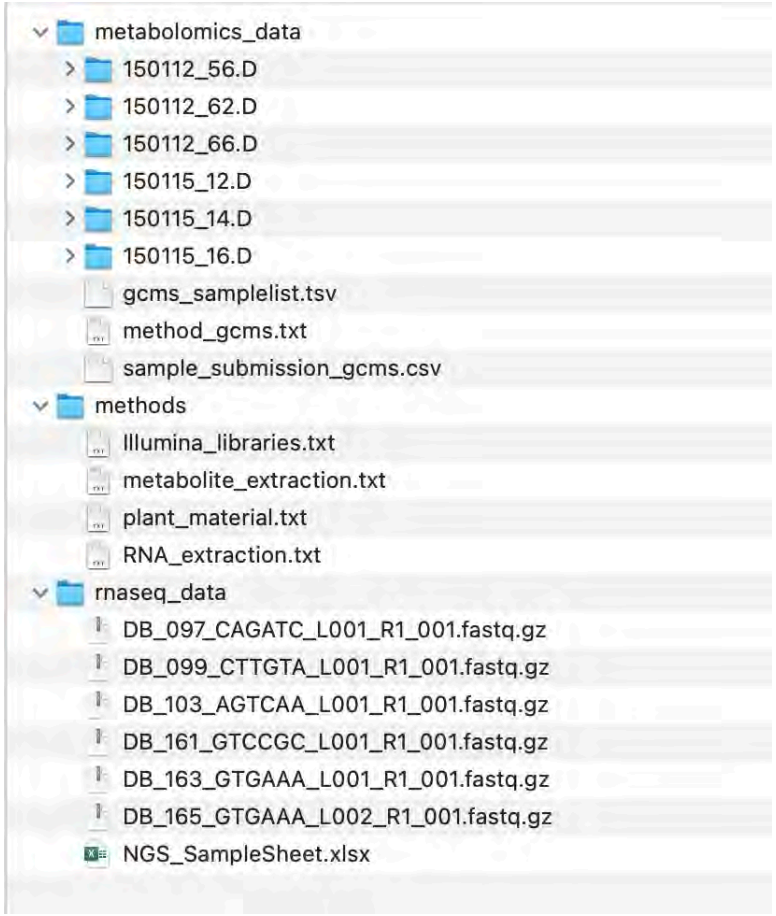
To save the changes, sync the ARC to the DataHUB including a message.

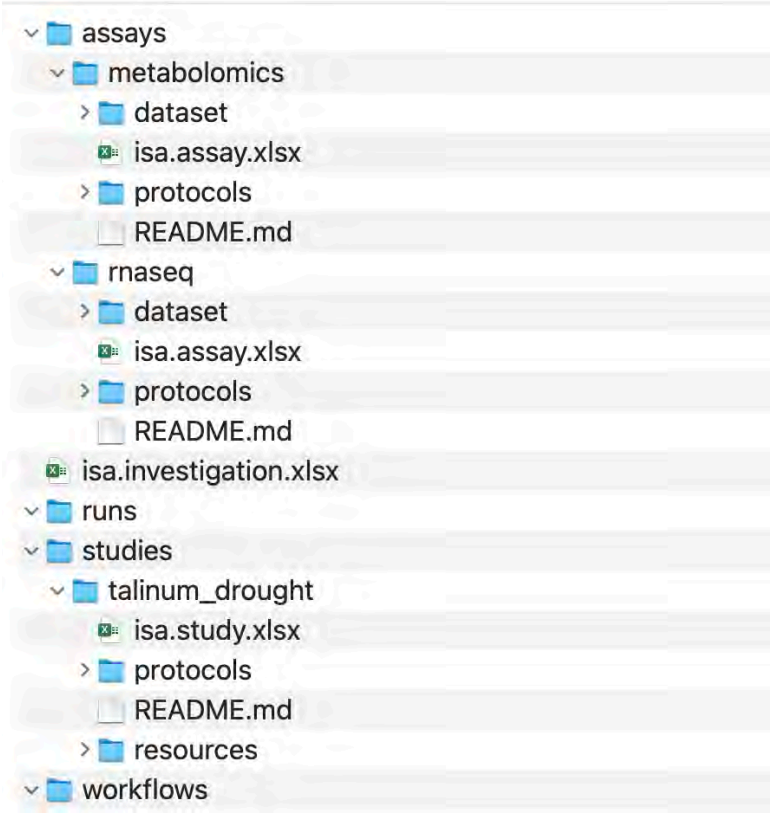
```
arc sync -m "sorted the demo data"
```

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

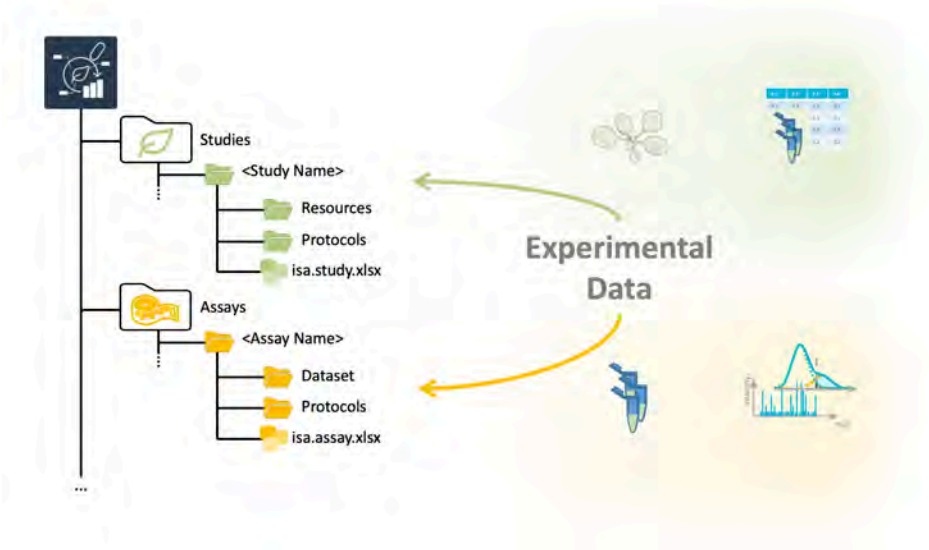
Please install the latest version of the ARCitect: https://github.com/nfdi4plants/ARCitect

🔥 (released September 20th, 2023) 🔥

Download the demo data

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

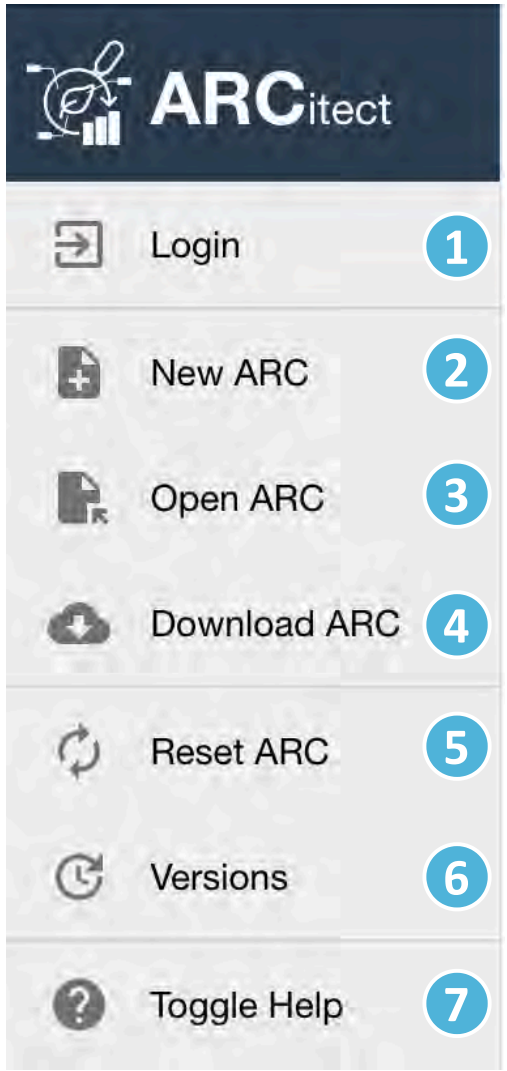
Sort Demo data in an ARC



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

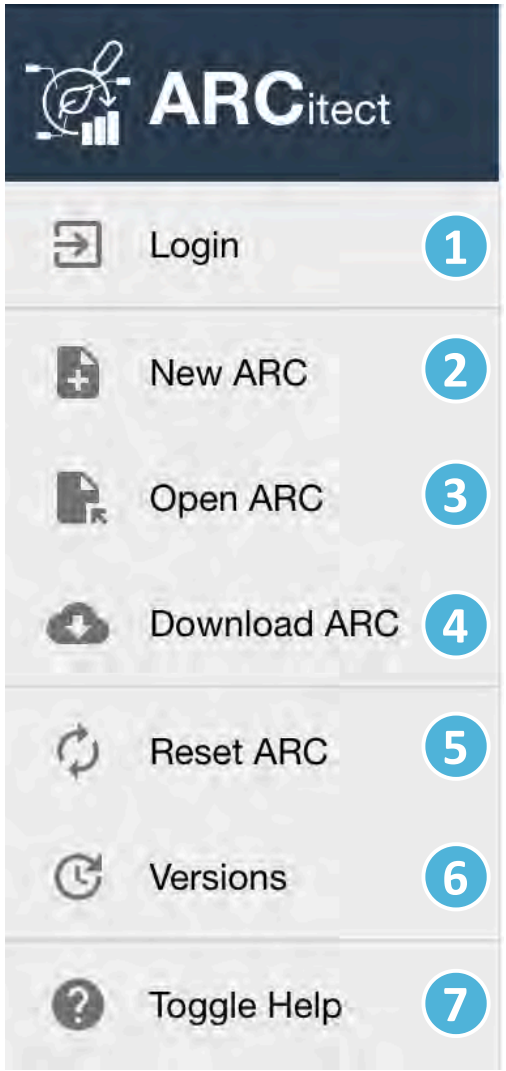
Open ARCitect

1. Login to DataHUB (1)



Initiate the ARC folder structure

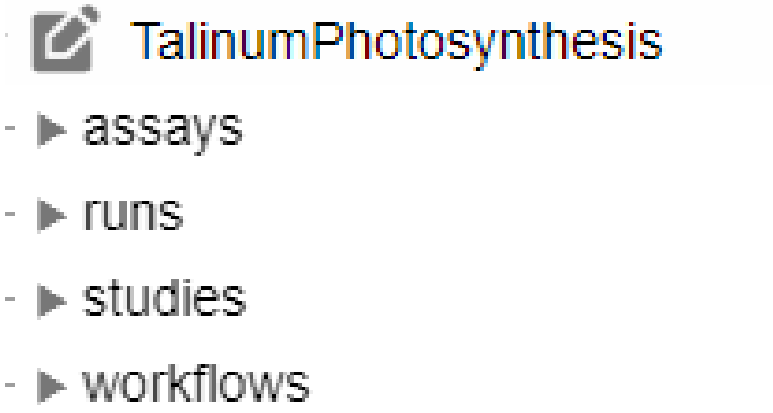
1. Create a **New ARC** (2)
2. Select a location and name it **TalinumPhotosynthesis**



Your ARC's name

💡 By default, your ARC's name will be used

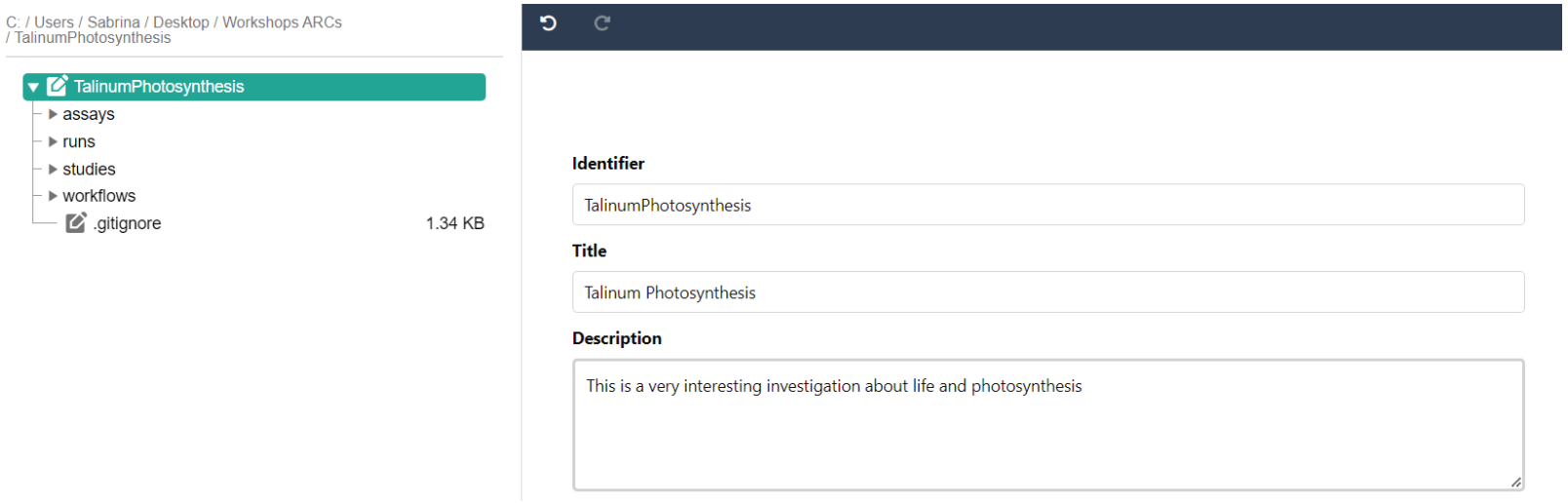
- for the ARC folder on your machine
- to create your ARC in the DataHUB at
  `https://git.nfdi4plants.org/<YourUserName>/<YourARC>`
  (see next steps)
- as the identifier for your investigation



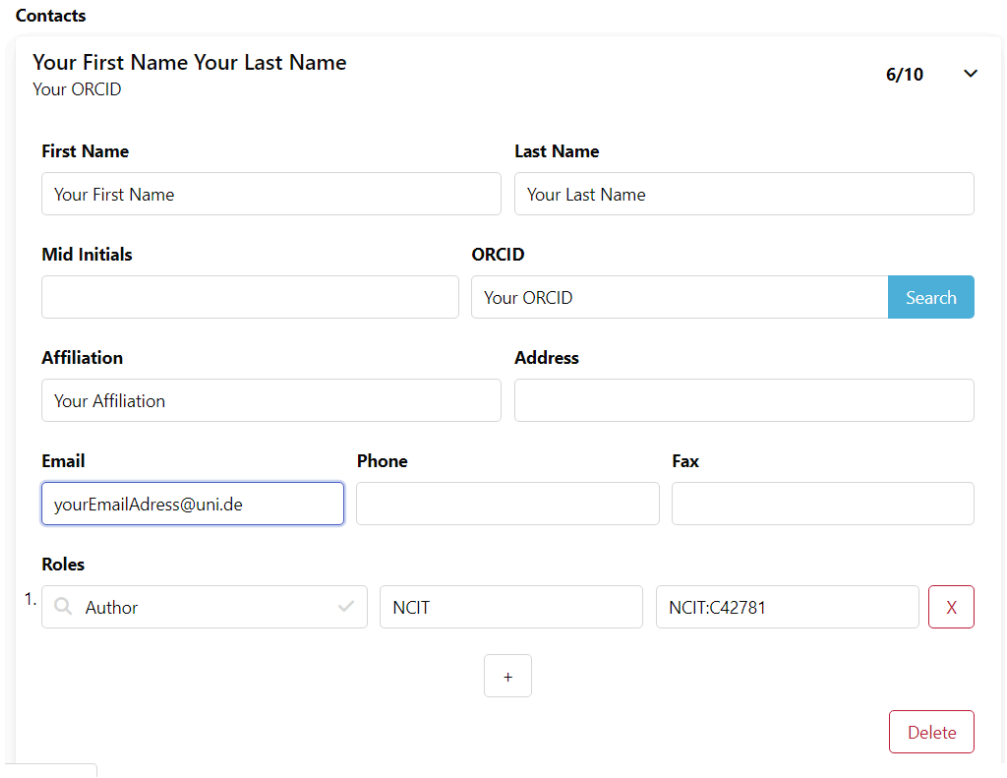
💡 Make sure that no ARC exists at
`https://git.nfdi4plants.org/<YourUserName>/<YourARC>`.
Otherwise you will sync to that ARC.

💡 Don't use spaces in ARC's name

Add a description to your investigation



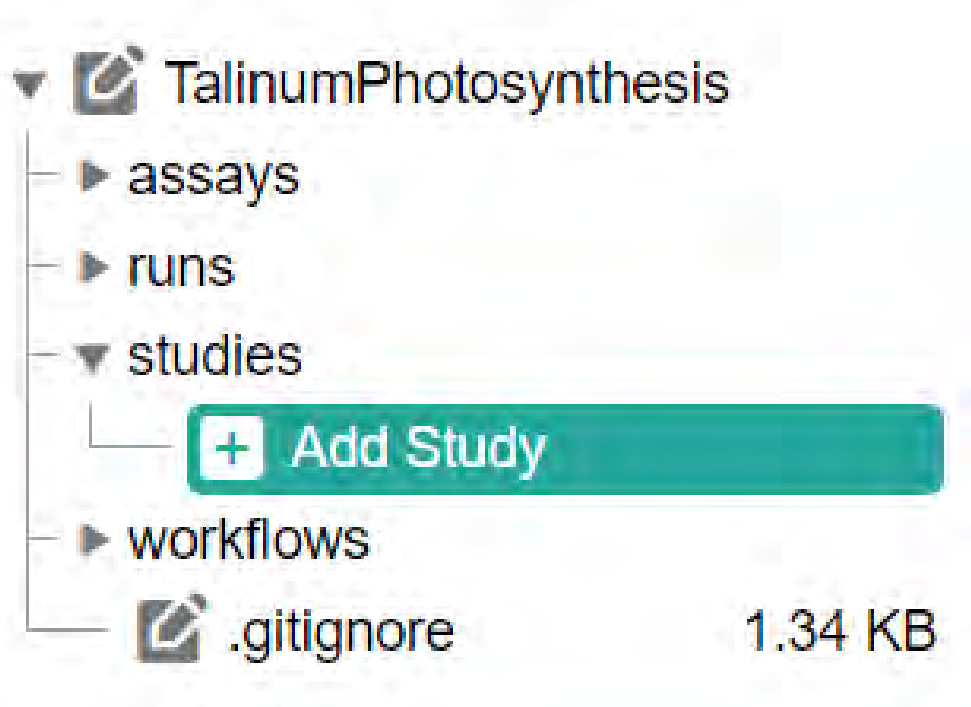
Add (at least one) contributor



Add a study

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

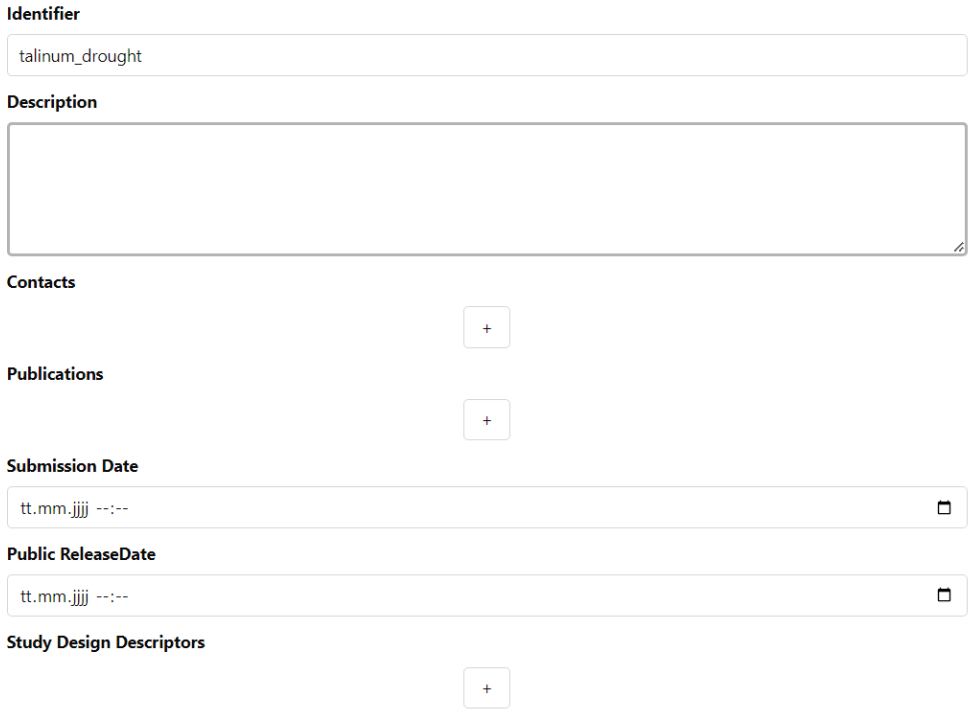
Use **talinum_drought** as an identifier



Study panel

In the study panel you can add

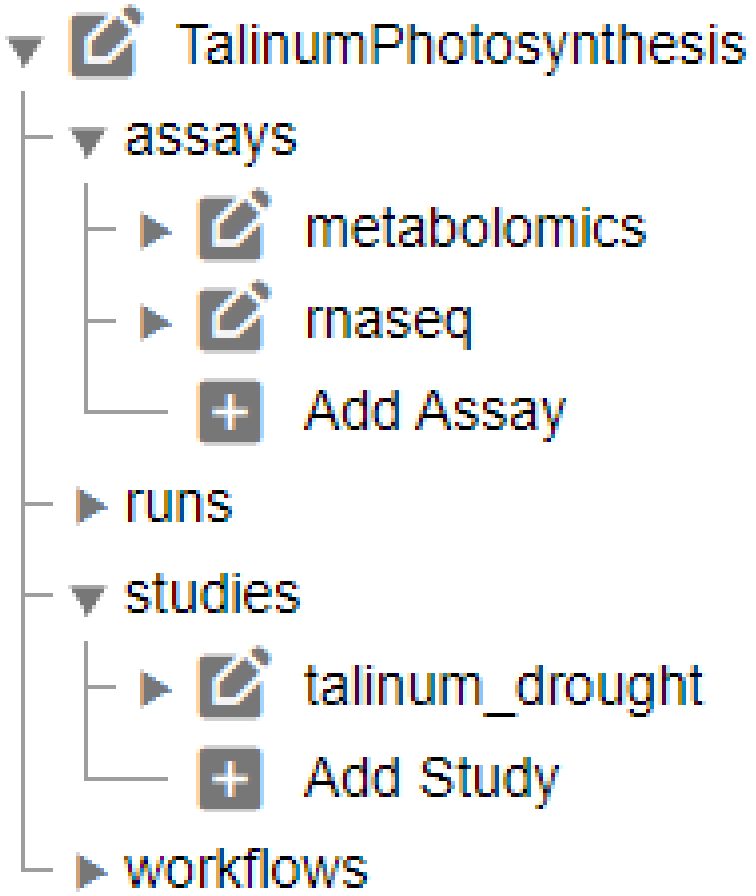
- general metadata,
- people, and
- publications
- data process information



Add an assay

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

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

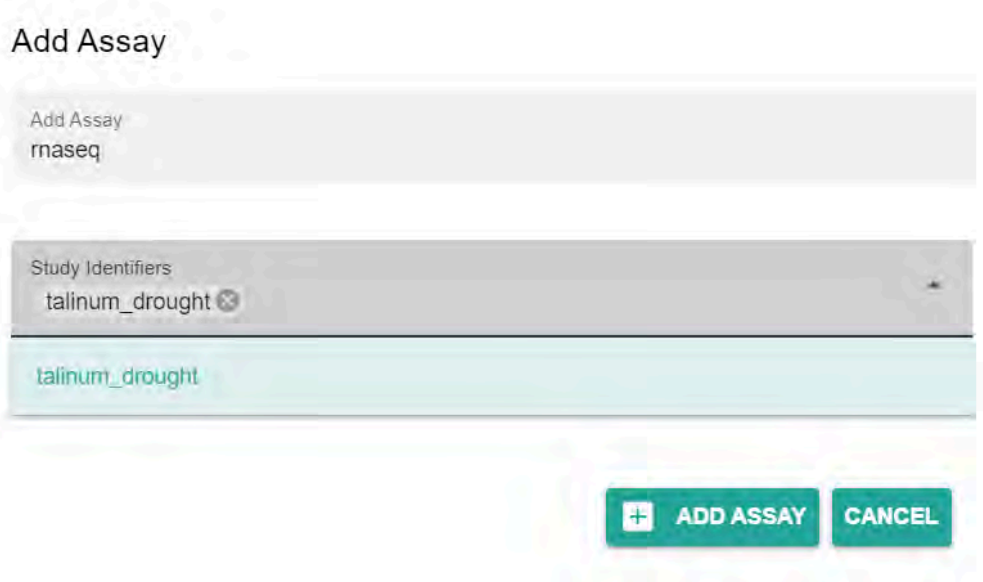


Link your assay to a study

You can either

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

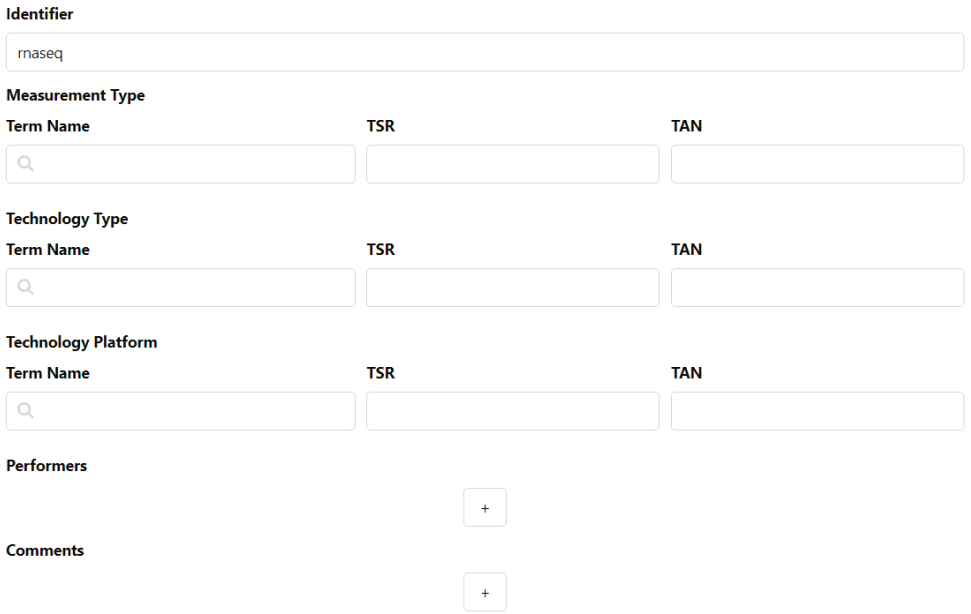
Link your assays to your **talinum_drought** study



Add information about your assay

In the assay panel you can

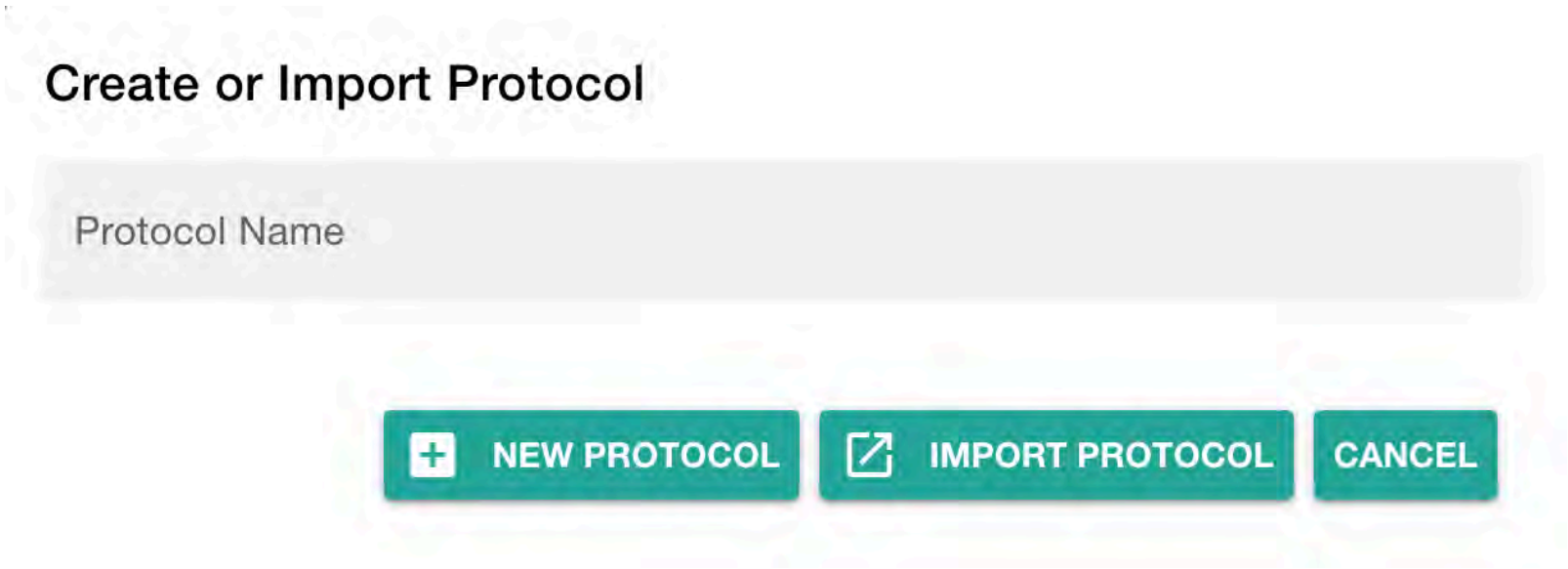
1. link or unlink the assay to studies, and
2. define the assay's
    - measurement type
    - technology type, and
    - technology platform.
3. add data process information



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.

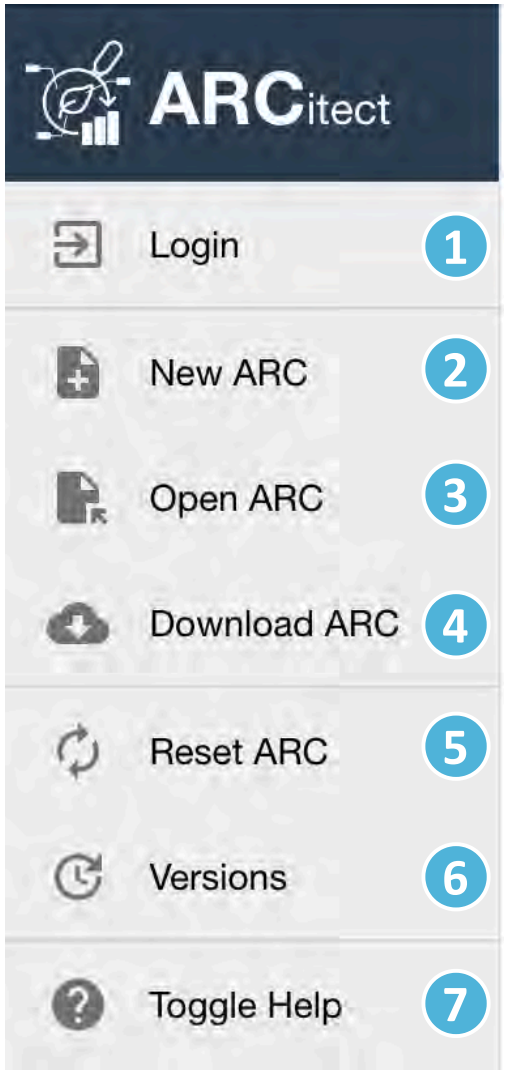
Sort Demo Data to your ARC

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

Login to the DataHUB

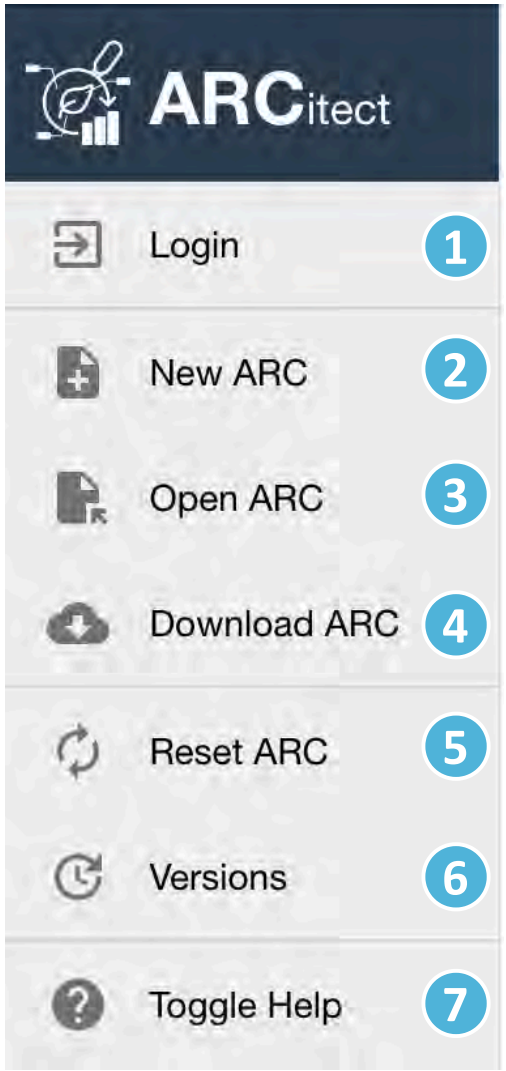
Click **Login** (1) in the sidebar to login to the DataHUB.

💡 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

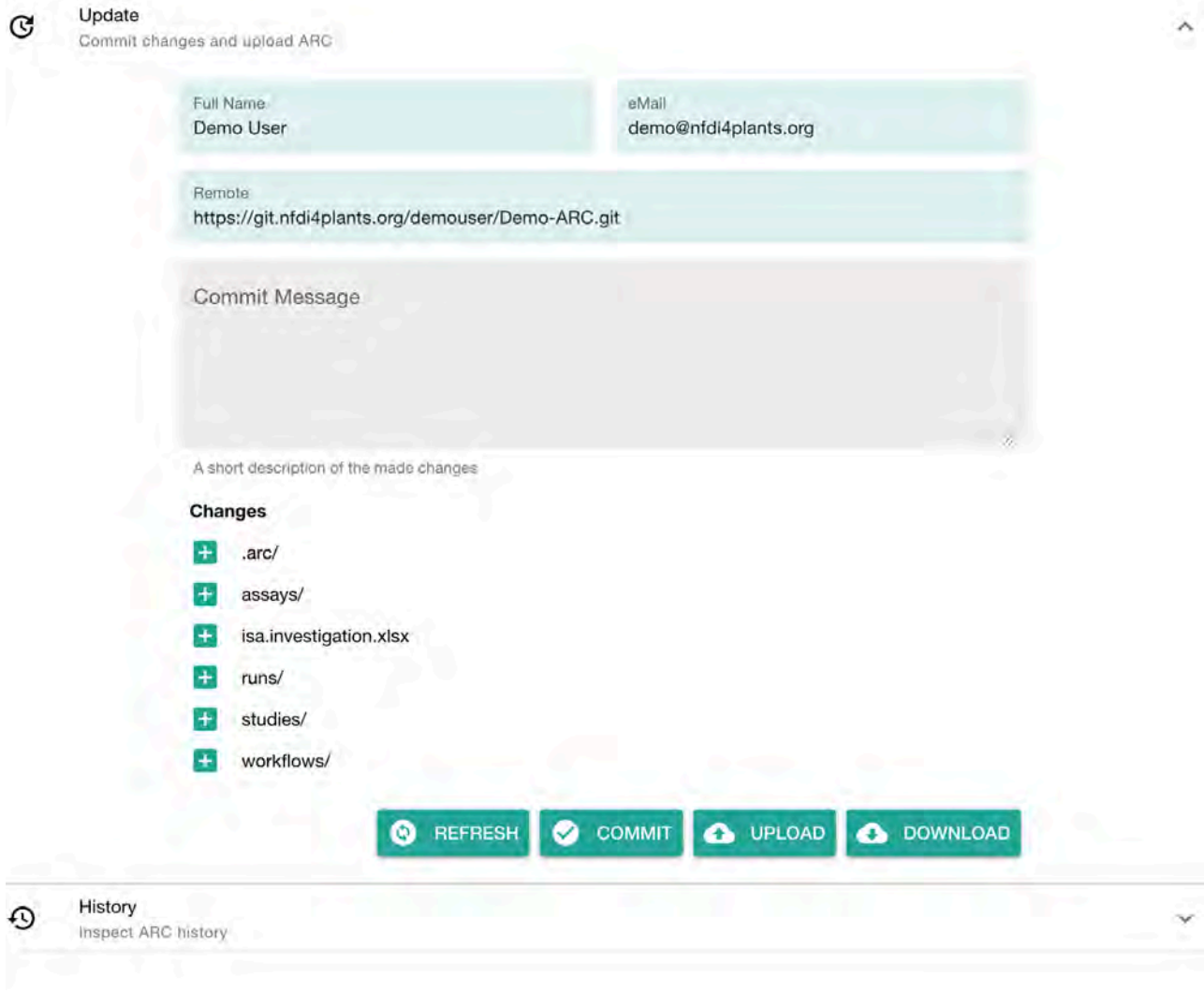
From the sidebar, navigate to **Versions** (6)



Versions

The versions panel allows you to

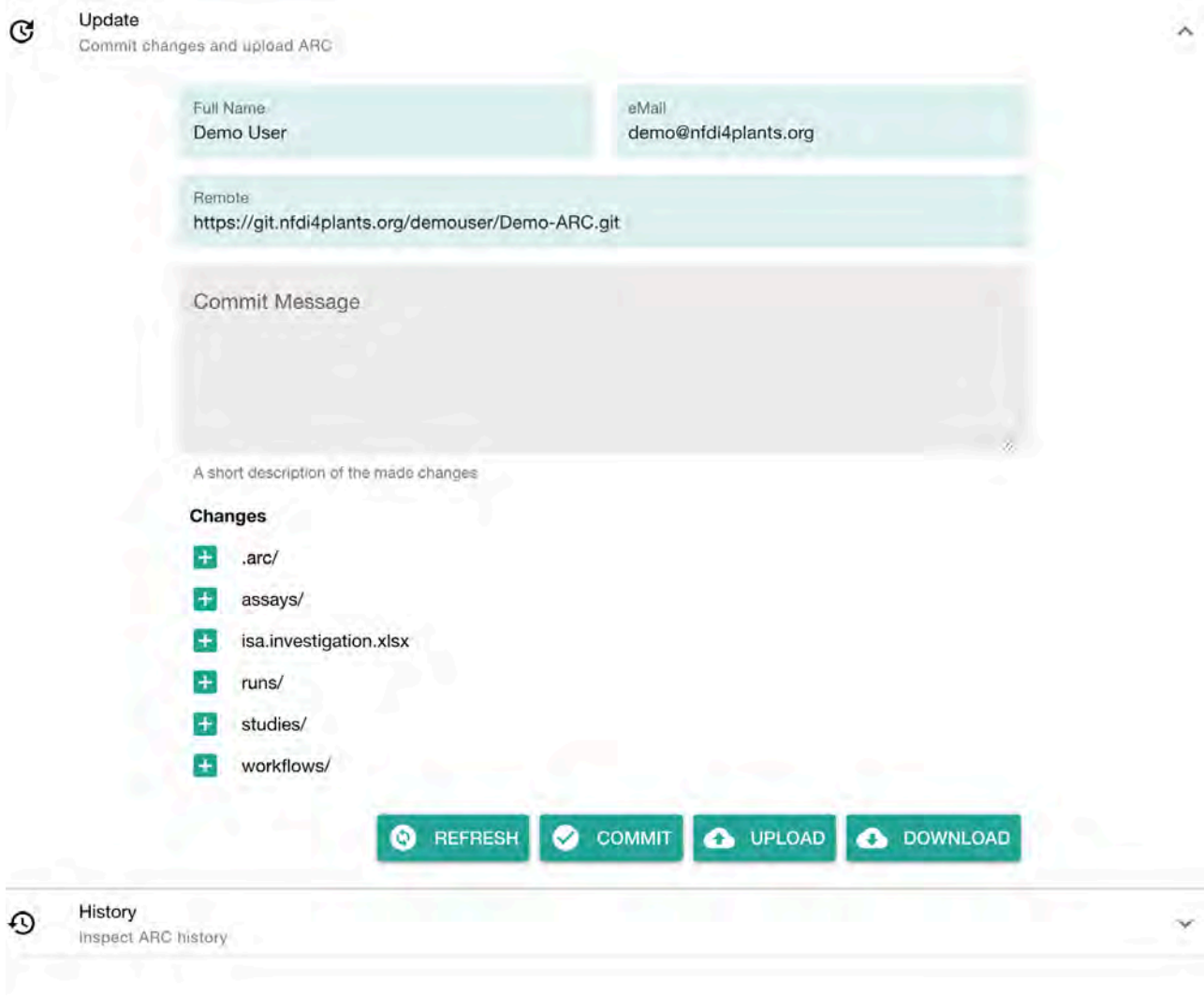
- store the local changes to your ARC in form of "commits",
- sync the changes to the DataHUB, and
- check the history of your ARC



Connection to the DataHUB

If you are logged in, the versions panel shows

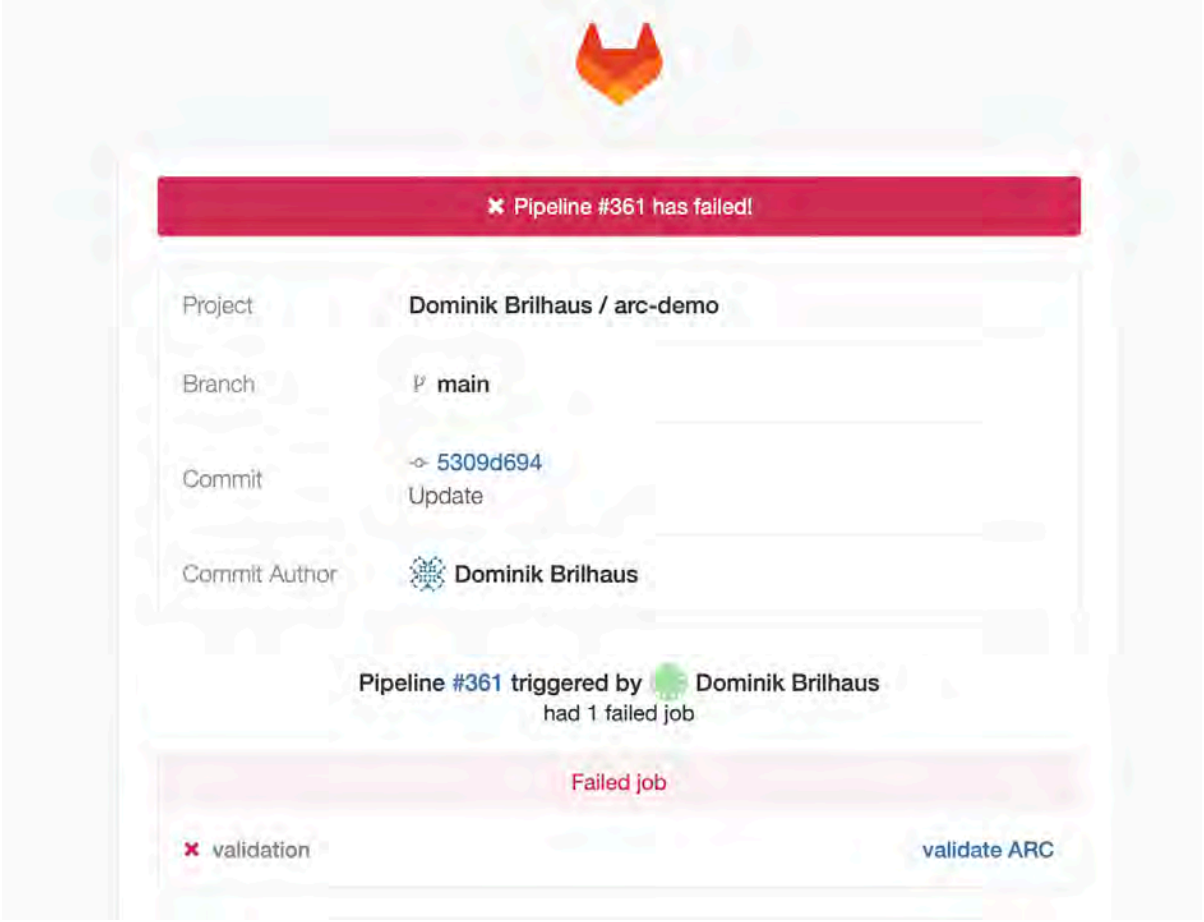
- your DataHUB's *Full Name* and *eMail*
- the URL of the current ARC in the DataHUB `https://git.nfdi4plants.org/<YourUserName>/<YourARC>`



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?

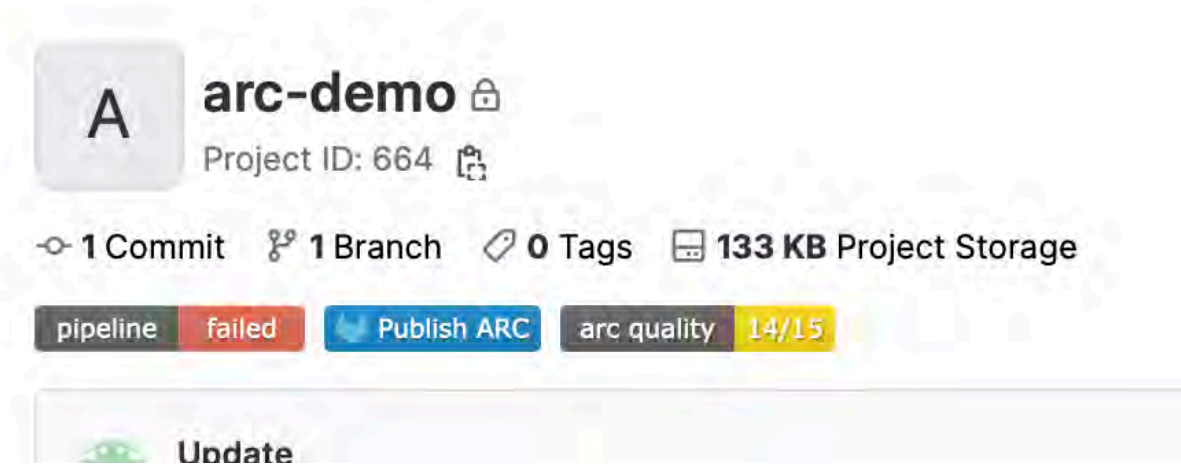


🔥 Don't worry 😄

Pipeline Failed

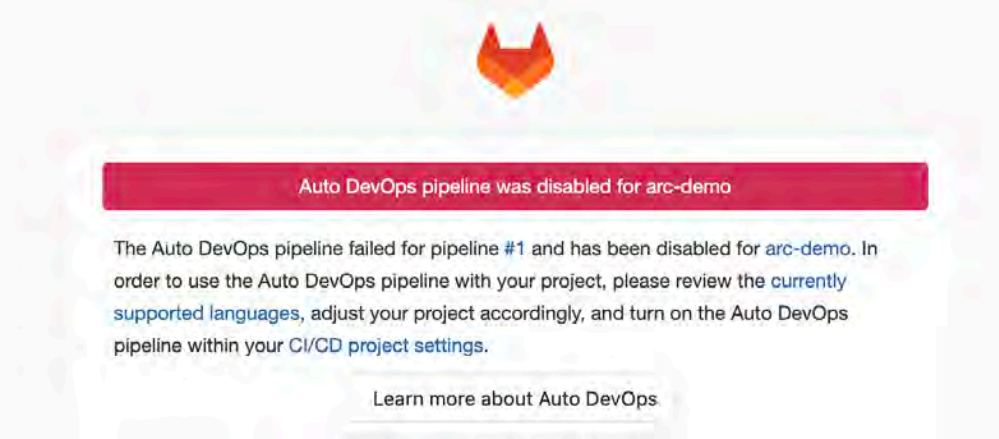
- a "continuous quality control" (CQC) pipeline validates your ARC
- This fails if one of the following metadata items is missing:

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



Pipeline Failed

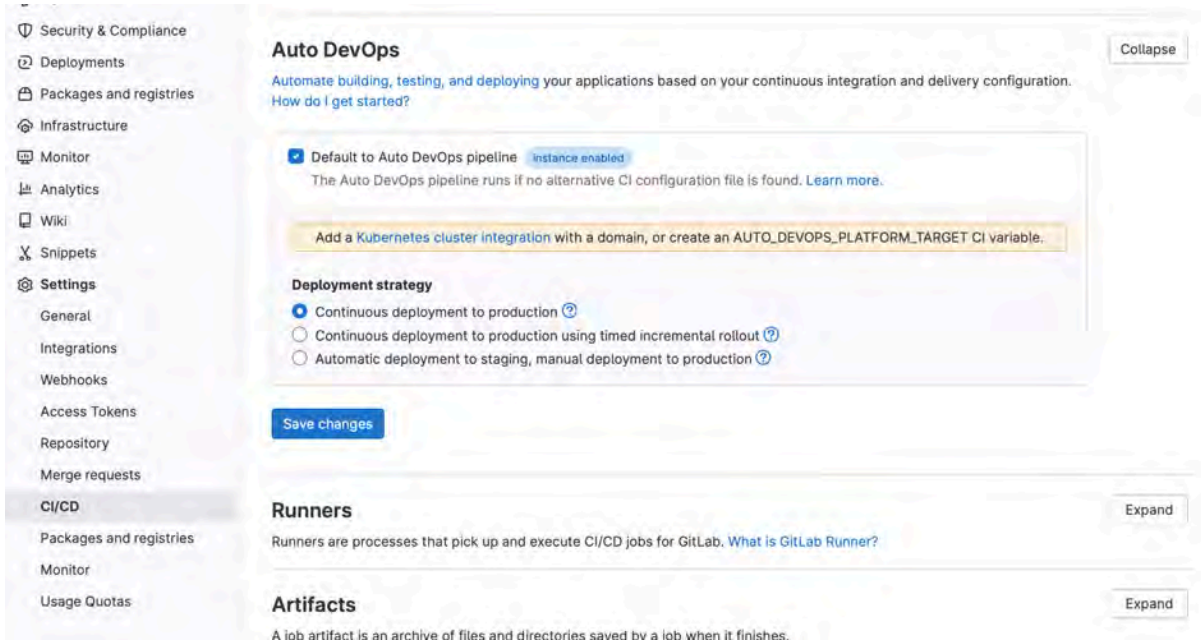
If the pipeline has failed once, it is disabled by default



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

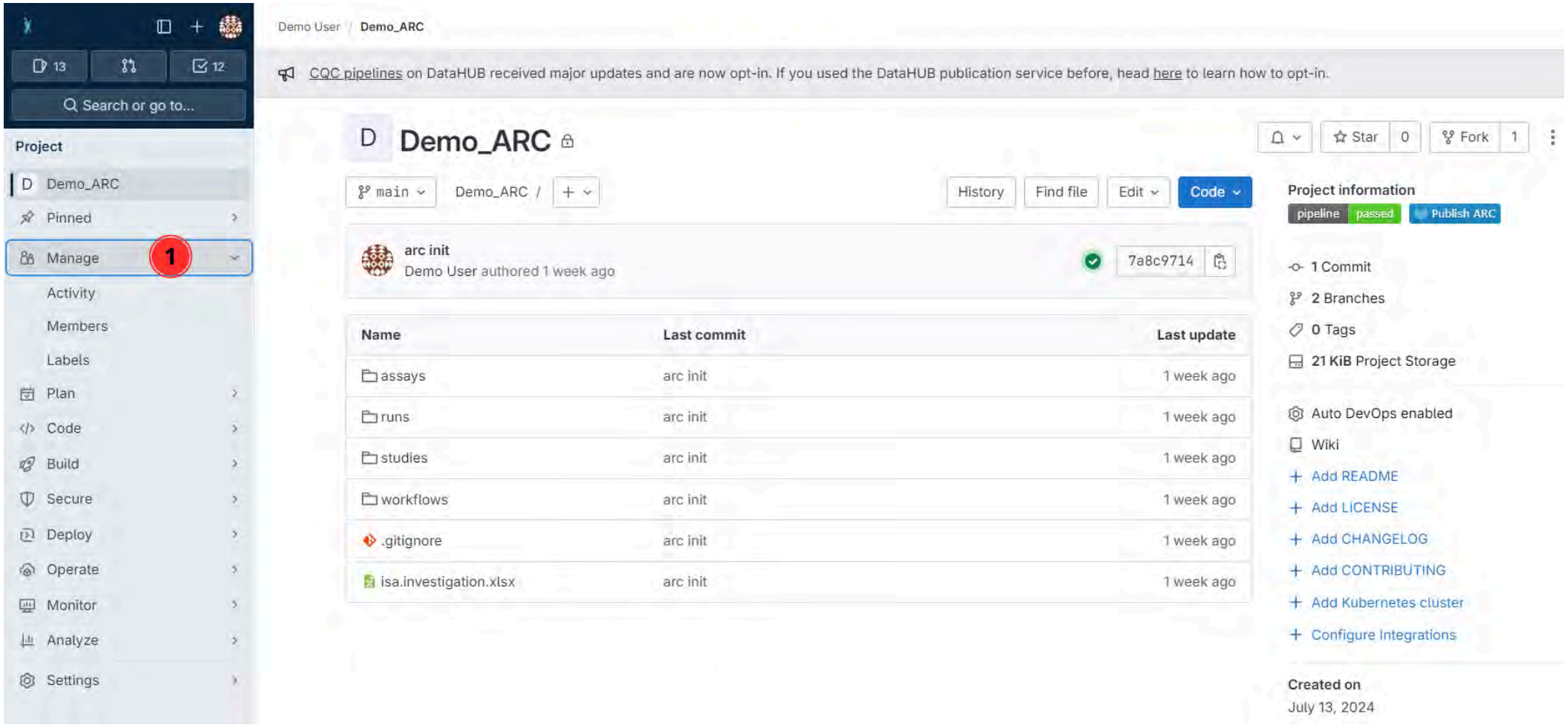


Collaborate and share

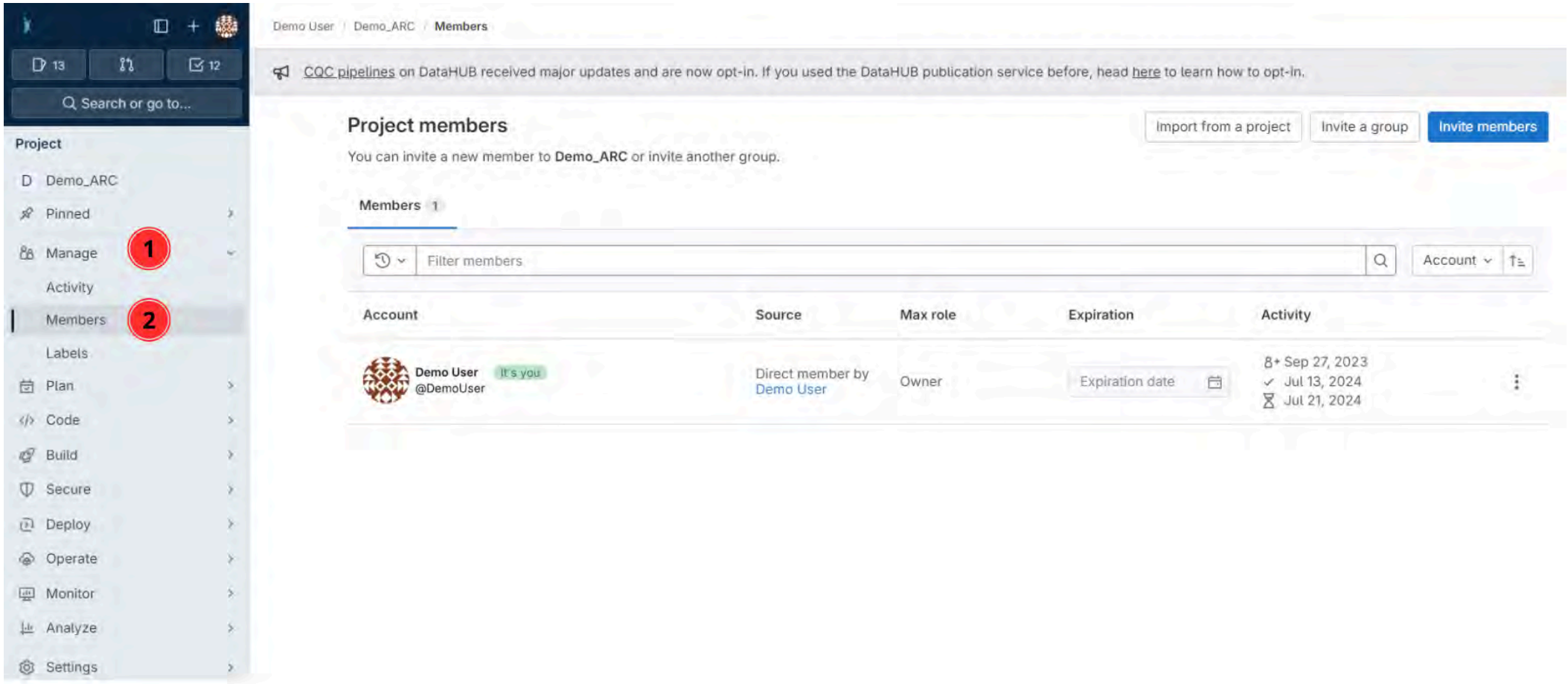
Invite collaborators

- 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.

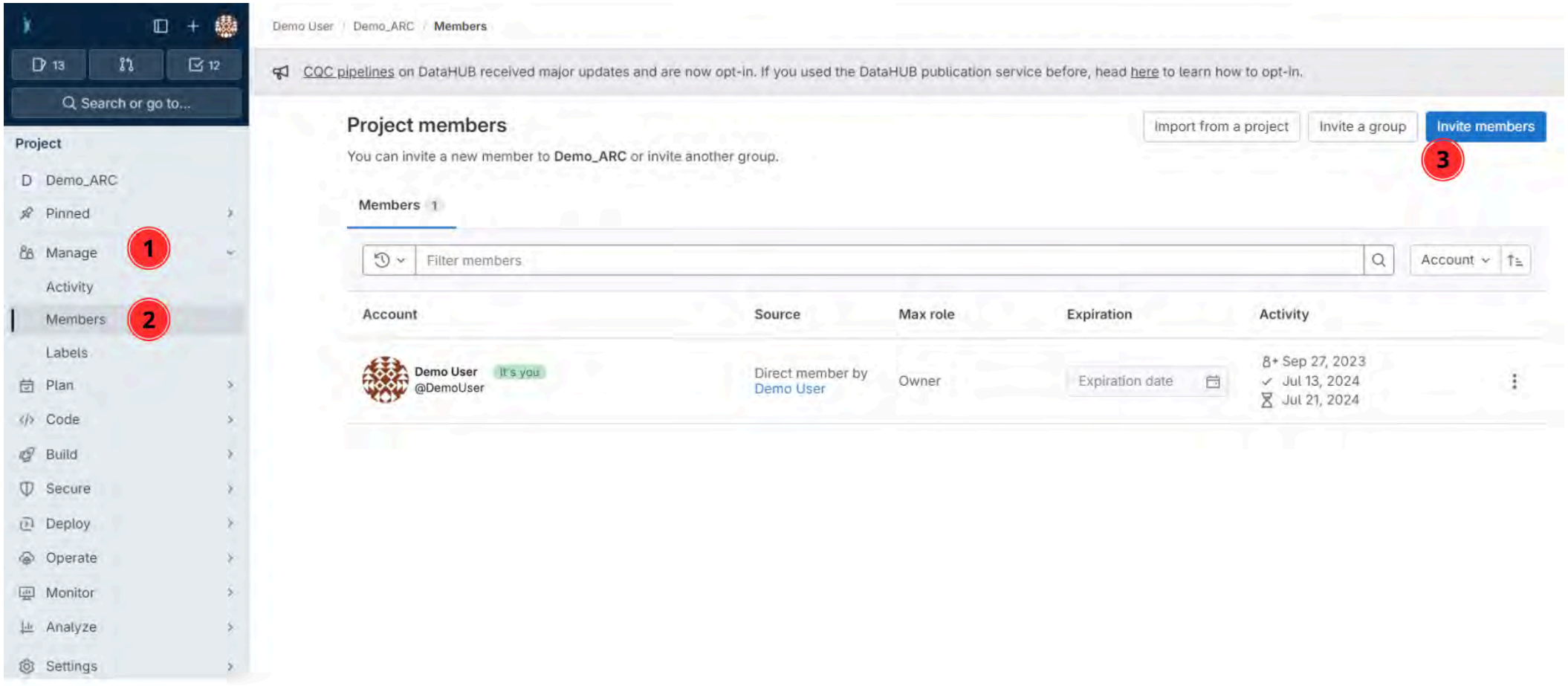
1. Click on **Project Information** in the left navigation panel



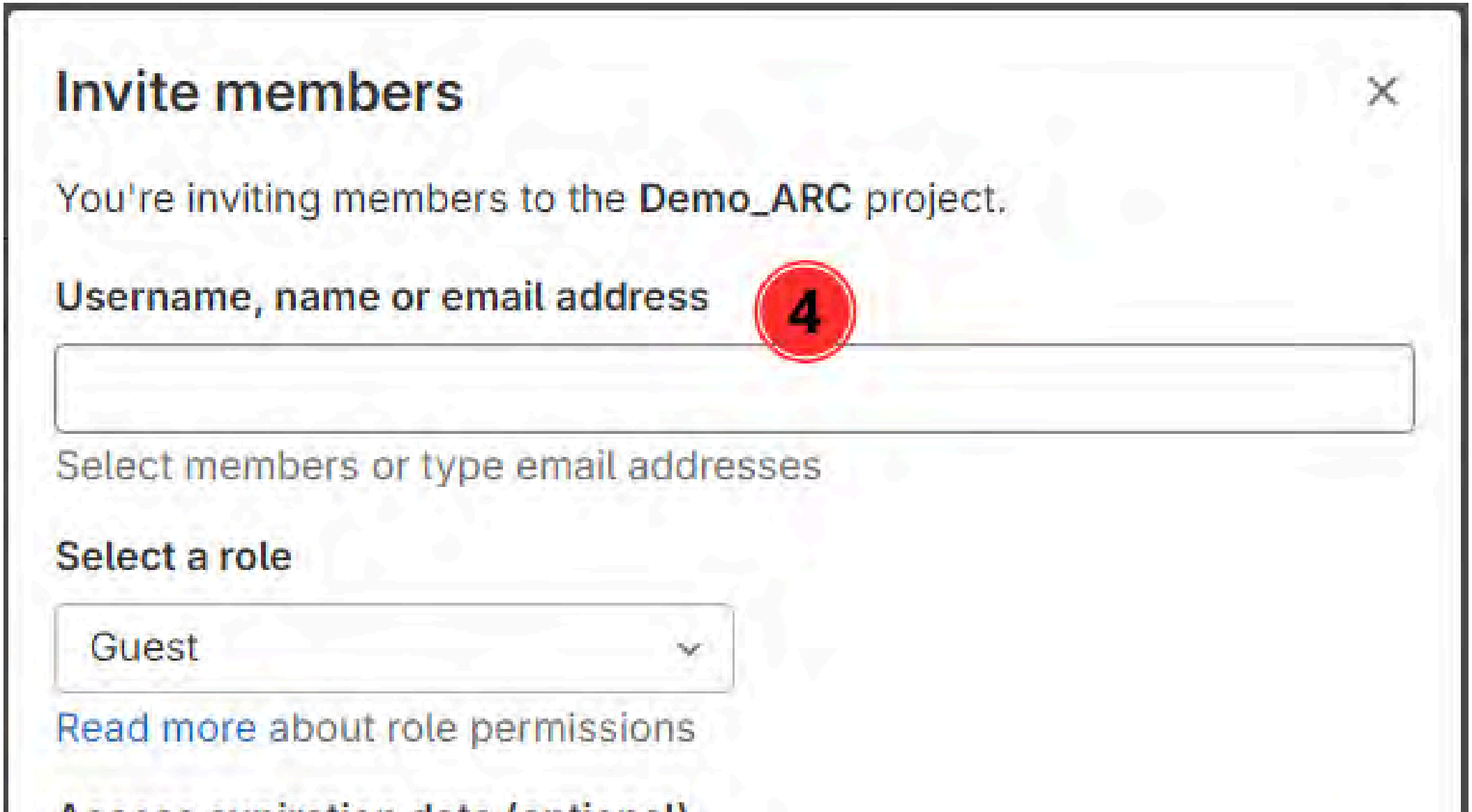
2. Click on **Members**



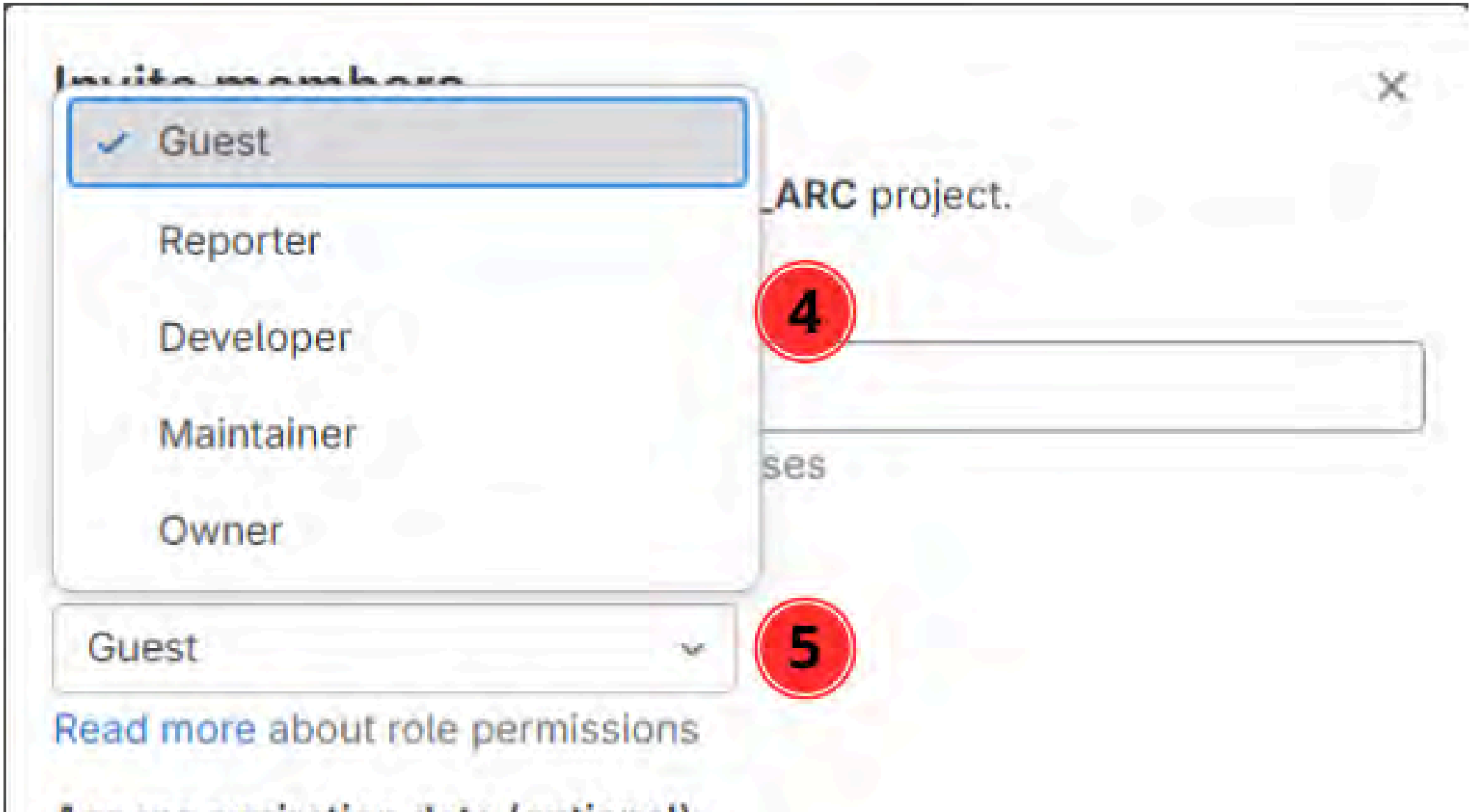
3. Click on **Invite members**



4. Search for potential collaborators



5. Select a role



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, rovenance,
- 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

Accessible

- information about origin
- location of storage
- access rights

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

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

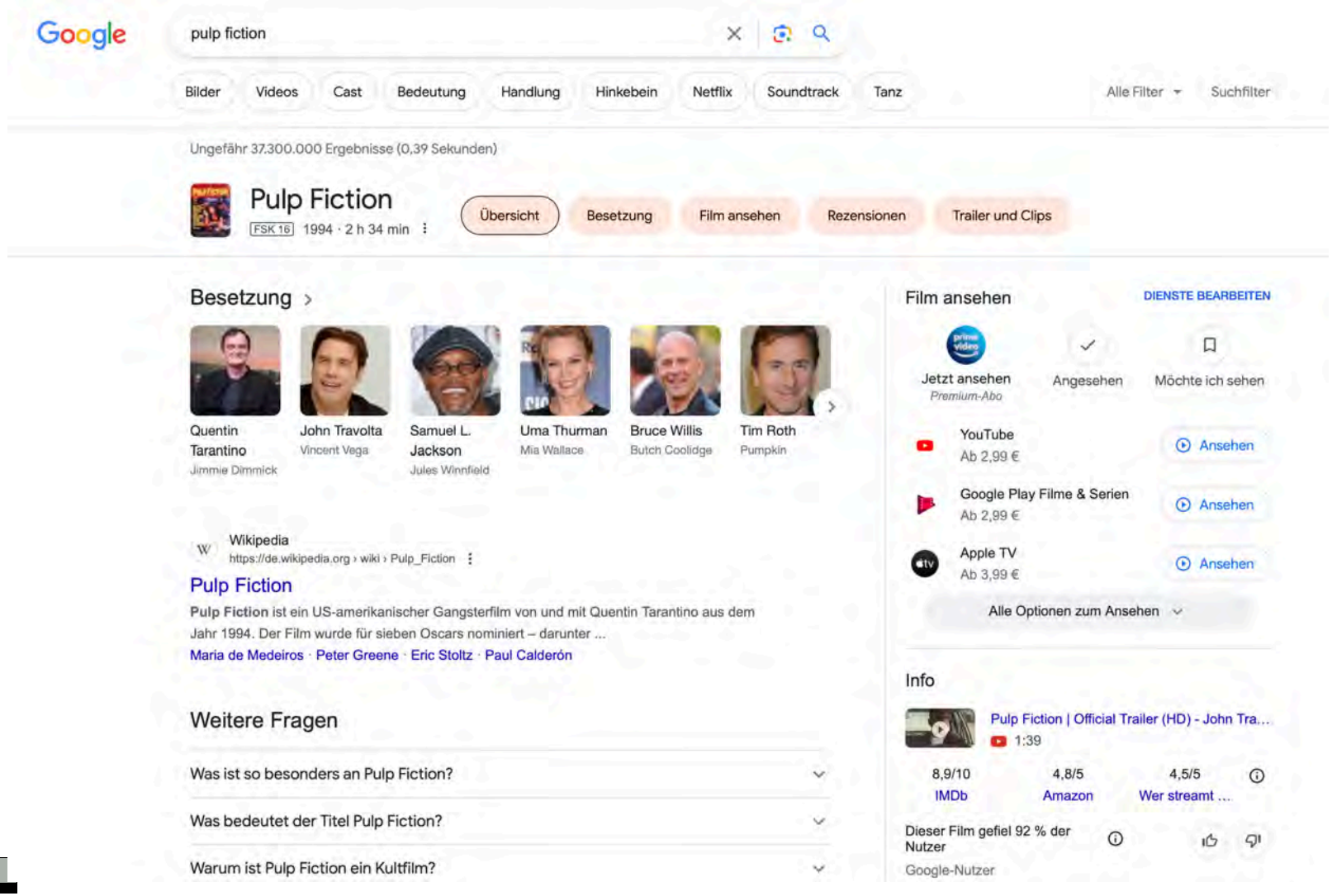
Metadata standards ≈ Checklists

- 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"?!



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 ambigous 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**

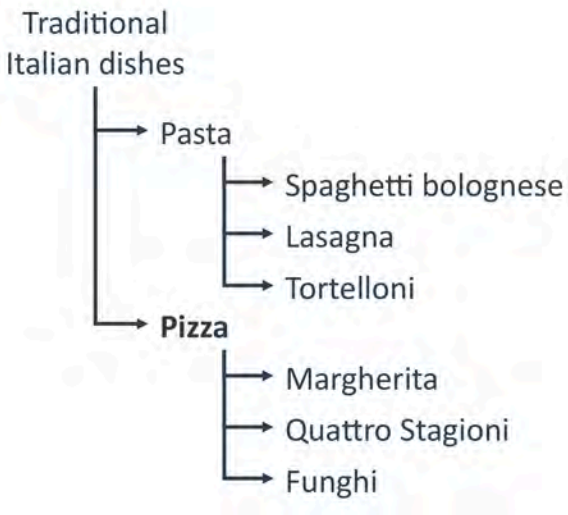
Dictionary

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

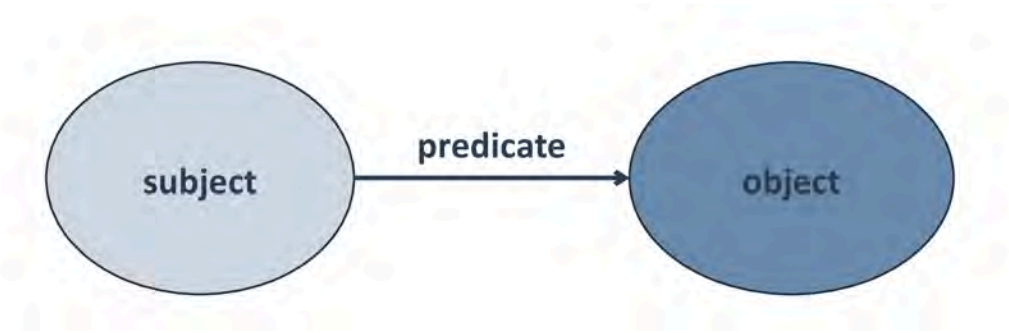
Dictionary of synonyms and relations

Pizza ≈ Lahmacun ≈ Focaccia ≈ Flammkuchen

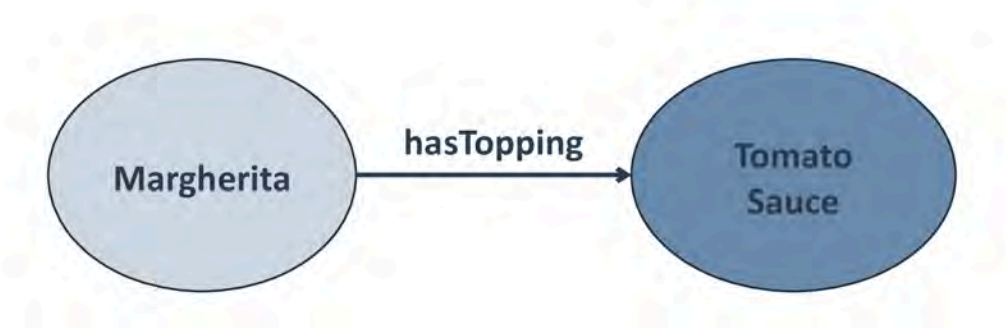
Ontology

- 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**

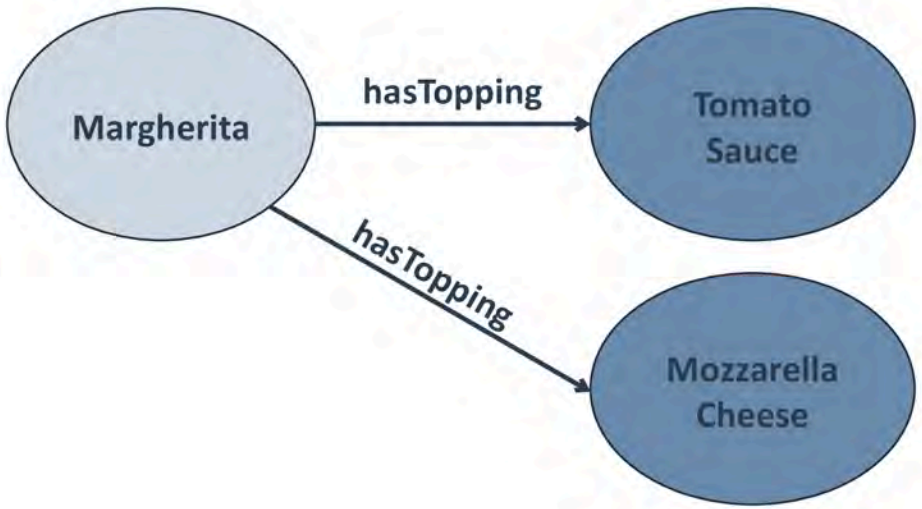
The semantic triple



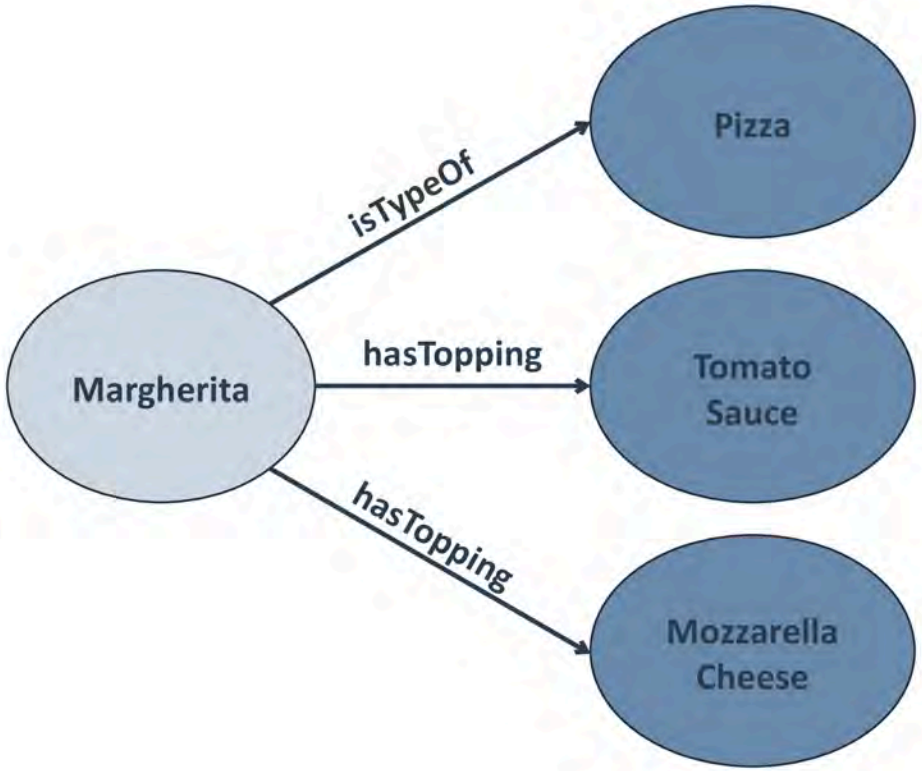
Modeling a pizza menu



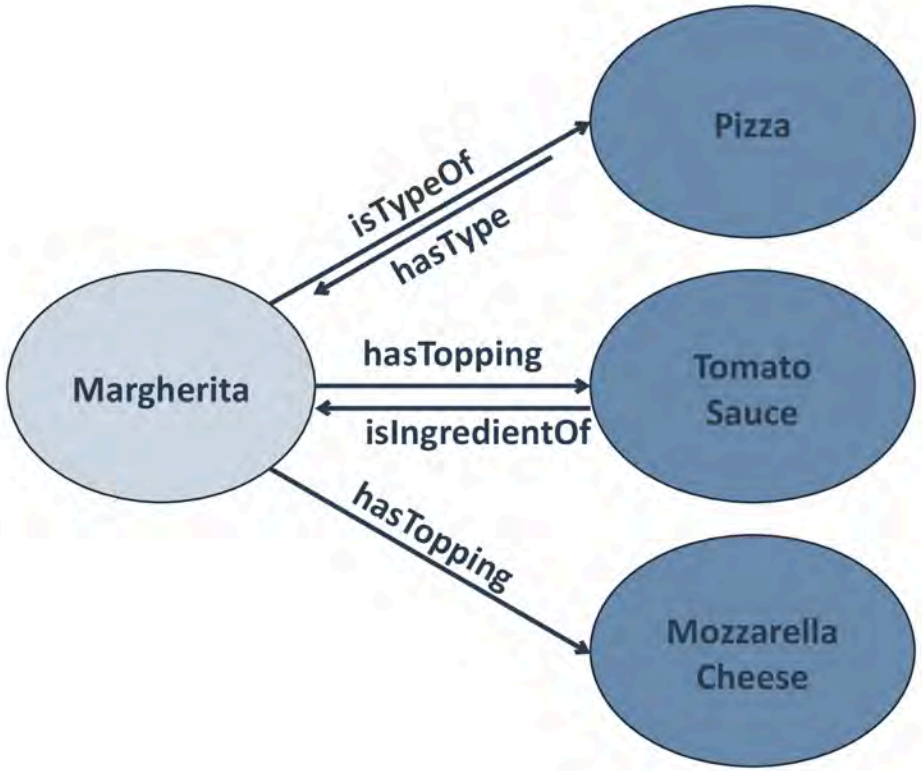
Modeling a pizza menu



Modeling a pizza menu

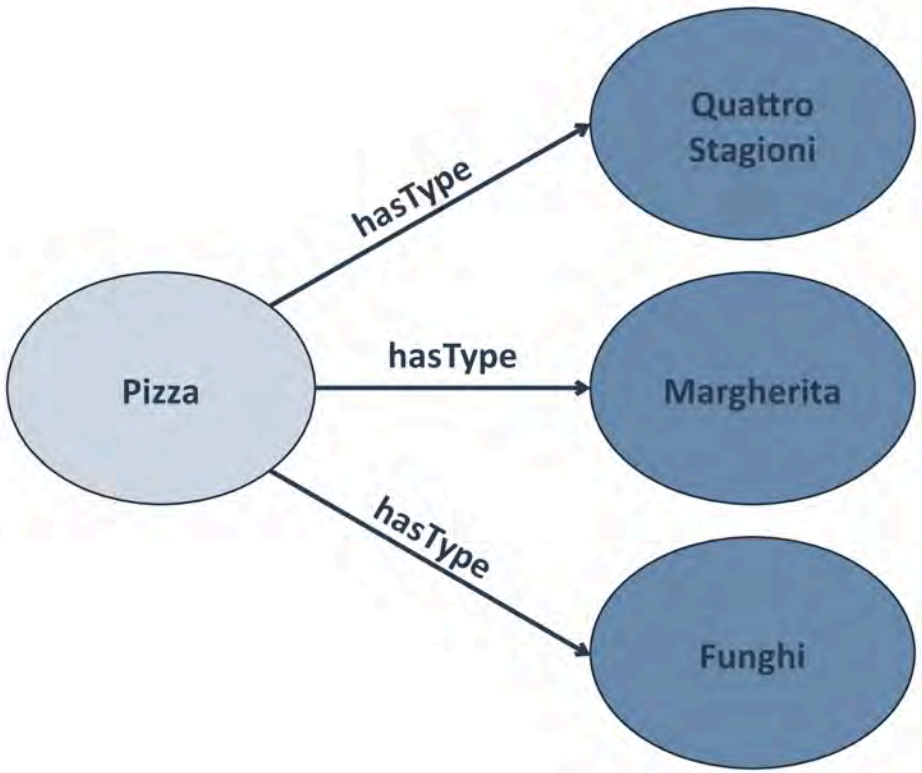


Predicates have two directions

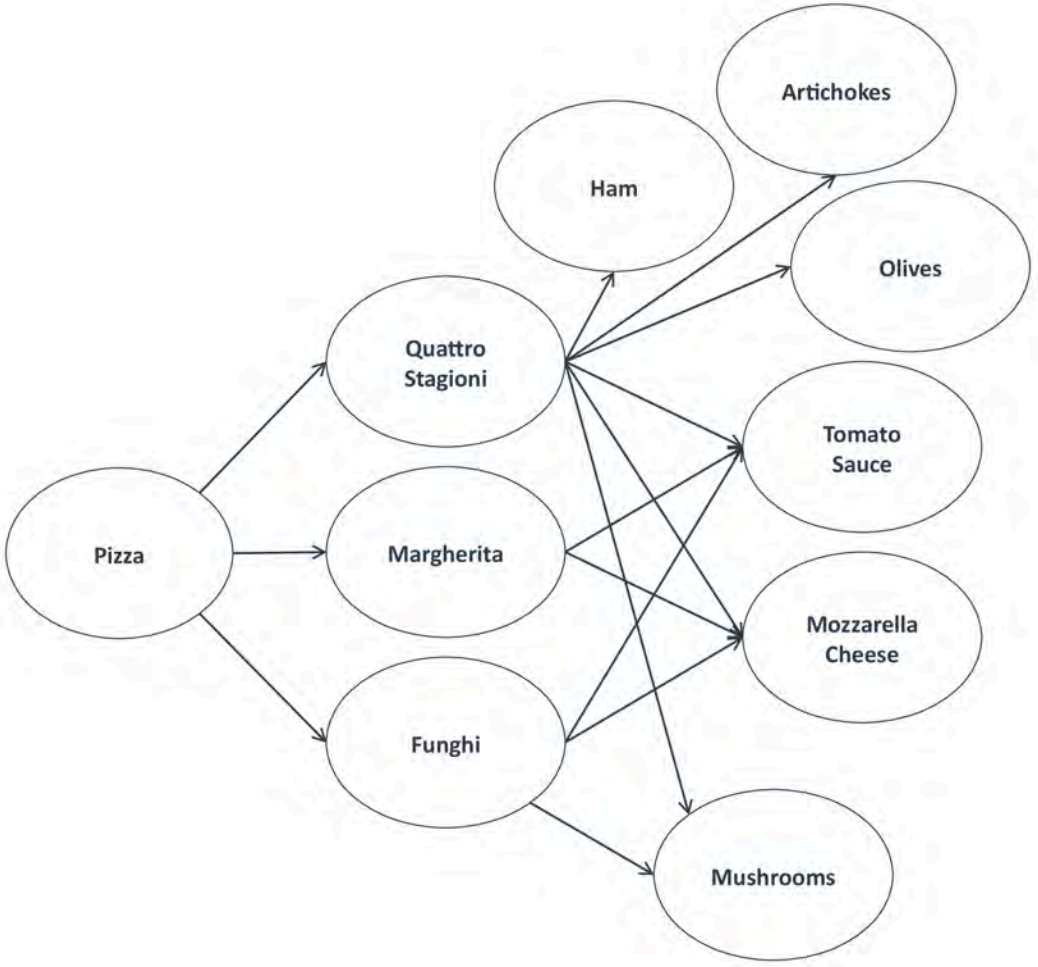


Looking at the menu from a different perspective

An object of one triplet can be the subject to another



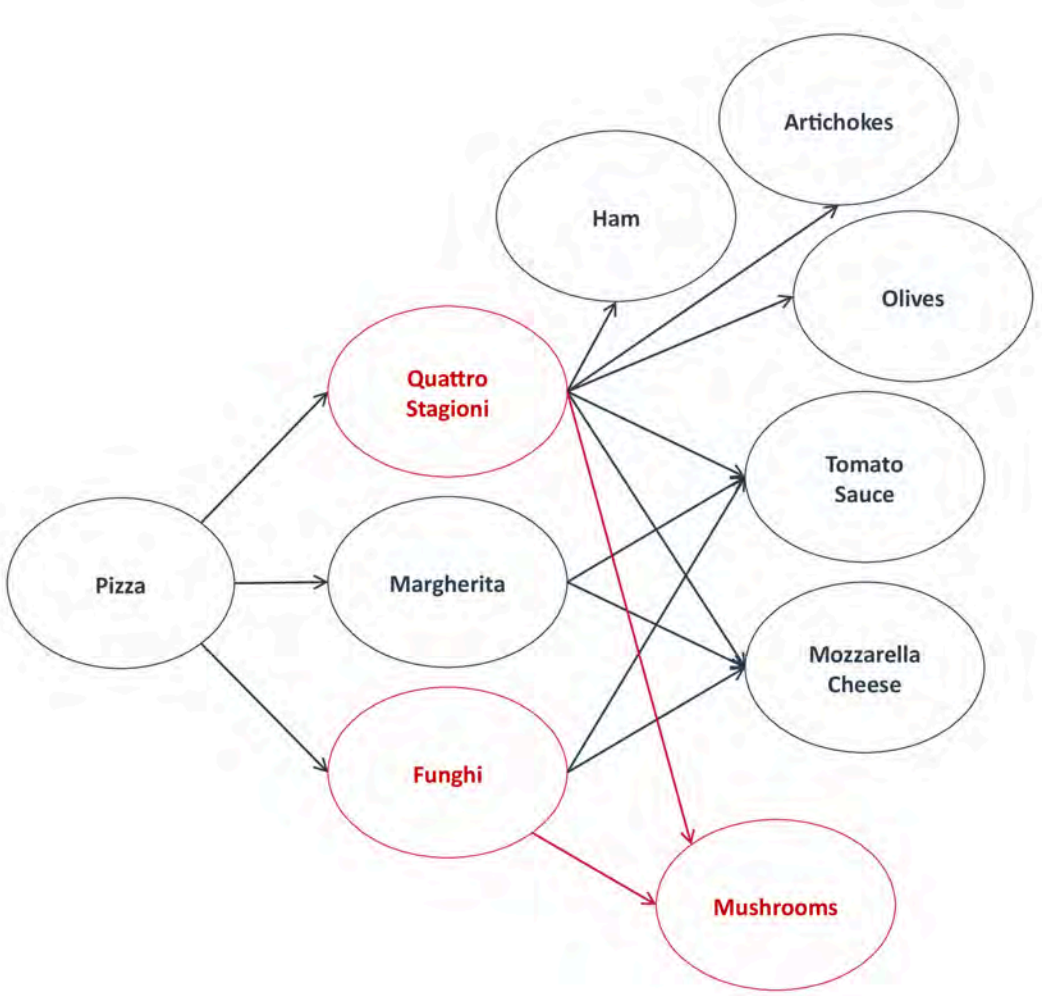
(Towards) a knowledge graph



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

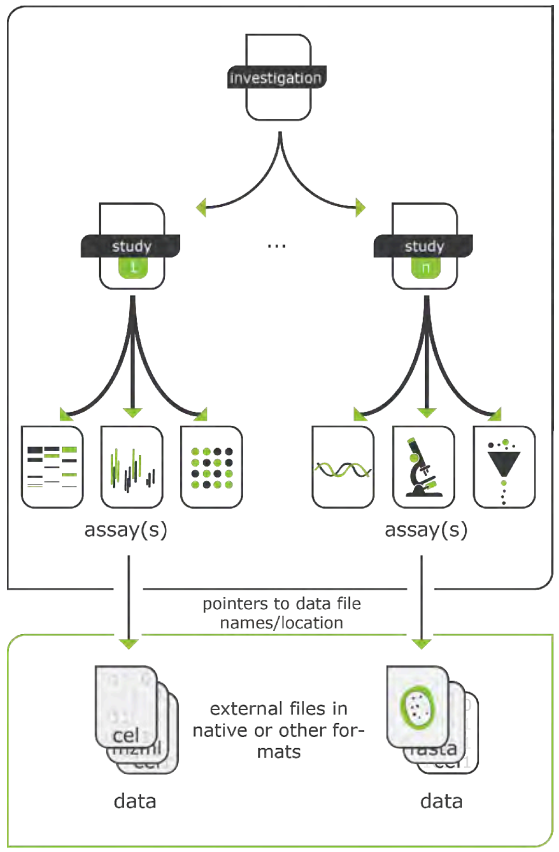
- Description: http://edamontology.org/page
- Browser: https://edamontology.github.io/edam-browser

PECO ontology

- Human-readble: 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

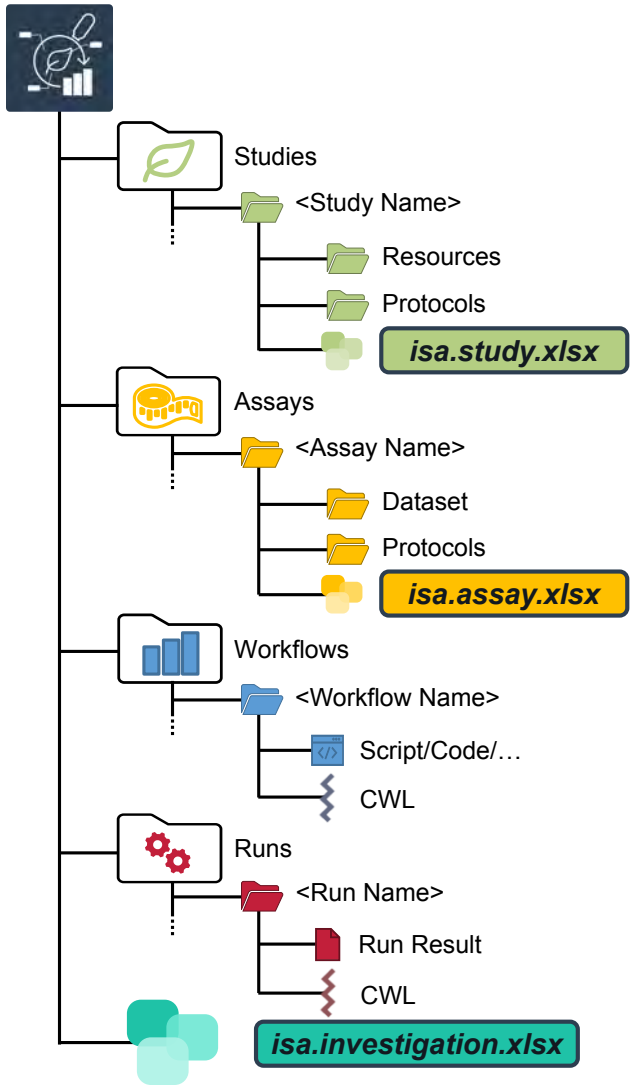
Overall goals
Scientific context

Study

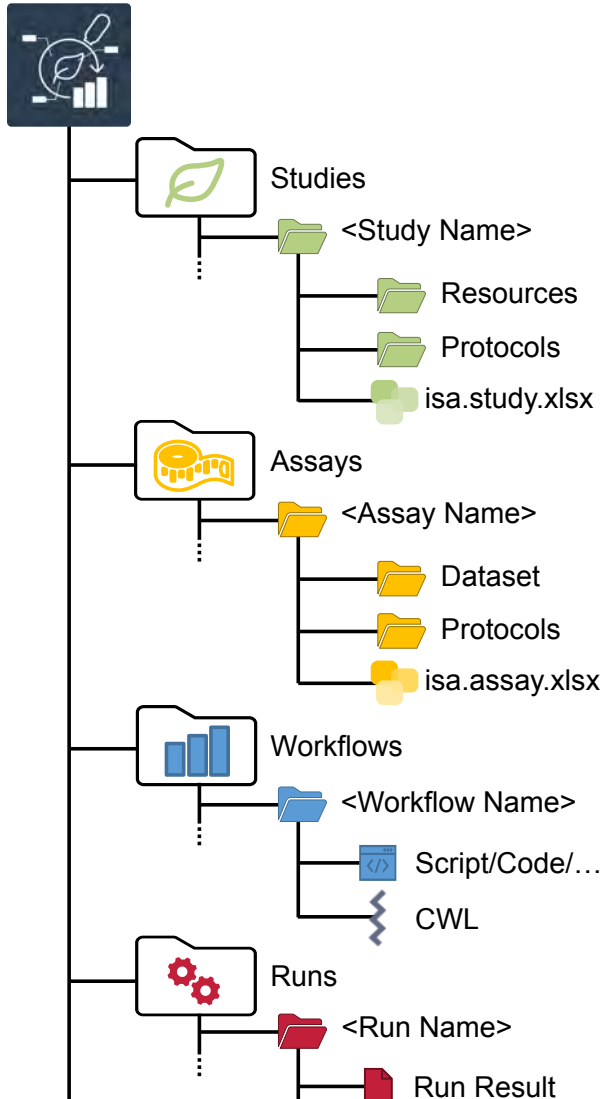
Experimental design

Assay

Leading to (raw) data



ARC builds on ISA



Metadata Annotations

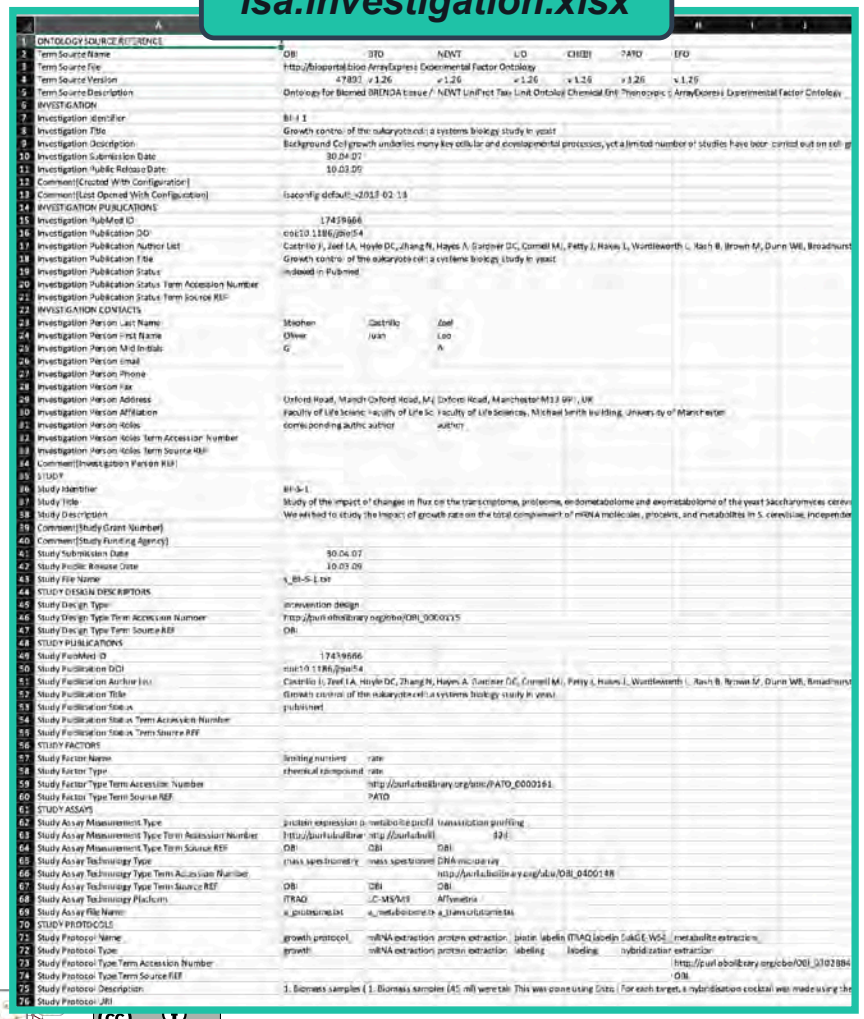
Source Name	Characteristics [Organism]	Sample Name
Ath001	*Arabidopsis thaliana*	Lre001
Ath002	*Arabidopsis thaliana*	Lre002
Ath003	*Arabidopsis thaliana*	Lre003
…	…	…

Source Name	Parameter [instrument model]	Raw Data File
Lre001	6130 Quadrupole LC/MS	Lre001.wiff
Lre002	6130 Quadrupole LC/MS	Lre002.wiff
Lre003	6130 Quadrupole LC/MS	Lre003.wiff
…	…	…

Investigation Title	
Investigation Description	
Investigation Person Last Name	

isa.<>.xlsx files within ARCs

isa.investigation.xlsx



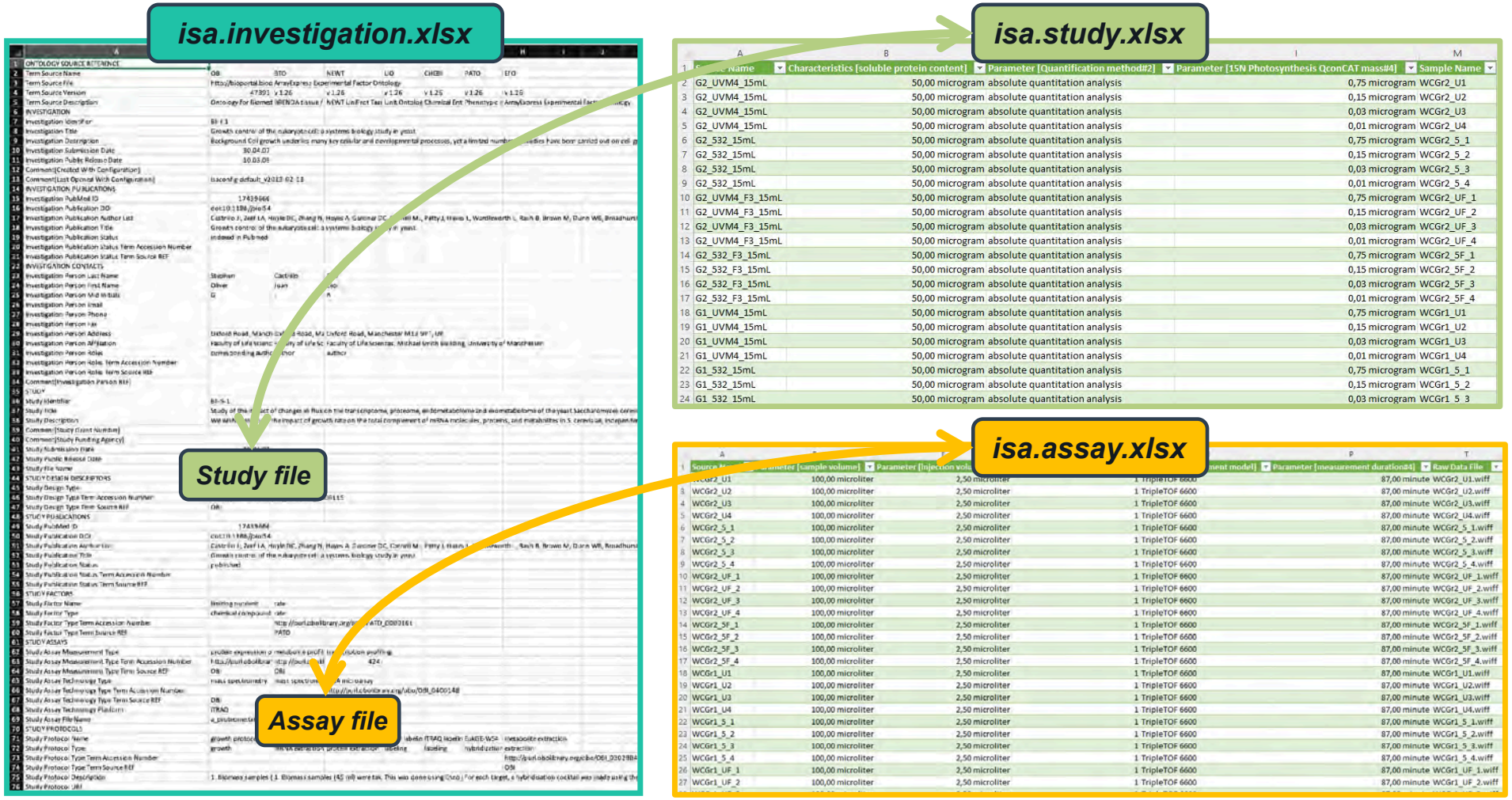
isa.study.xlsx

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.assay.xlsx

Source Name	Parameter [sample volume]	Parameter [injection volu	ment model]	Parameter [measurement duration#4]	Raw Data File
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_5F_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_1.wiff
WCGr2_5F_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_2.wiff
WCGr2_5F_3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_3.wiff
WCGr2_5F_4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_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

Study and assay files are registered in the investigation file



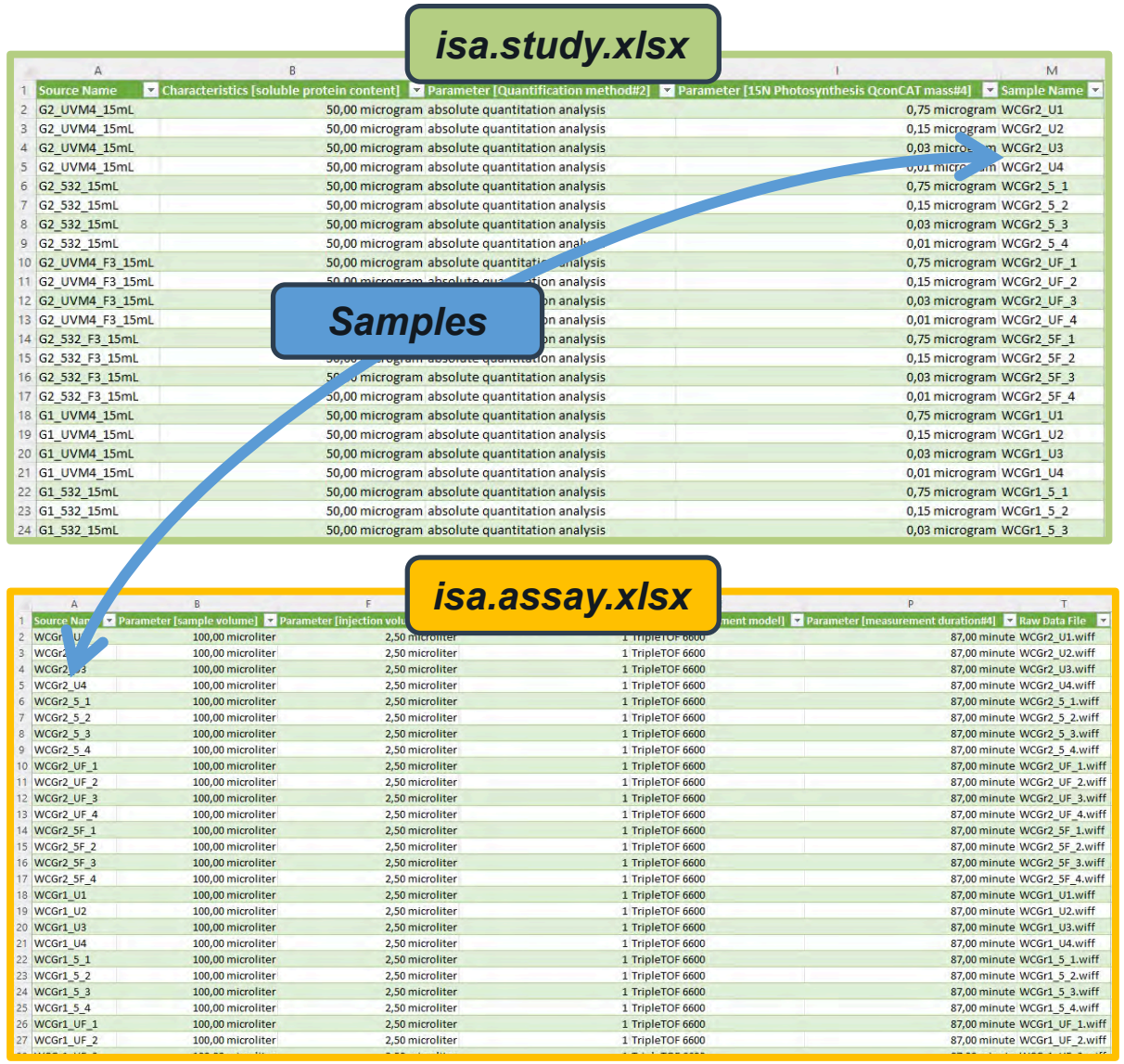
Sample Name	Characteristics [soluble protein content]	Parameter [Quantification method#2]	Parameter [15N Photosynthesis QconCAT mass#4]	Sample Name
G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,75 microgram	WCGr2_U1
G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis	0,15 microgram	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

Source Name	Parameter [sample volume]	Parameter [Injection volume]	[instrument model]	Parameter [measurement duration#4]	Raw Data File
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_5F_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_1.wiff
WCGr2_5F_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_2.wiff
WCGr2_5F_3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_3.wiff
WCGr2_5F_4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_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

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

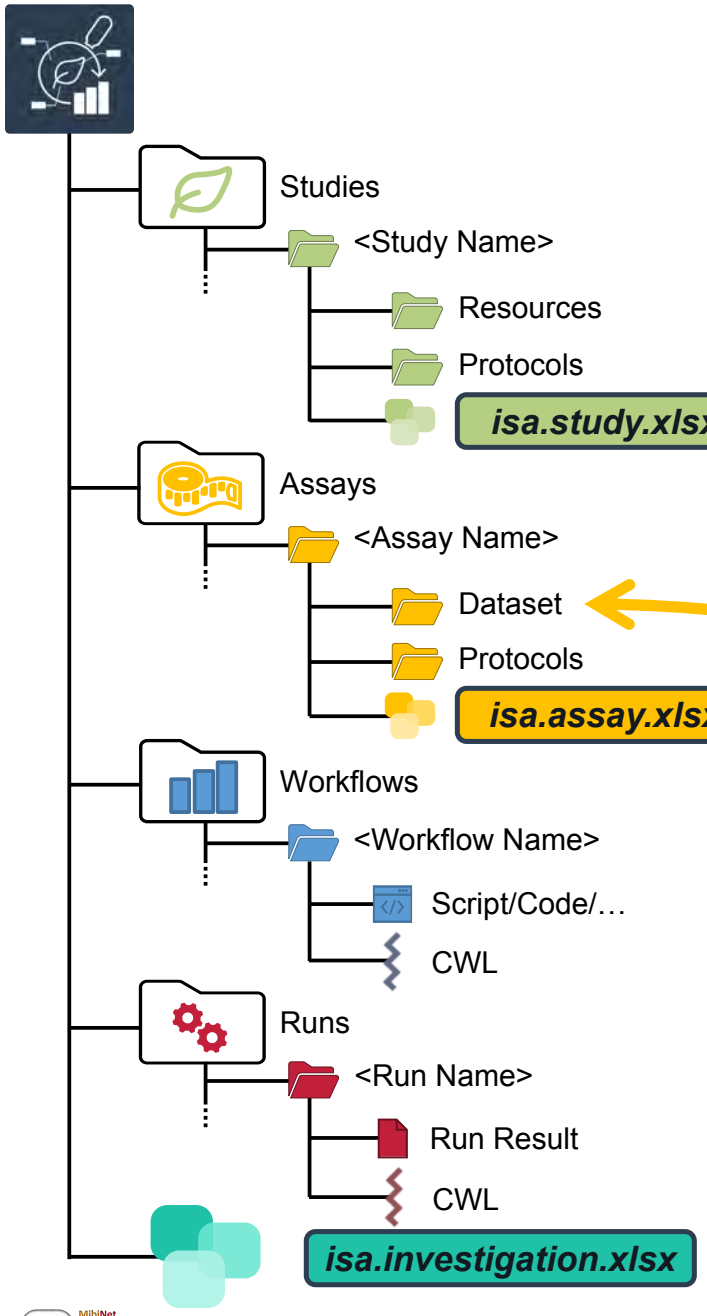
Output building blocks:

- Sample Name
- Raw Data File
- Derived Data File



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

Source Name	Parameter [sample volume]	Parameter [injection volume]	Parameter [measurement model]	Parameter [measurement duration#4]	Raw Data File
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_5F_1	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_1.wiff
WCGr2_5F_2	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_2.wiff
WCGr2_5F_3	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_3.wiff
WCGr2_5F_4	100,00 microliter	2,50 microliter	1 TripleTOF 6600	87,00 minute	WCGr2_5F_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.study.xlsx

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.assay.xlsx

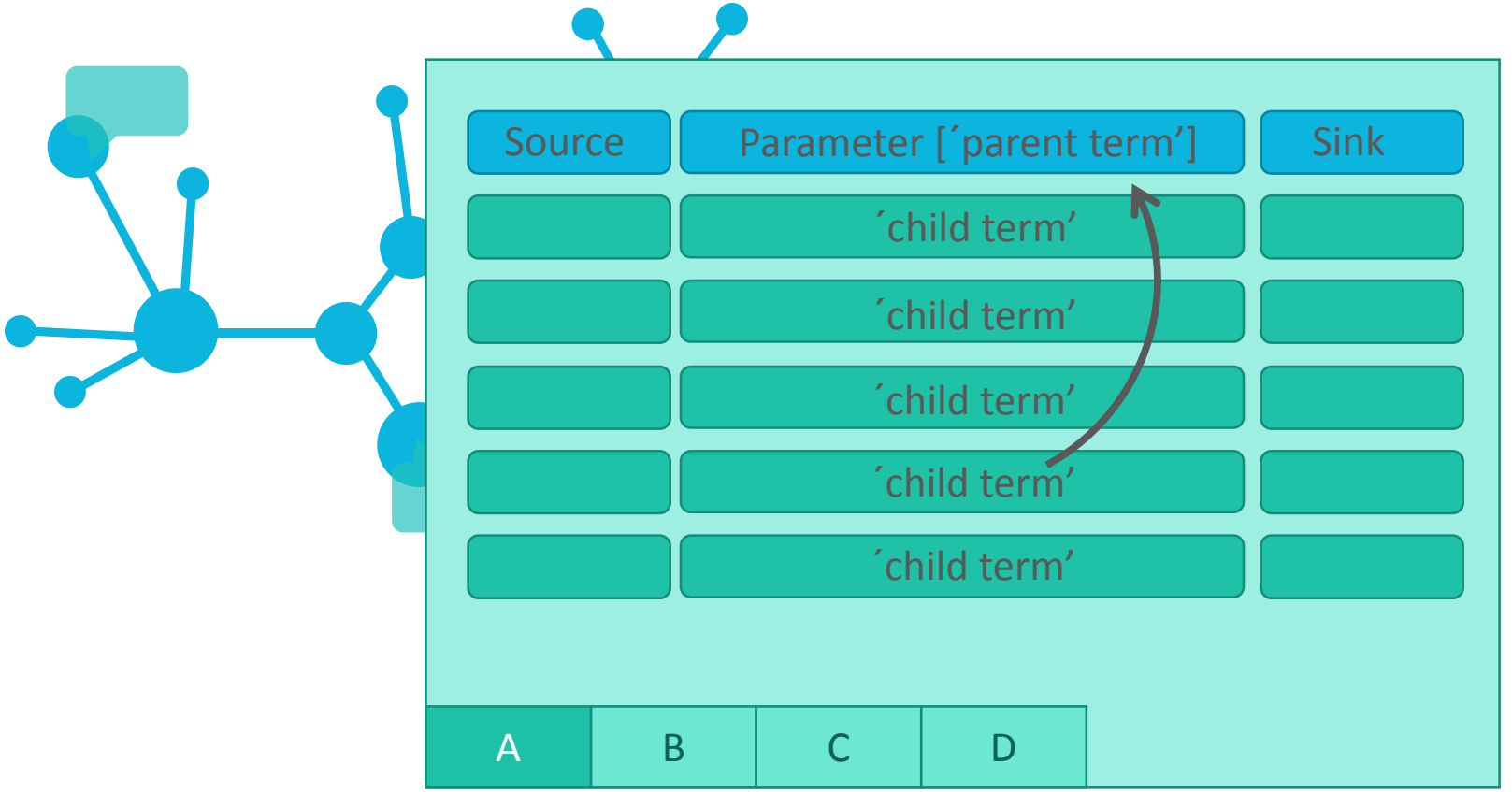
Source Name	Parameter [injection volume]	Parameter [measurement duration#4]	Raw Data File
WCGr2_U1	2,50 microliter	87,00 minute	WCGr2_U1.wiff
WCGr2_U2	2,50 microliter	87,00 minute	WCGr2_U2.wiff
WCGr2_U3	2,50 microliter	87,00 minute	WCGr2_U3.wiff
WCGr2_U4	2,50 microliter	87,00 minute	WCGr2_U4.wiff
WCGr2_5_1	2,50 microliter	87,00 minute	WCGr2_5_1.wiff
WCGr2_5_2	2,50 microliter	87,00 minute	WCGr2_5_2.wiff
WCGr2_5_3	2,50 microliter	87,00 minute	WCGr2_5_3.wiff
WCGr2_5_4	2,50 microliter	87,00 minute	WCGr2_5_4.wiff
WCGr2_UF_1	2,50 microliter	87,00 minute	WCGr2_UF_1.wiff
WCGr2_UF_2	2,50 microliter	87,00 minute	WCGr2_UF_2.wiff
WCGr2_UF_3	2,50 microliter	87,00 minute	WCGr2_UF_3.wiff
WCGr2_UF_4	2,50 microliter	87,00 minute	WCGr2_UF_4.wiff
WCGr2_5F_1	2,50 microliter	87,00 minute	WCGr2_5F_1.wiff
WCGr2_5F_2	2,50 microliter	87,00 minute	WCGr2_5F_2.wiff
WCGr2_5F_3	2,50 microliter	87,00 minute	WCGr2_5F_3.wiff
WCGr2_5F_4	2,50 microliter	87,00 minute	WCGr2_5F_4.wiff
WCGr1_U1	2,50 microliter	87,00 minute	WCGr1_U1.wiff
WCGr1_U2	2,50 microliter	87,00 minute	WCGr1_U2.wiff
WCGr1_U3	2,50 microliter	87,00 minute	WCGr1_U3.wiff
WCGr1_U4	2,50 microliter	87,00 minute	WCGr1_U4.wiff
WCGr1_5_1	2,50 microliter	87,00 minute	WCGr1_5_1.wiff
WCGr1_5_2	2,50 microliter	87,00 minute	WCGr1_5_2.wiff
WCGr1_5_3	2,50 microliter	87,00 minute	WCGr1_5_3.wiff
WCGr1_5_4	2,50 microliter	87,00 minute	WCGr1_5_4.wiff
WCGr1_UF_1	2,50 microliter	87,00 minute	WCGr1_UF_1.wiff
WCGr1_UF_2	2,50 microliter	87,00 minute	WCGr1_UF_2.wiff

WCGr2_5_1.wiff	WIFF File
WCGr2_5_2.wiff	WIFF File
WCGr2_5_3.wiff	WIFF File
WCGr2_5_4.wiff	WIFF File
WCGr2_U1.wiff	WIFF File
WCGr2_U2.wiff	WIFF File
WCGr2_U3.wiff	WIFF File
WCGr2_U4.wiff	WIFF File
WCGr2_UF_1.wiff	WIFF File
WCGr2_UF_2.wiff	WIFF File
WCGr2_UF_3.wiff	WIFF File
WCGr2_UF_4.wiff	WIFF File

Raw data

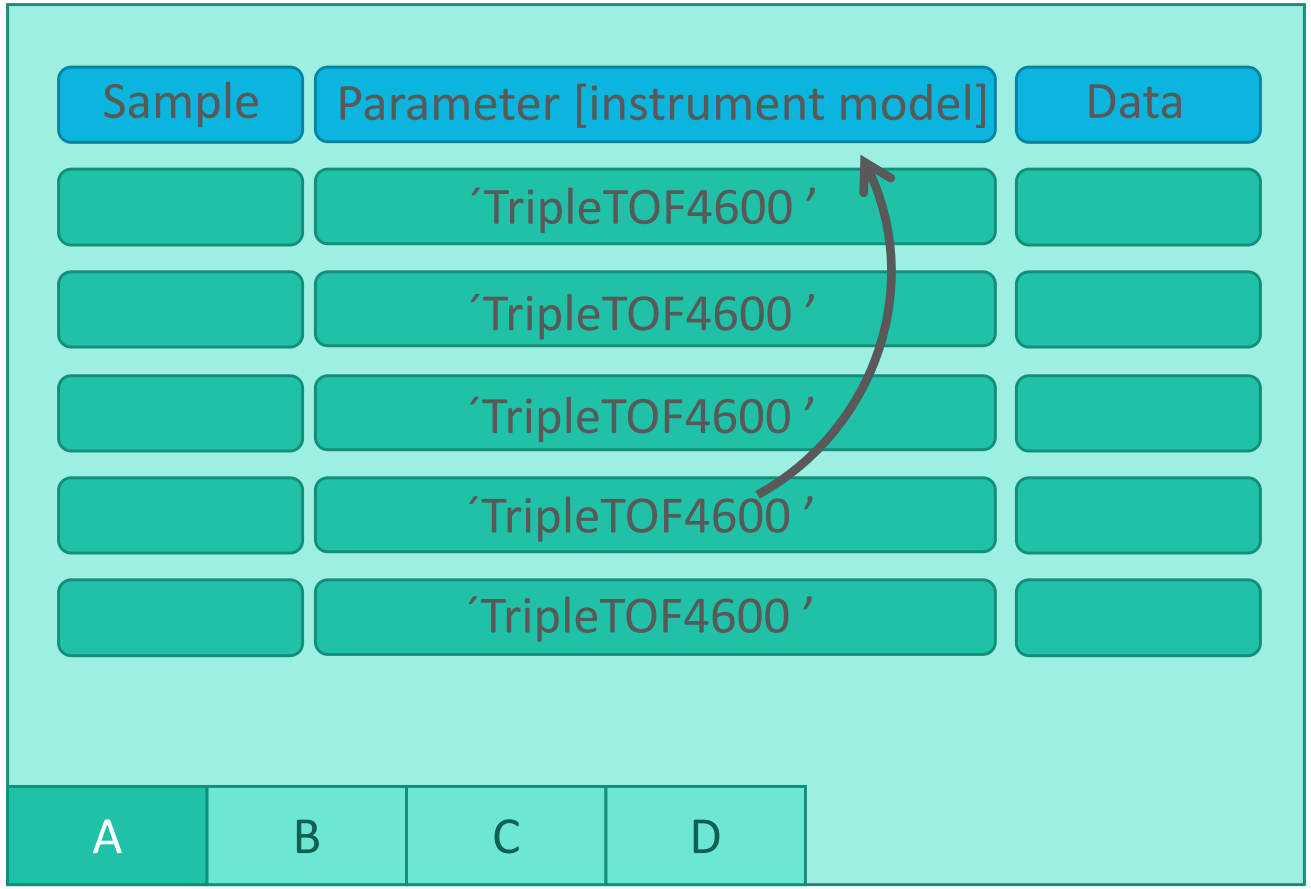
Swate

Annotation by flattening the knowledge graph



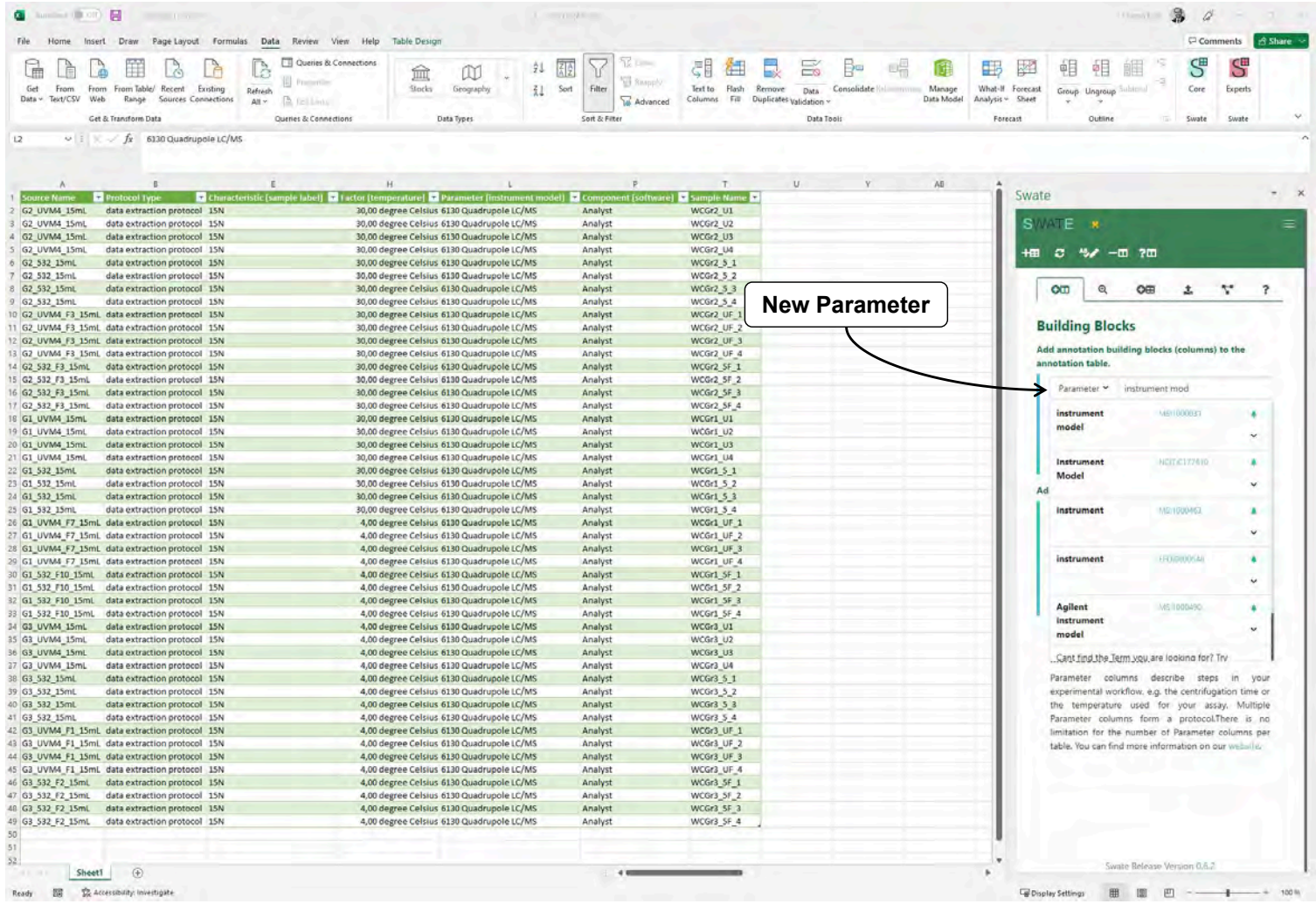
- Low-friction metadata annotation
- Familiar spreadsheet, row/column-based environment

Annotation principle



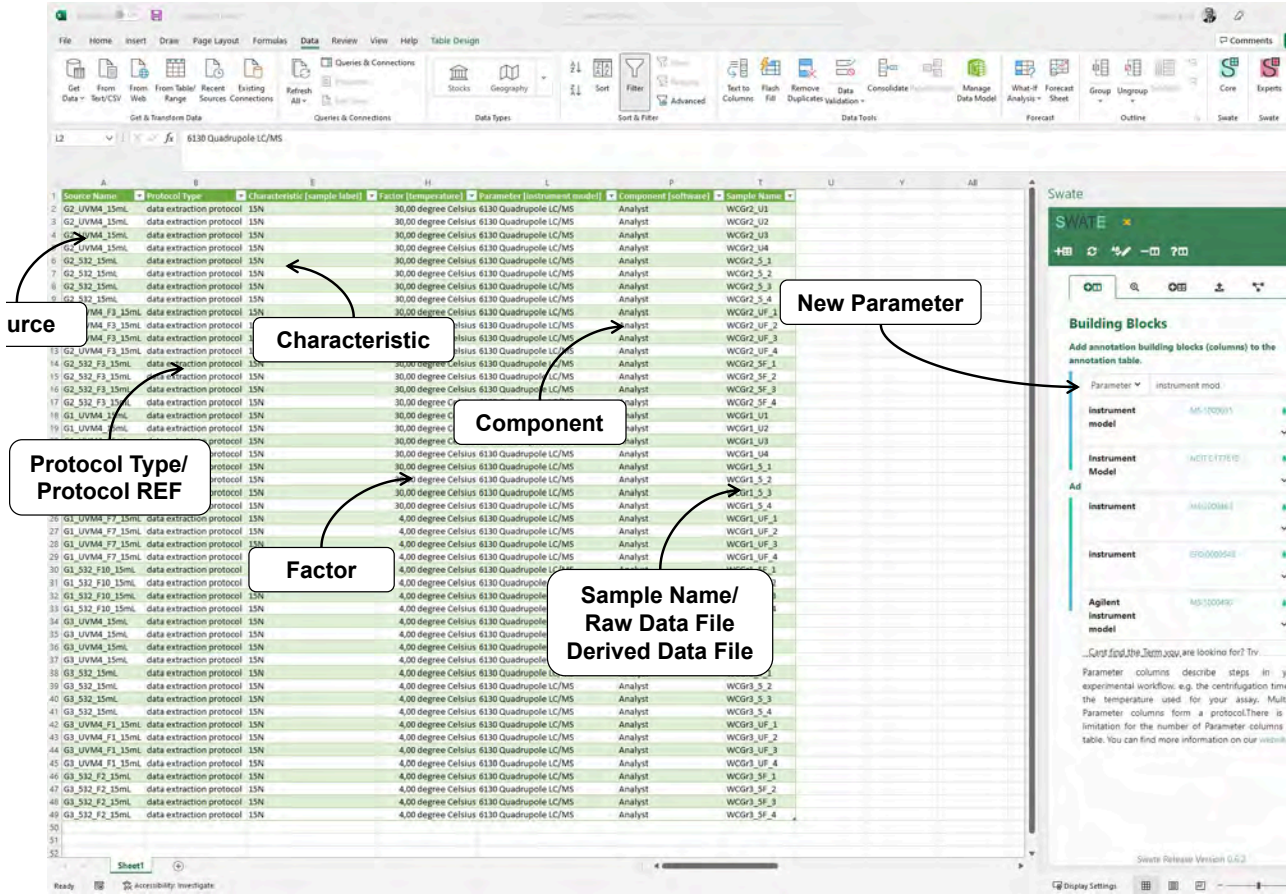
- Low-friction metadata annotation
- Familiar spreadsheet, row/column-based environment

Adding new building blocks (columns)



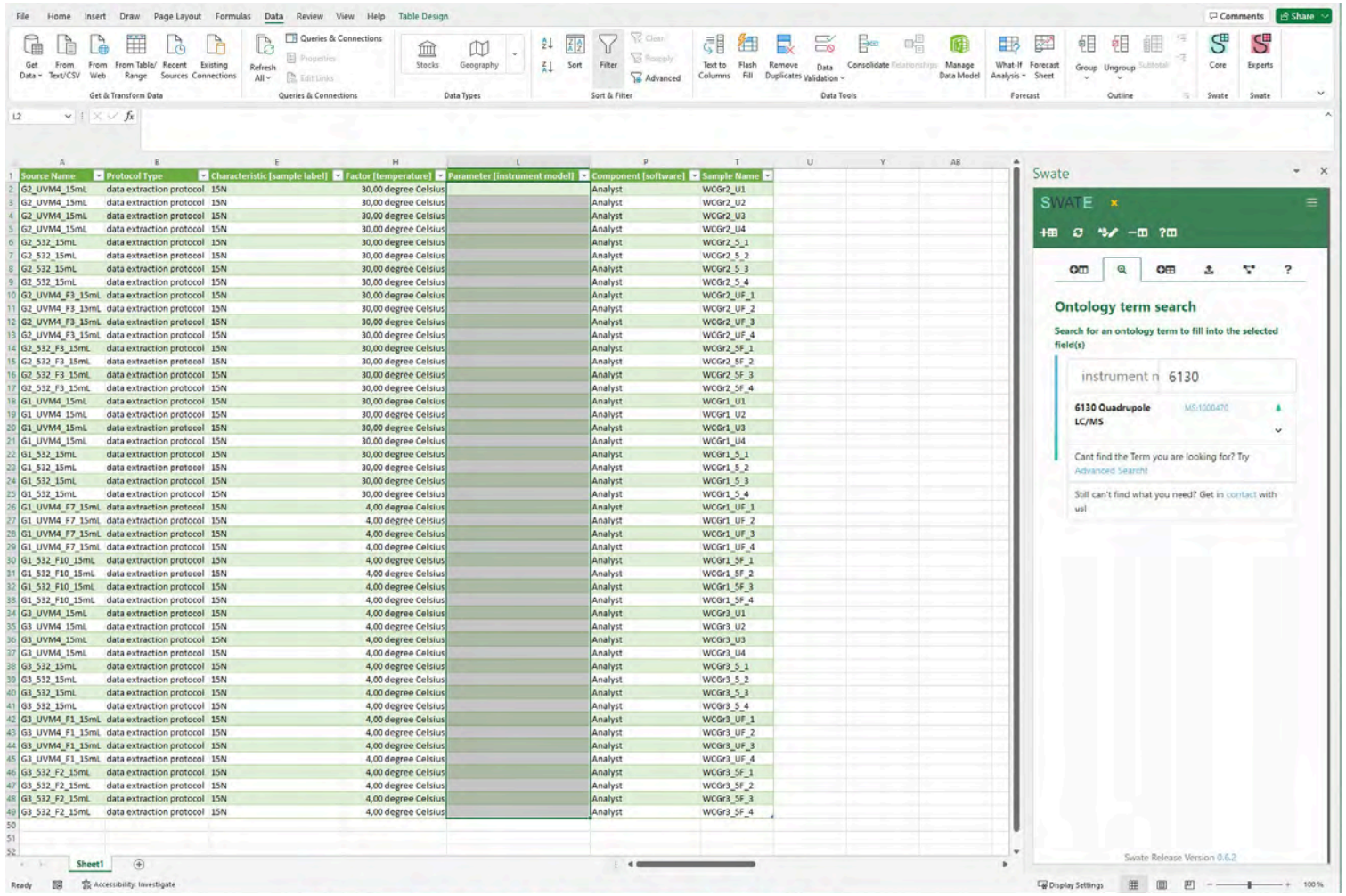
Annotation Building Block types

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



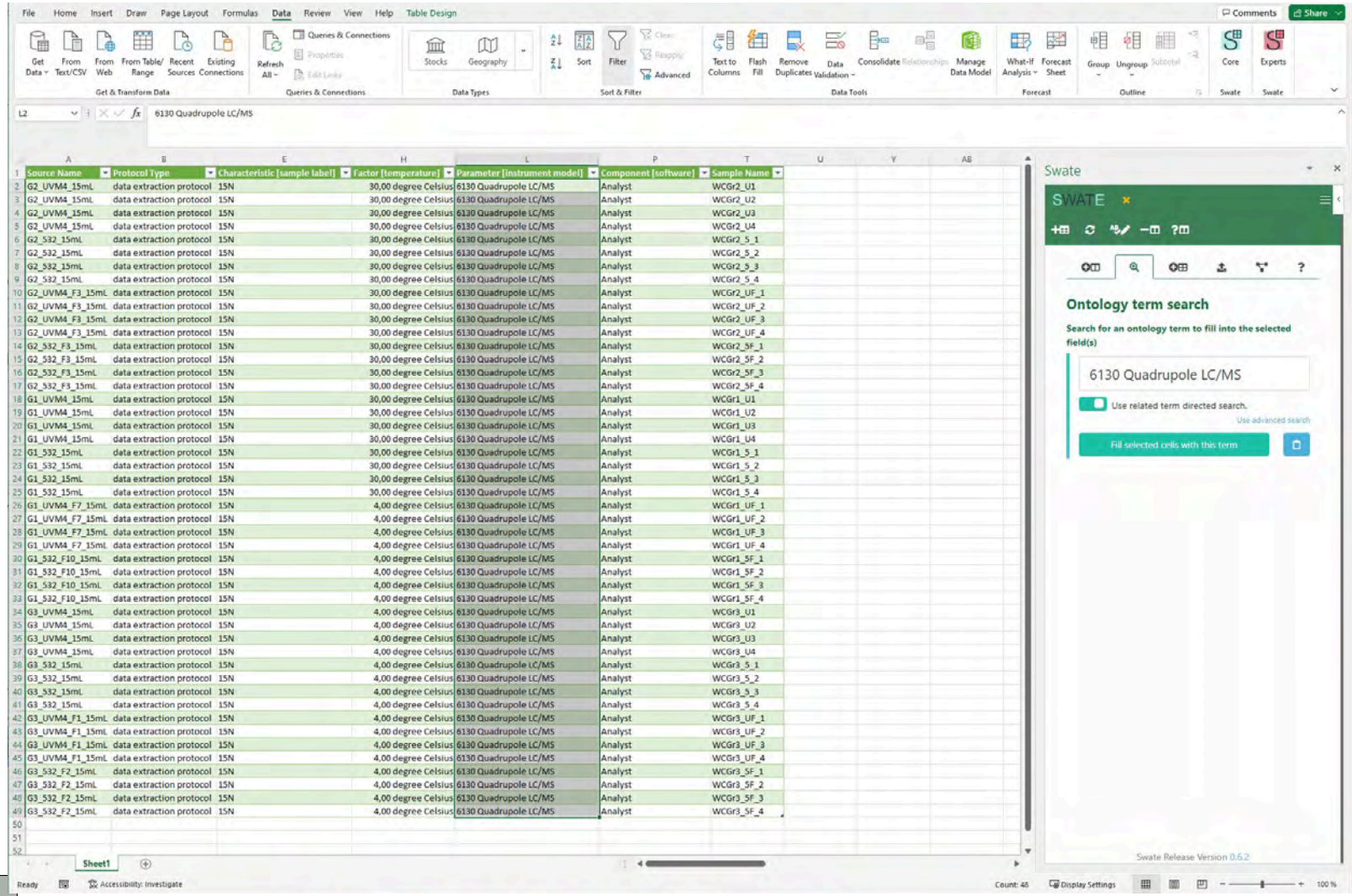
Let's take a detour on Annotation Principles | slides

Ontology term search

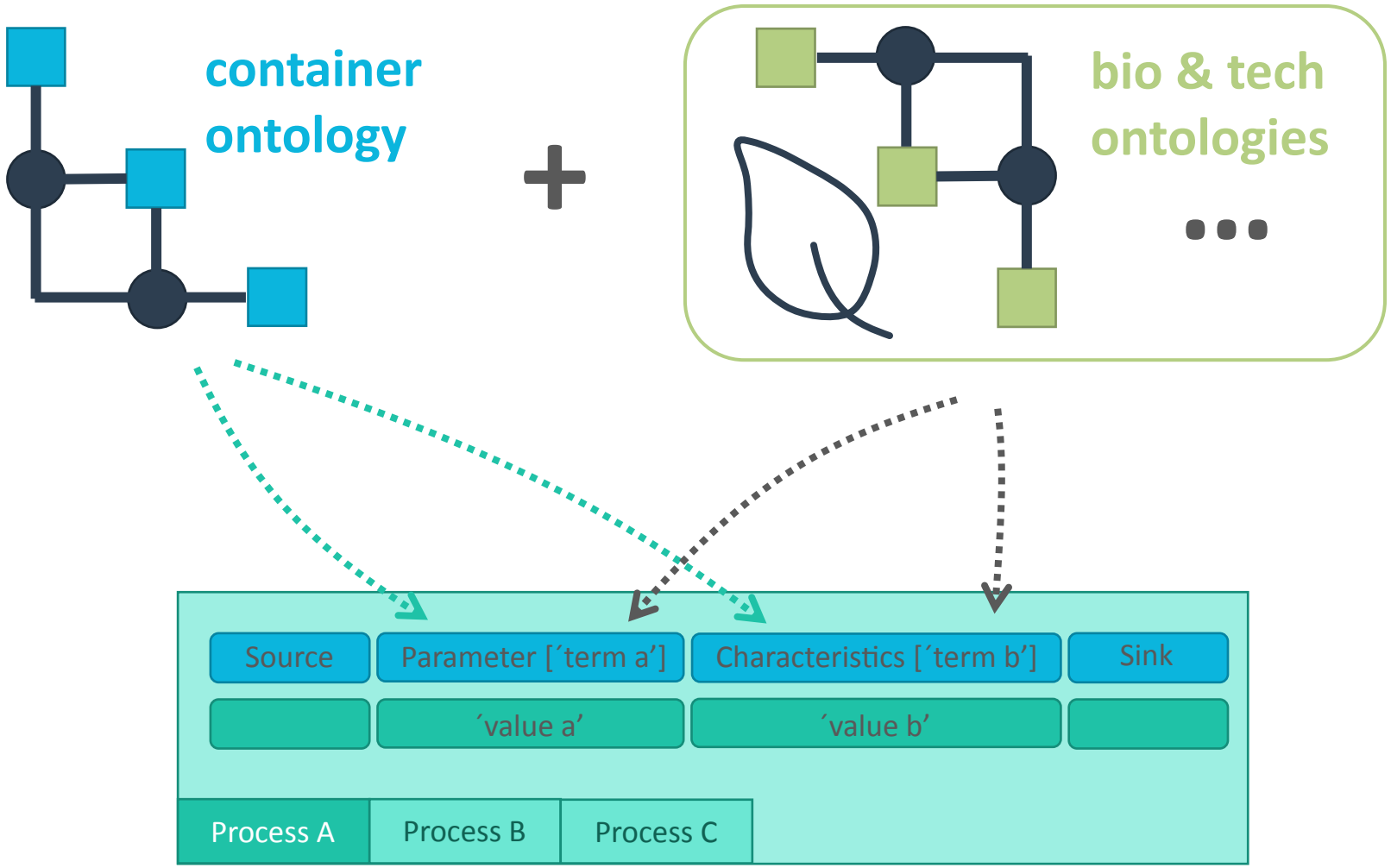


Enable **related term directed search** to directly fill cells with child terms

Fill your table with ontology terms



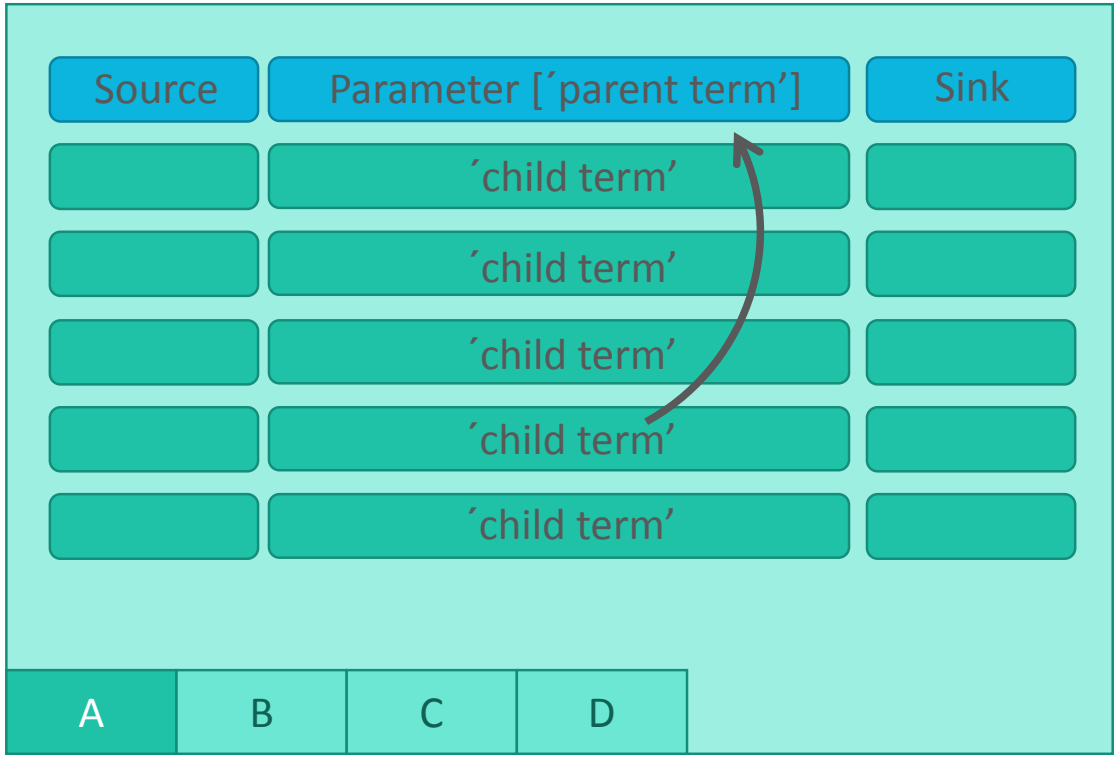
Hierarchical combination of ontologies



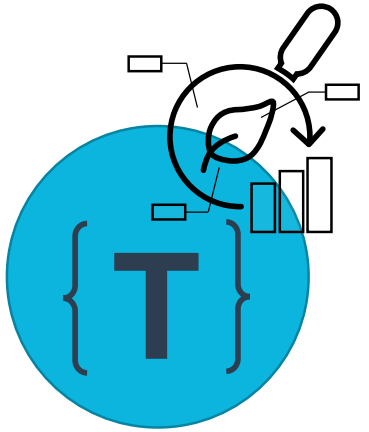
isa.study.xlsx or isa.assay.xlsx

Swate templates

Checklists and Templates

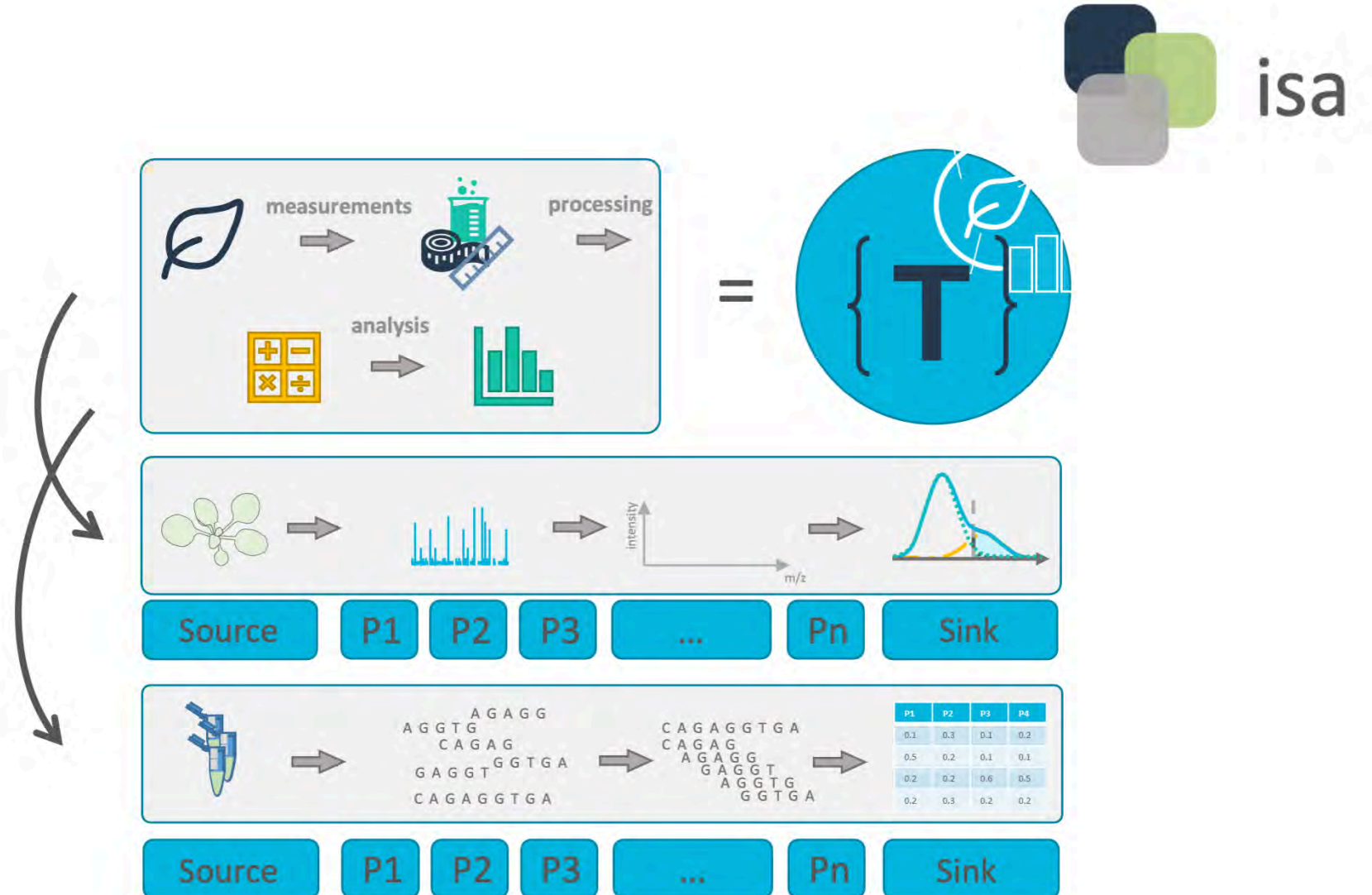


=



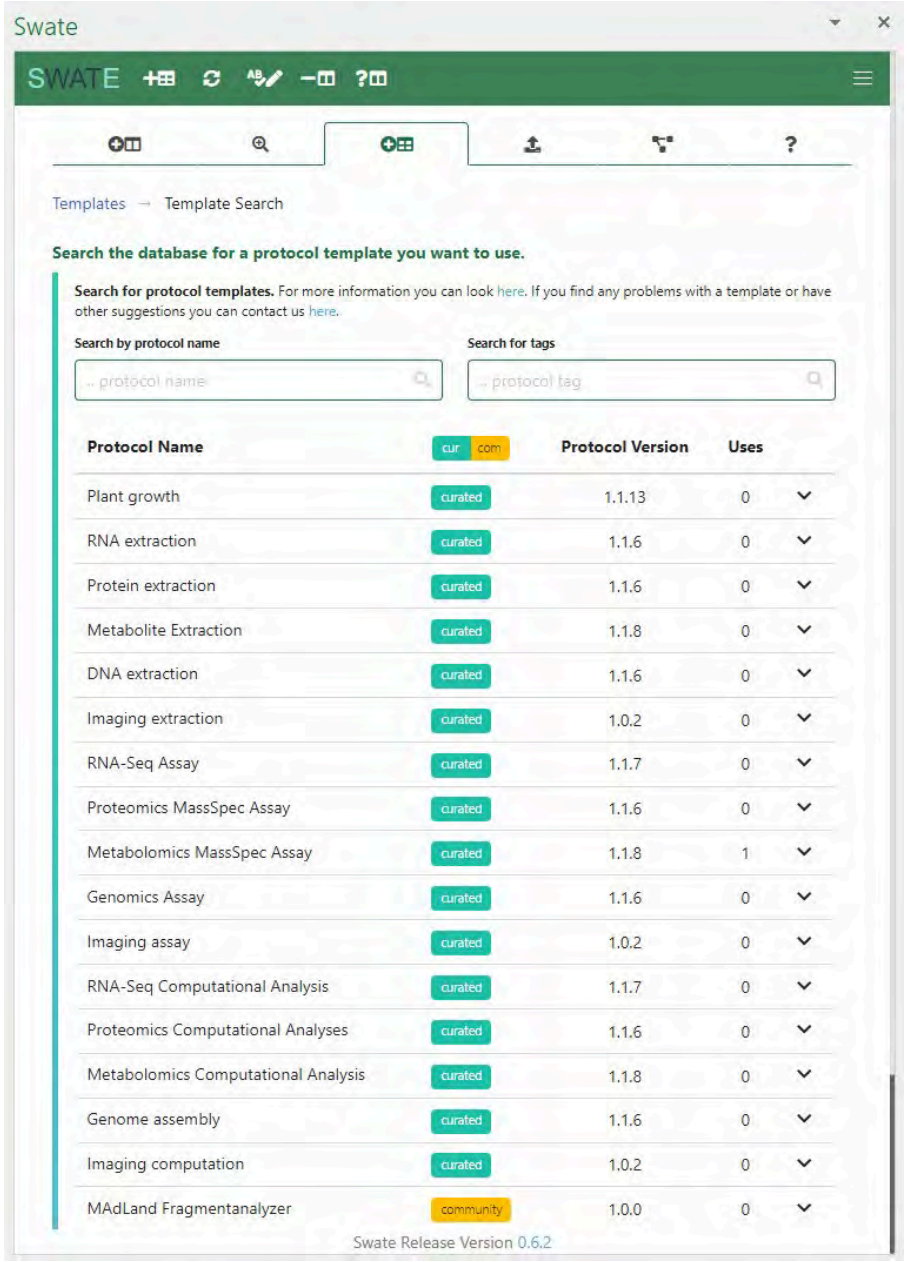
- Metadata standards or repository requirements can be represented as templates

Realization of lab-specific metadata templates



Directly import templates via Swate

- DataPLANT curated
- Community templates



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
- MacOS TextEdit

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

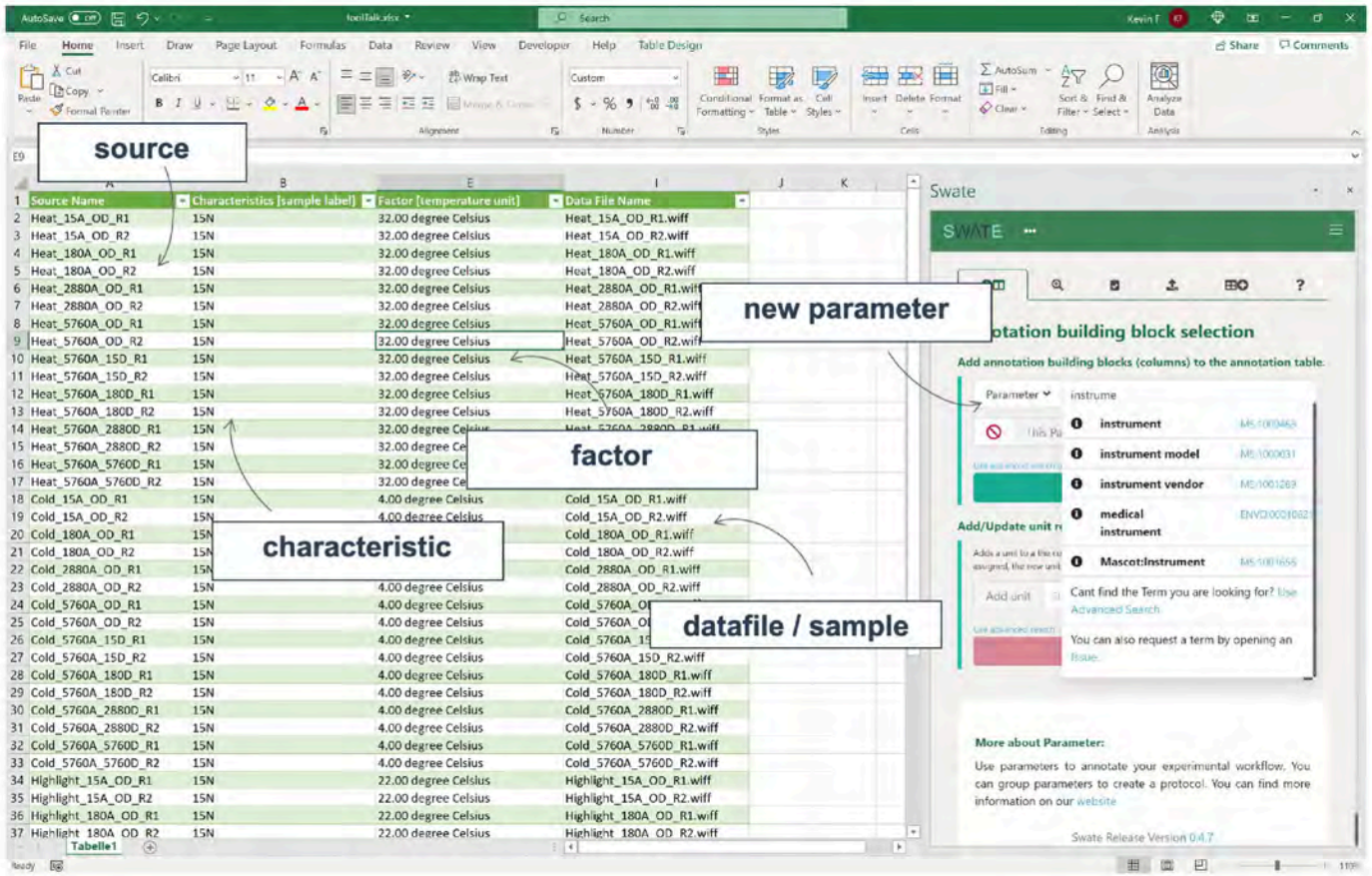
- Initiated an ARC
- Structured and ...
- Shared with collaborators

Today we want to

- ... **annotate the experimental data**

Swate hands-on with demo data

Swate Overview

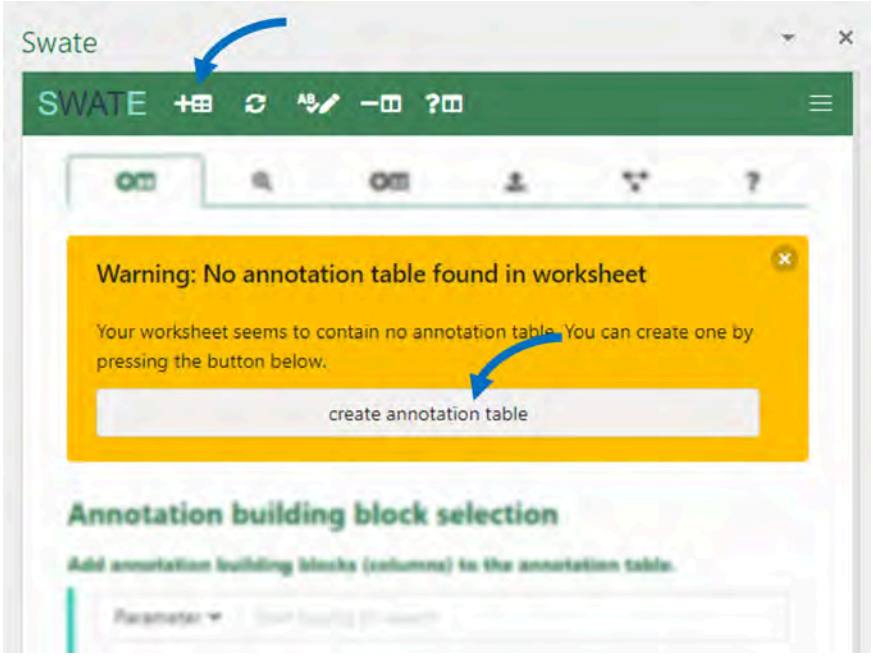


Let's annotate the plant samples first

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.

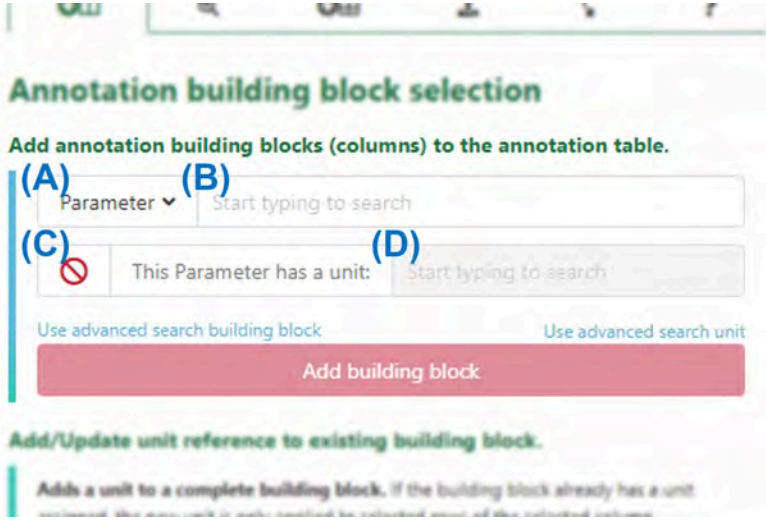


> 💡 Each table is by default created with one input (`Source Name`) and one output (`Sample Name`) column

> 💡 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.

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 `PECO:0007224`.
2. Check the box for *This Parameter has a unit* and search for `microeinstein per square meter per second` in the adjacent search bar.
3. Select `UO: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 microeinstein 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

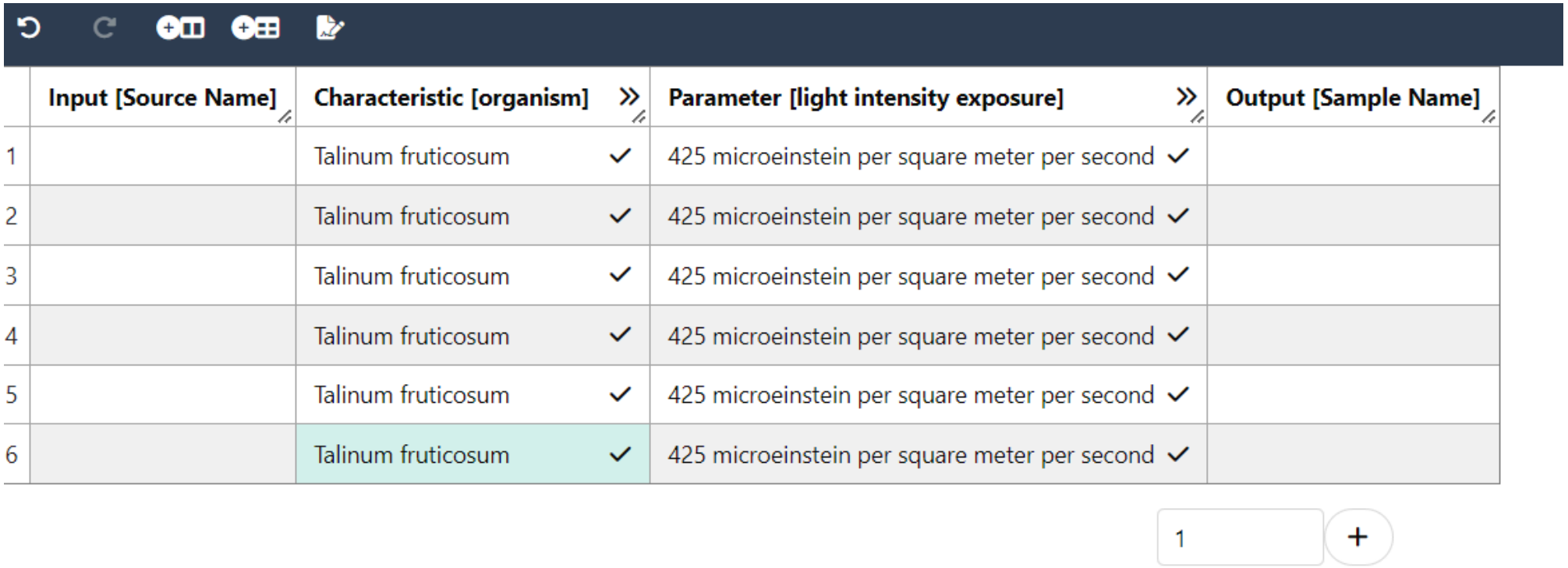
Update ontology reference columns

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 `microeinstein 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 microeinstein per square meter per second ✓	
2		Talinum fruticosum ✓	425 microeinstein per square meter per second ✓	
3		Talinum fruticosum ✓	425 microeinstein per square meter per second ✓	
4		Talinum fruticosum ✓	425 microeinstein per square meter per second ✓	
5		Talinum fruticosum ✓	425 microeinstein per square meter per second ✓	
6		Talinum fruticosum ✓	425 microeinstein per square meter per second ✓	

1 +

Hiding ontology reference columns

Click the `Autoformat Table` quick access button without holding `Ctrl` 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

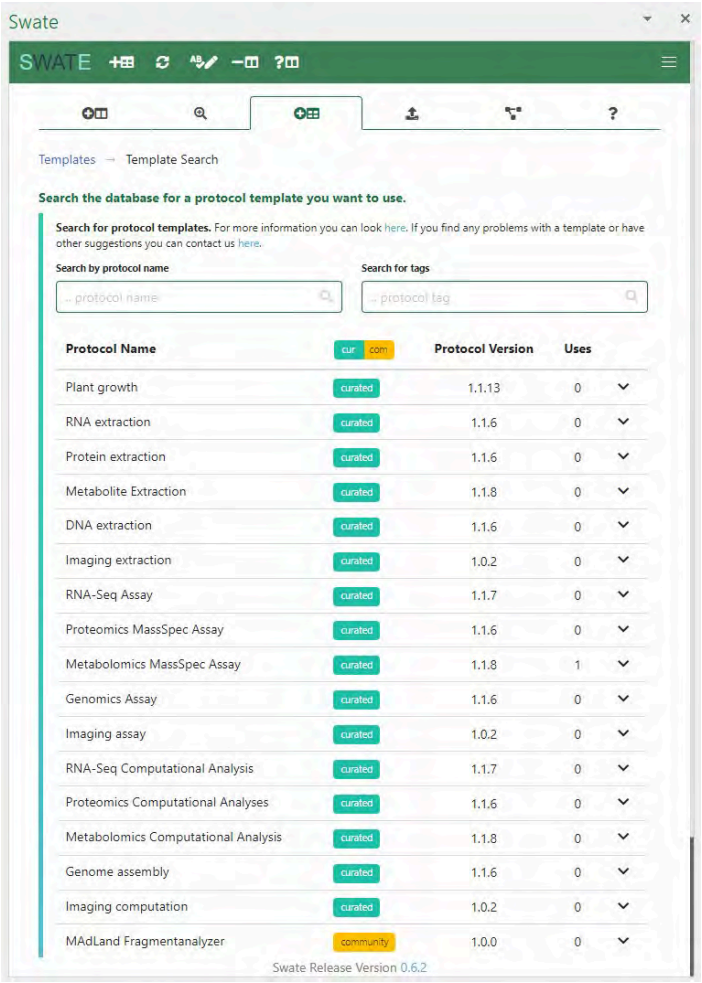
1. Navigate to the demo ARC.
2. Open the lab notes `assays/rnaseq/protocols/RNA_extraction.txt` in a text editor.
3. Open the empty `assays/rnaseq/isa.assay.xlsx)` workbook in Excel.

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 <u>do not exist yet</u>.



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 `Remove Building Block` 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:

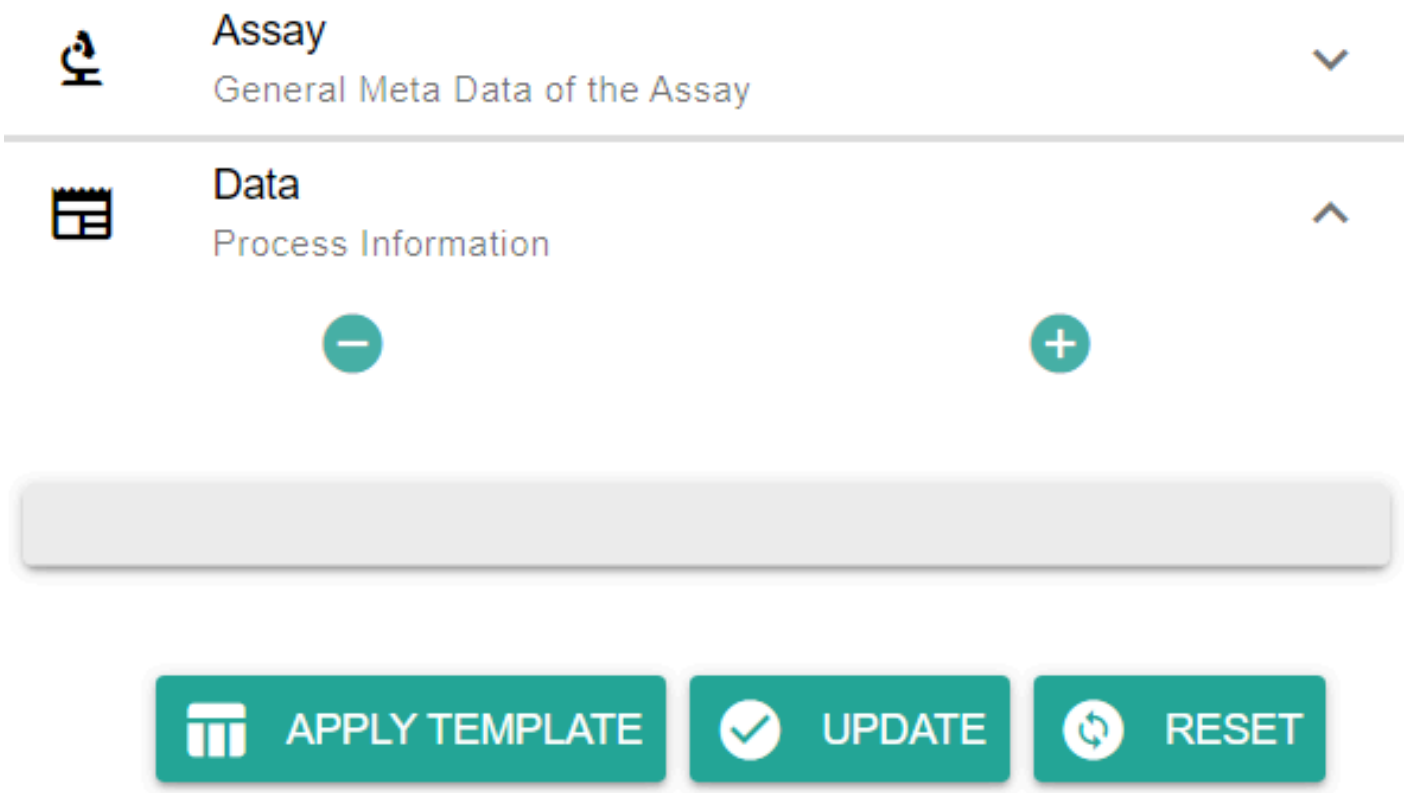
- Plant growth → sampling → extraction
- Measured data files → statistical analysis → result files

> 💡 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



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