



CEPLAS

Cluster of Excellence on Plant Sciences

The ARC Club

December 7th, 2023

Dominik Brilhaus, CEPLAS Data Science



House-keeping

Pad: https://pad.hhu.de/lhxZ_1GCRAa9ROHmj0zuYQ

- take notes
- ask questions
- copy / paste links, etc.

Get-to-know

- Lab
- CEPLAS / TRR / MibiNet / ?
- My favorite lab assay
- Used code / programming language before
- Have an ORCID
- My motivation / expectation

Materials

Slides will be shared via DataPLANT knowledge base

Tentative agenda

Day 1

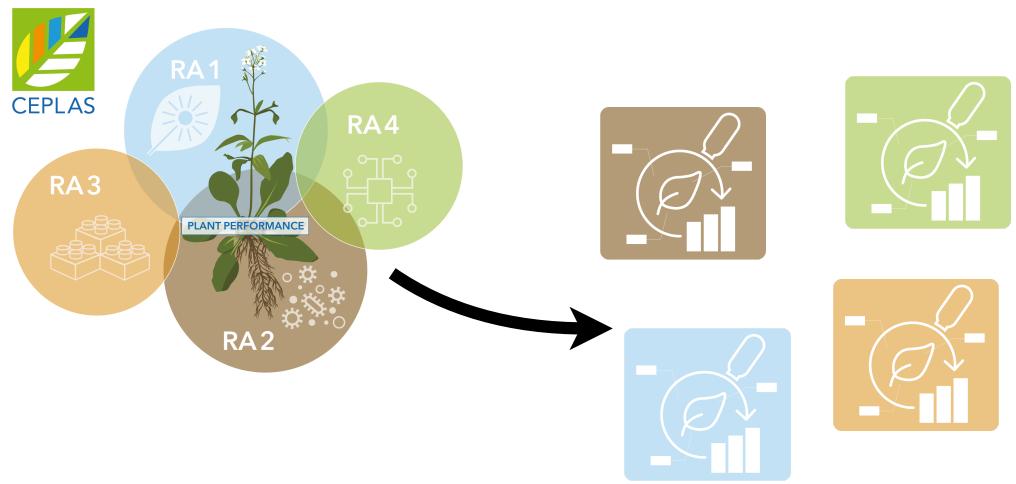
- Intro to ARC and demo
- ARCitect Hands-on
- Create your ARCs
- DataHUB Features

Day 2

- ARC recap session
- ISA and Metadata
- Swate Hands-on
- Annotate data in your ARC

The ARC Club – Goals

- Move existing datasets into ARCs
- Share them via the DataHUB
- First few steps into ARCs
- You or collaborators can pick them up from there



Perfect is the enemy of good

- There is no perfect ARC
- There is no complete ARC
- The only bad ARCs are those that don't exist yet.

🚀 Let's get started, the rest is easy 🚀

Let's draw a typical lab workflow from your lab



Resources



DataPLANT (nfdi4plants)

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

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

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

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

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



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

Contributors

Slides presented here include contributions by

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



CEPLAS

Cluster of Excellence on Plant Sciences

The ARC ecosystem

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



Data Stewardship between DataPLANT and the community

Community



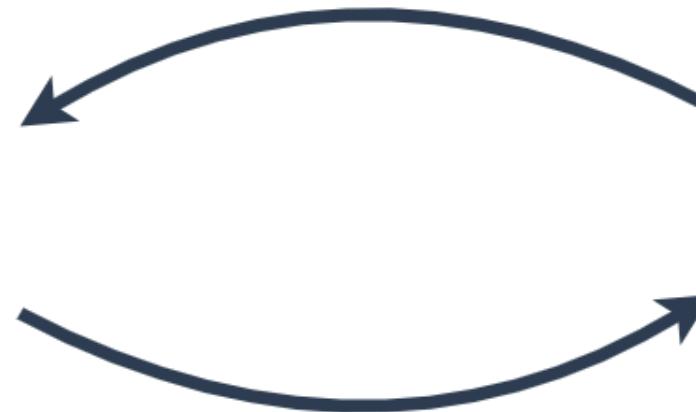
CEPLAS

Domain experts
User experience
Training

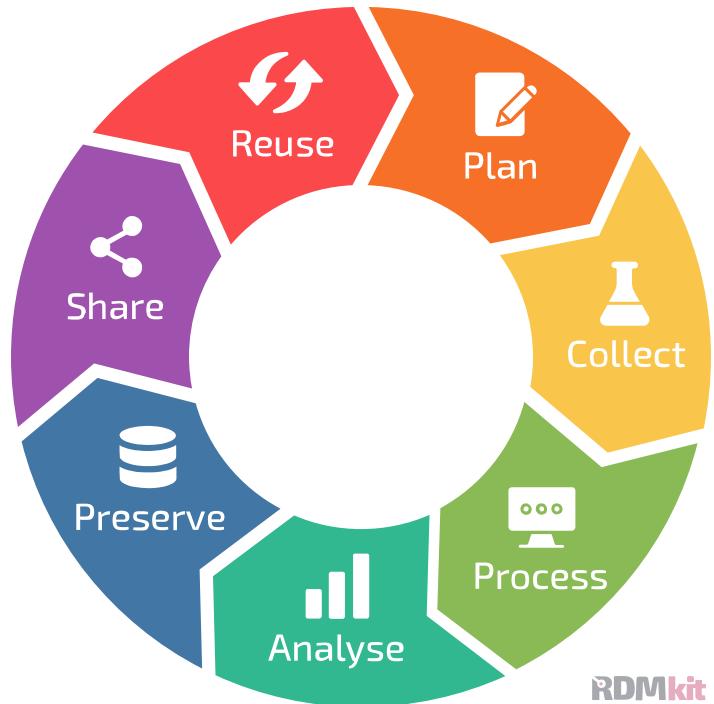
nfdi4plants



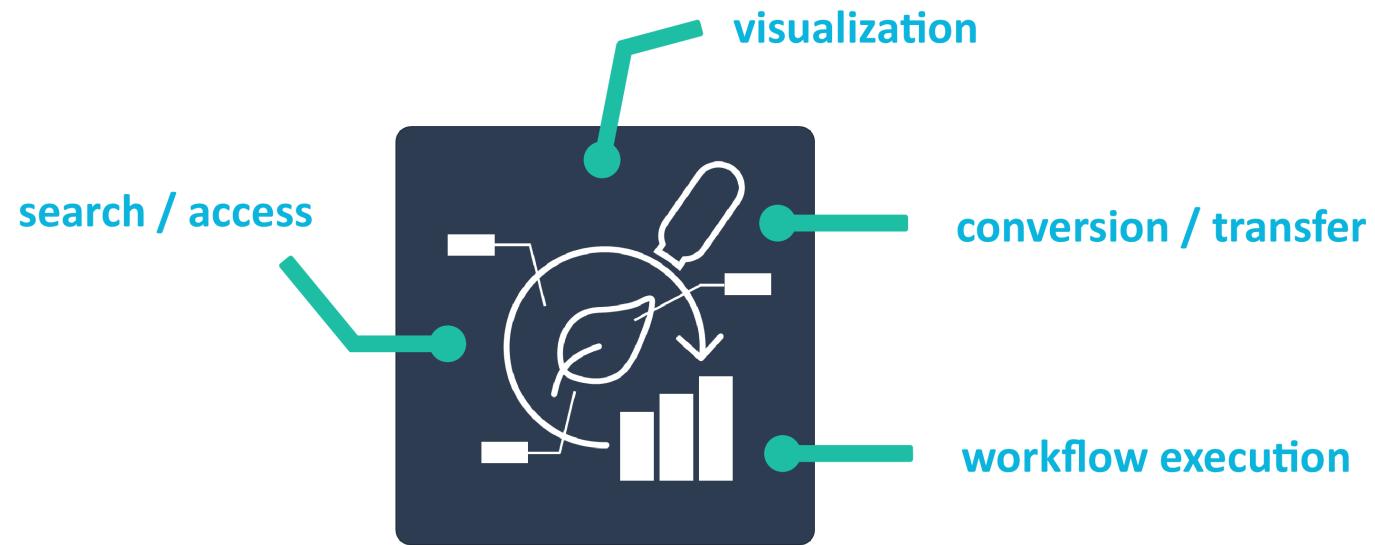
Service provider
Developers
Tech experts



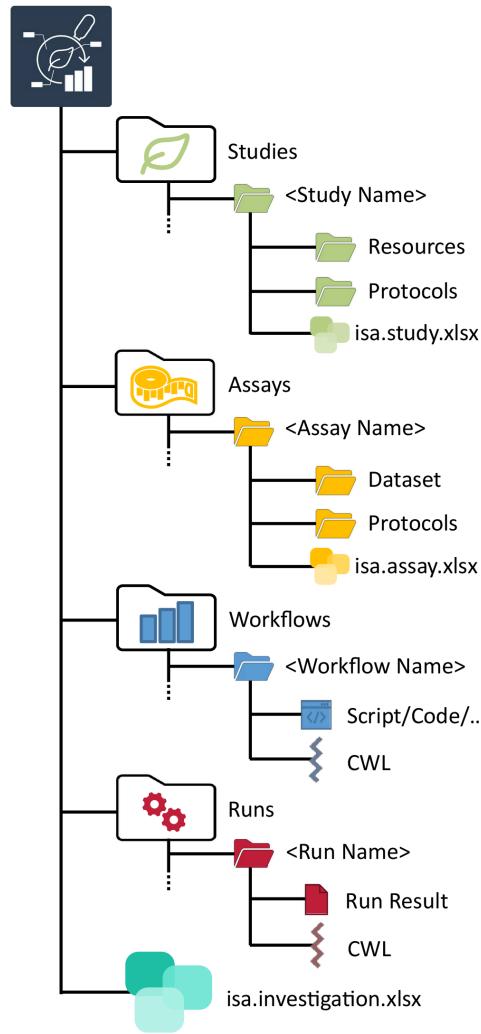
The research data life cycle



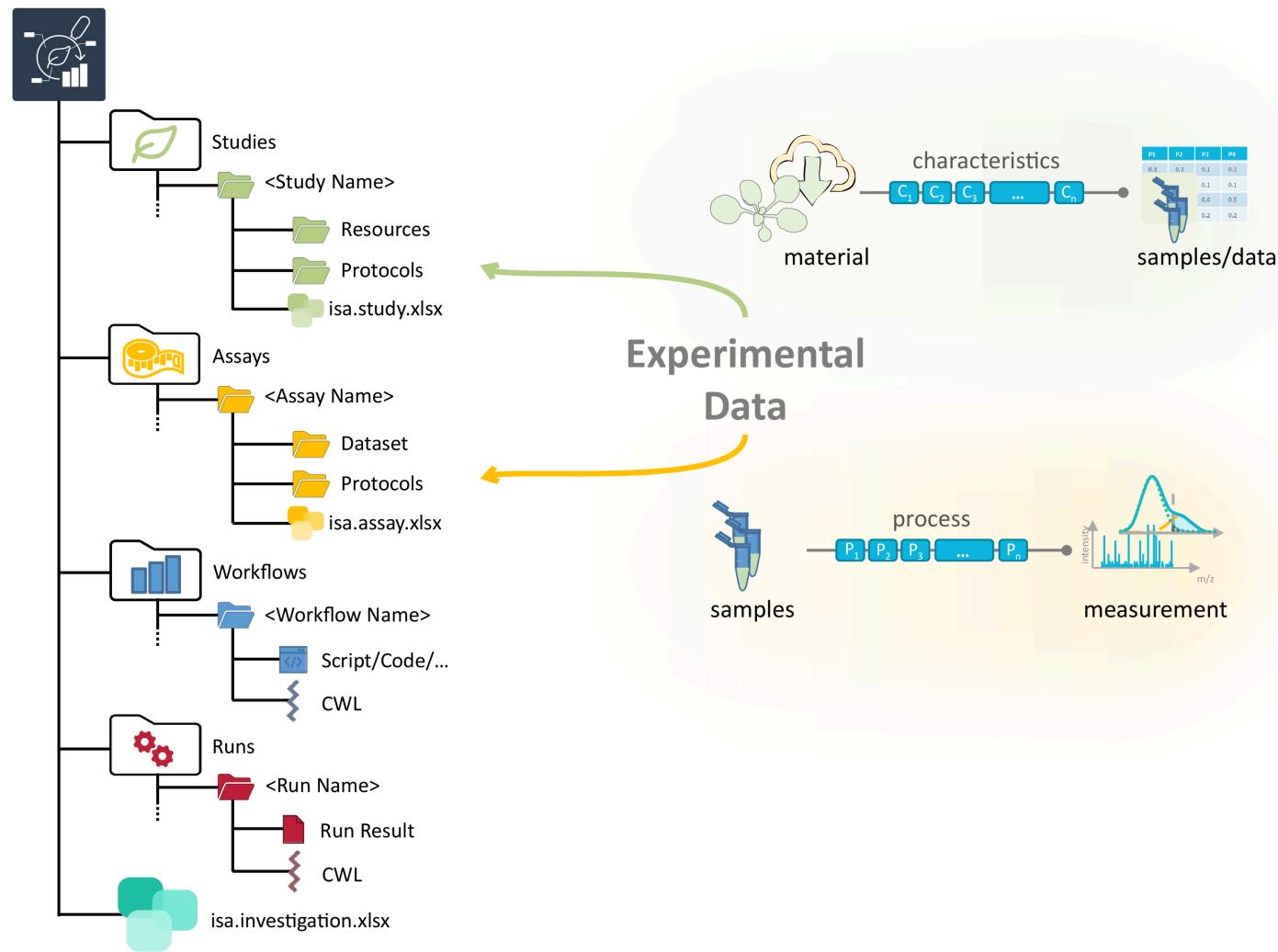
Annotated Research Context (ARC)



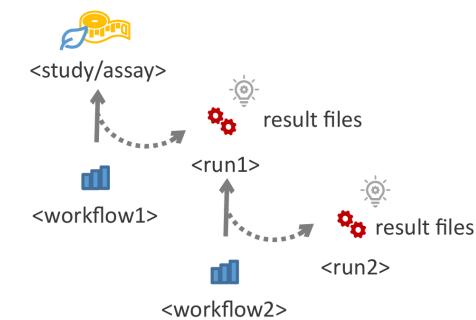
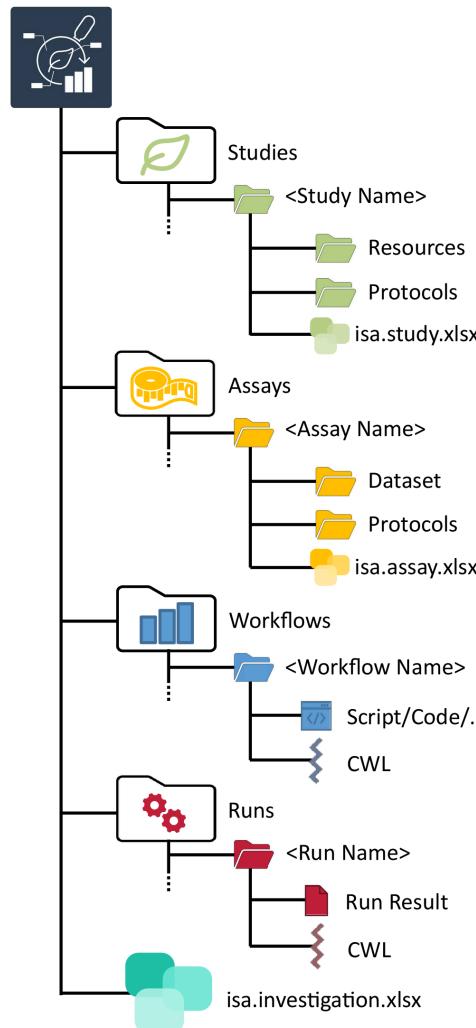
What does an ARC look like?



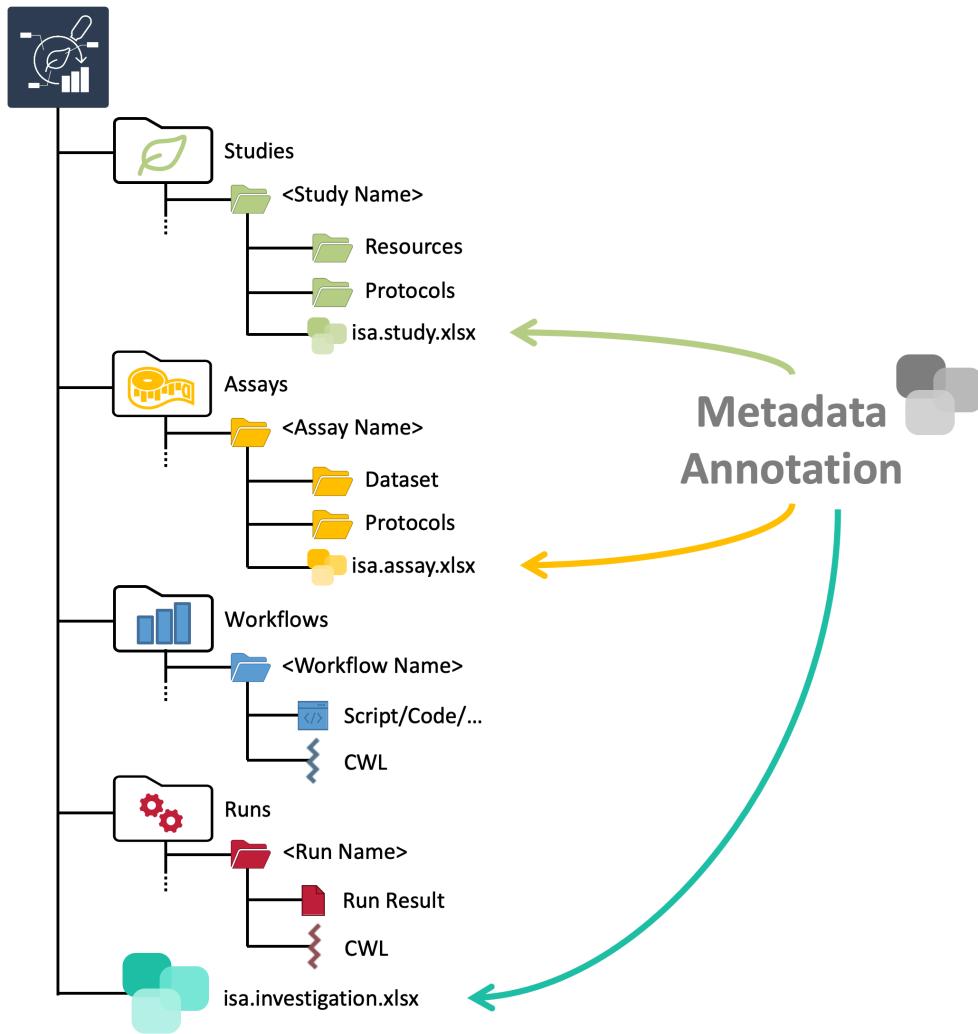
ARCs store experimental data



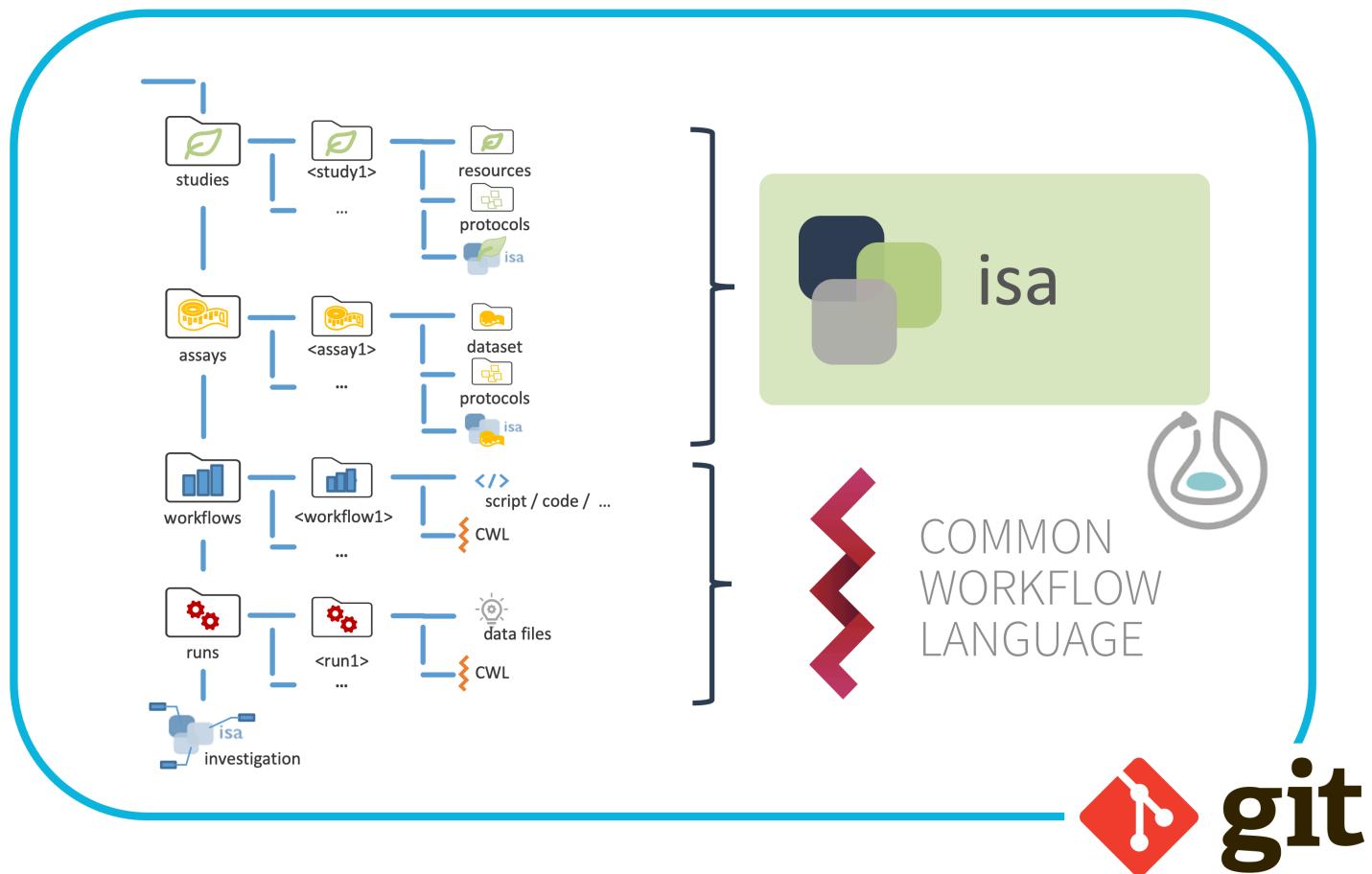
Computations can be run inside ARCs



ARCs come with comprehensive metadata



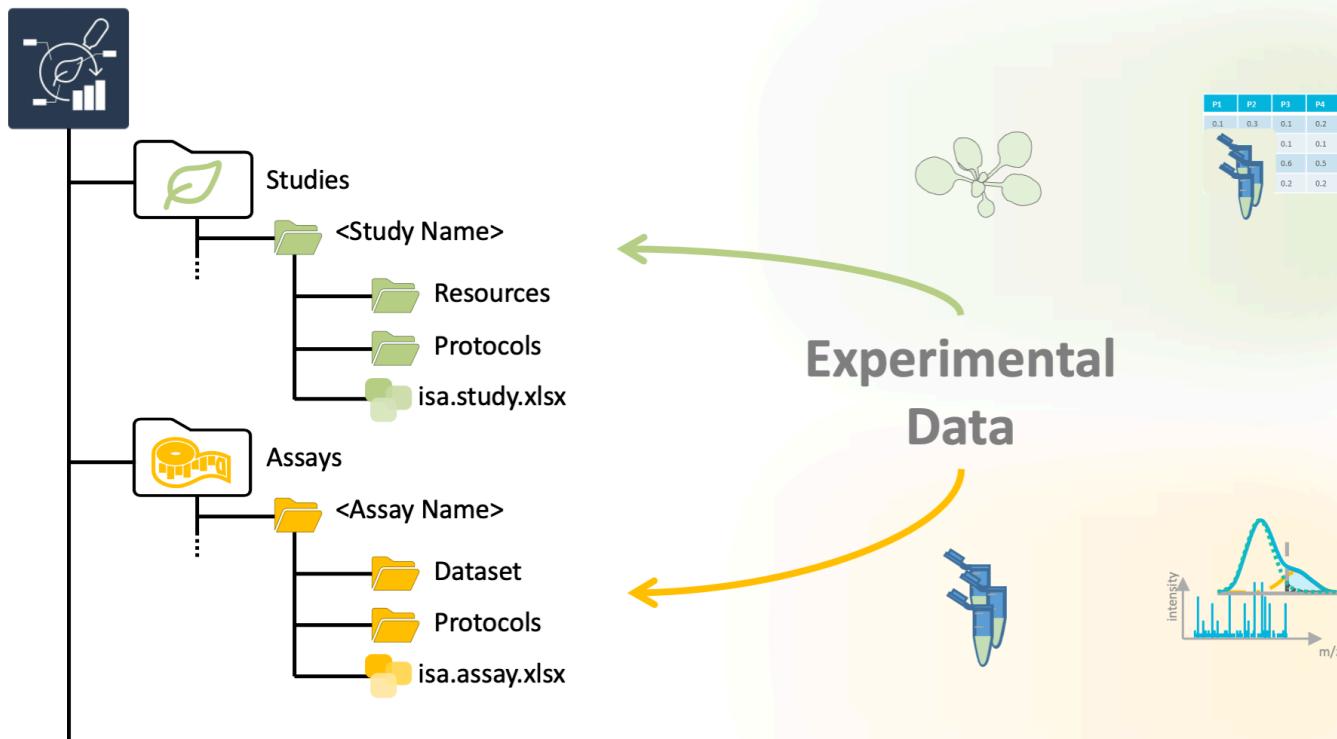
ARC builds on standards



<https://isa-tools.org/> | <https://www.commonwl.org/>
<https://www.researchobject.org/ro-crate/> | <https://git-scm.com>

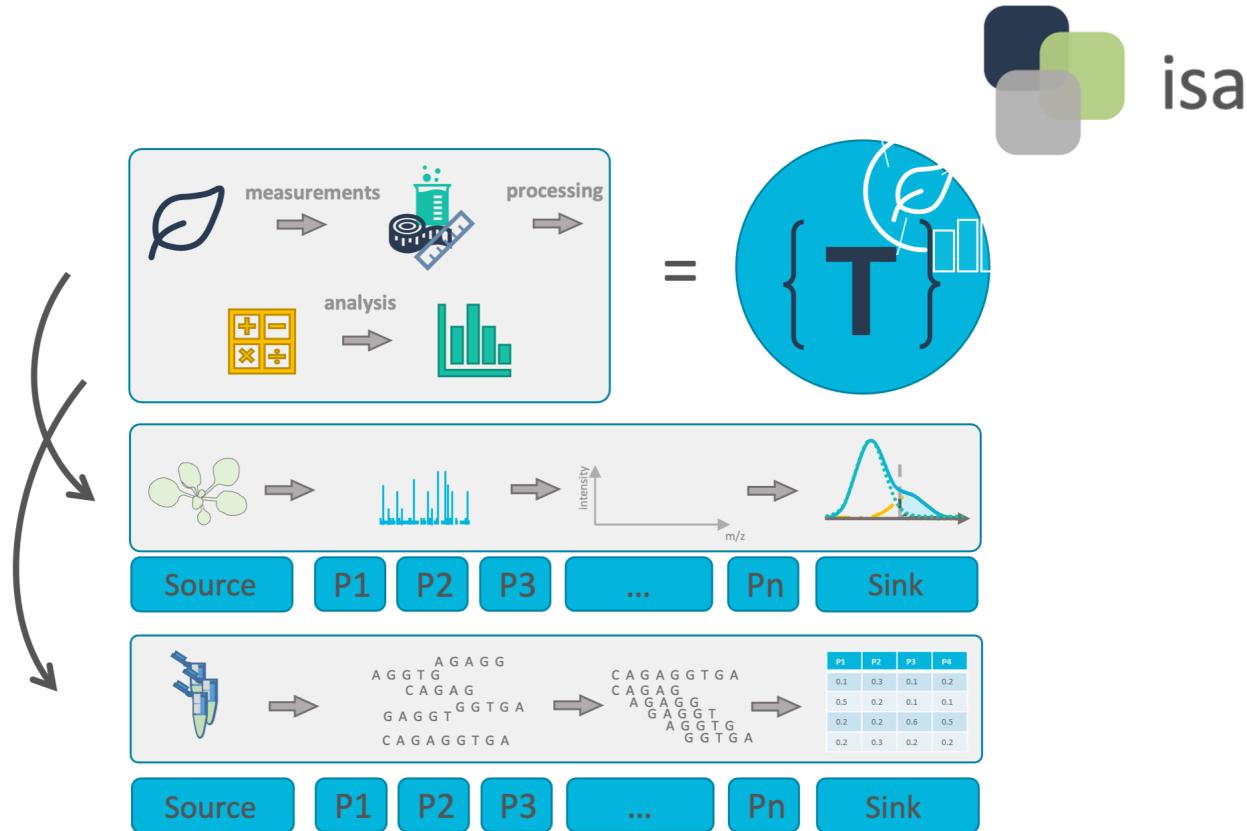


Collect



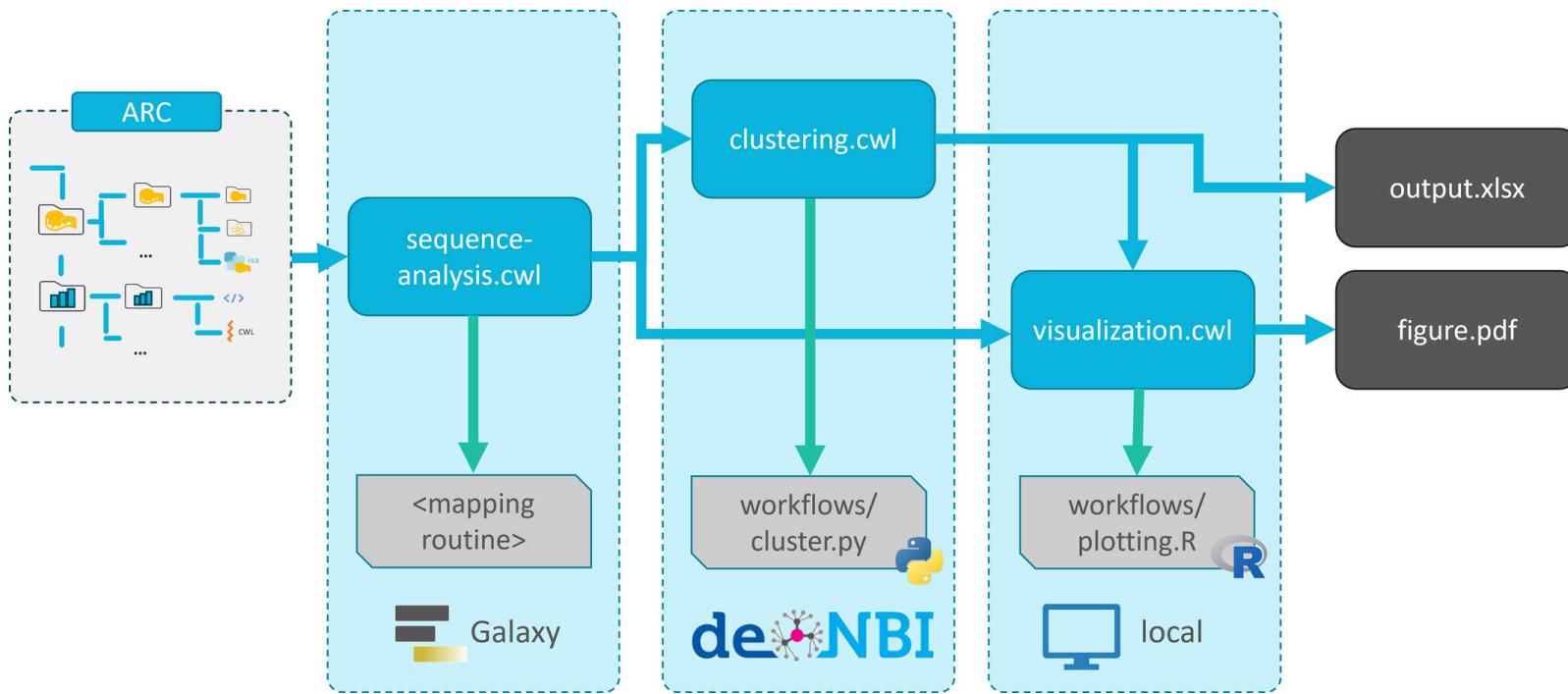


Process (e.g. annotate)



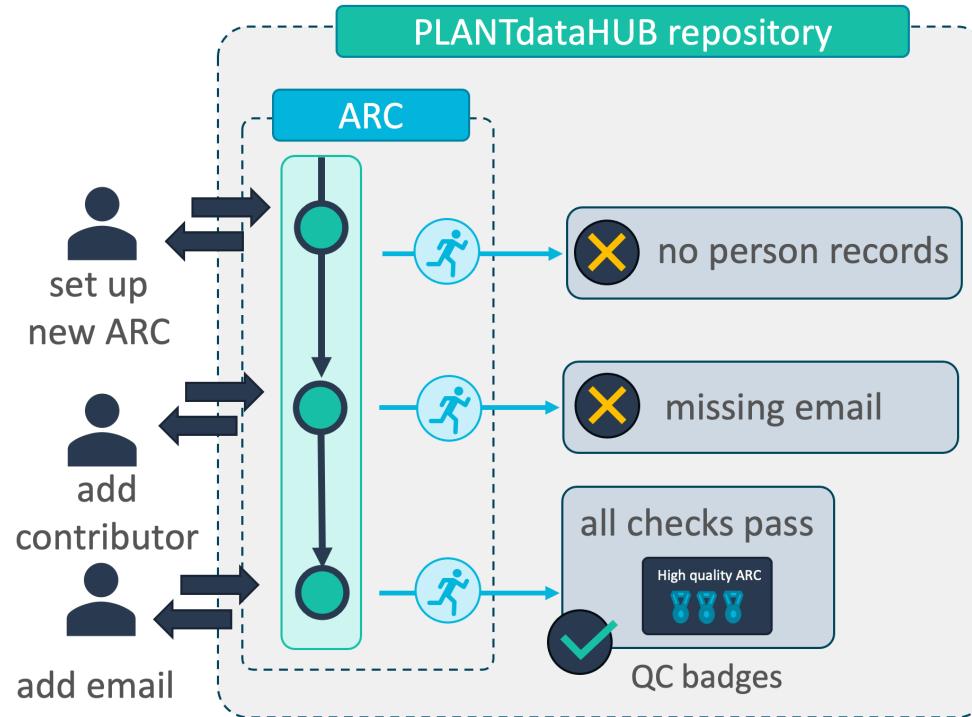


Analyse





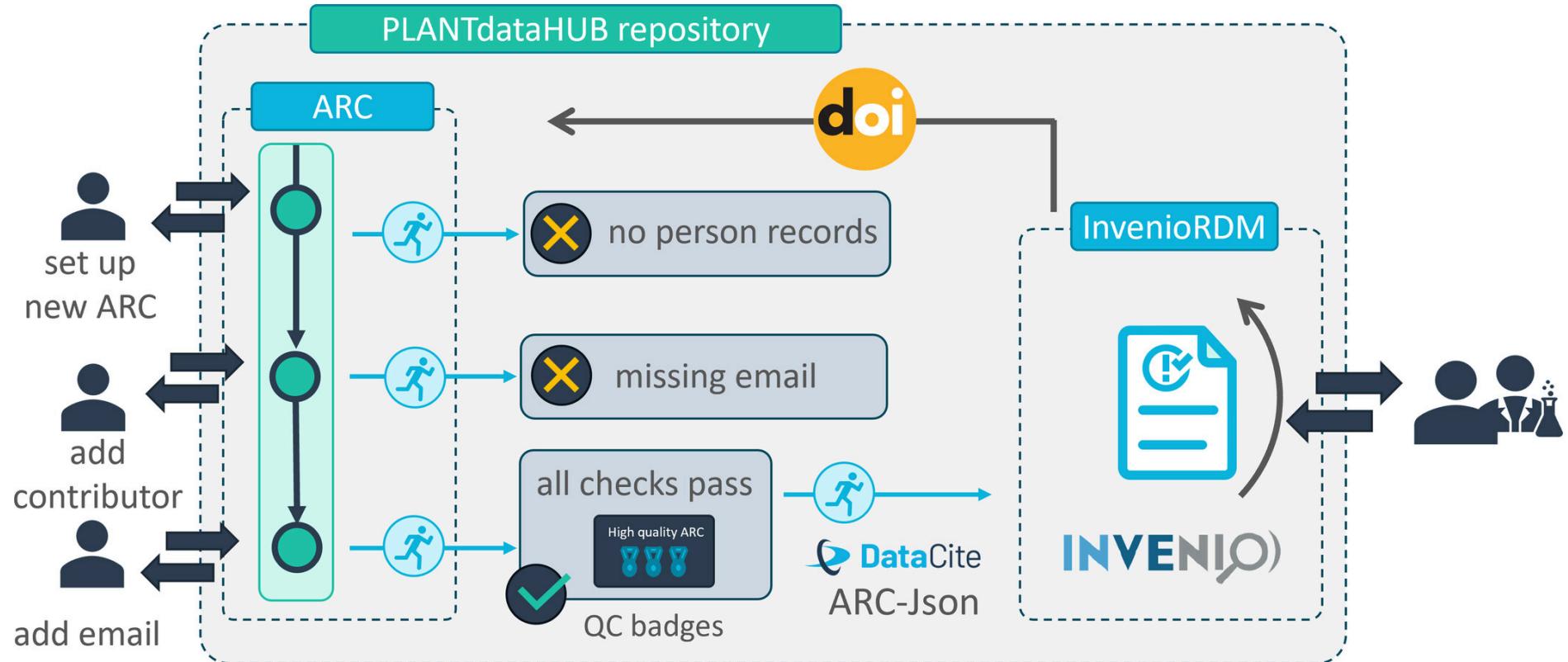
Preserve



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



Preserve and publish



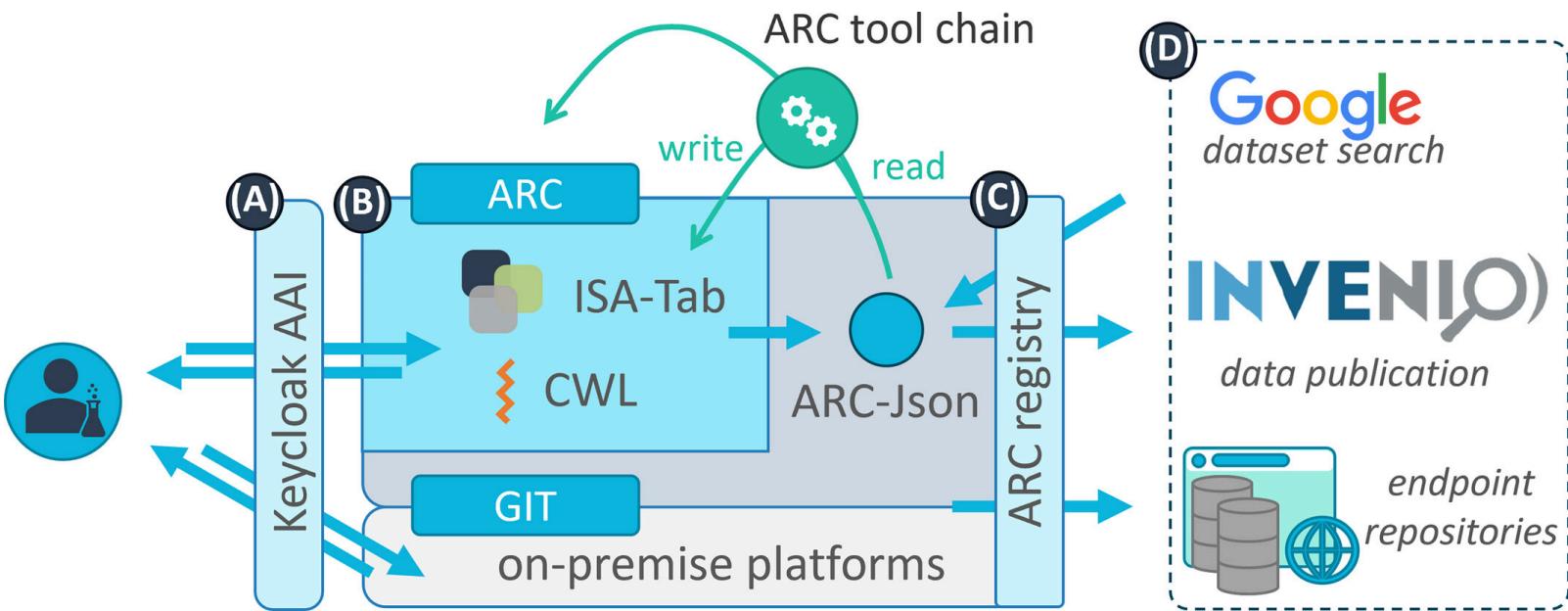


Share and collaborate

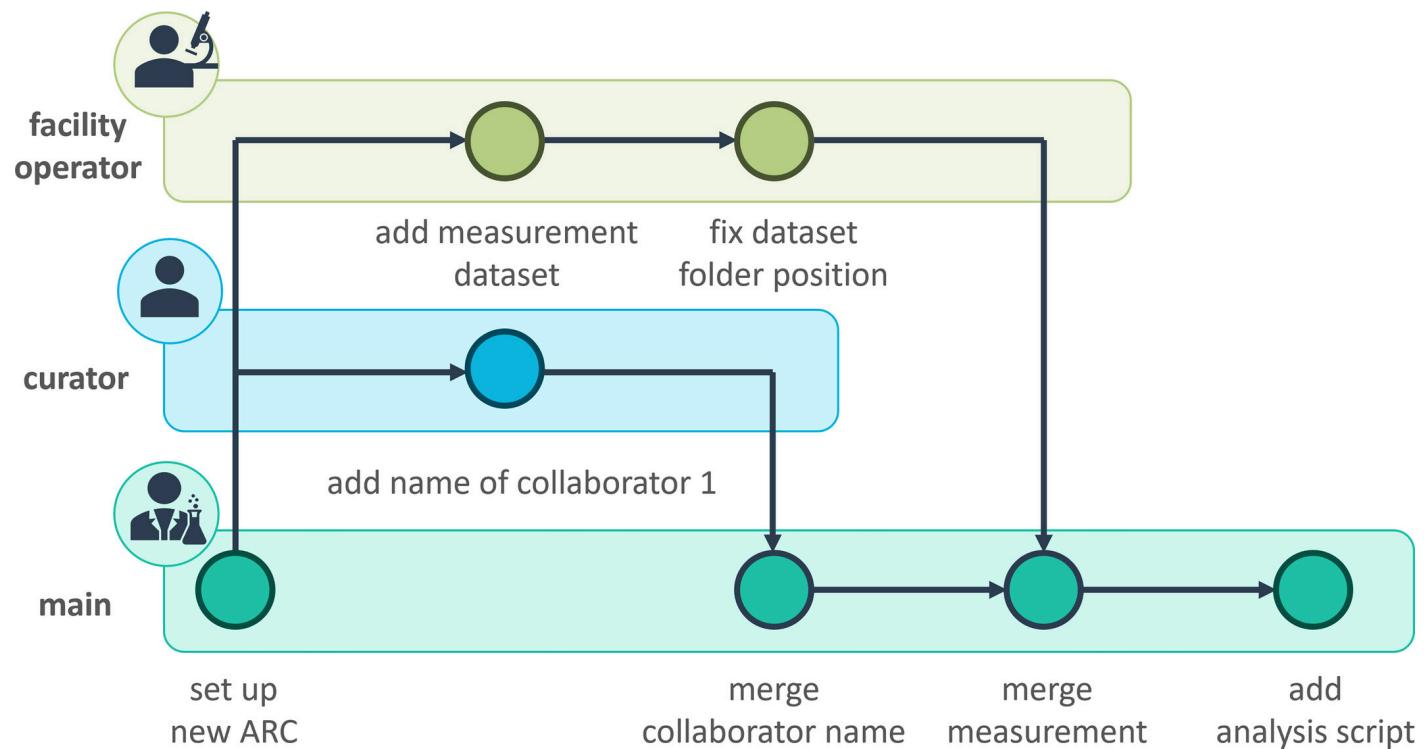




Reuse

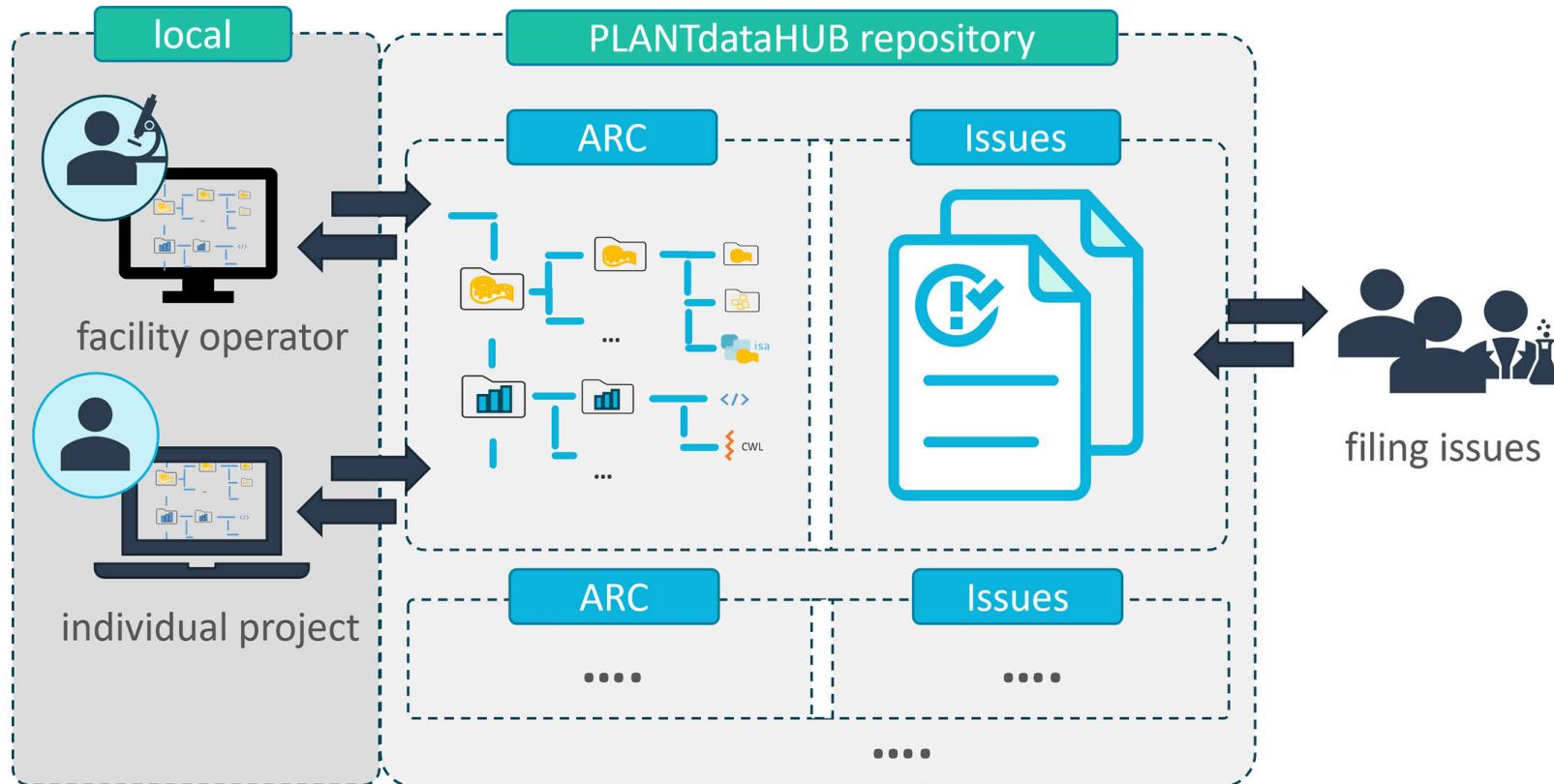


Mutable data life cycle





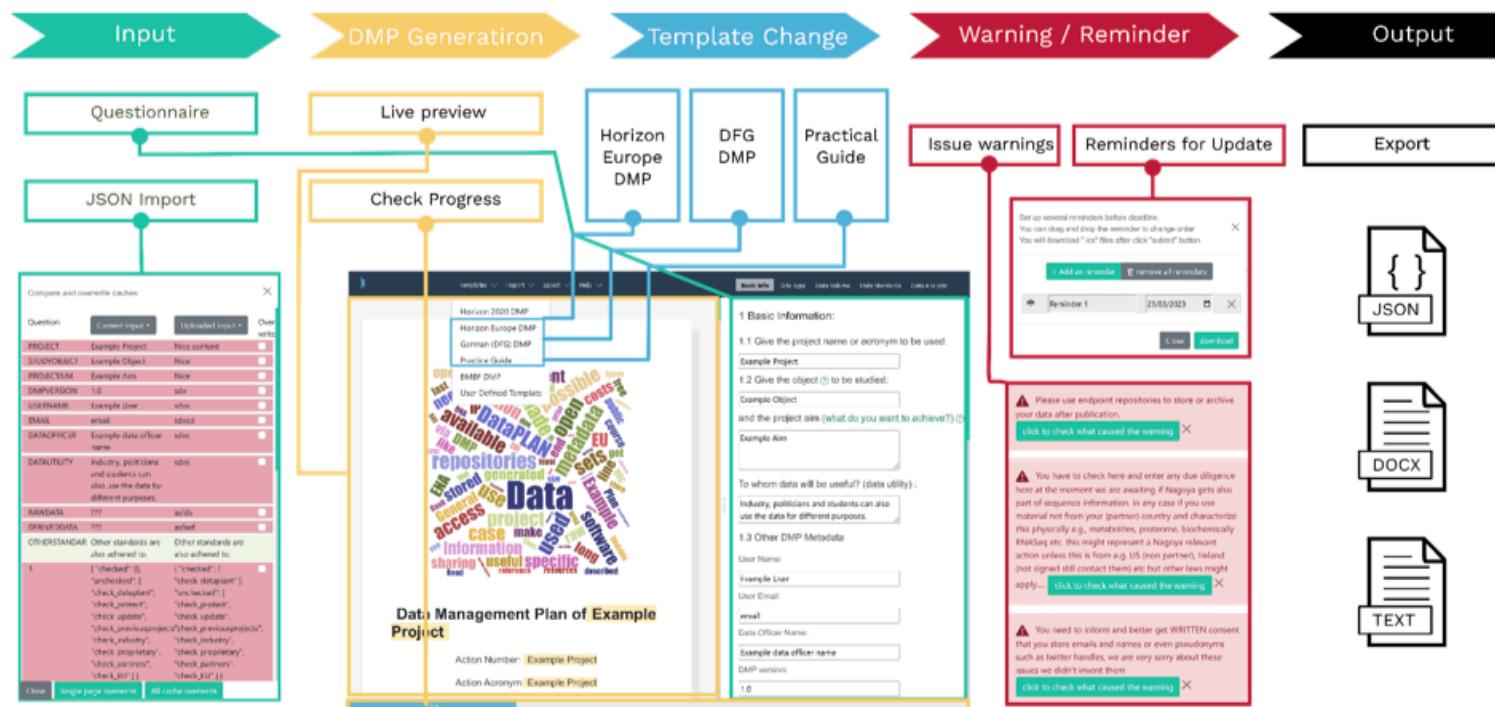
Plan – ARC scale





Plan – proposal scale

<https://dmpg.nfdi4plants.org>



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

The ARC ecosystem

Metadata Registry

Search ARCs by metadata

ARChigator

Publish & Archive

DataHUB

Share & Collaborate
Validate
(Hub to additional services)



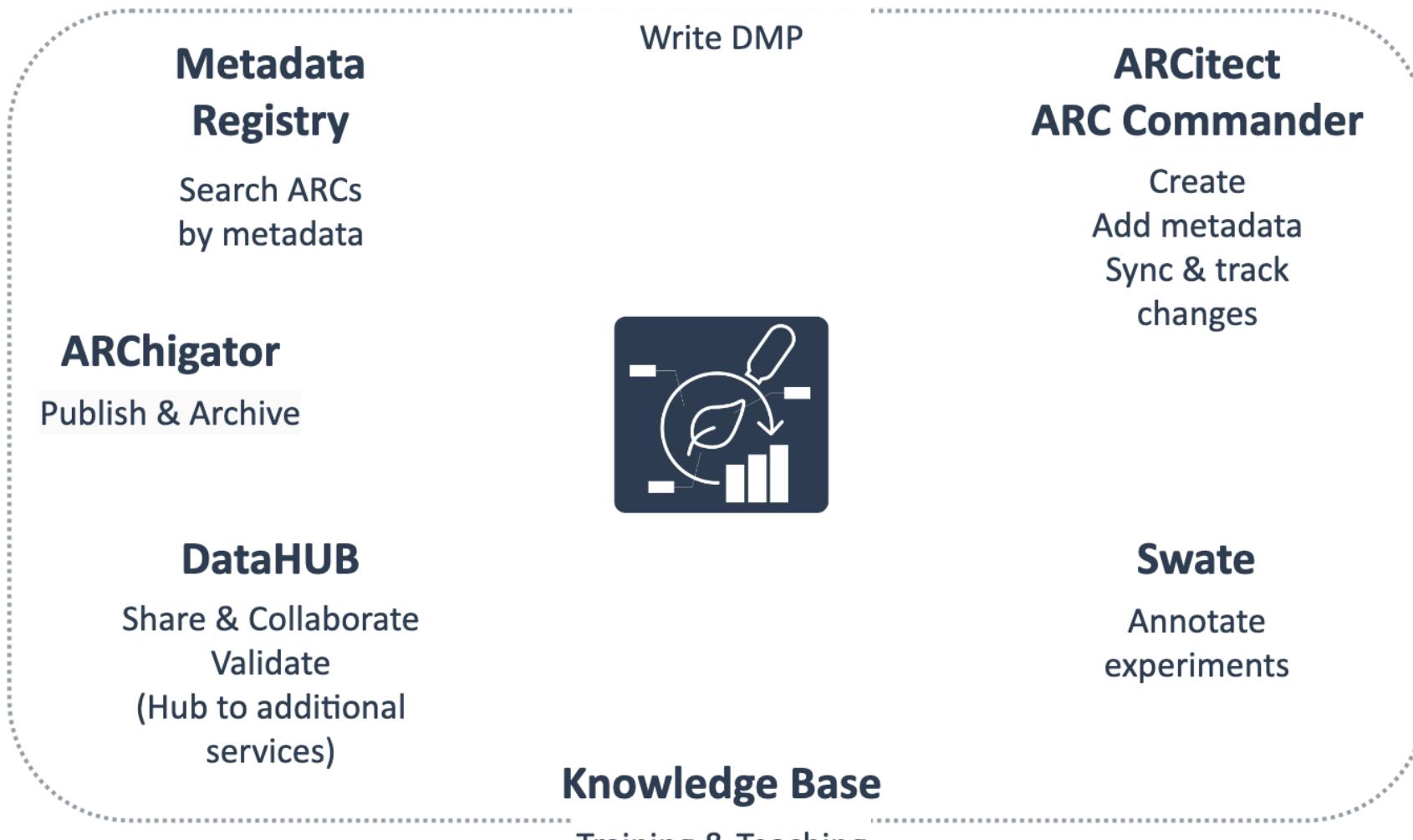
ARCitect ARC Commander

Create
Add metadata
Sync & track changes

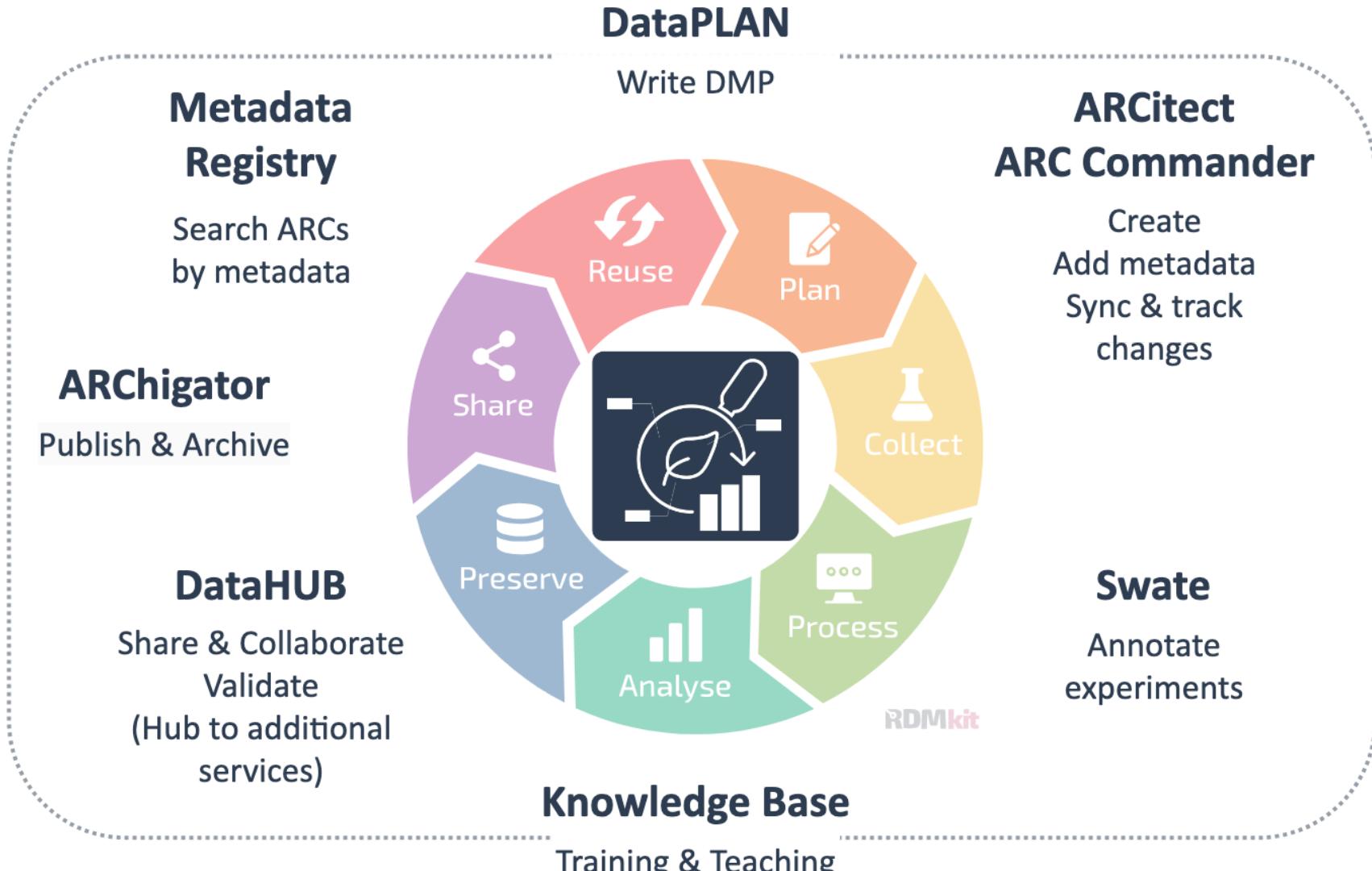
Swate

Annotate experiments

The ARC ecosystem



The ARC ecosystem



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>

ARCitect hands-on

ARCitect installation

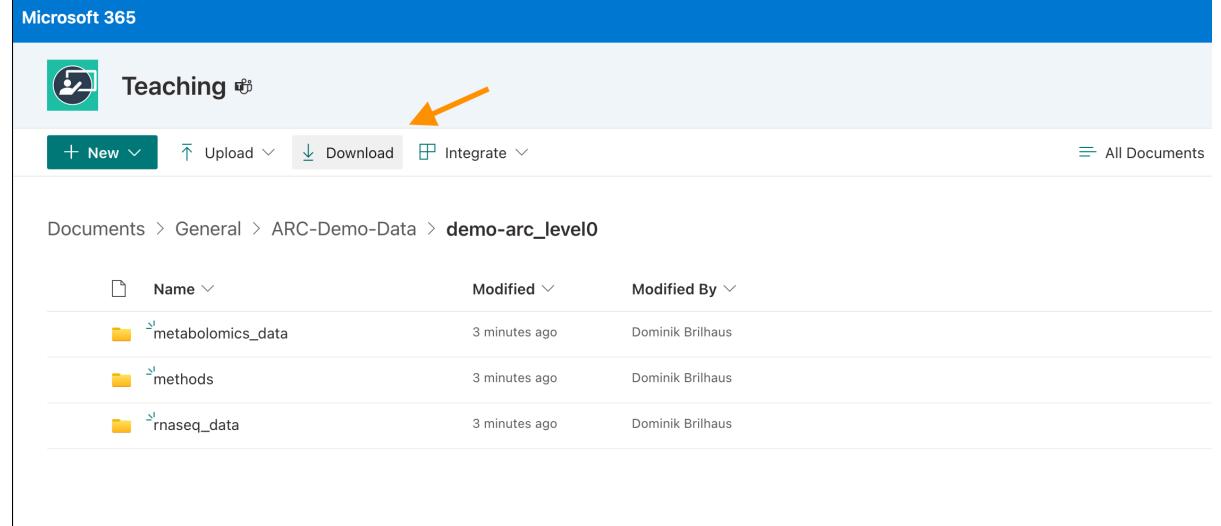
Please install version **v0.0.21** (or later) of the ARCitect:

<https://github.com/nfdi4plants/ARCitect/releases/tag/v0.0.21>

🔥 (released November 23rd, 2023) 🔥

Download the demo data

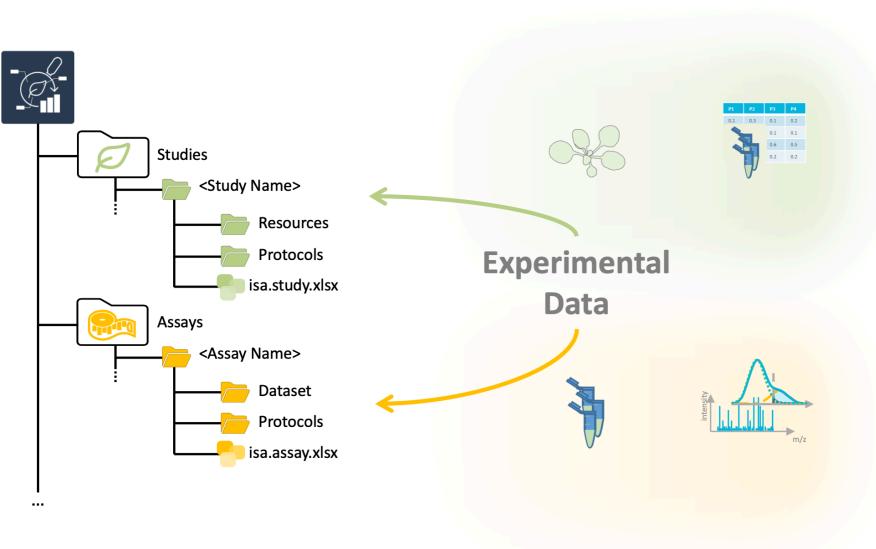
https://nfdi4plant.sharepoint.com/:f/s/Teaching/Eik7koJiMREgZ24kt07sIYBGxHmmZIS_Kzf7psk-5w-xg?e=u0sADd



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

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

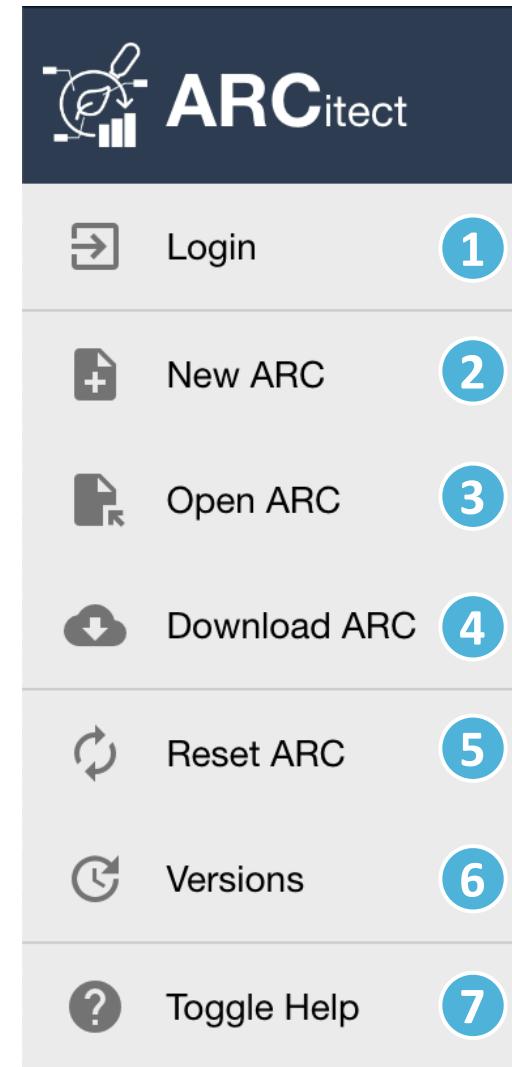
Sort Demo data in an ARC



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

Open ARCitect

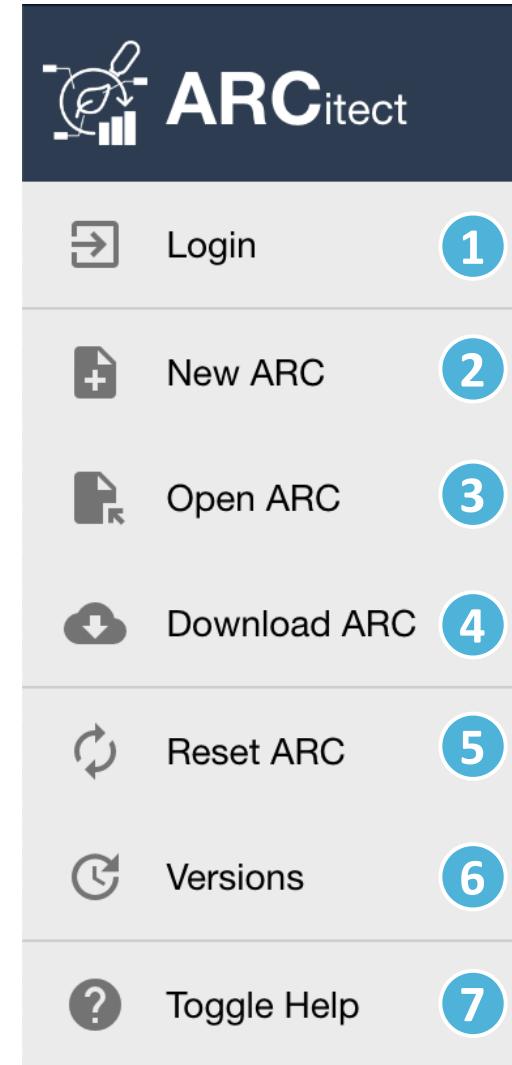
1. Login to DataHUB (1)



Initiate the ARC folder structure

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

TalinumPhotosynthesis



Your ARC's name

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

Add a description to your investigation

The image shows two side-by-side screenshots. On the left, a file tree is displayed in a terminal or file browser. The path is C:/Users/Sabrina/Desktop/Workshops ARCs/TalinumPhotosynthesis. The folder 'TalinumPhotosynthesis' is expanded, showing subfolders: assays, runs, studies, workflows, and a file named '.gitignore'. The total size of the folder is 1.34 KB. On the right, a form for adding a description to an investigation is shown. The form has three fields: 'Identifier' containing 'TalinumPhotosynthesis', 'Title' containing 'Talinum Photosynthesis', and 'Description' containing the text 'This is a very interesting investigation about life and photosynthesis'.

C:/Users/Sabrina/Desktop/Workshops ARCs
/TalinumPhotosynthesis

TalinumPhotosynthesis

- assays
- runs
- studies
- workflows

.gitignore

1.34 KB

Identifier

TalinumPhotosynthesis

Title

Talinum Photosynthesis

Description

This is a very interesting investigation about life and photosynthesis

Add a contributor

Contacts

Your First Name Your Last Name

Your ORCID

6/10 ▾

First Name

Your First Name

Last Name

Your Last Name

Mid Initials

Your Mid Initials

ORCID

Your ORCID

Search

Affiliation

Your Affiliation

Address

Your Address

Email

yourEmailAdress@uni.de

Phone

Your Phone

Fax

Your Fax

Roles

1. Author

NCIT

NCIT:C42781

X

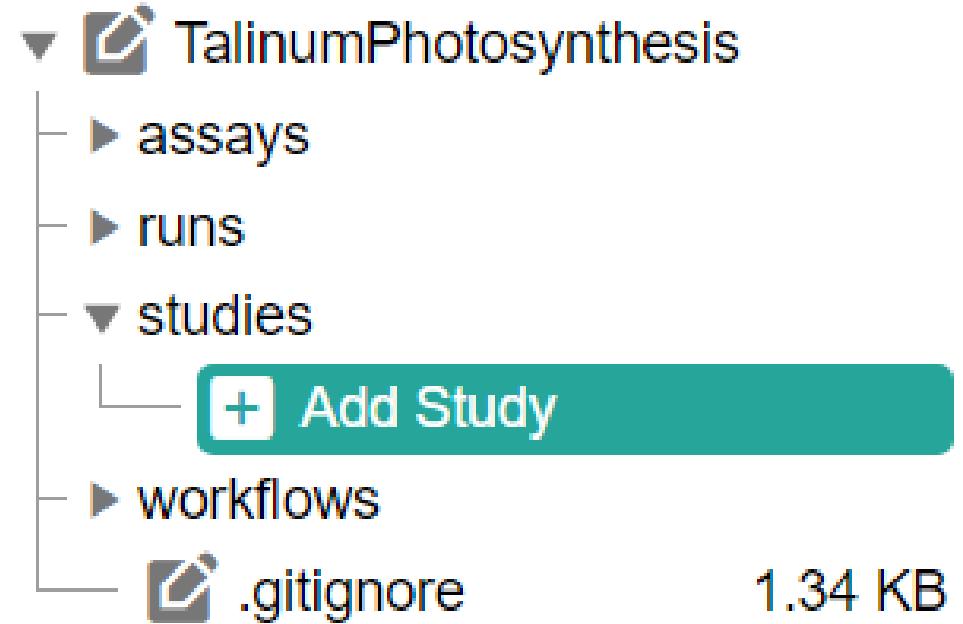
+

Delete

Add a study

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

Use **talinum_drought** as an identifier



Study panel

In the study panel you can add

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

Identifier
talinum_drought

Description

Contacts

Publications

Submission Date
tt.mm.jjjj --:--

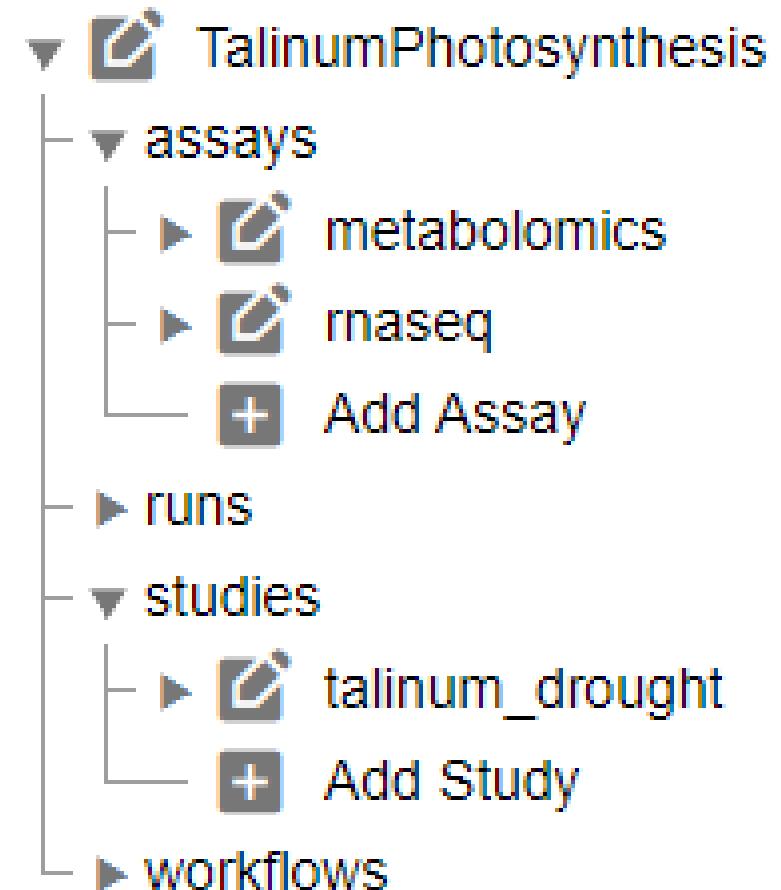
Public Release Date
tt.mm.jjjj --:--

Study Design Descriptors

Add an assay

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

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



Link your assay to a study

You can either

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

Link your assays to your
talinum_drought study

Add Assay

Add Assay
rnaseq

Study Identifiers
talinum_drought 

talinum_drought

 ADD ASSAY CANCEL

Add information about your assay

In the assay panel you can

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

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

Add protocols

You can either

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

Create or Import Protocol

Protocol Name



NEW PROTOCOL



IMPORT PROTOCOL

CANCEL

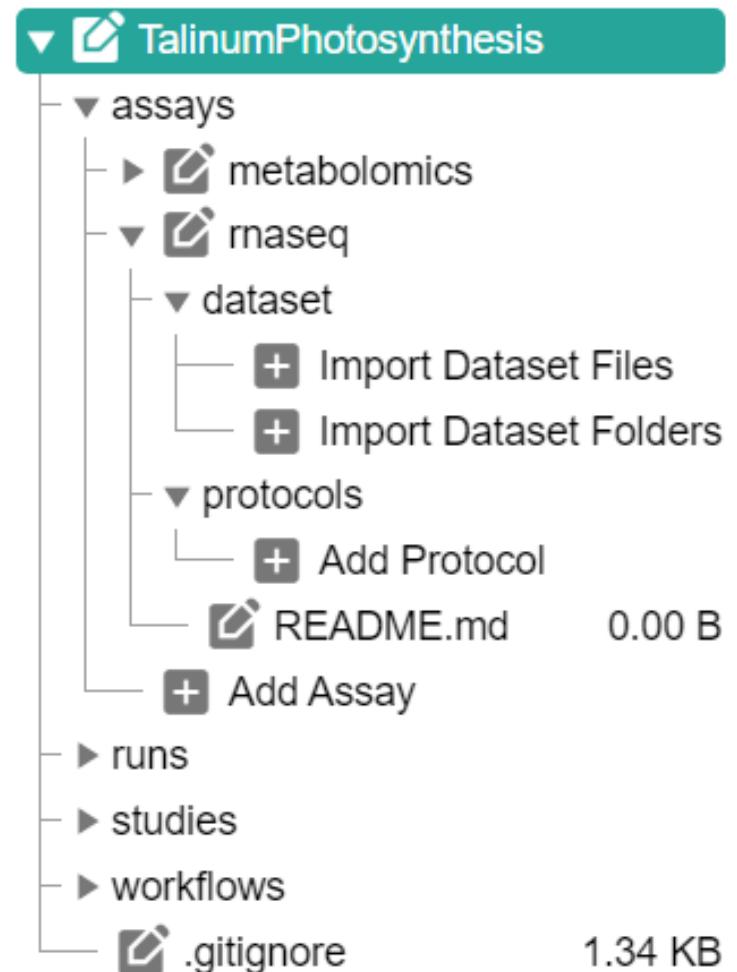
Add protocols and datasets

In the file tree you can

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

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

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

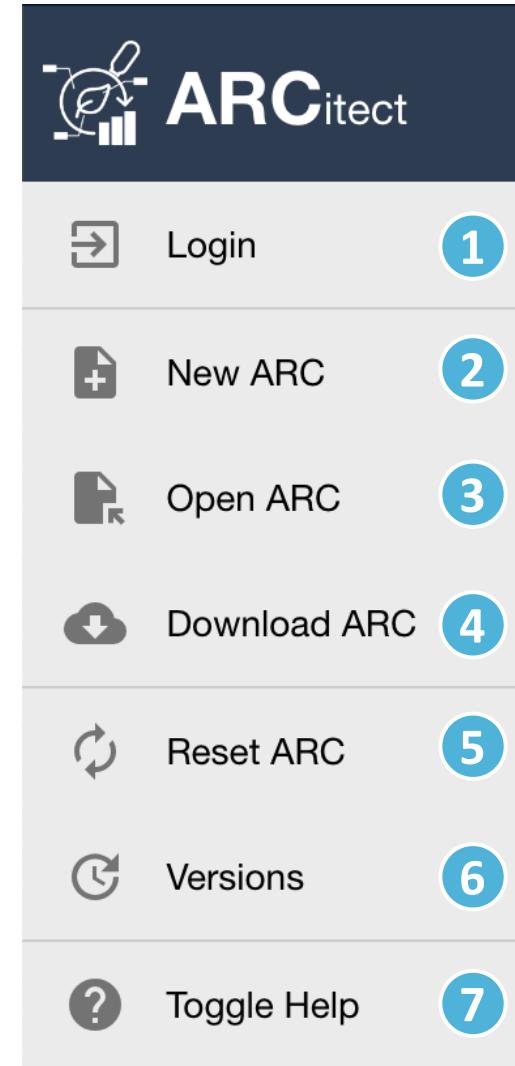


Sort Demo Data to your ARC

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

Upload your local ARC to the DataHUB

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



Versions

The versions panel allows you to

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

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

Update
Commit changes and upload ARC

Full Name
Demo User

eMail
demo@nfdi4plants.org

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

Commit Message

A short description of the made changes

Changes

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

REFRESH COMMIT UPLOAD DOWNLOAD

History
Inspect ARC history

Connection to the DataHUB

If you are logged in, the versions panel shows

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

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

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

Update
Commit changes and upload ARC

Full Name
Demo User

eMail
demo@nfdi4plants.org

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

Commit Message

A short description of the made changes

Changes

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

REFRESH COMMIT UPLOAD DOWNLOAD

History
Inspect ARC history

Check if your ARC is successfully uploaded

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

Your ARC is ready

 Initiated an ARC

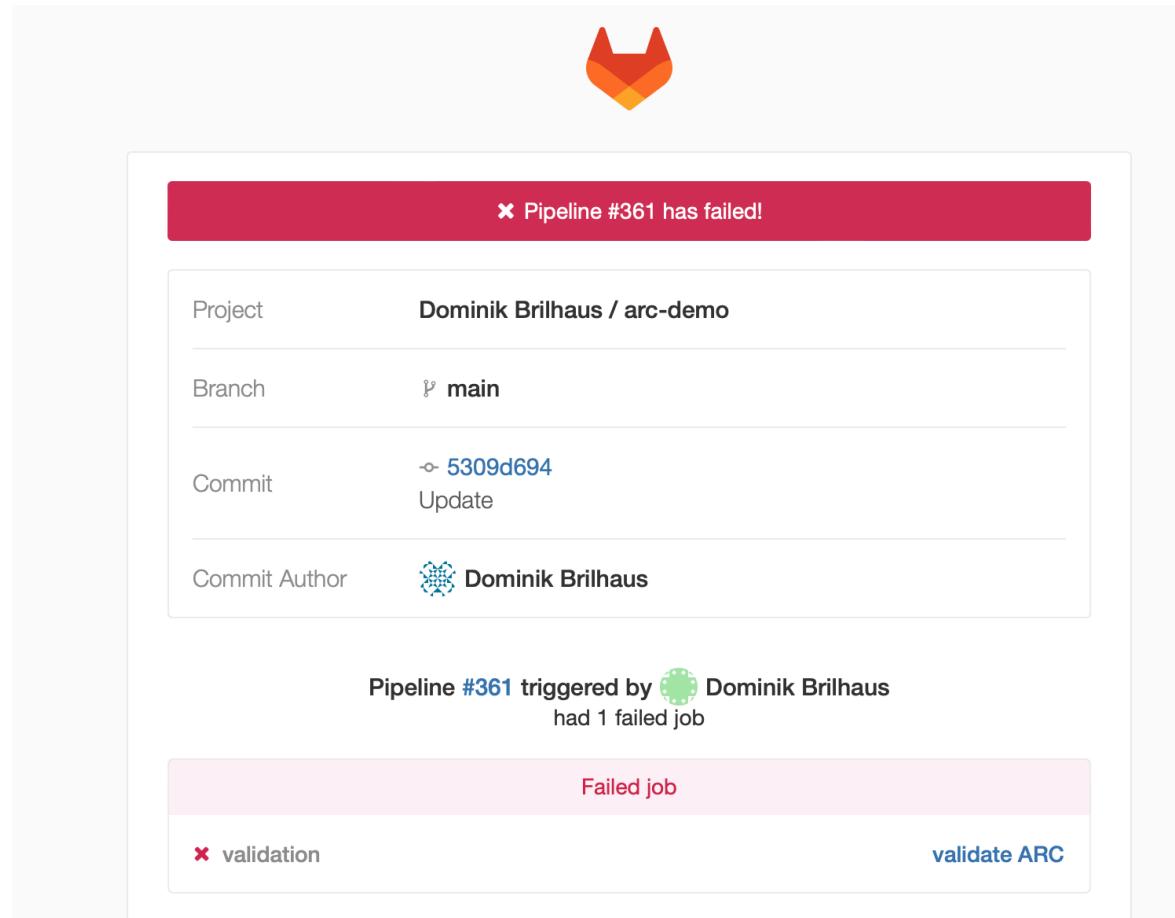
 Structured and ...

 ... annotated experimental data

 Shared with collaborators



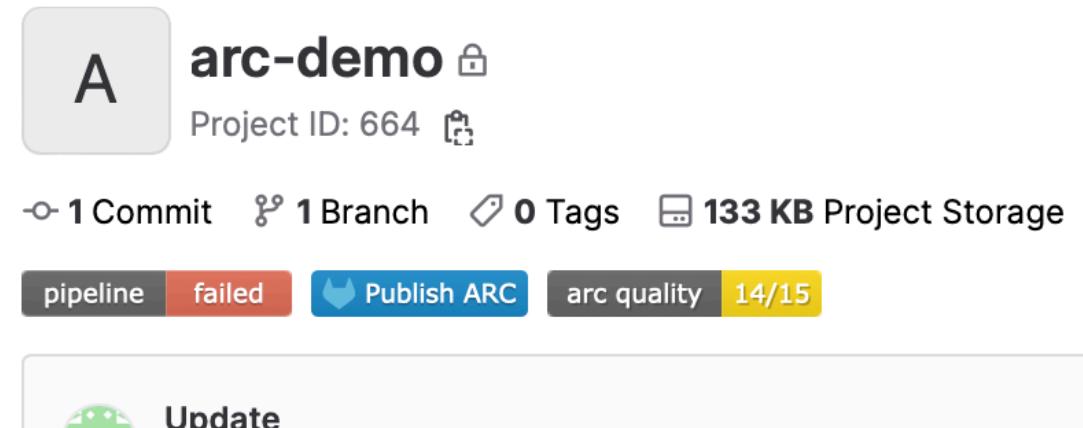
Received two emails from "GitLab" about a failed pipeline?



Pipeline Failed

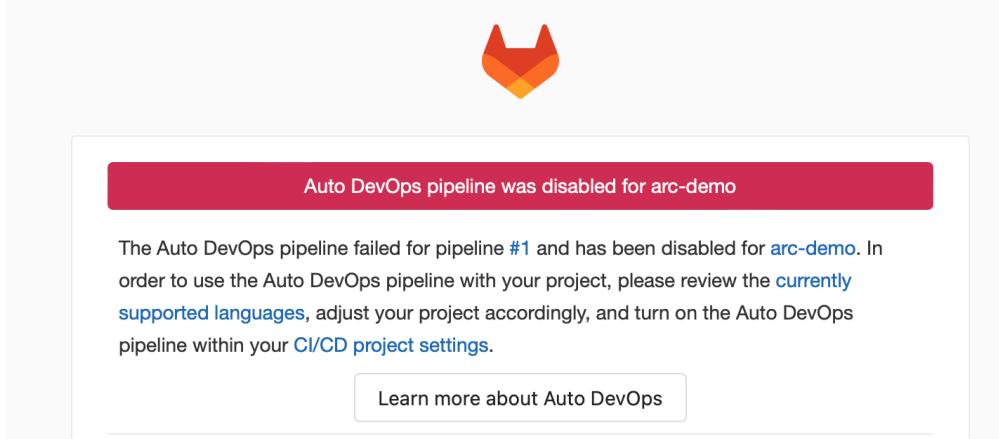
- a "continuous quality control" (CQC) pipeline validates your ARC
- This fails if one of the following metadata items is missing:

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



Pipeline Failed

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



The screenshot shows a failure message for an Auto DevOps pipeline. At the top right is a small orange and yellow logo. Below it, a red banner displays the text "Auto DevOps pipeline was disabled for arc-demo". The main message area contains text explaining the failure: "The Auto DevOps pipeline failed for pipeline #1 and has been disabled for [arc-demo](#). In order to use the Auto DevOps pipeline with your project, please review the [currently supported languages](#), adjust your project accordingly, and turn on the Auto DevOps pipeline within your [CI/CD project settings](#)." At the bottom right of the message area is a button labeled "Learn more about Auto DevOps".

Reactivate the CQC pipeline

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

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

The screenshot shows the GitLab CI/CD settings interface. On the left, there is a sidebar with various project management and development tools listed: Security & Compliance, Deployments, Packages and registries, Infrastructure, Monitor, Analytics, Wiki, Snippets, Settings (selected), General, Integrations, Webhooks, Access Tokens, Repository, Merge requests, CI/CD (selected), Packages and registries, Monitor, and Usage Quotas. The main content area is titled "Auto DevOps" and contains the following sections:

- Auto DevOps**: A brief description of what Auto DevOps does: "Automate building, testing, and deploying your applications based on your continuous integration and delivery configuration." It also includes a link to "How do I get started?"
- Default to Auto DevOps pipeline**: A checkbox labeled "instance enabled" is checked. Below it, a note states: "The Auto DevOps pipeline runs if no alternative CI configuration file is found. [Learn more.](#)"
- Add a Kubernetes cluster integration**: A note suggesting to "Add a Kubernetes cluster integration with a domain, or create an AUTO_DEVOPS_PLATFORM_TARGET CI variable."
- Deployment strategy**: Three radio button options are available:
 - Continuous deployment to production [?](#)
 - Continuous deployment to production using timed incremental rollout [?](#)
 - Automatic deployment to staging, manual deployment to production [?](#)
- Save changes**: A blue button at the bottom of the section.
- Runners**: A section describing runners and their role in executing CI/CD jobs. It includes a link to "What is GitLab Runner?" and an "Expand" button.
- Artifacts**: A section describing artifacts as archives saved by a job. It includes an "Expand" button.

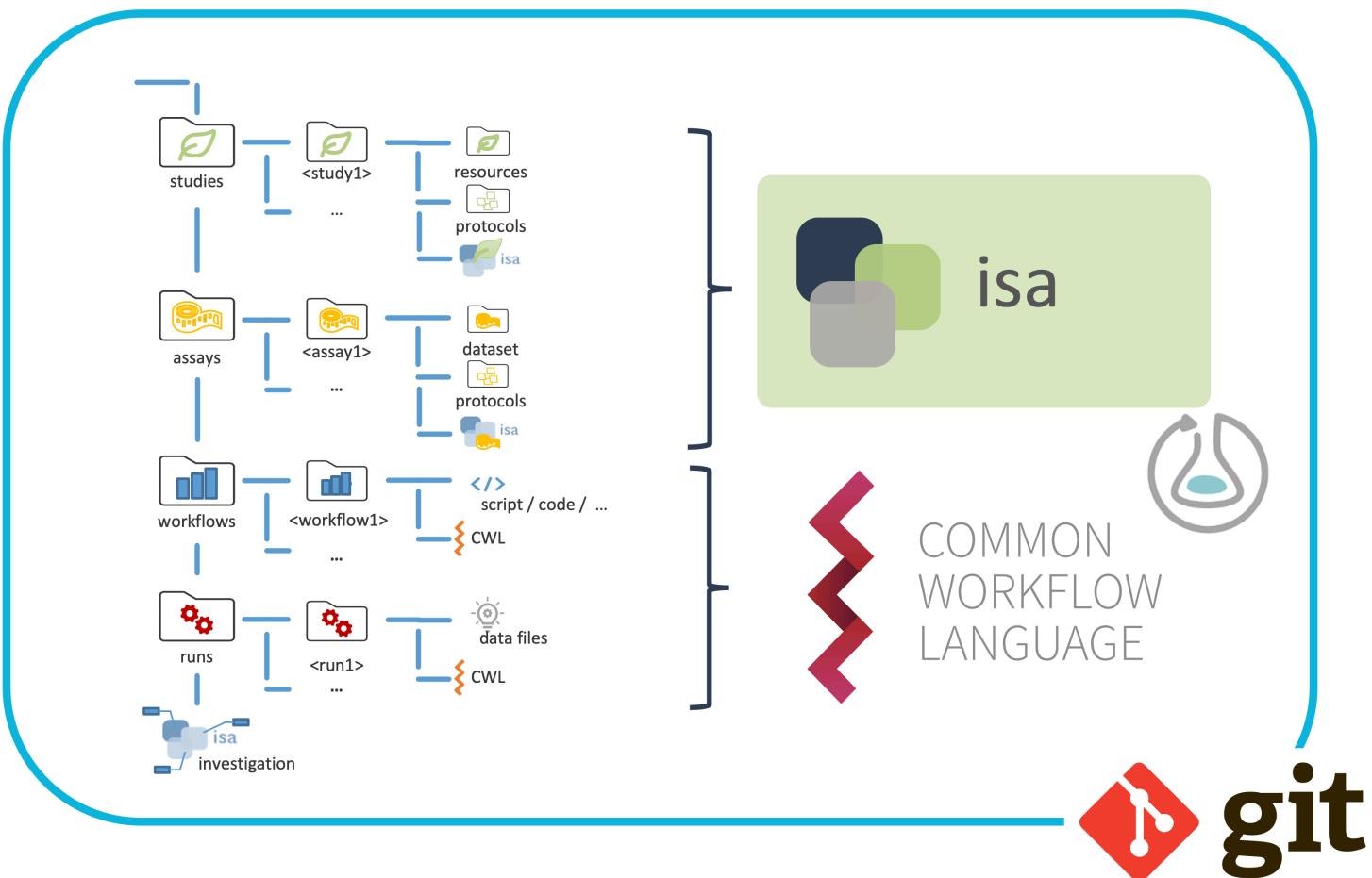
Contributors

Slides presented here include contributions by

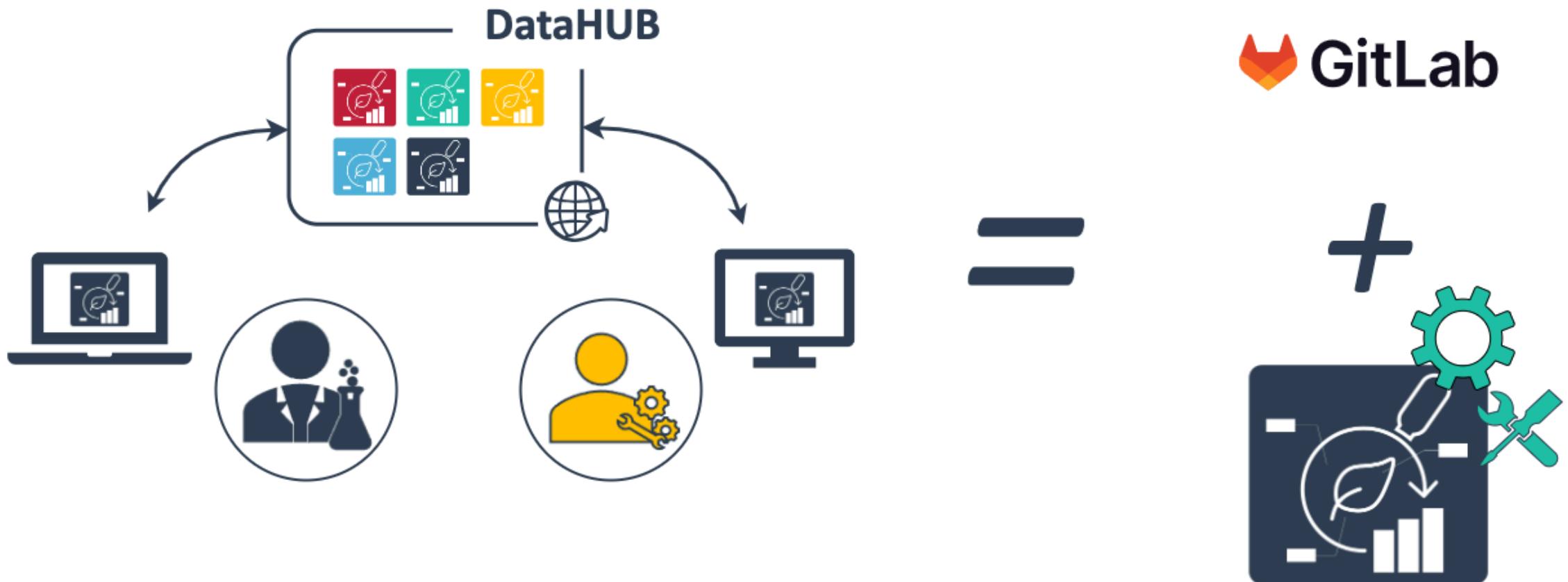
- name: Dominik Brilhaus
github: <https://github.com/brilator>
orcid: <https://orcid.org/0000-0001-9021-3197>
- name: Cristina Martins Rodrigues
github: <https://github.com/CMR248>
orcid: <https://orcid.org/0000-0002-4849-1537>
- name: Sabrina Zander
github: <https://github.com/SabrinaZander>
orcid: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

DataPLANT DataHUB

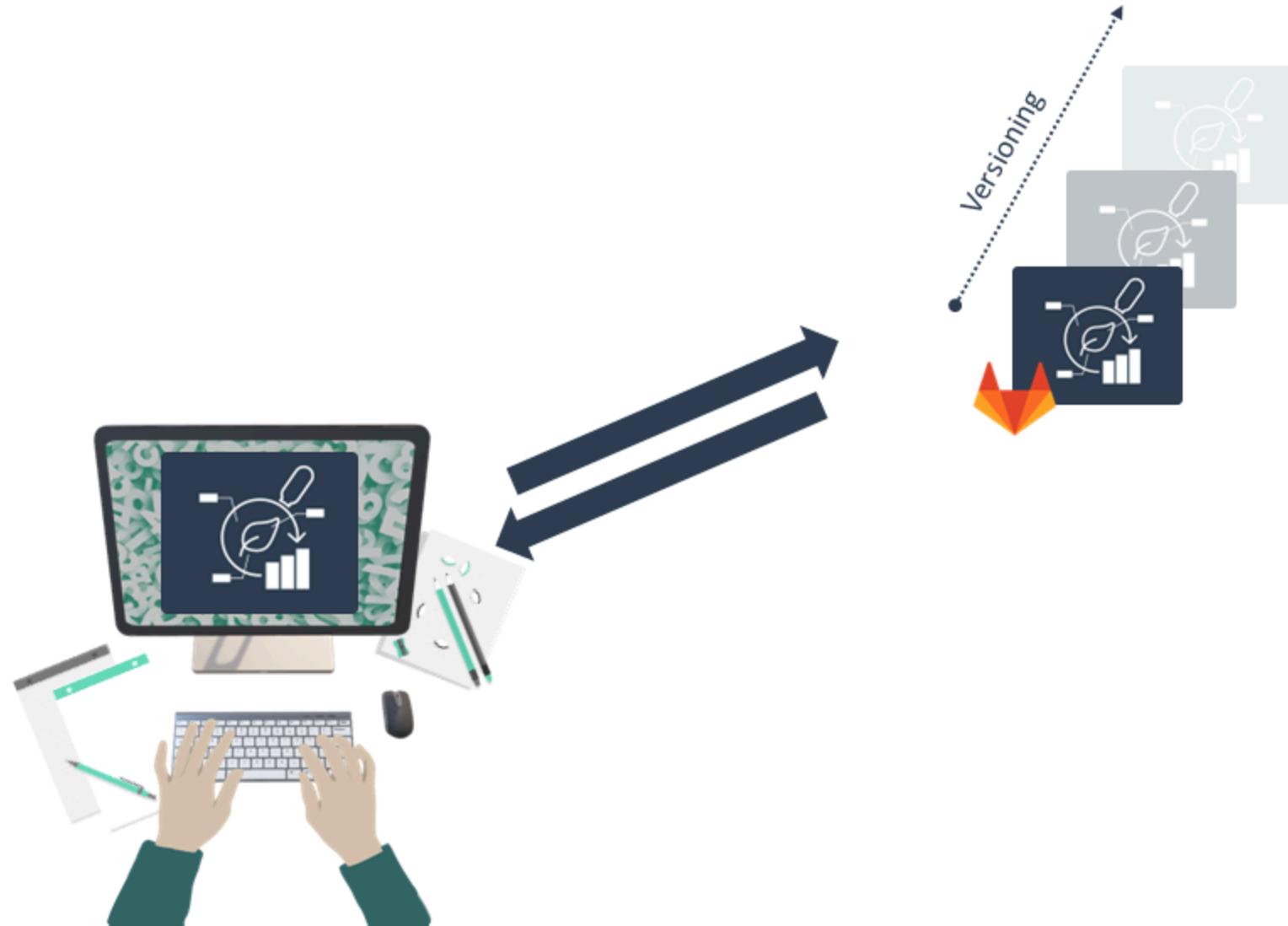
ARC builds on standards + Git



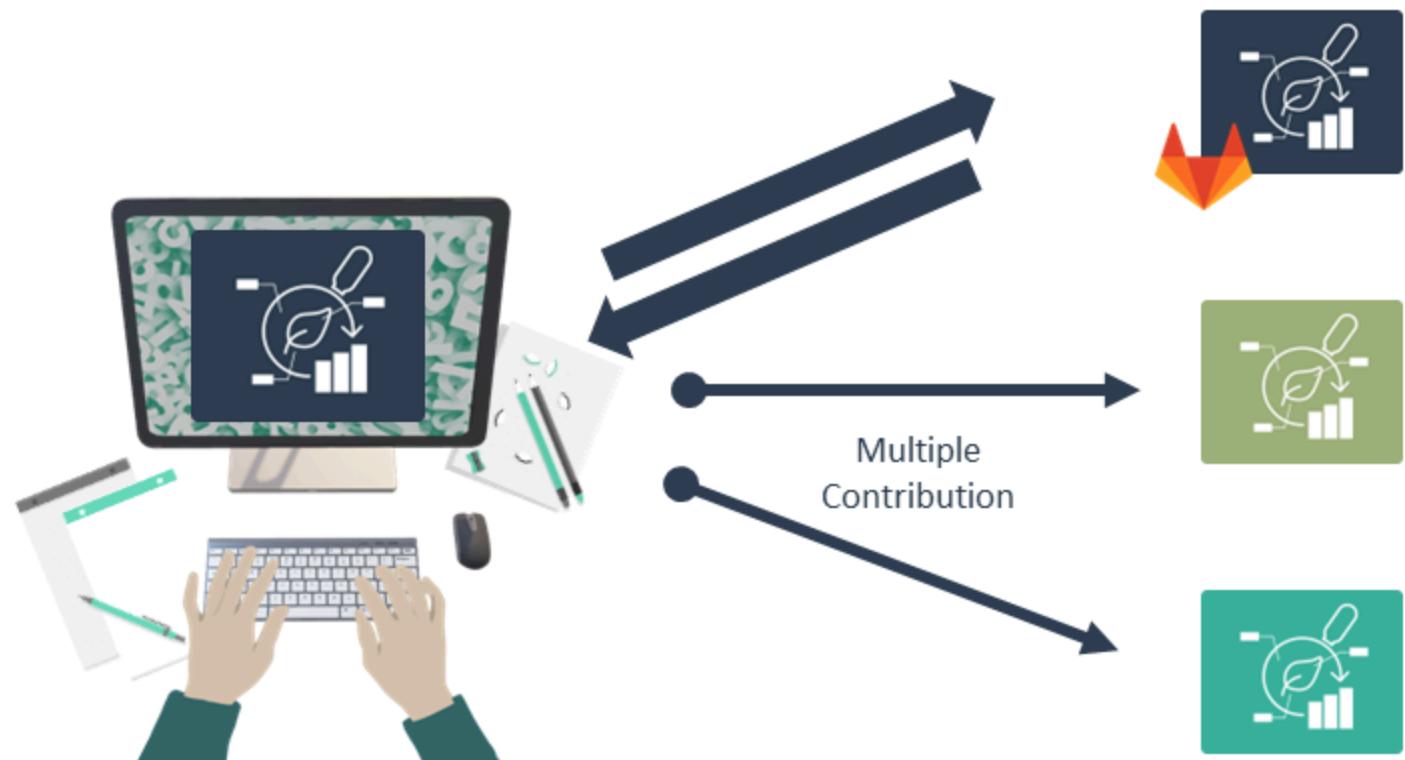
The DataPLANT DataHUB – a GitLab *Plus*



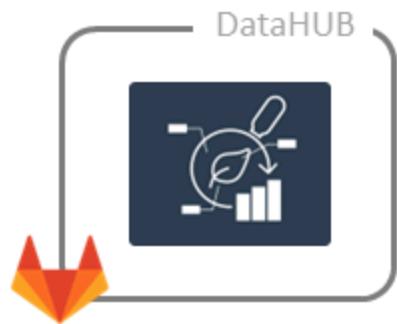






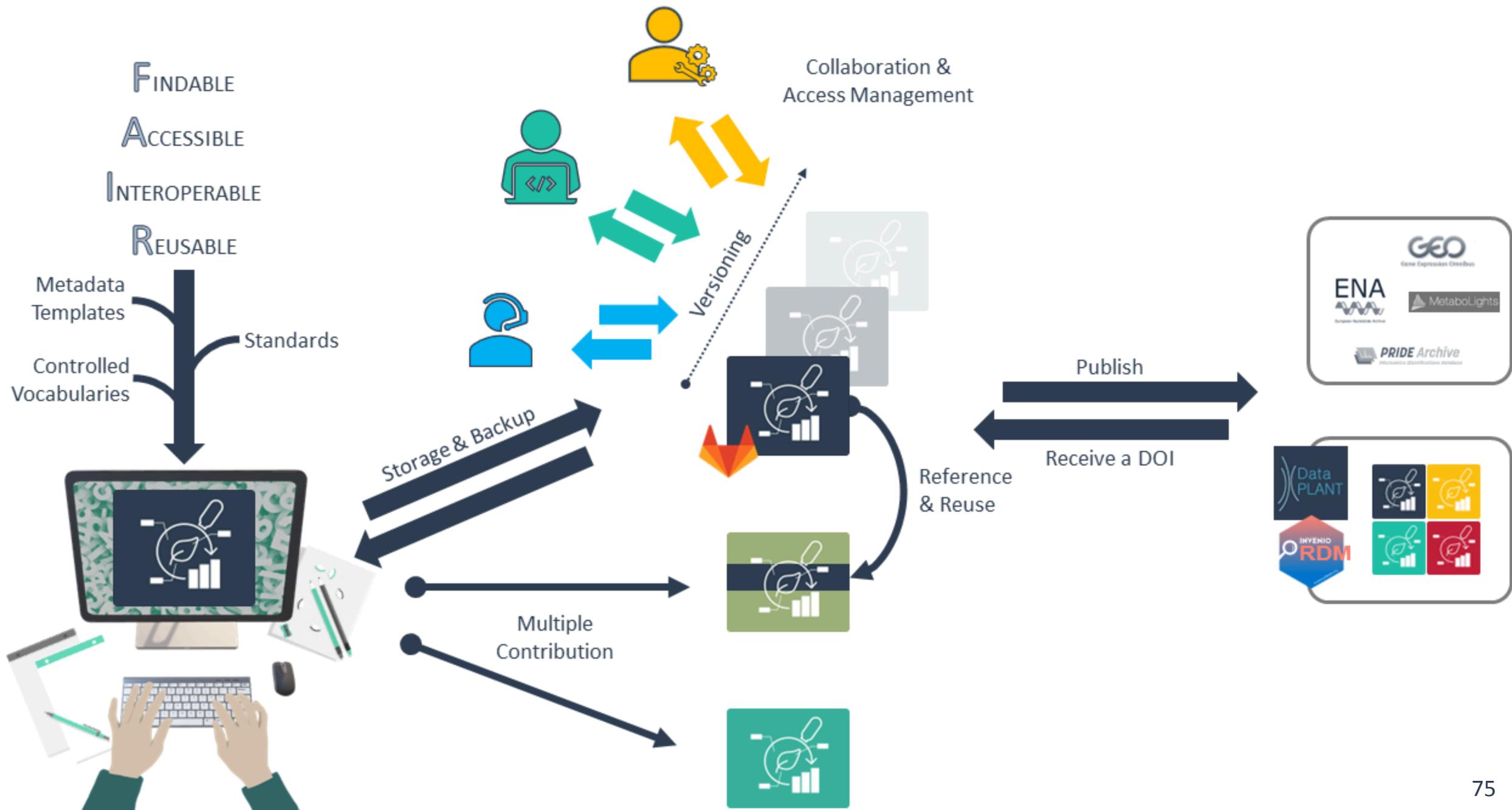




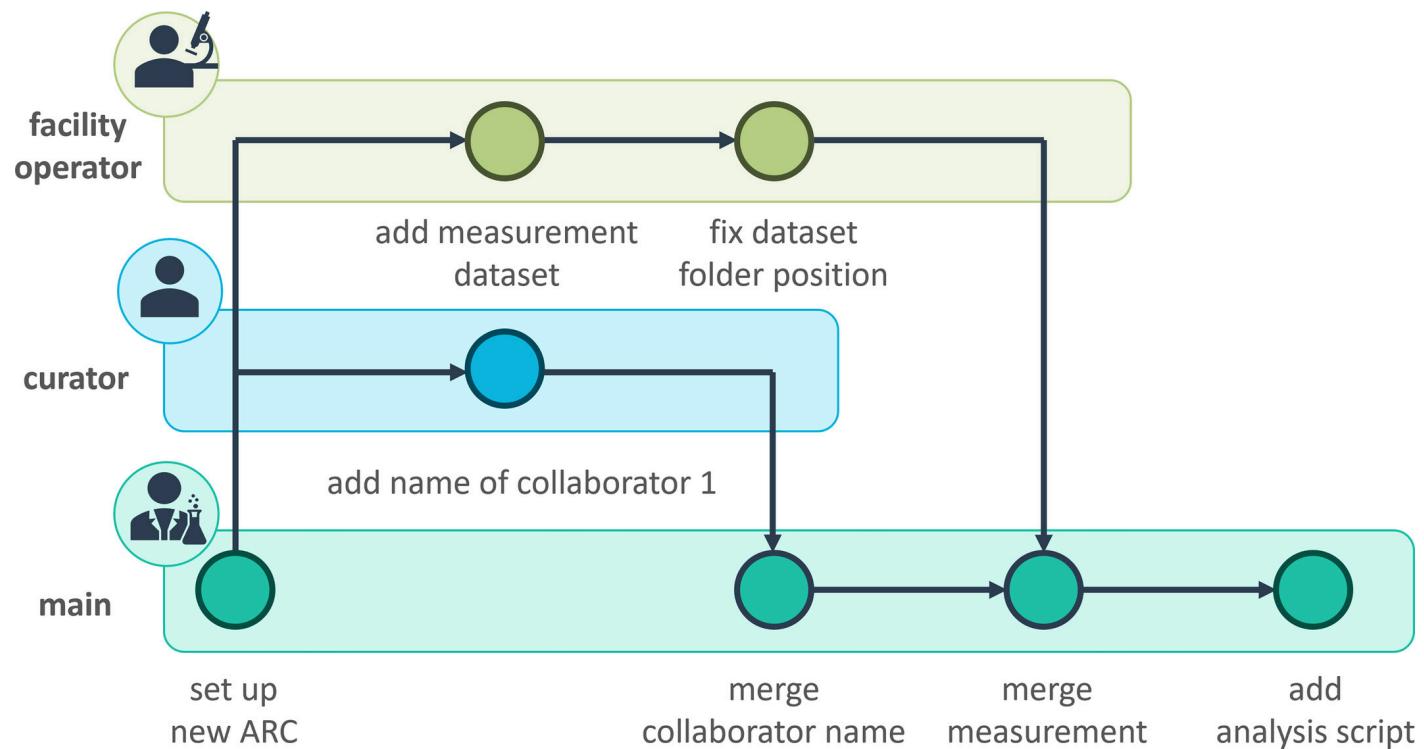


Publish
Receive a DOI

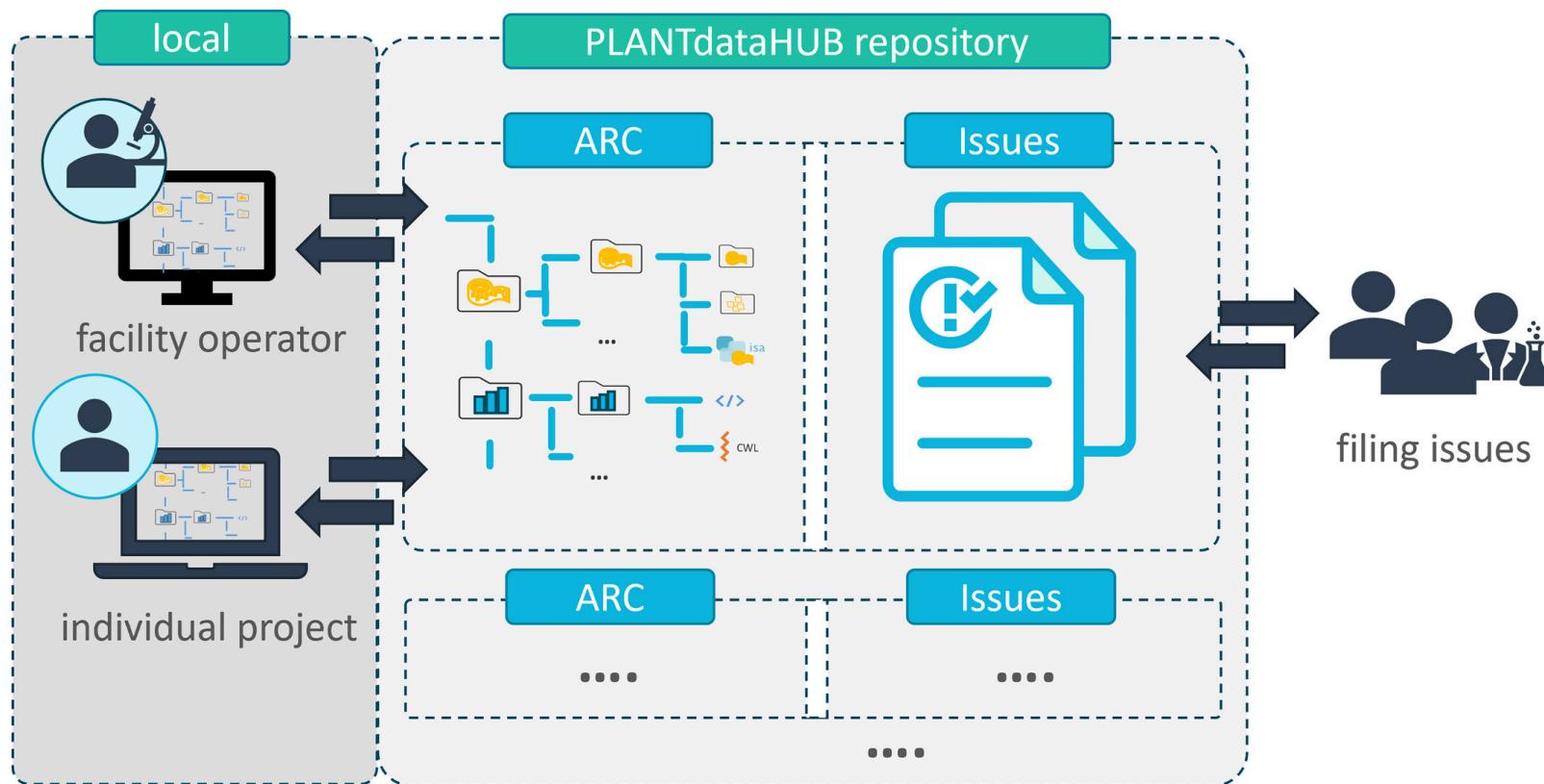




Mutable data life cycle



Project management



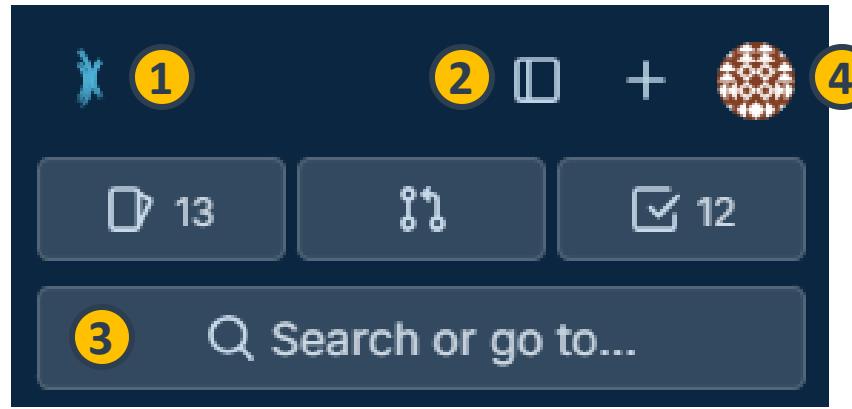
Contributors

Slides presented here include contributions by

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

DataHub Hands-On

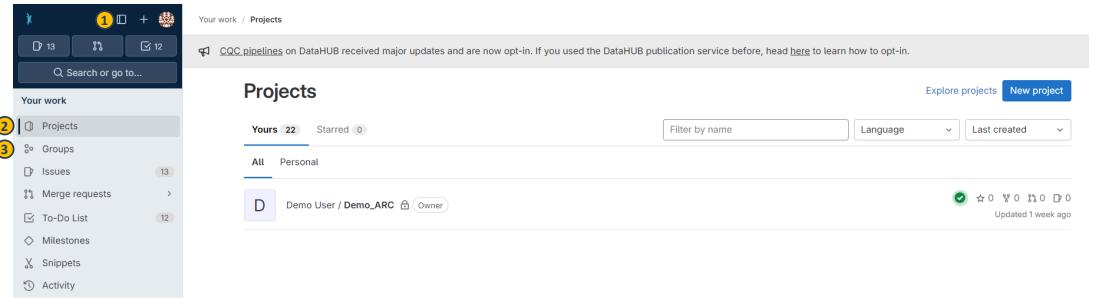
Navigation Bar



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

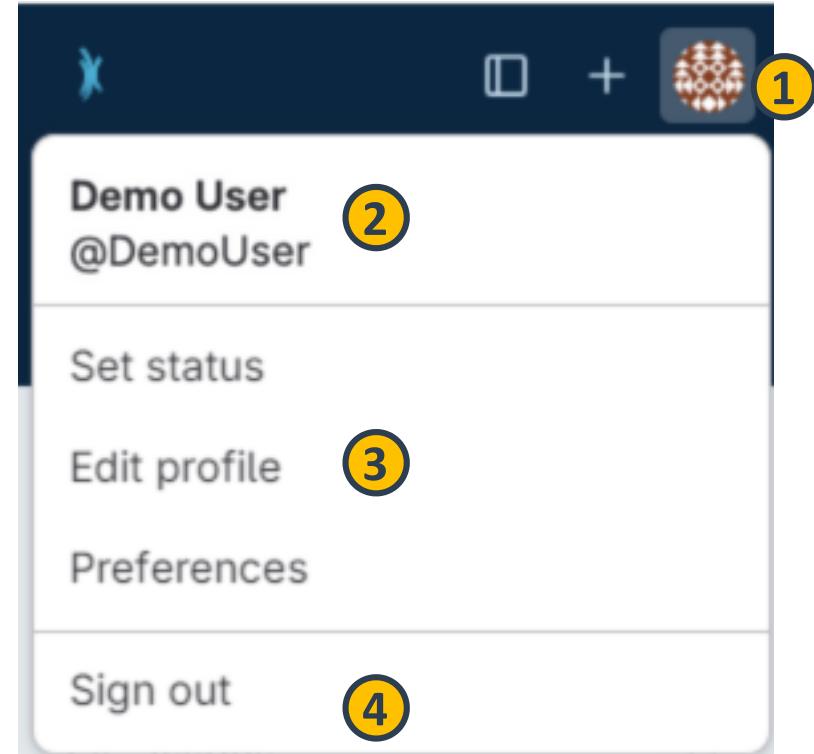
Hamburger Menu

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



Avatar Menu

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



Projects Panel

The screenshot shows the 'Your work / Projects' section of the DataHub interface. On the left, a sidebar lists 'Your work' options: Projects (selected), Groups, Issues (13), Merge requests, To-Do List (12), Milestones, Snippets, and Activity. The main area is titled 'Projects' and shows a summary: 'Yours 22' and 'Starred 0'. It includes filters for 'Filter by name', 'Language', and 'Last created'. A banner at the top right says: 'CQC pipelines on DataHUB received major updates and are now opt-in. If you used the DataHUB publication service before, head [here](#) to learn how to opt-in.' Six numbered callouts point to specific features: (1) 'Yours 22' tab, (2) project card for 'Demo User / Demo_ARC', (3) 'Owner' badge, (4) activity indicators (star, file, comment, reply), (5) 'Explore projects' button, and (6) 'New project' button.

ARC Panel

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

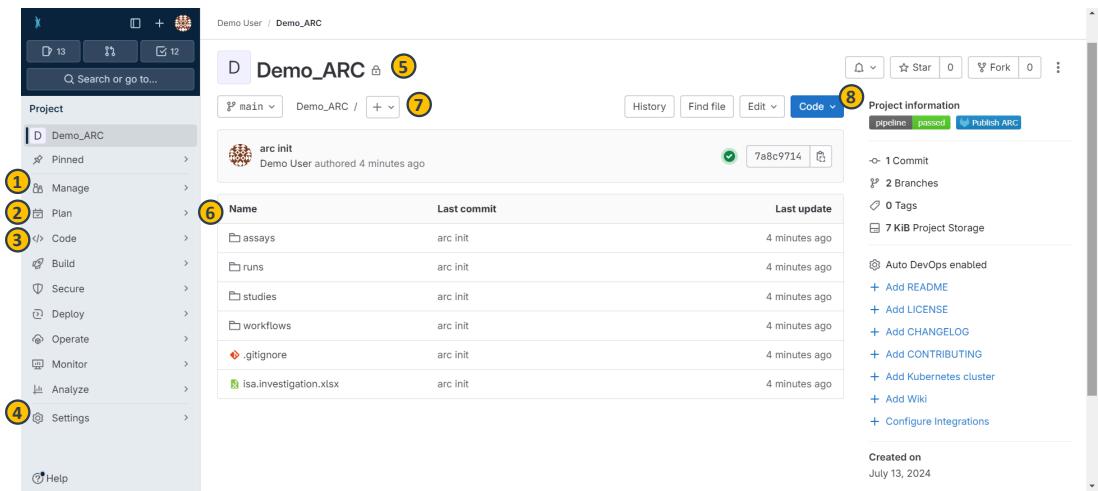
The screenshot shows the ARC Panel interface for a project named "Demo_ARC".

- Left Sidebar:** Contains navigation links for "Manage", "Plan", "Code", "Build", "Secure", "Deploy", "Operate", "Monitor", "Analyze", "Settings", and "Help".
- Project Header:** Shows the project name "Demo_ARC" with a lock icon and a badge with the number "5".
- Code Repository:** Displays a commit history for "arc init" by "Demo User" 4 minutes ago. The commit hash is 7a8c9714. A table below lists files and their last commits.
- Code Tab:** Active tab in the top right, showing options for "History", "Find file", "Edit", and "Code".
- Project Information:** Shows a green "pipeline passed" status and a "Publish ARC" button.
- Project Metrics:** Includes "1 Commit", "2 Branches", "0 Tags", and "7 KiB Project Storage".
- Auto DevOps:** Options to add README, LICENSE, CHANGELOG, CONTRIBUTING, Kubernetes cluster, Wiki, and Configure Integrations.
- Created On:** Date listed as July 13, 2024.

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

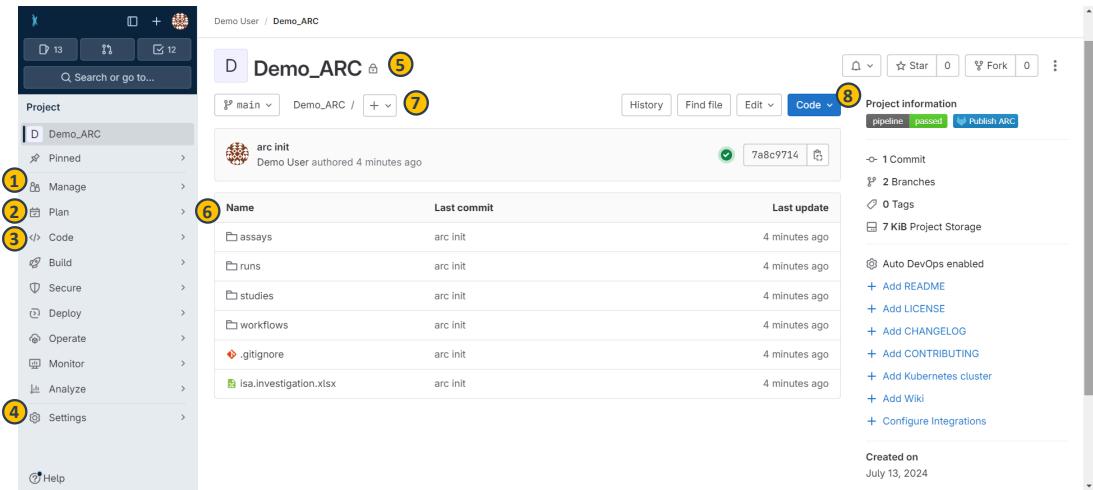
ARC Panel – sidebar

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



ARC Panel – main panel

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



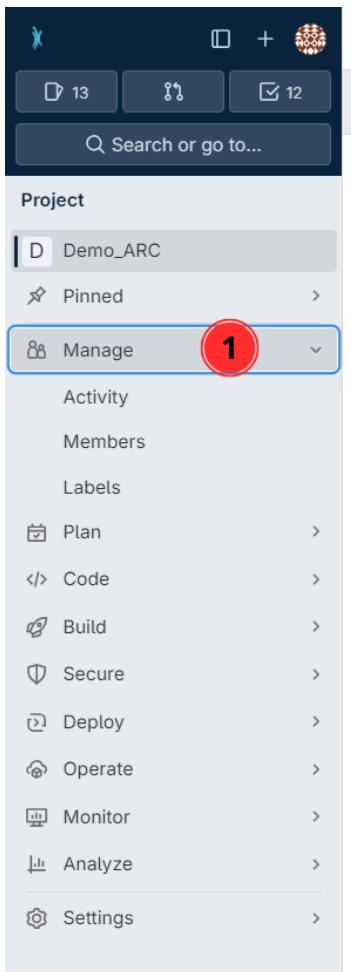
Collaborate and share



Invite collaborators

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

1. Click on Project Information in the left navigation panel



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

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

Project information

pipeline passed | Publish ARC

-o 1 Commit
2 Branches
0 Tags
21 KiB Project Storage

Auto DevOps enabled

Wiki
+ Add README
+ Add LICENSE
+ Add CHANGELOG
+ Add CONTRIBUTING
+ Add Kubernetes cluster
+ Configure Integrations

Created on
July 13, 2024

2. Click on Members

The screenshot shows the DataHub interface for a project named 'Demo_ARC'. The left sidebar has a 'Project' section with a pinned item 'Demo_ARC' and a 'Members' item highlighted with a red circle labeled '2'. Other items in the sidebar include 'Pinned', 'Manage', 'Activity', 'Labels', 'Plan', 'Code', 'Build', 'Secure', 'Deploy', 'Operate', 'Monitor', 'Analyze', and 'Settings'. The main content area is titled 'Project members' and displays one member: 'Demo User @DemoUser' (highlighted with a red circle labeled '1'). The member is listed under 'Account' with a circular profile icon, 'Source' as 'Direct member by Demo User', 'Max role' as 'Owner', and 'Expiration' set to '8+ Sep 27, 2023'. A 'Filter members' search bar and an 'Account' dropdown are also visible.

Demo User / Demo_ARC / Members

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

Project members

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

Members 1

Filter members Account ▾

Account	Source	Max role	Expiration	Activity
Demo User @DemoUser It's you	Direct member by Demo User	Owner	8+ Sep 27, 2023 ✓ Jul 13, 2024 ✗ Jul 21, 2024	⋮

3. Click on Invite members

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

Demo User / Demo_ARC / Members

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

Project members

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

Members 1

Filter members Account ▾

Account	Source	Max role	Expiration	Activity
Demo User @DemoUser It's you	Direct member by Demo User	Owner	Expiration date Sep 27, 2023 ✓ Jul 13, 2024 ✗ Jul 21, 2024	⋮

4. Search for potential collaborators

Invite members

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

Username, name or email address

4

Select members or type email addresses

Select a role

Guest

▼

[Read more about role permissions](#)

5. Select a role

A screenshot of a user interface showing a list of roles for a project. The 'Guest' role is selected and highlighted with a blue border. A red circle with the number '4' is positioned near the top right of the list area. A red circle with the number '5' is positioned near the bottom right of the list area. The list includes:

- Guest (selected)
- Reporter
- Developer
- Maintainer
- Owner

Below the list is a dropdown menu set to 'Guest'. At the bottom of the screen, there is a link to 'Read more about role permissions'.

Choosing the proper role

Guests

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

Reporters

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

Developers

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

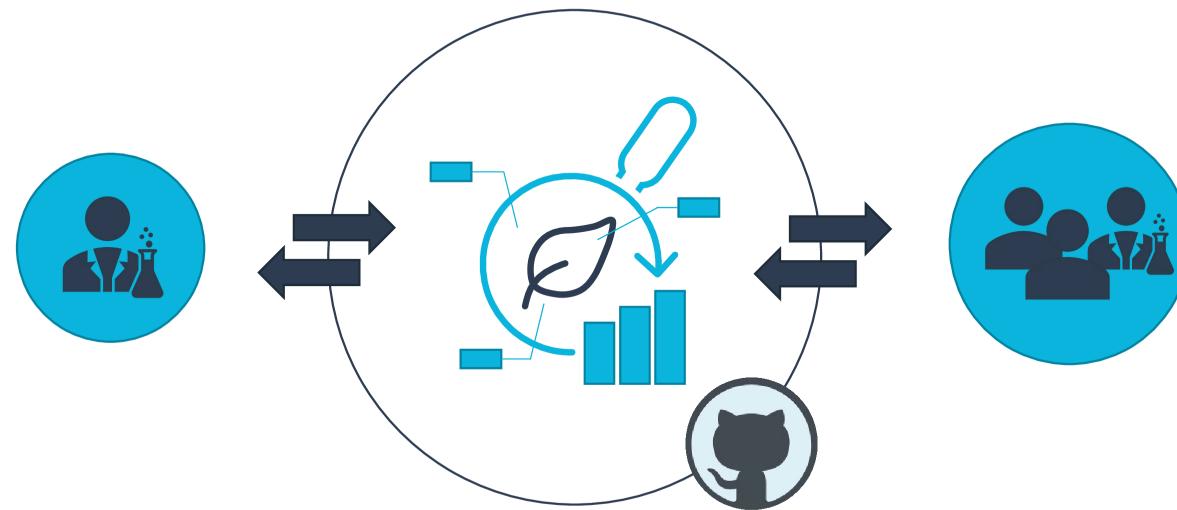
Maintainers

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

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

Congratulations!

You have just shared your ARC with a collaborator.



Version control

- Commit history

Project Management

- You can use issues

ARCs come with their own wiki space

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

The screenshot shows a DataHub interface for a 'Demo_ARC' project. The left sidebar includes sections for Project (Demo_ARC), Pinned, Manage, Plan, Issues (0), Issue boards, Milestones, Wiki (selected), Code, Build, Secure, and Help. The main content area is titled 'Home' and contains a message about CQC pipelines, a 'Last edited by Demo User just now' notice, and a placeholder text: 'This is the wiki to Demo_ARC. We will announce meeting schedules here.' On the right, there's a 'Pages' section with 2 items, a 'Home' link, a 'Meeting Schedule' section with three items, and an 'Ideas and drafts' section with one item.

Contributors

Slides presented here include contributions by

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

Metadata and ISA

**What is
metadata?**

Viola's PhD Project

Exercise: Take 5 minutes to note down the metadata

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

Metadata everywhere

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

Project metadata

project design

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

experimental processes

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

data-analytical processes

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

Other types of metadata

bibliographic

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

legal or administrative

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

technical

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

Metadata from a FAIR perspective

Findable

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

Interoperable

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

Reusable

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

Accessible

- information about origin
- location of storage
- access rights

Metadata "Standards"

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

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

Metadata standards ≈ Checklists

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

A small Interactive detour

-> favorite Movie

How does google "know"?!

Google X 🔍

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

Ungefähr 37.300.000 Ergebnisse (0,39 Sekunden)

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

Besetzung >



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

Jimmie Dimmick
Vincent Vega
Jules Winnfield
Mia Wallace
Butch Coolidge
Pumpkin

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

Pulp Fiction

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

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

Weitere Fragen

Was ist so besonders an Pulp Fiction? ▾

Was bedeutet der Titel Pulp Fiction? ▾

Warum ist Pulp Fiction ein Kultfilm? ▾

Film ansehen

DIENSTE BEARBEITEN

 Jetzt ansehen Premium-Abo  Angesehen  Möchte ich sehen

 Ab 2,99 €  Ansehen

 Ab 2,99 €  Ansehen

 Ab 3,99 €  Ansehen

[Alle Optionen zum Ansehen](#) ▾

Info

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

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

IMDb Amazon Wer streamt ...

Dieser Film gefiel 92 % der Nutzer  Google-Nutzer 

Schemas and machine-readability

Structured data and the internet

Schema.org

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

Structured data and the internet: Schema.org

<https://schema.org/Person>

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

JSON-LD

JSON-LD = JavaScript Object Notation for Linked Data

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

RDFa

RDFa = Resource Description Framework in Attributes

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

Standards

Dublin Core

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

DataCite Schema

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

DataCite Schema: Simple Example

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

```

Ontologies

Ontology

(Sometimes also referred to "semantic model")

An ontology combines features of

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

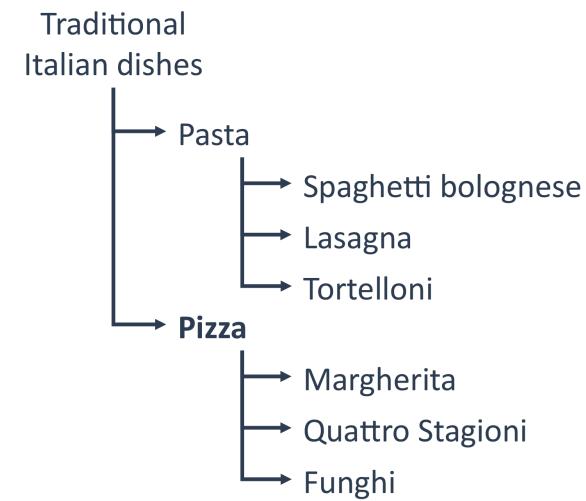
Dictionary

Alphabetically lists terms and their definitions

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

Taxonomy

Hierarchy or classification



Thesaurus

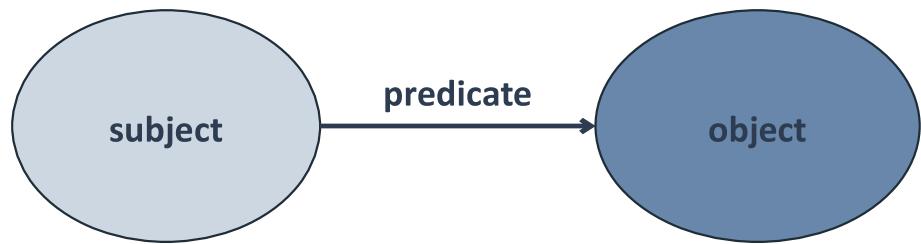
Dictionary of synonyms and relations

Pizza ≈ Lahmacun ≈ Focaccia ≈ Flammkuchen

Ontology

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

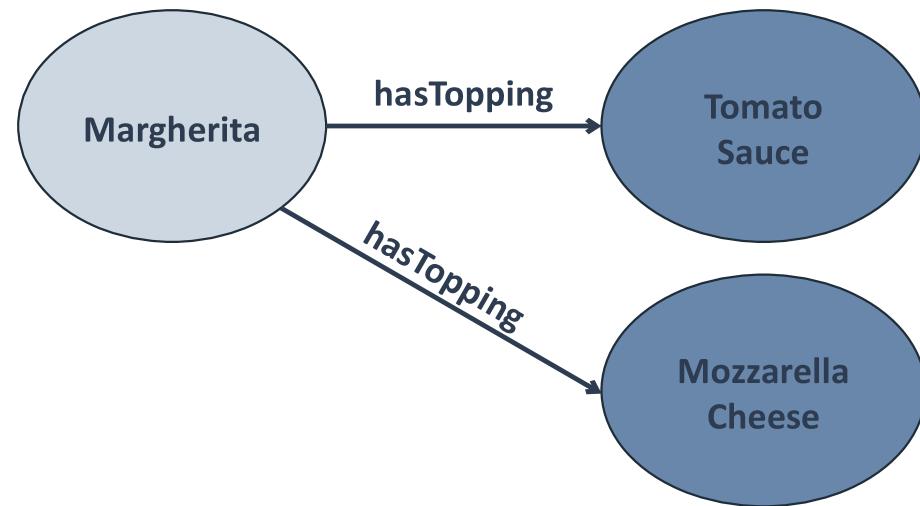
The semantic triple



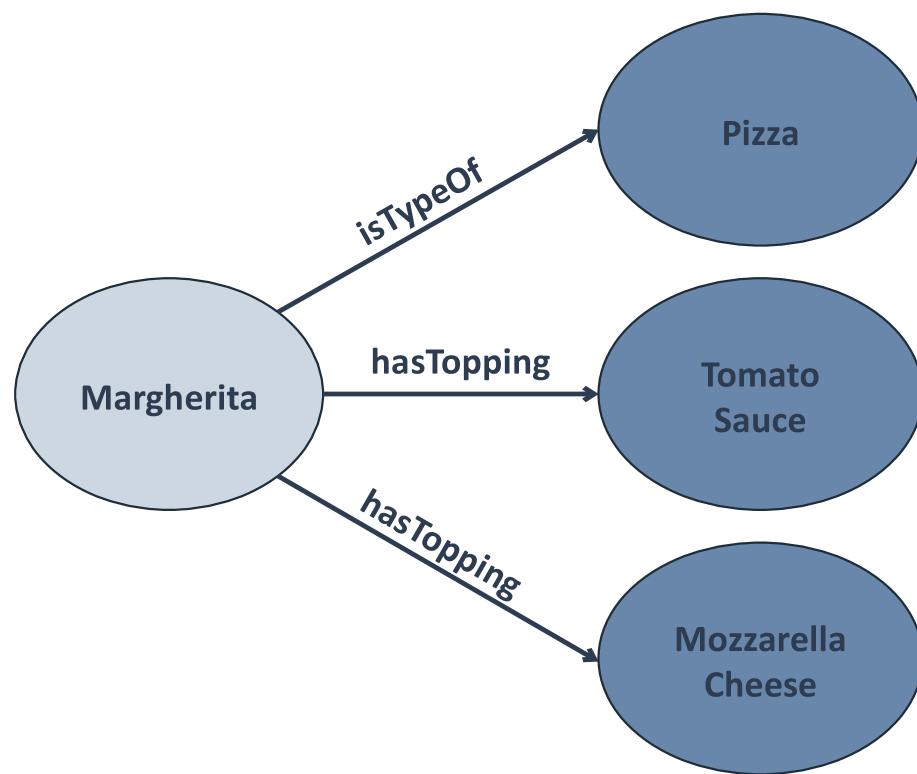
Modeling a pizza menu



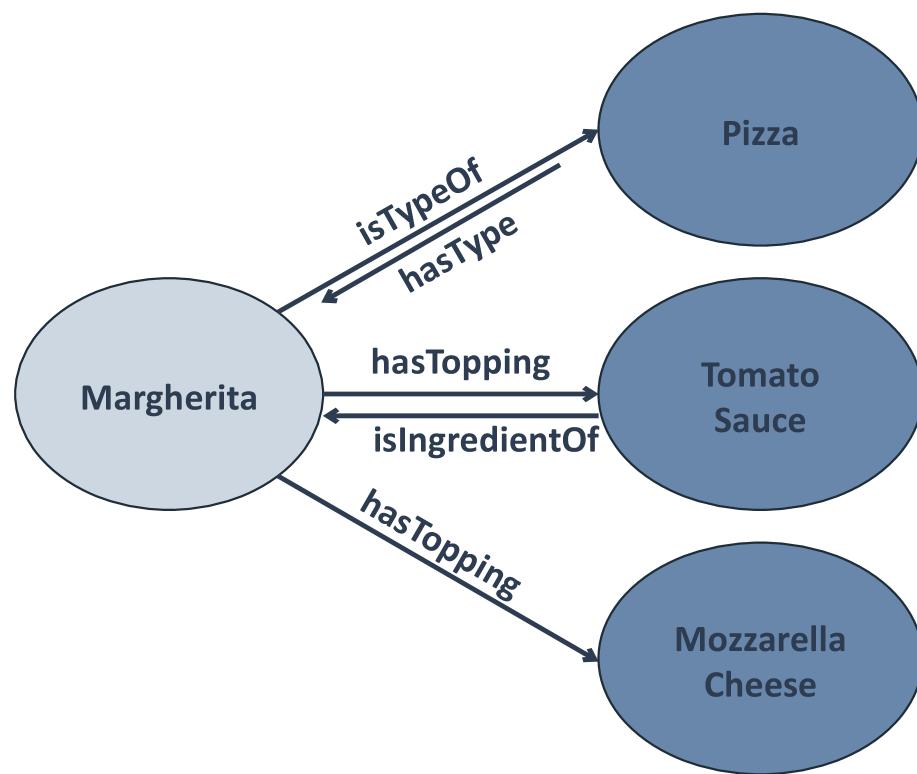
Modeling a pizza menu



Modeling a pizza menu

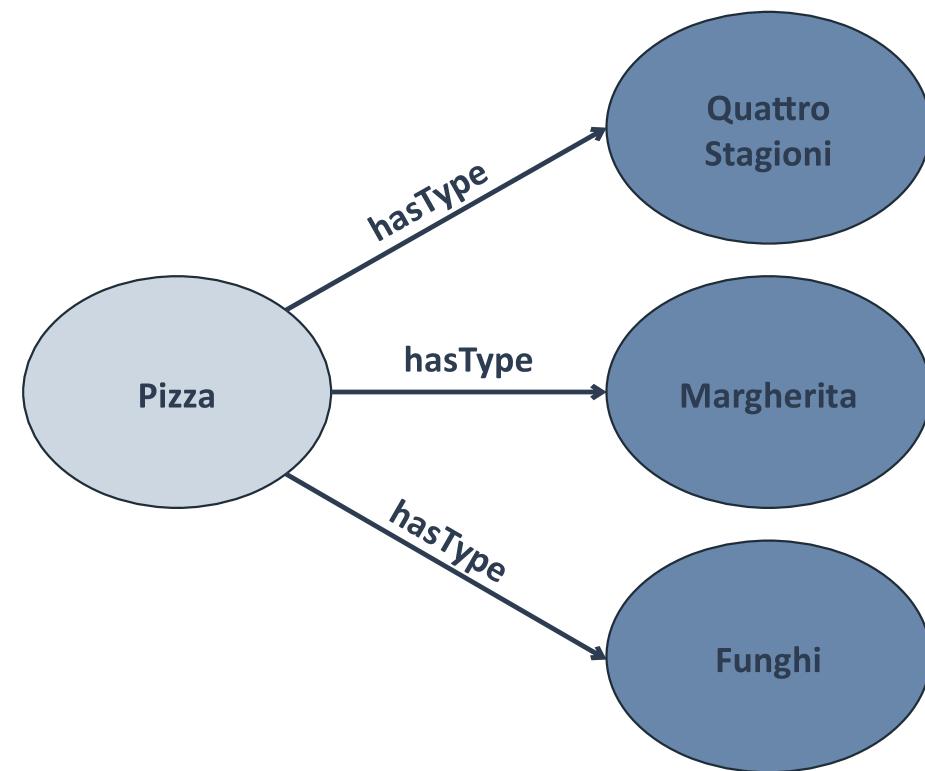


Predicates have two directions

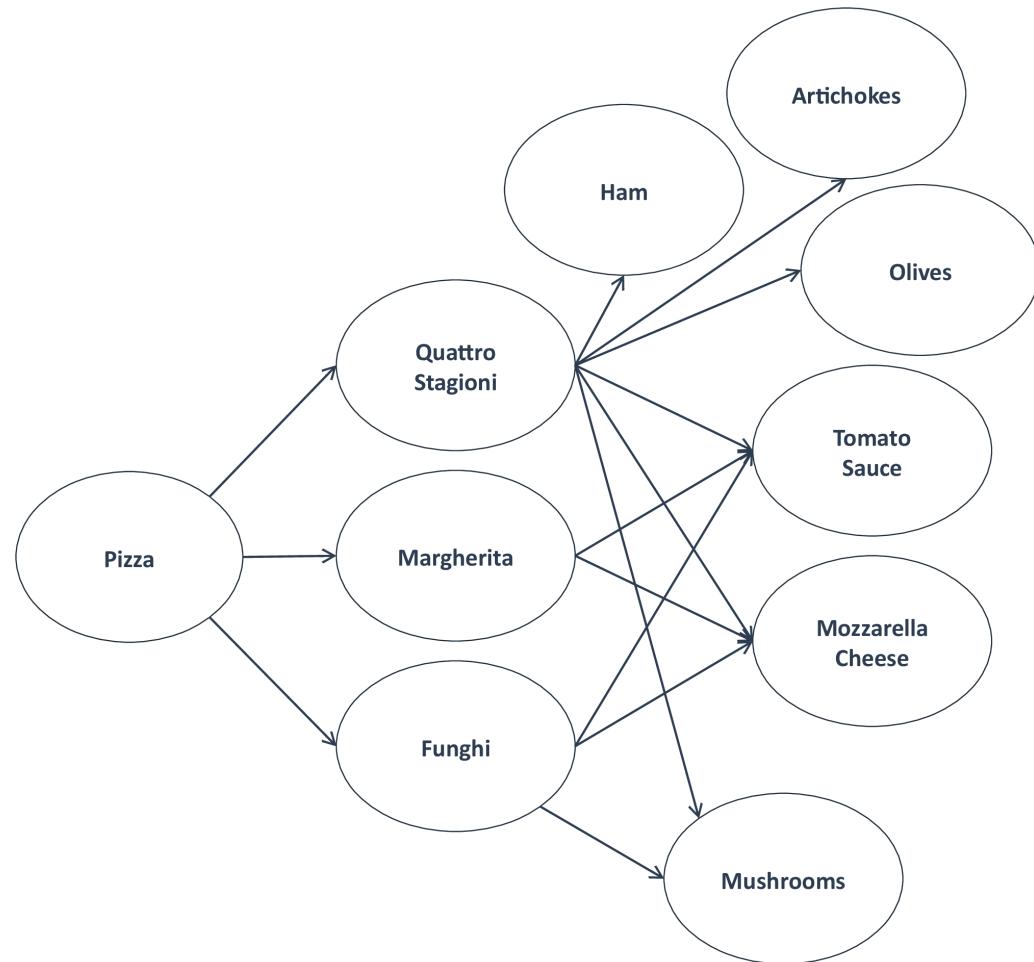


Looking at the menu from a different perspective

An object of one triplet can be the subject to another



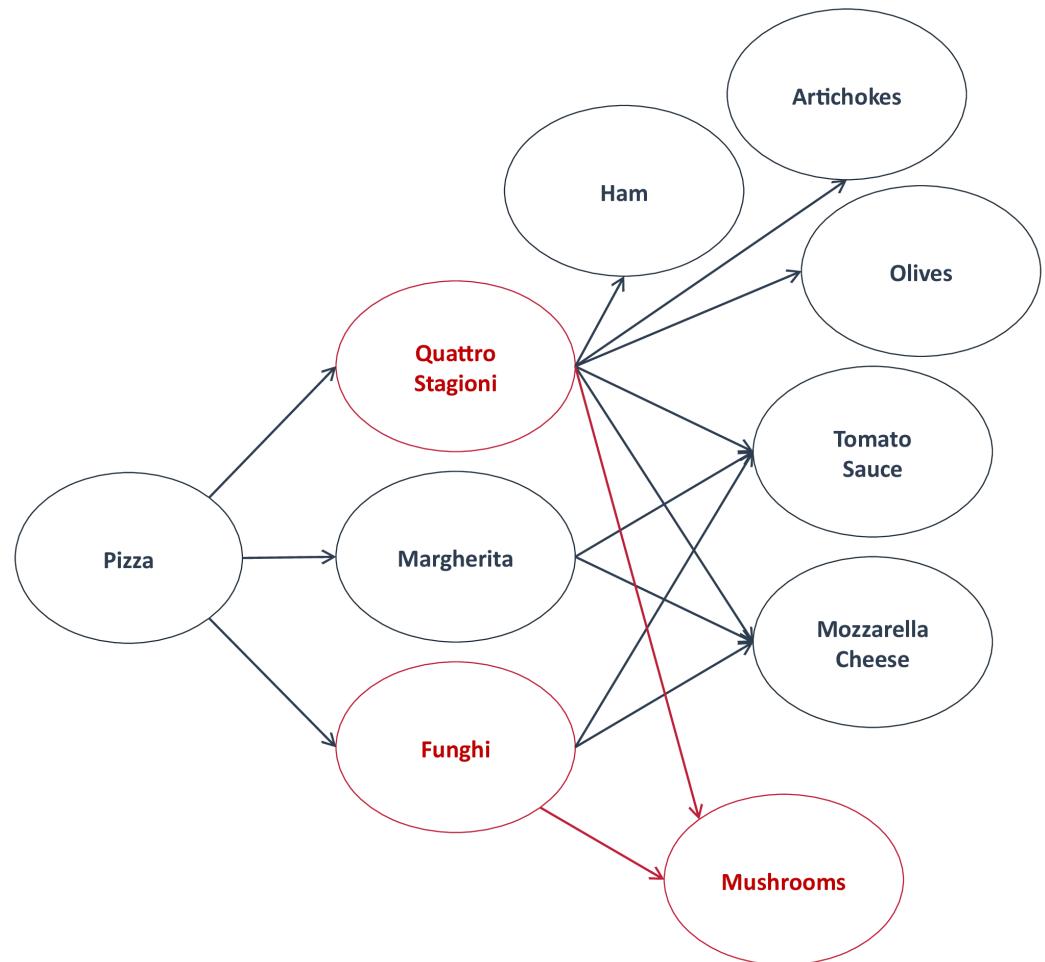
(Towards) a knowledge graph



Searching the menu

An ontology can be queried:

- *"name all pizzas with topping mushrooms"*



The Pizza Ontology

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

Example ontologies

EDAM ontology

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

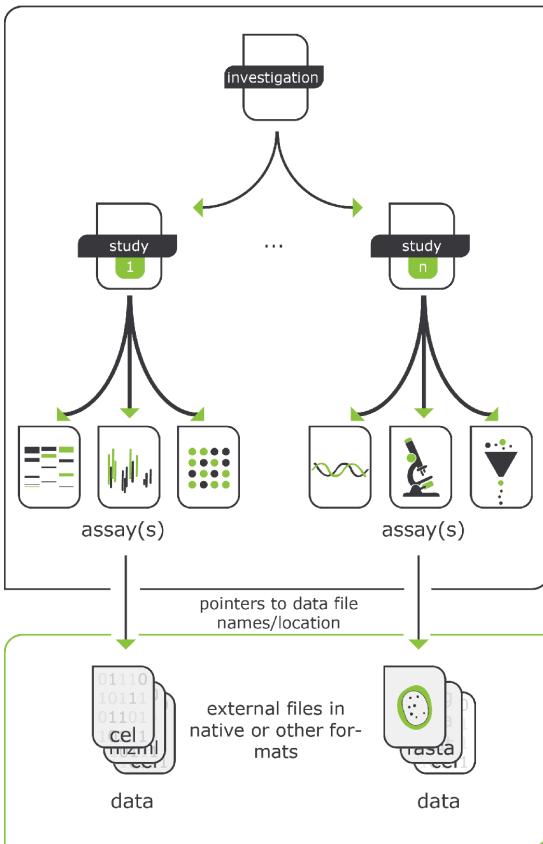
PECO ontology

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

Explore more examples

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

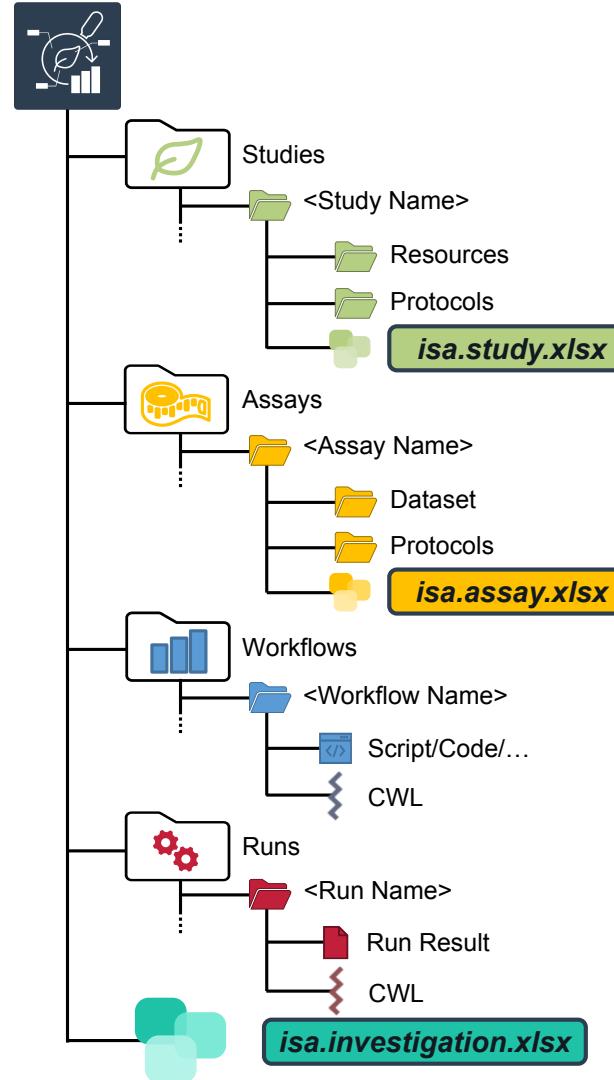
ARC builds on ISA



Investigation
Overall goals
Scientific context

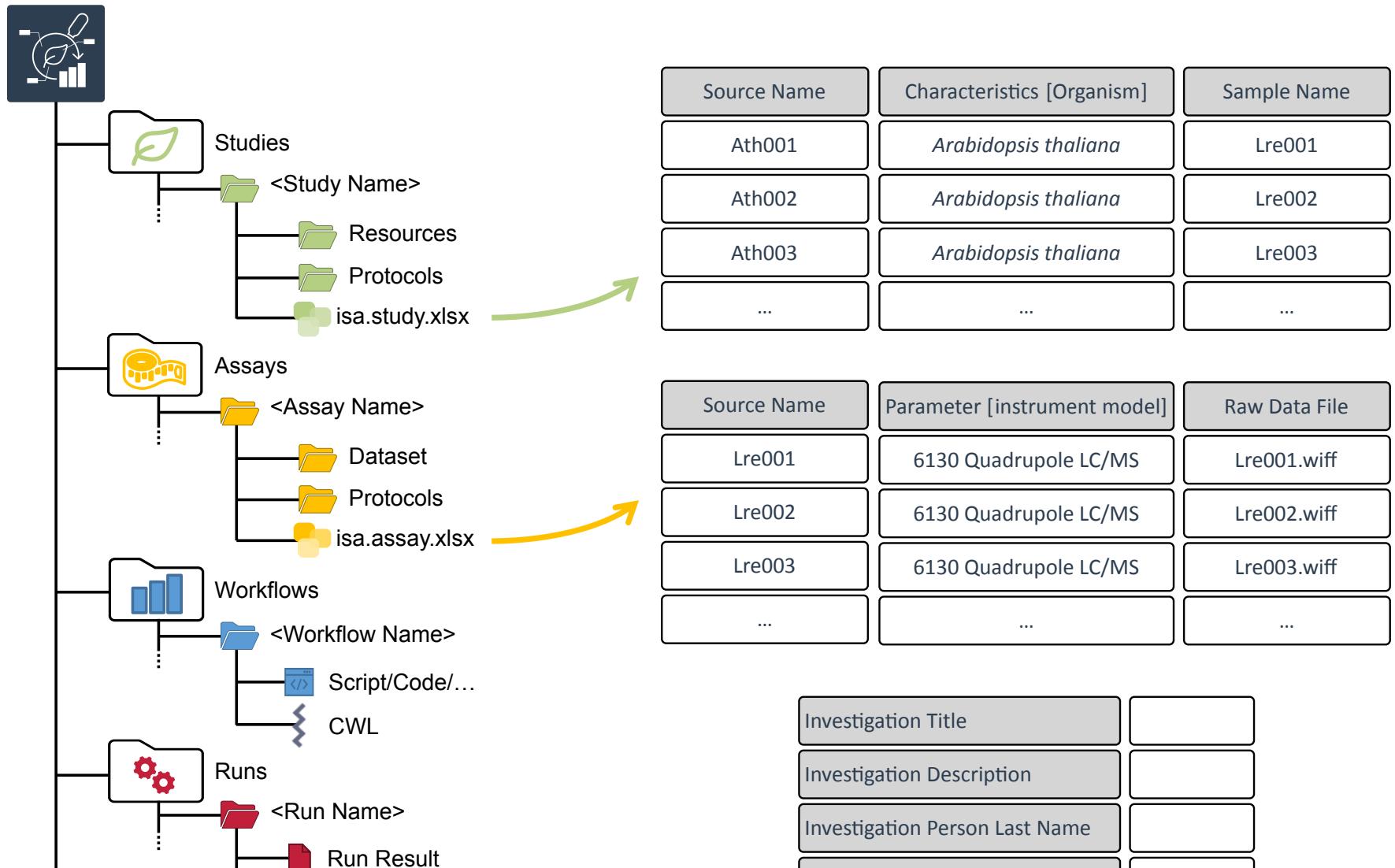
Study
Experimental design

Assay
Leading to (raw) data



ARC builds on ISA

Metadata Annotations



isa.<>.xlsx files within ARCs

ISA INVESTIGATION.XLSX								
1	ONTOLOGY SOURCE REFERENCE	OB	BTO	NEWT	UO	CHEBI	PATO	EFO
2	Term Source Name							
3	Term Source File	http://biportal.bio ArrayExpress Experimental Factor Ontology						
4	Term Source Version	47993	v1.26	v1.26	v1.26	v1.26	v1.26	v1.26
5	Term Source Description	Ontology for Biomed BRENDAs tissue / NEWT UniProt Tax Unit Ontology Chemical Ent PhenoTypic c ArrayExpress Experimental Factor Ontology						
6	INVESTIGATION							
7	Investigation Identifier							
8	Investigation Title	Growth control of the eukaryote cell: a systems biology study in yeast						
9	Investigation Description	Background Cell growth underlies many key cellular and developmental processes, yet a limited number of studies have been carried out on cell growth.						
10	Investigation Submission Date	30.04.07						
11	Investigation Public Release Date	10.03.09						
12	Comment [Created With Configuration]							
13	Comment [Last Opened With Configuration]	isaconfig default_v2013_02_13						
14	INVESTIGATION PUBLICATIONS							
15	Investigation Pubmed ID	17439666						
16	Investigation DOI	doi:10.1186/jbiol54						
17	Investigation Publication Author List	Castro I, Zeef LA, Hoyle DC, Zhang N, Hayes A, Gardner DC, Cornell MJ, Petty J, Hakes L, Wardleworth L, Rash B, Brown M, Dunn WB, Broadhurst						
18	Investigation Publication Title	Growth control of the eukaryote cell: a systems biology study in yeast.						
19	Investigation Publication Status	published						
20	Investigation Publication Status Term Accession Number							
21	Investigation Publication Status Term Source REF							
22	INVESTIGATION CONTACTS							
23	Investigation Person Last Name	Stephen	Castillo	Zeeb				
24	Investigation Person First Name	Oliver	Juan	Lao				
25	Investigation Person Mid Initials	G	I	A				
26	Investigation Person Email							
27	Investigation Person Phone							
28	Investigation Person Tax							
29	Investigation Person Address	Oxford Road, Manch Oxford Road, M5 Oxford Road, Manchester M13 9P', UK						
30	Investigation Person Affiliation	Faculty of Life Scienc Faculty of Life Sc Faculty of Life Sciences, Michael Smith Building, University of Manchester						
31	Investigation Person Roles	corresponding auth author						
32	Investigation Person Roles Term Accession Number							
33	Investigation Person Roles Term Source REF							
34	Comment [Investigation Person ERF]							
35	STUDY							
36	Study Identifier	BII-5-1						
37	Study Title	Study of the impact of changes in flux on the transcriptome, proteome, endometabolome and exometabolome of the yeast <i>Saccharomyces cerevisiae</i> .						
38	Study Description	We wished to study the impact of growth rate on the total complement of mRNA molecules, proteins, and metabolites in <i>S. cerevisiae</i> , independent						
39	Comment [Study Grant Number]							
40	Comment [Study Funding Agency]							
41	Study Submission Date	30.04.07						
42	Study Public Release Date	10.03.09						
43	Study File Name	s_BII-5-1.txt						
44	STUDY DESIGN DESCRIPTORS							
45	Study Design Type	Intervention design						
46	Study Design Type Term Accession Number	http://purl.obolibrary.org/obo/OBI_0000115						
47	Study Design Type Term Source REF	OBI						
48	STUDY PUBLICATIONS							
49	Study Pubmed ID	17439666						
50	Study Publication DOI	doi:10.1186/jbiol54						
51	Study Publication Author List	Castro I, Zeeb LA, Hoyle DC, Zhang N, Hayes A, Gardner DC, Cornell MJ, Petty J, Hakes L, Wardleworth L, Rash B, Brown M, Dunn WB, Broadhurst						
52	Study Publication Title	Growth control of the eukaryote cell: a systems biology study in yeast.						
53	Study Publication Status	published						
54	Study Publication Status Term Accession Number							
55	Study Publication Status Term Source REF							
56	STUDY FACTORS							
57	Study Factor Name	limiting nutrient	rate					
58	Study Factor Type	chemical compound						
59	Study Factor Type Term Accession Number	http://purl.obolibrary.org/obo/PATO_0001616						
60	Study Factor Type Term Source REF	PATO						
61	STUDY ASSAYS							
62	Study Assay Measurement Type	protein expression	metabolite profil	transcription profiling				
63	Study Assay Measurement Type Term Accession Number	http://purl.obolibrary.org/obo/OBI_0000111	http://purl.obolibrary.org/obo/OBI_0000111	424				
64	Study Assay Measurement Type Term Source REF	OBI	OBI	OBI				
65	Study Assay Technology Type	mass spectrometry	mass spectromet	DNA microarray				
66	Study Assay Technology Type Term Accession Number	http://purl.obolibrary.org/obo/OBI_0400148						
67	Study Assay Technology Type Term Source REF	OB	OB	OB				
68	Study Assay Technology Platform	iTRAQ	LC-MS/MS	Affymetrix				
69	Study Assay File Type	a_proteome.txt	a_proteome.txt	a_proteome.txt				
70	STUDY PROTOCOLS							
71	Study Protocol Name	growth protocol	mRNA extraction	protein extraction	biotin labelling	EukGE-W54	metabolite extraction	
72	Study Protocol Type	growth	mRNA extraction	protein extraction	labeling	hybridization	extraction	
73	Study Protocol Type Term Accession Number							
74	Study Protocol Type Term Source REF							
75	Study Protocol Description	1. Biomass samples	1. Biomass samples (45 ml) were tak	This was done using Enzo	For each target, a hybridisation cocktail was made using the			
76	Study Protocol URI							

A	B	C	D	E	F	G
1	Source Name	Characteristics [soluble protein content]	Parameter [Quantification method#2]	Parameter [1S1N Photosynthesis QconCAT mass#4]	Sample Name	
2	G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis		0,75 microgram	WGCr2_U1
3	G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis		0,15 microgram	WGCr2_U2
4	G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis		0,03 microgram	WGCr2_U3
5	G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis		0,01 microgram	WGCr2_U4
6	G2_S32_15mL	50,00 microgram	absolute quantitation analysis		0,75 microgram	WGCr2_5_1
7	G2_S32_15mL	50,00 microgram	absolute quantitation analysis		0,15 microgram	WGCr2_5_2
8	G2_S32_15mL	50,00 microgram	absolute quantitation analysis		0,03 microgram	WGCr2_5_3
9	G2_S32_15mL	50,00 microgram	absolute quantitation analysis		0,01 microgram	WGCr2_5_4
10	G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis		0,75 microgram	WGCr2_UF_1
11	G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis		0,15 microgram	WGCr2_UF_2
12	G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis		0,03 microgram	WGCr2_UF_3
13	G2_UVM4_F3_15mL	50,00 microgram	absolute quantitation analysis		0,01 microgram	WGCr2_UF_4
14	G2_S32_F3_15mL	50,00 microgram	absolute quantitation analysis		0,75 microgram	WGCr2_SF_1
15	G2_S32_F3_15mL	50,00 microgram	absolute quantitation analysis		0,15 microgram	WGCr2_SF_2
16	G2_S32_F3_15mL	50,00 microgram	absolute quantitation analysis		0,03 microgram	WGCr2_SF_3
17	G2_S32_F3_15mL	50,00 microgram	absolute quantitation analysis		0,01 microgram	WGCr2_SF_4
18	G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis		0,75 microgram	WGCr1_U1
19	G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis		0,15 microgram	WGCr1_U2
20	G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis		0,03 microgram	WGCr1_U3
21	G1_UVM4_15mL	50,00 microgram	absolute quantitation analysis		0,01 microgram	WGCr1_U4
22	G1_S32_15mL	50,00 microgram	absolute quantitation analysis		0,75 microgram	WGCr1_5_1
23	G1_S32_15mL	50,00 microgram	absolute quantitation analysis		0,15 microgram	WGCr1_5_2
24	G1_S32_15mL	50,00 microgram	absolute quantitation analysis		0,03 microgram	WGCr1_5_3

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
Source Name	Parameter [sample volume]	Parameter [injection vol]	Raw Data File																
2	WCGr2_U1	100.00 microliter	2.50 microliter		1	TripleTOF 8600													87,00 minute WCGr2_U1.wiff
3	WCGr2_U2	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr2_U2.wiff
4	WCGr2_U3	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr2_U3.wiff
5	WCGr2_U4	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr2_U4.wiff
6	WCGr2_S_1	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr2_S_1.wiff
7	WCGr2_S_2	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr2_S_2.wiff
8	WCGr2_S_3	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr2_S_3.wiff
9	WCGr2_S_4	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr2_S_4.wiff
10	WCGr2_UF_1	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr2_UF_1.wiff
11	WCGr2_UF_2	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr2_UF_2.wiff
12	WCGr2_UF_3	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr2_UF_3.wiff
13	WCGr2_UF_4	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr2_UF_4.wiff
14	WCGr2_SF_1	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr2_SF_1.wiff
15	WCGr2_SF_2	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr2_SF_2.wiff
16	WCGr2_SF_3	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr2_SF_3.wiff
17	WCGr2_SF_4	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr2_SF_4.wiff
18	WCGr1_U1	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr1_U1.wiff
19	WCGr1_U2	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr1_U2.wiff
20	WCGr1_U3	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr1_U3.wiff
21	WCGr1_U4	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr1_U4.wiff
22	WCGr1_S_1	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr1_S_1.wiff
23	WCGr1_S_2	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr1_S_2.wiff
24	WCGr1_S_3	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr1_S_3.wiff
25	WCGr1_S_4	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr1_S_4.wiff
26	WCGr1_UF_1	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr1_UF_1.wiff
27	WCGr1_UF_2	100.00 microliter	2.50 microliter		1	TripleTOF 6600													87,00 minute WCGr1_UF_2.wiff

Study and assay files are registered in the investigation file

isa.investigation.xlsx

Study file

Assay file

isa.study.xlsx

isa.assay.xlsx

Sample Name	Characteristics [soluble protein content]	Parameter [Quantification method#]	Parameter [15N Photosynthesis QconCAT mass#4]	Sample Name
G2_UVM4_15mL	50,00 microgram absolute quantitation analysis			0,75 microgram WGr2_U1
G2_UVM4_15mL	50,00 microgram absolute quantitation analysis			0,15 microgram WGr2_U2
G2_UVM4_15mL	50,00 microgram absolute quantitation analysis			0,03 microgram WGr2_U3
G2_UVM4_15mL	50,00 microgram absolute quantitation analysis			0,01 microgram WGr2_U4
G2_532_15mL	50,00 microgram absolute quantitation analysis			0,75 microgram WGr2_5_1
G2_532_15mL	50,00 microgram absolute quantitation analysis			0,15 microgram WGr2_5_2
G2_532_15mL	50,00 microgram absolute quantitation analysis			0,03 microgram WGr2_5_3
G2_532_15mL	50,00 microgram absolute quantitation analysis			0,01 microgram WGr2_5_4
G2_UVM4_F3_15mL	50,00 microgram absolute quantitation analysis			0,75 microgram WGr2_UF_1
G2_UVM4_F3_15mL	50,00 microgram absolute quantitation analysis			0,15 microgram WGr2_UF_2
G2_UVM4_F3_15mL	50,00 microgram absolute quantitation analysis			0,03 microgram WGr2_UF_3
G2_UVM4_F3_15mL	50,00 microgram absolute quantitation analysis			0,01 microgram WGr2_UF_4
G2_532_F3_15mL	50,00 microgram absolute quantitation analysis			0,75 microgram WGr2_5_F_1
G2_532_F3_15mL	50,00 microgram absolute quantitation analysis			0,15 microgram WGr2_5_F_2
G2_532_F3_15mL	50,00 microgram absolute quantitation analysis			0,03 microgram WGr2_5_F_3
G2_532_F3_15mL	50,00 microgram absolute quantitation analysis			0,01 microgram WGr2_5_F_4
G1_UVM4_15mL	50,00 microgram absolute quantitation analysis			0,15 microgram WGr1_U2
G1_UVM4_15mL	50,00 microgram absolute quantitation analysis			0,03 microgram WGr1_U3
G1_UVM4_15mL	50,00 microgram absolute quantitation analysis			0,01 microgram WGr1_U4
G1_UVM4_15mL	50,00 microgram absolute quantitation analysis			0,01 microgram WGr1_U4
G1_UVM4_15mL	50,00 microgram absolute quantitation analysis			0,75 microgram WGr1_5_1
G1_532_15mL	50,00 microgram absolute quantitation analysis			0,15 microgram WGr1_5_2
G1_532_15mL	50,00 microgram absolute quantitation analysis			0,03 microgram WGr1_5_3

Source Name	Parameter [sample volume]	Parameter [injection volume]	Parameter [measurement model#]	Parameter [measurement duration#4]	Raw Data File
WGr2_U1	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr2_U1.wiff
WGr2_U2	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr2_U2.wiff
WGr2_U3	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr2_U3.wiff
WGr2_U4	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr2_U4.wiff
WGr2_5_1	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr2_5_1.wiff
WGr2_5_2	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr2_5_2.wiff
WGr2_5_3	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr2_5_3.wiff
WGr2_5_4	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr2_5_4.wiff
WGr2_UF_1	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr2_UF_1.wiff
WGr2_UF_2	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr2_UF_2.wiff
WGr2_UF_3	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr2_UF_3.wiff
WGr2_UF_4	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr2_UF_4.wiff
WGr2_SF_1	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr2_SF_1.wiff
WGr2_SF_2	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr2_SF_2.wiff
WGr2_SF_3	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr2_SF_3.wiff
WGr2_SF_4	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr2_SF_4.wiff
WGr1_U1	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr1_U1.wiff
WGr1_U2	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr1_U2.wiff
WGr1_U3	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr1_U3.wiff
WGr1_U4	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr1_U4.wiff
WGr1_5_1	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr1_5_1.wiff
WGr1_5_2	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr1_5_2.wiff
WGr1_5_3	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr1_5_3.wiff
WGr1_5_4	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr1_5_4.wiff
WGr1_UF_1	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr1_UF_1.wiff
WGr1_UF_2	100,00 microliter	2,50 microliter	1 TripleTOF 6000	87,00 minute	WGr1_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

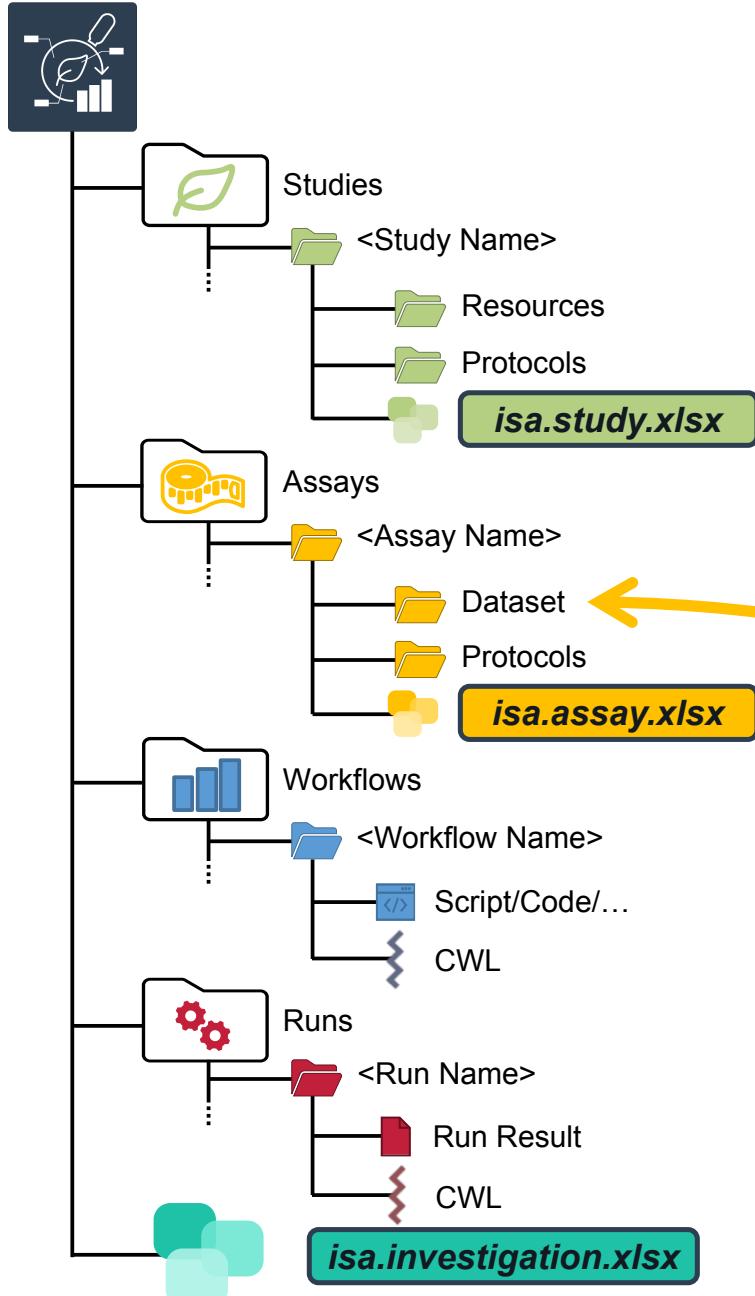
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Source Name	Characteristics [soluble protein content]	Parameter [Quantification method#2]	Parameter [15N Photosynthesis QconCAT mass#4]	Sample Name								
2	G2_UVM4_15mL	50,00 microgram	absolute quantitation analysis										0,75 microgram WCGr2_U1
3	G2_UVM4_15mL	50,00 microgram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

isa.study.xlsx

Samples

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

isa.assay.xlsx



The screenshot shows three Excel spreadsheets corresponding to the components in the IDMO diagram:

- isa.study.xlsx**: A table with columns A through M. It lists various source names and their characteristics, parameters, and sample names.
- isa.assay.xlsx**: A table with columns A through T. It lists source names, parameters (sample volume and injection volume), and raw data files (TripleTOF 6600 and WIFF files).
- isa.investigation.xlsx**: A table with columns A through T. It lists source names, parameters (measurement duration), and raw data files (TripleTOF 6600 and WIFF files). A yellow box highlights the "Raw data" section.

isa.study.xlsx Data Preview:

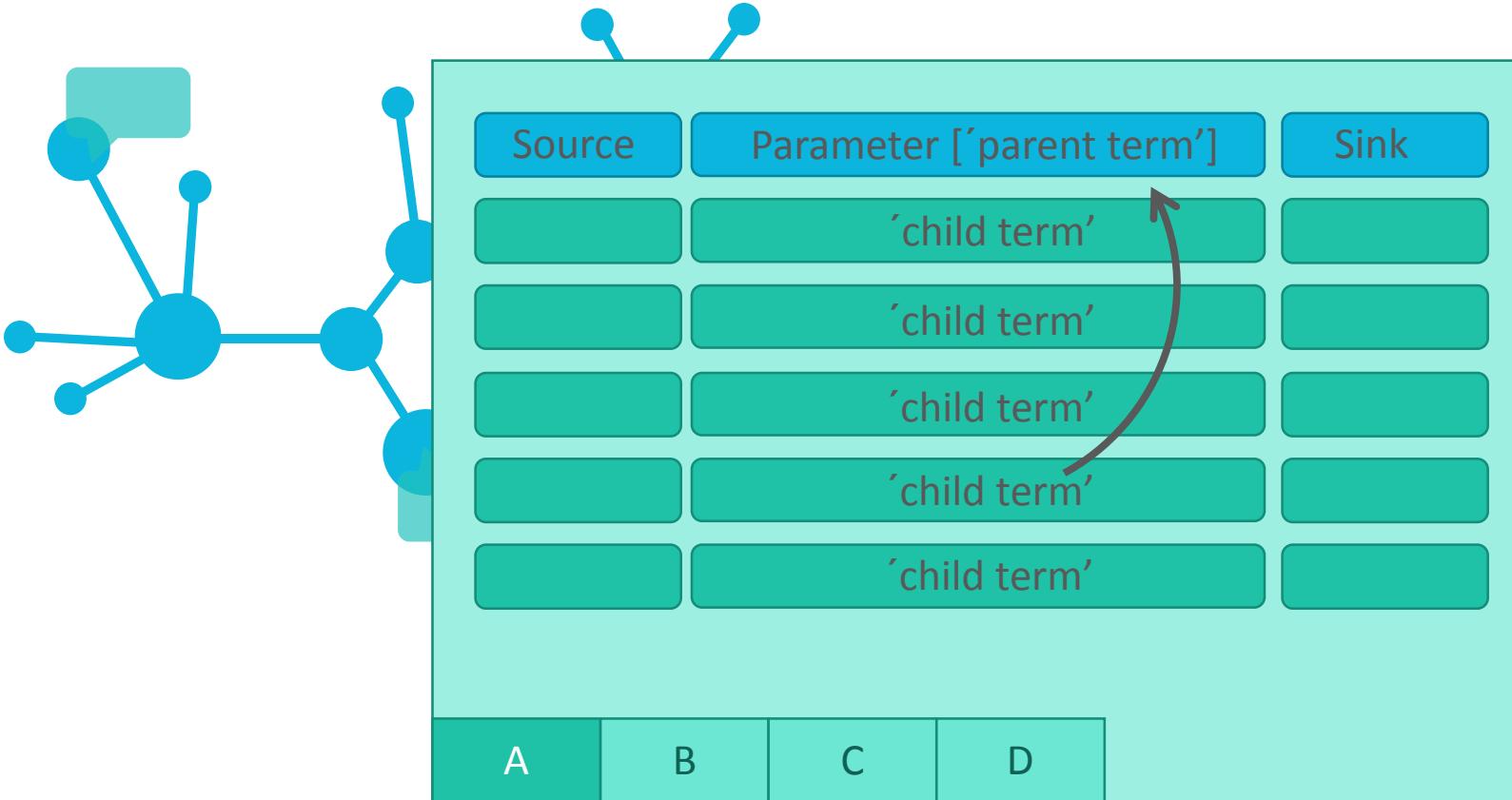
A	B	C	D	E	F	G	H	I	J	K	L	M
1	Source Name	Characteristics [soluble protein content]	Parameter [Quantification method#2]	Parameter [15N Photosynthesis QconCAT mass#4]	Sample Name							
2	G2_UVM4_15mL	50,00 microgram absolute quantitation analysis			0,75 microgram WCGr2_U1							
3	G2_UVM4_15mL	50,00 microgram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

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

A	B	C	D	E	F	G	H	I	J	K	L	M
1	Source Name	Parameter [sample volume]	Parameter [injection vol.									
2	WCGr2_U1	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
3	WCGr2_U2	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
4	WCGr2_U3	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
5	WCGr2_U4	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
6	WCGr2_5_1	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
7	WCGr2_5_2	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
8	WCGr2_5_3	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
9	WCGr2_5_4	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
10	WCGr2_SF_1	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
11	WCGr2_SF_2	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
12	WCGr2_SF_3	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
13	WCGr2_SF_4	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
14	WCGr1_U1	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
15	WCGr1_U2	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
16	WCGr1_U3	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
17	WCGr1_U4	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
18	WCGr1_5_1	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
19	WCGr1_5_2	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
20	WCGr1_5_3	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
21	WCGr1_5_4	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
22	WCGr1_UF_1	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
23	WCGr1_UF_2	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
24	WCGr1_UF_3	100,00 microliter	2,50 microliter	1	TripleTOF 6600							
25	WCGr1_UF_4	100,00 microliter	2,50 microliter	1	TripleTOF 6600							

Swate

Annotation by flattening the knowledge graph



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

Annotation principle

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

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

Adding new building blocks (columns)

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

On the right side of the screen, the "Swate" application is open. It displays a "Building Blocks" interface with a sidebar containing "Add annotation building blocks (columns) to the annotation table." Below this, there is a list of instrument models:

- instrument model MS:1000031
- Instrument Model NCITC177610
- instrument MS:1000463
- instrument EFO:0000548
- Agilent instrument model MS:1000490

A search bar at the bottom of the Swate window says "...Can't find the term you are looking for? Try...".

At the bottom of the Excel window, there is a note: "Parameter columns describe steps in your experimental workflow, e.g. the centrifugation time or the temperature used for your assay. Multiple Parameter columns form a protocol. There is no limitation for the number of Parameter columns per table. You can find more information on our website."

Annotation Building Block types

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

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

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

The 'Building Blocks' pane on the right lists various parameters and instruments, with a specific entry for 'MS:1000031' highlighted. The status bar at the bottom right indicates 'Swate Release Version 0.6.2'.

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

Ontology term search

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

A modal window titled "Swate" is open on the right side of the screen. The title bar of the window says "SWATE". The main area of the window is titled "Ontology term search" and contains the text "Search for an ontology term to fill into the selected field(s)". Below this, there is a search input field with the value "instrument n 6130" and a dropdown menu showing "6130 Quadrupole MS:1000470 LC/MS". There is also a link "Advanced Search!" and a note "Still can't find what you need? Get in contact with us!"

The Excel ribbon at the top includes tabs for File, Home, Insert, Draw, Page Layout, Formulas, Data, Review, View, Help, and Table Design. The Data tab is currently selected. The ribbon also features various icons for data manipulation, such as Sort & Filter, Text to Columns, Flash Fill, Remove Duplicates, Consolidate, Relationships, Manage Data Model, What-If Analysis, Forecast Sheet, Group, Ungroup, Subtotal, Outline, and Swate.

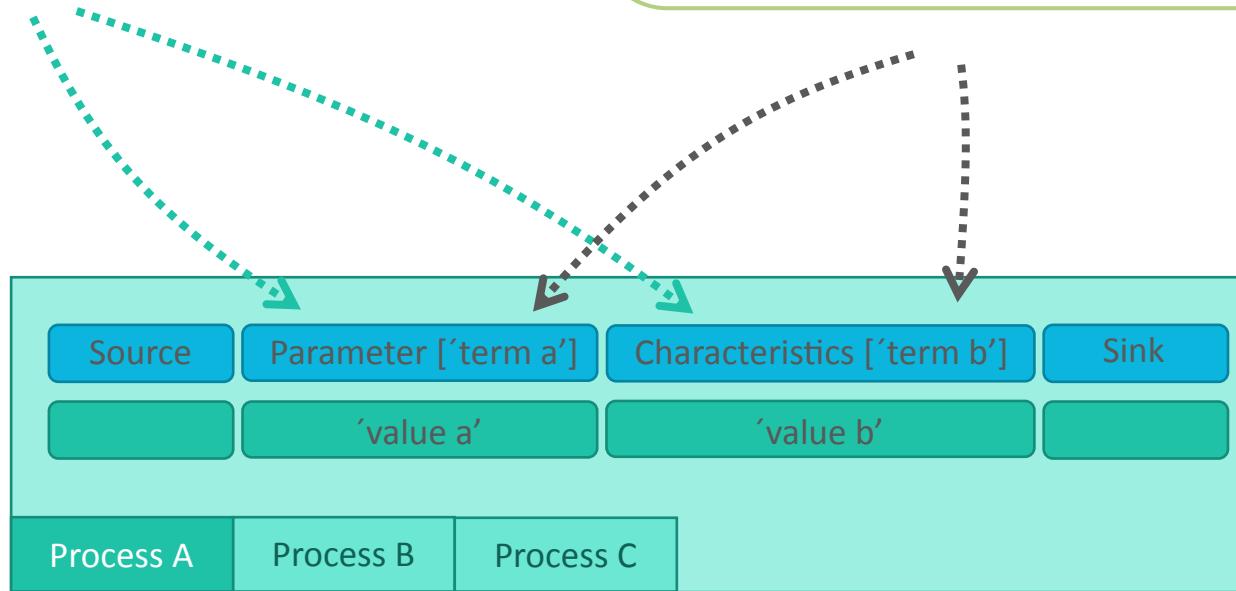
Fill your table with ontology terms

This screenshot shows a Microsoft Excel spreadsheet titled "6130 Quadrupole LC/MS" open in the "Swate" add-in. The spreadsheet contains a large dataset of experimental runs, with columns including Source Name, Protocol Type, Characteristic [sample label], Factor [temperature], Parameter [instrument model], Component [software], and Sample Name.

The "Swate" ribbon tab is selected, providing various data management and analysis tools. A floating "Ontology term search" window is visible, prompting the user to search for an ontology term to fill a selected field. The search bar contains the text "6130 Quadrupole LC/MS".

The bottom right corner of the screen displays the "Swate Release Version 0.6.2" watermark.

Hierarchical combination of ontologies



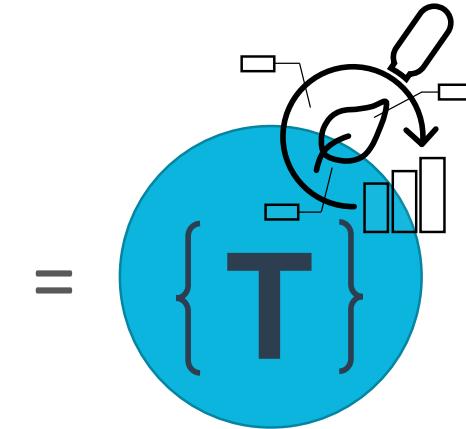
[isa.study.xlsx or isa.assay.xlsx](#)

Swate templates

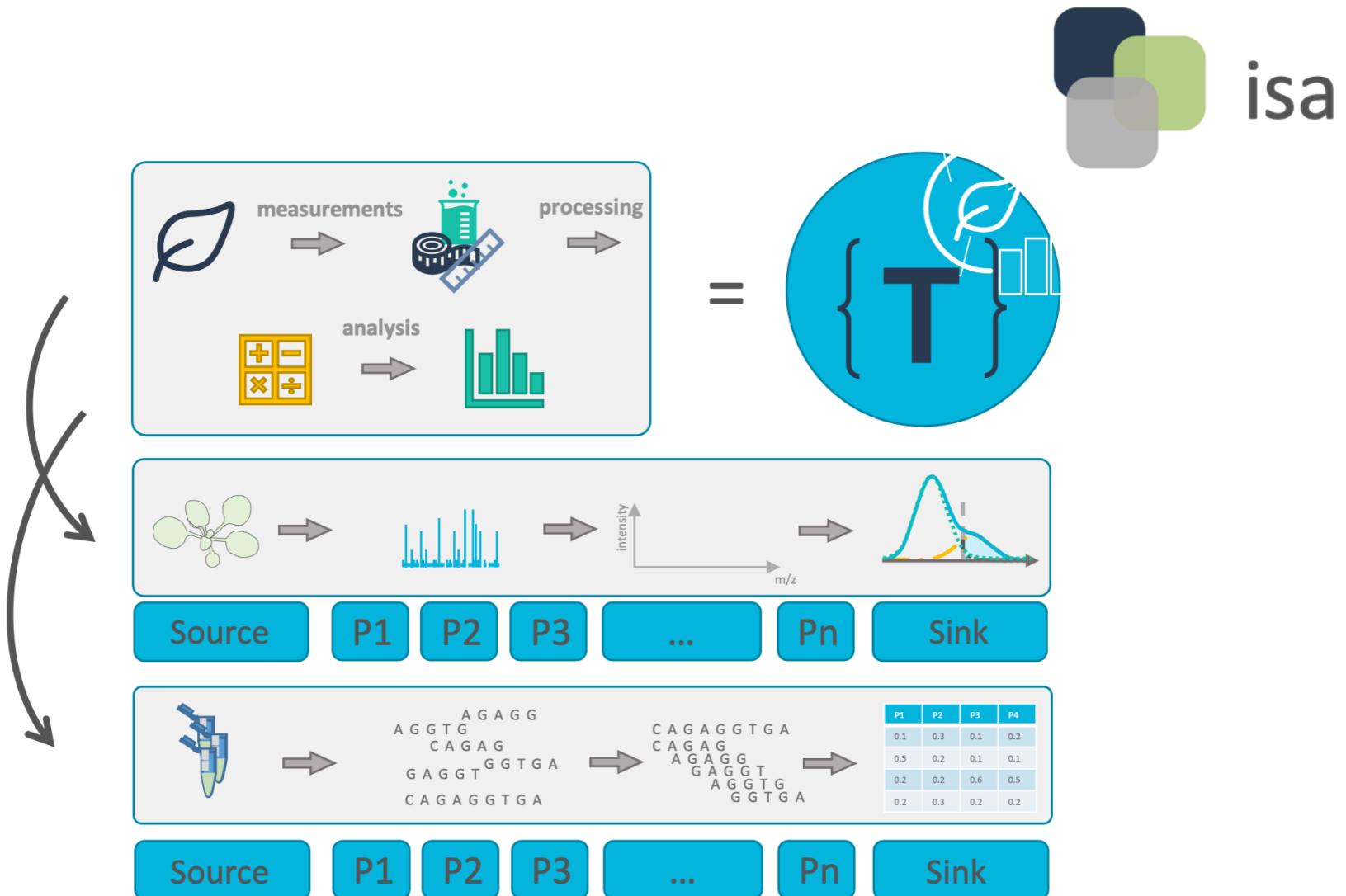
Checklists and Templates

Source	Parameter ['parent term']	Sink
	'child term'	

A B C D

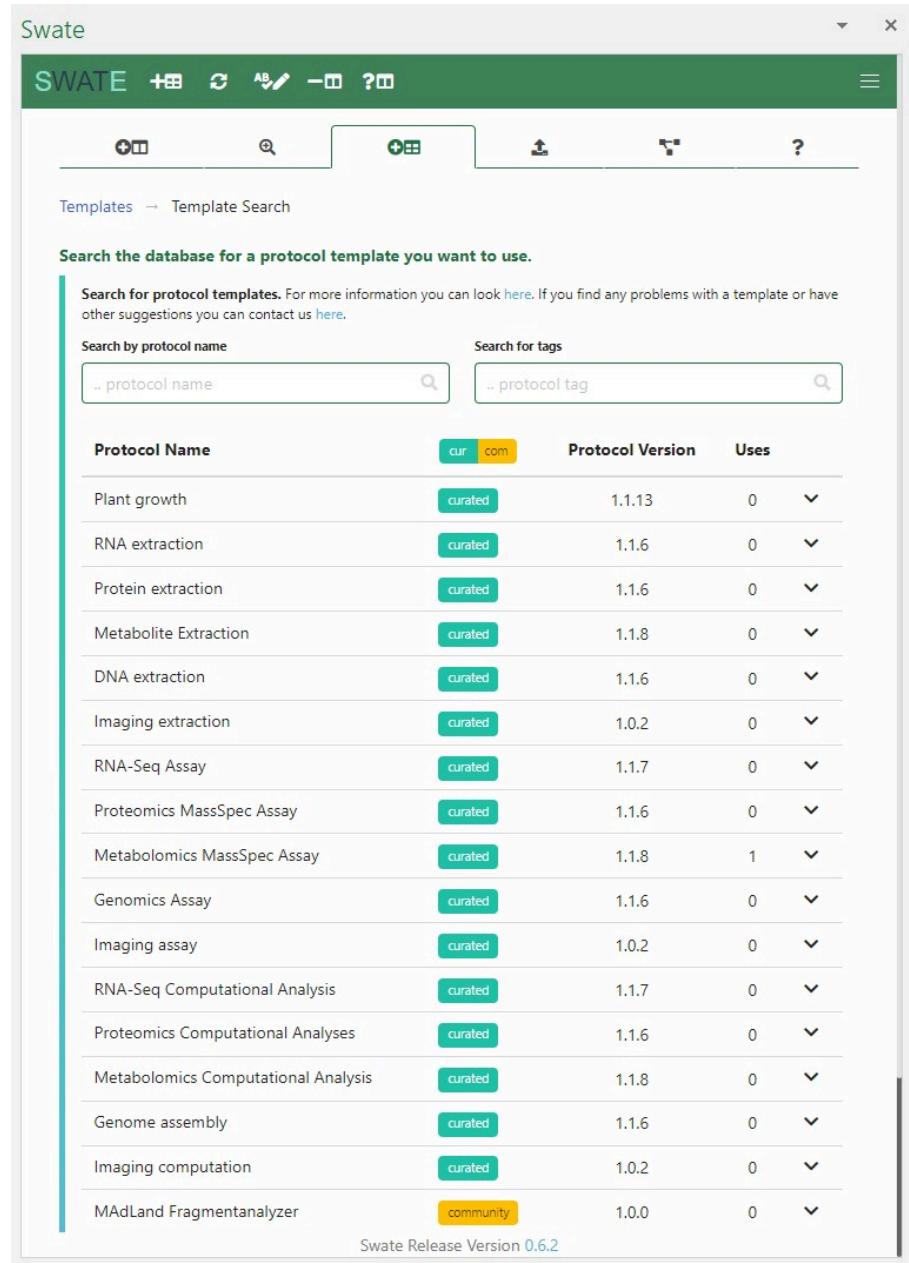


Realization of lab-specific metadata templates



Directly import templates via Swate

- DataPLANT curated
- Community templates



The screenshot shows the Swate software interface. At the top, there's a green header bar with the word "Swate" and various icons. Below the header is a toolbar with buttons for creating new templates, searching, and other functions. The main area is titled "Templates → Template Search" and contains a search bar and a "Search by protocol name" input field. To the right of these fields is a "Search for tags" input field. The main content area displays a table of protocol templates:

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

At the bottom right of the main area, it says "Swate Release Version 0.6.2".

Contributors

Slides presented here include contributions by

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

Swate hands-on

Goals

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

Check Swate installation

 Make sure [Swate is installed](#):

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

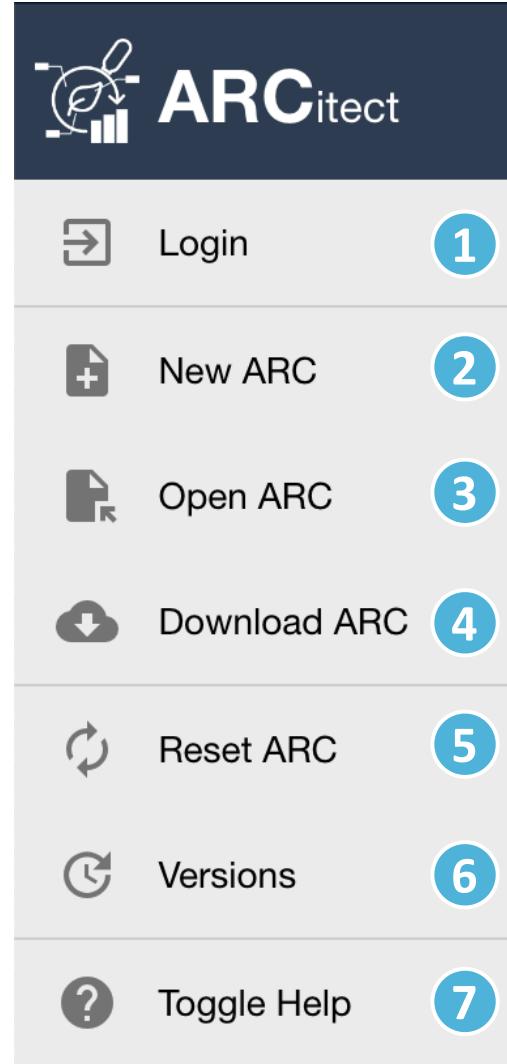
 Alternatively, you can use [Swate standalone](#)

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

Have a simple text editor ready

- Windows Notepad
- MacOS TextEdit

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



Download the demo data

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

Download the demo data

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

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

The screenshot shows the ARCitect application interface. On the left is a sidebar with the following menu items:

- New ARC
- Open ARC
- Download ARC
- Save ARC
- Explorer
- Commit
- DataHUB Sync
- History

In the center, there is a large magnifying glass icon with the text "Open ARC" below it. To the right, a search results page titled "Download ARC" is displayed. The search bar contains the identifier "Talinum-CAM-Photosynthesis". Below the search bar, there are checkboxes for "LFS" and "Host git.nfdi4plants.org". A green "C" button is also present. The search results list one item: "Talinum-CAM-Photosynthesis [2023-10-11T09:24:10.208Z] Teaching". To the right of this item are a green search icon and a blue download icon.

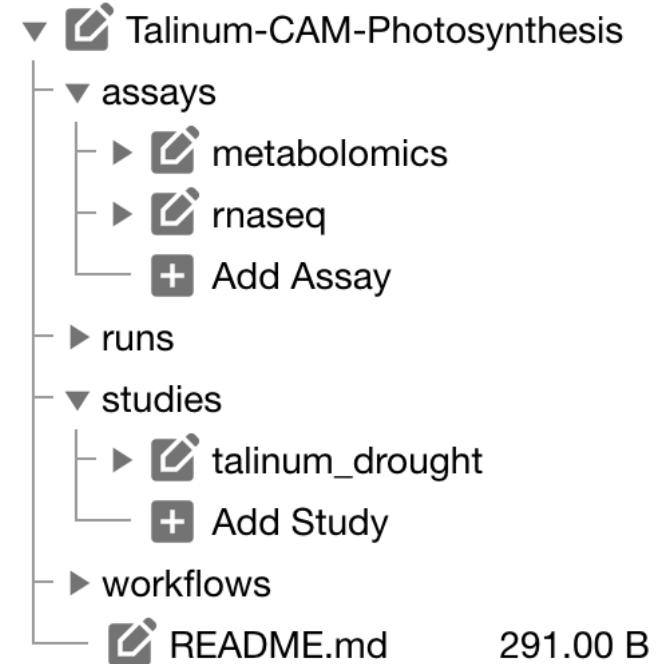
 This is basically the ARC we created last session.

Where we left off last time

 Initiated an ARC

 Structured and ...

 Shared with collaborators



Today we want to

 ... annotate the experimental data

Swate hands-on with demo data

Swate Overview

The screenshot shows a Microsoft Excel spreadsheet titled "toolTalk.xlsx" with a table named "Tabelle1". The table has columns labeled "Source Name", "Characteristics [sample label]", "Factor [temperature unit]", and "Data File Name". The "Characteristics" column contains values like "Heat_15A_OD_R1", "Heat_15A_OD_R2", etc. The "Factor" column contains values like "32.00 degree Celsius", "4.00 degree Celsius", etc. The "Data File Name" column contains file names such as "Heat_15A_OD_R1.wiff", "Heat_15A_OD_R2.wiff", etc.

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

- A callout labeled "source" points to the header of the first column.
- A callout labeled "new parameter" points to the "Parameter" dropdown in the Swate ribbon.
- A callout labeled "otation building block selection" points to the "otation building block selection" section in the Swate ribbon.
- A callout labeled "factor" points to the "Factor" column header.
- A callout labeled "characteristic" points to the "Characteristics" column header.
- A callout labeled "datafile / sample" points to the "Data File Name" column header.

The Swate ribbon is visible at the top, showing tabs for Home, Insert, Page Layout, Formulas, Data, Review, View, Developer, Help, and Table Design. The "Developer" tab is selected. The "Table Design" tab is also visible. The "Swate" ribbon tab is open, showing sections for "Add annotation building blocks (columns) to the annotation table.", "Add/Update unit", and "More about Parameter:".

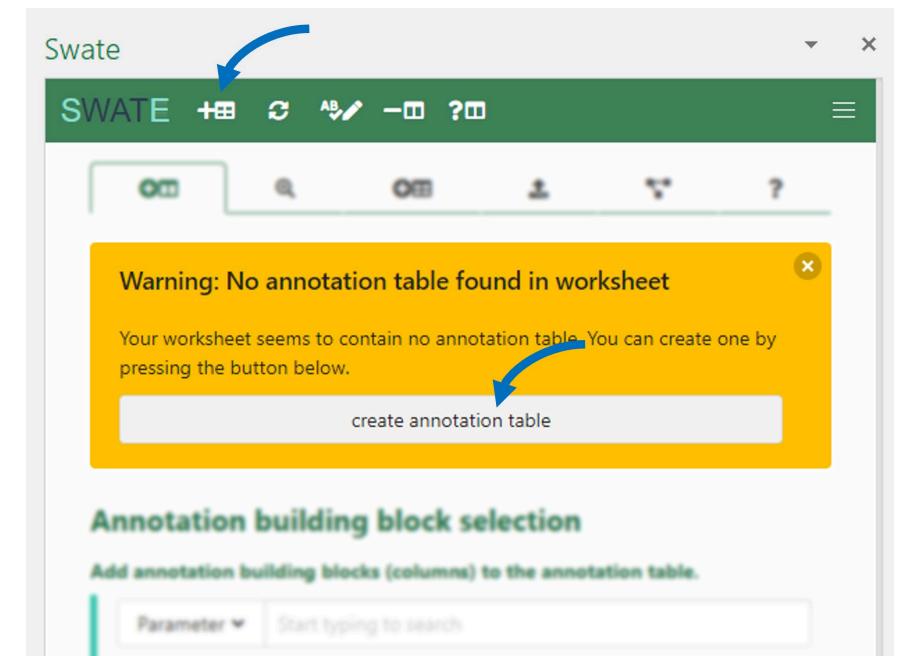
Let's annotate the plant samples first

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

Create an annotation table

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

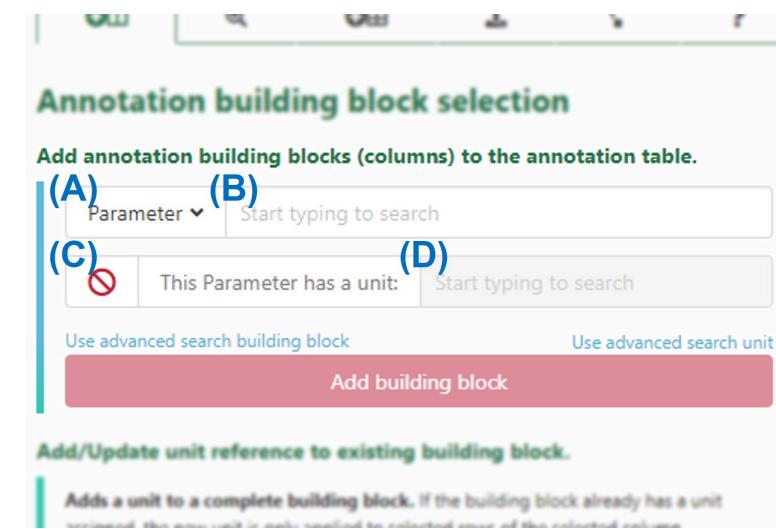
- 💡 Each table is by default created with one input (**Source Name**) and one output (**Sample Name**) column
- 💡 Only one annotation table can be added per Excel sheet



Add a building block

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

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



Annotation building block selection

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

(A) Parameter (B) Start typing to search

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

Use advanced search building block Use advanced search unit

Add building block

Add/Update unit reference to existing building block.

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

Insert values to annotate your data

1. Navigate to the *Terms* tab in the Navbar
2. In the annotation table, select any number of cells below **Characteristic**
[organism]
3. Click into the search field in Swate.
 - 💡 You should see **organism** showing in a field in front of the search field
 - 💡 The search will now yield results related to **organism**
4. In the search field, search for "Talinum fruticosum"
5. Select the first hit and click **Fill selected cells with this term**

Add a building block with a unit

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

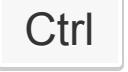
 This adds four columns to your table, one visible and **three** hidden.

Insert unit-values to annotate your data

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

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

Showing ontology reference columns

Hold  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 microeinsteins per square meter per second	
2	Talinum fruticosum	425 microeinsteins per square meter per second	
3	Talinum fruticosum	425 microeinsteins per square meter per second	
4	Talinum fruticosum	425 microeinsteins per square meter per second	
5	Talinum fruticosum	425 microeinsteins per square meter per second	
6	Talinum fruticosum	425 microeinsteins per square meter per second	

1

Hiding ontology reference columns

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

Exercise



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

Add a factor building block

1. In the *Building Blocks* tab, select *Factor*, search for `watering exposure` and select the term with id `PEC0:0007383`.
 2. Click `Add building block`.
 3. Add the drought treatment ("no water for 12 days", "re-water for 2 days") to the respective samples
-  There are different options to add the drought treatment.

Link the protocol to the isa table

1. In the *Building Blocks* tab, select *Protocol Columns* -> *Protocol REF*.
 2. Click  .
 3. Add the name of the protocol file (plant_material.txt) to the *Protocol REF* column.
-  This allows you to reference the free-text, human-readable protocol.

Fill out source name and sample name

Transfer the sample ids from the protocol.

1. Invent names for **Source Name** (we do not have this information)
2. Use the sample names (DB_*) as **Sample Name**

Let's annotate the RNA Seq data

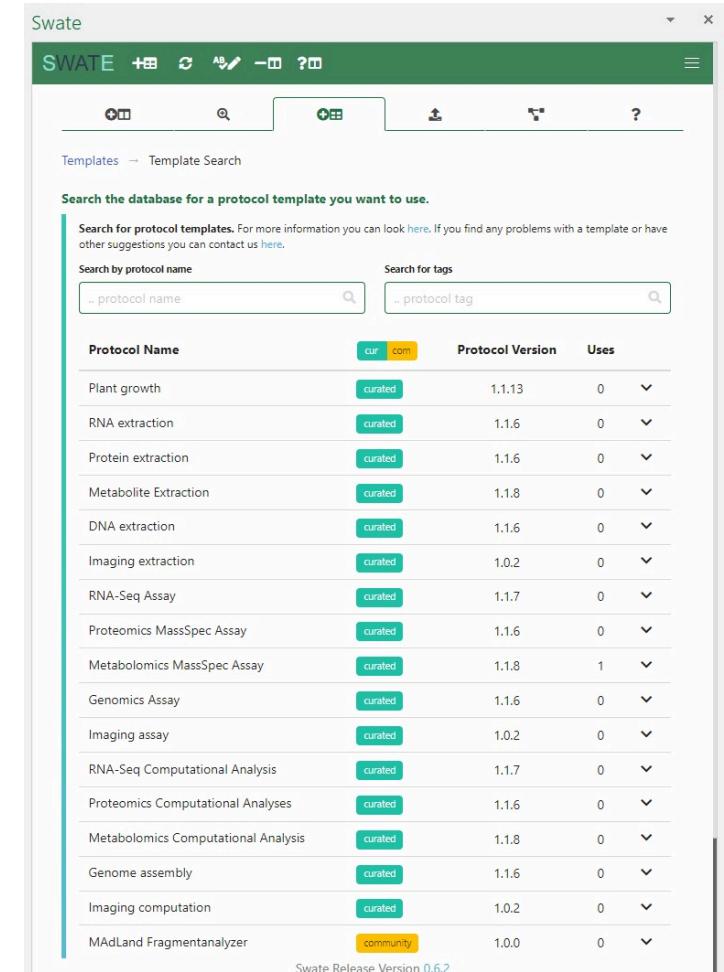
1. Navigate to the demo ARC.
2. Open the lab notes `assays/rnaseq/protocols/RNA_extraction.txt` in a text editor.
3. Open the empty `assays/rnaseq/isa.assay.xlsx` workbook in Excel.

Use a template

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

 Here you can find community created workflow annotation templates

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



The screenshot shows the Swate software interface with the title "Swate" at the top. Below the title is a toolbar with various icons. The main area is titled "Template Search" and contains a search bar with placeholder text "Search the database for a protocol template you want to use." and two search fields: "Search by protocol name" and "Search for tags". Below these fields is a table titled "Protocol Name" with columns for "Protocol Name", "Protocol Version", and "Uses". The table lists several templates, each with a status indicator (cur or com) and a dropdown arrow. The listed templates include: Plant growth, RNA extraction, Protein extraction, Metabolite Extraction, DNA extraction, Imaging extraction, RNA-Seq Assay, Proteomics MassSpec Assay, Metabolomics MassSpec Assay, Genomics Assay, Imaging assay, RNA-Seq Computational Analysis, Proteomics Computational Analyses, Metabolomics Computational Analysis, Genome assembly, Imaging computation, and MADLand Fragmentanalyzer. At the bottom right of the table, it says "Swate Release Version 0.6.2".

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

Adding / Updating unit references

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

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

Remove building blocks

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

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

New process, new worksheet

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

Exercise



Try to fill the two sheets with the protocol details:

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

Link samples across studies and assays

1. Use the **Sample Name** of studies/talinum_drought/isa.study.xlsx as the **Source Name** to **rna-extraction**.
2. Use the **Sample Name** of **rna-extraction** as the **Source Name** to **illumina-libraries**.



Link dataset files to samples

1. In the *Building Blocks* tab, select *Output Columns -> Raw Data File*.

2. Click **Add building block**.

 You see a warning about a changed output column.

3. Click **Continue**.

4. Go to the *File Picker* tab and click **Pick file names**.

5. Select and open the *fastq.gz files from the dataset folder.

6. Copy / paste them to the **Raw Data File**.

 This allows you to link your samples to the resulting raw data files.

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.

Re-use a protocol (process.json)

1. Open the empty

assays/metabolomics/isa.assay.xlsx) workbook in Excel.

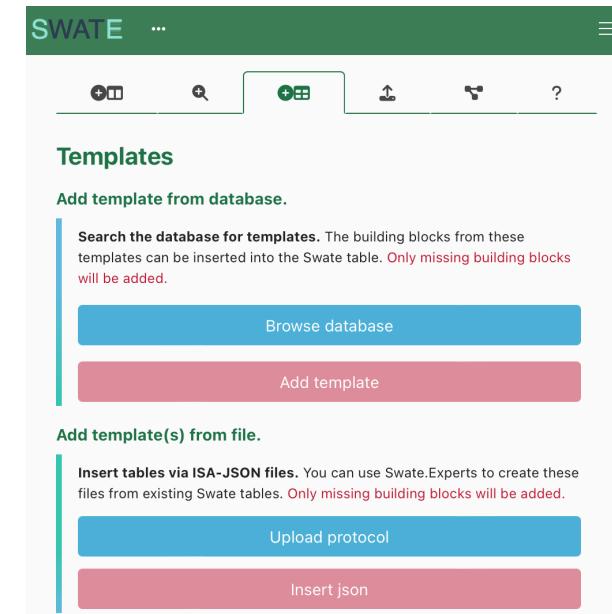
2. Navigate to *Templates* in the Navbar and scroll down to "Add template(s) from file."

3. Click **Upload protocol**

4. Select the file "swate_agilent_gc.json" from the demo data.

5. Click **Insert json**

 This adds not only an empty template, but a filled out table with keys (headers) and values (cells).

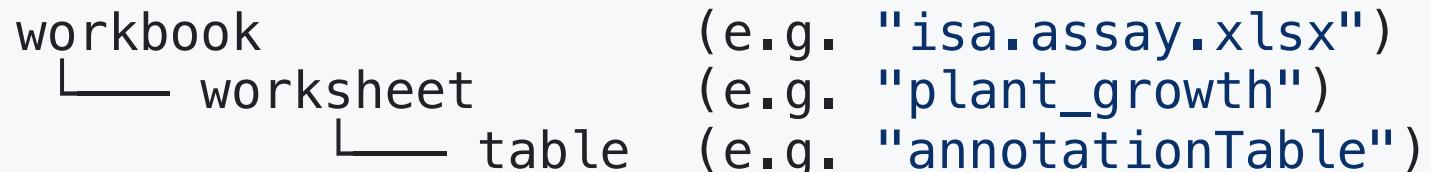


A small detour on "Excel Tables"

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.



Known issues with ARCitect and Swate (Nov 2023)

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

Contributors

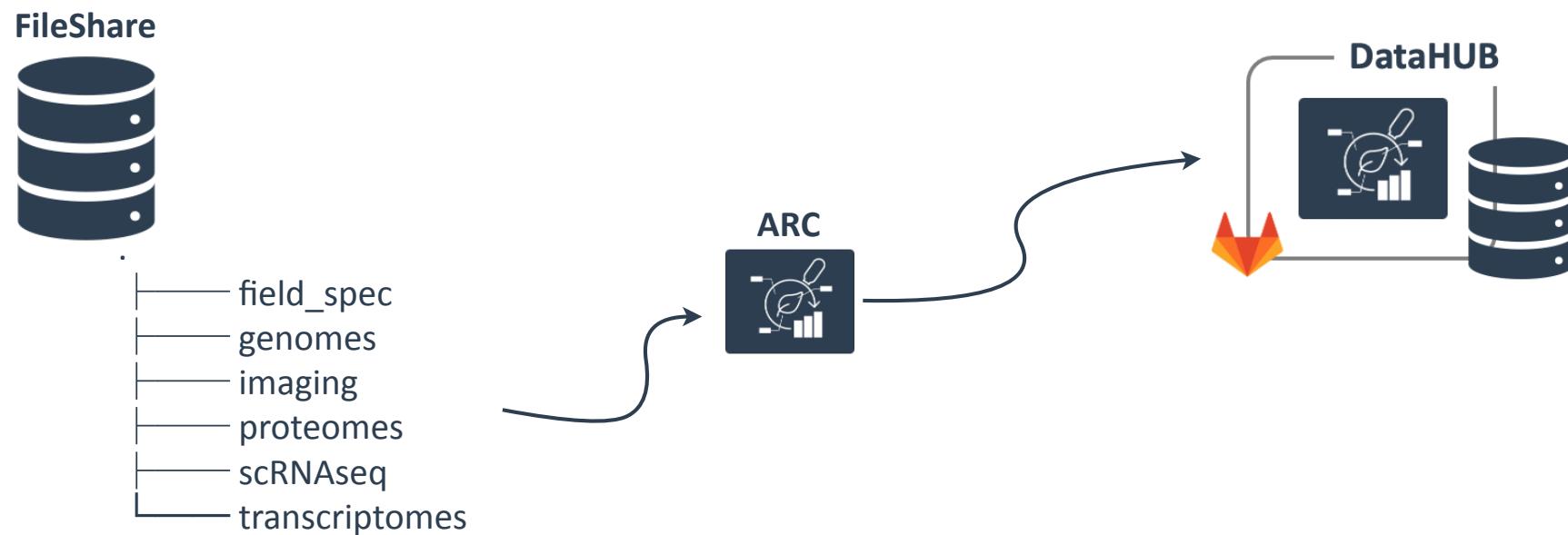
Slides presented here include contributions by

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

The ARC Club

Happy ARCing

Moving from FileShare to DataHUB – *via* ARCs



Assign projects

Rough routine for each project

1. Identify the available data and resources
2. Create the ARC
3. Add metadata and data
4. Share via DataHUB group

Low(er) hanging fruits: published projects

1. Add the authors
2. Add the publication(s)
 - i. Add citation and DOI
 - ii. Add supplemental
 - iii. Convert M&M to *protocols*
3. Reference data in public repositories
4. Add large data (e.g. from file share)
5. Set ARC to **public!**

More challenging ARCs

- (unpublished) left-overs of colleagues who have since moved
- unclear status

Collect / derive as much info about the investigation as possible

MUST haves

Investigation Identifier

Investigation Title

...

Investigation Publication Status

...

Investigation Person Last Name

Investigation Person First Name

- 💡 This and more investigation-level info can be collected in the ARC's `isa.investigation.xlsx`

Create and share the ARC

```
arc init  
arc sync -f -r https://git.nfdi4plants.org/<GroupName>/<InvestigationID>
```

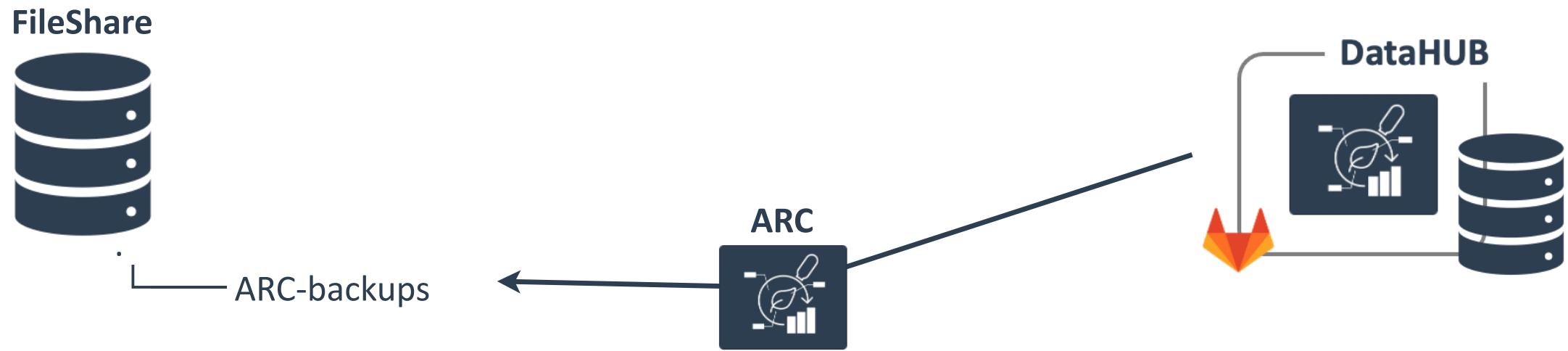
Copy data

1. **Copy** data to the ARC, do not **move** data from original source
(we'll take care of that later)
2. Ideally use `rsync` rather than copying manually
3. Ideally use `md5` or `md5sum` to check for correct file transfer

 Ask the coders for help!

Perspective and administration in the future

Administration / Backup



Structure your current project folder into an ARC

Sketch your laboratory workflows

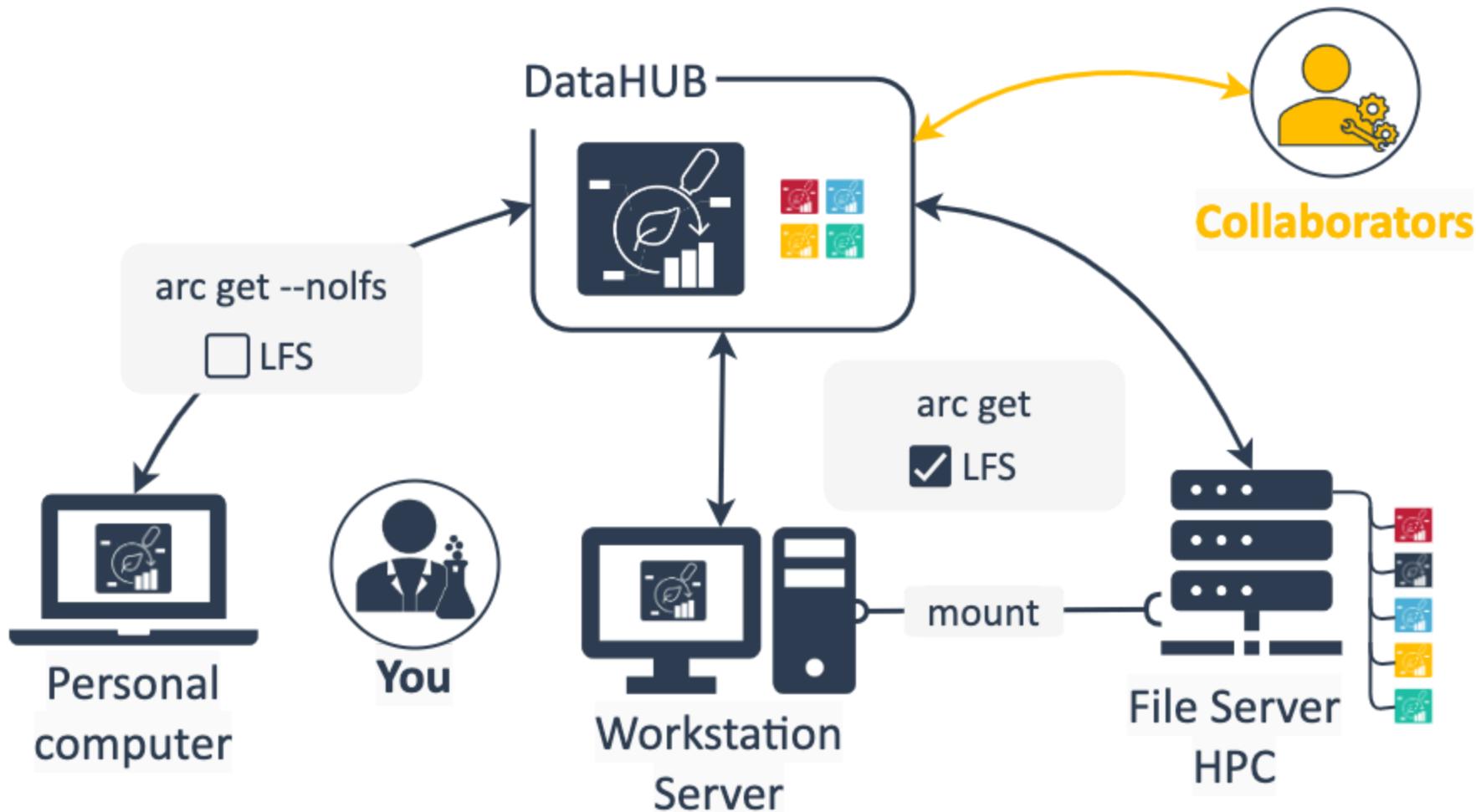
The ARC helps

- to tell, which finding or result originated from which biological experiment
- link dataset files to the individual sample
- to follow a path of *processes* with *inputs* and *outputs*
- allow inputs and outputs want to be reused or reproduced, some of the processes want to be applied to other inputs

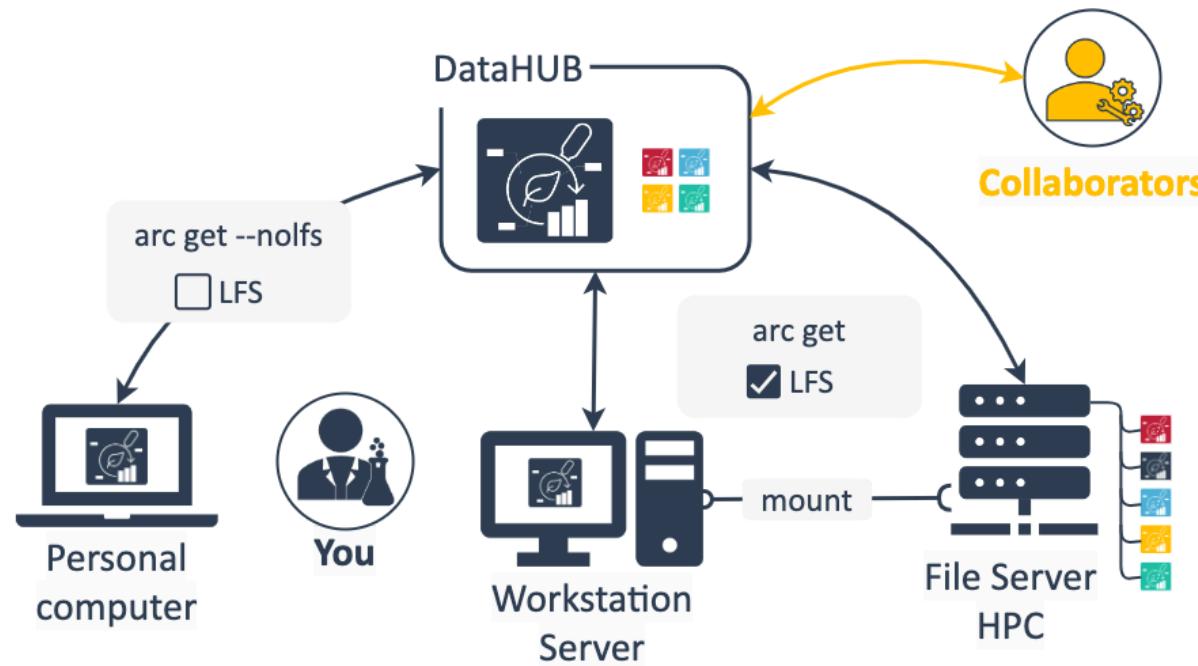
Green-house to gene expression

Where do I store my ARC?

ARC storage and sharing



ARC storage and sharing



- DataHUB as "ground truth" / original clone
- You can sync and communicate all changes to your ARC via the DataHUB

💡 ARCitect and ARC commander provide options to avoid syncing large files (LFS = Large file storage)

- Personal computer
 - work on small files
 - annotate metadata
 - add scripts
 - add protocols
- Workstation / Server
 - work on large files
 - run analyses
- FileShare
 - mount to local machine and arc sync
- HPC
 - direct connection HPC to DataHUB (depends on security settings)
 - or mount to local machine and arc sync from there
