



CEPLAS

Cluster of Excellence on Plant Sciences

Good Data Management Practices

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

November 10th, 2024

Vittorio Tracanna, Dominik Brilhaus
CEPLAS Data



House-keeping

Pad: <https://pad.hhu.de/0NdPK05LQ5CHBRN2iuG91Q>

Materials

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

Goals

- Appreciate FAIR principles
- Learn tools and services for FAIR data management
- Effectively manage your own research data

Introduce yourself

- Used code / programming language before
- Experience with Git / GitLab / GitHub?
- Have an ORCID
- My motivation / expectation
- My favorite lab assay

Let's draw a typical lab workflow



Why Research Data Management (RDM)?

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

How is your data analysis going?

Can't understand the data

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

That is terrible and so
cruel !

Who is it, who collected the
data ?

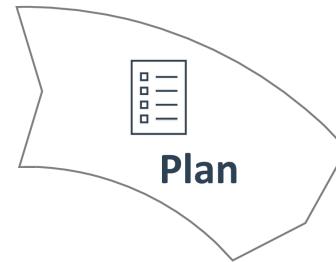
I did... 3 years ago



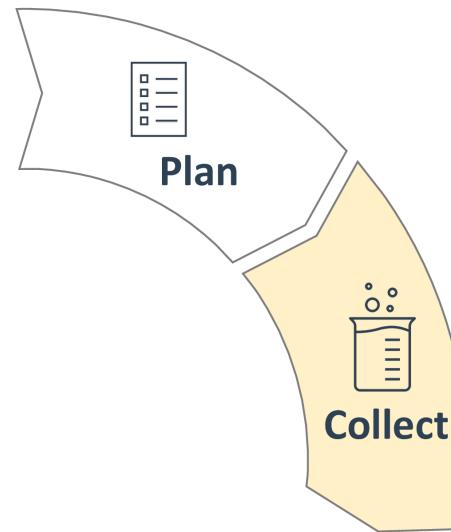
Your first collaborators
are your future selves,
be nice to them !

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

The Research Data Lifecycle



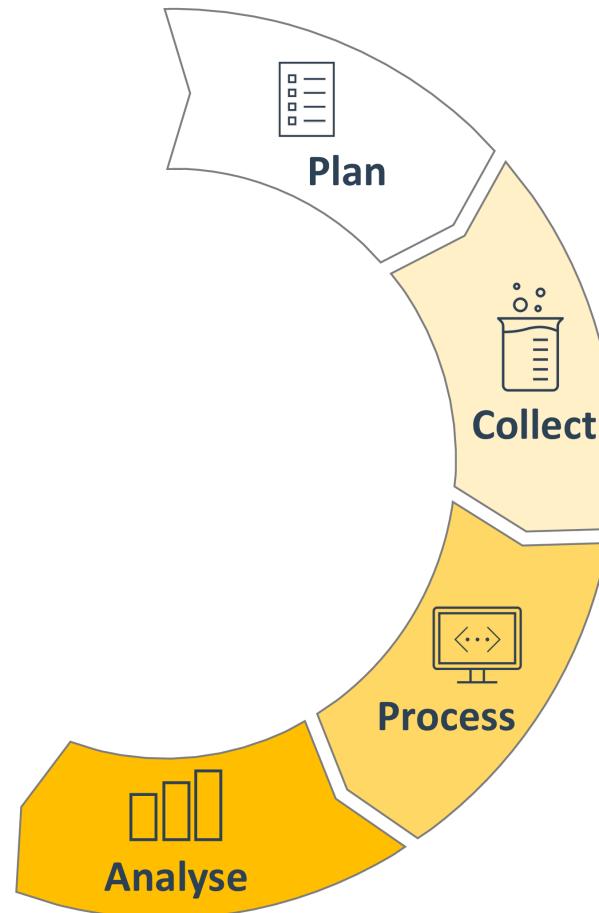
The Research Data Lifecycle



The Research Data Lifecycle



The Research Data Lifecycle



The Research Data Lifecycle



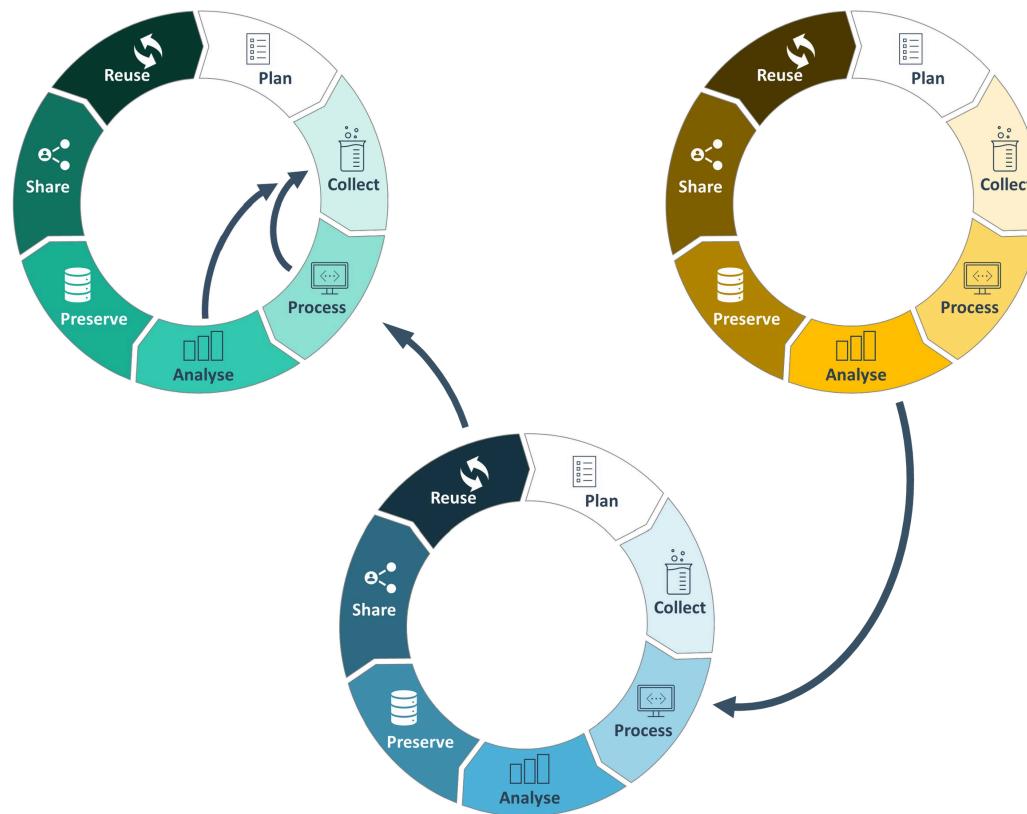
The Research Data Lifecycle



The Research Data Lifecycle



The Research Data Lifecycle *is mutable*



FAIR

- Findable
- Accessible
- Interoperable
- Reusable

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

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

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

The FAIR Guiding Principles for scientific data management and stewardship

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

— Show fewer authors

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

The FAIR principles

<p>Findable Accessible Interoperable Reusable</p>	<p>Easier collaboration & sharing</p> <pre>graph TD; A(()) --> B(()); A(()) --> C(()); A(()) --> D(()); B(()) --> E(()); C(()) --> E(()); D(()) --> E(());</pre>	<p>Increased findability and visibility</p>	<p>Reproducibility</p>																								
<p>Added-value to the research community</p> <p>nfdi NCBI EMBL-EBI</p>	<p>Compliance with funding policies</p> <pre>graph LR; A[Checkmark Document] --- B[DFG]; A --- C[EU Flag]</pre>	<p>Receive due credit</p> <p>FAIR</p> <p>Reuse</p> <p>Citations</p> <table border="1"><caption>Data from graph</caption><thead><tr><th>Time</th><th>Reuse</th><th>Citations</th></tr></thead><tbody><tr><td>Initial</td><td>Low</td><td>Low</td></tr><tr><td>Middle</td><td>Medium</td><td>Medium</td></tr><tr><td>Final</td><td>High</td><td>High</td></tr></tbody></table>	Time	Reuse	Citations	Initial	Low	Low	Middle	Medium	Medium	Final	High	High	<p>Saves time & workload</p> <p>Time wasted</p> <table border="1"><caption>Data from graph</caption><thead><tr><th>Time</th><th>Red Line (Time wasted)</th><th>Blue Line (Saves time)</th></tr></thead><tbody><tr><td>Initial</td><td>High</td><td>Low</td></tr><tr><td>Middle</td><td>Medium</td><td>Medium</td></tr><tr><td>Final</td><td>Low</td><td>High</td></tr></tbody></table>	Time	Red Line (Time wasted)	Blue Line (Saves time)	Initial	High	Low	Middle	Medium	Medium	Final	Low	High
Time	Reuse	Citations																									
Initial	Low	Low																									
Middle	Medium	Medium																									
Final	High	High																									
Time	Red Line (Time wasted)	Blue Line (Saves time)																									
Initial	High	Low																									
Middle	Medium	Medium																									
Final	Low	High																									

Is your data FAIR?

Findable | Accessible | Interoperable | Reusable

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



Findable

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

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

Accessible

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

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

Interoperable

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

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

Reusable

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

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

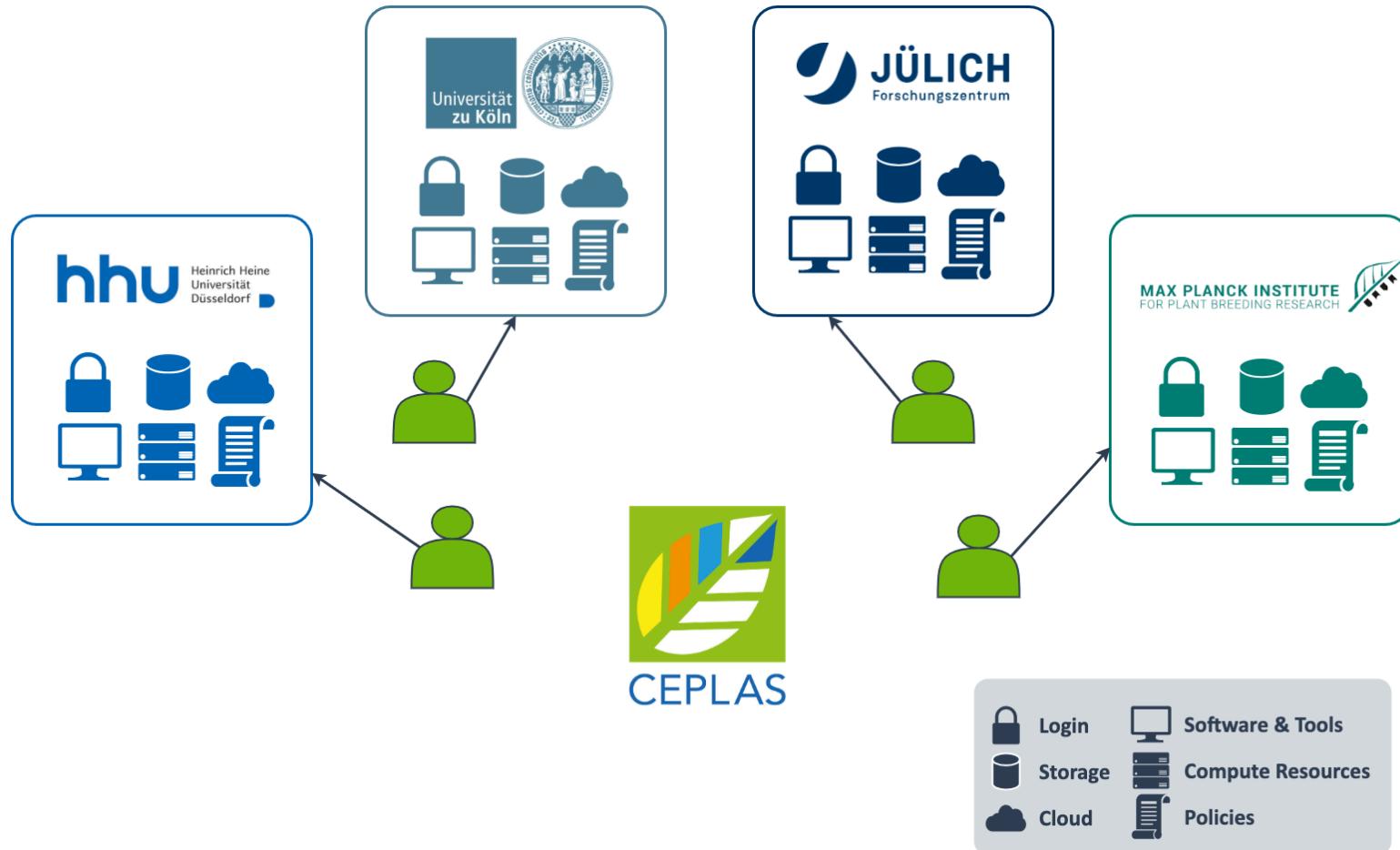
FAIR on multiple layers

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

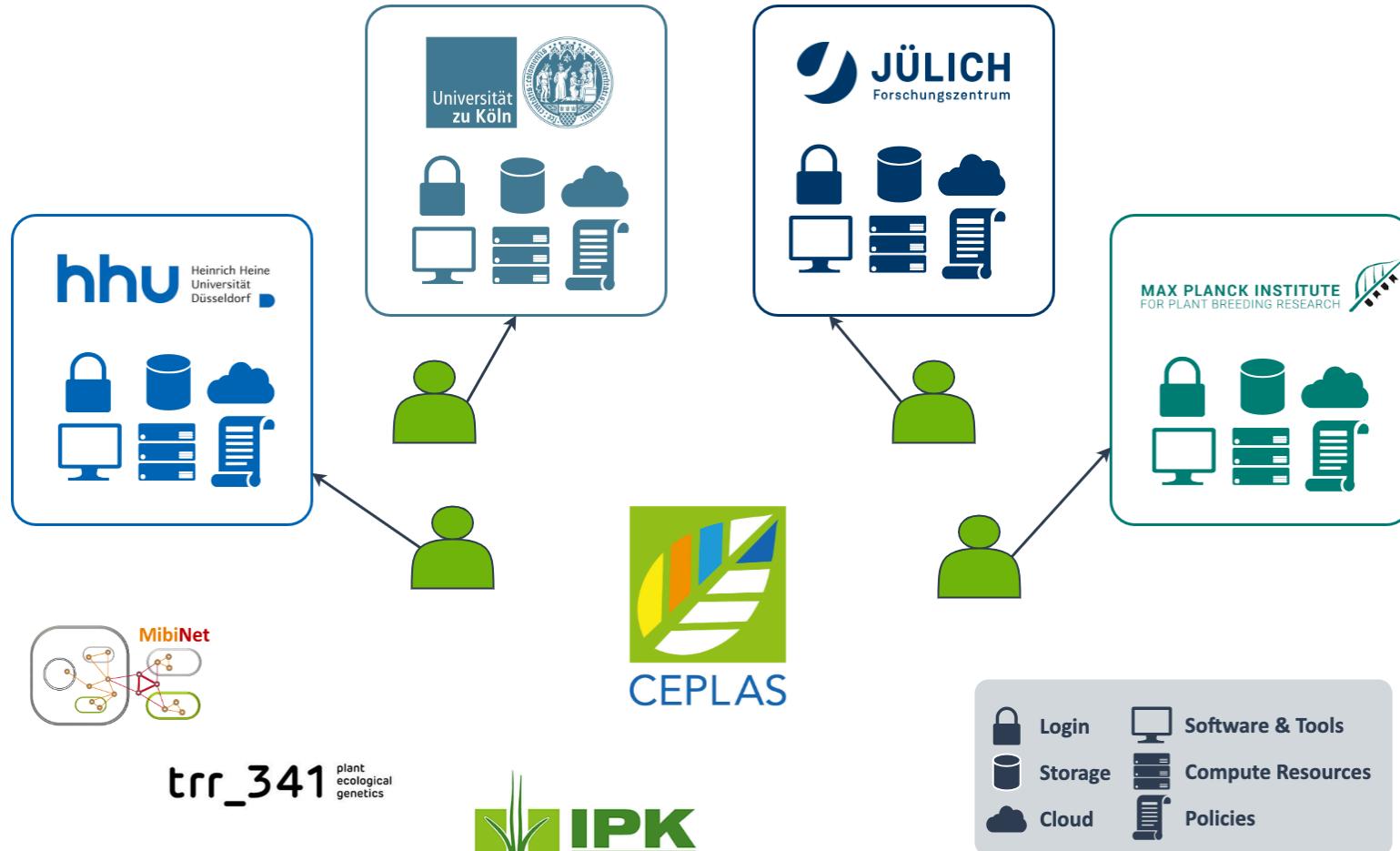
CEPLAS – One cluster, four locations



Data silos impede collaboration

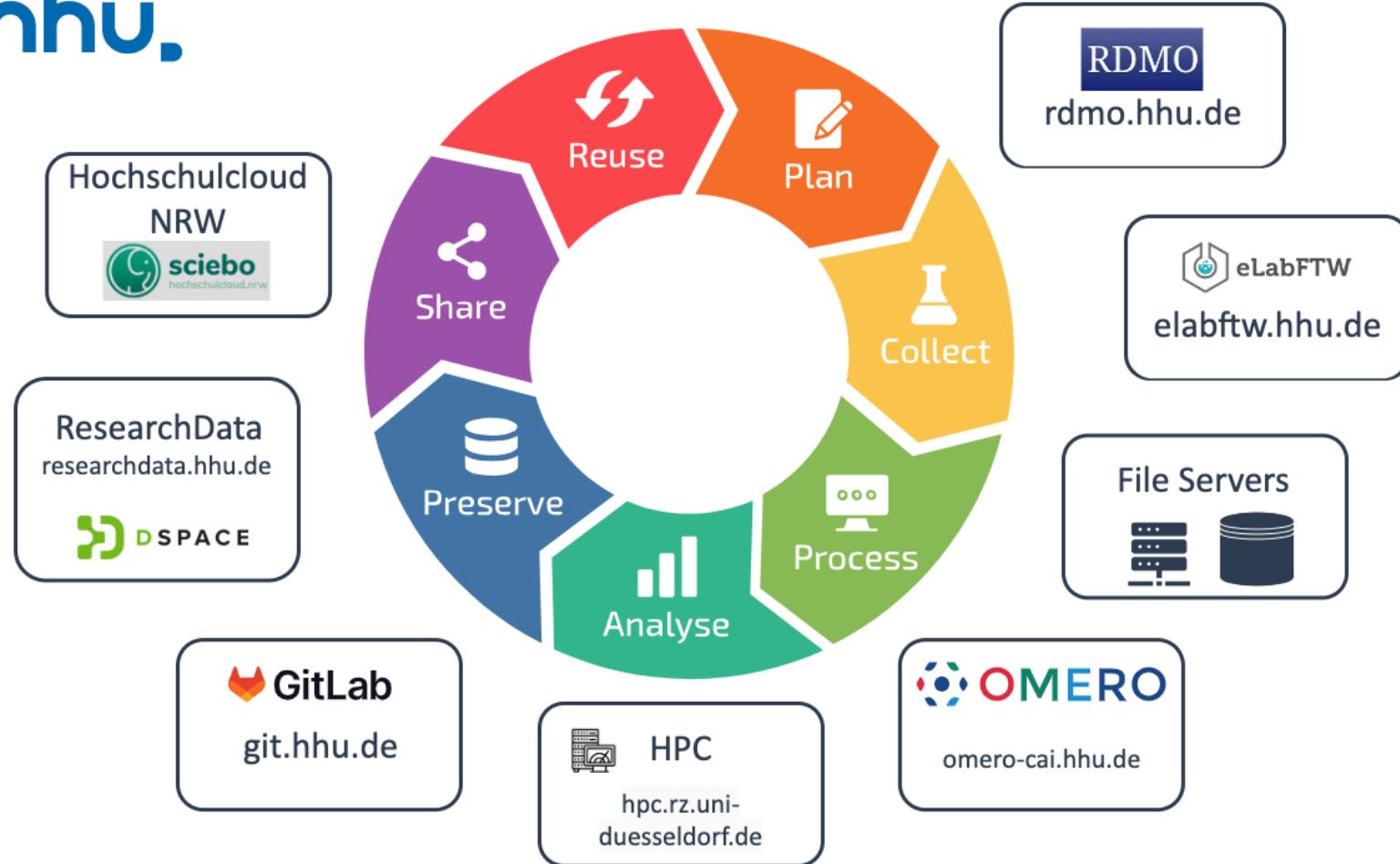


Data silos impede collaboration



Missing interfaces impede collaboration

hhu.





Understand your colleague's project



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

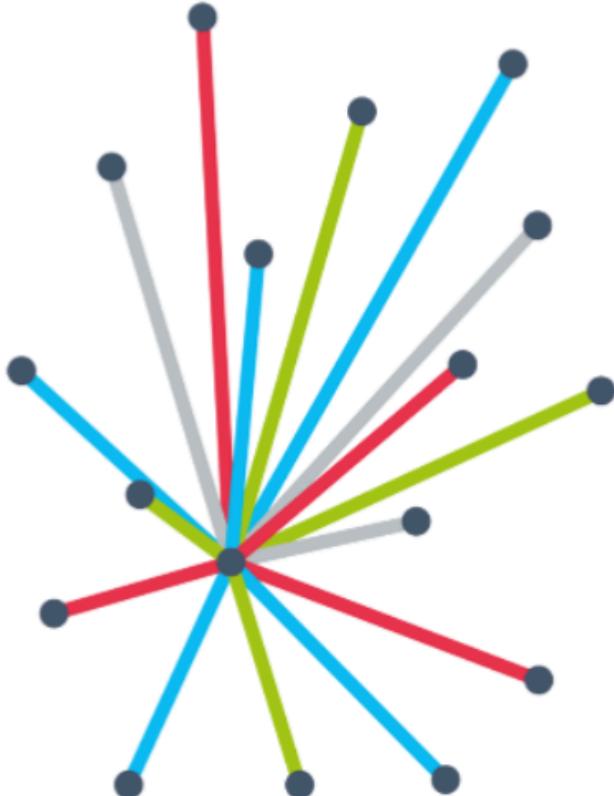
Assignment

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

Let's discuss your findings



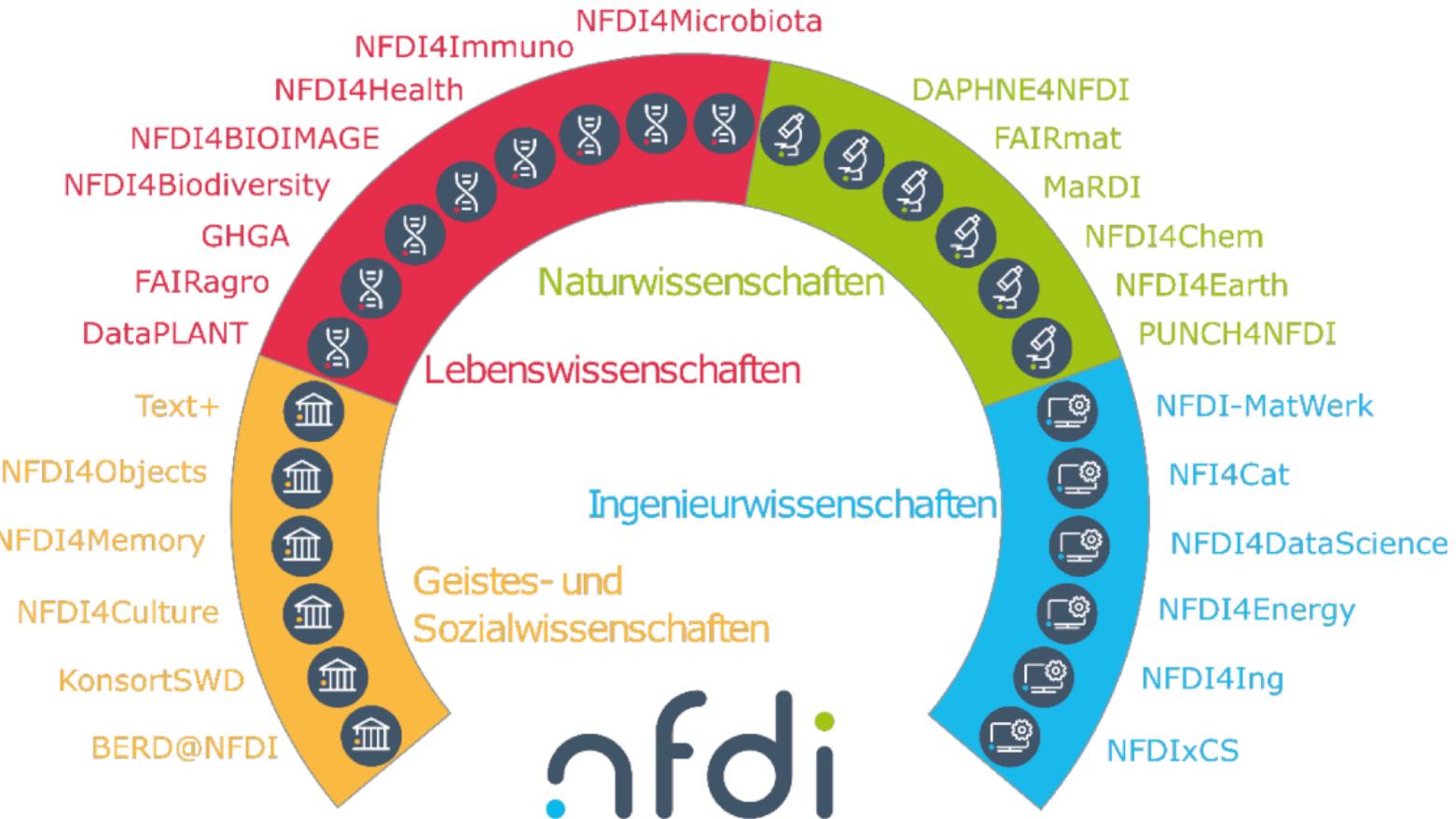
NFDI – Nationale Forschungsdaten Infrastruktur



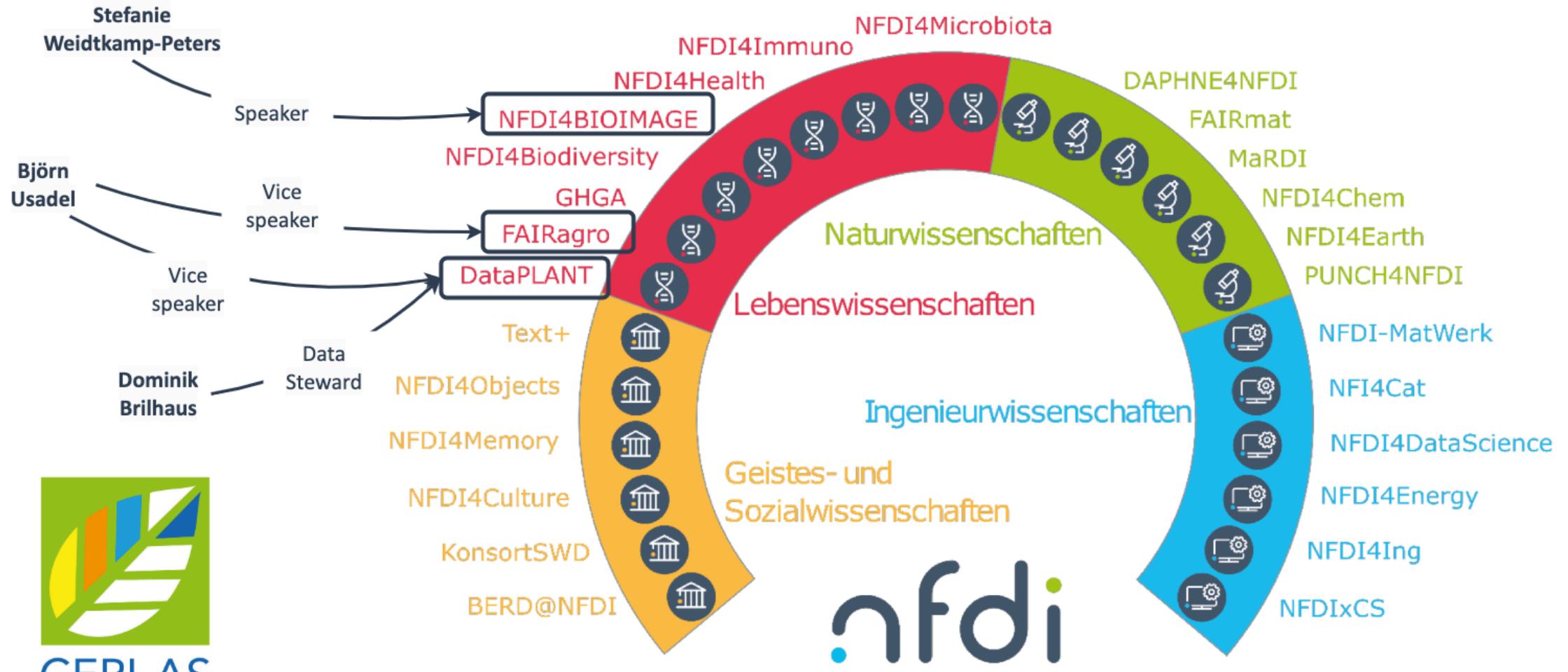
The aim of the national research data infrastructure (NFDI) is to systematically manage scientific and research data, provide long-term data storage, backup and accessibility, and network the data both nationally and internationally.

nfdi.de

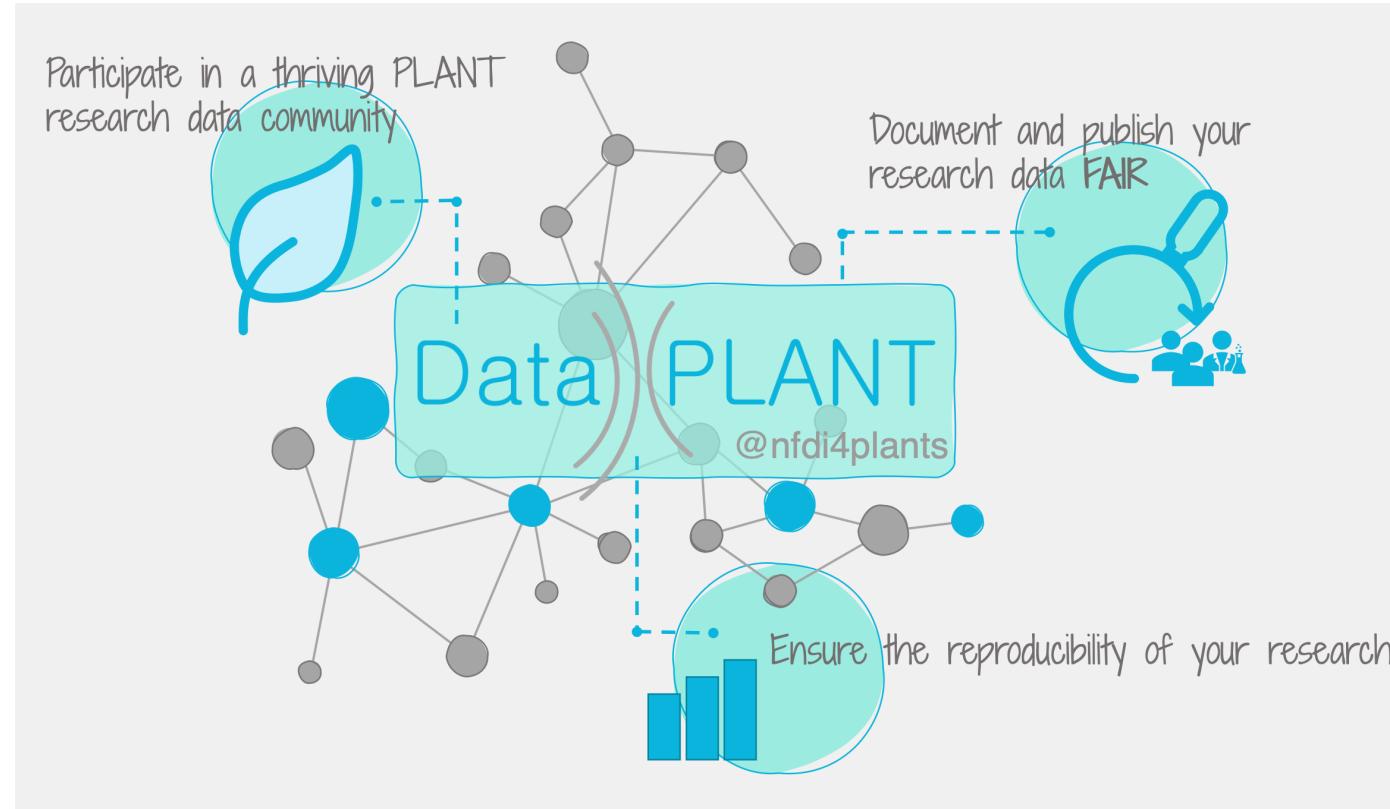
NFDI – A community-targeted approach for RDM



CEPLAS connection to the NFDI



DataPLANT – NFDI4plants



- <https://nfdi4plants.org>
- <https://arc-rdm.org>

Data Stewardship between DataPLANT and the community

Community

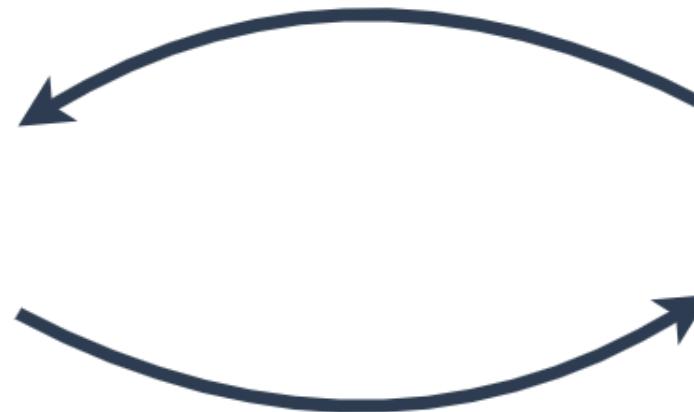


Domain experts
User experience
Training

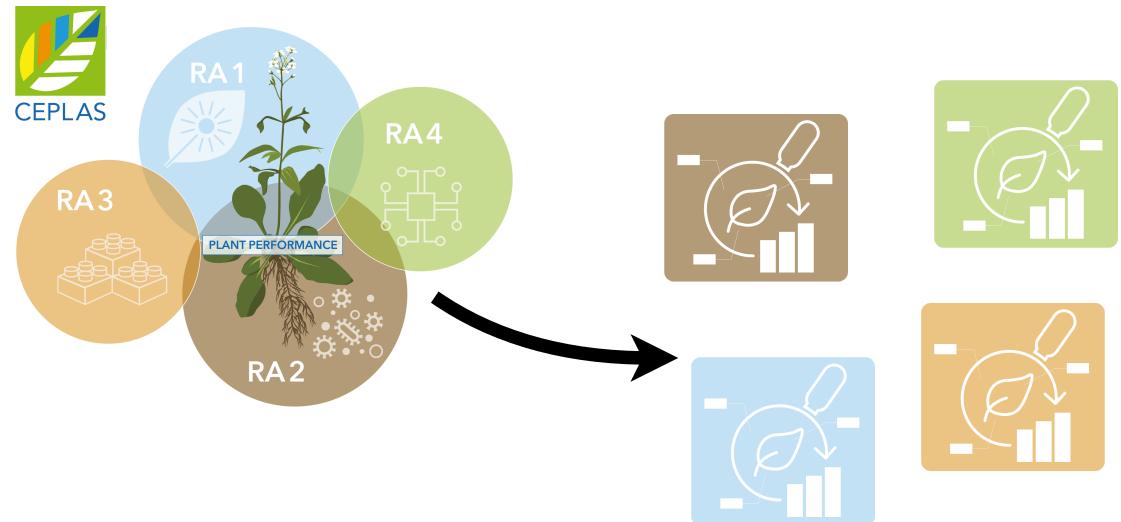
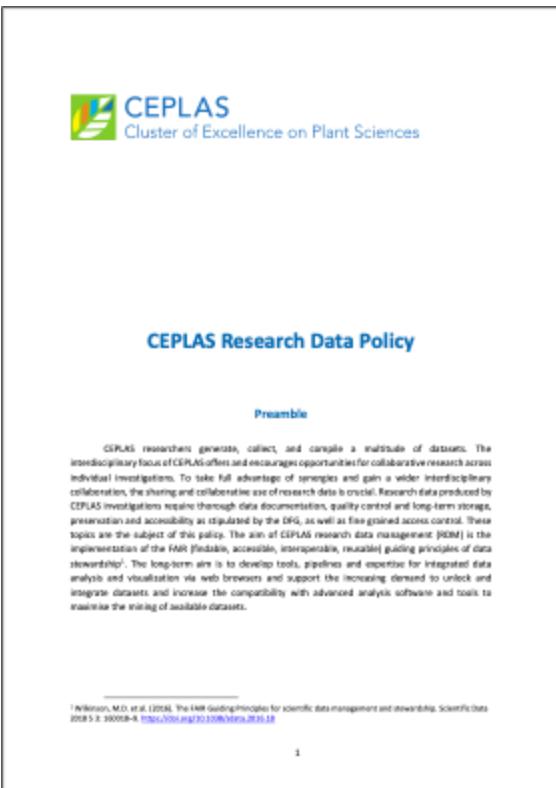
nfdi4plants



Service provider
Developers
Tech experts

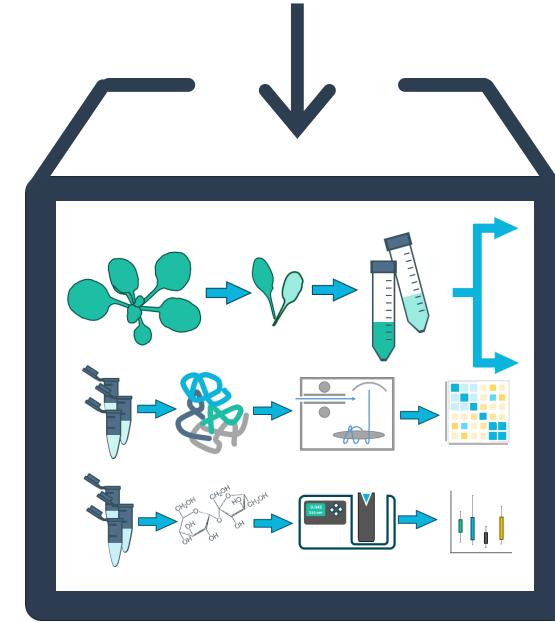
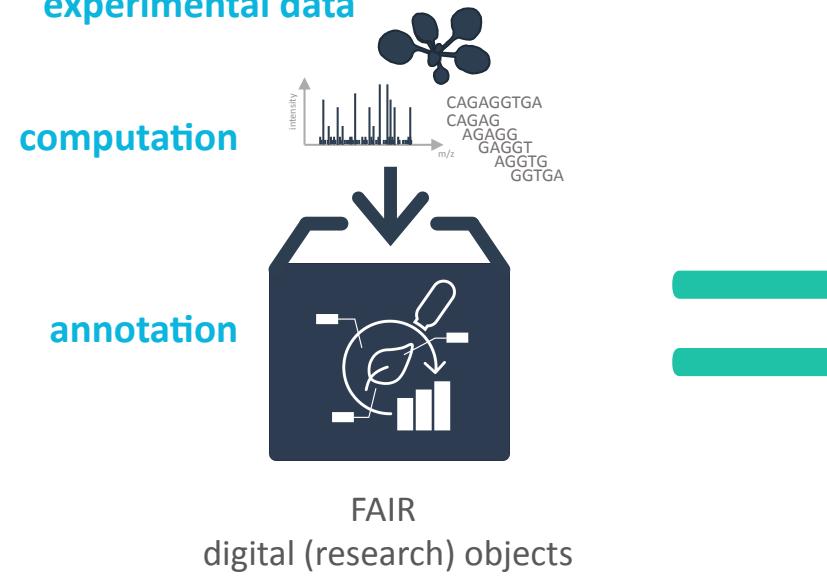


CEPLAS Research Data Policy



The policy is available via [CEPLAS intranet](#)

Annotated Research Context (ARC)

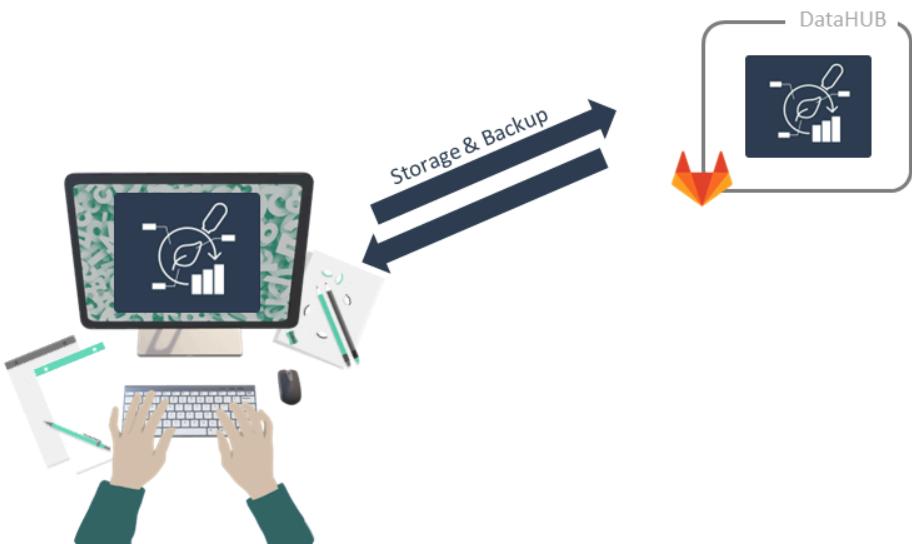


Your entire investigation in a single unified bag

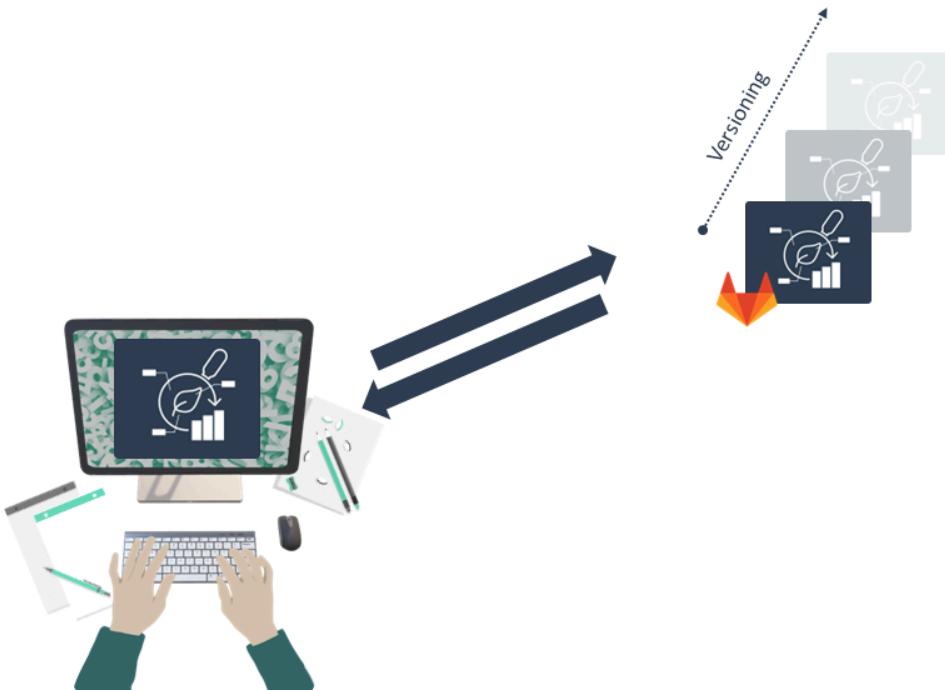
You can store your ARC in the DataHUB



You can store your ARC in the DataHUB



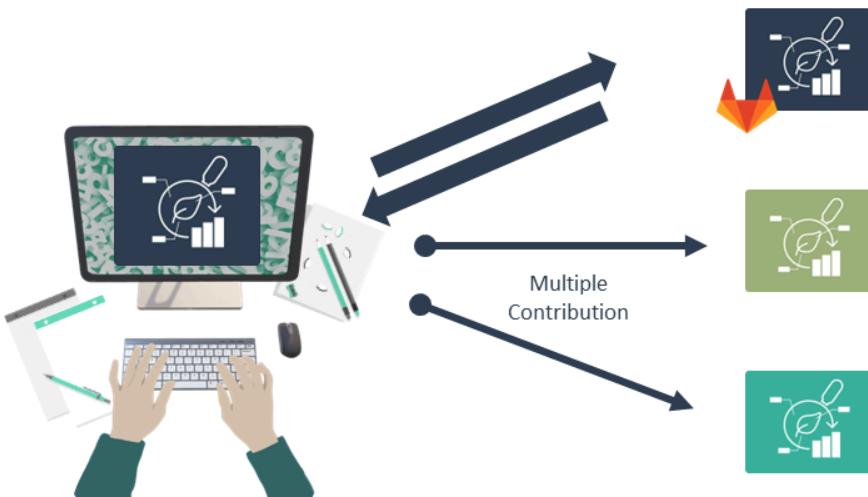
ARCs are versioned



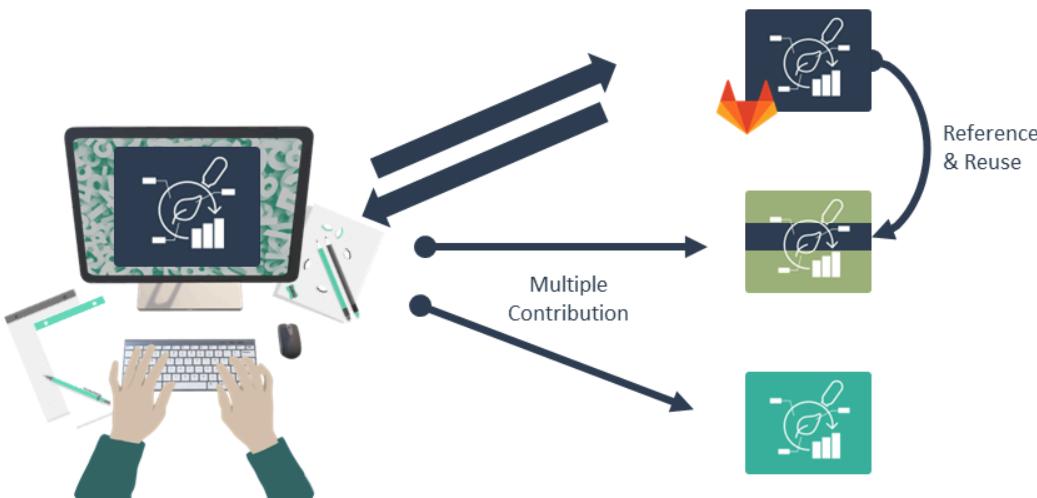
You can invite collaborators



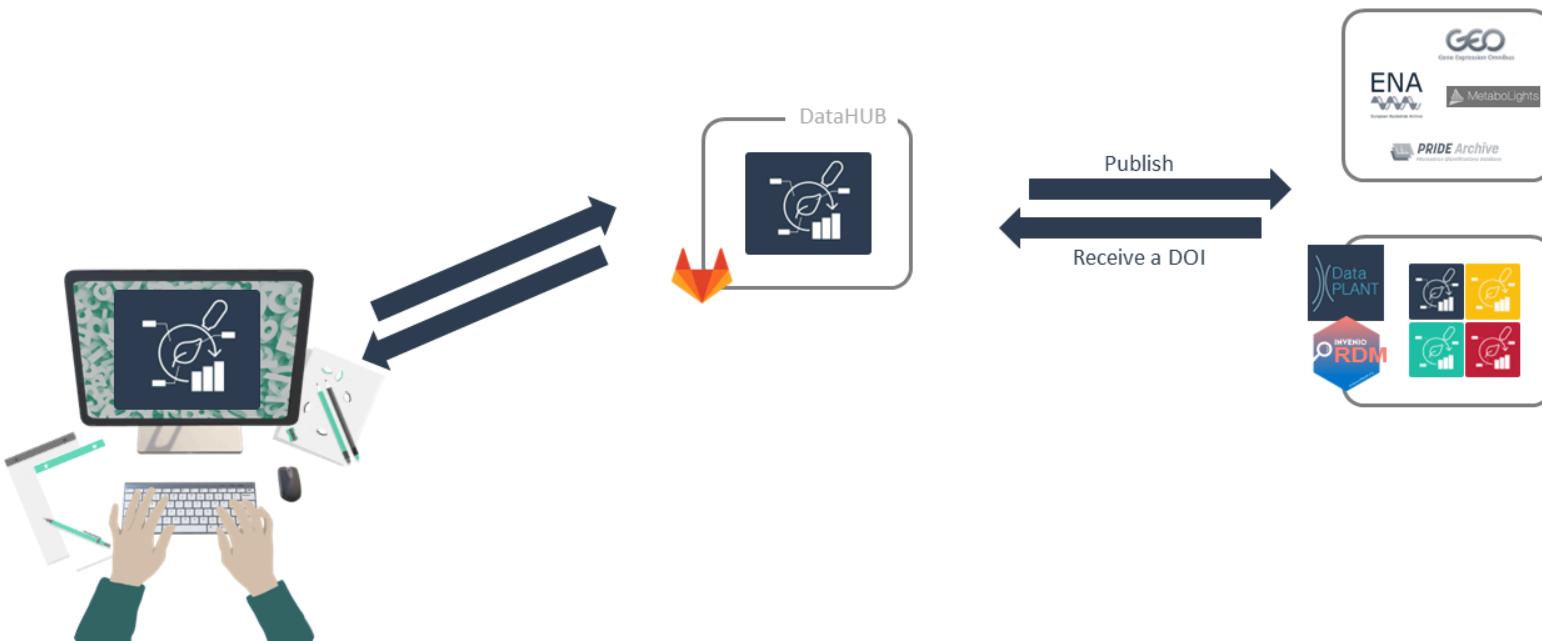
Collaborate and contribute



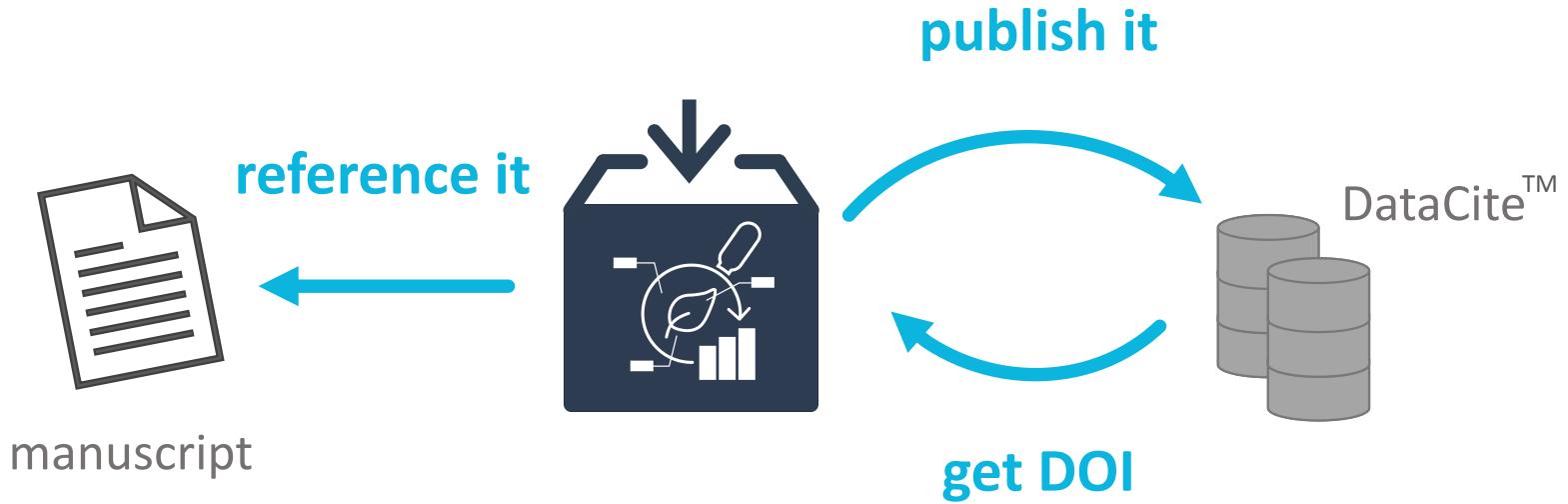
Reuse data in ARCs



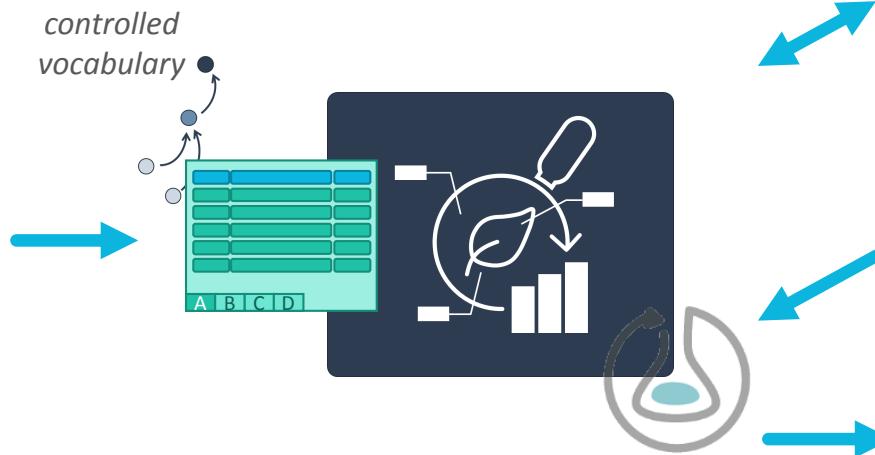
Publish your ARC



Publish your ARC, get a DOI



ARC as single-entry point



specialized endpoints

ENA
European Nucleotide Archive

GEO
Gene Expression Omnibus

EBI: MetaboLights

PRIDE Archive
Proteomics Identifications Database

BioImage Archive

dataset search

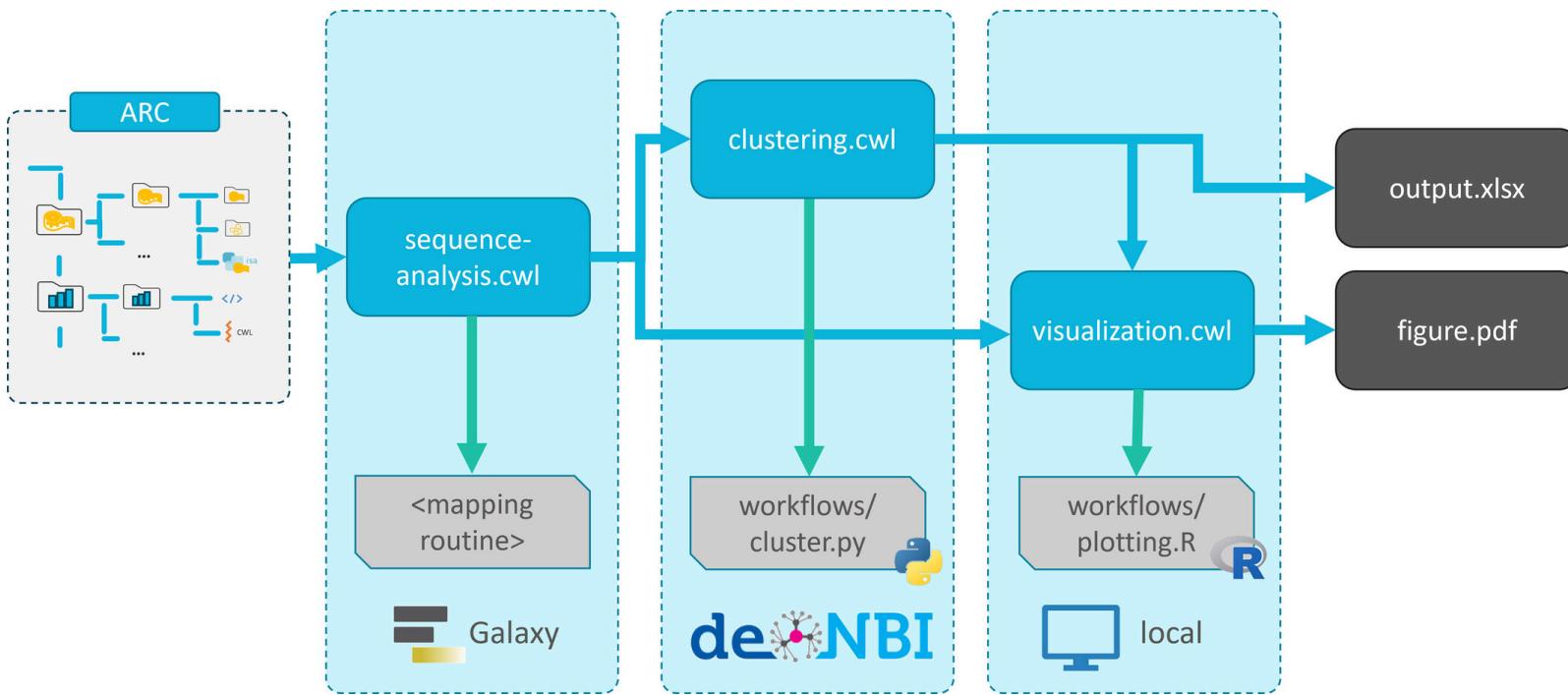
Google
OpenAIRE

The
Dataverse®
Project

INVENIO

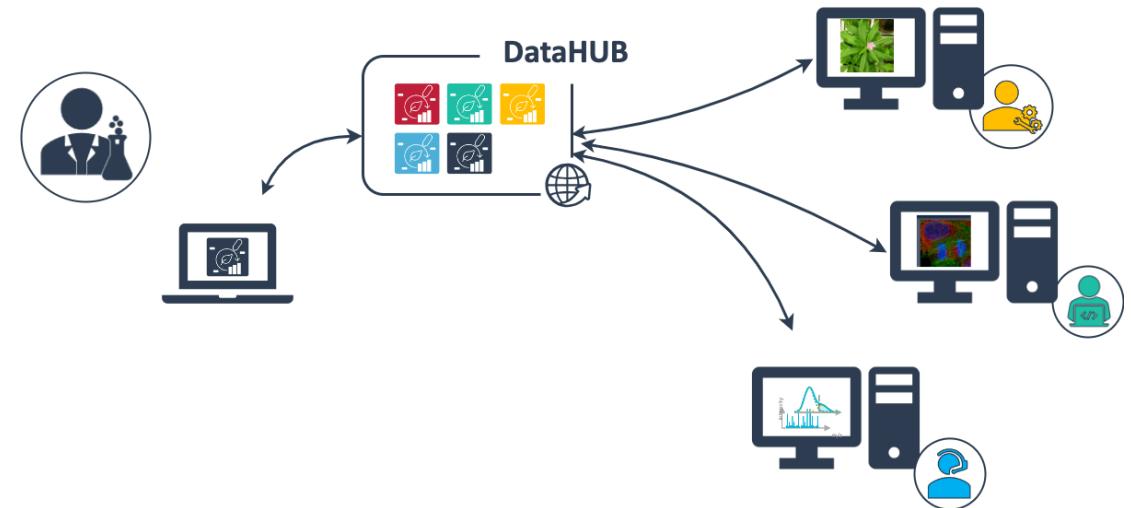
data publication

Data analysis and workflows

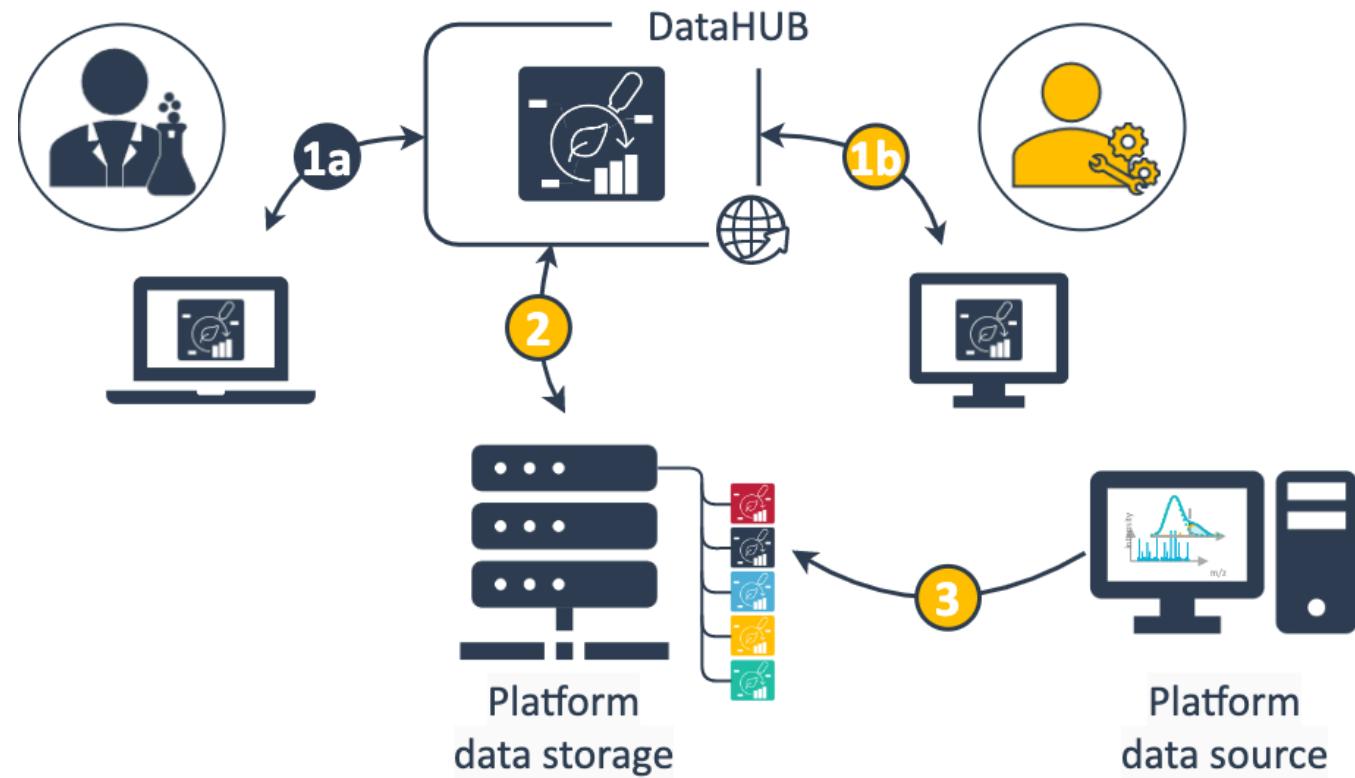


Enabling platforms

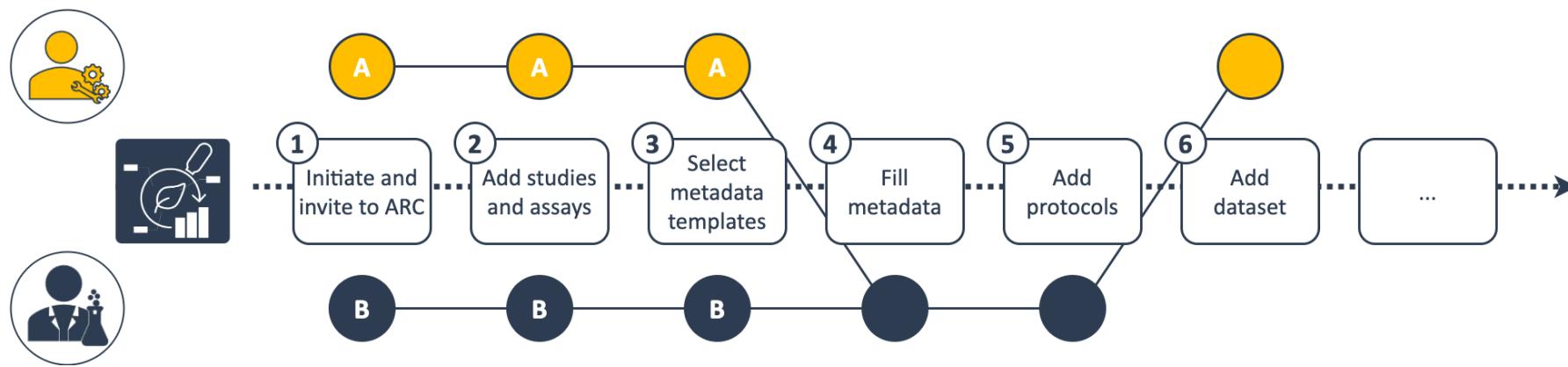
- Streamlined exchange of (meta)data
- Communication and project management



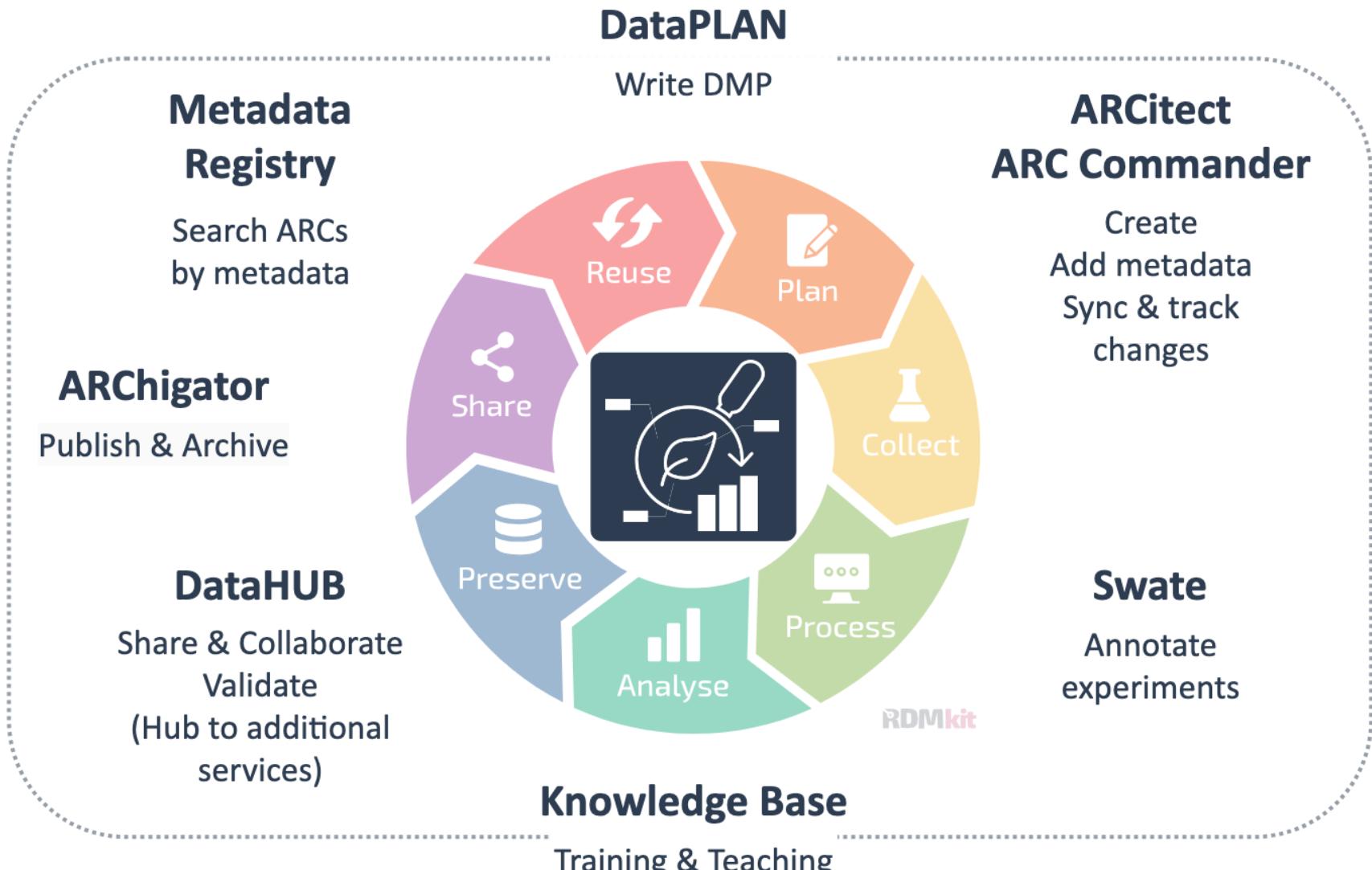
Streamlined data exchange



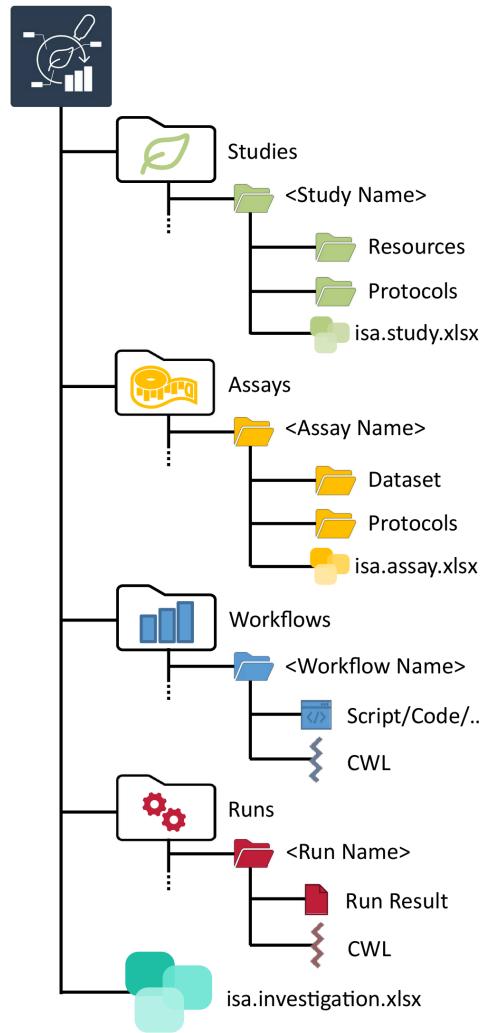
Meet your collaborators in an ARC



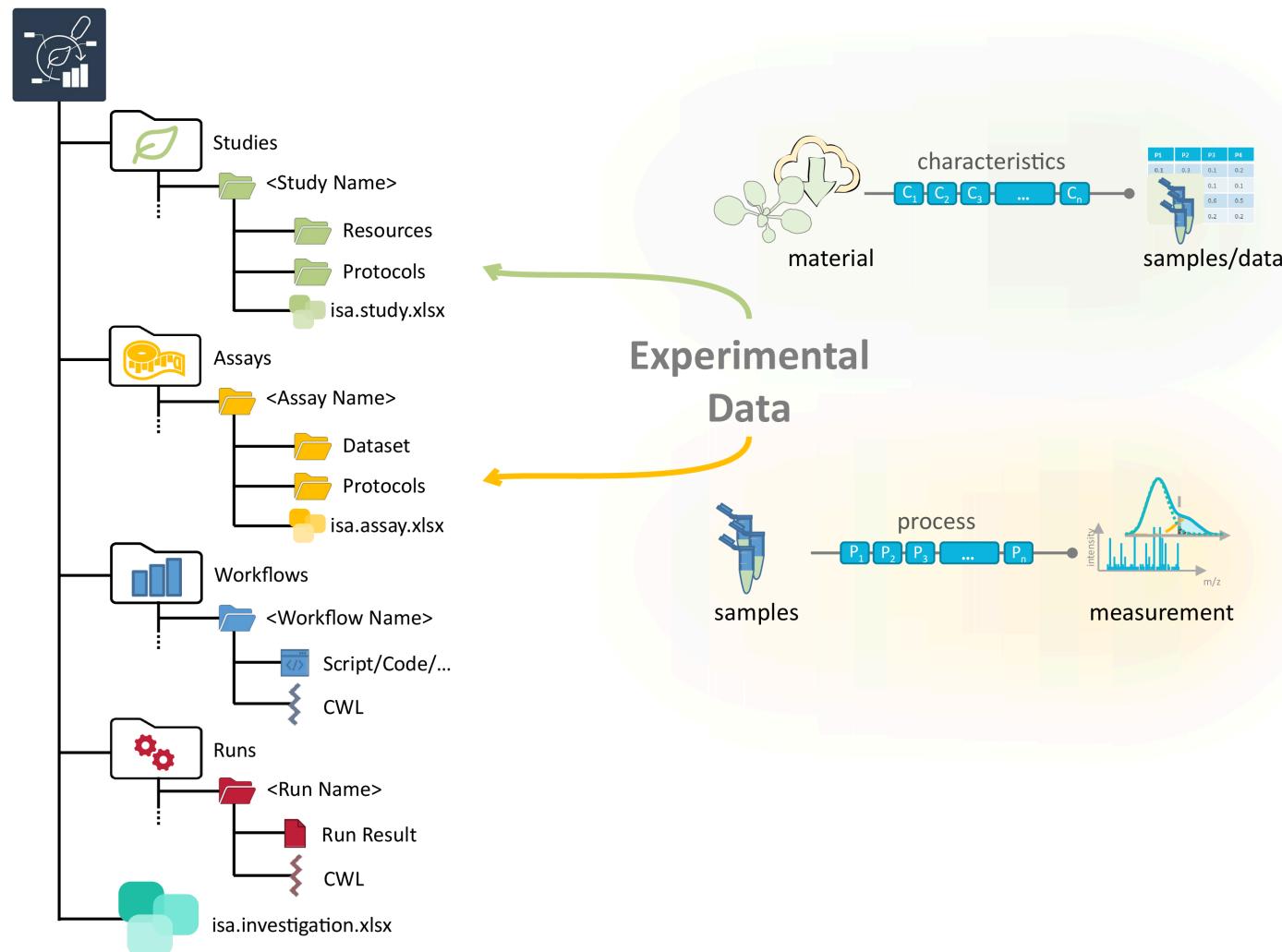
The ARC ecosystem



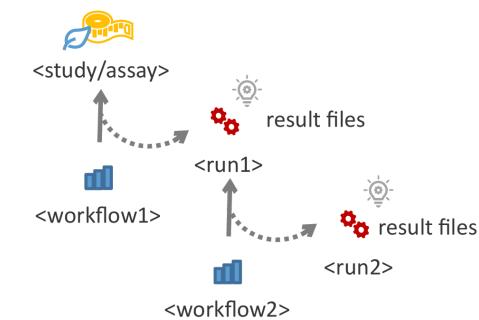
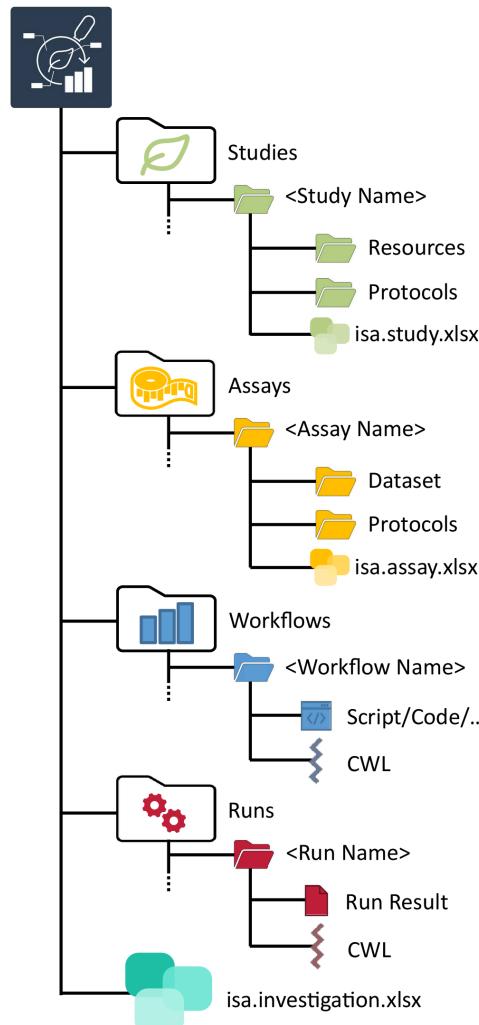
What does an ARC look like?



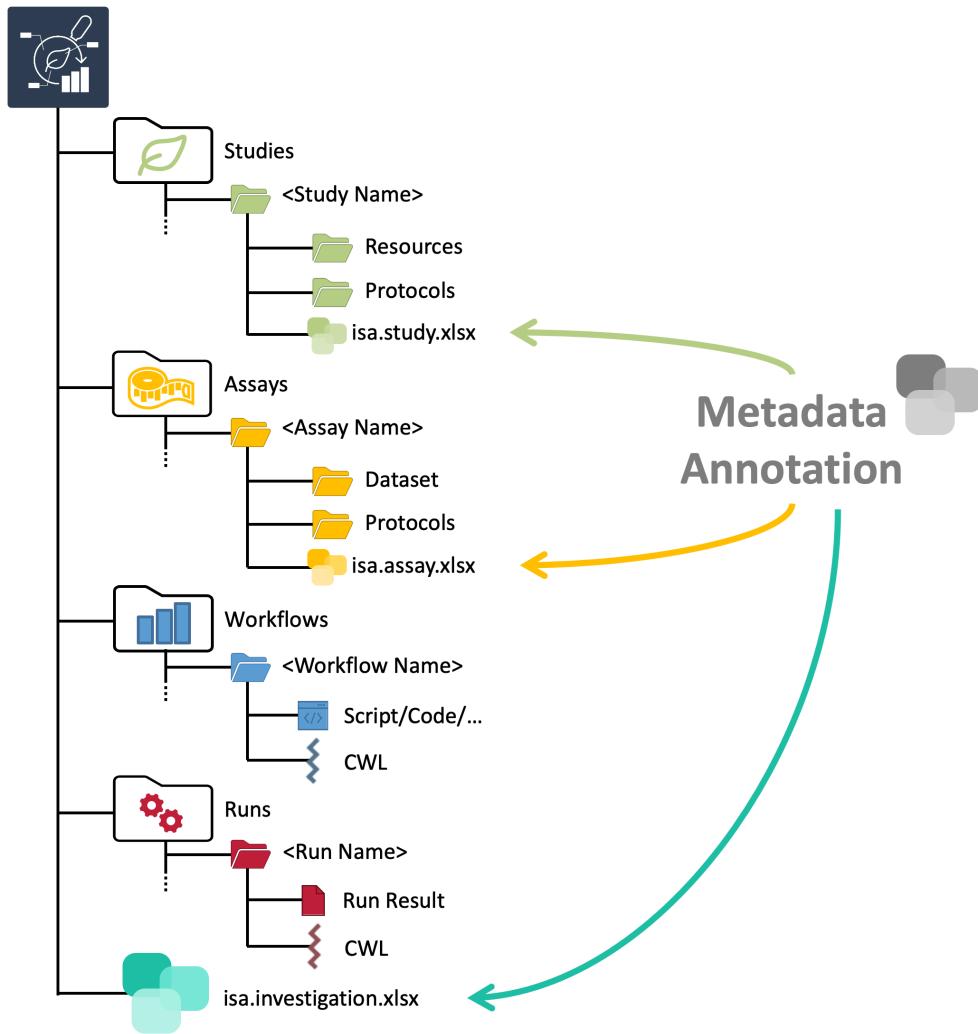
ARCs store experimental data



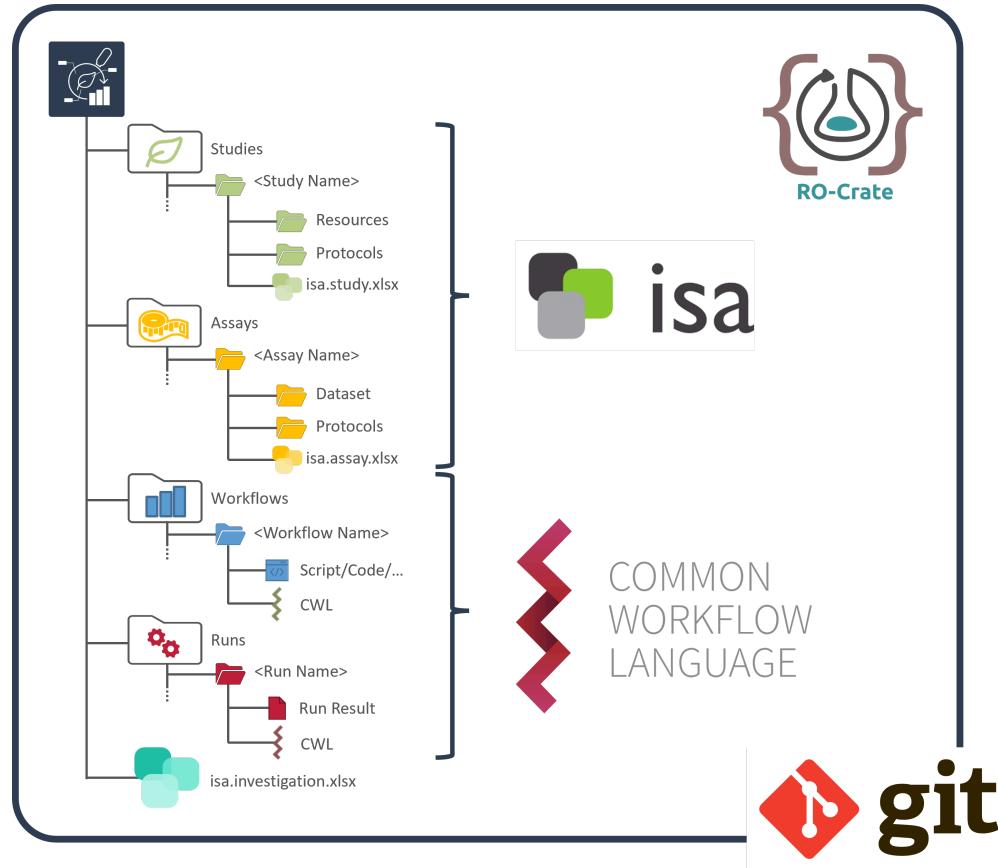
Computations can be run inside ARCs



ARCs come with comprehensive metadata



ARC builds on standards



RO-Crate

- standardized exchange
- <https://www.researchobject.org/ro-crate/>

ISA

- structured, machine-readable metadata
- <https://isa-tools.org/>

CWL

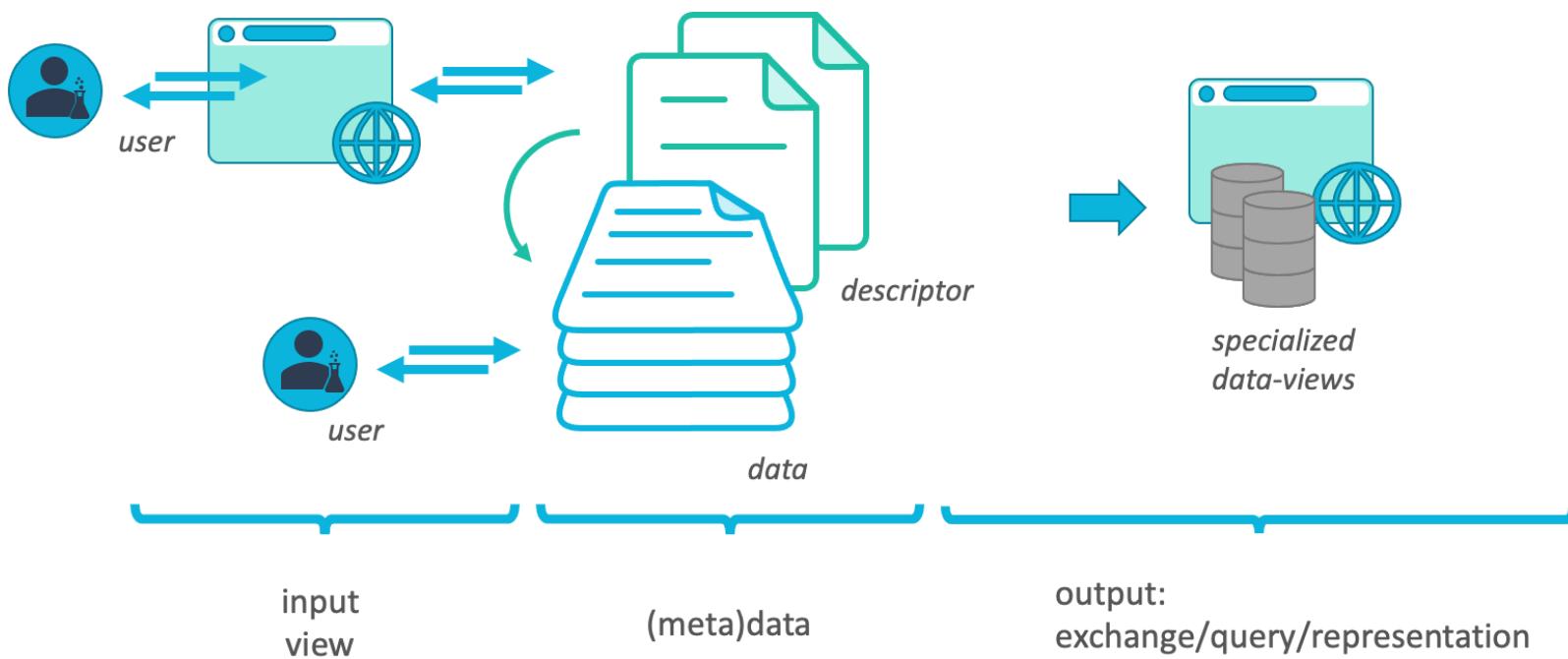
- reproducible, re-usable data analysis
- <https://www.commonwl.org/>

Git

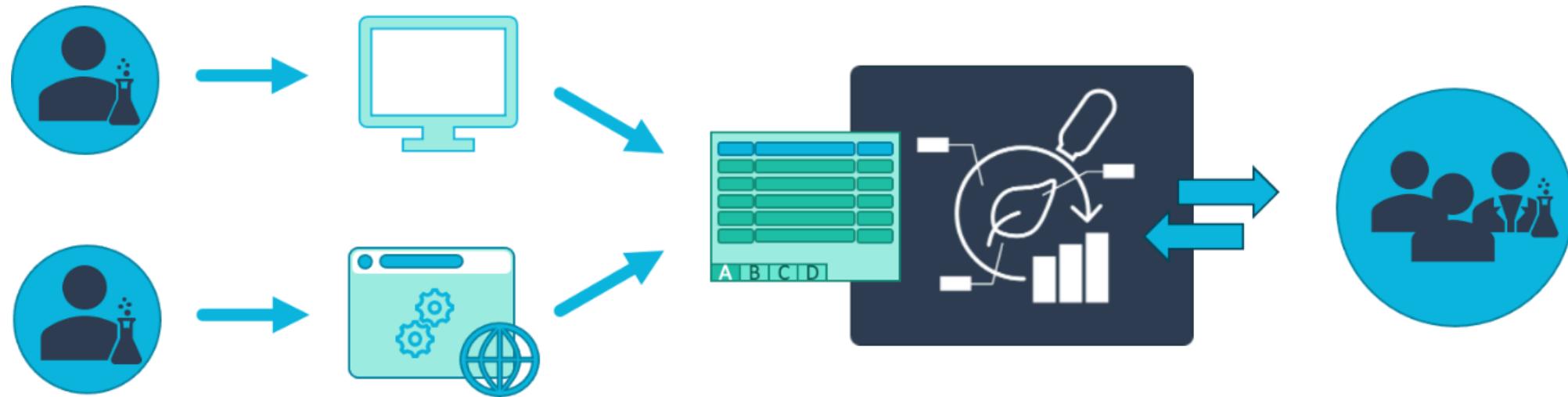
- version control
- <https://git-scm.com>

Everything is a file

The ARC is a **data-centric** approach to RDM



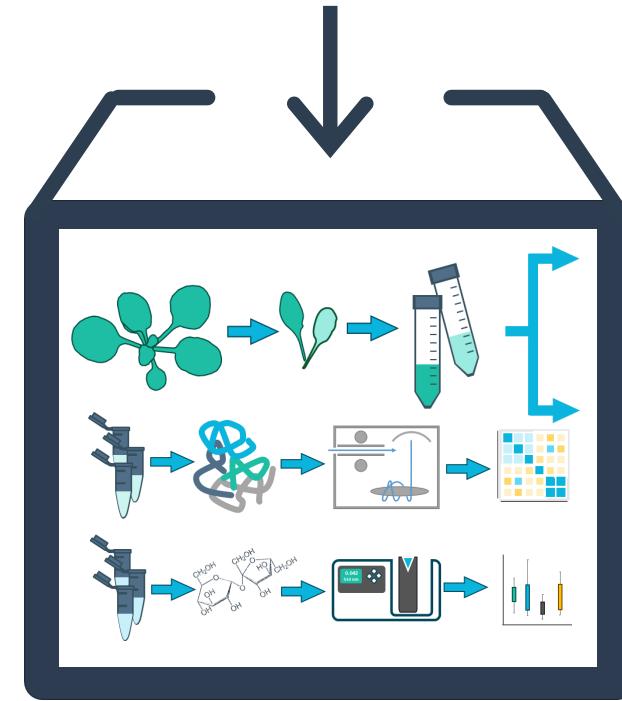
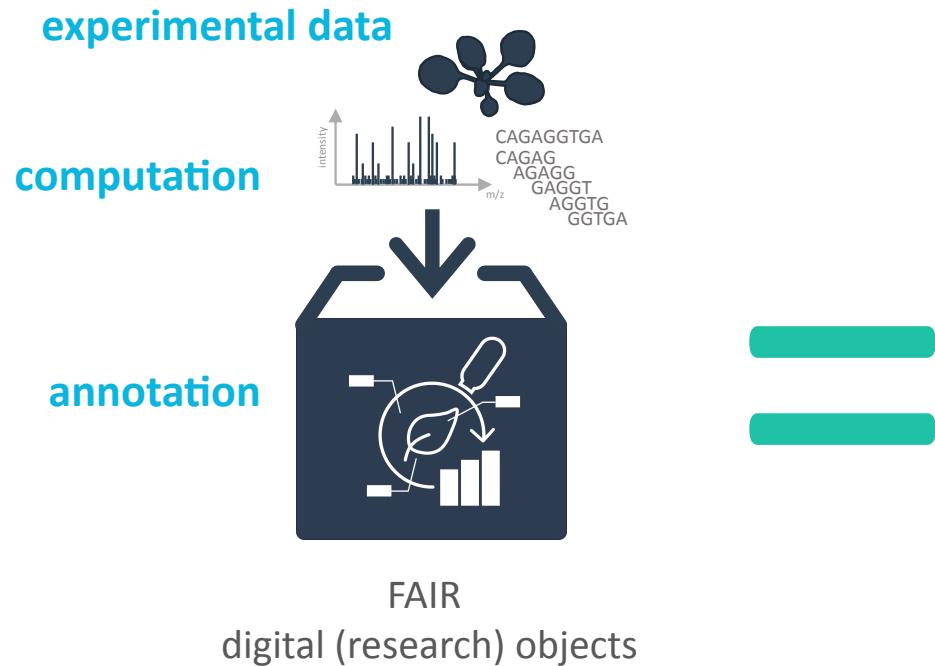
No technical lock-in



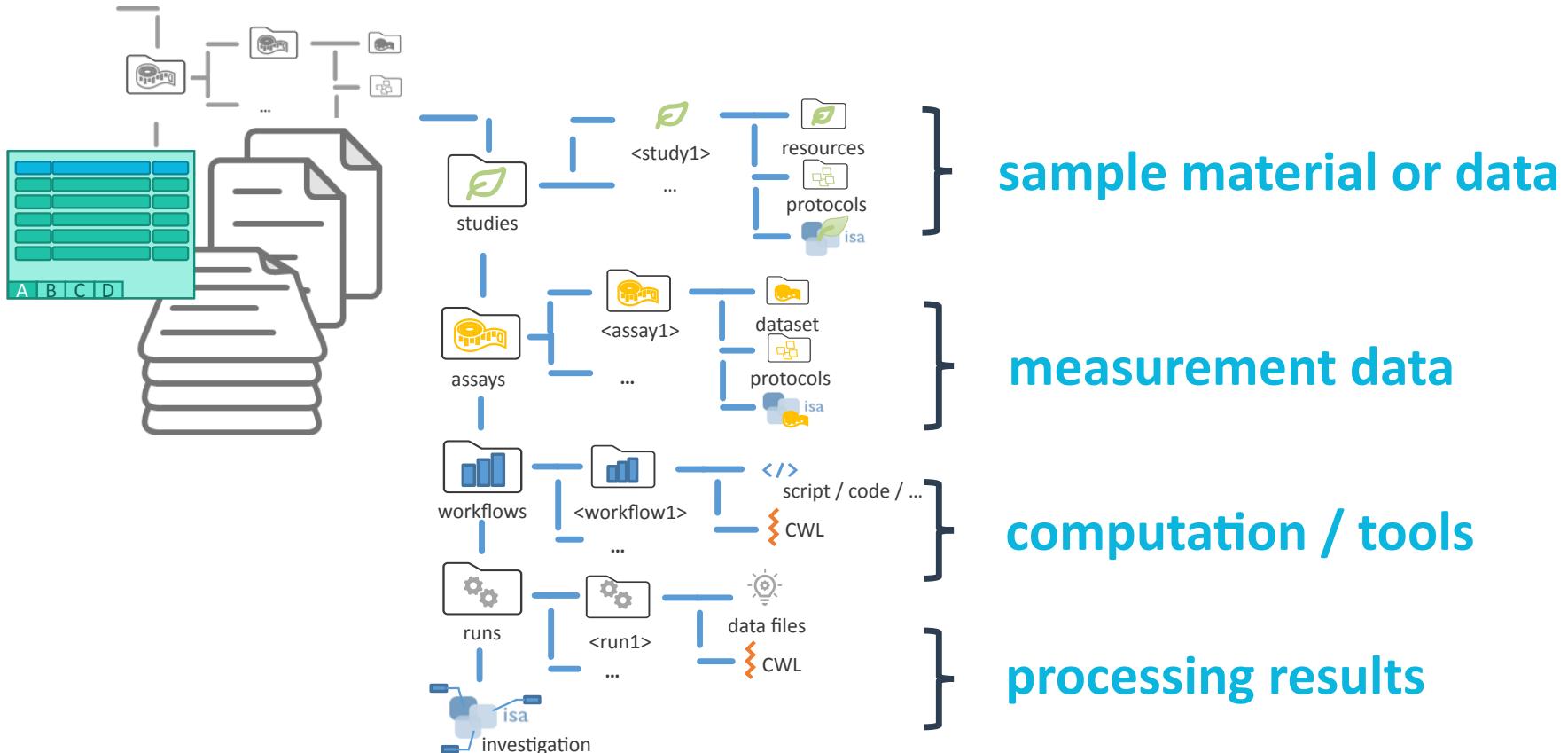
(Meta)data transparency with tool assistance but **no technical lock-in**

Hands-on part 1: Setup and ARCitect

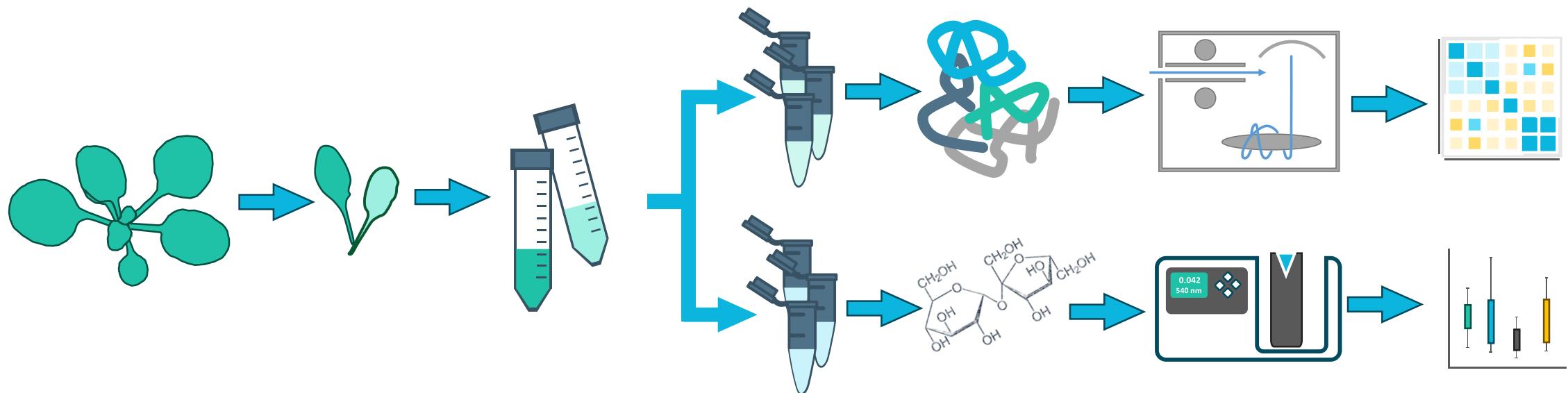
ARC: Annotated research context



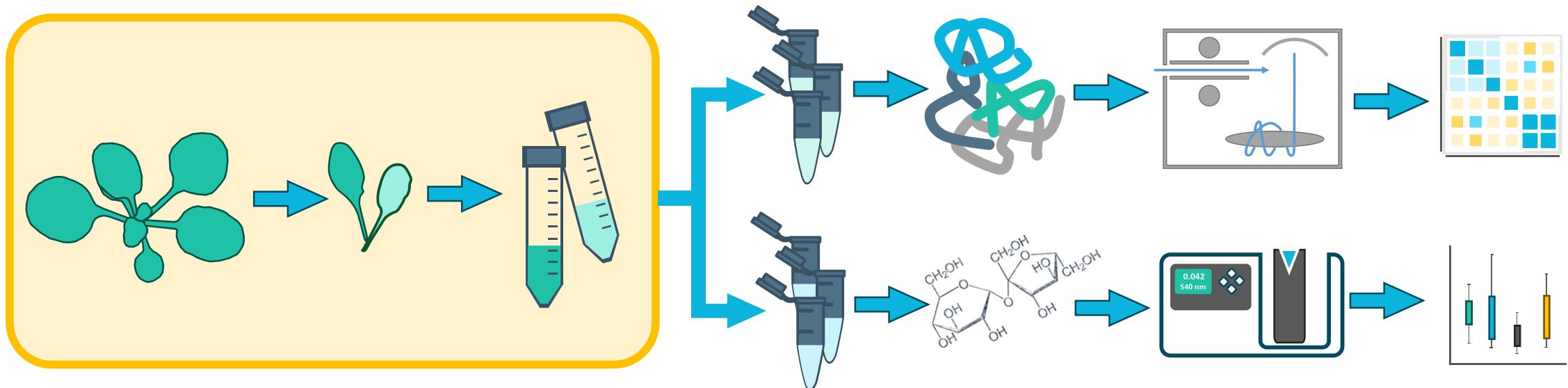
The ARC scaffold structure



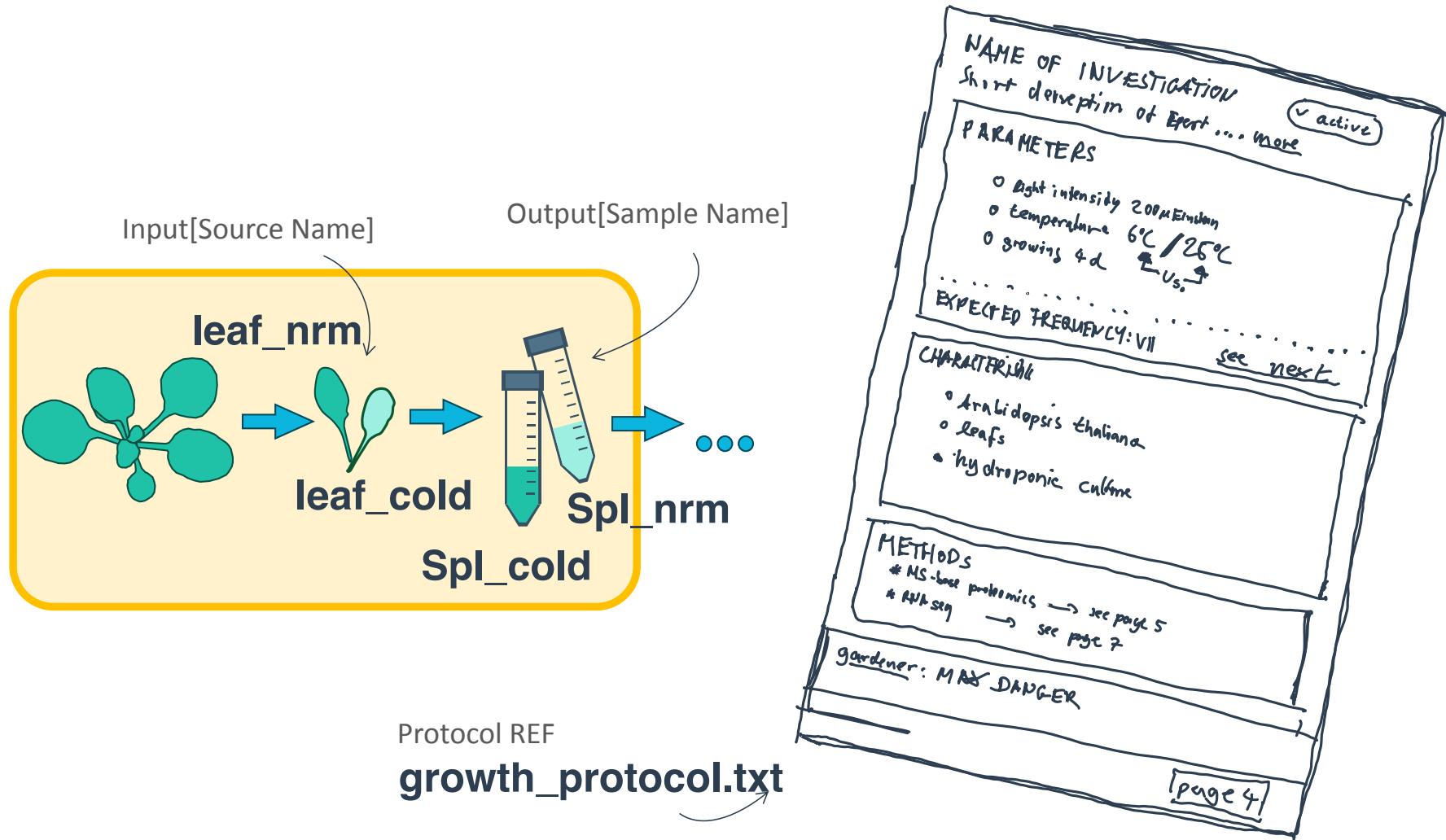
A small prototypic project



Divide and conquer for reproducibility



Identifying the 'study' part



A table-based organization schema

Input[Source Name]		Output[Sample Name]
leaf_nrm		spl_nrm
leaf_cold		spl_cold
A	B	C
D		

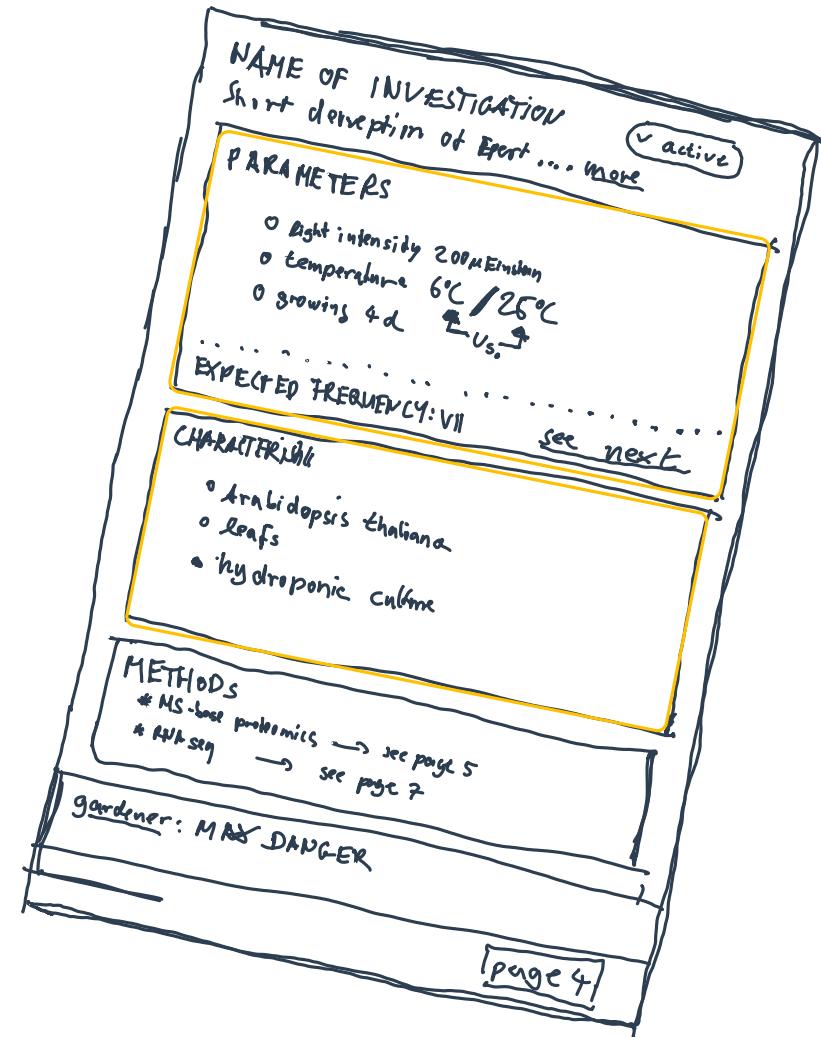
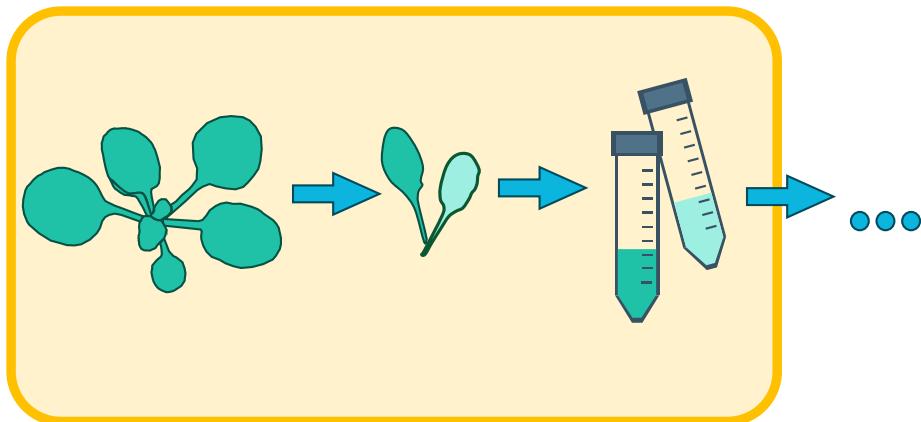
Referencing a protocol

This allows you to reference the free-text, human-readable protocol.

Input[Source Name]	Protocol REF	Output[Sample Name]
leaf_nrm	growth_protocol.txt	spl_nrm
leaf_cold	growth_protocol.txt	spl_cold
A	B	C
D		

- 💡 It is recommended that the protocol is in an open format (.md|.txt|.docx|...)
- 💡 But everything is possible also an URI to an electronic lab notebook

Parameterizing the 'study'



Finding the right metadata vocabulary

Parameters []

- Light intensity 200 µEinstein
- Temperature 6°C / 25°C
- Growing 4d

Characteristics []

- *Arabidopsis thaliana*
- Leaf
- Hydroponic culture
- Columbia

OLS: Finding the right metadata vocabulary

Temperature Dependence [Temperature:Dependence_Annotation](#)

http://purl.uniprot.org/core/Temperature_Dependence_Annotation

Indicates the optimum temperature for enzyme activity and/or the variation of enzyme activity with temperature variation; the thermostability/thermolability of the enzyme is also mentioned when it is known.

Ontology: [UNIPROT RDFS](#)

temperature [AFO:/result#AFR_0001584](#)

http://purl.allotrope.org/ontologies/result#AFR_0001584

A temperature (datum) is a quantity facet that quantifies some temperature. [Allotrope]

Ontology: [AFO](#)

temperature [FBcv:0000466](#)

http://purl.obolibrary.org/obo/FBcv_0000466

Mutation caused by exposure to a temperature that is higher or lower than 25 degrees Celsius.

Ontology: [FBCV](#)

temperature [PATO:0000146](#)

http://purl.obolibrary.org/obo/PATO_0000146

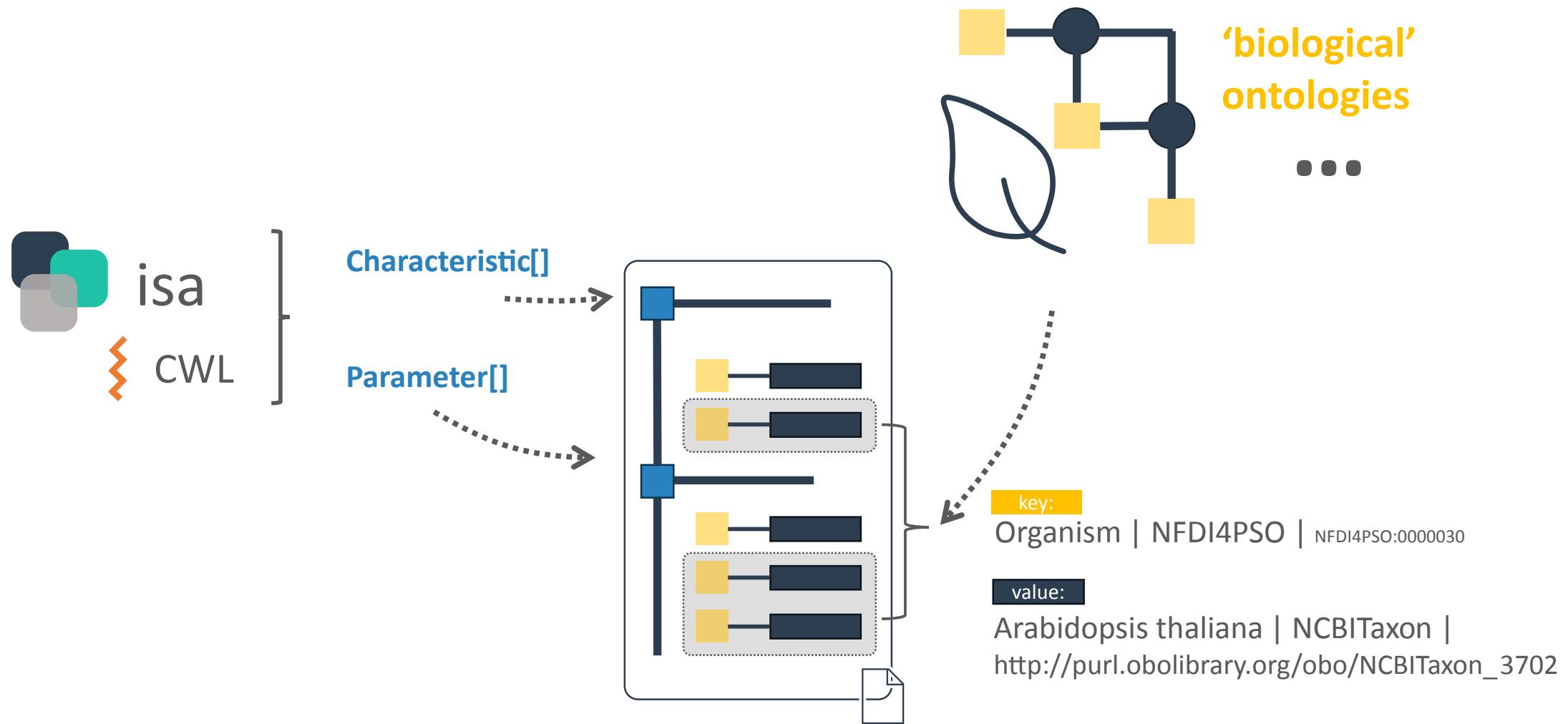
A physical quality of the thermal energy of a system.

Ontology: [PATO](#)

Also appears in: [NGBO](#) [HTN](#) [CAO](#) [ZP](#) [AGRO](#) [OMIABIS](#) [OBIB](#) [MONDO](#) [TXPO](#) [MCO](#) +

Ontology Lookup Service (OLS): <https://www.ebi.ac.uk/ols4/>

Finding the metadata vocabulary and descriptors



Finding the metadata vocabulary and descriptors

Parameters []

-   Light intensity 200 µEinstein
-   Temperature 6°C / 25°C
-   Growing 4d

Characteristics []

-   Arabidopsis thaliana
-   Leaf
-   Hydroponic culture
-   Columbia

Finding the metadata vocabulary and descriptors

Parameters []

- Light intensity
 - 200 µEinstein
- Temperature
 - 6°C / 25°C
- Growth time
 - 4d

Characteristics []

- Organism
 - Arabidopsis thaliana
- Tissue
 - Leaf
- Growth medium
 - Hydroponic culture
- Ecotype
 - Columbia

Hands-on part 1: Setup and ARCitect

Follow the [Start Here guide](#) in the DataPLANT knowledge base.



Until step **Add a study**

Metadata and standards

**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**

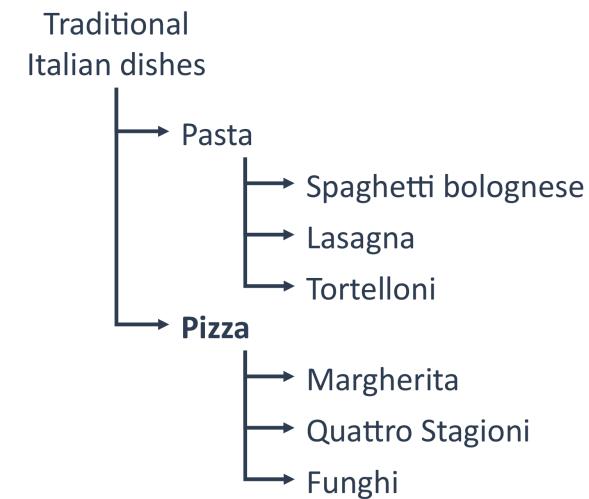
Dictionary

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**

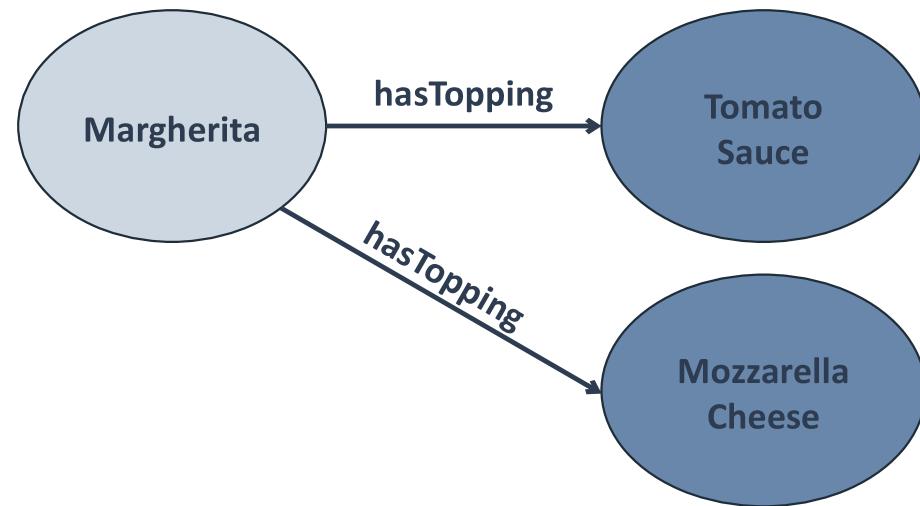
The semantic triple



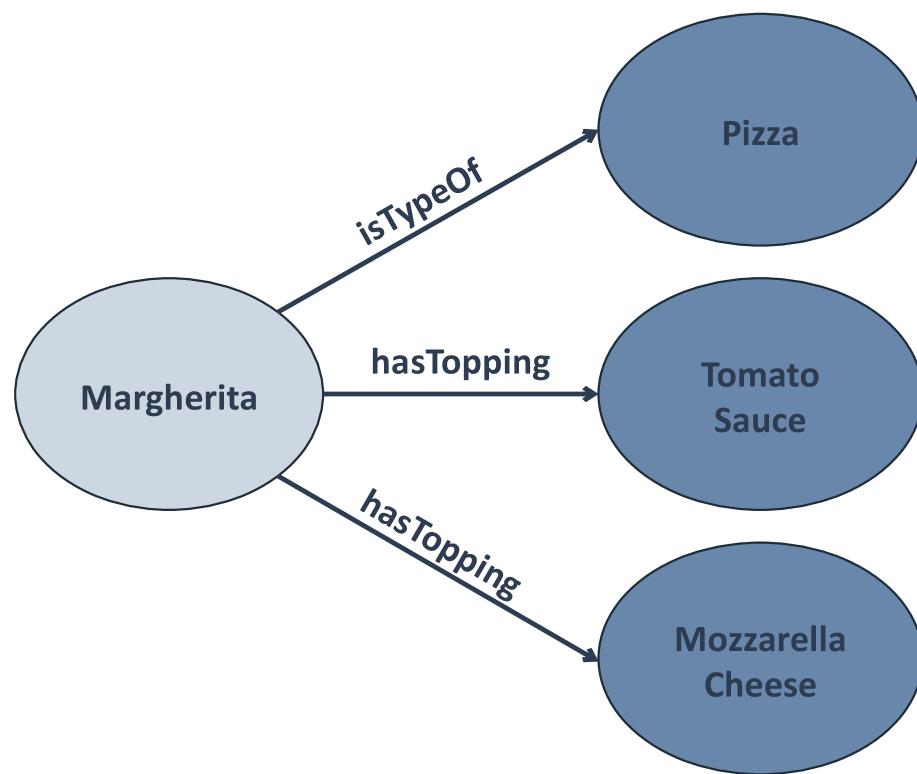
Modeling a pizza menu



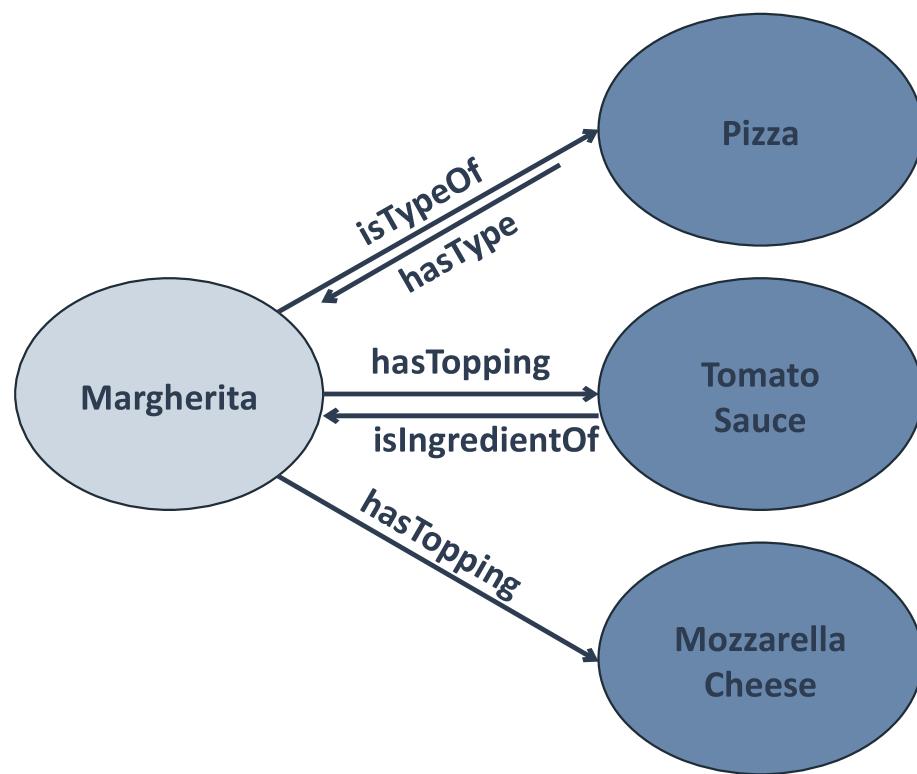
Modeling a pizza menu



Modeling a pizza menu

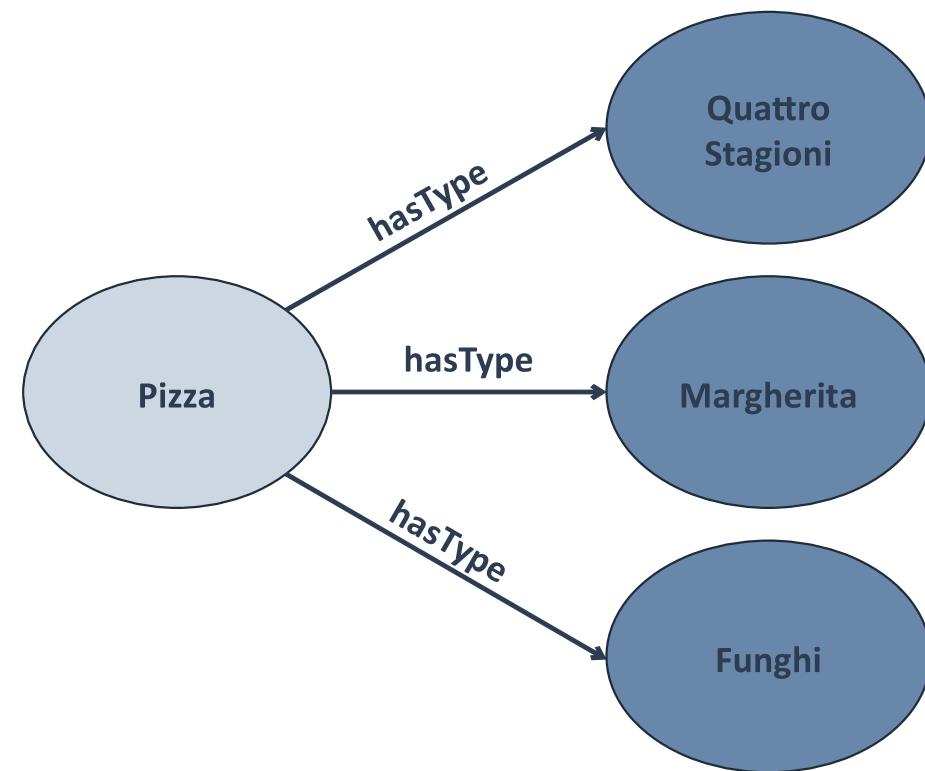


Predicates have two directions

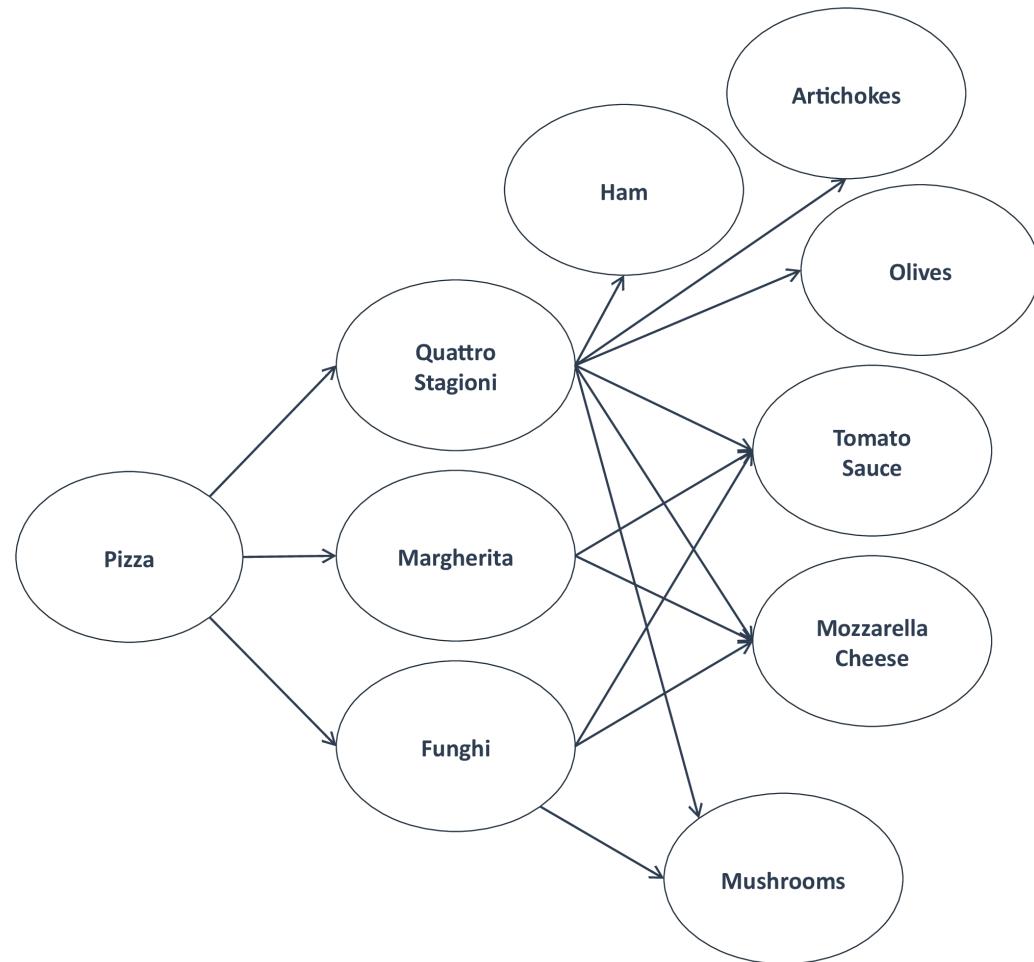


Looking at the menu from a different perspective

An object of one triplet can be the subject to another



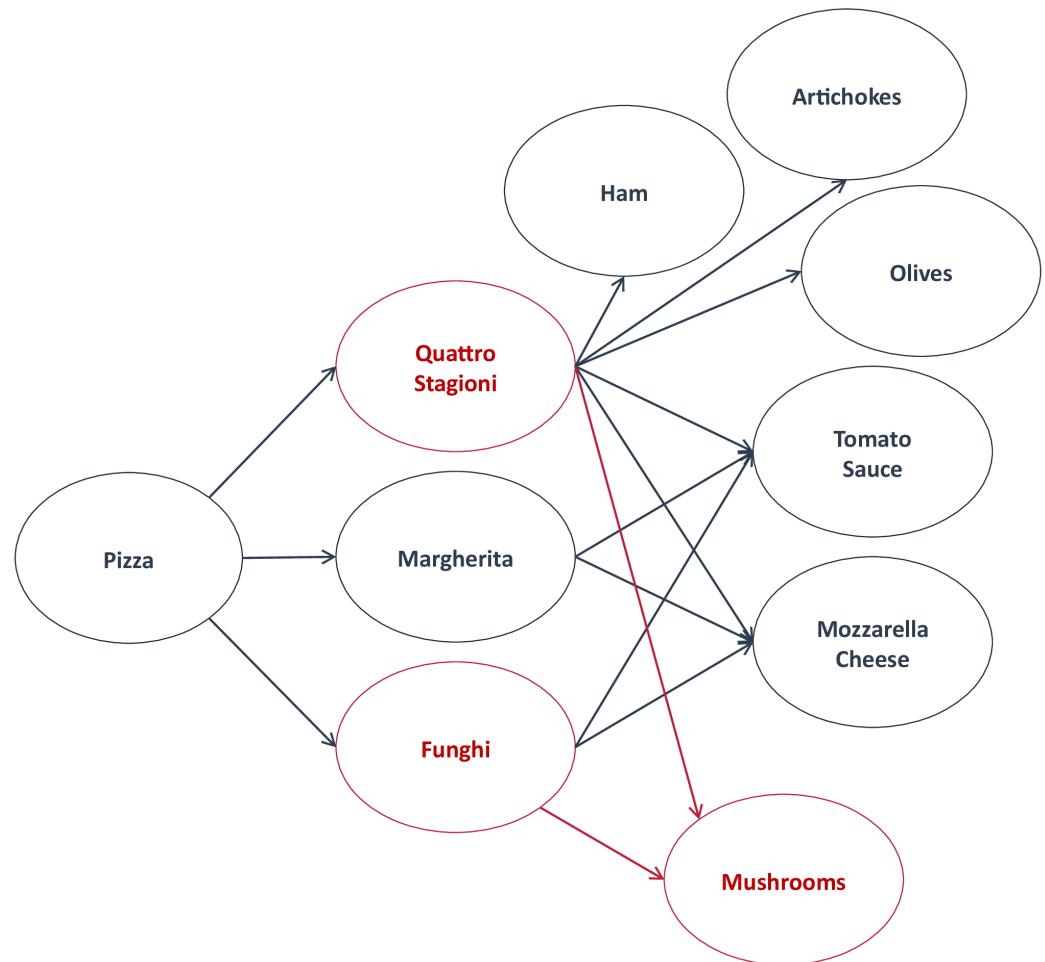
(Towards) a knowledge graph



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>

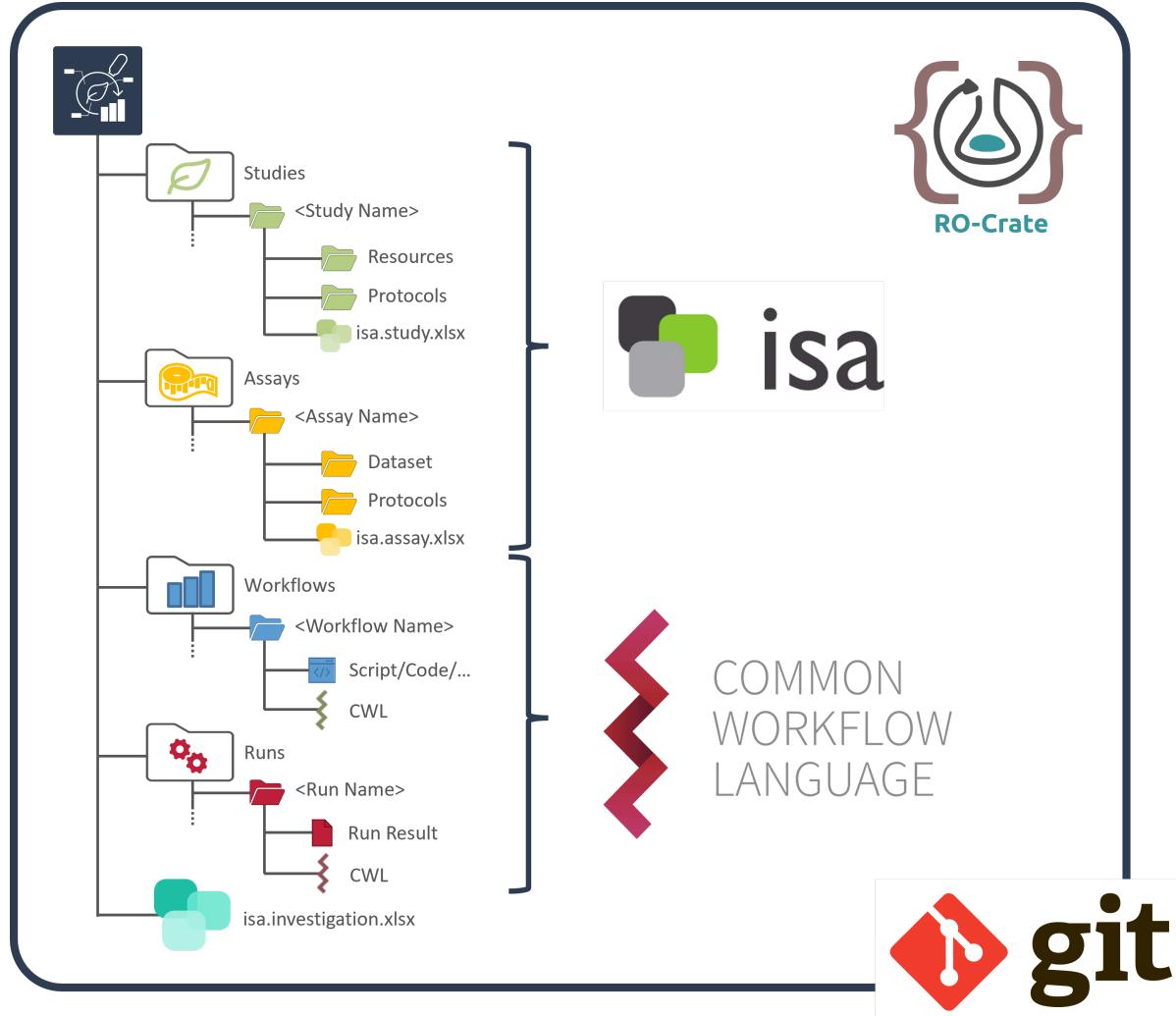
PECO ontology

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

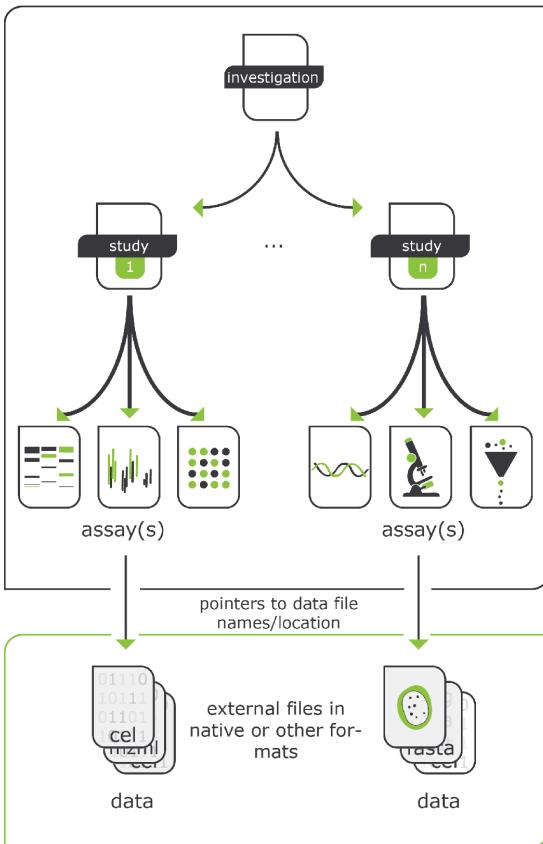
Explore more examples

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

ARC builds on standards



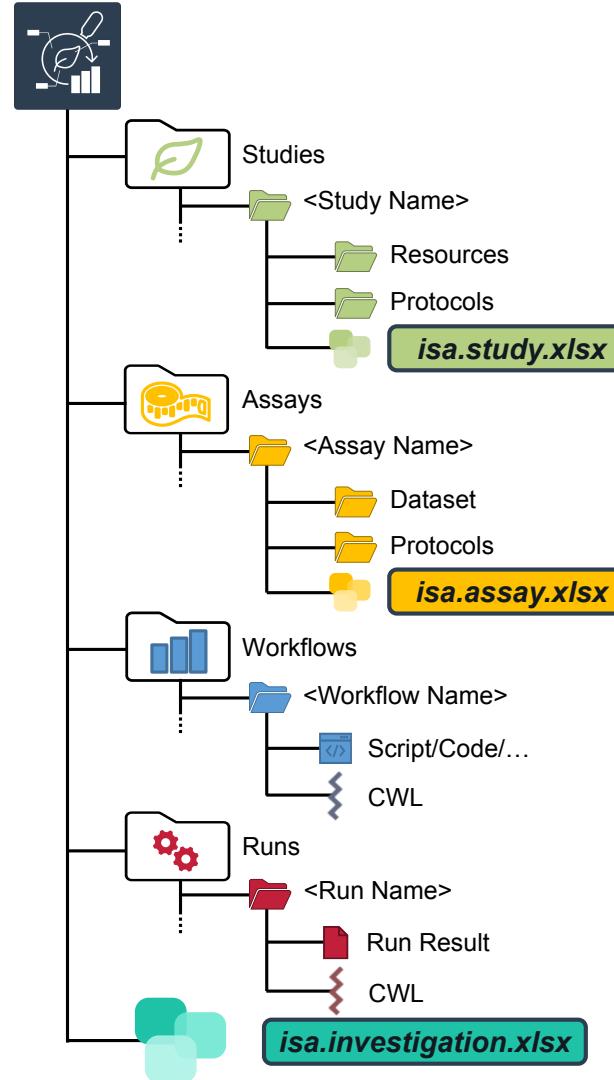
ARC builds on ISA



Investigation
Overall goals
Scientific context

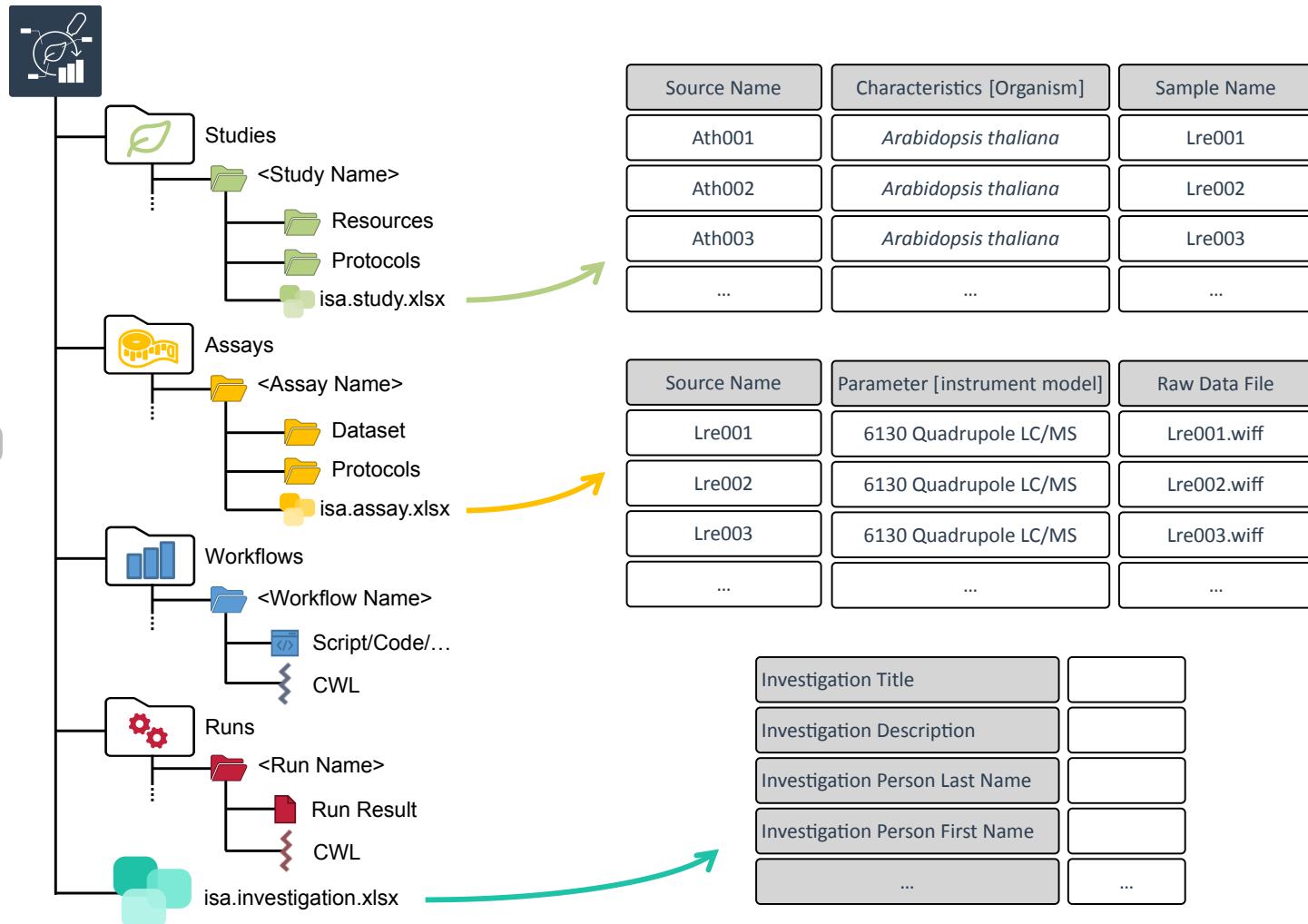
Study
Experimental design

Assay
Leading to (raw) data

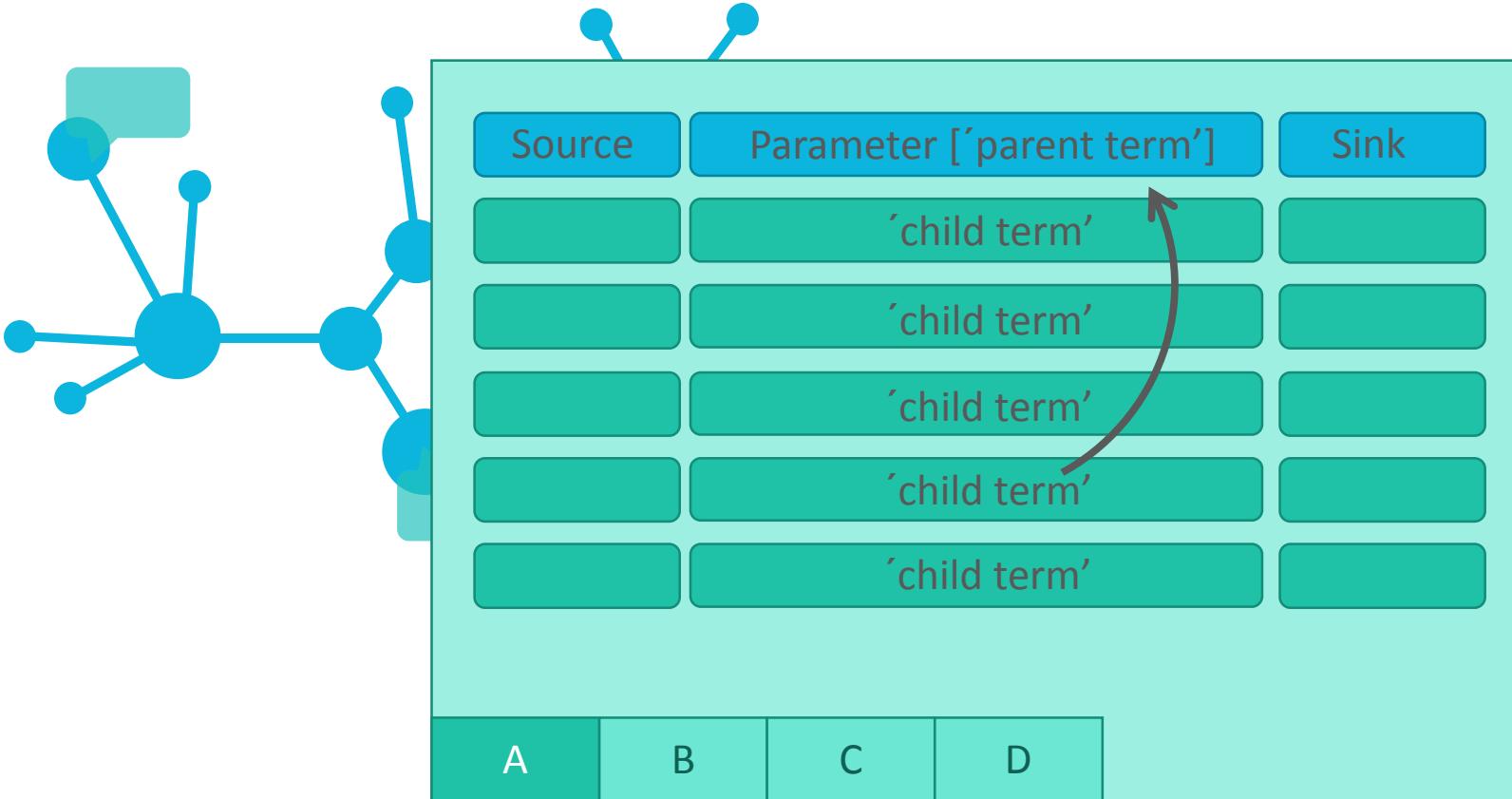


ARC builds on ISA

Metadata Annotations



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

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 table of data. Overlaid on the table are several callout boxes with arrows pointing to specific columns:

- Characteristic**: Points to the 'Characteristic [sample label]' column.
- Protocol Type/Protocol REF**: Points to the 'Protocol Type' and 'Protocol Ref' columns.
- Factor**: Points to the 'Factor [temperature]' column.
- Component**: Points to the 'Component [instrument model]' column.
- Sample Name/Raw Data File/Derived Data File**: Points to the 'Sample Name' column.
- New Parameter**: Points to a column header labeled 'Parameter' in the adjacent 'Building Blocks' pane.

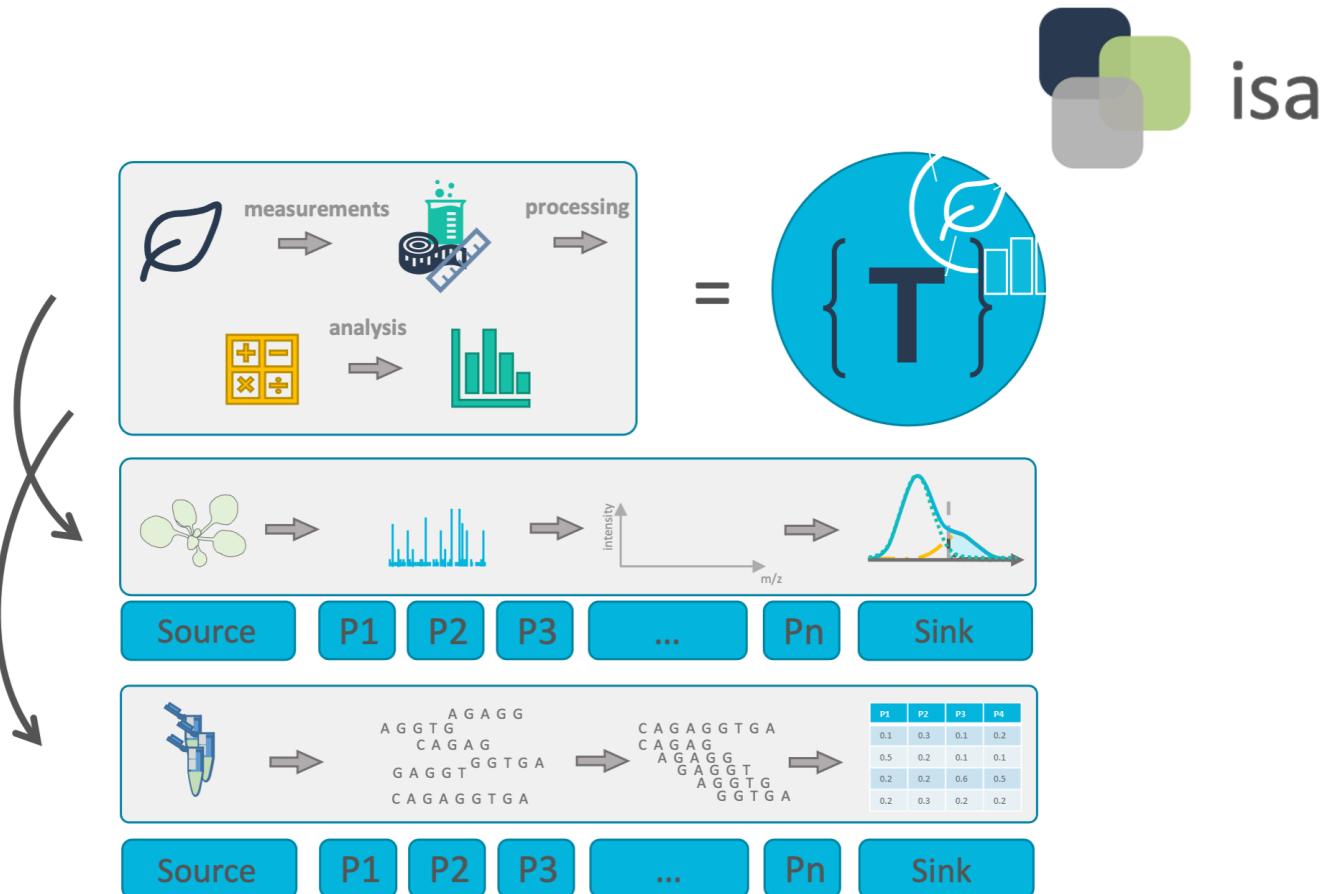
The 'Building Blocks' pane on the right lists various annotations:

- Instrument mod: MS1000031
- Instrument Model: NCIT:C177610
- Instrument: MS1000463
- Instrument: EFO:0000548
- Agilent instrument model: MS1000490

A note at the bottom of the pane says: "Can't find the term you are looking for? Try 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 table. You can find more information on our website."

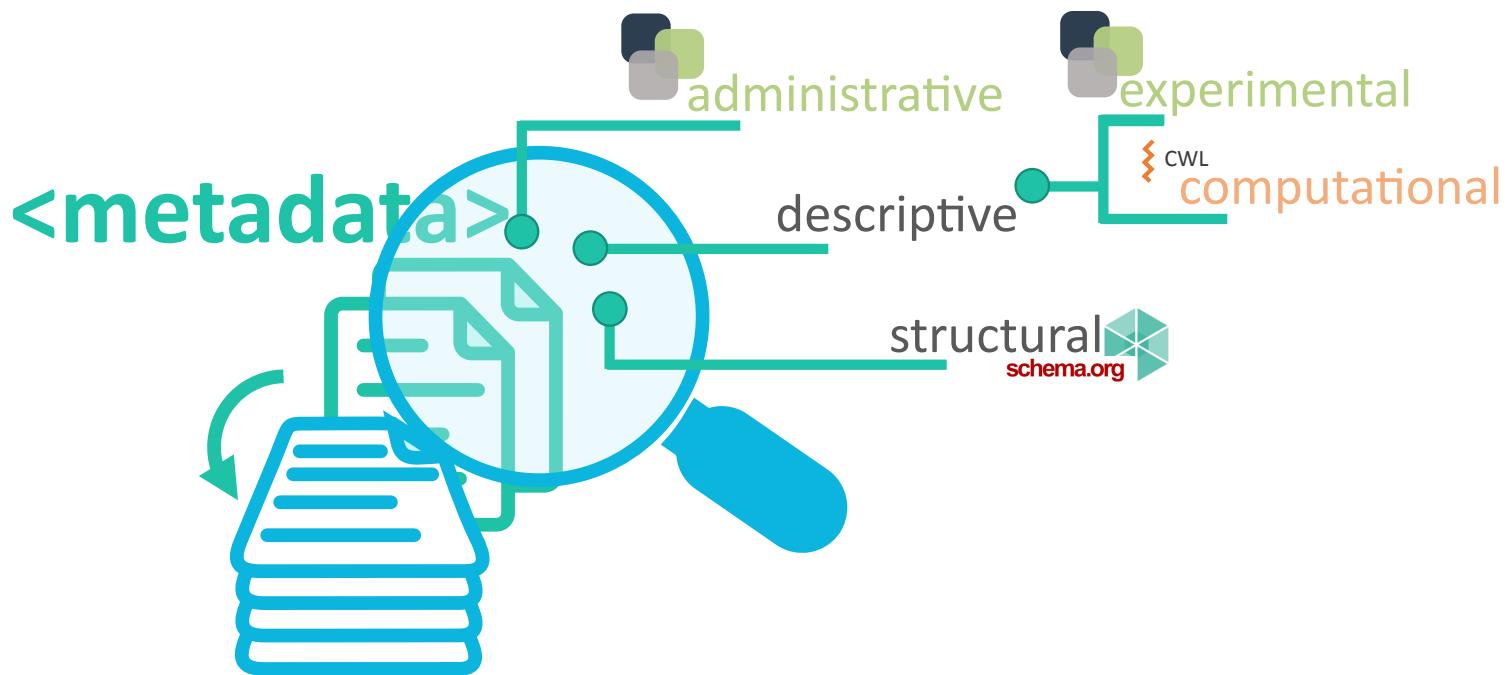
Let's take a detour on [Annotation Principles | slides](#)

Realization of lab-specific metadata templates



Facilities can define their most common workflows as templates

Build on existing well-established standards



Integration of established (meta)data standards ensures compatibility by design

ISA abstract model in a nutshell



isa

Investigation
administrative (meta)data

- Summary
 - Titel
 - Description
- Person
- Organisation
- Publication reference

Study
descriptive (meta)data
information on the subject

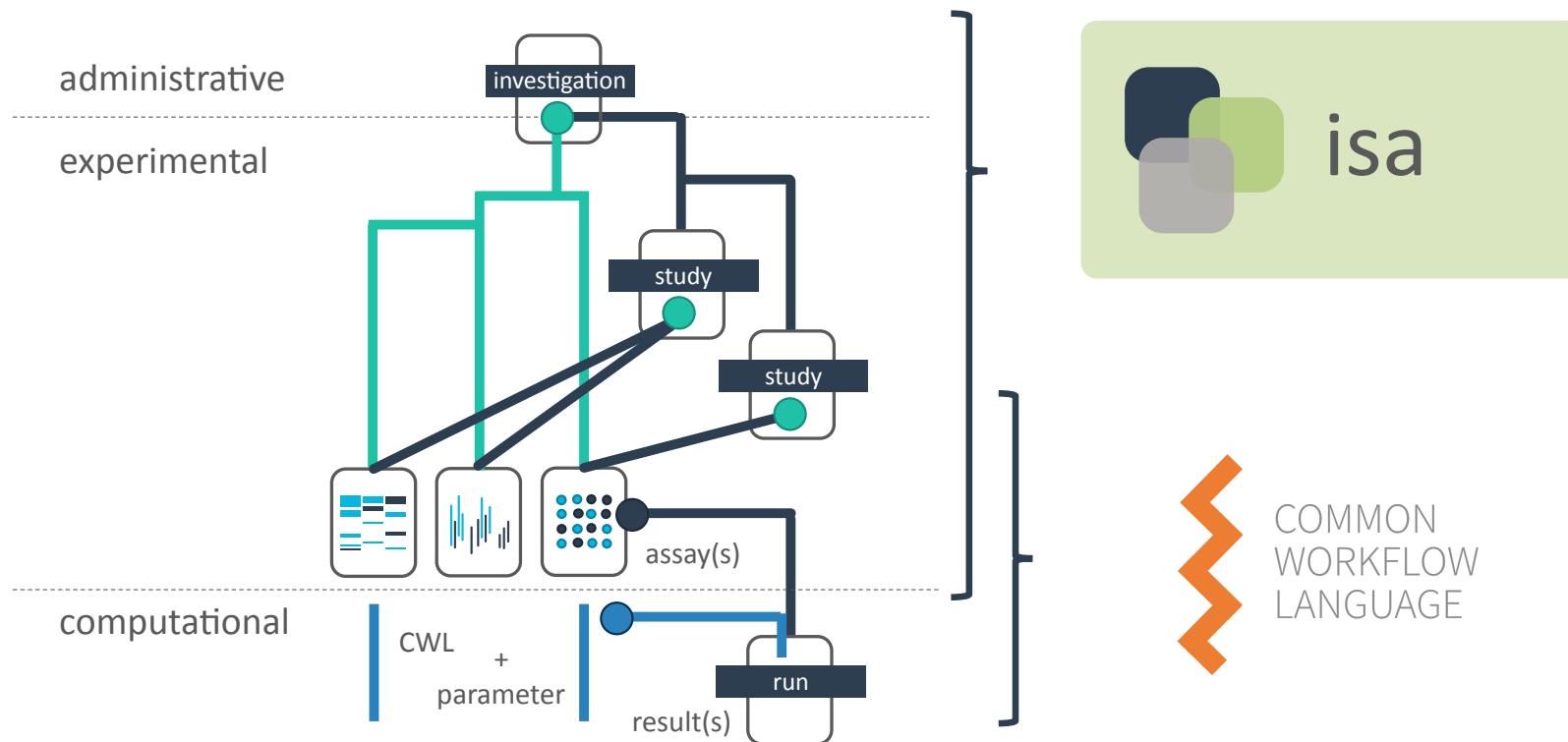
- Characteristics
- Parameters
- Components
- Factors

Assay

descriptive (meta)data
information on the measurement

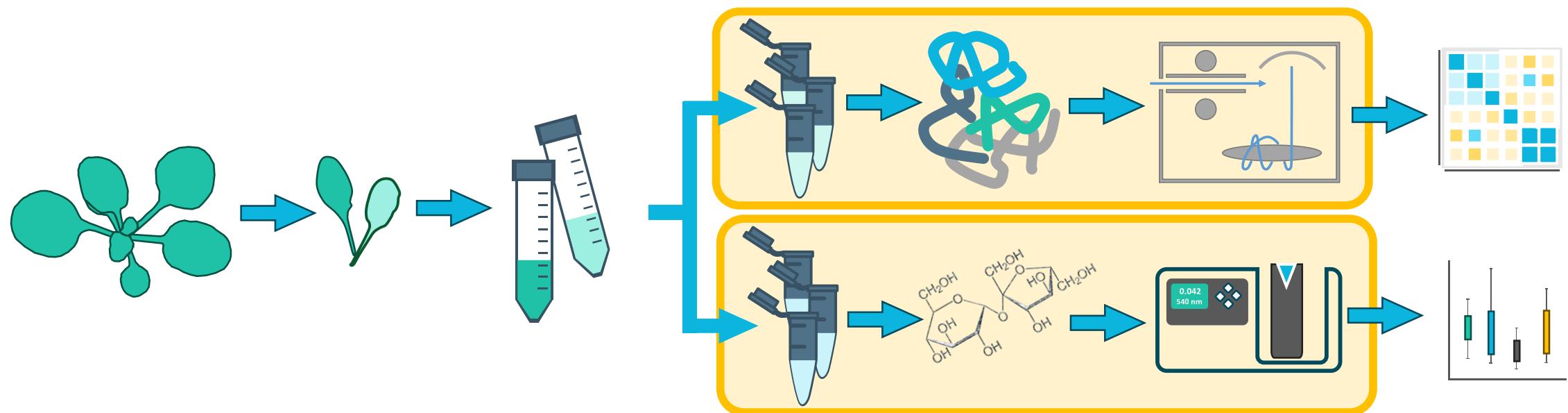
- Characteristics
- Parameters
- Components
- Factors

ISA and CWL – Connected by similarity

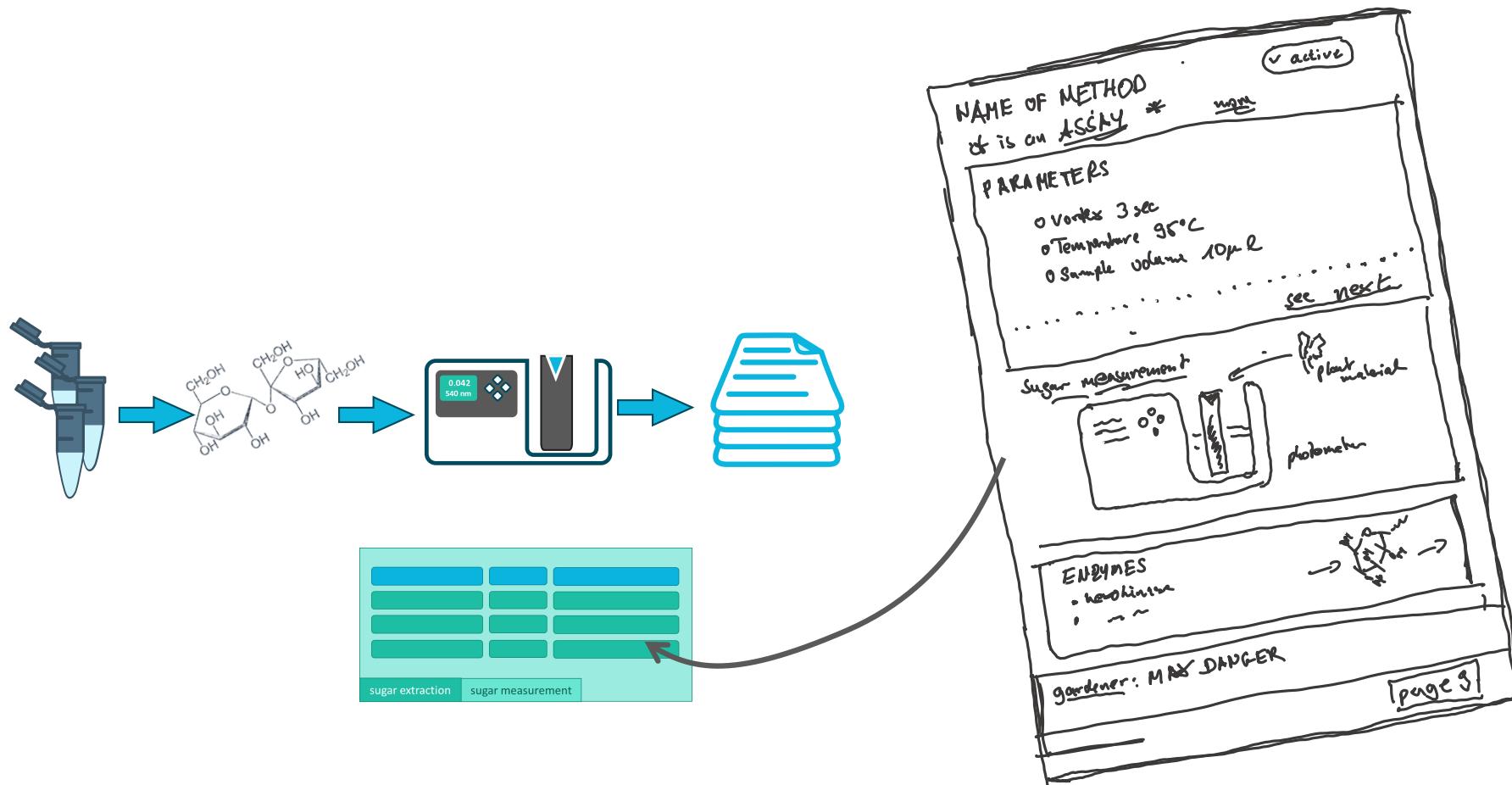


Hands-on part 2: ARCitect (and Swate)

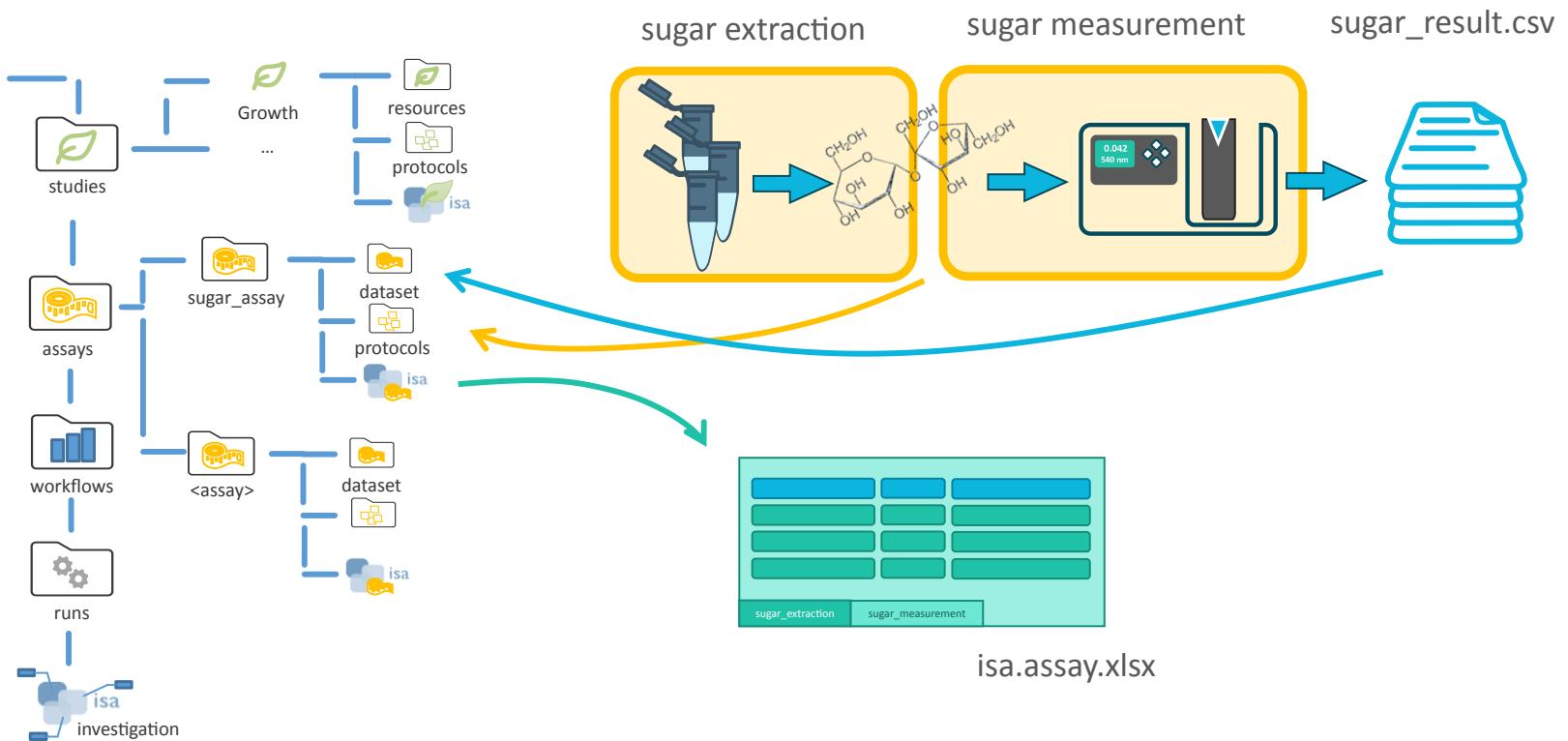
Identifying assays



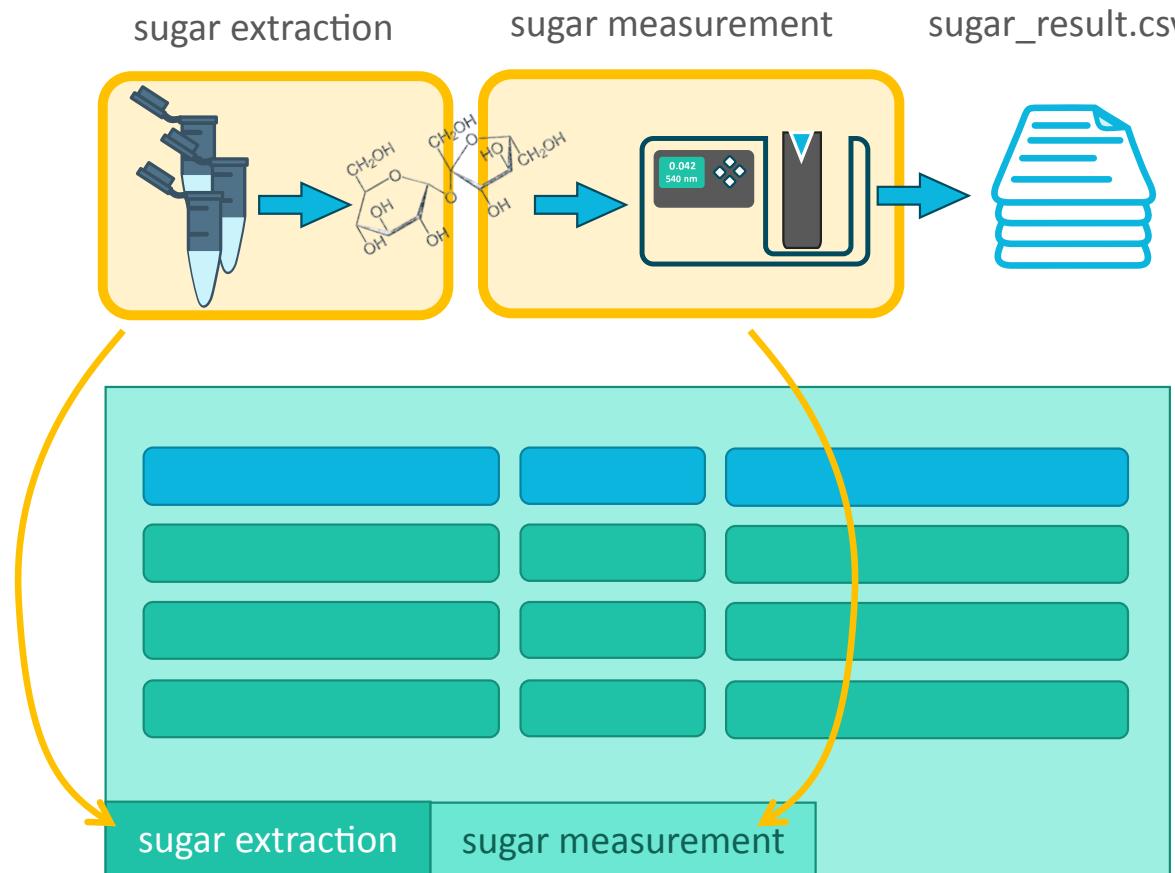
Assay for sugar measurement



Separating different assay elements



Isolating the lab processes in an assay



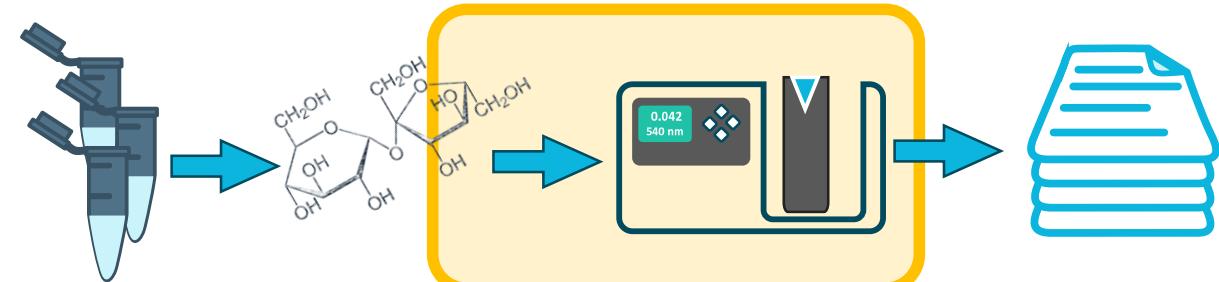
Parameteterization: sugar extraction

- Vortex Mixer
 - 3 seconds
- Temperature
 - 95 degree celsius



Parameteterization: sugar measurement

- █ technical replicate
 - █ 1,2,3,...
- █ sample volume
 - █ 10 microliter
- █ buffer volume
 - █ 190 microliter
- █ cycle count
 - █ 5



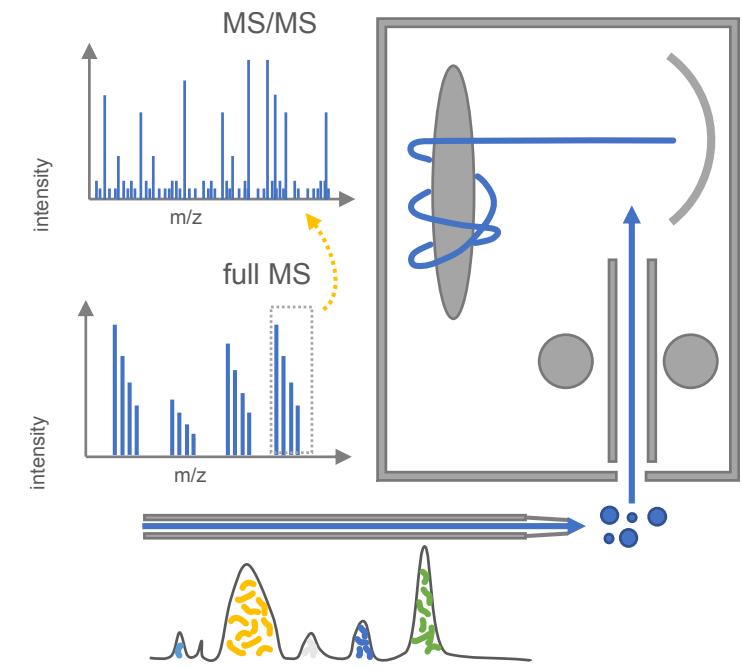
Save time using standard methods and SOPs

Parameter []

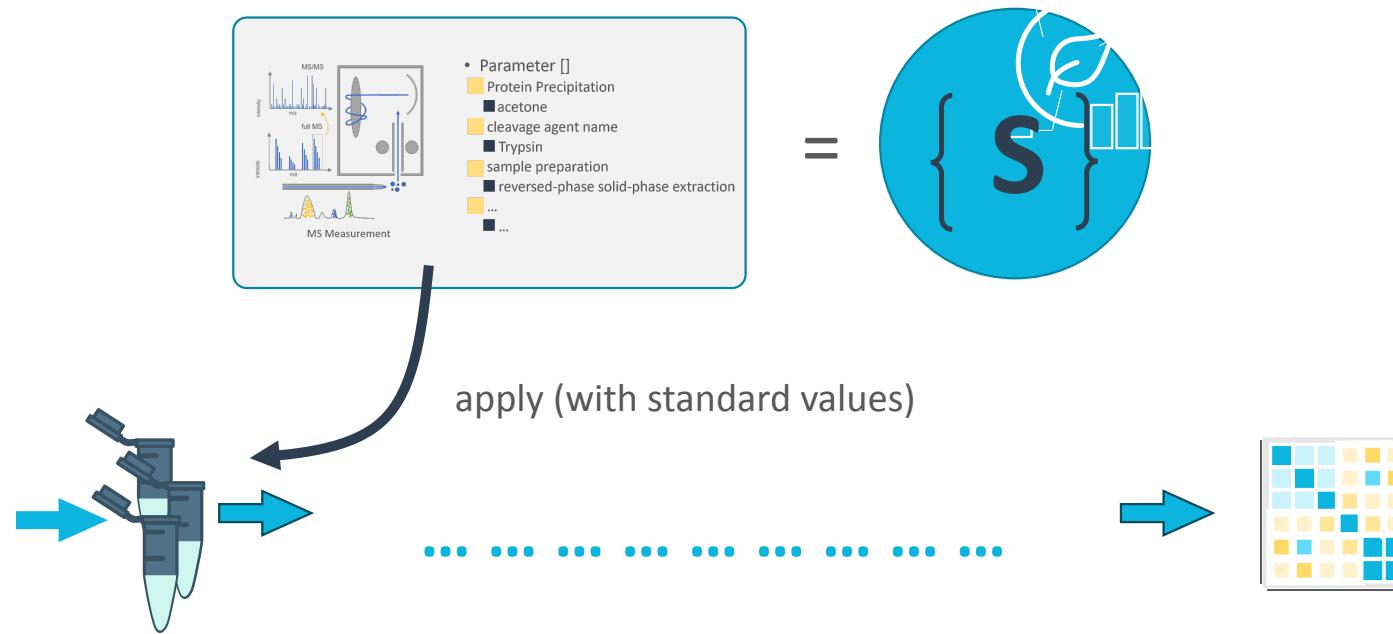
- █ Protein Precipitation
 - █ acetone
- █ cleavage agent name
 - █ Trypsin
- █ sample preparation
 - █ reversed-phase solid-phase extraction
- ...

Component []

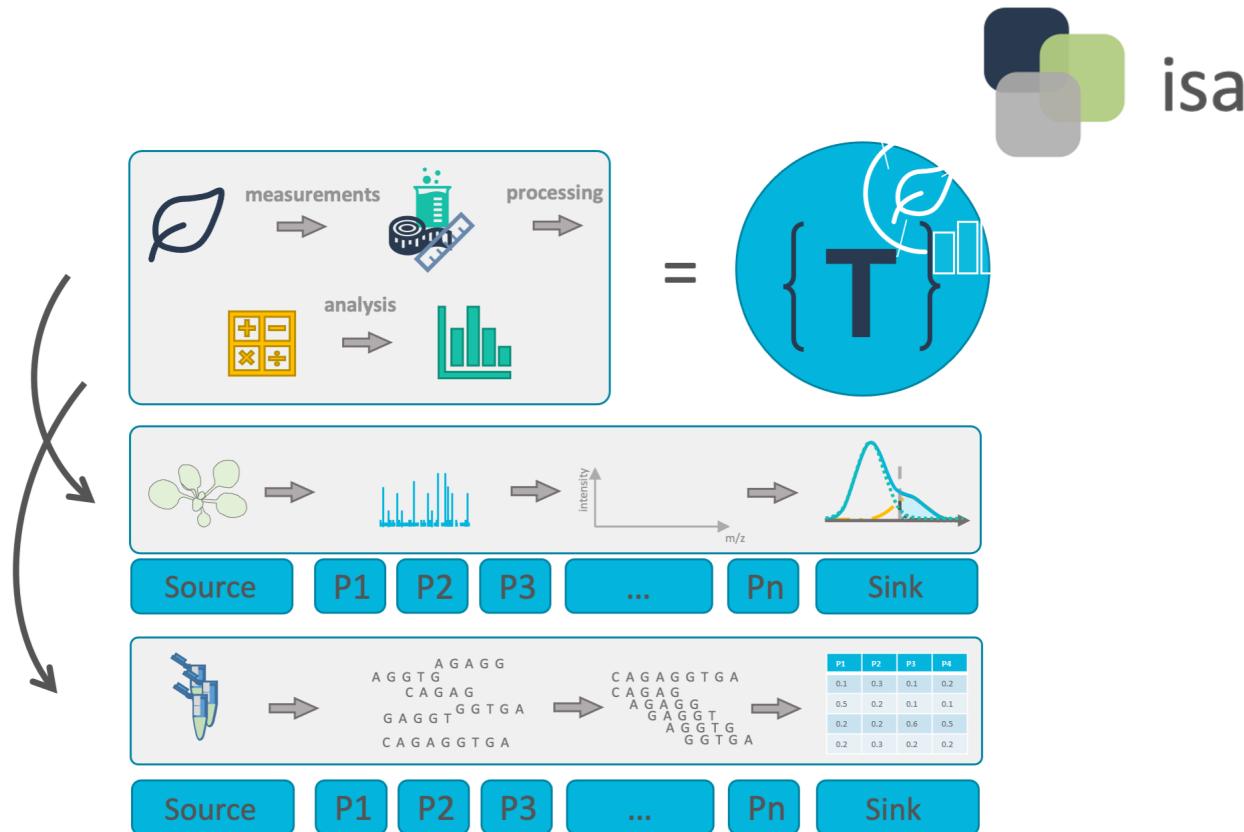
- █ chromatography instrument model
 - █ nanoElute2
- █ chromatography column model
 - █ PepSep C18 1.9 μ , 25cm x 75 μ m
- ...
...



Applying standard procedures to sample record



Realization of lab-specific metadata with templates



Facilities can define their most common workflows as templates

Hands-on part 2: ARCitect (and Swate)

Continue the [Start Here](#) guide in the DataPLANT knowledge base.



Until step **Add assay data**

Code Notebooks and IDEs

- Interactive (good start for non-coders)
- Document and comment code
- Often offer integrated version control (e.g. git plugin)
- Executable code + "result preview"

Examples

- Visual Studio Code: <https://code.visualstudio.com/>
- RMarkdown: <http://rmarkdown.rstudio.com>
- Jupyter Notebooks: <https://jupyter.org/>

Workflow languages

- Describe analysis workflows and tools
- Make them portable and scalable
- Across a variety of environments (software and hardware)

Examples

<https://www.commonwl.org>

<https://www.nextflow.io>

<https://nf-co.re/>

<https://snakemake.github.io>

Common Workflow Language

CWL workflow document (*.cwl)

```
1 --> 2 --> 3
#!/usr/bin/env cwl-runner

cwlVersion: v1.0
class: Workflow
inputs:
  tarball: File
  name_of_file_to_extract: string

outputs:
  compiled_class:
    type: File
    outputSource: compile/classfile

steps:
  untar:
    run: tar-param.cwl
    in:
      tarfile: tarball
      extractfile: name_of_file_to_extract
    out: [extracted_file]

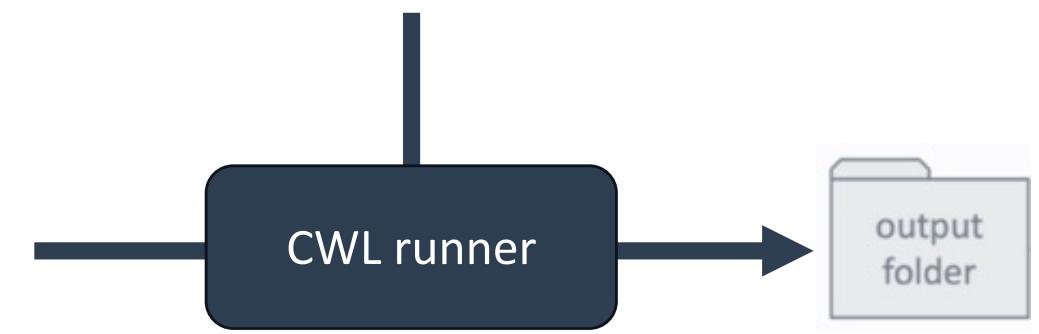
  compile:
    run: arguments.cwl
    in:
      src: untar/extracted_file
    out: [classfile]
```

CWL
tool descriptors (*.cwl)

CWL job parameter (*.yaml)

```
job
yaml

file: fastq
param: 5
workflow: wf.cwl
output_folder:
  /temp
```



Common Workflow Language

```
cwlVersion: v1.2
class: CommandLineTool
requirements:
  - class: InitialWorkDirRequirement
    listing:
      - entryname: heatmap.py
        entry:
          $include: heatmap.py
baseCommand: [python, heatmap.py]
inputs:
  MeasurementTableCSV:
    type: File
    inputBinding:
      position: 1
  FigureFileName:
    type: string
    inputBinding:
      position: 2

outputs:
  output:
    type: File
    outputBinding:
      glob: "*.svg"
```

Galaxy

Platform that makes using code-based tools easy to use in a graphical user interface

Resources

- <https://usegalaxy.eu>
- <https://galaxyproject.eu/>

Galaxy integration: Extra value for plant research



- Full ARC compatibility
- Automated metadata generation
- Specialized tools and workflows for 'omics processing and analysis
- Public repository compatibility
- Galaxy teaching resource for data analysis

e.g. <https://plants.usegalaxy.eu>

Software Containers

- Big step towards reproducibility **and** reusability
- Help installing software (OS-agnostic)
- Help managing and documenting package and library dependencies

Examples:

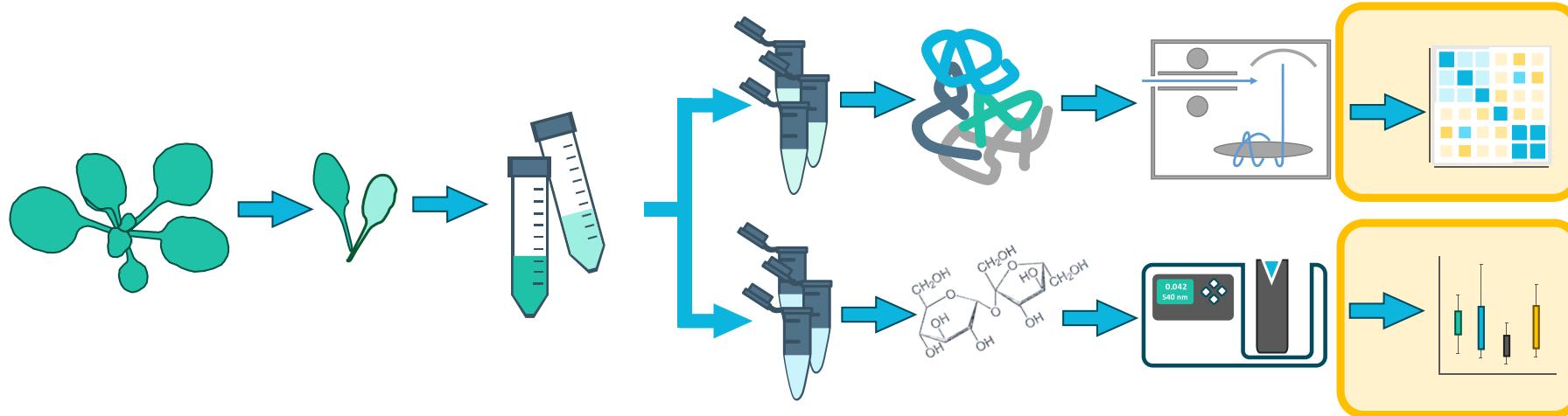
- <https://www.docker.com>
- <https://podman.io>

Resources

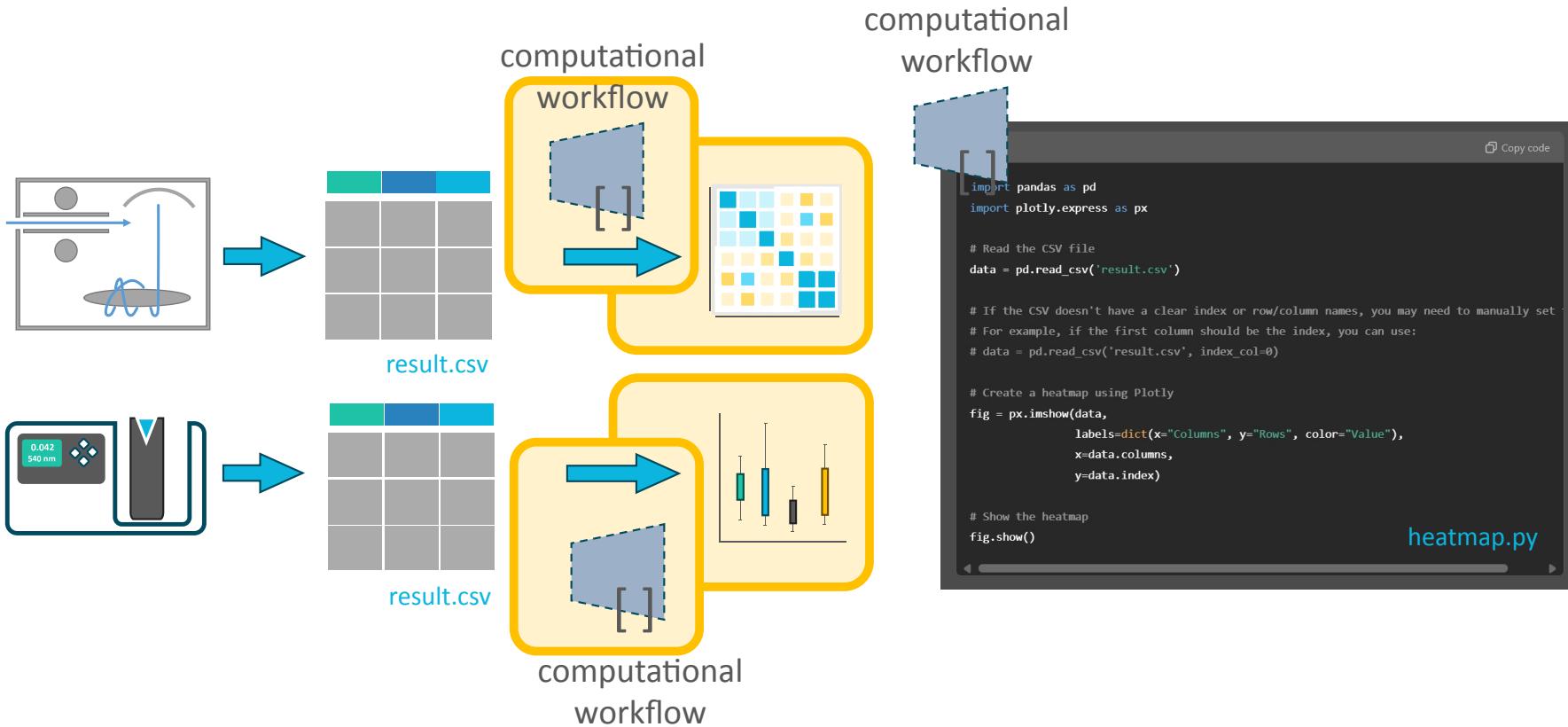
- <https://www.bioconductor.org>
- BioContainers: <https://github.com/BioContainers/>

Hands-on part 3: Data

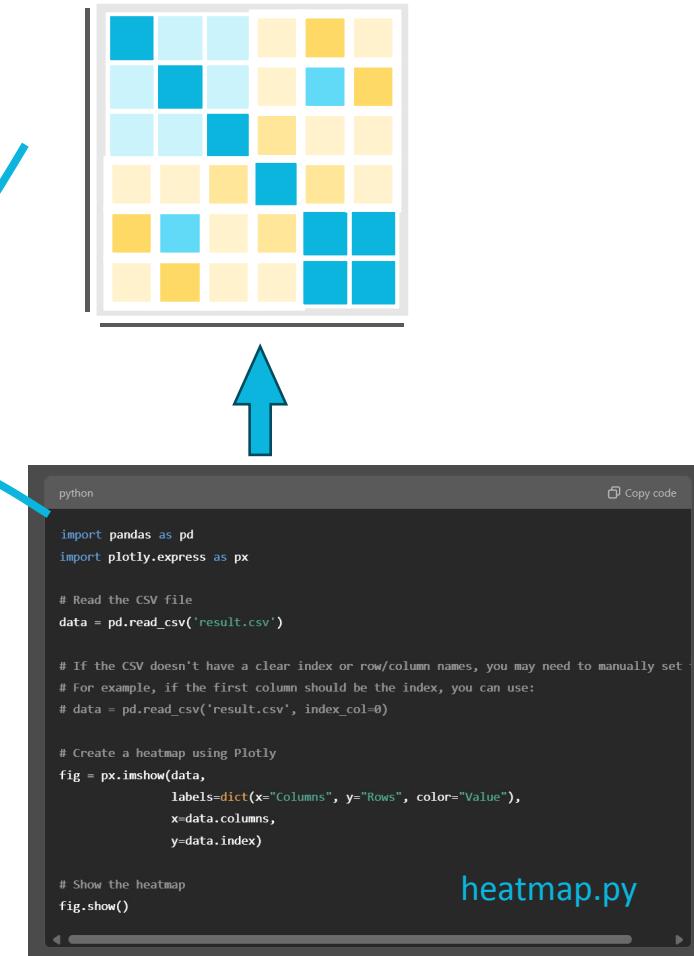
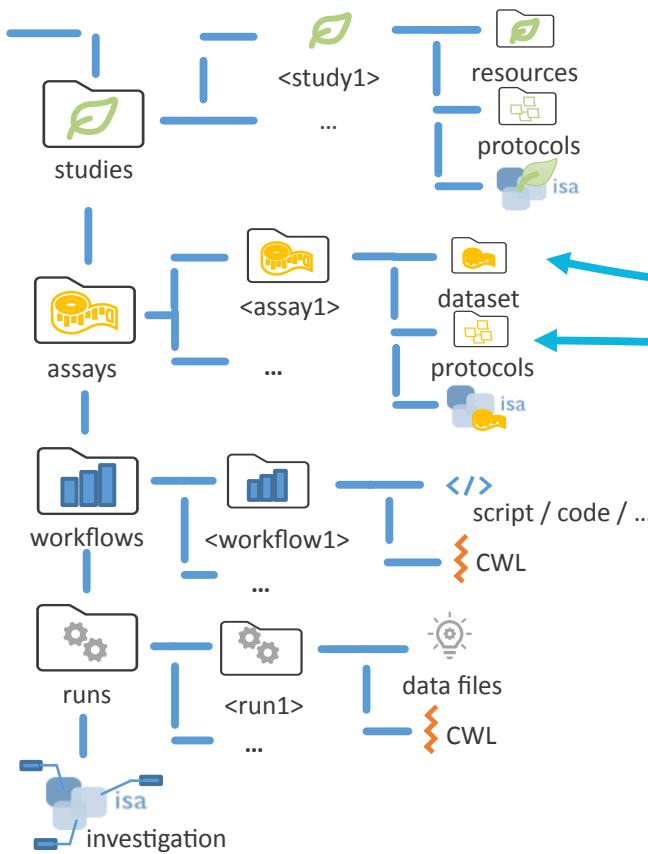
Options to annotate the data analysis



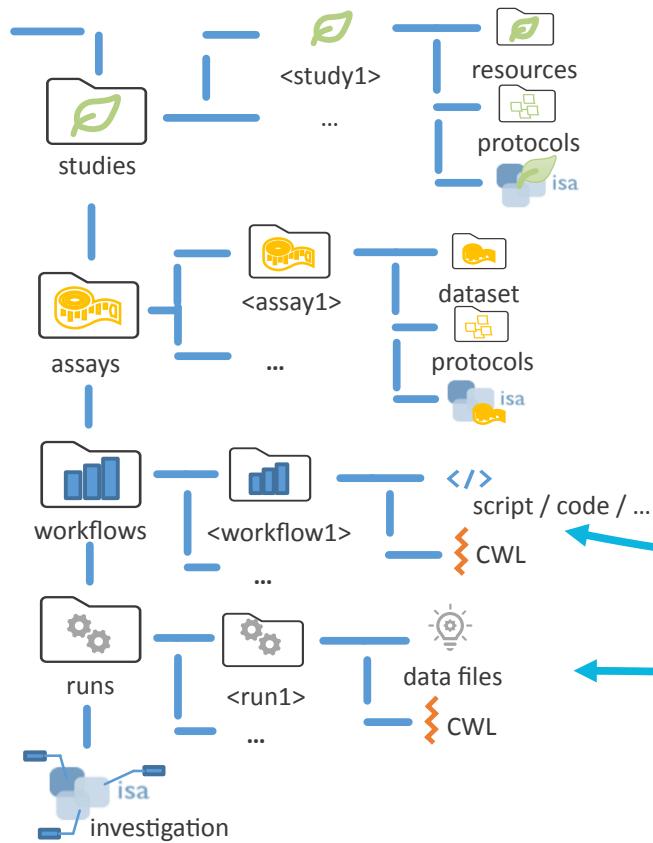
A computational workflow is like a protocol



Option I: Create a virtual assay



Option II: Create a workflow and run



A screenshot of a code editor window titled "heatmap.py" containing the following Python script:

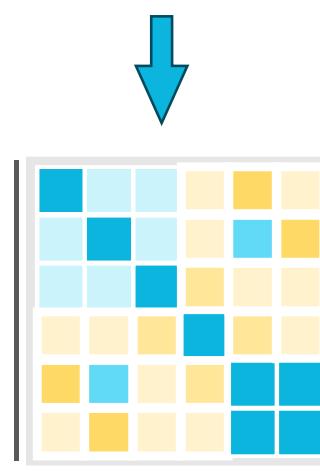
```
python
import pandas as pd
import plotly.express as px

# Read the CSV file
data = pd.read_csv('result.csv')

# If the CSV doesn't have a clear index or row/column names, you may need to manually set -
# For example, if the first column should be the index, you can use:
# data = pd.read_csv('result.csv', index_col=0)

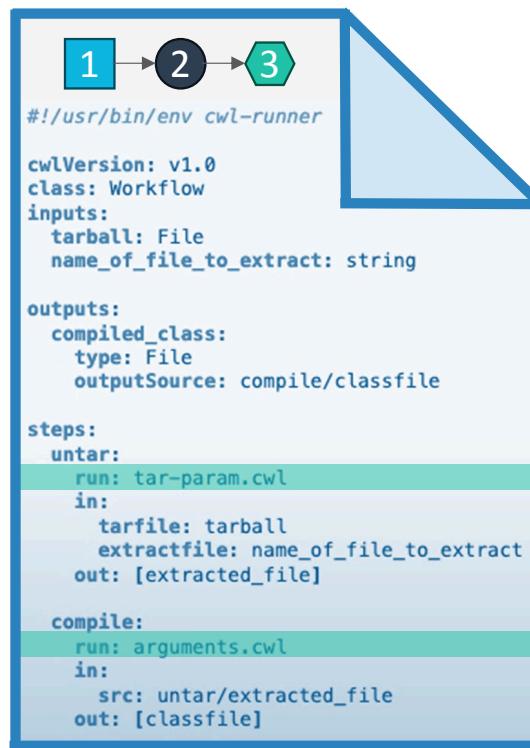
# Create a heatmap using Plotly
fig = px.imshow(data,
                 labels=dict(x="Columns", y="Rows", color="Value"),
                 x=data.columns,
                 y=data.index)

# Show the heatmap
fig.show()
```



Use CWL to wrap your workflow

CWL workflow document (*.cwl)



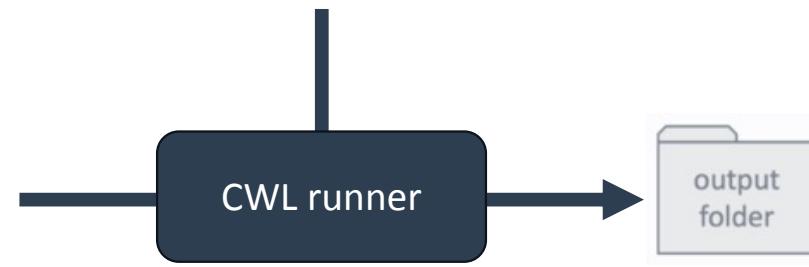
```
graph LR; 1[1] --> 2(( )); 2 --> 3[3];  
#!/usr/bin/env cwl-runner  
cwlVersion: v1.0  
class: Workflow  
inputs:  
  tarball: File  
  name_of_file_to_extract: string  
  
outputs:  
  compiled_class:  
    type: File  
    outputSource: compile/classfile  
  
steps:  
  untar:  
    run: tar-param.cwl  
    in:  
      tarfile: tarball  
      extractfile: name_of_file_to_extract  
    out: [extracted_file]  
  
  compile:  
    run: arguments.cwl  
    in:  
      src: untar/extracted_file  
    out: [classfile]
```

CWL
tool descriptors (*.cwl)

CWL job parameter (*.yaml)



```
job.yaml  
file: fastq  
param: 5  
workflow: wf.cwl  
output_folder:  
  /temp
```



Hands-on part 3: Data

Continue the [Start Here](#) guide in the DataPLANT knowledge base.



Until step **Data analysis**

Q&A and Wrap-up Day1

DataPLANT Resources – join the open source movement

DataPLANT 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...

 NEW! ARC website: <https://arc-rdm.org>

Continuous support

Data managers in Düsseldorf, Cologne, Jülich and close by (CEPLAS, MibiNet, TRR341) offer support.

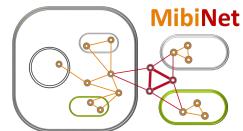
1. Slack Workspace for ad hoc support
2. Monthly user meeting (2nd Friday of the month)

→ [Details](#).



trr_341

plant
ecological
genetics



Overview of Institutional services at UoC and HHU

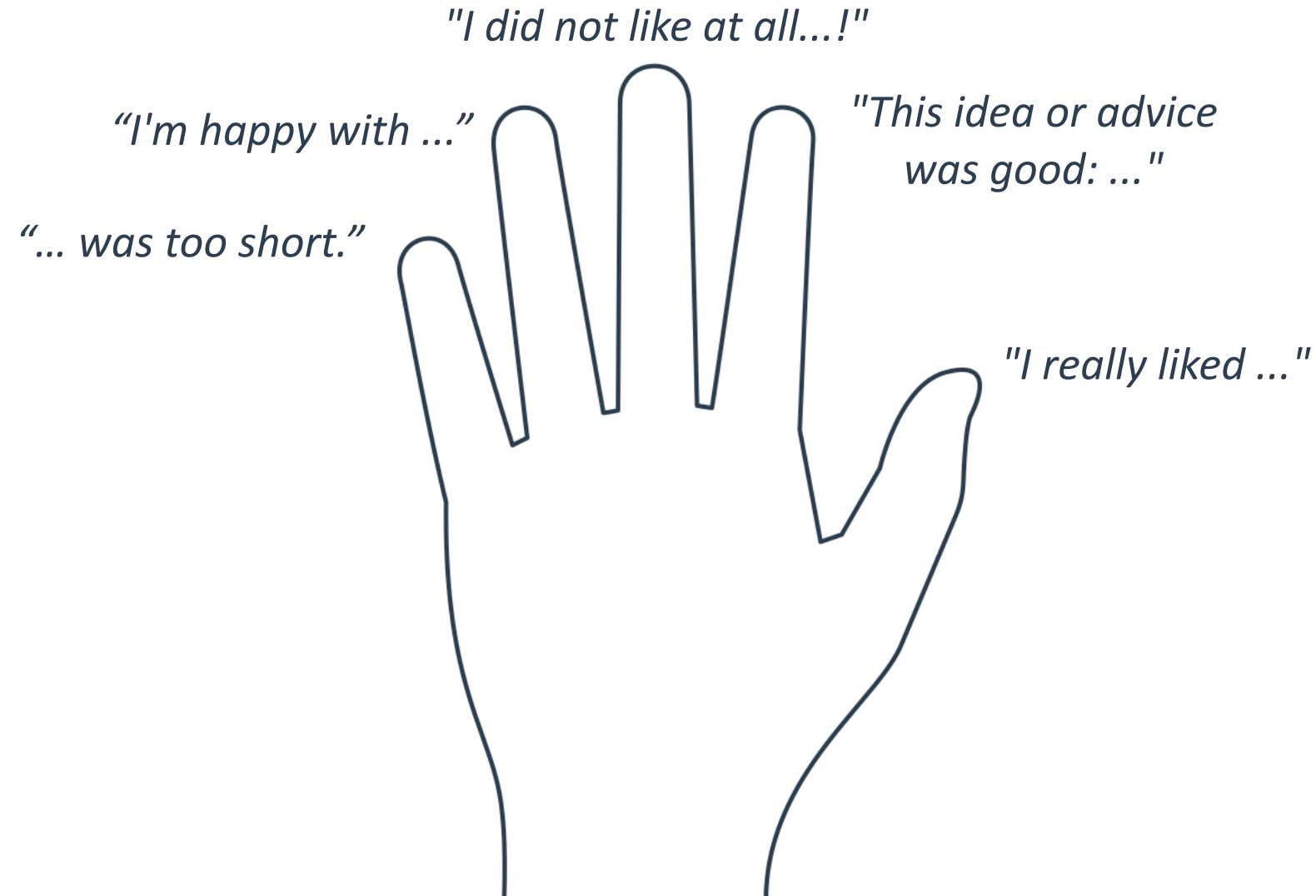
UoC

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

HHU

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

Five-Finger-Feedback





CEPLAS

Cluster of Excellence on Plant Sciences

Good Data Management Practices

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

November 12th, 2024

Vittorio Tracanna, Dominik Brilhaus
CEPLAS Data



House-keeping

Pad: <https://pad.hhu.de/0NdPK05LQ5CHBRN2iuG91Q>

Points to discuss from and since day 1

Data Storage and Versioning

Data stores

Local hard disks



Institute server



University server



Cloud services



Mail



labfolder

Electronic lab
notebooks



Git repositories

Wiki, Project management

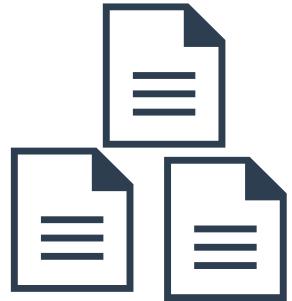


Backup vs. Archive

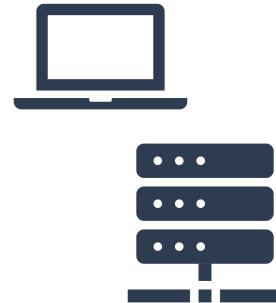
	Backup	Archive
Storage type	Short-, mid-term	Long-term
Purpose	Disaster recovery	Long-term storage, compliance
Reason	Duplication	Migration
Usage	Work in progress	Cold, Unused data
Changes	Short-term updates	No updates
Trend	Cyclic, Replacement	Growing
Latency	Short/Costly	High/Cheaper

3-2-1 backup rule

*3 copies
of data*



*2 storage
media*



*1 copy
off-site*



Version control and track changes

It's good practice to document:

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

Types of Version Control

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

Which files need to be "versioned"?



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

Concept of Git and git-based platforms

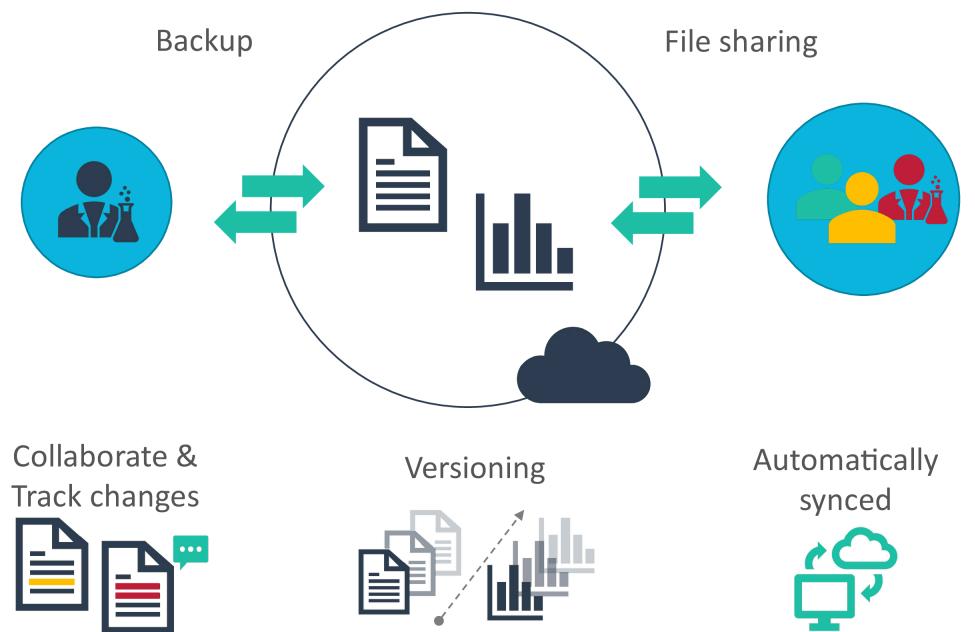
Cloud Services

- ✓ Documents
- ✓ Small data
- ✓ Presentations

X Code

X Data analytical projects

X Big (“raw”) data



Git and git platforms

- ~ Documents
- ✓ Small data
- ~ Presentations

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

Why git? => Why code?

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

A simple example: RNASeq project

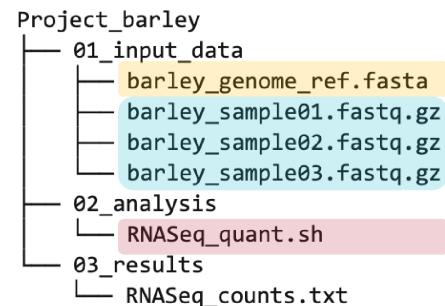
```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
└── 02_analysis
    └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

A simple example: RNASeq project

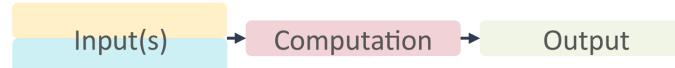
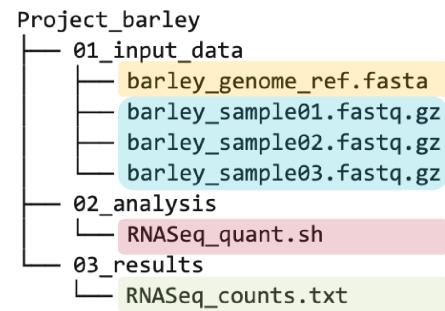
```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
└── 02_analysis
    └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

Input(s)

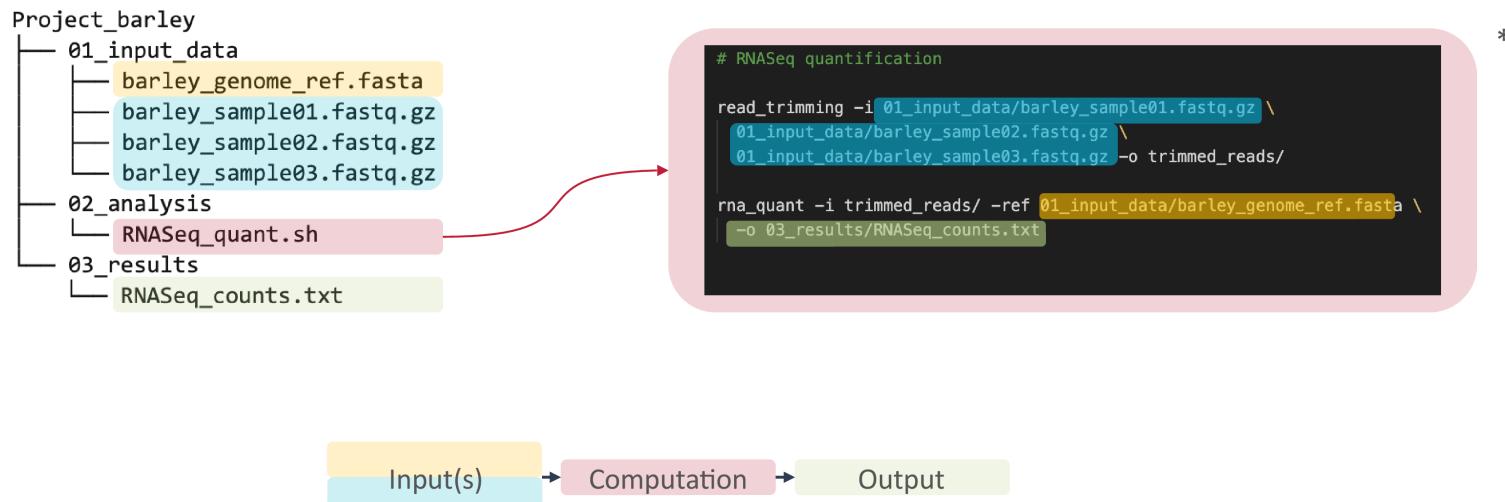
A simple example: RNASeq project



A simple example: RNASeq project



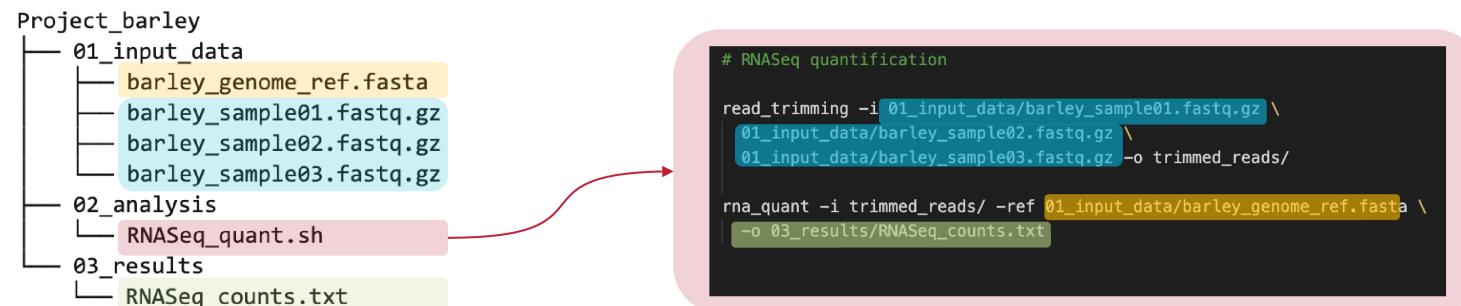
A simple example: RNASeq project



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

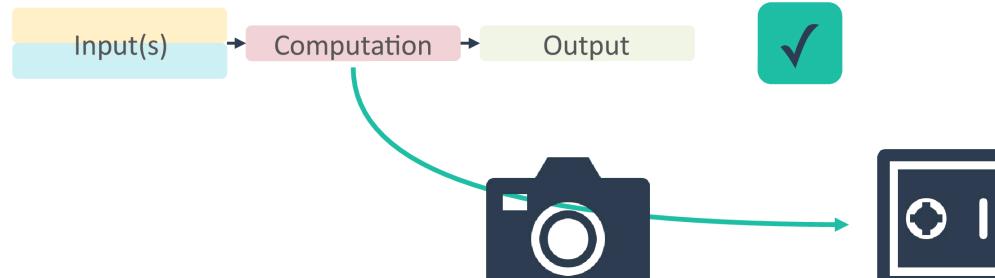
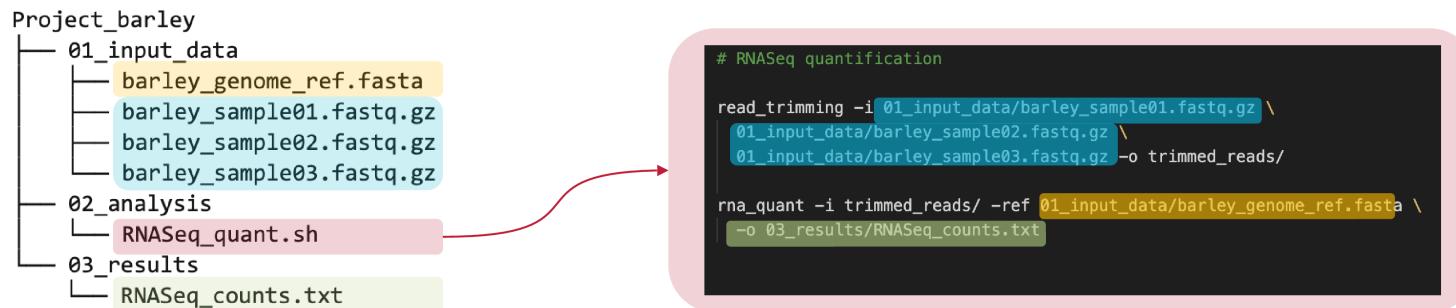
Take snapshots of your code work...

(... as long as it works)



Take snapshots of your code work...

(... as long as it works)



Scenario 1: More data

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

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

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

Scenario 1: More data

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

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

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

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

Scenario 1: More data

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

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

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

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

Let git track changes and keep things clean

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

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

“version 1”

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

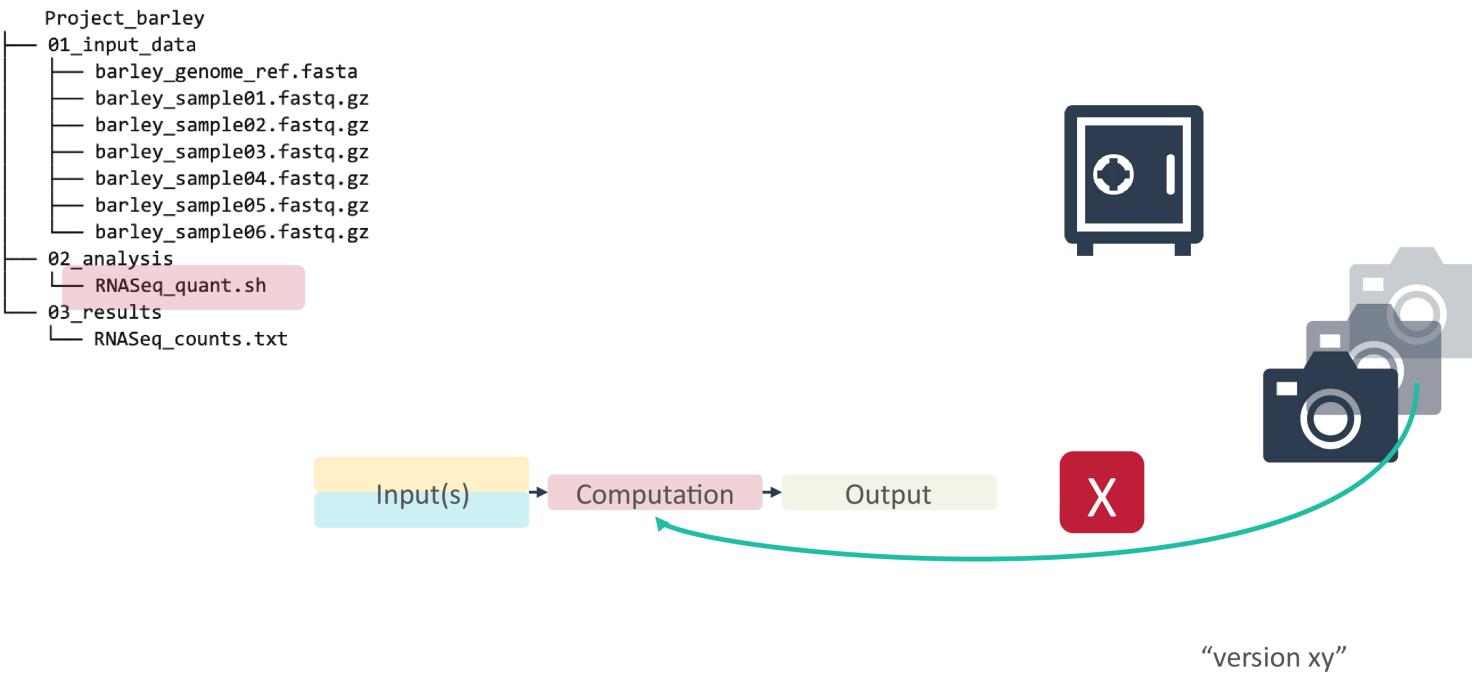
“version 2”

Scenario 2: Pipeline breaks

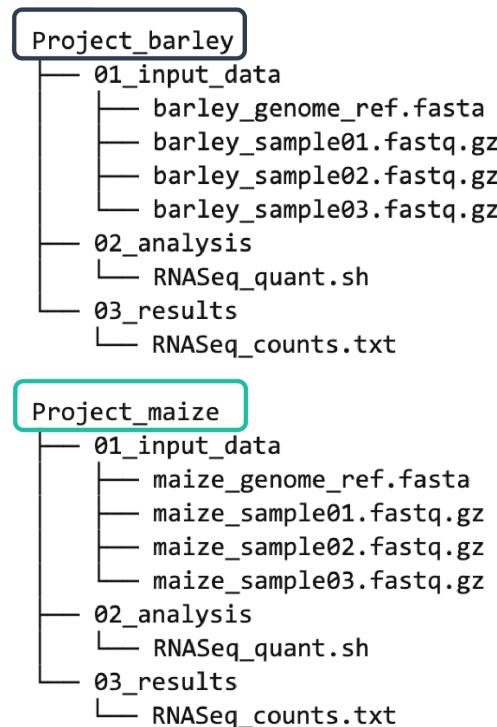
```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   ├── barley_sample03.fastq.gz
│   ├── barley_sample04.fastq.gz
│   ├── barley_sample05.fastq.gz
│   └── barley_sample06.fastq.gz
└── 02_analysis
    └── RNASeq_quant.sh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



Revert to snapshot



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



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

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

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

```
# RNASeq quantification

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

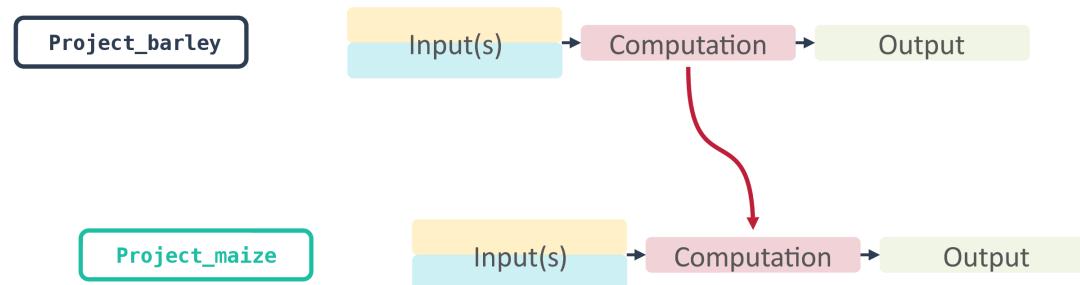
rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
-o 03_results/RNASeq_counts.txt
```

```
# RNASeq quantification

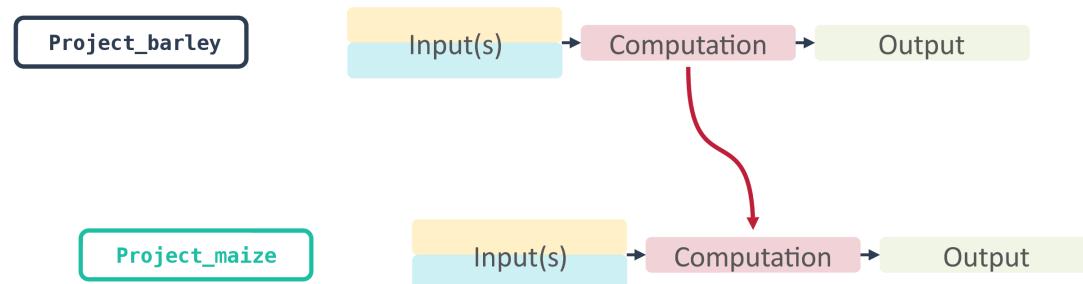
read_trimming -i 01_input_data/maize_sample01.fastq.gz \
  01_input_data/maize_sample02.fastq.gz 01_input_data/maize_sample03.fastq.gz \
-o trimmed_reads/

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

Re-use code



Re-use code



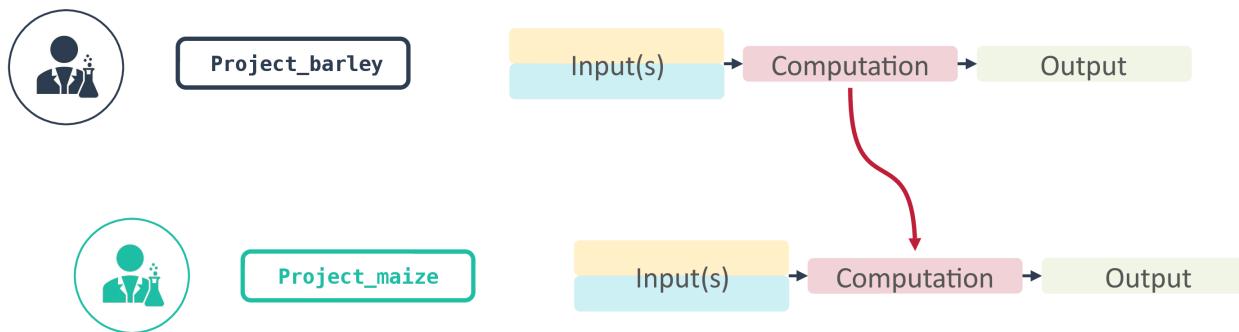
```
1 # RNASeq quantification
2
3 read_trimming -i 01_input_data/barley_sample01.fastq.gz \
4 01_input_data/barley_sample02.fastq.gz \
5 01_input_data/barley_sample03.fastq.gz \
6 -o trimmed_reads/
7
8 rna_quant -i trimmed_reads/ -ref 01_input_data/barley_genome_ref.fasta \
9 -o 03_results/RNASeq_counts.txt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

“version barley”

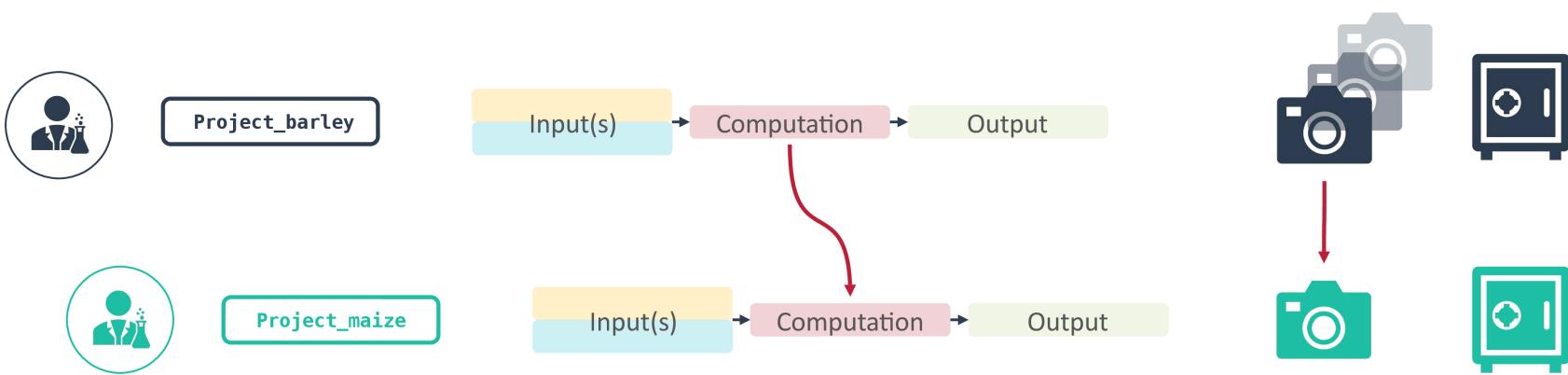
```
1 # RNASeq quantification
2
3+ read_trimming -i 01_input_data/maize_sample01.fastq.gz \
4+ 01_input_data/maize_sample02.fastq.gz 01_input_data/maize_sample03.fastq.gz
5 -o trimmed_reads/
6
7+ rna_quant -i trimmed_reads/ -ref 01_input_data/ maize_genome_ref.fasta -o 03_results/RNASeq_counts.txt
```

“version maize”

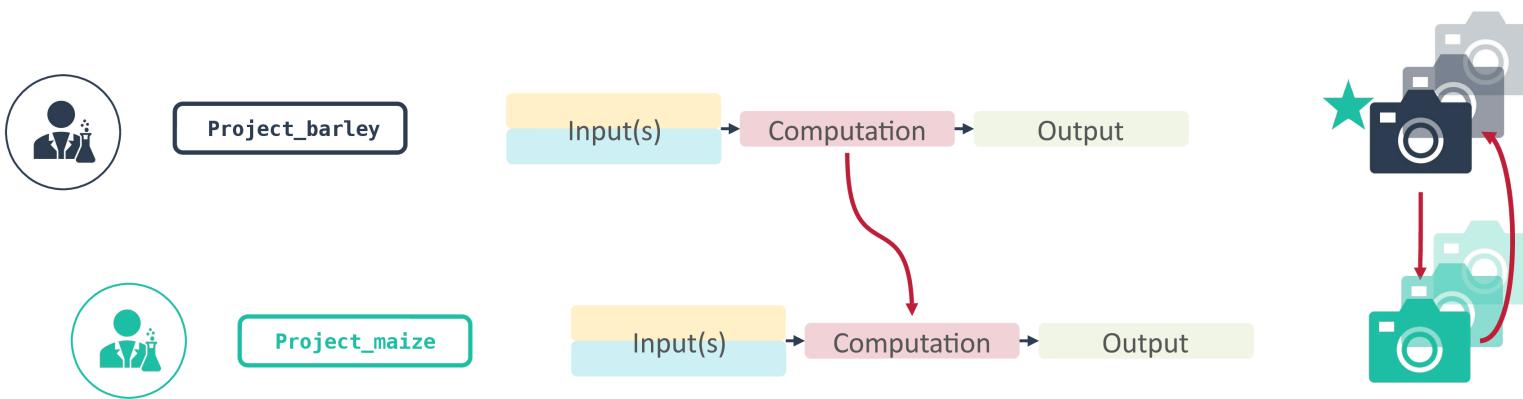
Re-use code – People have done this



Re-use code – People have done this



Re-use code – Link and contribute



Git: summary

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

GitHub and GitLab

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

GitHub and Gitlab team projects

Simplifies concurrent work & merging changes

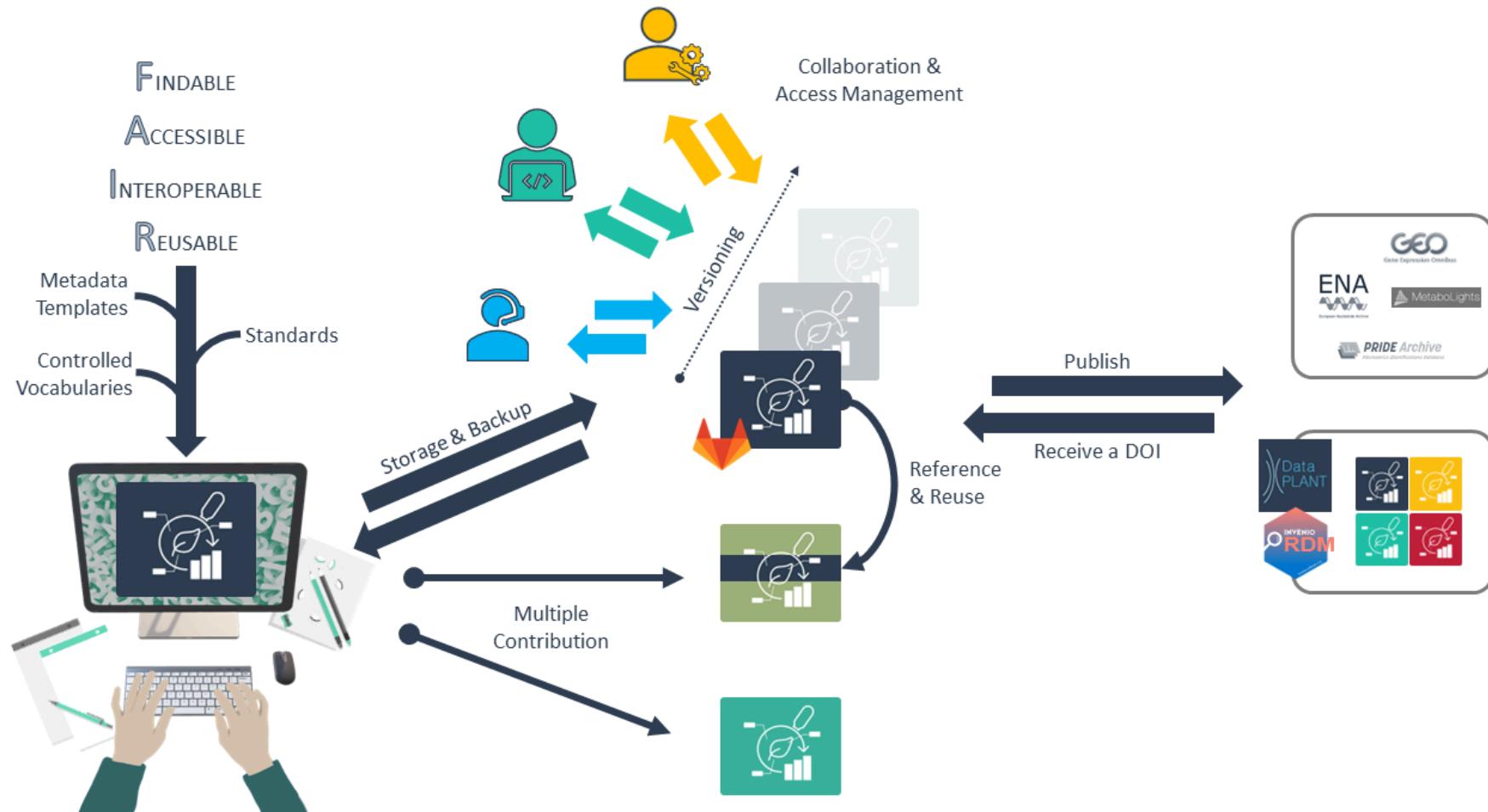
- Online service to host our projects
- Share code with other developers
- Others can download our projects, work on and contribute to them
- They can upload their changes and merge them with the main project

Cloud vs. Git

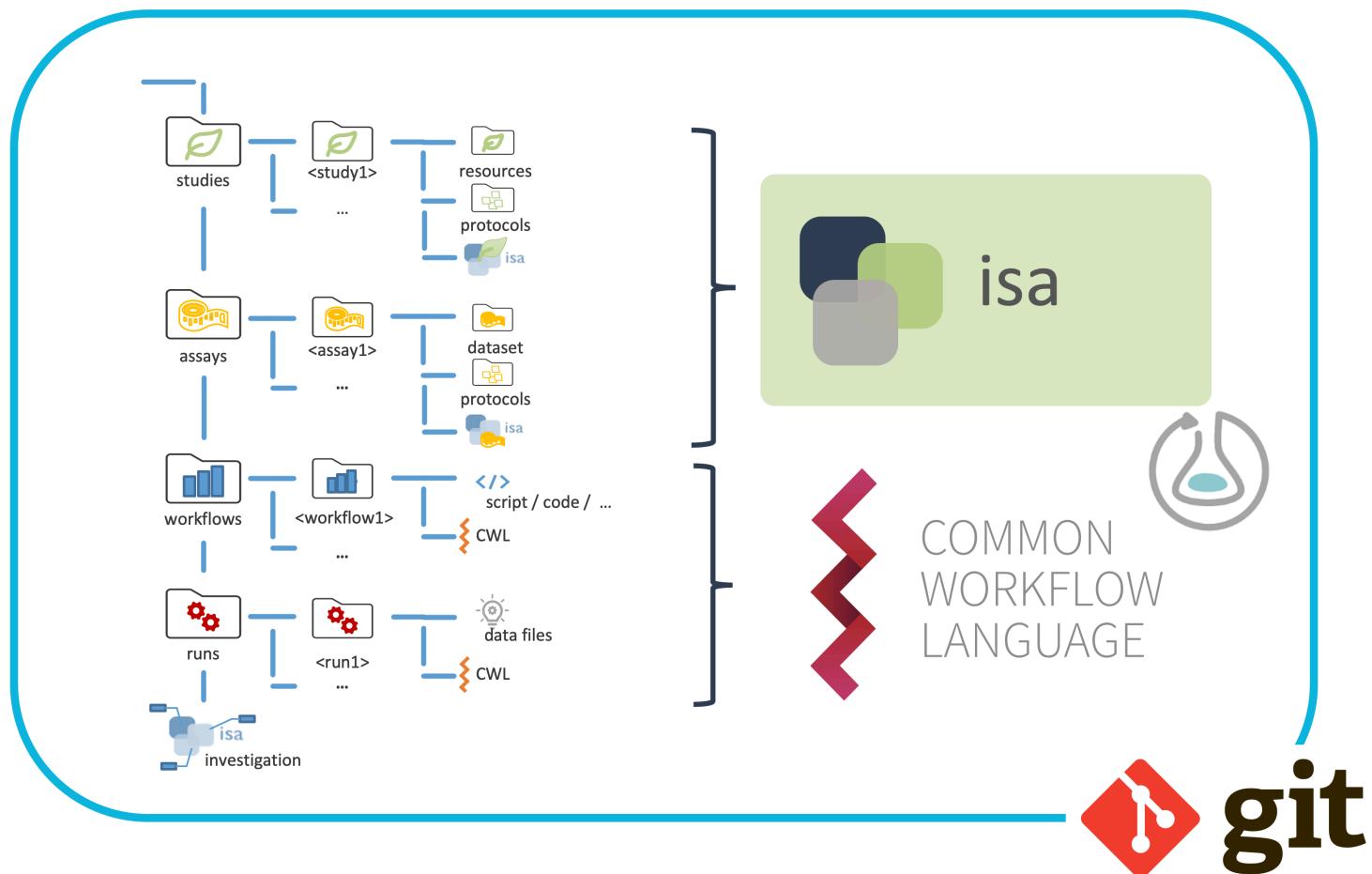
	Cloud services	Git / GitHub / GitLab
Track changes		  
Collaboration		 
Versioning		issue tracker, tracked contribution
Syncing		Well-documented (commit history)
Access		Active / controlled by user
Data security		Easily collaborate across institutions
	Oftentimes only within organization / institution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

Hands-on part 4: DataHUB

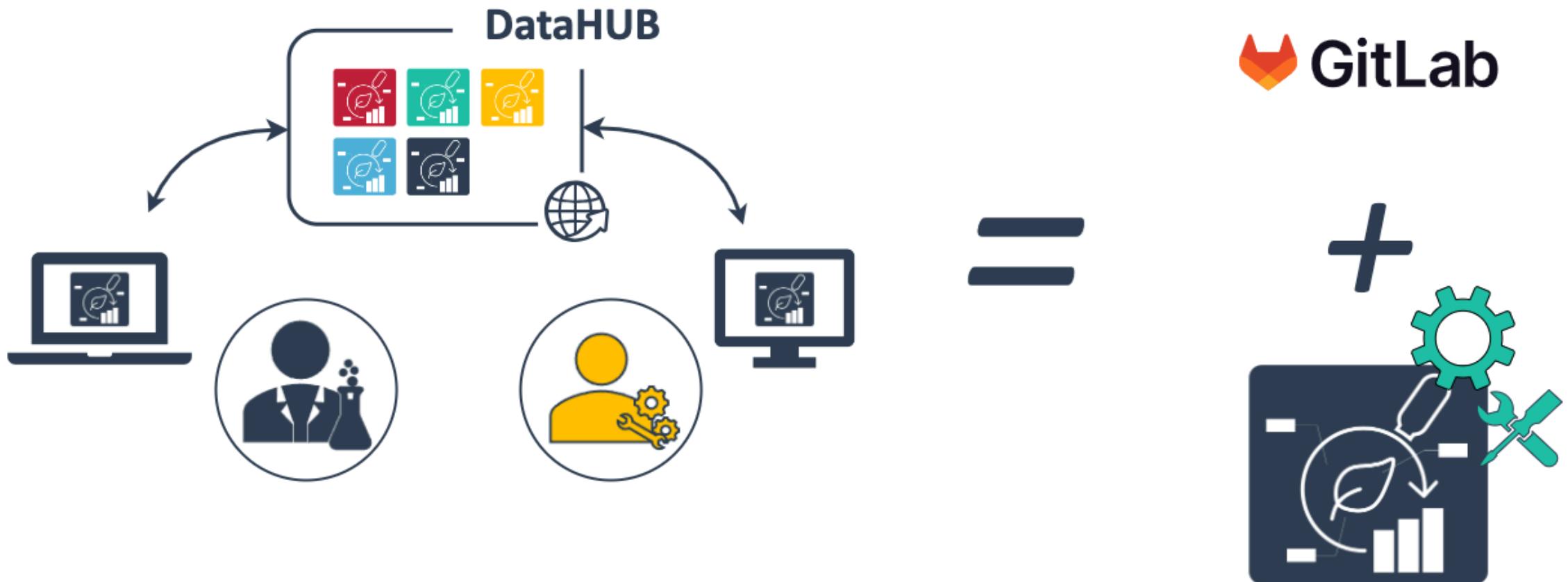
DataHUB overview



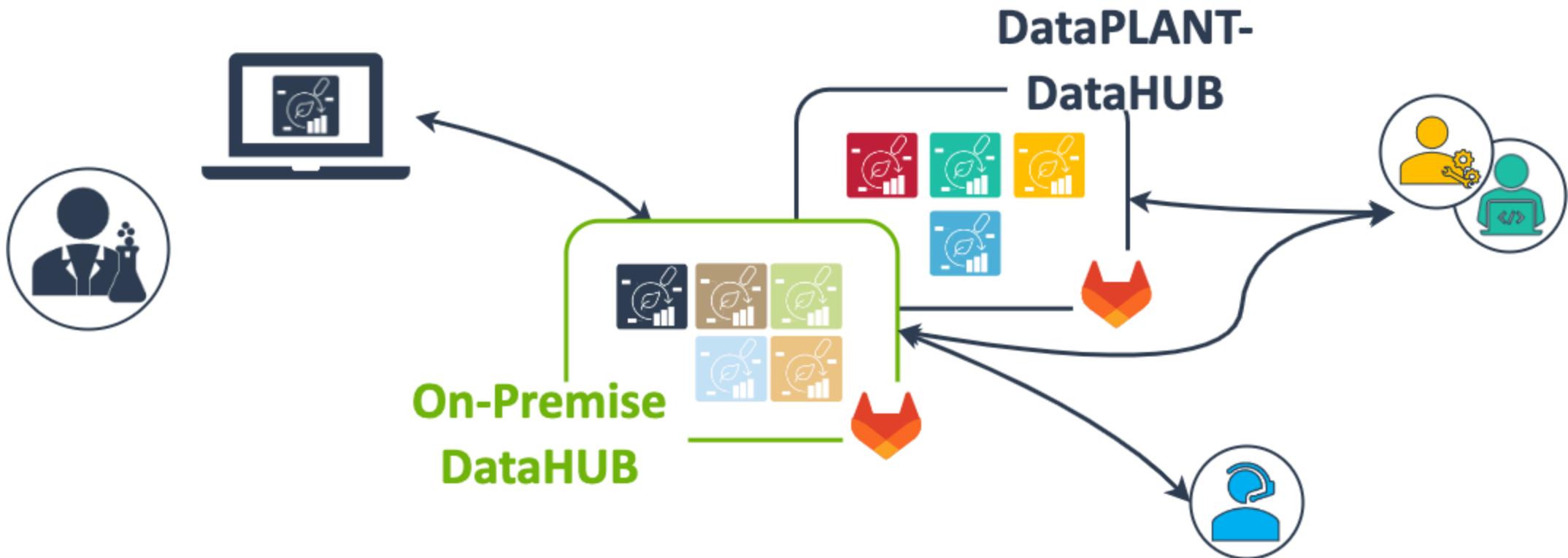
ARC builds on standards + Git



The DataPLANT DataHUB – a GitLab *Plus*

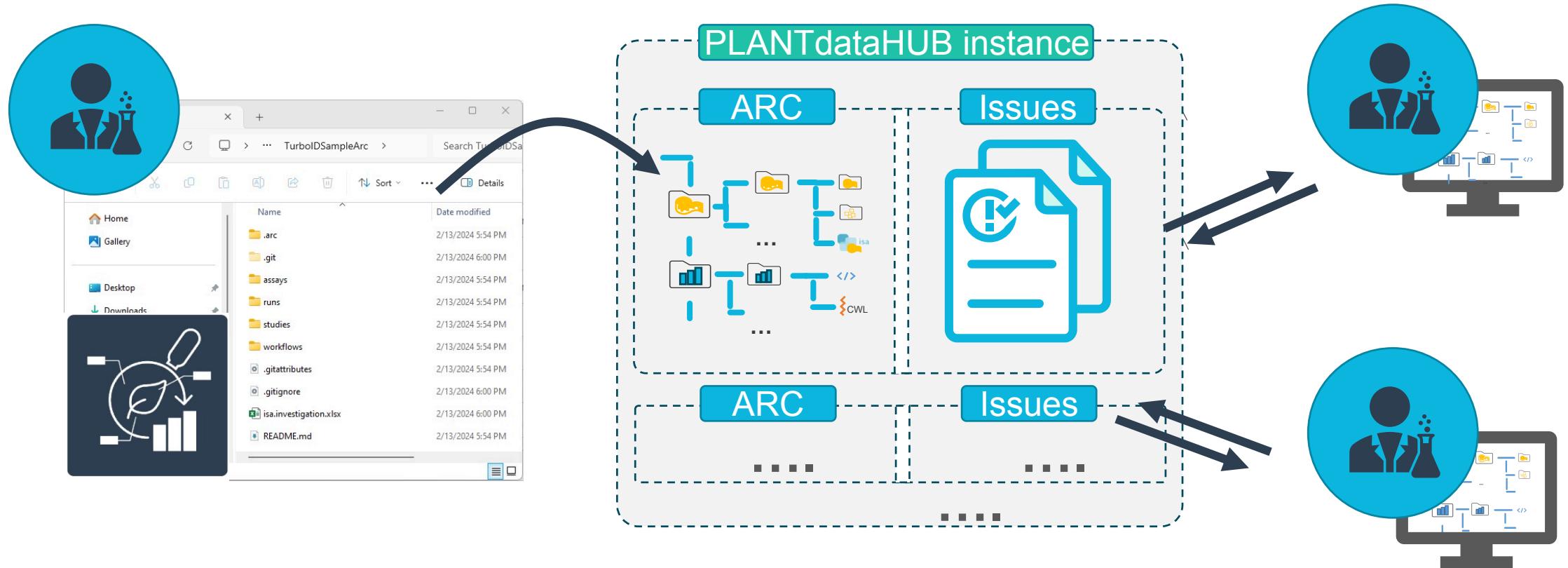


On-premise DataHUBs

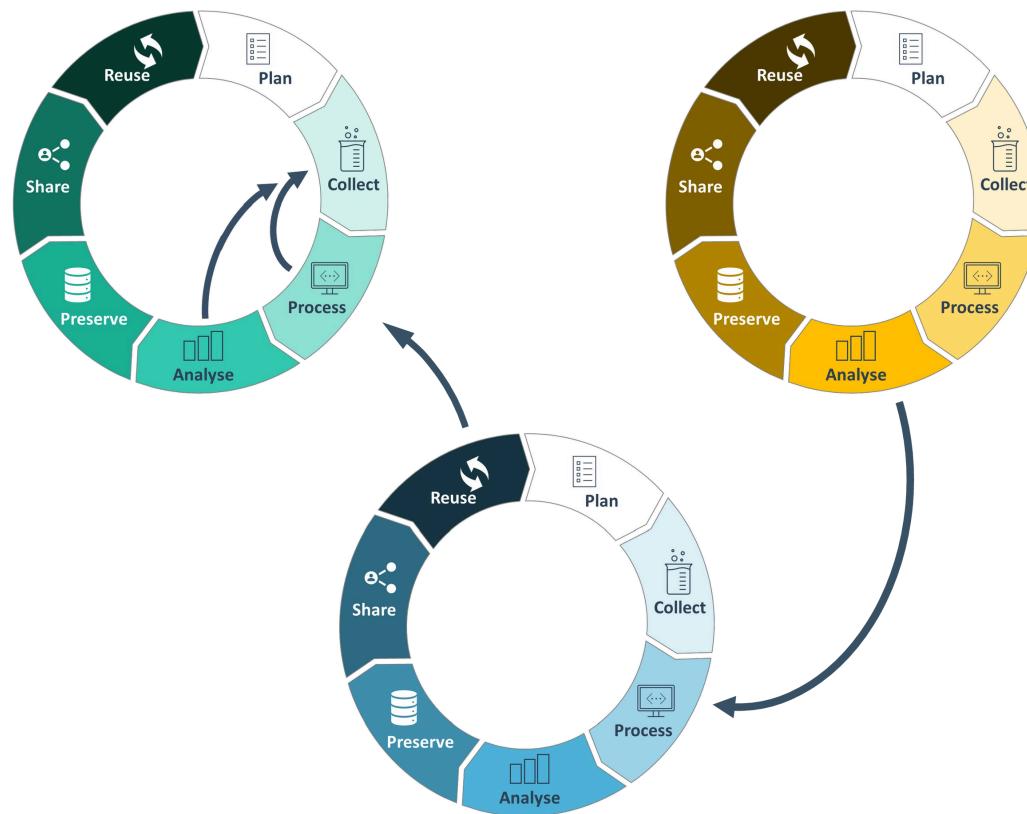


ARC services are available as on-premise option

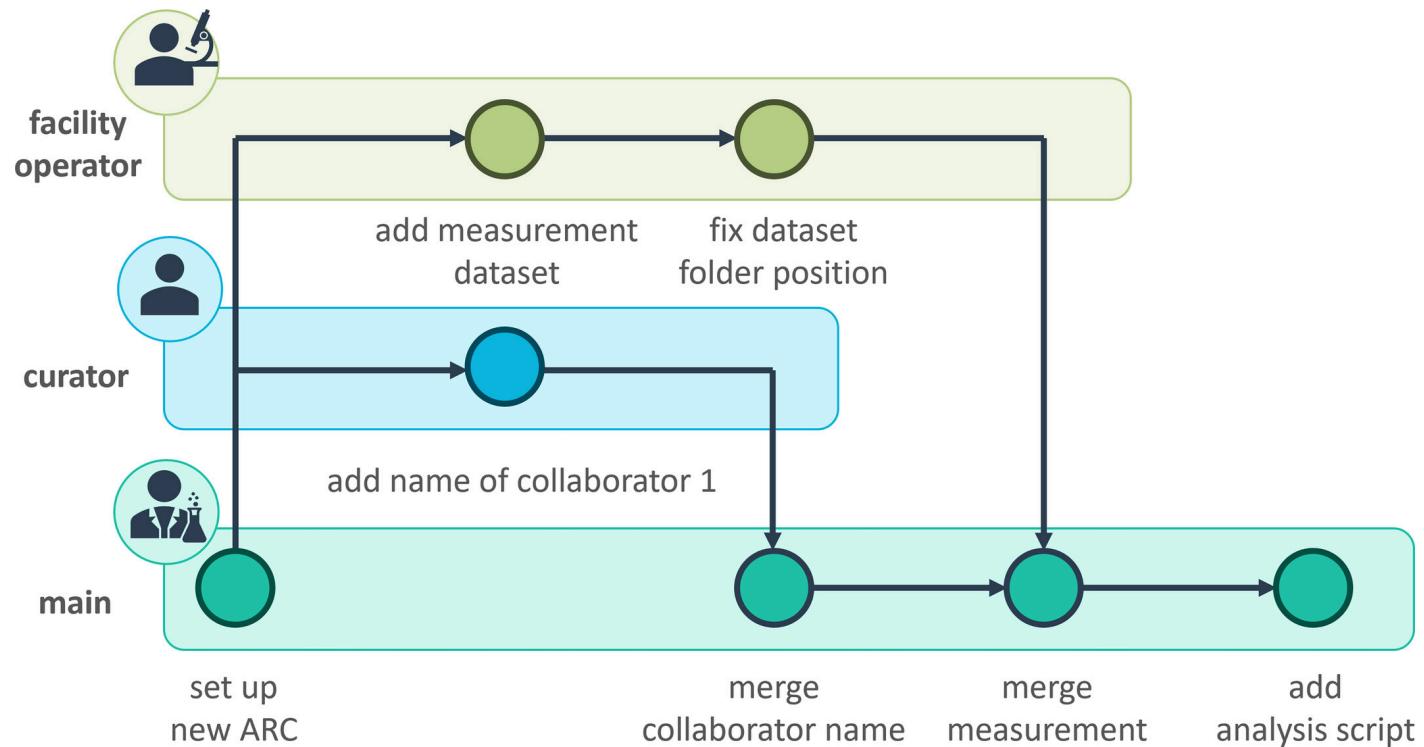
Using the DataHUB to collaborate



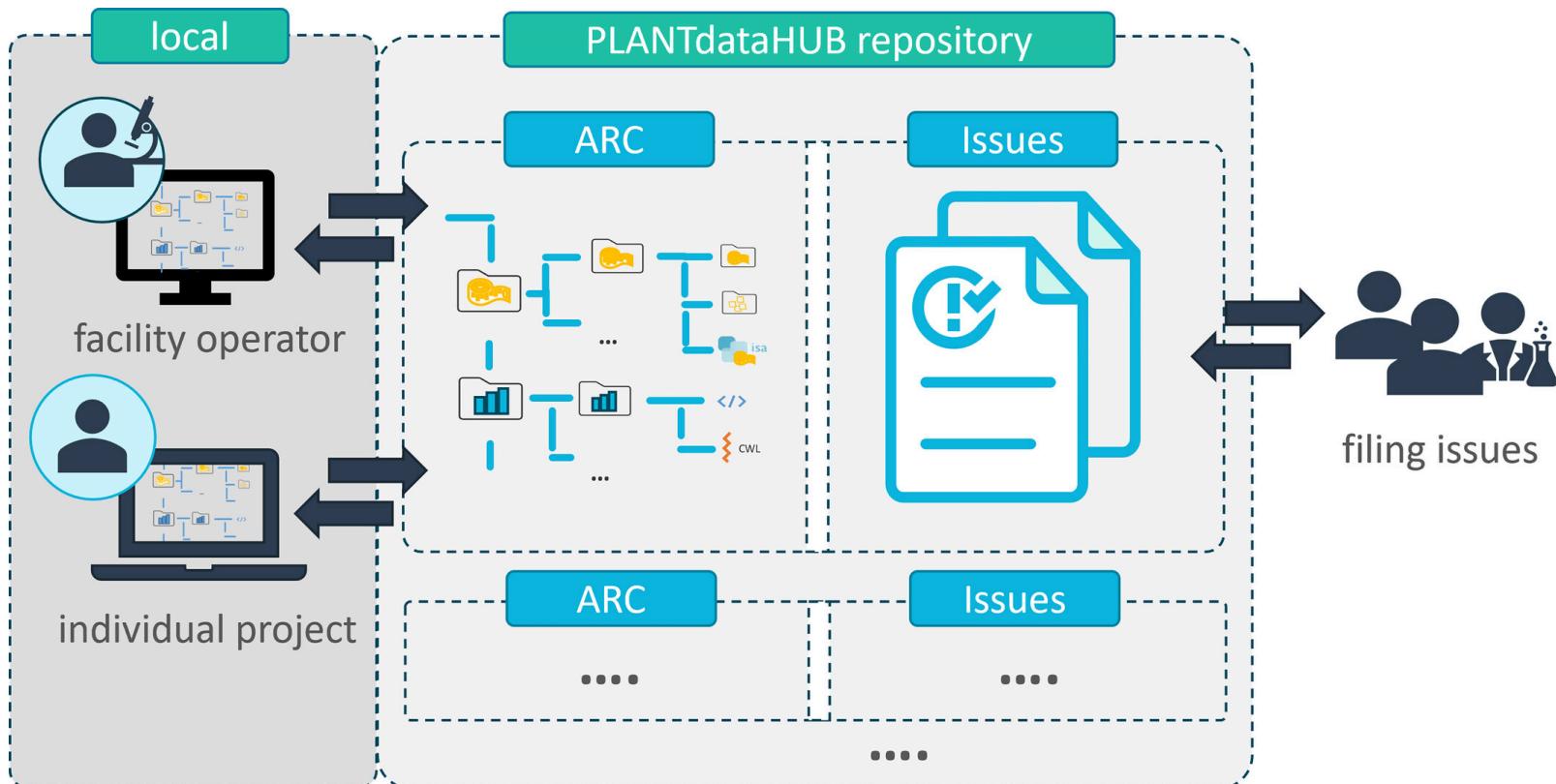
The Research Data Lifecycle *is mutable*



Mutable data life cycle



Project management



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. On the left is a sidebar with various project management options: Demo_ARC (selected), Pinned, Manage, Plan, Issues (0), Issue boards, Milestones, Wiki (selected), Code, Build, Secure, and Help. The main content area is the 'Home' page of the Demo_ARC wiki. It features a message about CQC pipelines, a 'Pages' section (2 pages), and sections for Home, Meeting Schedule, and Ideas and drafts.

Project

- D Demo_ARC
- Pinned
- Manage
- Plan
- Issues (0)
- Issue boards
- Milestones
- Wiki (selected)
- Code
- Build
- Secure
- Help

Home

Last edited by Demo User just now

This is the wiki to Demo_ARC. We will announce meeting schedules here.

Pages 2

Meeting Schedule

- 2024-06-12 Kick-off
- 2024-06-27 Proposal discussion
- 2024-07-04 Discuss RNA-seq pipeline

Ideas and drafts

- Golden Gate protocol

Hands-on part 4: DataHUB

Continue the [Start Here](#) guide in the DataPLANT knowledge base.

-  Until the final **complete** step

DataHUB terminology and data sharing

Choosing the proper role

When inviting new members to an ARC or group, you can choose between different levels.

Permissions & Roles

Roles are assigned when adding a user to an ARC or to a group. This is a very simplified summary.

Guest – Can only see the ARC's wiki

Reporter – Can read, but not add or edit data

Developer – Reporter permissions + can read, add, and edit data

Maintainer – Developer permissions + can add new members

Owner – Maintainer + can delete ARC, manage memberships and permissions

 By default you are **Owner** of an ARC you create or upload to the DataHUB.

Projects and Groups are not the same

- "Project" = ARC
- "Groups" = Group of users

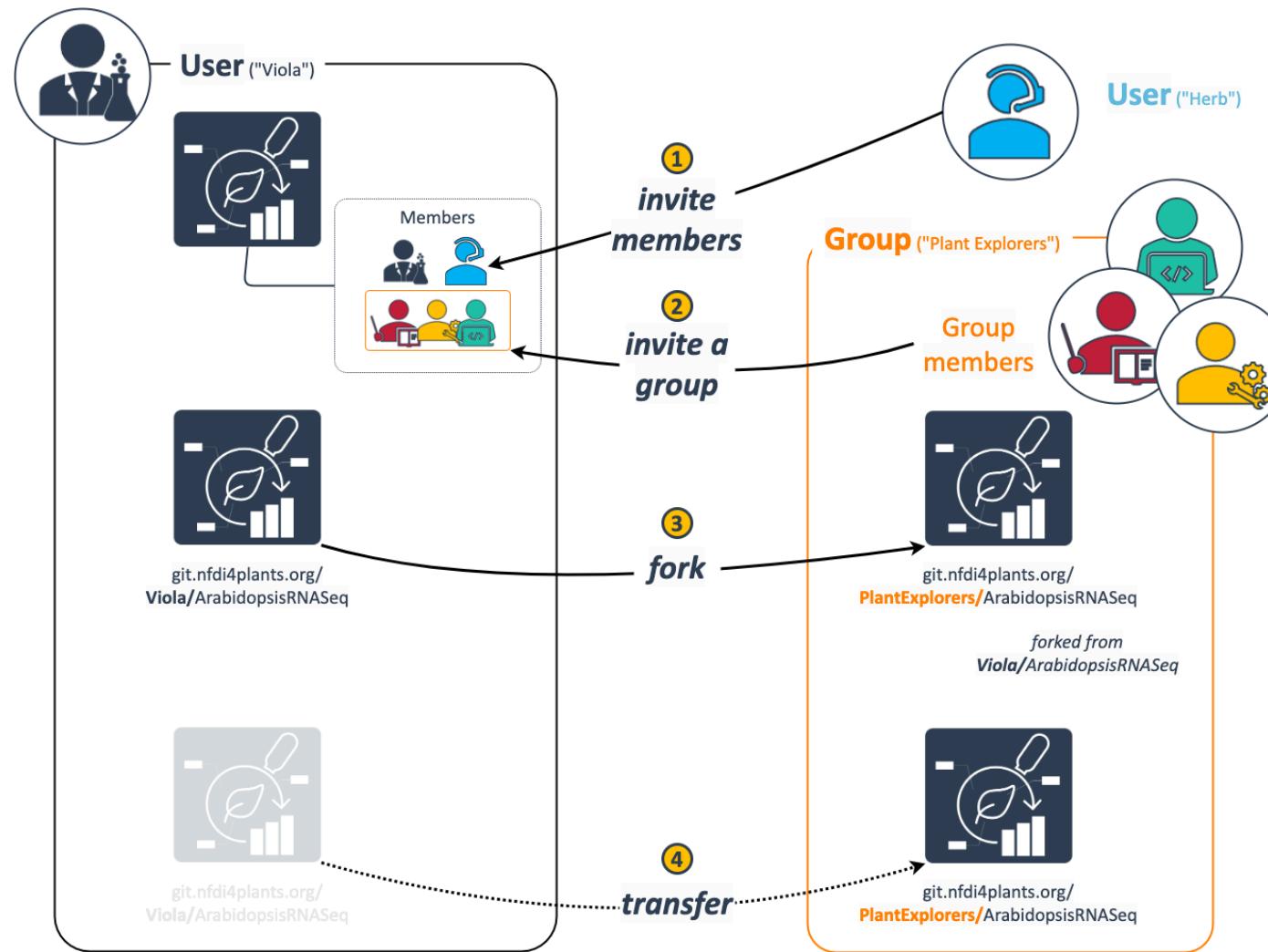
Project = ARC

- In the DataHUB, ARCs are called "projects"; they are the same.
- An ARC can be shared with individual users (invited as "members") or a group.

DataHUB Groups = Group of members (e.g. lab)

- A "Group" is a group of users with specific permissions
- A group can share ARCs
- A group can be invited to an ARC
- Groups can have subgroups

Options to share an ARC via the DataHUB



Namespaces

- Every user has a personal namespace, where they can upload or create new ARCs
- Every group and subgroup has its own namespace

Type	URL	Namespace	Name
A personal ARC	https://git.nfdi4plants.org/brilator/Facultative-CAM-in-Talinum	brilator	Dominik Brilhaus
An group-shared ARC	https://git.nfdi4plants.org/hhu-plant-biochemistry/Samuilov-2018-BOU-PSP	hhu-plant-biochemistry	HHU Plant Biochemistry

 **Personal** is not the same as **private**

Visibility

The visibility of ARCs and groups can be managed individually for each ARC or group

Visibility

The visibility of each ARC can be managed in the settings of the ARC



Private – ARC access must be granted explicitly to each user or group.



Internal – ARC can be accessed by any logged in user.



Public – ARC can be accessed without authentication.



By default every ARC and every group is set to **private**.

ARC DataHUB members // ARC Investigation contacts

Project members

You can invite a new member to **Samuilov-2018-BOU-PSP** or invite another group.

Account	Source	Max role	Expiration	Activity
Adriano Nunes-Nesi @nunesnesi	HHU Plant Biochemistry by Sebastian Triesch	Maintainer	Expiration date	User created: Jul 05, 2023 Access granted: Jul 10, 2023 Last activity: Aug 09, 2023
Andreas Weber @andreas.weber	HHU Plant Biochemistry by Sebastian Triesch	Maintainer	Expiration date	User created: Mar 10, 2023 Access granted: Jul 31, 2023 Last activity: Sep 11, 2023
Dominik Brilhaus It's you @brilator	Direct member by Dominik Brilhaus	Owner	Expiration date	User created: Feb 21, 2022 Access granted: Dec 06, 2023 Last activity: Mar 26, 2024
Franziska Fichtner @franziska.fichtner	HHU Plant Biochemistry by Sebastian Triesch	Maintainer	Expiration date	User created: Aug 11, 2023 Access granted: Aug 11, 2023 Last activity: Aug 11, 2023

DataHUB: ARC members

https://git.nfdi4plants.org/hhu-plant-biochemistry/Samuilov-2018-BOU-PSP/-/project_members

Identifier: Samuilov-2018-BOU-PSP

Title:

Description:

Contacts:

Contact	Score
Sladjana Samuilov <orcid>	4/10
Nadine Rademacher <orcid>	3/10
Samantha Flachbart <orcid>	3/10
Leila Arab <orcid>	3/10
Saleh Alfarraj <orcid>	3/10
Franziska Kuhnert <orcid>	3/10
Stanislav Kopriva <orcid>	3/10
Andreas P. M. Weber <orcid>	4/10
Tabea Mettler-Altmann <orcid>	3/10

ARCIctect: Investigation Contacts

Investigation contacts are not automatically invited as members to the ARC.

Version control

Check out the **commit history** of your ARC via Repository (2) or directly via commits (7)

The screenshot shows the Data PLANT ARC interface. On the left, there's a sidebar with navigation links: Manage (1), Plan (2), Code (3), Build, Secure, Deploy, Operate, Monitor, Analyze, Settings (4), and Help. The main area shows a project named "Demo_ARC" (5). A commit history table (6) lists the following files and their last commits:

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

At the top right, there's a repository header with branches (main), a code dropdown (7), and a project information section (8) showing a pipeline status of "passed". The bottom right shows the project was created on July 13, 2024.



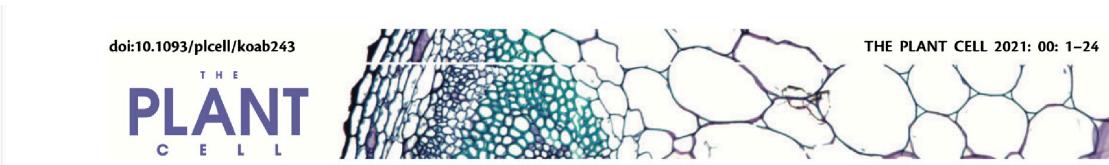
Structure your project as an ARC



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

Data publication and repositories

Spot the PIDs



Research Article

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

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

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

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

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

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

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

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

‡Senior authors

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

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

Abstract

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

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

Persistent Identifiers (PIDs)

Globally unique, stable, persistent

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



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



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



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



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



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



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

Properties of a PID

Ideally, PIDs are

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

Additional resources

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

Institutional publication guidelines

HHU Düsseldorf recommends use of ORCID and other PIDs

Publikationsrichtlinie der Heinrich-Heine-Universität Düsseldorf vom 09.11.2023:

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

Domain-specific data repositories

Good

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

Intermediate

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

Domain-specific data repositories

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

Choosing a data repository

Domain-specific >> Generic >> Institutional

Find repositories at:

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

Generic data repositories

Good

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



<https://zenodo.org>



<https://datadryad.org/>

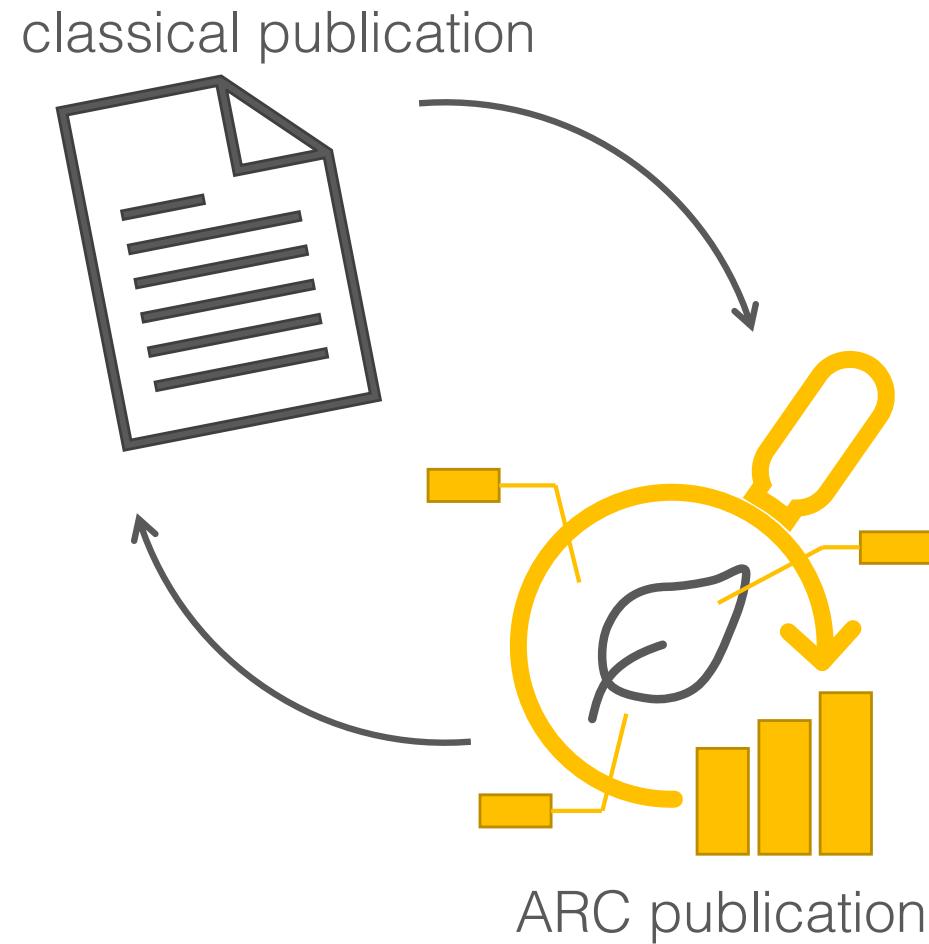
Intermediate

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

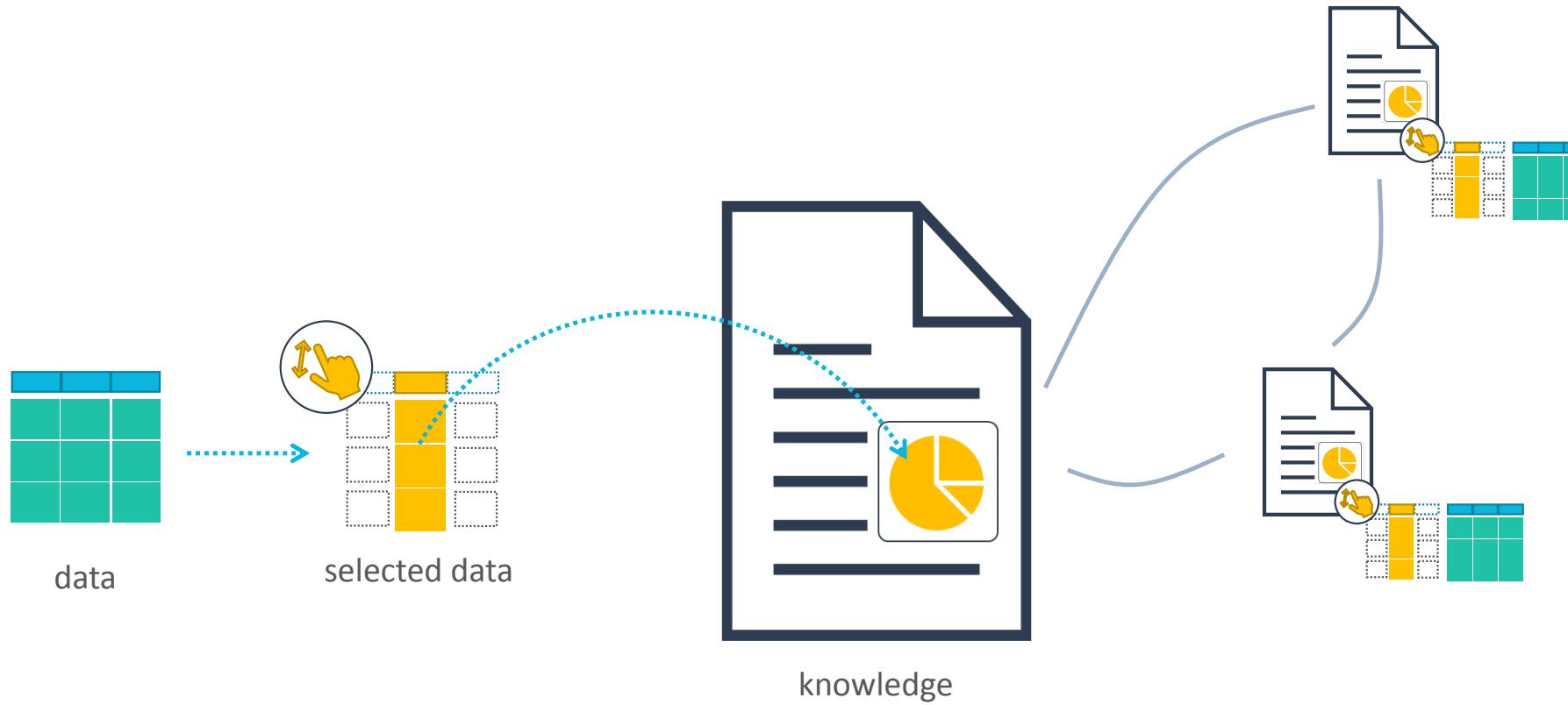


<https://figshare.com>

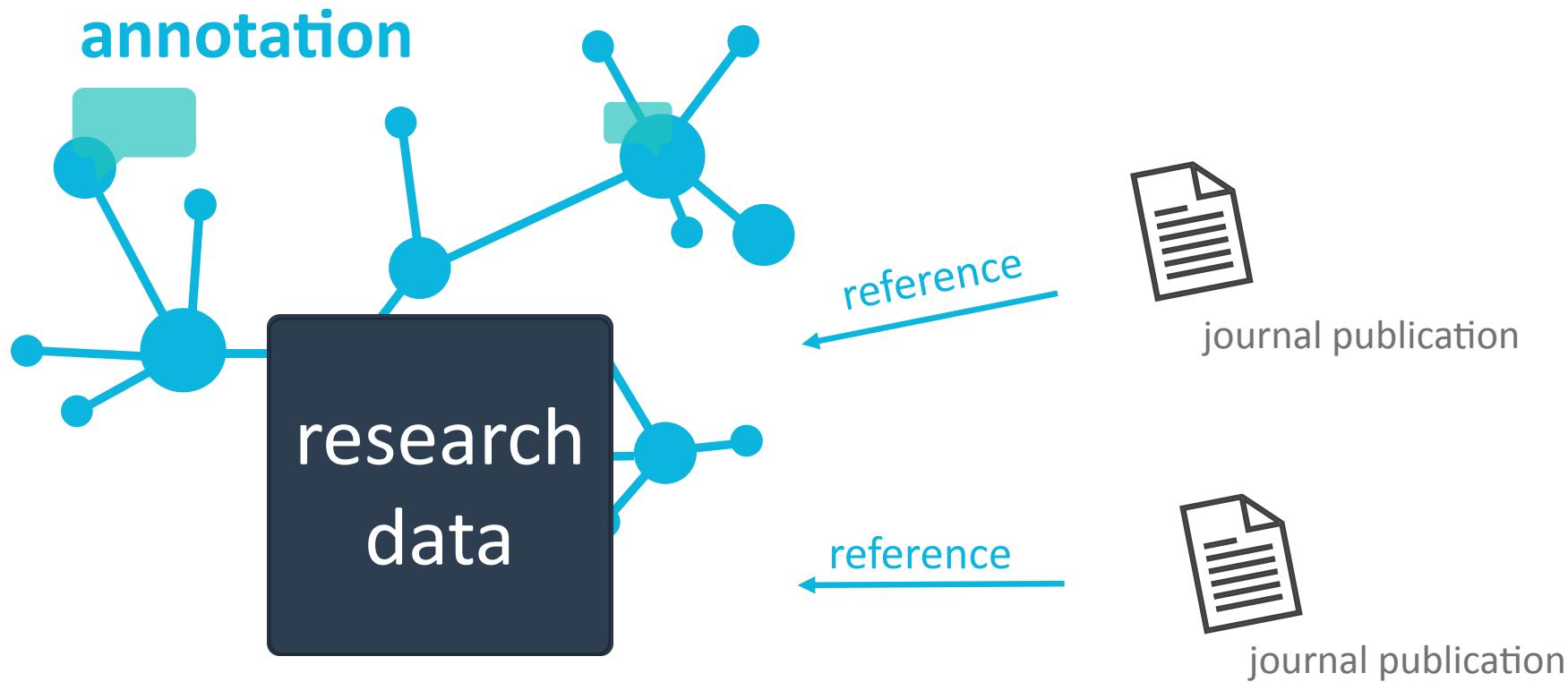
Hands-on part 5: ARC data publication



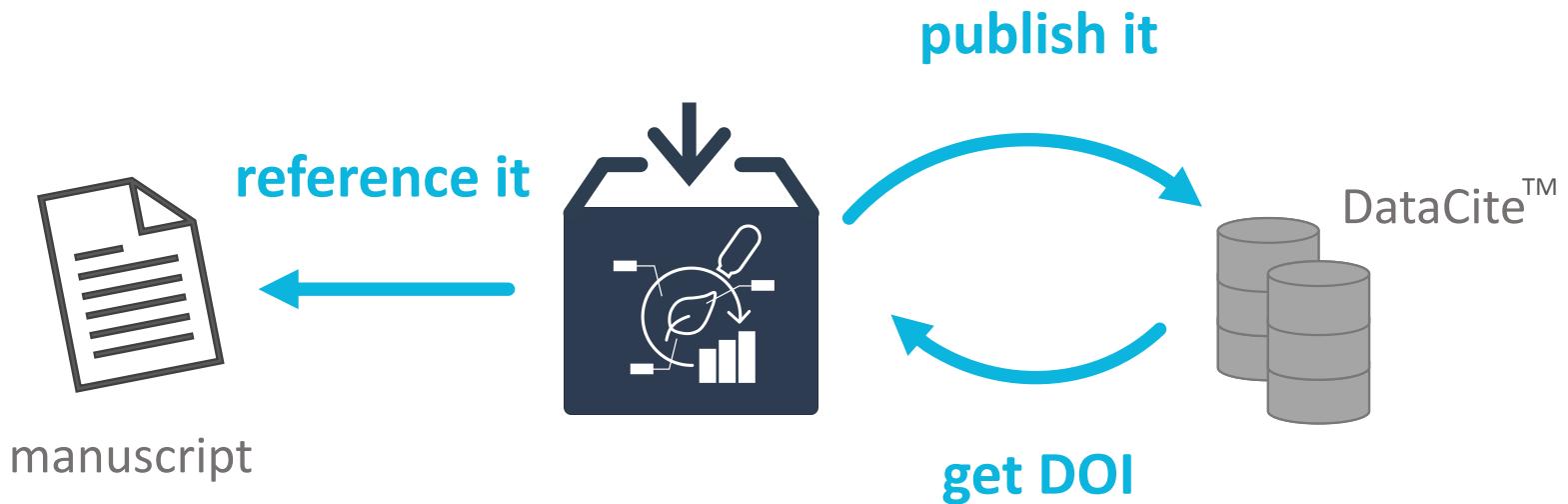
Moving from paper to FAIR data publication



Moving from paper to data publications



Publish your ARC, get a DOI



Publish your ARC with a few clicks

The screenshot shows a bioRxiv project page for 'Ru_ChlamyHeatstress'. The project icon features a green cell with a鞭毛 (flagella) and a red thermometer icon. The project ID is 122. The page displays 53 commits, 1 branch, 0 tags, and 293.9 GB of project storage. Topics include Chlamydomonas, abiotic stress, proteomics, and more. A note states: "Algae cultures were grown mixotrophically (TAP). After 24h of 35°C/40°C the cells were shifted back to room temperature for 48h. 'omics samples were taken." At the bottom, there are three numbered buttons: 1. pipeline (passed), 2. Publish ARC (highlighted in blue), and 3. arc quality (301/301).

Ru_ChlamyHeatstress

Project ID: 122

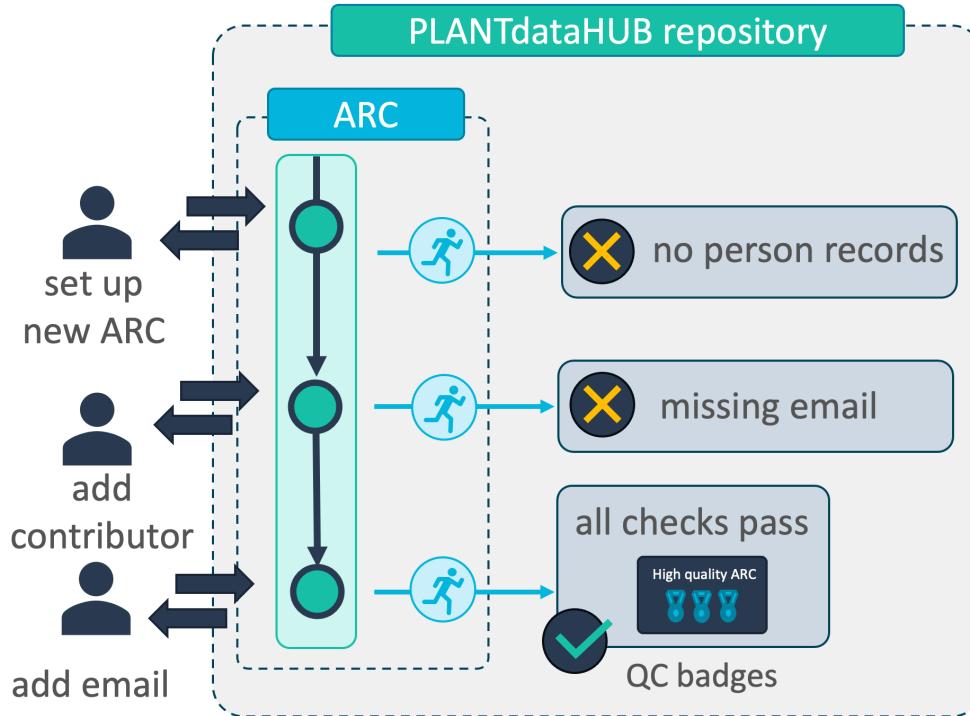
53 Commits 1 Branch 0 Tags 293.9 GB Project Storage

Topics: Chlamydomonas, abiotic stress, proteomics, + 1 more

Algae cultures were grown mixotrophically (TAP). After 24h of 35°C/40°C the cells were shifted back to room temperature for 48h. 'omics samples were taken.

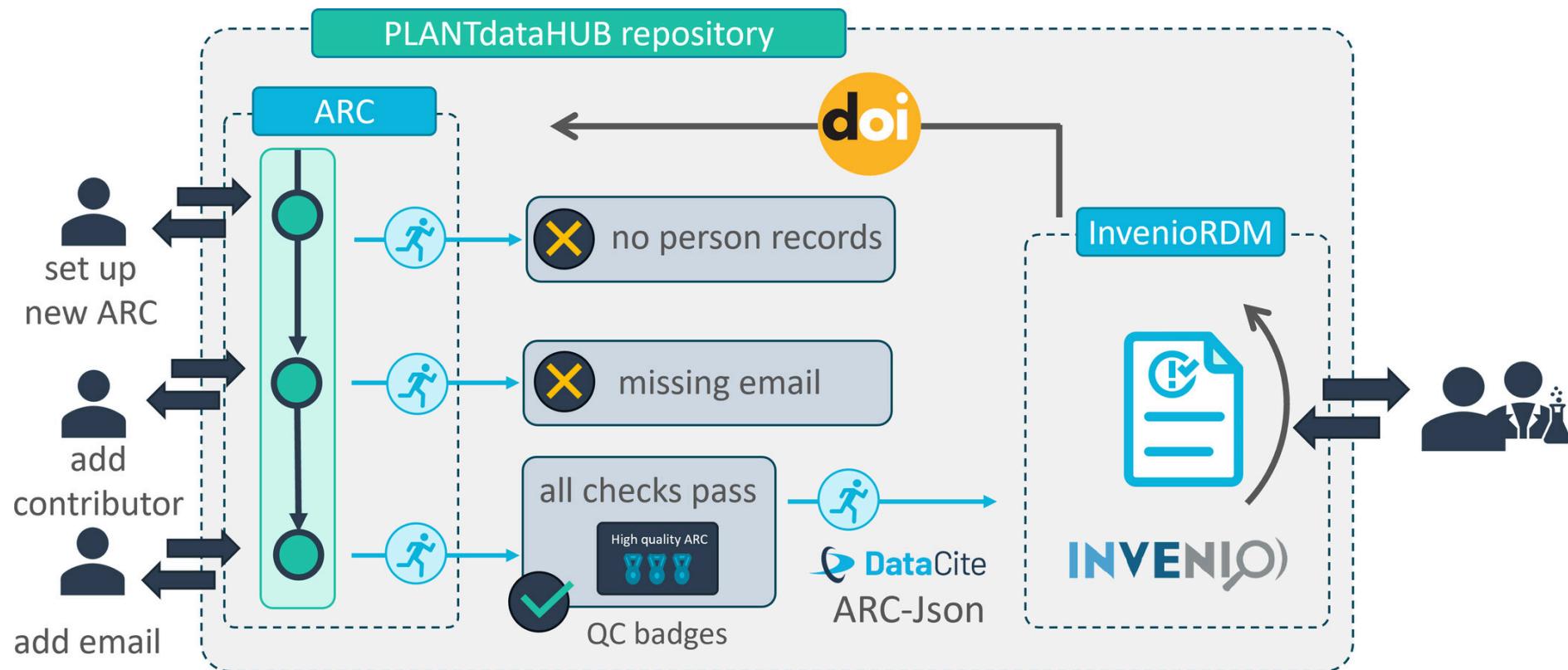
1 pipeline passed 2 Publish ARC 3 arc quality 301/301

Validation



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

Validate & publish



Validation towards publication

Ru_ChlamyHeatstress

main Ru_ChlamyHeatstress History Find file Code

add validation_packages.yml Kevin Schneider authored 2 months ago da71d3c9

Name	Last commit	Last update
.arc	add validation_packages.yml	2 months ago
assays	Add missing data files	1 year ago
publication	add publication, add zScores	2 years ago
runs	add tpm data	1 year ago
studies	add publication information	1 year ago
workflows	Update	2 years ago
.gitattributes	rename missing samples; #2	1 year ago
.gitignore	add gitignore	1 year ago
README.md	Add doi badge	1 year ago
isa.investigation.xlsx	add author emails and adapt title	1 year ago

README.md

Chlamydomonas reinhardtii heat stress time course experiment

DOI 10.60534/9e5jx-75d83

Star 2 ·

Project information

Algae cultures were grown mixotrophically (TAP). After 24h of 35°C/40°C the cells were shifted back to room temperature for 48h. 'omics samples were taken.

Chlamydomonas abiotic stress
proteomics + 1 more

pipeline passed invenio 10/10

55 Commits 2 Branches 0 Tags

README Auto DevOps enabled

Created on July 11, 2022

Receive a DOI

Published September 7, 2023 | Version v1

Dataset 

Systems-wide investigation of responses to moderate and acute high temperatures in the green alga Chlamydomonas reinhardtii.

Zhang, Ningning¹; Mattoon, Erin¹; McHargue, Will¹ ; Venn, Benedict² ; Zimmer, David² ; Pecani, Kresti³; Jeong, Jooyeon¹; Anderson, Cheyenne¹; Chen, Chen⁴; Berry, Jeffrey¹; Xia, Ming¹; Tzeng, Shin-Cheng¹ ; Becker, Eric¹; Pazouki, Leila¹; Evans, Bradley¹; Cross, Fred³; Cheng, Jianlin⁴; Czymmek, Kirk¹ ; Schröda, Michael⁵ ; Mühlhaus, Timo² ; Zhang, Ru¹ 

Show affiliations

Style

APA

1
Citation

Zhang, N., Mattoon, E., McHargue, W., Venn, B., Zimmer, D., Pecani, K., Jeong, J., Anderson, C., Chen, C., Berry, J., Xia, M., Tzeng, S.-C., Becker, E., Pazouki, L., Evans, B., Cross, F., Cheng, J., Czymmek, K., Schröda, M., ... Zhang, R. (2023). Systems-wide investigation of responses to moderate and acute high temperatures in the green alga Chlamydomonas reinhardtii. [Data set]. DataPLANT. <https://doi.org/10.60534/9e5jx-75d83>

Description

hosted on: <https://git.nfd4plants.org/projects/122>

Files

arc-summary.md

[Data set] Systems-wide investigation of responses to moderate and acute high temperatures in the green alga Chlamydomonas reinhardtii.

File contents:

- root
 - isa.investigation.xlsx
 - README.md
 - runs

2
Edit
New version
Share

Versions

Version v1 Sep 7, 2023
10.60534/9e5jx-75d83

Details

DOI
[DOI 10.60534/9e5jx-75d83](https://doi.org/10.60534/9e5jx-75d83)

Resource type
Dataset

Publisher
DataPLANT

Export

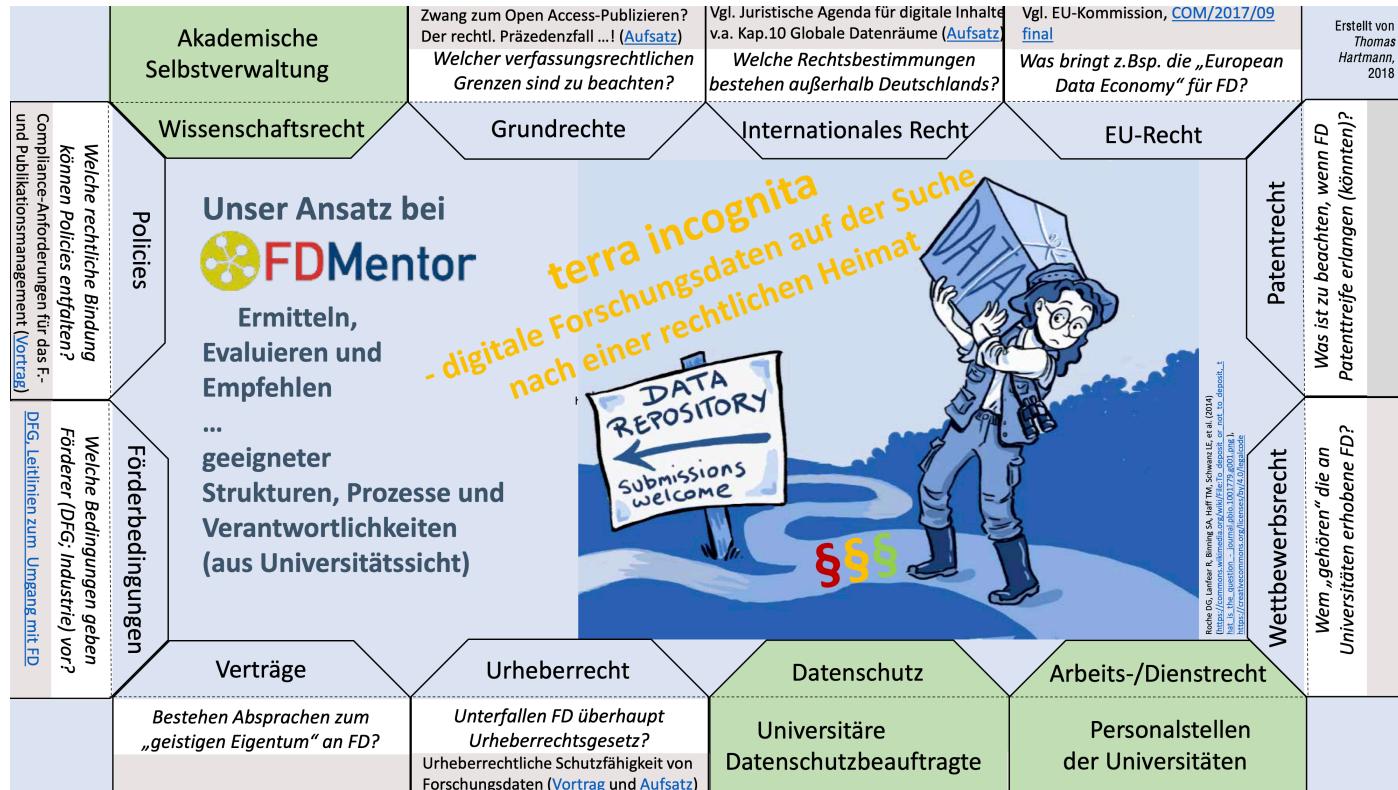
JSON Export

Hands-on part 5: Data publication

1. ARC validation
2. (Towards) ARC publication

Legal aspects of RDM

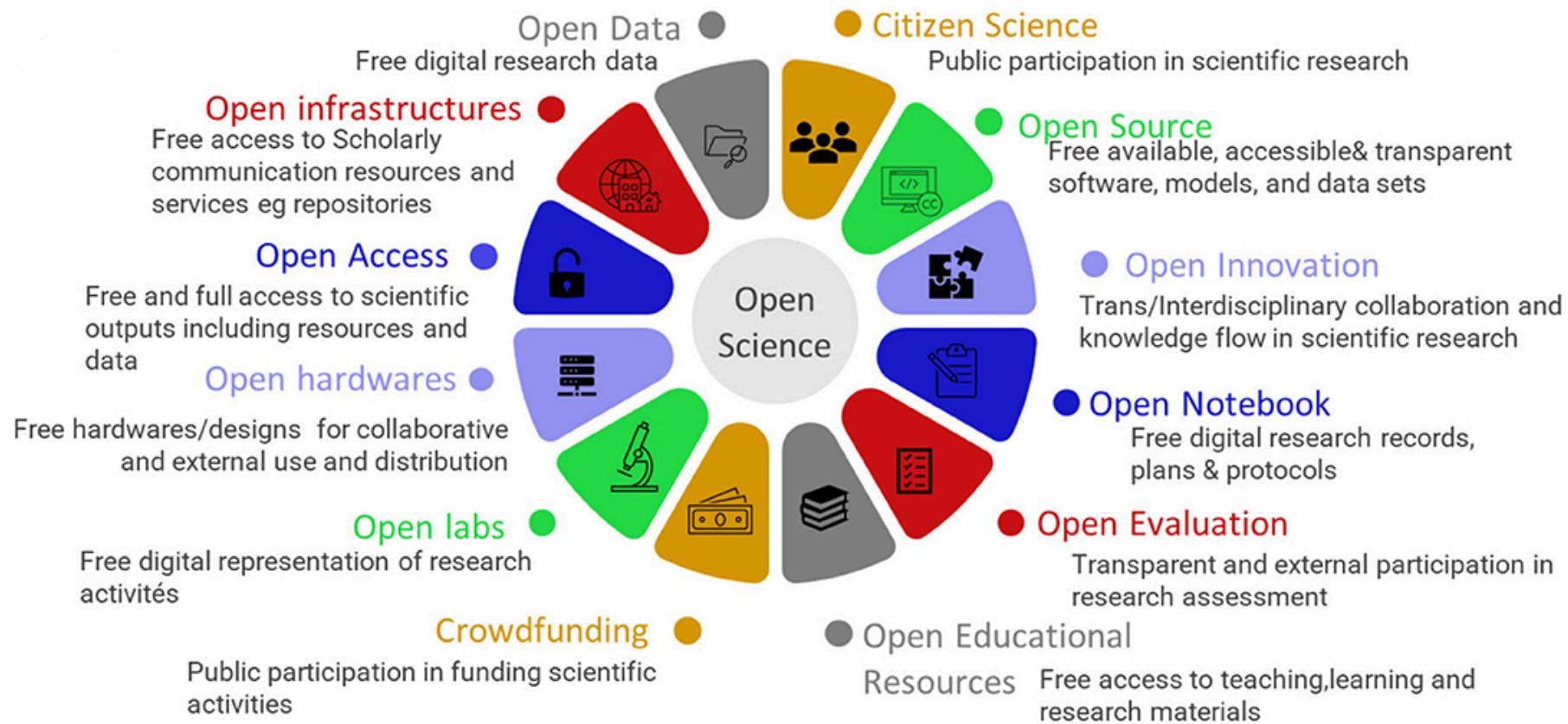
Different laws touched by RDM



Open Access (OA) categories

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

Open Science is more than Open Access



Creative commons

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



Attribution
CC BY



Attribution – ShareAlike
CC BY-SA



Attribution – NoDerivs
CC BY-ND



Attribution – NonCommercial
CC BY-NC



Attribution – NonCommercial – ShareAlike
CC BY-NC-SA



Attribution – NonCommercial – NoDerivs
CC BY-NC-ND

Data protection

GDPR: General Data Protection Regulation

DS-GVO (german): Datenschutz-Grundverordnung

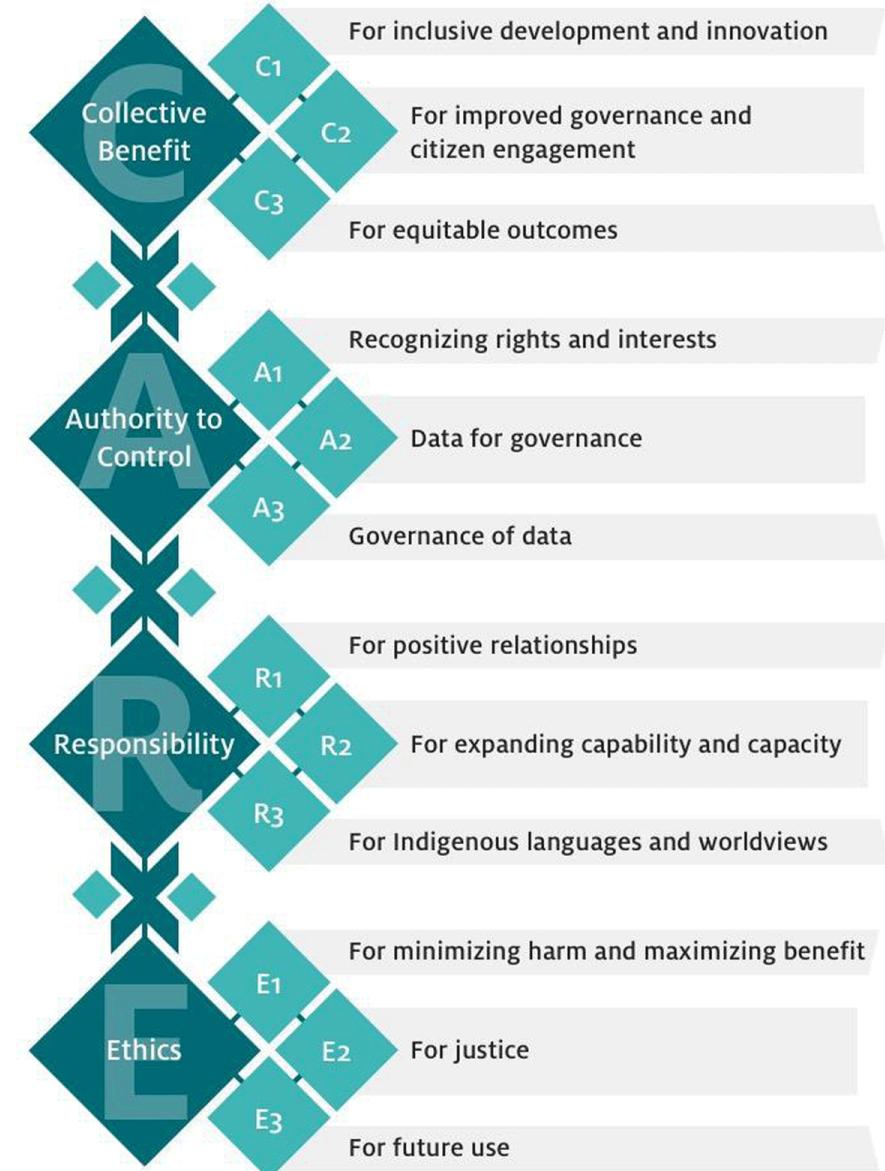
Use of biological materials

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

FAIR and CARE



CARE principles



Research Data policies



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

CEPLAS relevant data handling guidelines & policies

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

The Data Management Plan (DMP)

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

DMP tools

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

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



Share your ARC

Invite your colleague to your ARC



Understand your colleague's ARC

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

Assignment

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

Contributors

If not referenced otherwise, figures and slides presented here were created by members of DataPLANT (<https://nfdi4plants.org>).

Additional slides were contributed 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: Hajira Jabeen
github: <https://github.com/HajiraJabeen>
orcid: <https://orcid.org/0000-0003-1476-2121>
- name: Kevin Frey
github: <https://github.com/Freymaurer>
orcid: <https://orcid.org/0000-0002-8493-1077>
- name: Sabrina Zander
orcid: <https://orcid.org/0009-0000-4569-6126>

