

# Research data management

# The data life cycle



- ▶ Acquiring and analyzing data is fundamental in research to discover new knowledge
- ▶ Successful data management requires to distinguishing between data, information and knowledge

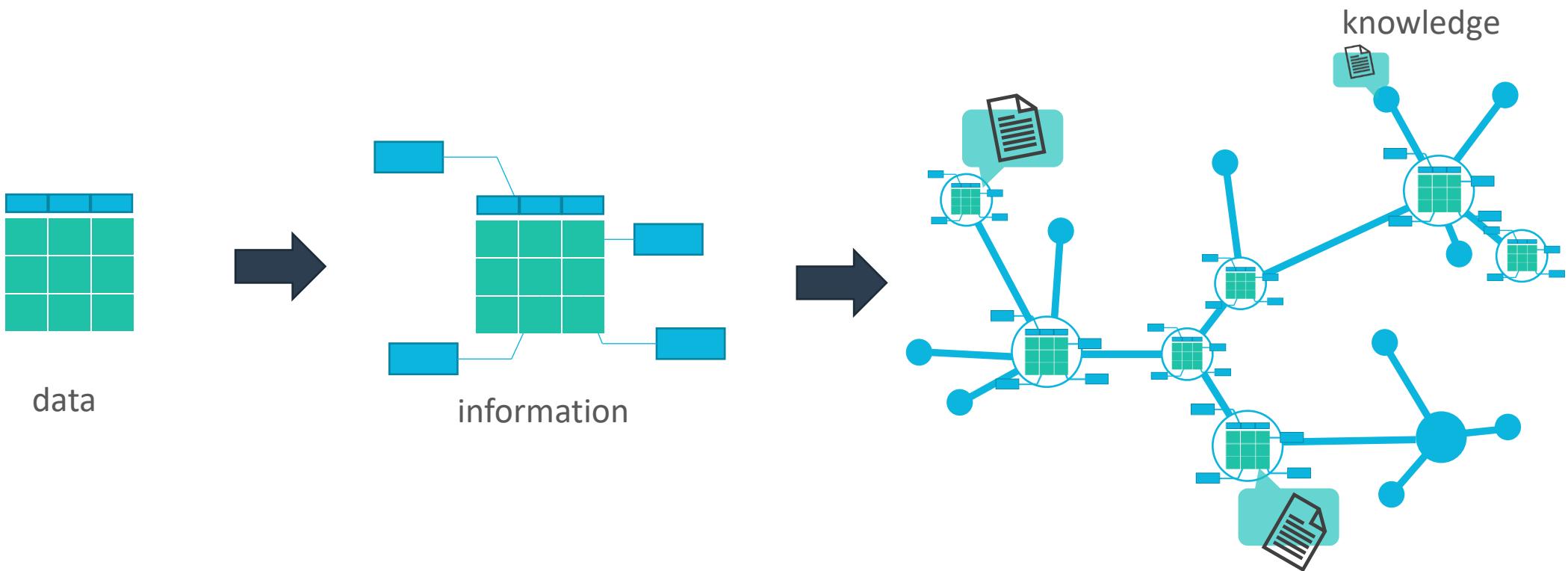
# From data and information to knowledge

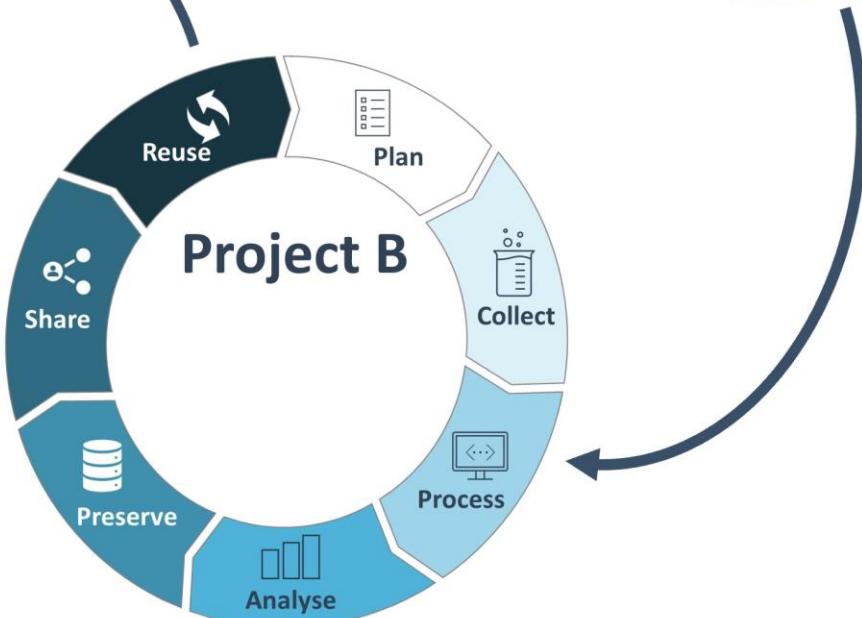
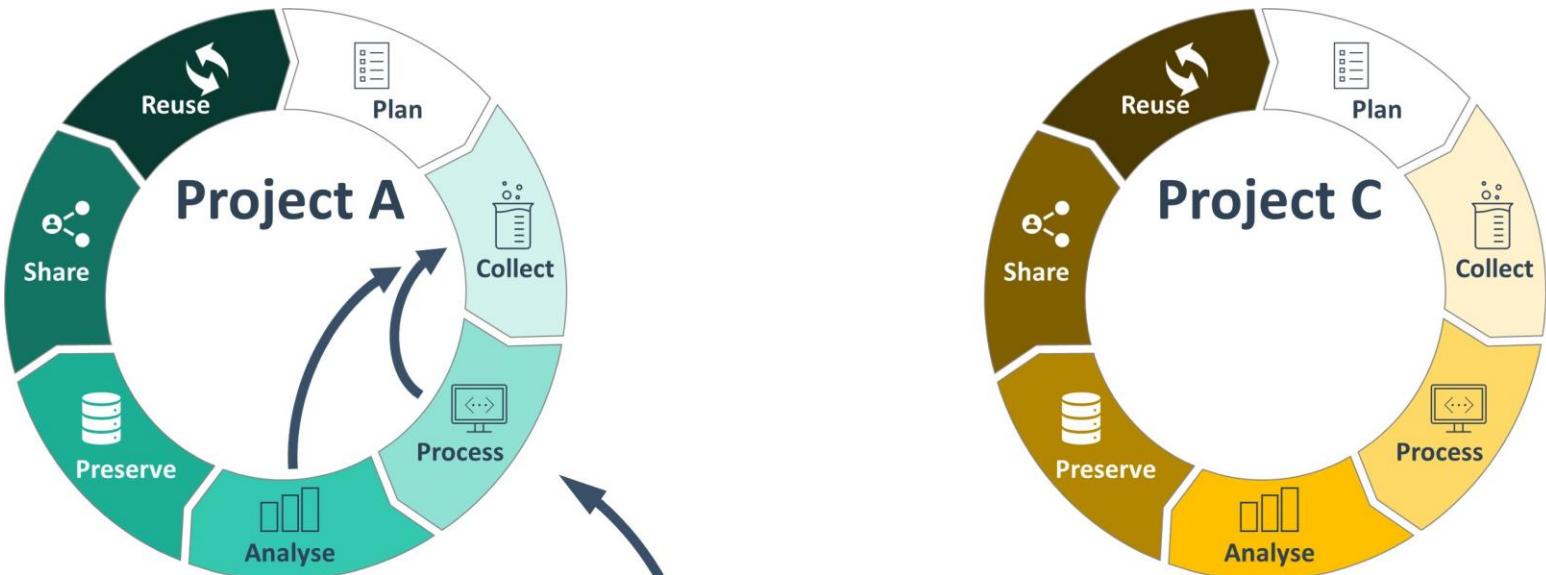


- Connected information becomes knowledge

# Data: a primary product in science

- ▶ Data itself has value for science:
  - ▶ It can be reused to gain further knowledge
  - ▶ New techniques or theories may require it to be re-examined

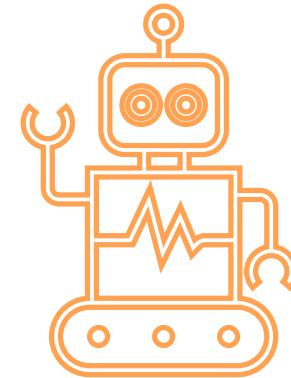
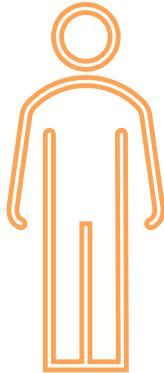




# The FAIR data principles

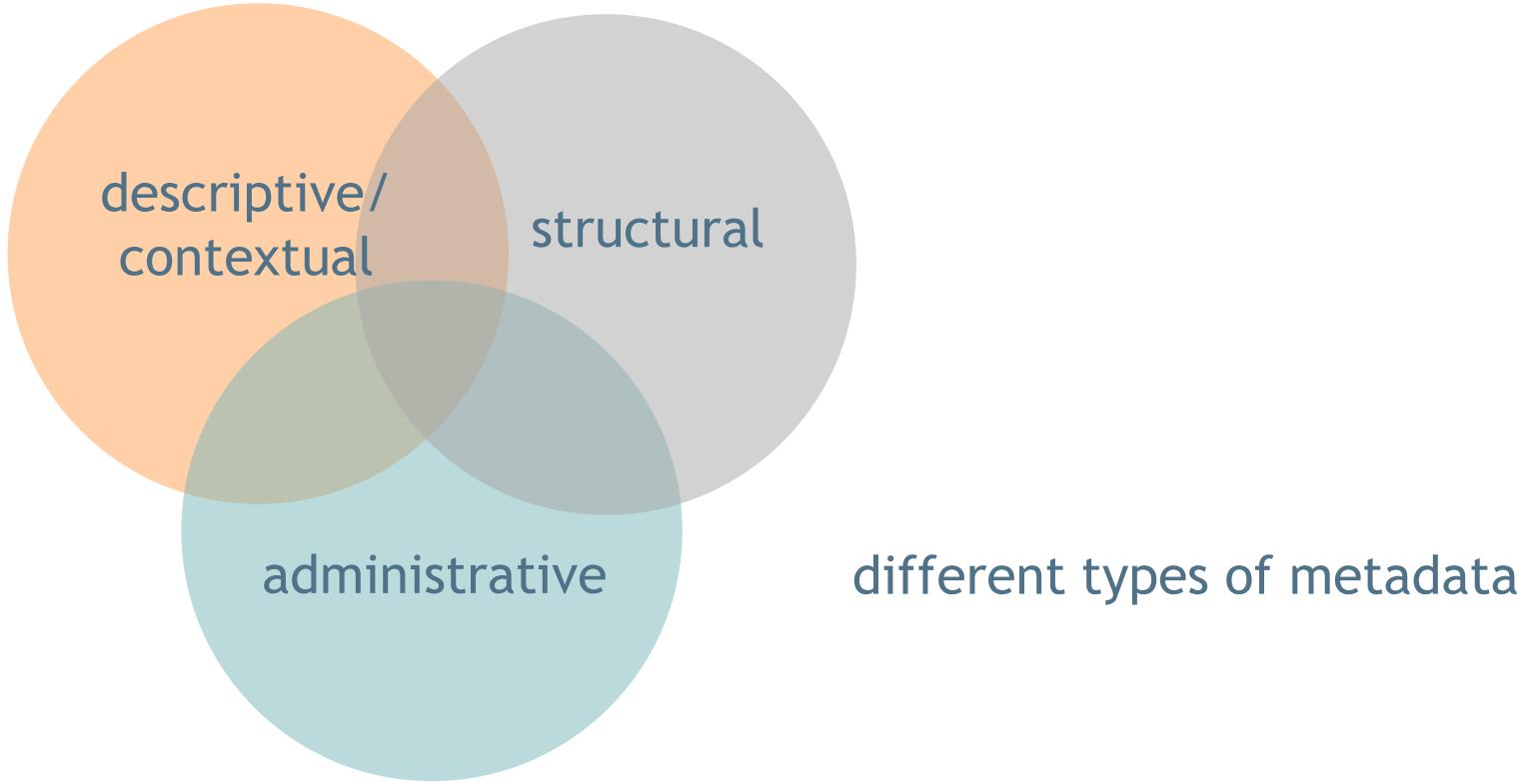
High level principles to make data:

- ▶ **F**indable
- ▶ **A**ccessible
- ▶ **I**nteroperable
- ▶ **R**e-usable



... for humans and machines

# Data context or (meta)data



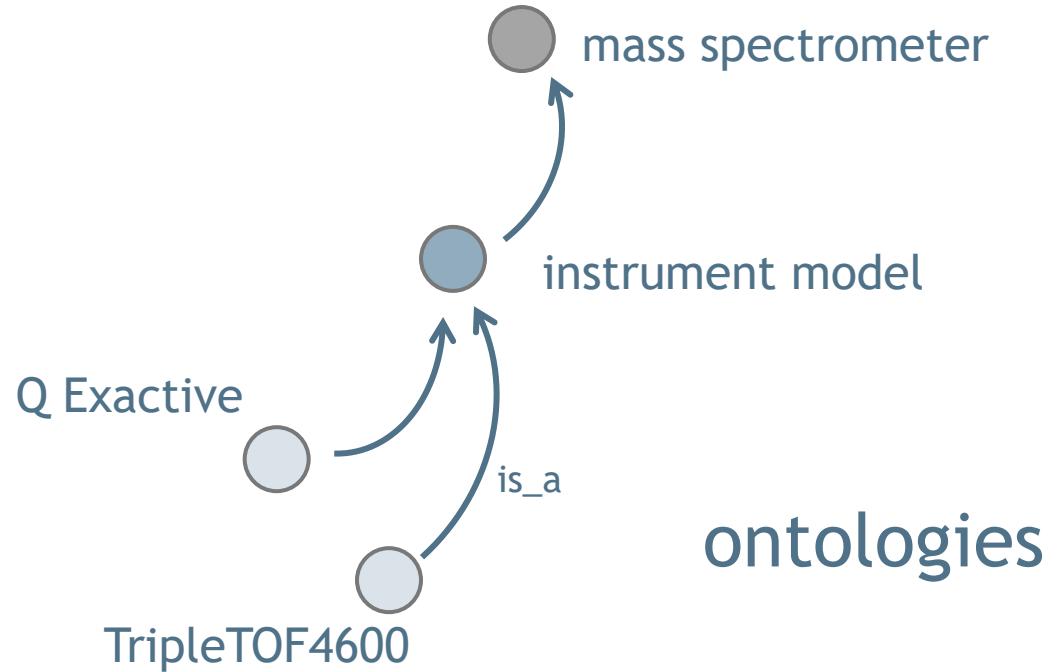
- ▶ Metadata represents data about data
- ▶ Metadata enriches the data with context that makes it usable, and also easier to find and manage

# Controlled vocabularies



dictionaries

=



ontologies

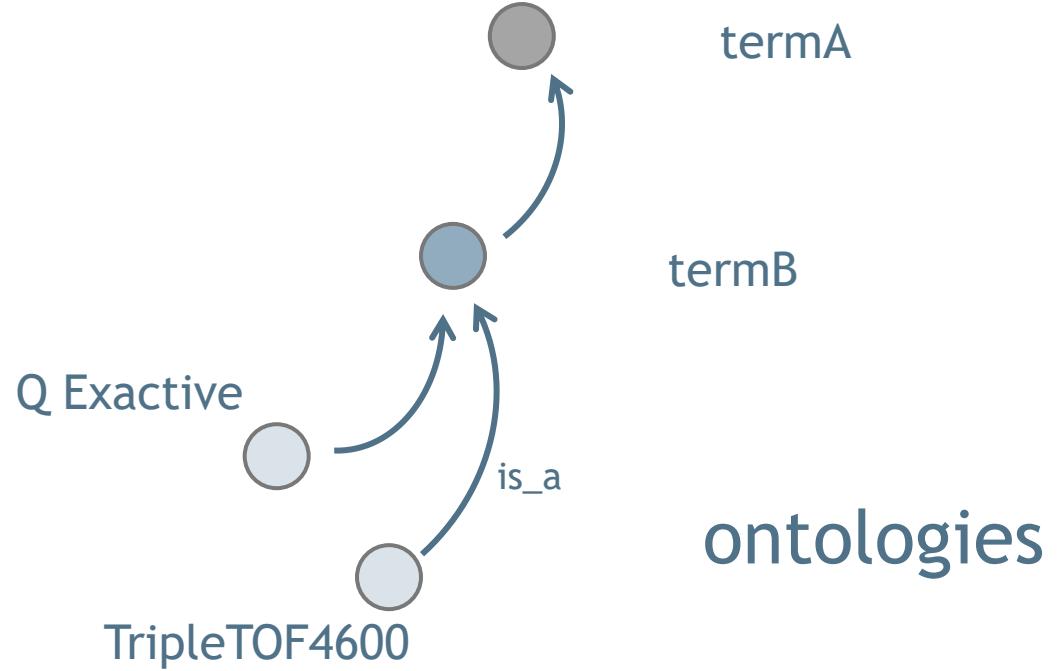
Meta data need to “speak“ the same language

# Controlled vocabularies



dictionaries

=



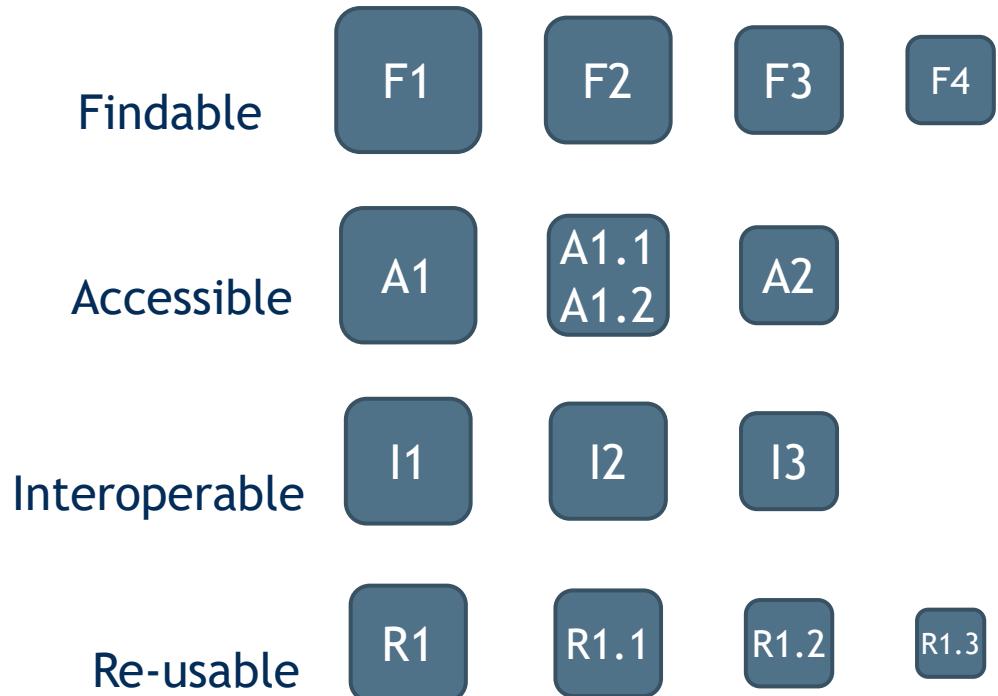
termA

termB

ontologies

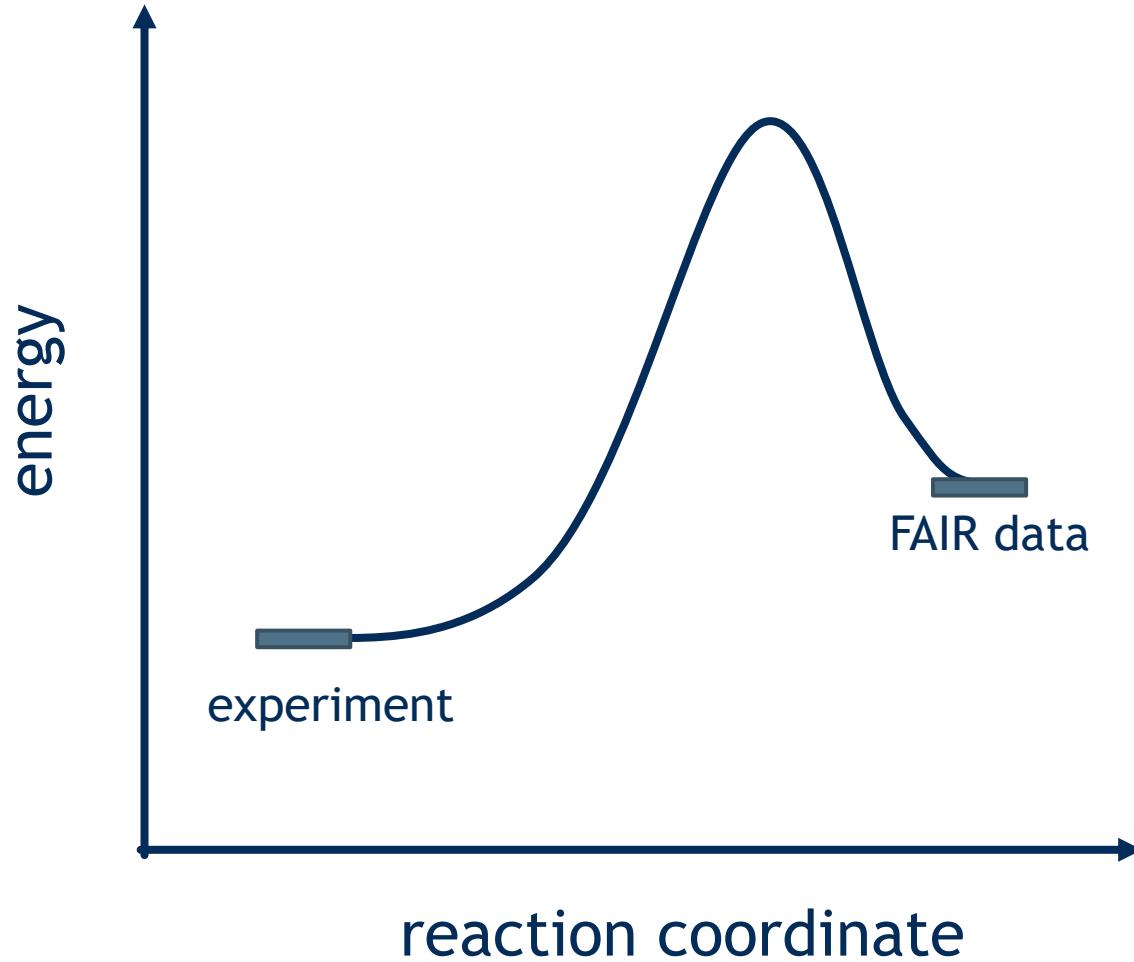
Meta data need to “speak“ the same language

# FAIRification is an incremental process

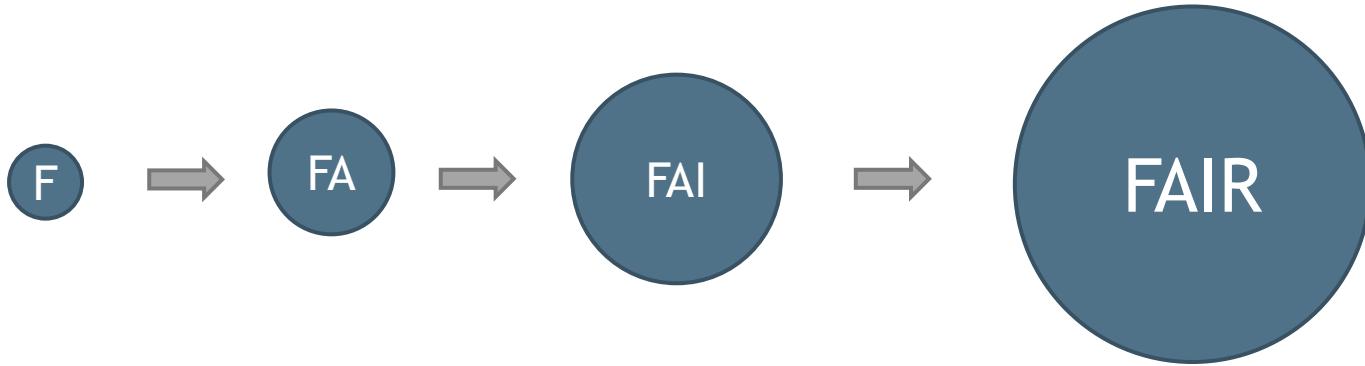


- ▶ FAIRness can be gradually improved
- ▶ FAIRness is a compromise between agile development and scholar stability

# You need to invest energy

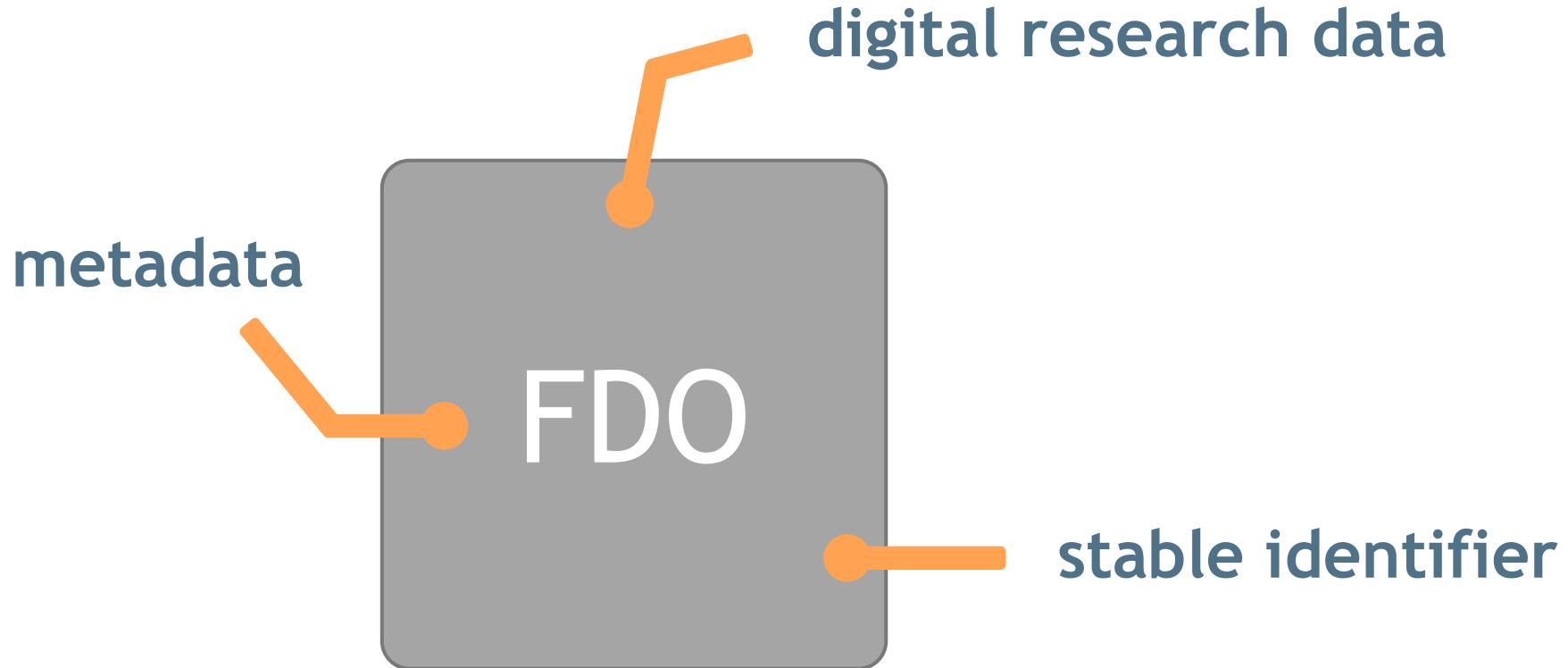


# Do the best you can...



- ▶ FAIRness is a spectrum, and FAIRer is a step forward
- ▶ Reach out for help
- ▶ Things will get easier

# FAIR Digital Objects



# Annotated Research Context

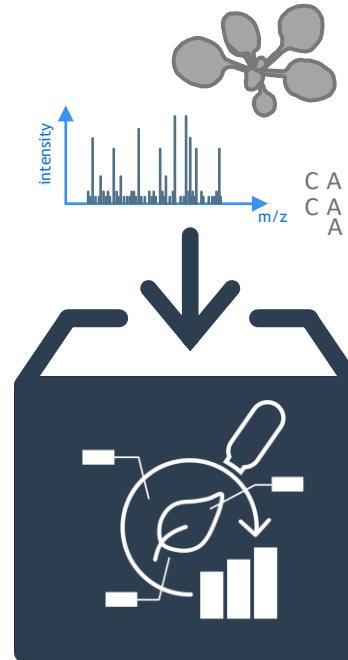
+ ARC Toolchain



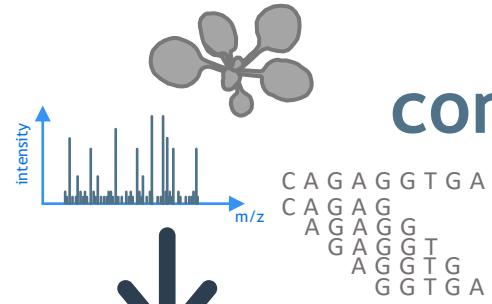
# Bag your research in an ARC *(Annotated Research Context)*



=



experimental data

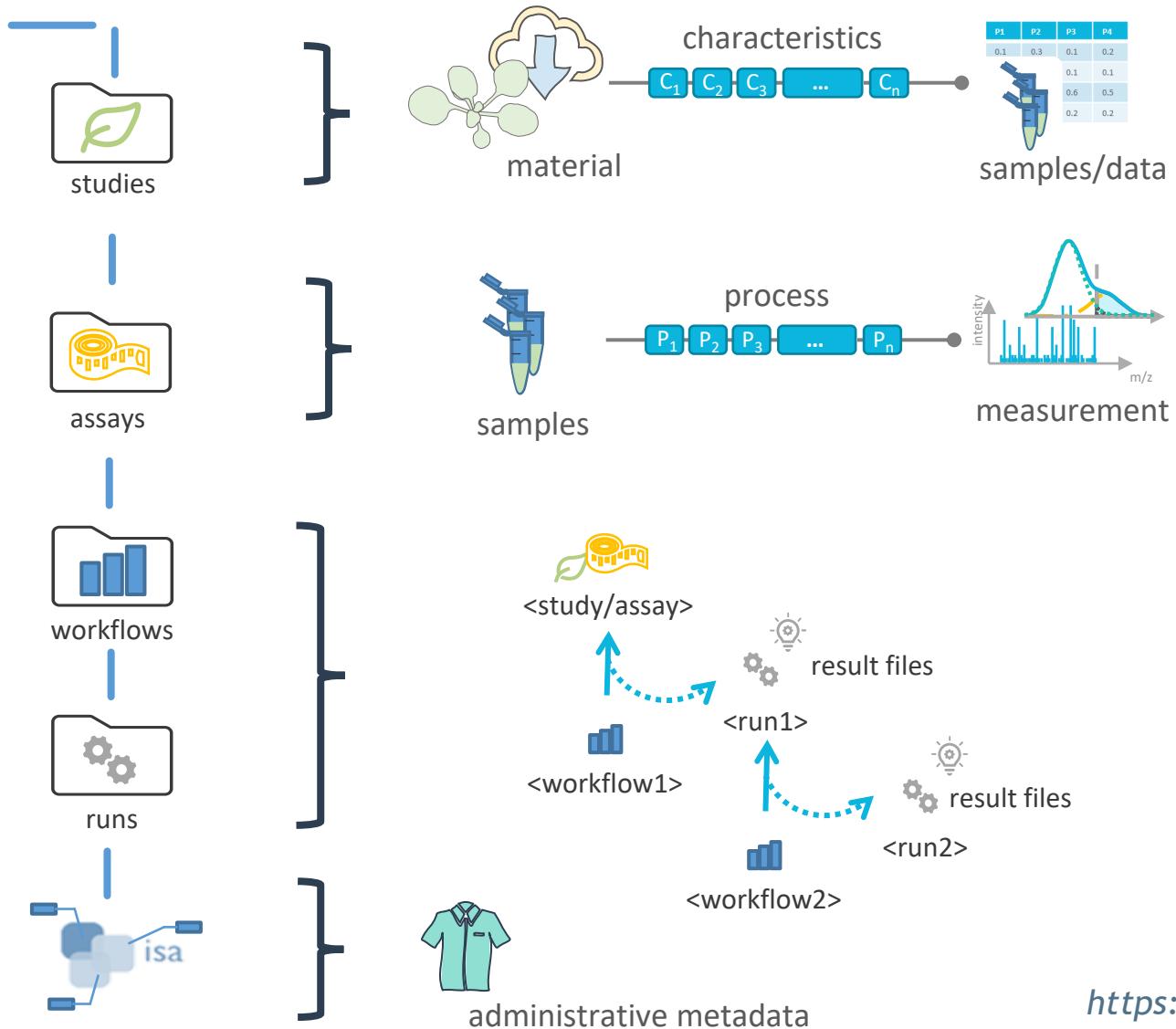


computation

annotation

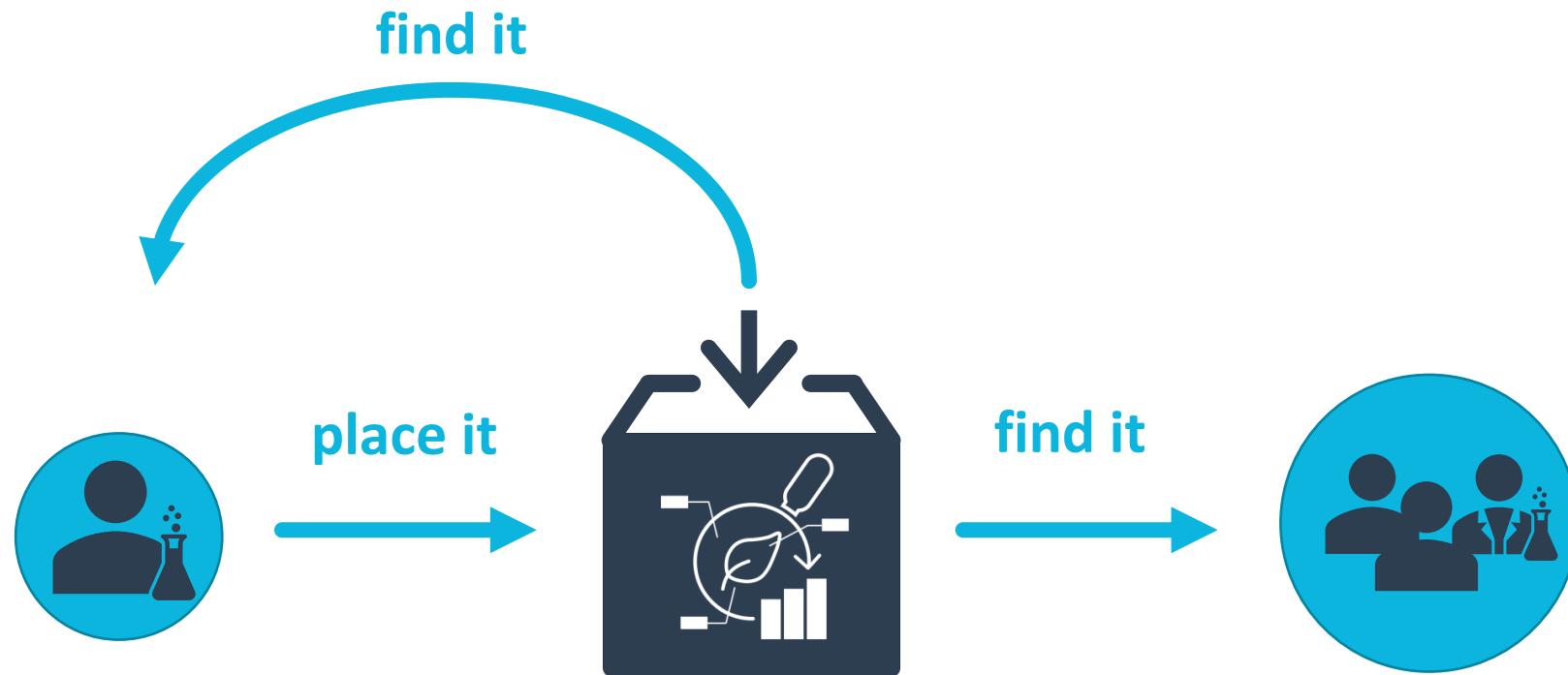
- Your whole investigation in a single unified bag

# Bag your research in an ARC



<https://github.com/nfdi4plants/ARC-specification>

# ARCs are “understandable”



FAIRness for everybody

# Find ARCs on the DataHUB

Search GitLab /

Projects

## Projects

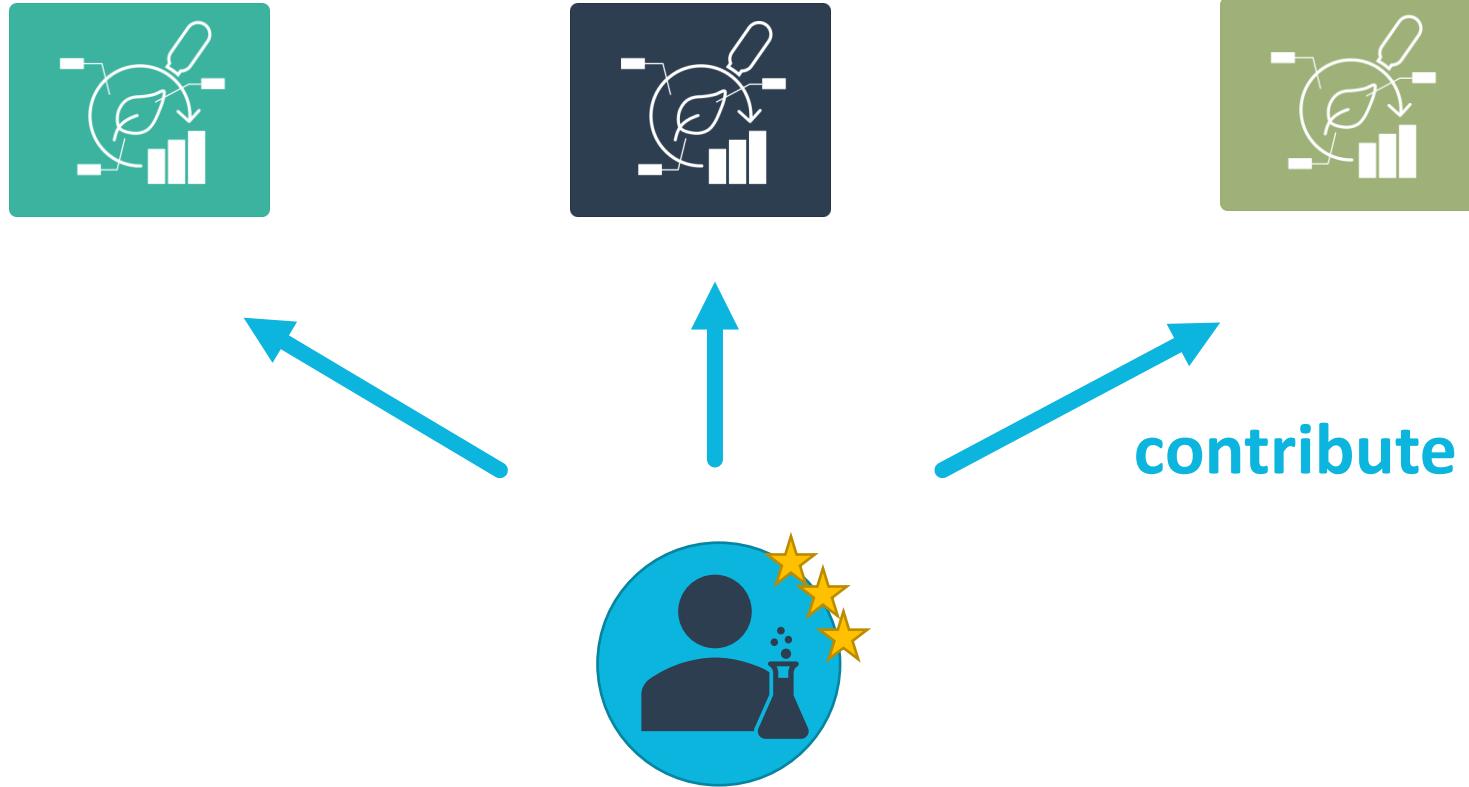
Yours 18 Starred 0 Explore Topics

Filter by name Updated date

All Most stars Trending Visibility: Any

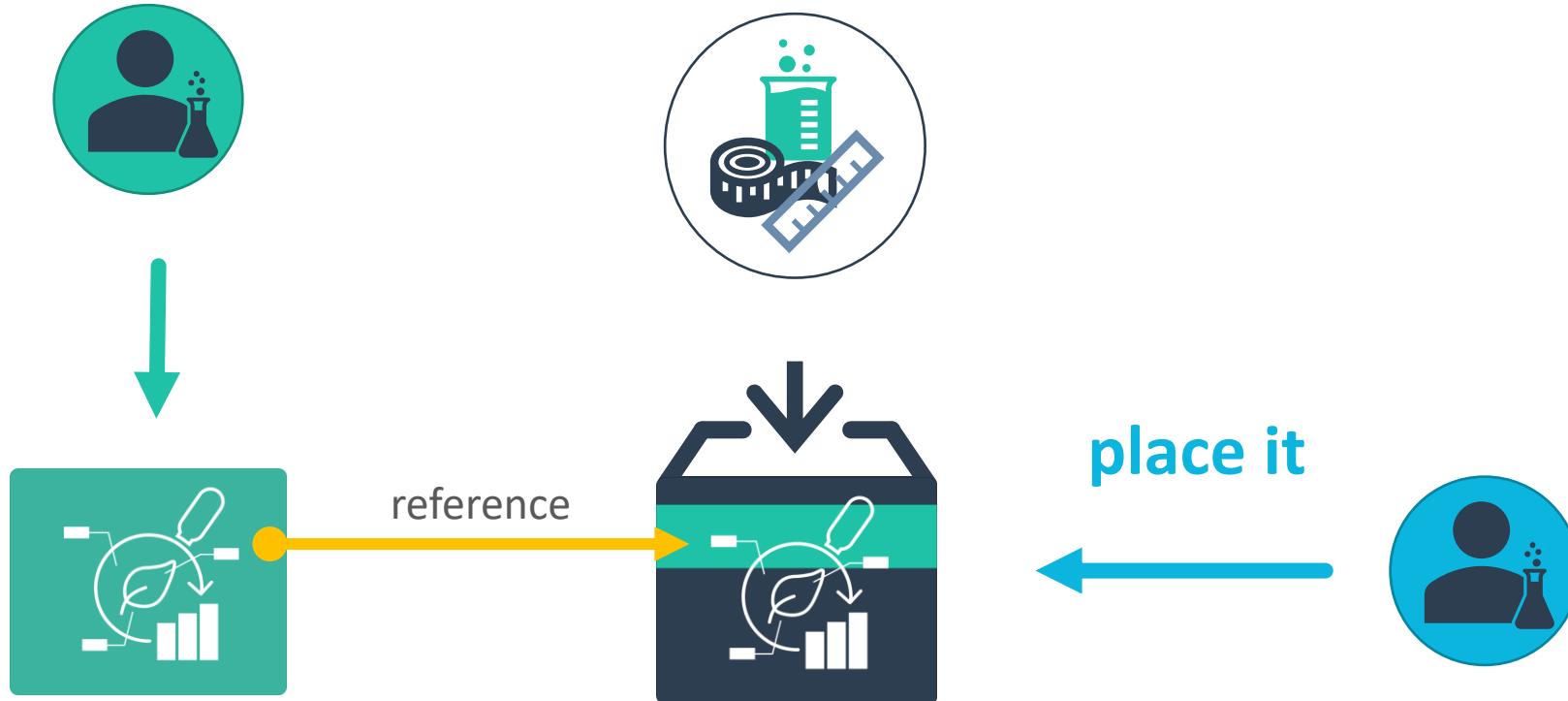
Owner	Name	Description	Stars	Forks	Issues	Commits	Last updated
G	Heinrich Lukas Weil / GeoSampleArc	Maintainer	0	0	0	0	Updated 18 hours ago
D	Teaching / demo-arc_start-data	Owner	0	0	0	0	Updated 4 days ago
T	Martin Kuhl / testVM	Owner	0	0	0	0	Updated 4 days ago
T	Dominik Brilhaus / Talinum fruticosum genome	🌐	0	0	0	0	Updated 5 days ago
A	Oliver Maus / AbsoluteQuantificationChlamyProteins	🛡 Maintainer	0	1	0	0	Updated 5 days ago
H	Felix Jung / HeatshockTuberculosisStressitance	🌐	0	0	0	0	Updated 5 days ago
T	Vivek Srivastav / Triticum	🌐	0	0	0	0	Updated 2 weeks ago
CAM	Dominik Brilhaus / SampleARC_RNAseq	🌐	3	2	0	1	Updated 2 weeks ago
S	Louisa Perelo / sampleARC_nfcore	🌐	0	0	0	0	Updated 2 weeks ago
D	Felix Jung / deepSTABp	🌐	0	0	0	0	Updated 2 weeks ago

# Contribute to multiple ARCs



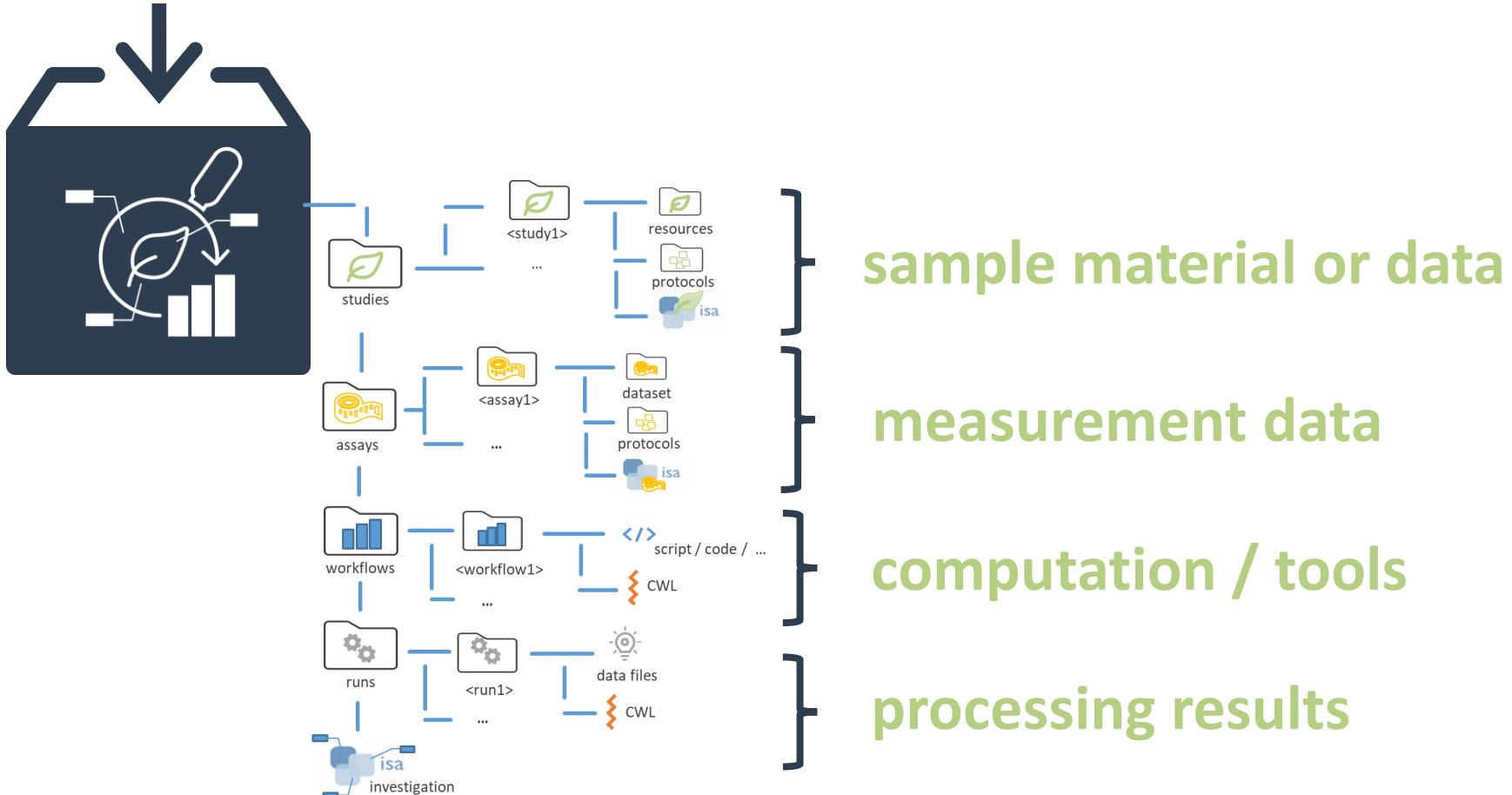
- Your contribution is tracked on ARCs but also visible in your profile

# ARCs can inherit information from other ARCs



- ▶ Gain more visibility by making your data and workflows reusable

# A real-world ARC is a folder structure



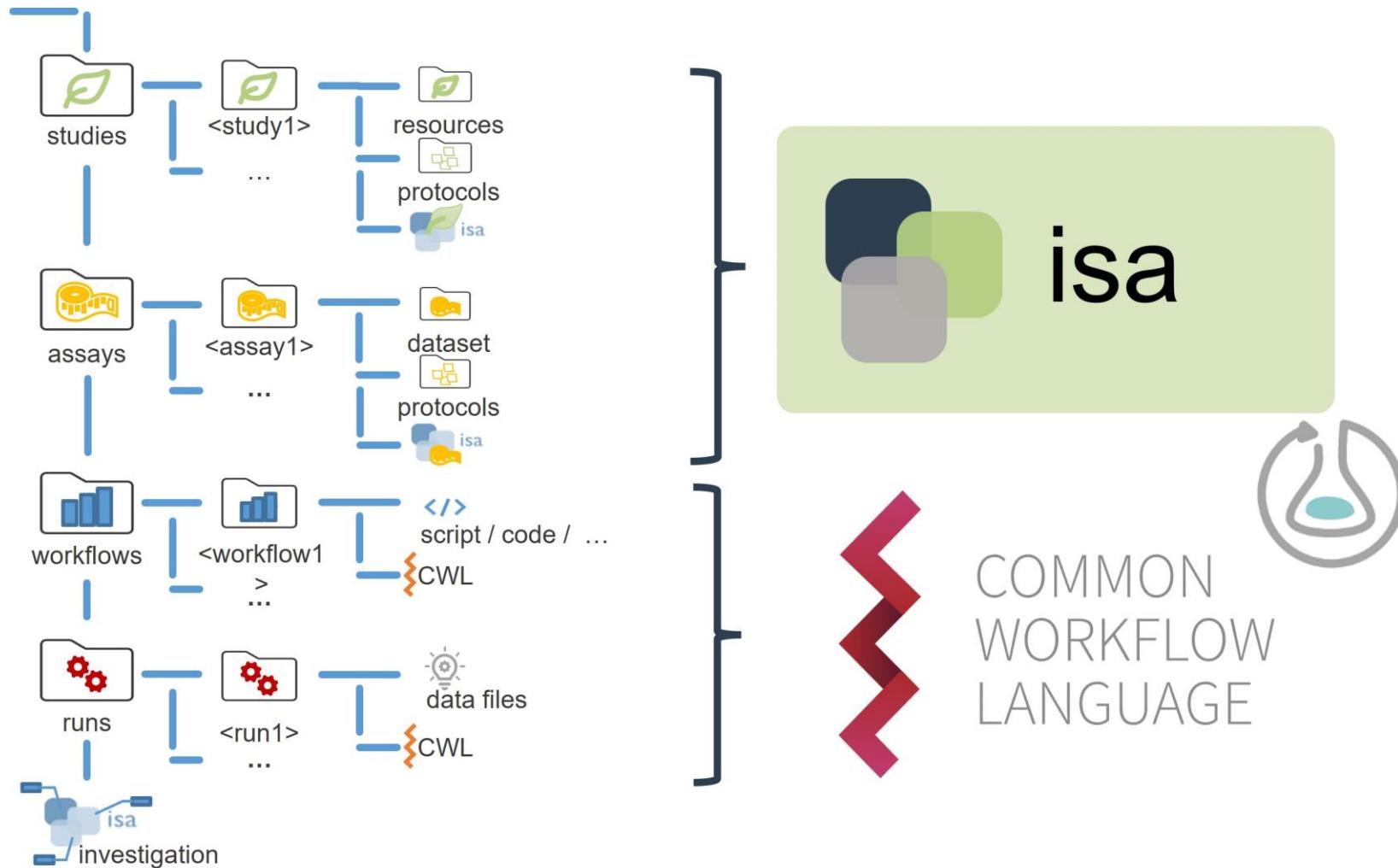
- User-friendly scaffold for research data and analyses, including a change tracking system to ensure provenance

# Version control system (git) tracks changes

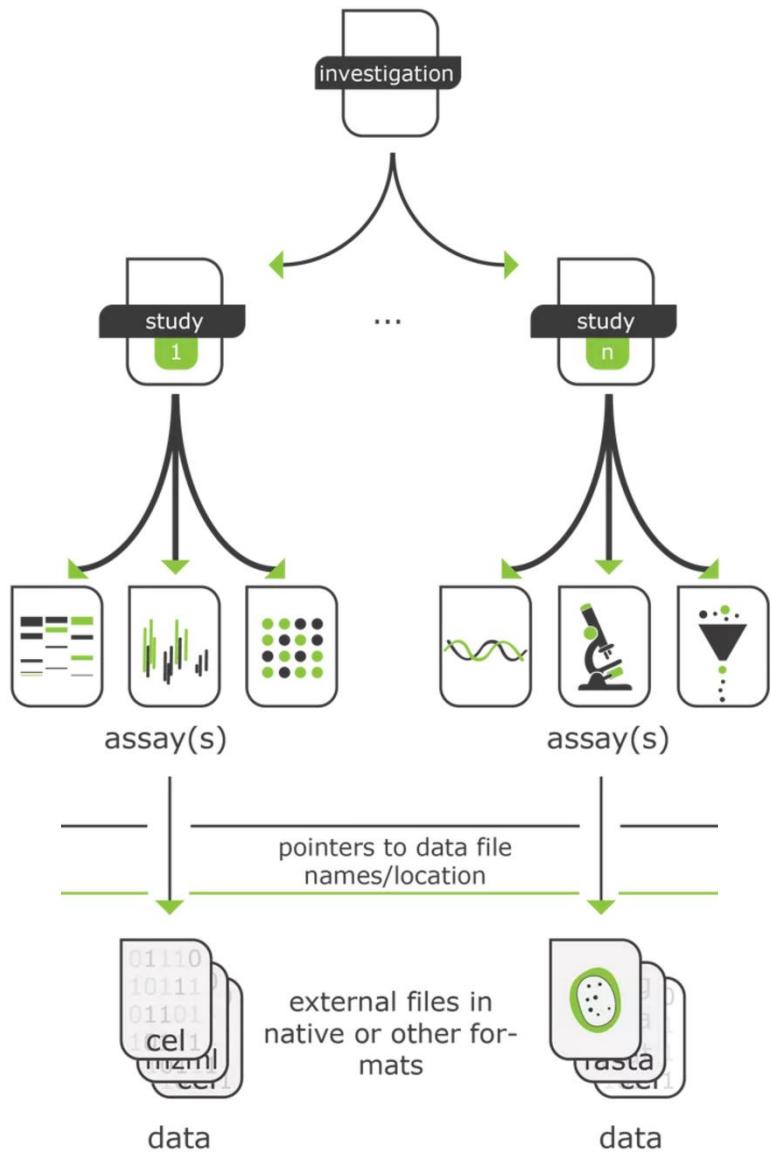


- ▶ Allows publishing of ARC snapshots

# ARC is built on top of existing standards



# The ISA metadata tracking framework



## investigation

high level concept to link related studies

## study

the central unit, containing information on the subject under study, its characteristics and any treatments applied.

*a study has associated assays*

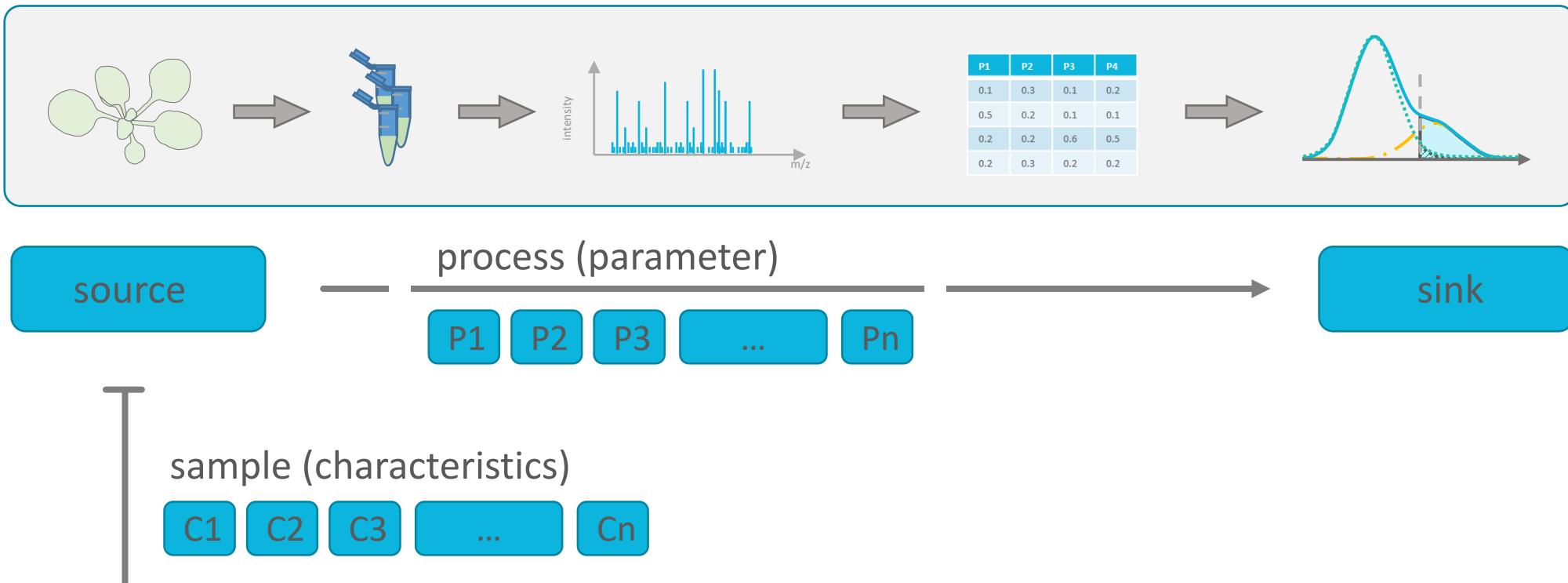
## assay

test performed either on material taken from the subject or on the whole initial subject, which produce qualitative or quantitative measurements (data)

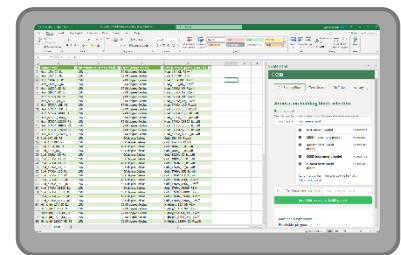
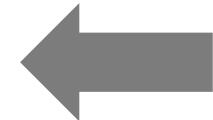
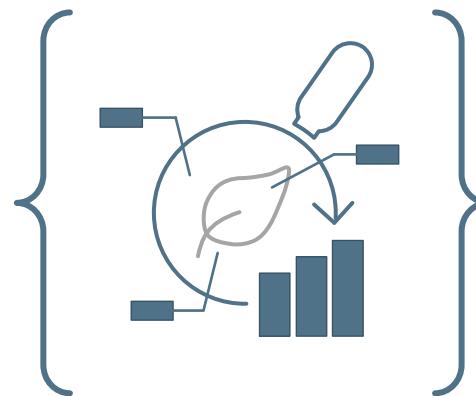
# Investigation for administrative metadata

ONTOLOGY SOURCE REFERENCE			
Term Source Name	PECO	UO	NFDI4PSO
Term Source File	<a href="http://purl.obolibrary.org/obo/peco.owl">http://purl.obolibrary.org/obo/peco.owl</a>	<a href="http://purl.obolibrary.org/obo/uo.owl">http://purl.obolibrary.org/obo/uo.owl</a>	<a href="https://github.com/nfdi4plants/nfdi4plants_ontology/blob/main/nfdi4plants_onto">https://github.com/nfdi4plants/nfdi4plants_ontology/blob/main/nfdi4plants_onto</a>
Term Source Version	releases/2020-08-21	releases/2020-03-10	
Term Source Description	Plant Experimental Conditions Ontology	Unit ontology	NFDI for Plant Science Ontology
INVESTIGATION			
Investigation Identifier	samplearc_metabolomics		
Investigation Title			
Investigation Description			
Investigation Submission Date	July 07, 2015		
Investigation Public Release Date	Nov 05, 2015		
INVESTIGATION PUBLICATIONS			
Investigation Publication PubMed ID		26530316	
Investigation Publication DOI		10.1104/pp.15.01076	
Investigation Publication Author List		Brilhaus, Dominik; Bräutigam, Andrea; Mettler-Altmann, Tabea; Winter, Klaus; Weber, Andreas P M	
Investigation Publication Title		Reversible Burst of Transcriptional Changes during Induction of Crassulacean Acid Metabolism in <i>Talinum triangulare</i> .	
Investigation Publication Status		Published	
Investigation Publication Status Term Accession Number			
Investigation Publication Status Term Source REF			
INVESTIGATION CONTACTS			
Investigation Person Last Name	Brilhaus	Bräutigam	Mettler-Altmann
Investigation Person First Name	Dominik	Andrea	Tabea
Investigation Person Mid Initials			
Investigation Person Email	dominik.brilhaus@hhu.de		
Investigation Person Phone	0211-8115523		
Investigation Person Fax			
Investigation Person Address	HHU Düsseldorf, 22.07.U1.43, Universitätsstr. 1	Universitätsstr. 1, 40225 Düsseldorf	Universitätsstr. 1, 40225 Düsseldorf
Investigation Person Affiliation	Institute of Plant Biochemistry, HHU Düsseldorf	Institute of Plant Biochemistry, HHU Düsseldorf	Institute of Plant Biochemistry, HHU Düsseldorf
Investigation Person Roles	research assistant	research assistant	research assistant
Investigation Person Roles Term Accession Number	<a href="http://purl.org/spar/scoro/research-assistant">http://purl.org/spar/scoro/research-assistant</a>	<a href="http://purl.org/spar/scoro/research-assistant">http://purl.org/spar/scoro/research-assistant</a>	<a href="http://purl.org/spar/scoro/research-assistant">http://purl.org/spar/scoro/research-assistant</a>
Investigation Person Roles Term Source REF	scoro	scoro	scoro

# Study and assay for dataset context (experimental metadata)



# Tools to (easily) work with ARCs

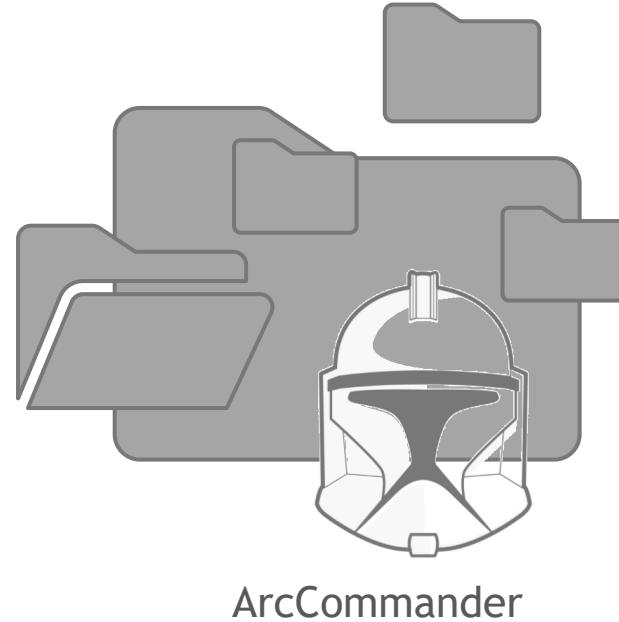


SWATE

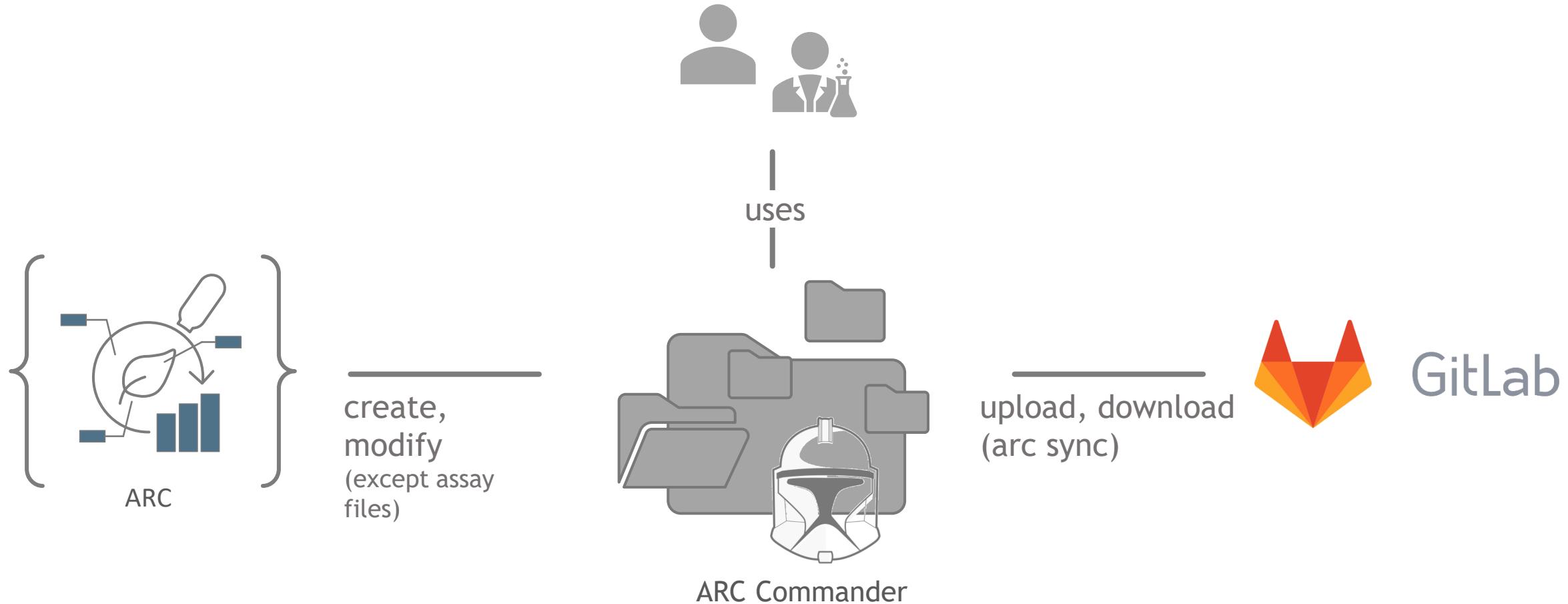


# ARC Commander

- ▶ Command line tool
- ▶ Facilitates managing the ARC structure and administrative metadata
- ▶ Connection to the DataHUB



# ARC Commander

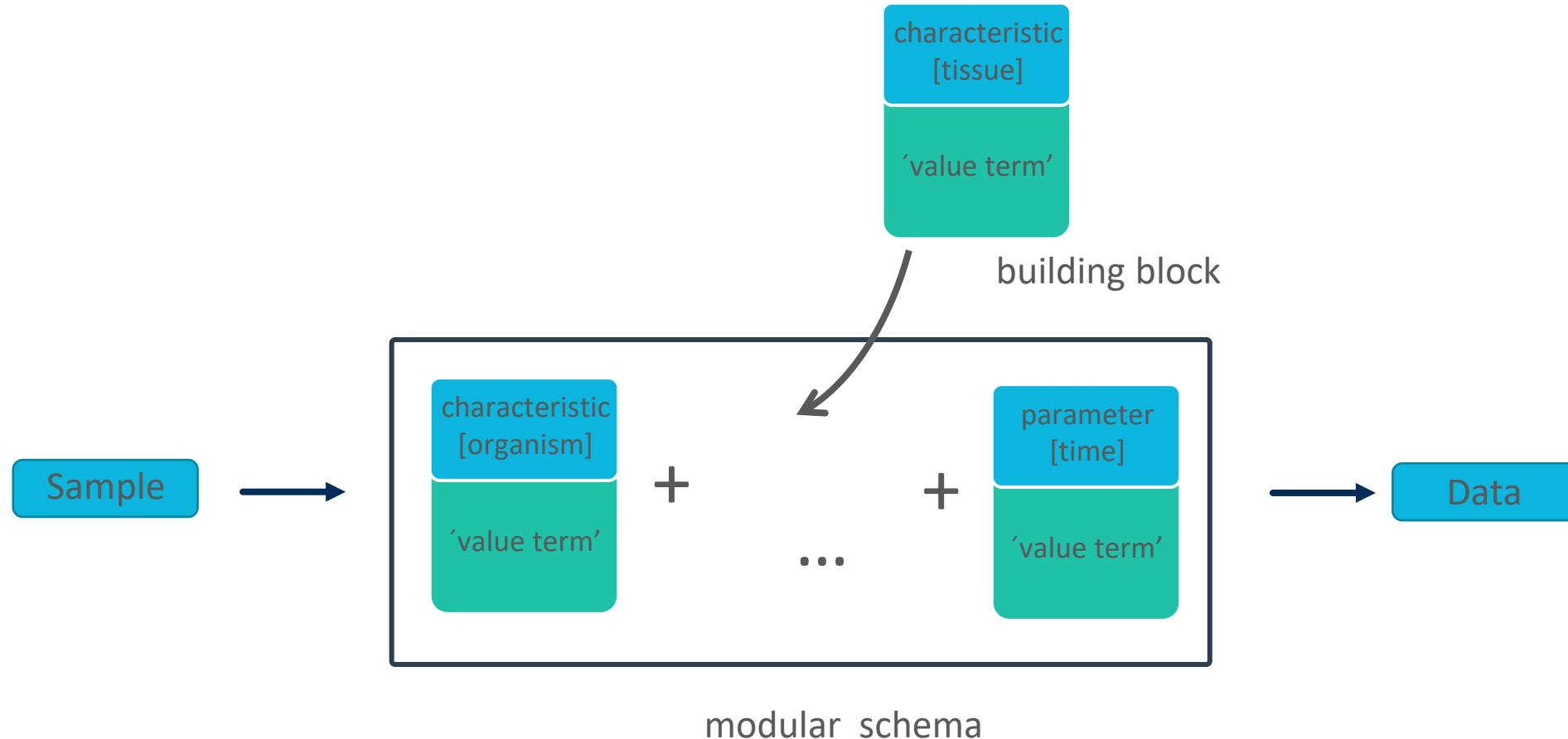


# Swate

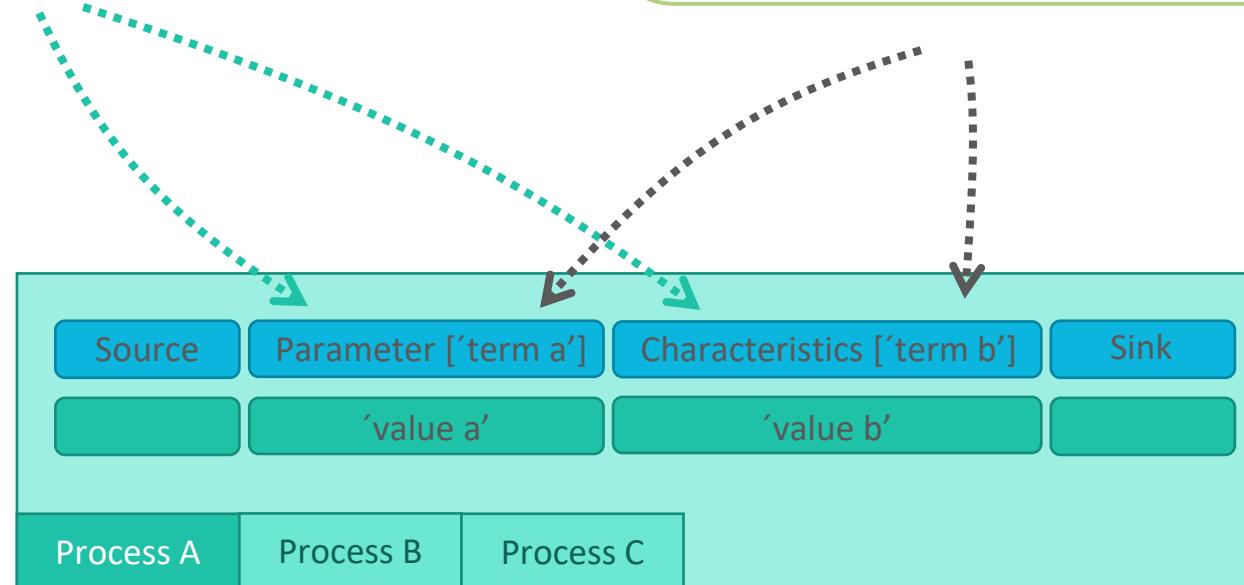
- ▶ (Currently) Excel Add-In
- ▶ Facilitates creation and modification of ISA annotation tables in study/assay files
- ▶ Library of templates to insert into your tables
  - ▶ Fred created a template for his MS workflow

The screenshot shows the Swate interface integrated into an Excel spreadsheet. On the left, there is a table with columns A, B, and C. Column A is labeled 'Source Name' with dropdown menus for 'Characteristic [Organism]' and 'Characteristic [Genotype]'. The data in column A consists of rows from 1 to 33, each containing a unique identifier followed by 'Chlamydomonas reinhardtii' and 'wild type genotype'. To the right of the table is the Swate application window. The title bar says 'Swate' and 'SWATE'. The main area is titled 'Building Blocks' with the sub-instruction 'Add annotation building blocks (columns) to the annotation table.' Below this are search and search unit input fields, and a prominent red button labeled 'Add building block'. Another section below is titled 'Add/Update unit reference to existing building block.' with the instruction 'Adds a unit to a complete building block. If the building block already has a unit assigned, the new unit is only applied to selected rows of the selected column.' It includes 'Add unit' and 'Start typing to search' fields, and a red button labeled 'Update unit for cells'. At the bottom right of the application window, it says 'Swate Release Version 0.7.1'.

# Combining annotation building blocks

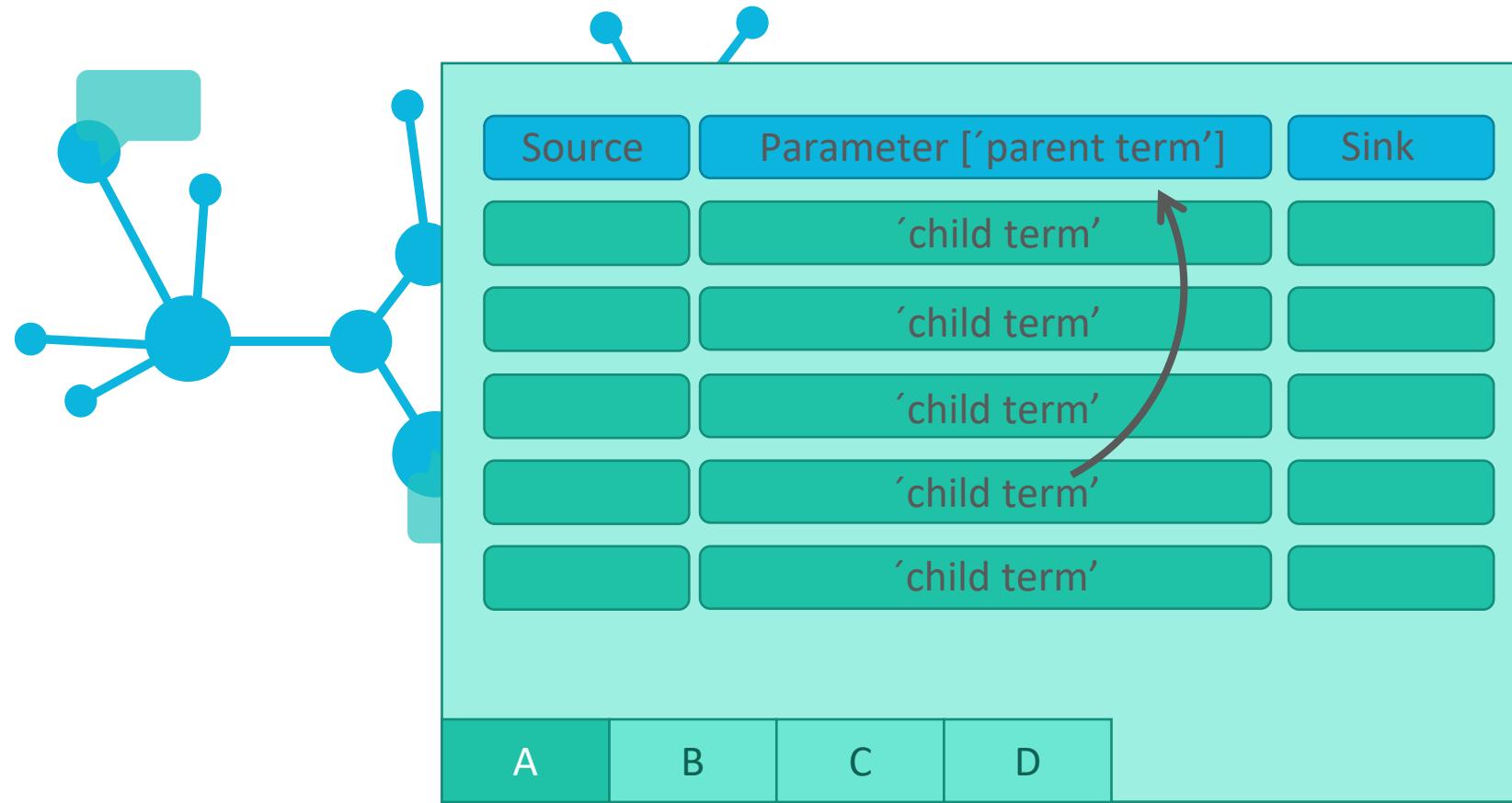


# Hierarchical combination of ontologies



**document (ISA assay)**

# Annotation by flattening the knowledge graph



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

# Annotation by flattening the knowledge graph

Sample	Component [instrument model]	Data
	'TripleTOF4600'	

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

# Building block types

The screenshot shows the SWATe software interface with the title 'Swate'. The main window is titled 'Building Blocks' and contains a search bar and a list of building block types: Source Name, Parameter, Factor, Characteristic, Component, Protocol Columns, and Output Columns. A tooltip for 'Parameter' indicates it describes steps in the experimental workflow. A red box highlights the 'Add building block' button for 'Parameter'. Below the list, there is a section titled 'More about Parameter:' with a detailed description of what parameter columns represent.

Swate

SWATE + - ?

Building Blocks

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

Parameter Start typing to search

Source Name as a unit: Start typing to search

Parameter lock Use advanced search unit

Add building block

Factor existing building block.

Characteristic

Component

Protocol Columns →

Output Columns →

Update unit for cells

More about Parameter:

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](#).

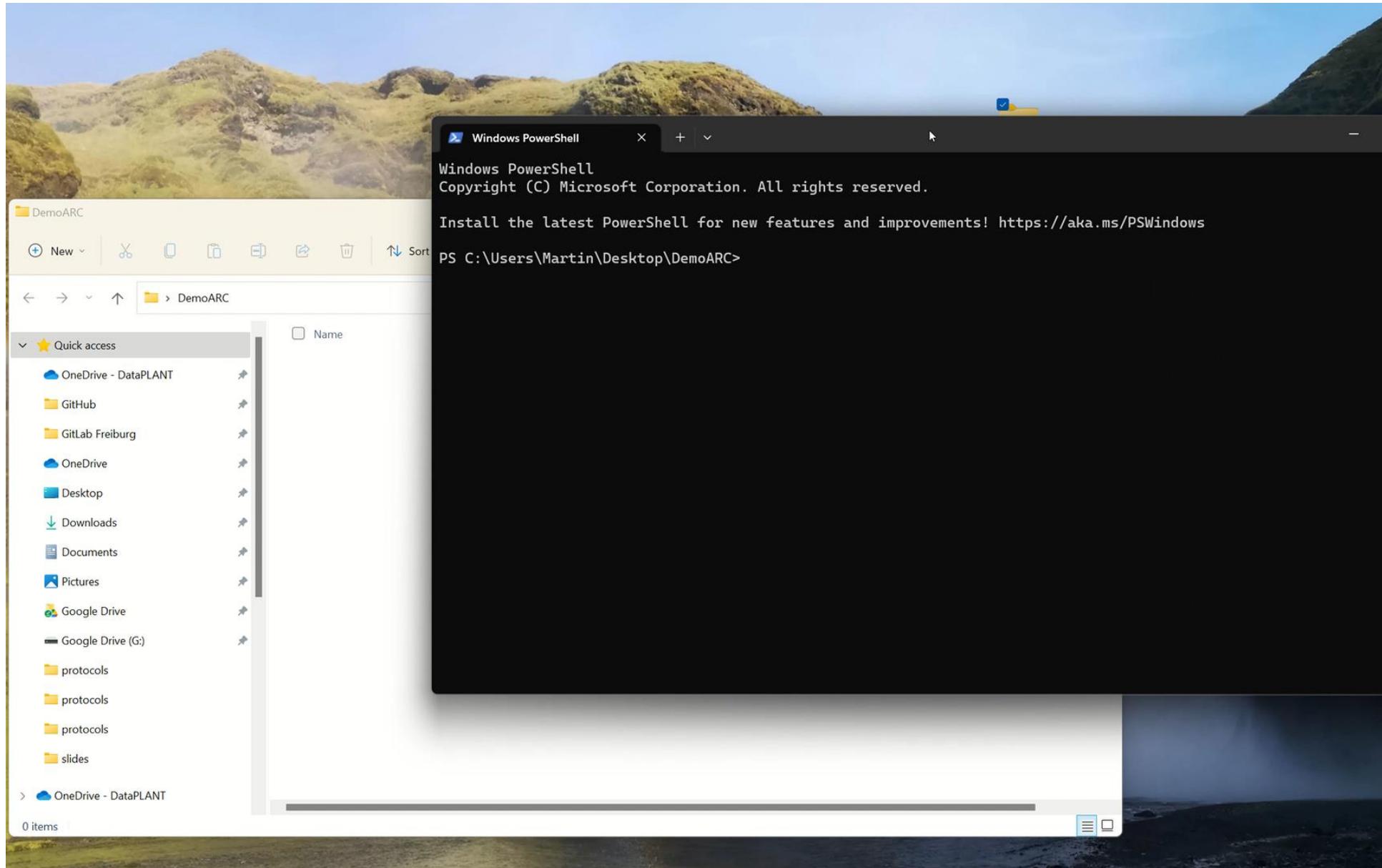
- ▶ **Source Name:** Input of your table (unique identifier)
- ▶ **Characteristic:** Inherent properties of the source material (organism, strain, etc.)
- ▶ **Parameter:** steps in experimental workflow (light intensity, temperature, etc.)
- ▶ **Component:** anything physical (instruments, reagents, software)
- ▶ **Factor:** experimental design, most important for computational analysis
- ▶ **Output columns (unique identifier)**
  - ▶ Sample Name, Raw Data File, Derived Data File

# Setting up an ARC

# Check ARC Commander functionality



# Initialize an ARC

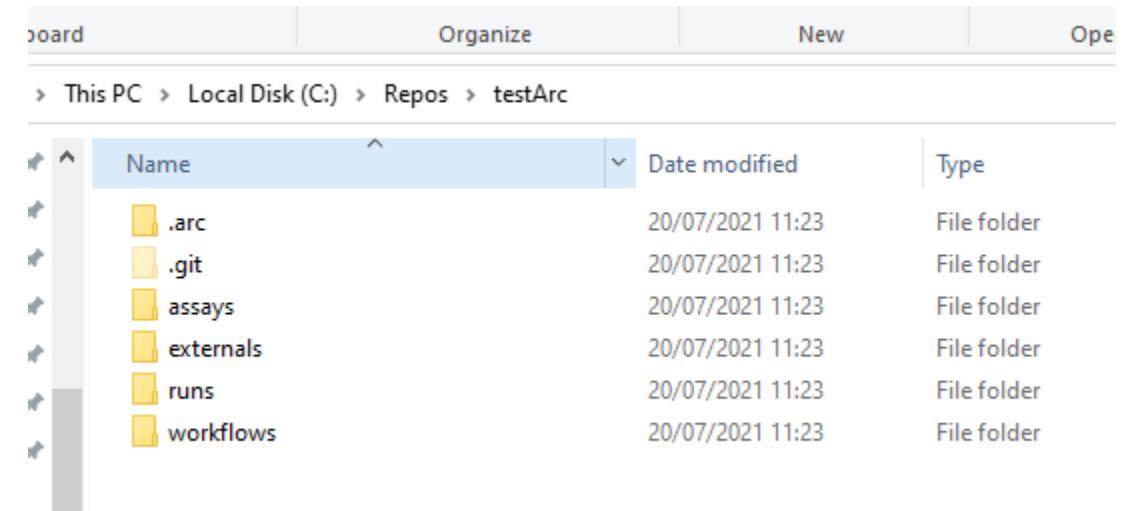


# ARC Commander functions

## ► *arc init*

```
PS C:\Repos\testArc> arc init
Start processing command with the arguments
  EditorPath:Field ""
  GitLFSByteThreshold:Field ""
  Owner:Field ""
  RepositoryAdress:Field ""

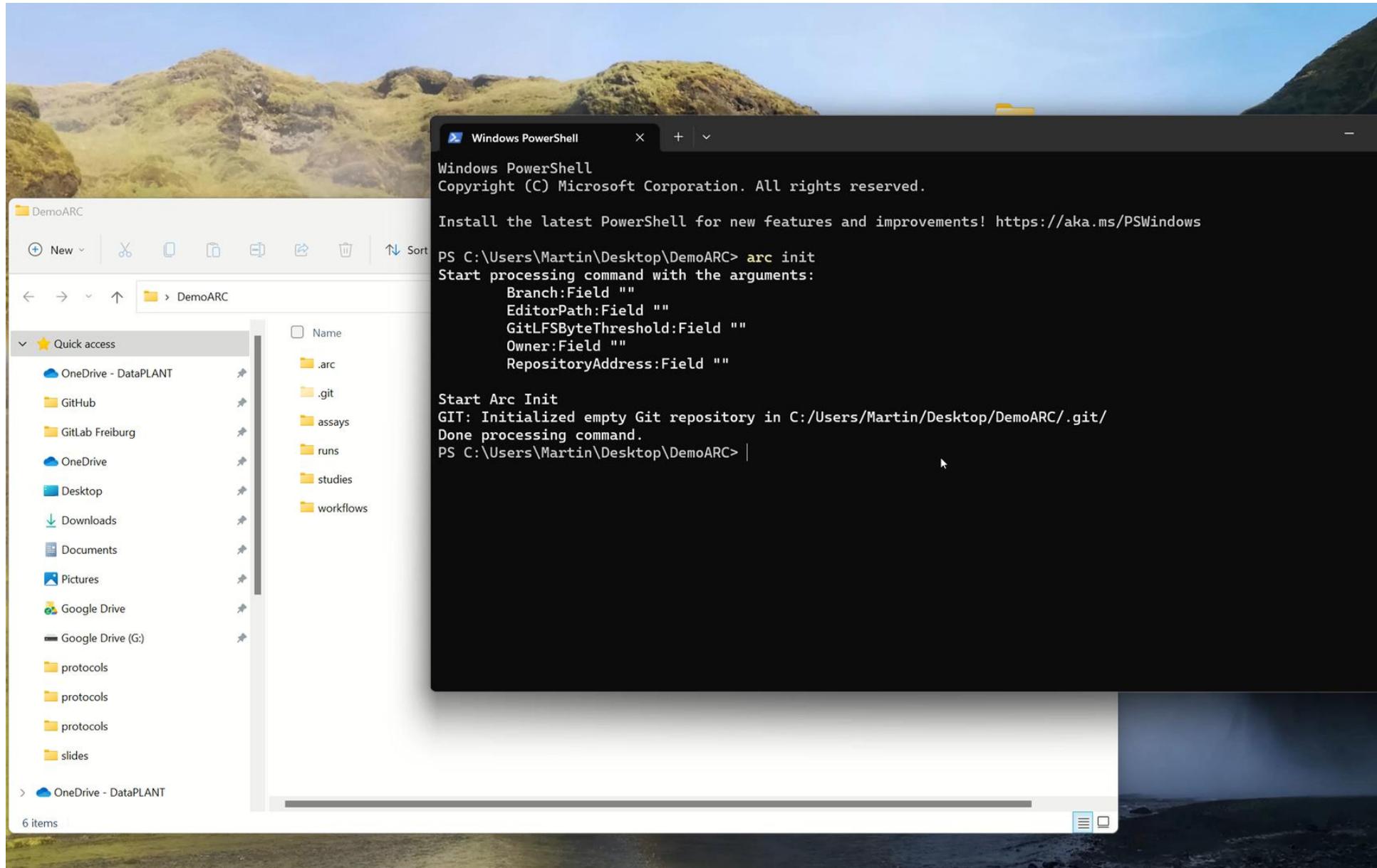
Start Arc Init
Initialized empty shared Git repository in C:/Repos/testArc/.git/
Done processing command
```



A screenshot of a Windows File Explorer window. The address bar shows the path: This PC > Local Disk (C:) > Repos > testArc. The main area displays a list of files and folders:

Name	Date modified	Type
.arc	20/07/2021 11:23	File folder
.git	20/07/2021 11:23	File folder
assays	20/07/2021 11:23	File folder
externals	20/07/2021 11:23	File folder
runs	20/07/2021 11:23	File folder
workflows	20/07/2021 11:23	File folder

# Add administrative metadata (isa.investigation.xlsx)

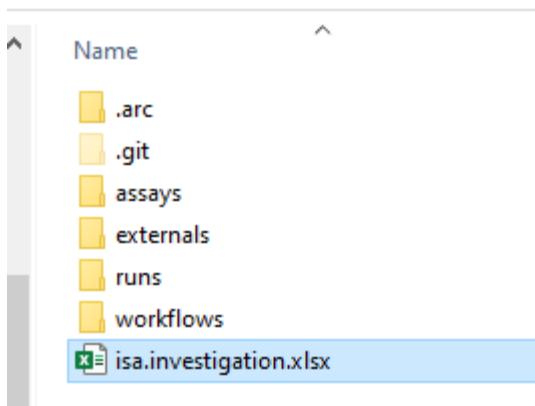


# ARC Commander functions

## ► *arc i create*

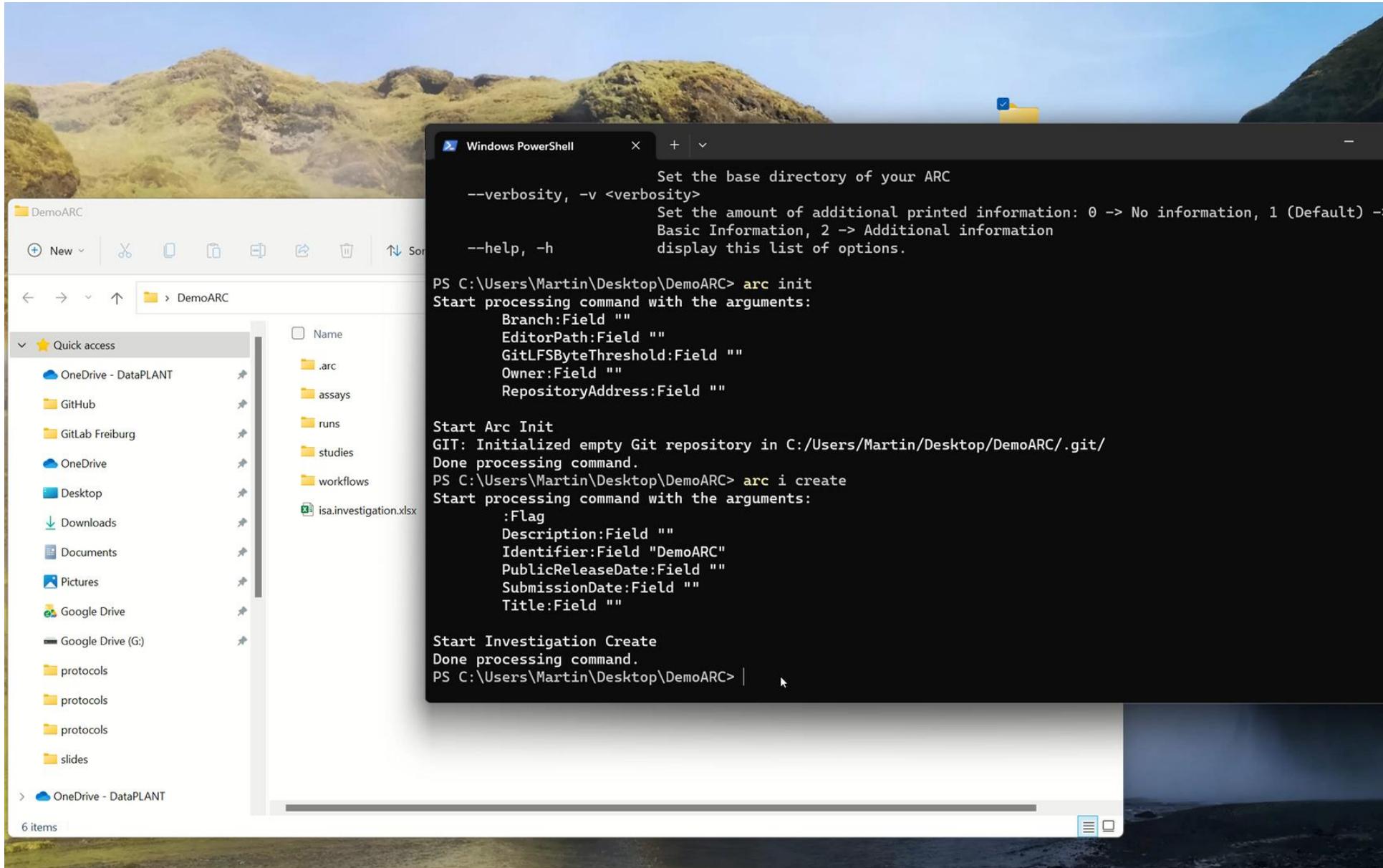
```
PS C:\Repos\testArc> arc i create
Start processing command with the arguments
  :Flag
  Description:Field "myDescription"
  Identifier:Field "---"
  PublicReleaseDate:Field "in the future"
  SubmissionDate:Field "in the future"
  Title:Field "myTitle"

Start Investigation Create
Done processing command
```



A	B
1	ONTOLOGY SOURCE REFERENCE
2	Term Source Name
3	Term Source File
4	Term Source Version
5	Term Source Description
5	INVESTIGATION
7	Investigation Identifier
8	---
3	Investigation Title
9	myTitle
9	Investigation Description
0	myDescription
0	Investigation Submission Date
1	in the future
1	Investigation Public Release Date
2	in the future
2	INVESTIGATION PUBLICATIONS
3	Investigation Publication PubMed ID
4	Investigation Publication DOI
5	Investigation Publication Author List
6	Investigation Publication Title
7	Investigation Publication Status
8	Investigation Publication Status Term Accession Number
9	Investigation Publication Status Term Source REF
0	INVESTIGATION CONTACTS
1	Investigation Person Last Name
2	Investigation Person First Name
3	Investigation Person Mid Initials
4	Investigation Person Email
5	Investigation Person Phone
6	Investigation Person Fax
7	Investigation Person Address
8	Investigation Person Affiliation
9	Investigation Person Roles
0	Investigation Person Roles Term Accession Number
1	Investigation Person Roles Term Source REF

# Add a study to your ARC

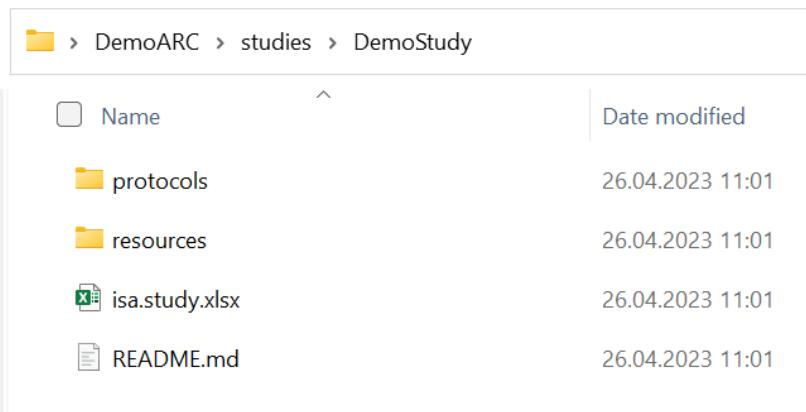


# ARC Commander functions

## ► *arc s add*

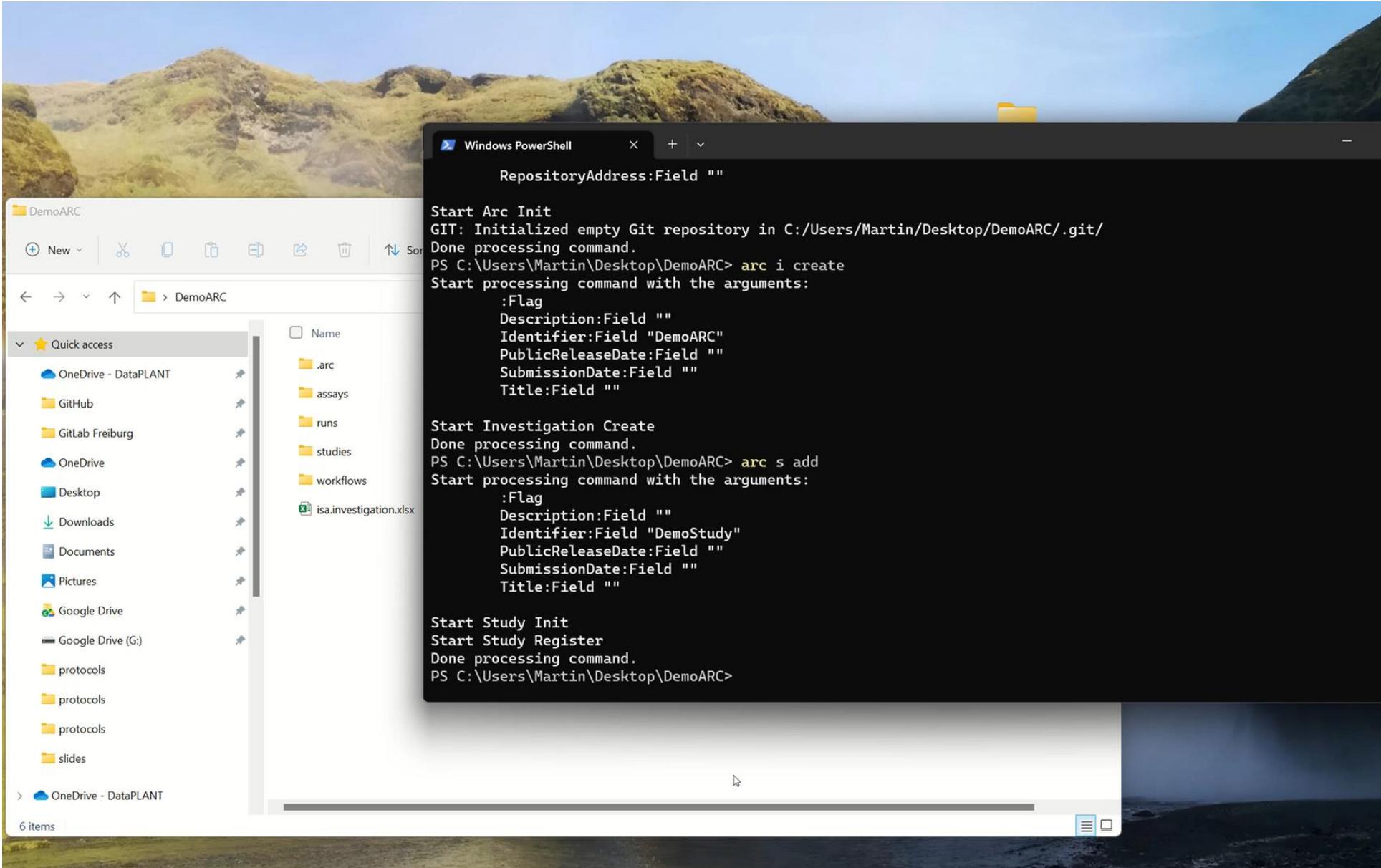
```
PS C:\Repos\testArc> arc s add
Start processing command with the arguments
:Flag
Description:Field "myStudyDescription"
Identifier:Field "myStudyID"
PublicReleaseDate:Field "myStudyPRDate"
SubmissionDate:Field "myStudySubDate"
Title:Field "myStudyTitle"
```

```
Start Study Init
Start Study Register
Done processing command
```



1	INVESTIGATION PERSON ROLES TERM SOURCE REF	
2	STUDY	
3	Study Identifier	myStudyID
4	Study Title	myStudyTitle
5	Study Description	myStudyDescription
6	Study Submission Date	myStudySubDate
7	Study Public Release Date	myStudyPRDate
8	Study File Name	myStudyID.study.xlsx
9	STUDY DESIGN DESCRIPTORS	
0	Study Design Type	
1	Study Design Type Term Accession Number	
2	Study Design Type Term Source REF	
3	STUDY PUBLICATIONS	
4	Study Publication PubMed ID	
5	Study Publication DOI	
6	Study Publication Author List	
7	Study Publication Title	
8	Study Publication Status	
9	Study Publication Status Term Accession Number	
0	Study Publication Status Term Source REF	

# Add an assay to your study

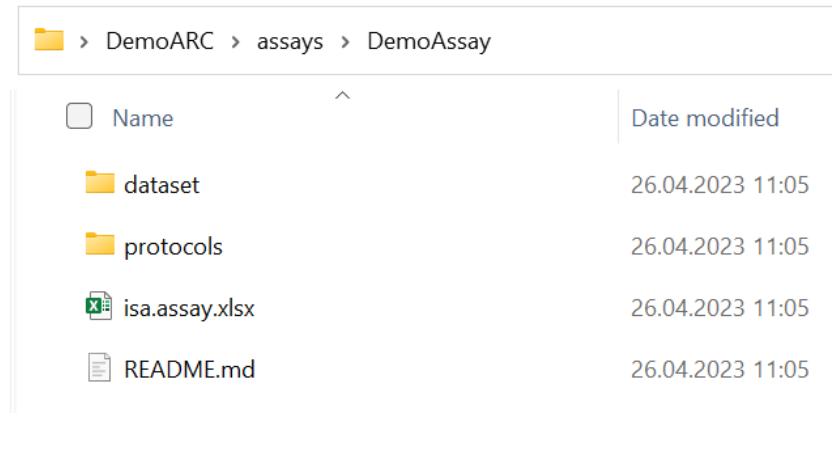


# ARC Commander functions

## ► *arc a add*

```
PS C:\Repos\testArc> arc a add
Start processing command with the arguments
  :Flag
  AssayIdentifier:Field "myAssayID"
  MeasurementType:Field "myMeasurementType"
  MeasurementTypeTermAccessionNumber:Field "myMTTAN"
  MeasurementTypeTermSourceREF:Field "myMTTSR"
  StudyIdentifier:Field "myStudyID"
  TechnologyPlatform:Field "myTechPlatform"
  TechnologyType:Field "myTechType"
  TechnologyTypeTermAccessionNumber:Field "myTTTAN"
  TechnologyTypeTermSourceREF:Field "myTTTSR"

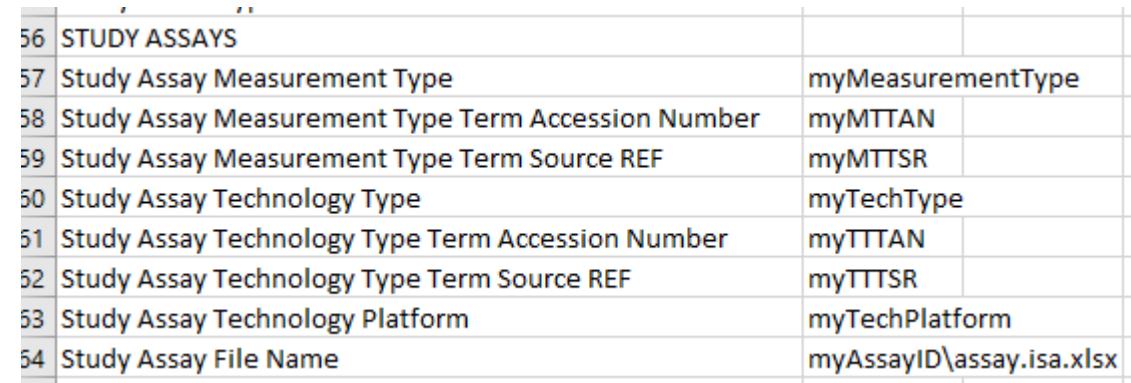
Start Assay Init
Start Assay Register
Done processing command
```



The screenshot shows a Windows File Explorer window with the following directory structure:

```
DemoARC > assays > DemoAssay
```

Name	Date modified
dataset	26.04.2023 11:05
protocols	26.04.2023 11:05
isa.assay.xlsx	26.04.2023 11:05
README.md	26.04.2023 11:05

The screenshot shows an Excel spreadsheet with the following data:

	STUDY ASSAYS	
56	Study Assay Measurement Type	myMeasurementType
57	Study Assay Measurement Type Term Accession Number	myMTTAN
58	Study Assay Measurement Type Term Source REF	myMTTSR
59	Study Assay Technology Type	myTechType
60	Study Assay Technology Type Term Accession Number	myTTTAN
61	Study Assay Technology Type Term Source REF	myTTTSR
62	Study Assay Technology Platform	myTechPlatform
63	Study Assay File Name	myAssayID\assay.isa.xlsx

# ARC Commander functions

## ► always useful:

- *arc -h*
- *arc <command> -h*
- *arc <command> <subcommand> -h*

```
PS C:\Repos\testArc> arc --help
USAGE: arc.exe [--help] [--workingdir <working directory>] [--verbosity <verbosity>] [<subcommand> [<options>]]

SUBCOMMANDS:

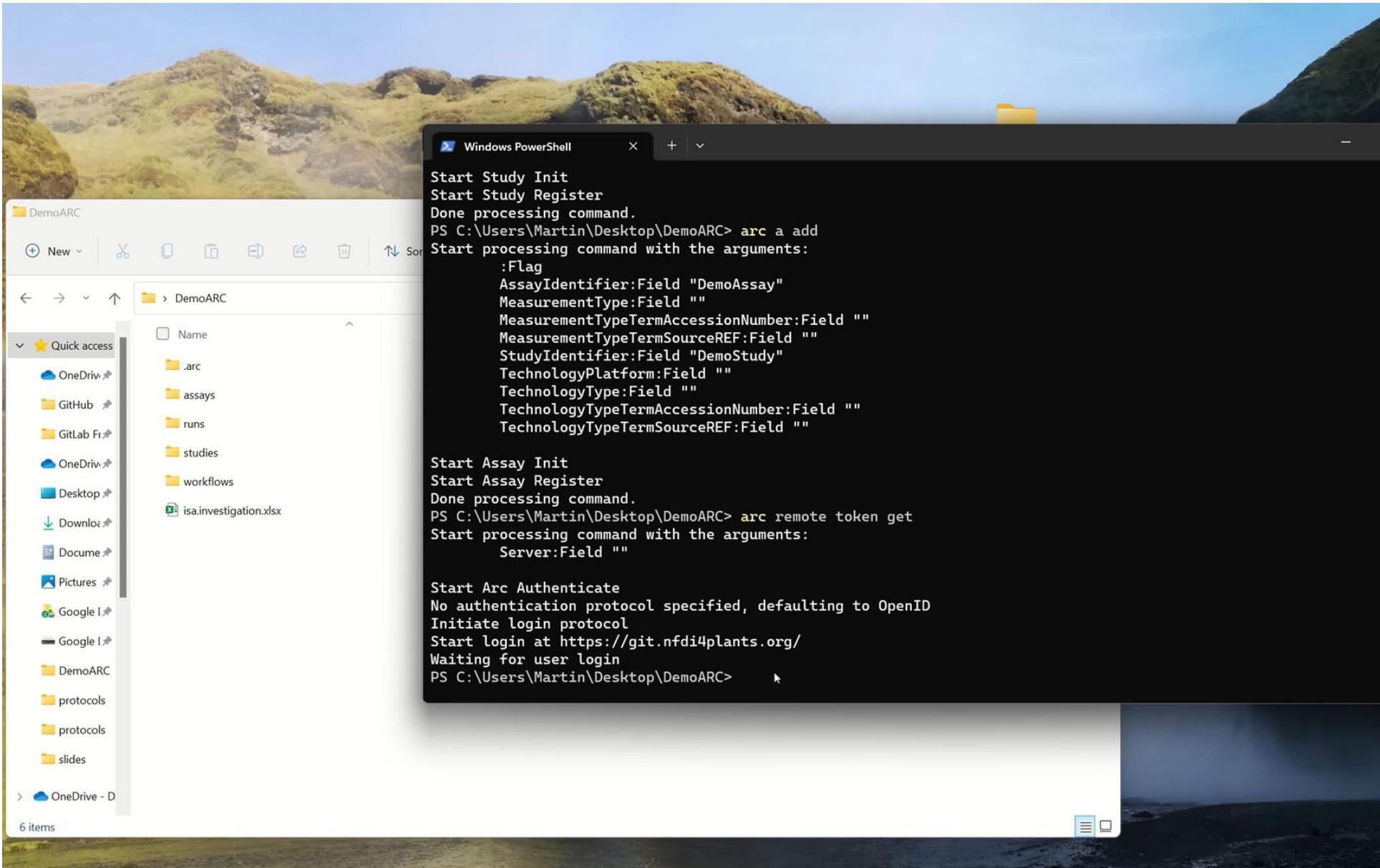
    init <init args>      Initializes basic folder structure
    synchronize, sync     Synchronize ISA and other items
    investigation, i <verb and args>
                          Investigation file functions
    study, s <verb and args>
                          Study functions
    assay, a <verb and args>
                          Assay functions
    configuration, config <verb and args>
                          Configuration editing
    git <verb and args>   Git related

    Use 'arc.exe <subcommand> --help' for additional information.

OPTIONS:

    --workingdir, -p <working directory>
                          Set the base directory of your ARC
    --verbosity, -v <verbosity>
                          Sets the amount of additional printed information: 0->No information, 1 (Default) -> Basic
                                         Information, 2 -> Additional information
    --help                display this list of options.
```

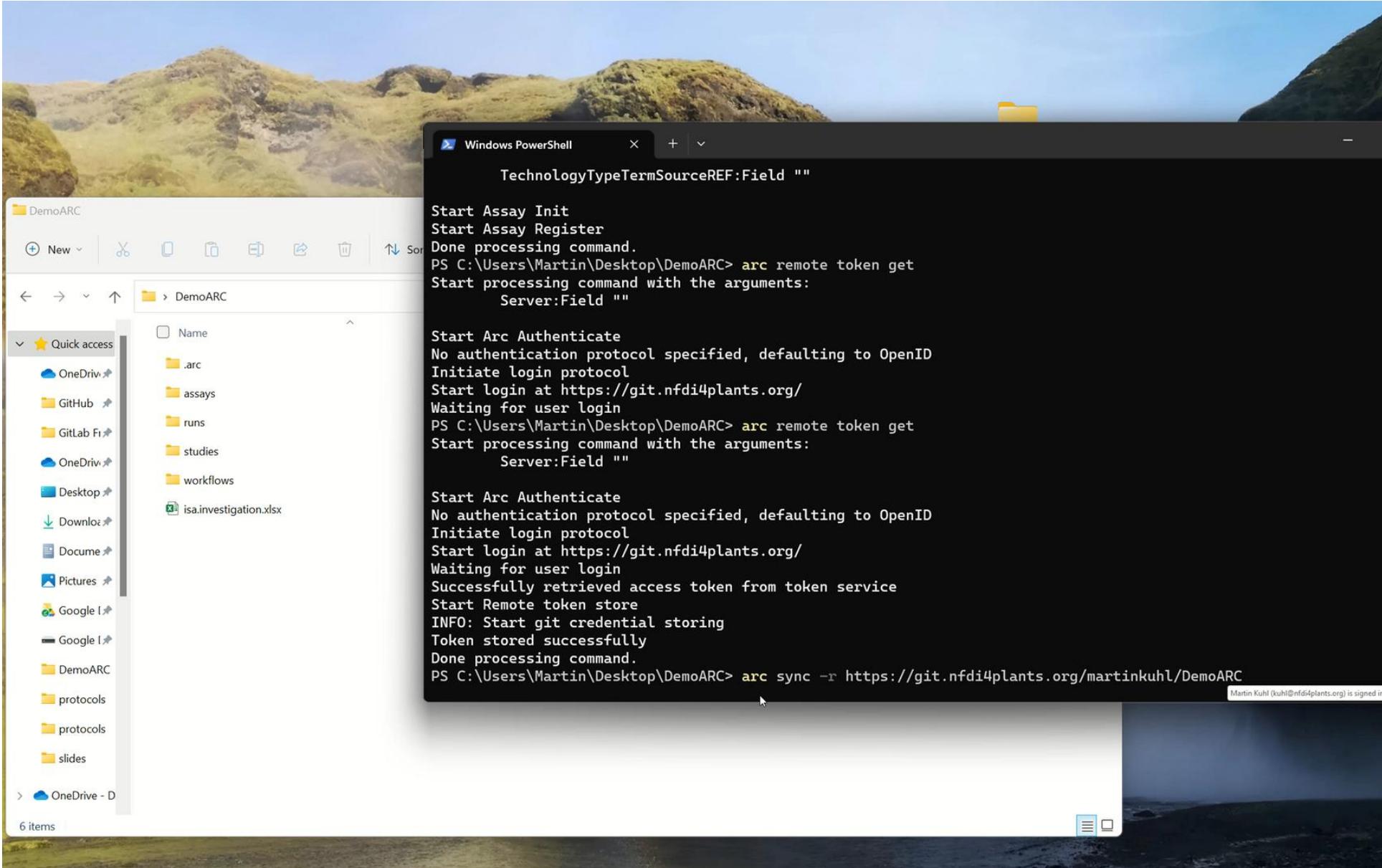
# Synchronize your ARC to the DataHUB



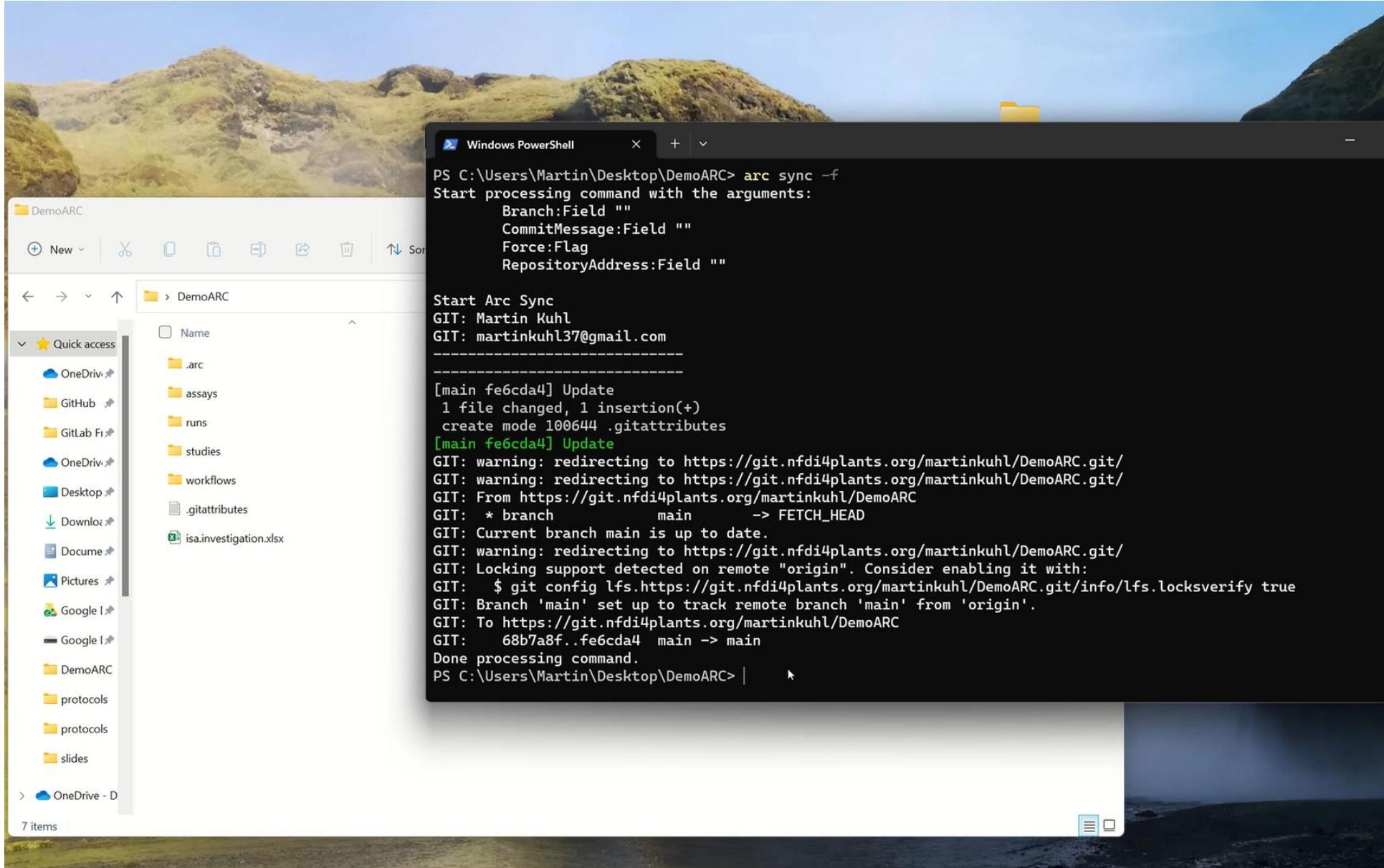
# ARC Commander functions

- ▶ *arc remote token get*
  - ▶ Retrieves an access token from the DataHUB
  - ▶ Enables the ARC Commander to up- and download new versions of the ARC to the DataHUB
- ▶ Arc sync -r <https://git.nfdi4plants.org/>*UserNAME/NameOfARC*

# Synchronize your ARC to the DataHUB



# Synchronize your ARC to the DataHUB



# Synchronize your ARC to the DataHUB

1. *arc sync -r https://git.nfdi4plants.org/<YourUserName>/<YourARC>*
2. If no repository exists under the given URL, the ARC Commander will produce an error ensuring that you spelled the URL correctly. To force synchronization, use  
→ *arc sync -f*

```
[master (root-commit) f0/20e7] Update
ERROR: Remote repo was set, but does not exist.
Check whether it was spelled correctly. If not, you can run "arc sync" again using the --repositoryAddress argument.
If you want to create a new remote repository instead. You can run "arc sync -f" to force push the local repository to a new upstream.
Done processing command.
PS C:\Users\Martin\Desktop\Quickstart> arc sync -f
Start processing command with the arguments:
    Branch:Field ""
    CommitMessage:Field ""
    Force:Flag
    RepositoryAddress:Field ""

Start Arc Sync
GIT: Martin Kuhl
GIT: martinkuhl137@gmail.com
-----
GIT: Tracking "**/dataset/**"
-----
On branch main
Untracked files:
  (use "git add <file>..." to include in what will be committed)
    .gitattributes
```

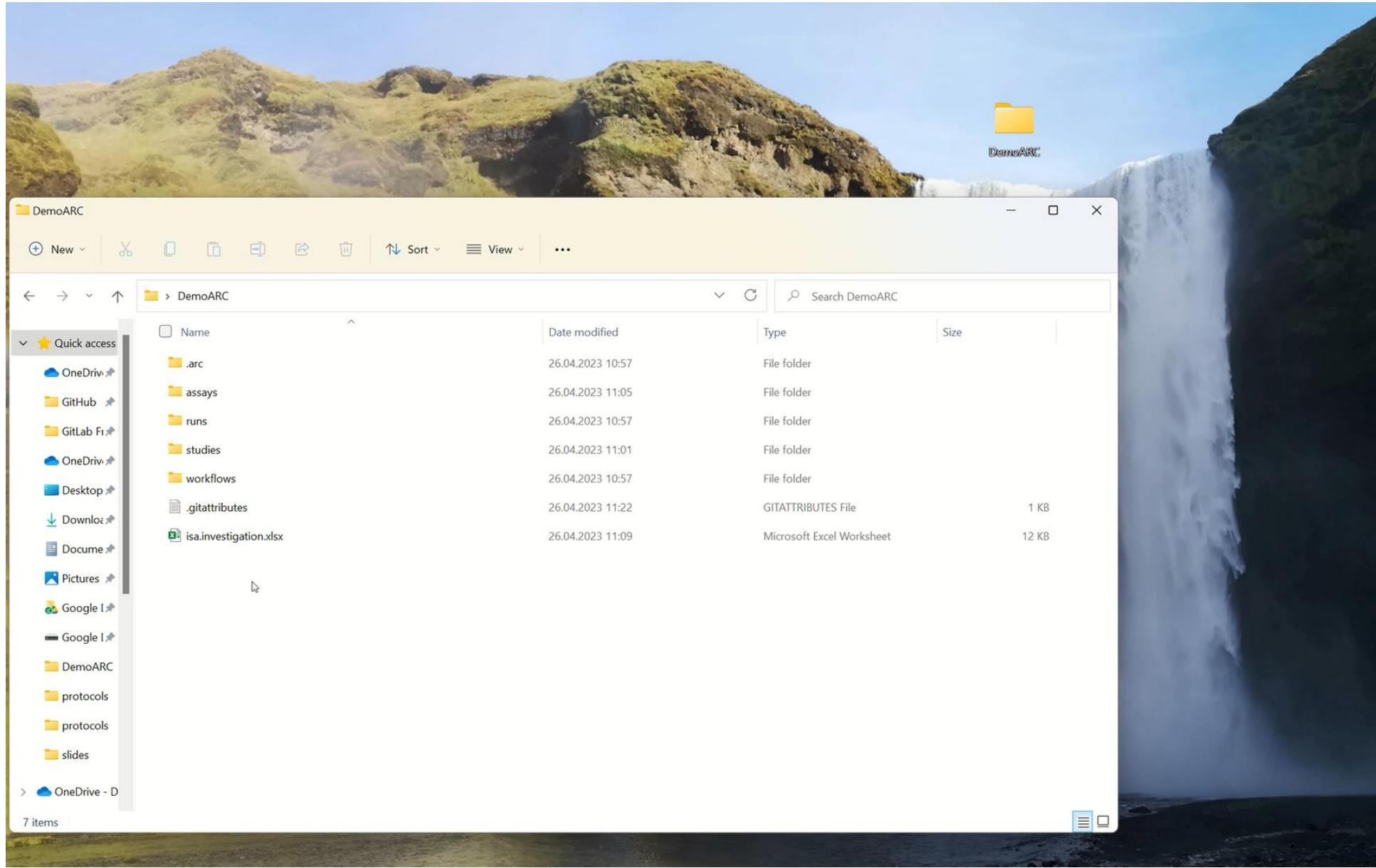
3. Check, if the upload was successful by visiting the respective URL  
→ <https://git.nfdi4plants.org/<YourUserName>/<YourARC>>

# Invite collaborators to your ARC

The screenshot shows a web browser window displaying a GitLab project named 'DemoARC'. The URL in the address bar is <https://git.nfdi4plants.org/martinkuhl/DemoARC>. The sidebar on the left contains links for Project information, Repository, Issues (0), Merge requests (0), CI/CD, Security & Compliance, Deployments, Packages and registries, Infrastructure, Monitor, Analytics, Wiki, Snippets, and Settings. The main content area shows the 'DemoARC' project details: Project ID 405, 1 Commit, 1 Branch, 0 Tags, and 123 KB Project Storage. A recent commit by Martin Kuhl is listed with the commit hash fe6cda47. Below the commit, there are buttons for Find file, Web IDE, Clone, and more. A table lists files in the repository, all of which have been updated 5 minutes ago except for one which was updated 4 minutes ago.

Name	Last commit	Last update
.arc	Update	5 minutes ago
assays	Update	5 minutes ago
runs	Update	5 minutes ago
studies	Update	5 minutes ago
workflows	Update	5 minutes ago
.gitattributes	Update	4 minutes ago

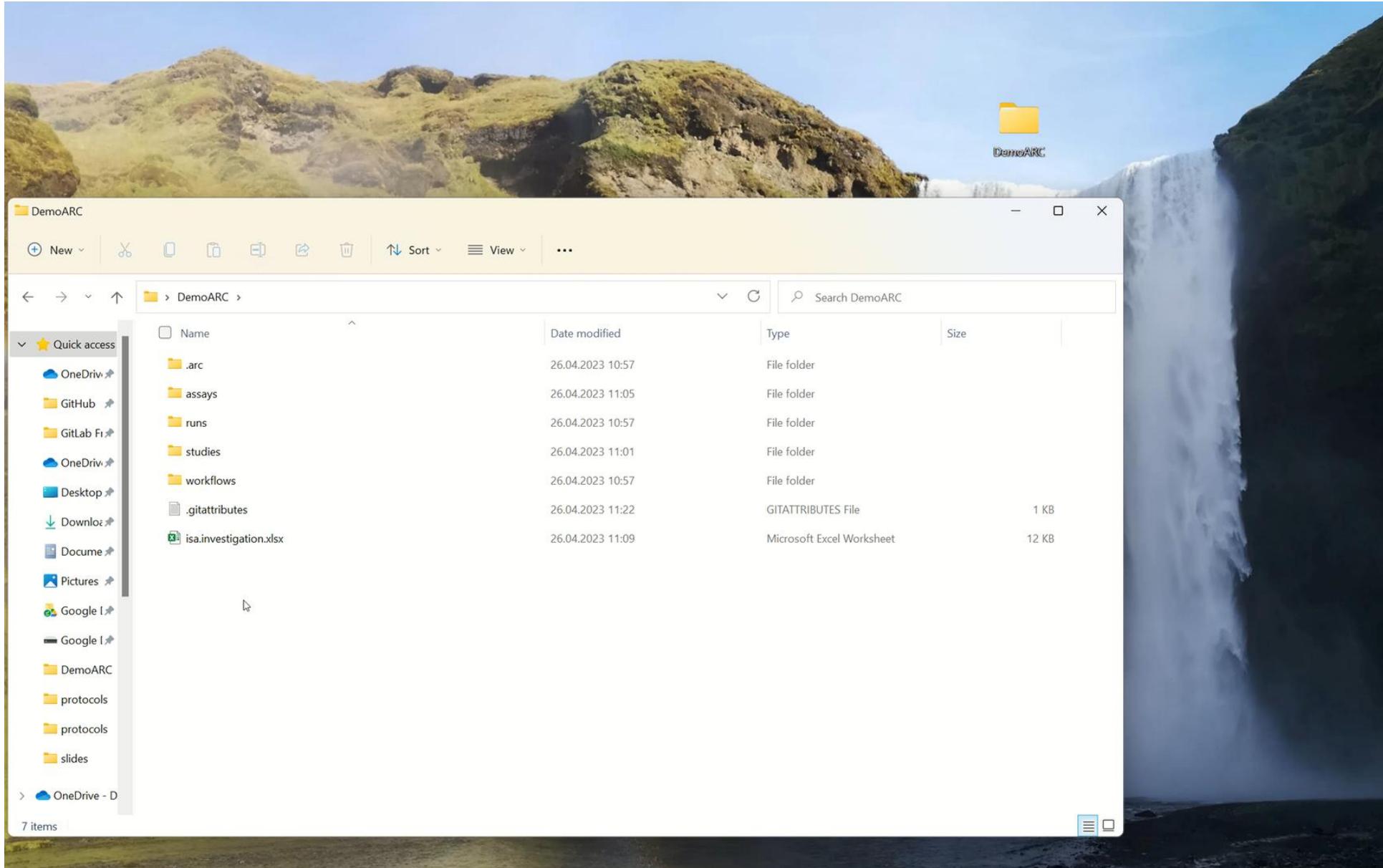
# Add metadata to your Study - building blocks



# Add metadata to your Study - ontology term search

The screenshot shows a Microsoft Excel spreadsheet titled "isa.study.xlsx" with the "Data" tab selected. The ribbon menu includes "File", "Home", "Insert", "Draw", "Page Layout", "Formulas", "Data", "Review", "View", and "Help". The "Data" tab has several groups: "Get & Transform Data", "Queries & Connections", "Data Types", "Sort & Filter", "Data Tools", "Forecast", "Swate", and "Swate". The "Sort & Filter" group contains icons for "Sort", "Filter", and "Advanced". The "Data Tools" group contains "Text to Columns", "What-If Analysis", "Forecast Sheet", "Outline", "Core", and "Experts". The "Forecast" group contains "What-If Analysis", "Forecast Sheet", "Outline", "Core", and "Experts". The "Swate" group contains "Swate" and "Swate". The "Swate" add-in is currently active, displaying a sidebar titled "Swate" with a green header. The sidebar features a search bar with the placeholder "Start typing to search for" and a button labeled "Fill selected cells with this term". Below the search bar are buttons for "Use related term directed search." and "Use advanced search". The main Excel window shows a table with columns A through H. Column A contains the header "Source Name" and a dropdown menu. Column B contains the header "Characteristic [organism]" and a dropdown menu. Column E contains the header "Characteristic [genetic population background information]" and a dropdown menu. Column H contains the header "Characteristic [Knock-out]" and a dropdown menu. The "Swate" sidebar is positioned on the right side of the Excel window.

# Add metadata to your Assay - units



# Add metadata to your Assay - units

The screenshot shows a Microsoft Excel spreadsheet titled "isa.assay.xlsx". The "Table Design" tab is selected in the ribbon. A table is present in the worksheet with columns labeled "Source Name", "Parameter [temperature]", and "Sample Name". The "Parameter [temperature]" column has a dropdown arrow indicating it is a parameter. The "Swate" add-in is open as a sidebar, showing the "Building Blocks" section. It lists "Parameter" as "temperature" and indicates that "This Parameter has a unit: degree celsius". There is a button to "Add building block". Below this, there is a section for "Add/Update unit reference to existing building block". The Swate sidebar also displays "Swate Release Version 0.7.1".

Annotation Table

Source Name Parameter [temperature] Sample Name

Parameter

This Parameter has a unit: degree celsius

Add building block

Add unit Start typing to search

Swate Release Version 0.7.1

# Add metadata to your Assay - workflow

The screenshot shows a Microsoft Excel spreadsheet titled "isa.assay.xlsx" with the "Table Design" tab selected in the ribbon. A table is being edited, with the first two rows visible:

Source Name	Parameter [temperature]	Sample Name
	32,00 degree celsius	

The "Table Style Options" group in the ribbon is expanded, showing various style options like "Header Row", "First Column", "Filter Button", etc., with "Header Row" and "Filter Button" checked. To the right, a "Table Styles" gallery shows several color-coded styles.

A "Swate" add-in window is open on the right side of the screen. It has a green header bar with the word "SWATE". Below it is a toolbar with icons for search, refresh, and help. The main area is titled "Building Blocks" and contains the following text:

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

Parameter temperature  
This Parameter has a unit: degree celsius

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 assigned, the new unit is only applied to selected rows of the selected column.

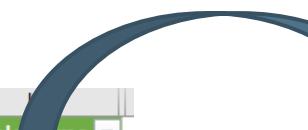
Add unit      Start typing to search

Swate Release Version 0.7.1

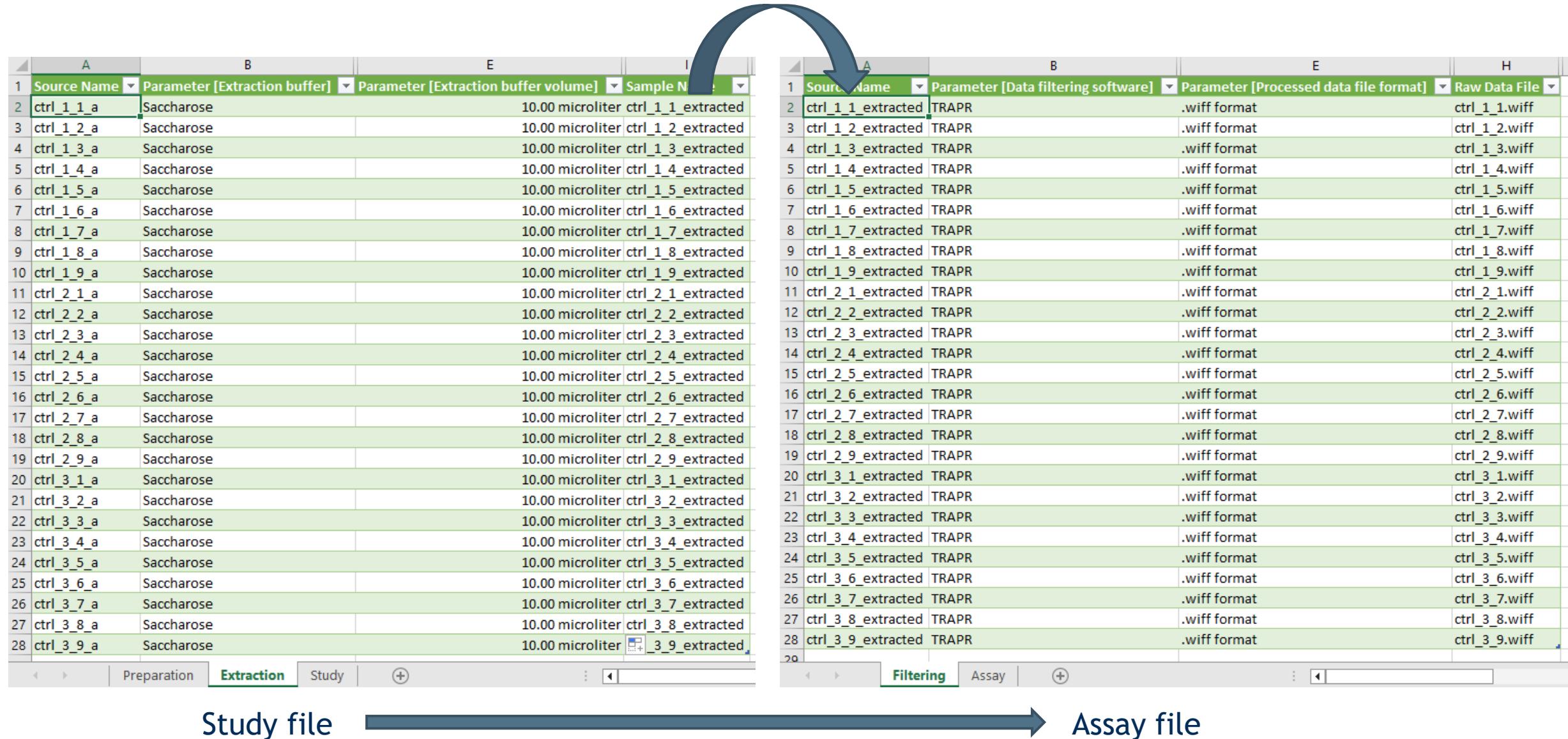
# Best practice usage: Source -> Sink

A	B	C	E	F	G	H	I	
1	Source Name	Characteristic [Organism]	Characteristic [Genotype]	Sample Name				
2	ctrl_1_1	Chlamydomonas reinhardtii	wild type genotype	ctrl_1_1_a				
3	ctrl_1_2	Chlamydomonas reinhardtii	wild type genotype	ctrl_1_2_a				
4	ctrl_1_3	Chlamydomonas reinhardtii	wild type genotype	ctrl_1_3_a				
5	ctrl_1_4	Chlamydomonas reinhardtii	wild type genotype	ctrl_1_4_a				
6	ctrl_1_5	Chlamydomonas reinhardtii	wild type genotype	ctrl_1_5_a				
7	ctrl_1_6	Chlamydomonas reinhardtii	wild type genotype	ctrl_1_6_a				
8	ctrl_1_7	Chlamydomonas reinhardtii	wild type genotype	ctrl_1_7_a				
9	ctrl_1_8	Chlamydomonas reinhardtii	wild type genotype	ctrl_1_8_a				
10	ctrl_1_9	Chlamydomonas reinhardtii	wild type genotype	ctrl_1_9_a				
11	ctrl_2_1	Chlamydomonas reinhardtii	wild type genotype	ctrl_2_1_a				
12	ctrl_2_2	Chlamydomonas reinhardtii	wild type genotype	ctrl_2_2_a				
13	ctrl_2_3	Chlamydomonas reinhardtii	wild type genotype	ctrl_2_3_a				
14	ctrl_2_4	Chlamydomonas reinhardtii	wild type genotype	ctrl_2_4_a				
15	ctrl_2_5	Chlamydomonas reinhardtii	wild type genotype	ctrl_2_5_a				
16	ctrl_2_6	Chlamydomonas reinhardtii	wild type genotype	ctrl_2_6_a				
17	ctrl_2_7	Chlamydomonas reinhardtii	wild type genotype	ctrl_2_7_a				
18	ctrl_2_8	Chlamydomonas reinhardtii	wild type genotype	ctrl_2_8_a				
19	ctrl_2_9	Chlamydomonas reinhardtii	wild type genotype	ctrl_2_9_a				
20	ctrl_3_1	Chlamydomonas reinhardtii	wild type genotype	ctrl_3_1_a				
21	ctrl_3_2	Chlamydomonas reinhardtii	wild type genotype	ctrl_3_2_a				
22	ctrl_3_3	Chlamydomonas reinhardtii	wild type genotype	ctrl_3_3_a				
23	ctrl_3_4	Chlamydomonas reinhardtii	wild type genotype	ctrl_3_4_a				
24	ctrl_3_5	Chlamydomonas reinhardtii	wild type genotype	ctrl_3_5_a				
25	ctrl_3_6	Chlamydomonas reinhardtii	wild type genotype	ctrl_3_6_a				
26	ctrl_3_7	Chlamydomonas reinhardtii	wild type genotype	ctrl_3_7_a				
27	ctrl_3_8	Chlamydomonas reinhardtii	wild type genotype	ctrl_3_8_a				
28	ctrl_3_9	Chlamydomonas reinhardtii	wild type genotype	ctrl_3_9_a				

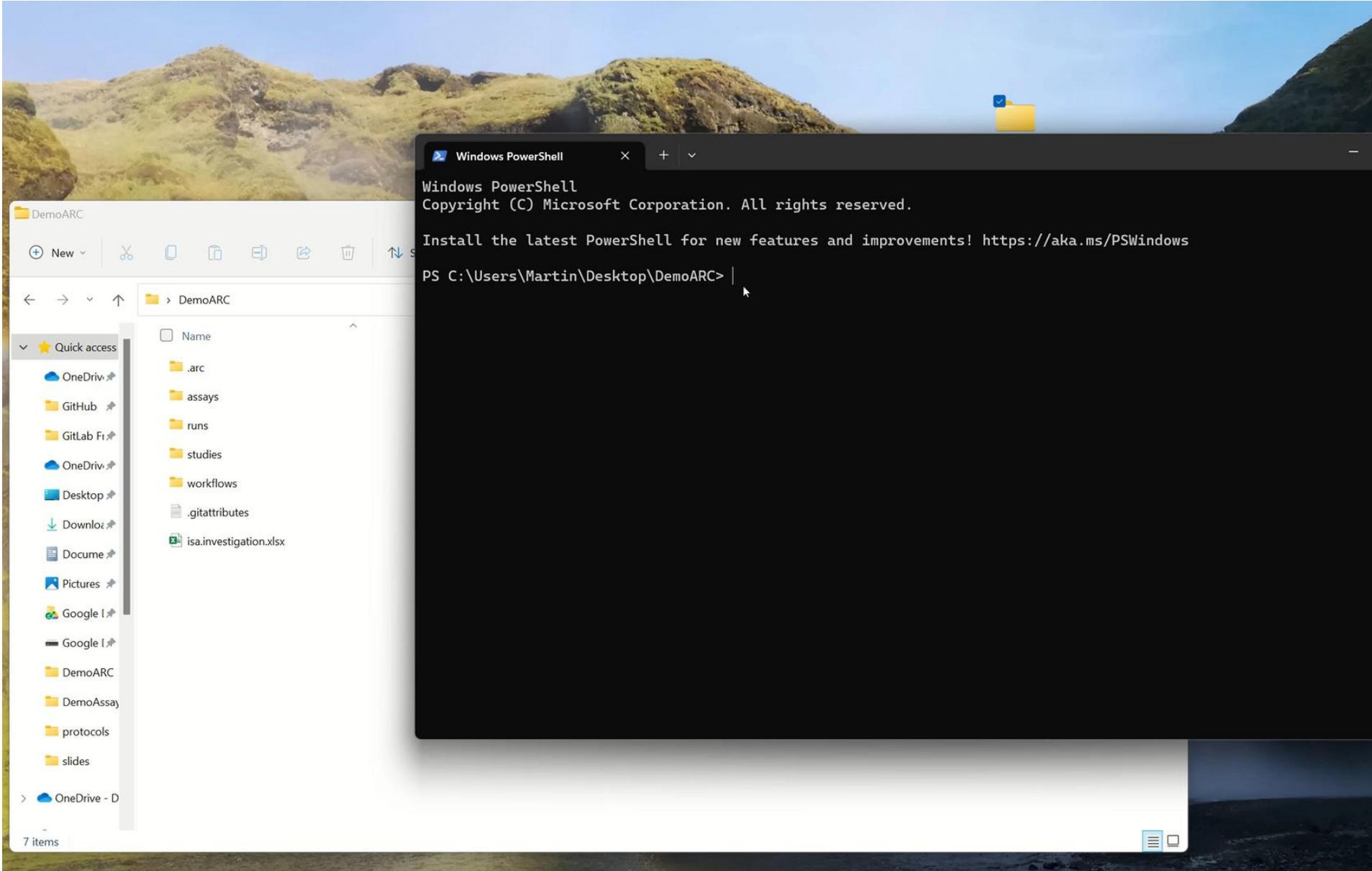
A	B	C	E	F	G	H	I	
1	Source Name	Parameter [Extraction buffer]	Parameter [Extraction buffer volume]	Sample Name				
2	ctrl_1_1_a	Saccharose	10.00 microliter	ctrl_1_1_extracted				
3	ctrl_1_2_a	Saccharose	10.00 microliter	ctrl_1_2_extracted				
4	ctrl_1_3_a	Saccharose	10.00 microliter	ctrl_1_3_extracted				
5	ctrl_1_4_a	Saccharose	10.00 microliter	ctrl_1_4_extracted				
6	ctrl_1_5_a	Saccharose	10.00 microliter	ctrl_1_5_extracted				
7	ctrl_1_6_a	Saccharose	10.00 microliter	ctrl_1_6_extracted				
8	ctrl_1_7_a	Saccharose	10.00 microliter	ctrl_1_7_extracted				
9	ctrl_1_8_a	Saccharose	10.00 microliter	ctrl_1_8_extracted				
10	ctrl_1_9_a	Saccharose	10.00 microliter	ctrl_1_9_extracted				
11	ctrl_2_1_a	Saccharose	10.00 microliter	ctrl_2_1_extracted				
12	ctrl_2_2_a	Saccharose	10.00 microliter	ctrl_2_2_extracted				
13	ctrl_2_3_a	Saccharose	10.00 microliter	ctrl_2_3_extracted				
14	ctrl_2_4_a	Saccharose	10.00 microliter	ctrl_2_4_extracted				
15	ctrl_2_5_a	Saccharose	10.00 microliter	ctrl_2_5_extracted				
16	ctrl_2_6_a	Saccharose	10.00 microliter	ctrl_2_6_extracted				
17	ctrl_2_7_a	Saccharose	10.00 microliter	ctrl_2_7_extracted				
18	ctrl_2_8_a	Saccharose	10.00 microliter	ctrl_2_8_extracted				
19	ctrl_2_9_a	Saccharose	10.00 microliter	ctrl_2_9_extracted				
20	ctrl_3_1_a	Saccharose	10.00 microliter	ctrl_3_1_extracted				
21	ctrl_3_2_a	Saccharose	10.00 microliter	ctrl_3_2_extracted				
22	ctrl_3_3_a	Saccharose	10.00 microliter	ctrl_3_3_extracted				
23	ctrl_3_4_a	Saccharose	10.00 microliter	ctrl_3_4_extracted				
24	ctrl_3_5_a	Saccharose	10.00 microliter	ctrl_3_5_extracted				
25	ctrl_3_6_a	Saccharose	10.00 microliter	ctrl_3_6_extracted				
26	ctrl_3_7_a	Saccharose	10.00 microliter	ctrl_3_7_extracted				
27	ctrl_3_8_a	Saccharose	10.00 microliter	ctrl_3_8_extracted				
28	ctrl_3_9_a	Saccharose	10.00 microliter	ctrl_3_9_extracted				



# Best practice usage: Source -> Sink



# ARC creation is an iterative process



# ARCification of Hammel et al., 2018

# Absolute Quantification of Major Photosynthetic Protein Complexes in *Chlamydomonas reinhardtii* Using Quantification Concatamers (QconCATs)

OPEN ACCESS

**Edited by:**

Wolfgang Hohenwarter,  
Leibniz-Institut für Pflanzenbiochemie  
(IPB), Germany

**Reviewed by:**

Dirk Dobritsch,  
Martin Luther University  
of Halle-Wittenberg, Germany  
Michael Hippler,  
Universität Münster, Germany

**\*Correspondence:**

Timo Mühlhaus  
[muelhaus@bio.uni-kl.de](mailto:muehlhaus@bio.uni-kl.de)  
Michael Schröda  
[schroda@bio.uni-kl.de](mailto:schroda@bio.uni-kl.de)

<sup>†</sup>These authors have contributed  
equally to this work

Alexander Hammel<sup>†</sup>, David Zimmer<sup>†</sup>, Frederik Sommert<sup>†</sup>, Timo Mühlhaus\* and Michael Schröda\*

Molekulare Biotechnologie und Systembiologie, Technische Universität Kaiserslautern, Kaiserslautern, Germany

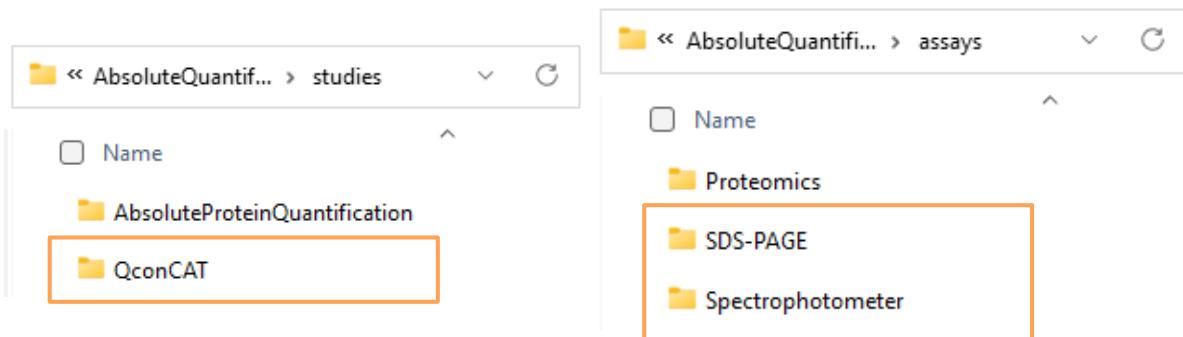
For modeling approaches in systems biology, knowledge of the absolute abundances of cellular proteins is essential. One way to gain this knowledge is the use of quantification concatamers (QconCATs), which are synthetic proteins consisting of proteotypic peptides derived from the target proteins to be quantified. The QconCAT protein is labeled with a heavy isotope upon expression in *E. coli* and known amounts of the purified protein are spiked into a whole cell protein extract. Upon tryptic digestion, labeled and unlabeled peptides are released from the QconCAT protein and the native proteins, respectively, and both are quantified by LC-MS/MS. The labeled Q-peptides then serve as standards for determining the absolute quantity of the native peptides/proteins. Here, we have applied the QconCAT approach to *Chlamydomonas reinhardtii* for the absolute quantification of the major proteins and protein complexes

# Scope of publication

- ▶ Determine absolute quantity of native peptides/proteins with quantitative peptides (Q-Peptides)
  - ▶ Expression, purification, and quantification of heavy-labeled photosynthesis QconCATs in *E.coli*
- ▶ Protein extraction and precipitation of *C. reinhardtii* whole cell extracts
  - ▶ QconCATs spiked into this extract
- ▶ Tryptic digest and quantification using LC-MS/MS

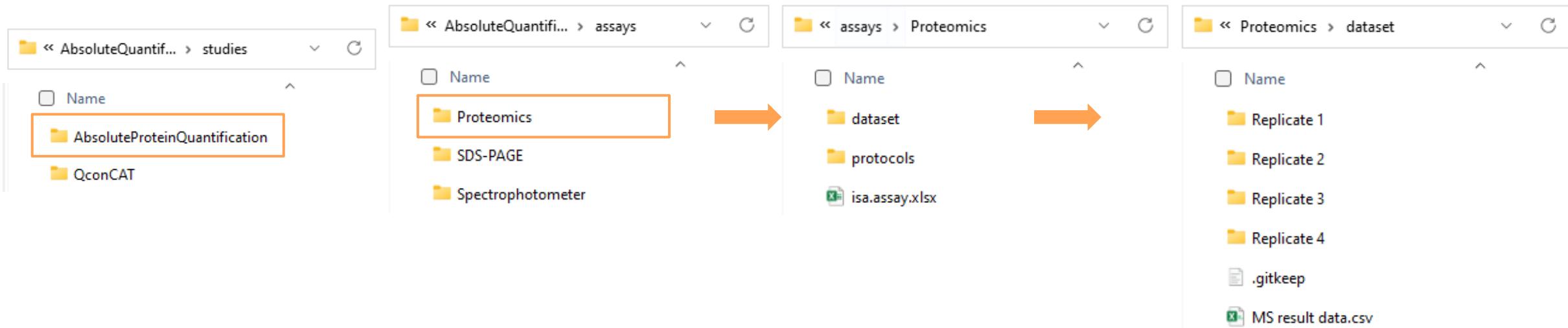
# Potential ARC layout

- ▶ Study about QconCAT expression
  - ▶ SDS-PAGE assay for QconCAT purification
  - ▶ Spectrophotometer assay for QconCAT quantification



# Potential ARC layout

- ▶ Second study about *Chlamydomonas* growth and absolute protein quantification
  - ▶ Proteomics assay to this study and place the MS data in the correct location

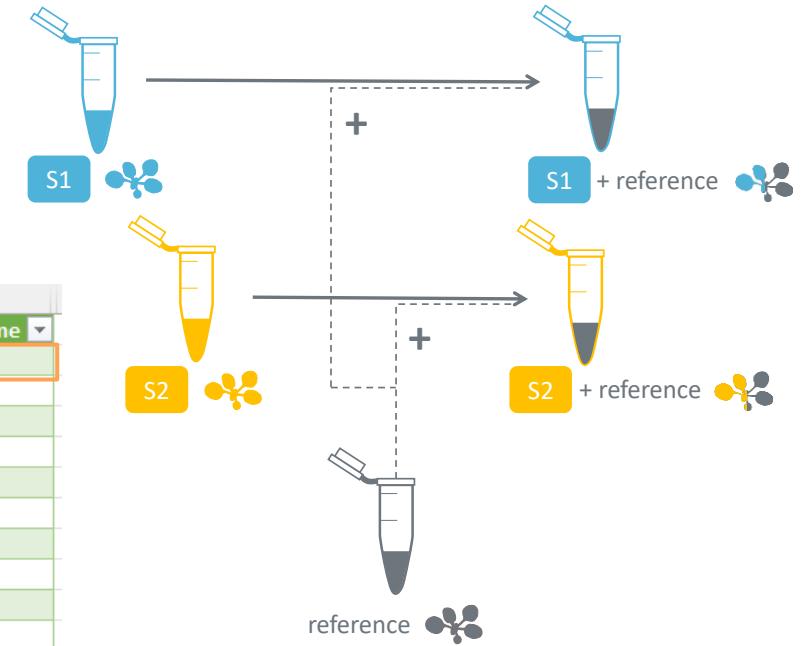


# Spiking samples

A	B	E	H
Source Name	Parameter [Cellular Compartment Analysis]	Parameter [sample state]	Parameter [Protein amount]
1 C1-1-1	whole cell	solution	50,25 microgram
2 C1-1-1	whole cell	solution	50,50 microgram
3 C1-1-1	whole cell	solution	52,50 microgram
4 C1-1-1	whole cell	solution	55,00 microgram
5 C1-2-1	whole cell	solution	50,20 microgram
6 C1-2-1	whole cell	solution	50,39 microgram
7 C1-2-1	whole cell	solution	51,95 microgram
8 C1-2-1	whole cell	solution	53,90 microgram
9 C1-3-1	whole cell	solution	50,25 microgram
11 C1-3-1	whole cell	solution	50,50 microgram
12 C1-3-1	whole cell	solution	52,50 microgram
13 C1-3-1	whole cell	solution	55,00 microgram
14 C1-4-1	whole cell	solution	50,25 microgram
15 C1-4-1	whole cell	solution	50,50 microgram
16 C1-4-1	whole cell	solution	52,50 microgram
17 C1-4-1	whole cell	solution	55,00 microgram
18 QconCAT	whole cell	solution	50,25 microgram
19 QconCAT	whole cell	solution	50,50 microgram
20 QconCAT	whole cell	solution	52,50 microgram
21 QconCAT	whole cell	solution	55,00 microgram
22 QconCAT	whole cell	solution	50,20 microgram
23 QconCAT	whole cell	solution	50,39 microgram
24 QconCAT	whole cell	solution	51,95 microgram
25 QconCAT	whole cell	solution	53,90 microgram
26 QconCAT	whole cell	solution	50,25 microgram
27 QconCAT	whole cell	solution	50,50 microgram
28 QconCAT	whole cell	solution	52,50 microgram
29 QconCAT	whole cell	solution	55,00 microgram
30 QconCAT	whole cell	solution	50,25 microgram
31 QconCAT	whole cell	solution	50,50 microgram
32 QconCAT	whole cell	solution	52,50 microgram
33 QconCAT	whole cell	solution	55,00 microgram

AC	AG
Parameter [urea]	Sample Name
8,00 molar	S1-1-1
8,00 molar	S1-1-2
8,00 molar	S1-1-3
8,00 molar	S1-1-4
8,00 molar	S1-2-1
8,00 molar	S1-2-2
8,00 molar	S1-2-3
8,00 molar	S1-2-4
8,00 molar	S1-3-1
8,00 molar	S1-3-2
8,00 molar	S1-3-3
8,00 molar	S1-3-4
8,00 molar	S1-4-1
8,00 molar	S1-4-2
8,00 molar	S1-4-3
8,00 molar	S1-4-4
8,00 molar	S1-1-1
8,00 molar	S1-1-2
8,00 molar	S1-1-3
8,00 molar	S1-1-4
8,00 molar	S1-2-1
8,00 molar	S1-2-2
8,00 molar	S1-2-3
8,00 molar	S1-2-4
8,00 molar	S1-3-1
8,00 molar	S1-3-2
8,00 molar	S1-3-3
8,00 molar	S1-3-4
8,00 molar	S1-4-1
8,00 molar	S1-4-2
8,00 molar	S1-4-3
8,00 molar	S1-4-4

...



# Initialize an ARC with a study on *Chlamydomonas* growth and protein quantification

1. *arc init*
2. *arc i create*
3. *arc s add*
  - ▶ Give your study a unique and precise identifier in the editor
4. *arc a add*
  - ▶ Specify the study to which you want to add the assay
  - ▶ Give your assay a unique and precise identifier
5. Place the MS data in the correct location within the ARC

# Add a second study with two assays to the ARC on QconCAT expression and purification

1. Add the study
  - ▶ Give your study a unique and precise identifier in the editor
2. Add an assay to the study
  - ▶ Specify the study to which you want to add the assay
  - ▶ Give your assay a unique and precise identifier
3. Repeat and add a second assay to the study
4. Place the protein sequence, the western blot and MS data in the correct location within the ARC

# Synchronize your ARC to the DataHUB

1. *arc remote token get*
2. *arc sync -m „added study and assays“*

# Some useful links

- **DataPLANT homepage:** <https://nfdi4plants.org/>
- **DataHUB:** <https://git.nfdi4plants.org/>
- **Knowledge Base:** <https://nfdi4plants.org/nfdi4plants.knowledgebase/>
- **ARC User Journey:** [DataPLANT documentation - ARC User Journey \(nfdi4plants.org\)](#)
- **ARC Commander QuickStart:** [DataPLANT documentation - ARC Commander QuickStart \(nfdi4plants.org\)](#)

Thank you for your attention