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

Welcome

House-keeping

Pad: https://pad.hhu.de/0NdPKO5LQ5CHBRN2iuG91Q

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

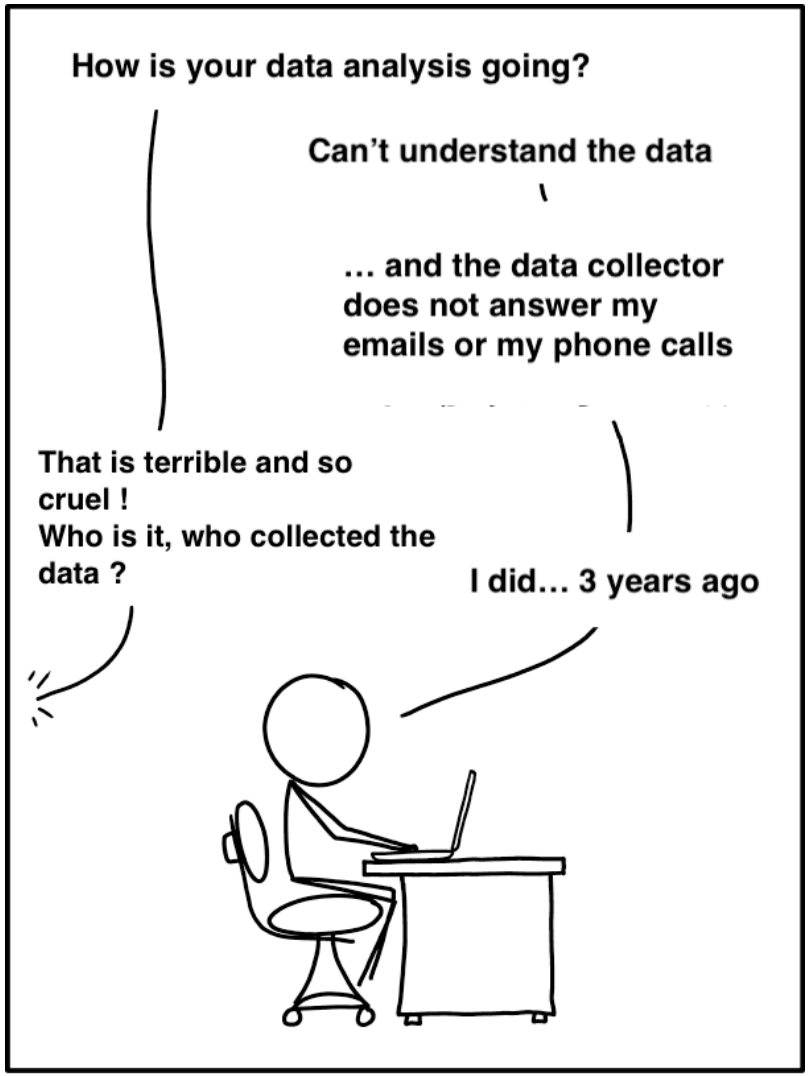
Introduce yourself

- 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)?

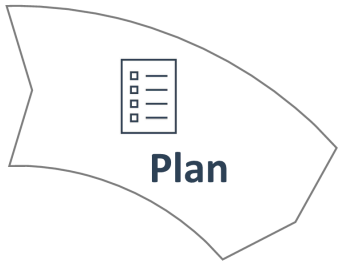
- Increase transparency
- Make data accessible
- Save time (writing, reusing)
- Reduce the risk of data loss
- Optimize the costs
- Facilitate future reuse and sharing
- Improve citations



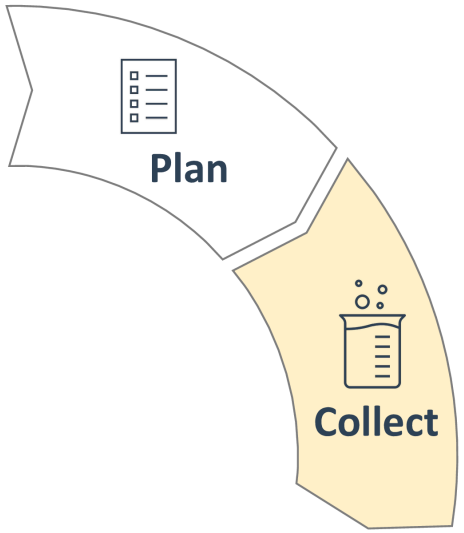
Your first collaborators are your future selves, be nice to them !

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

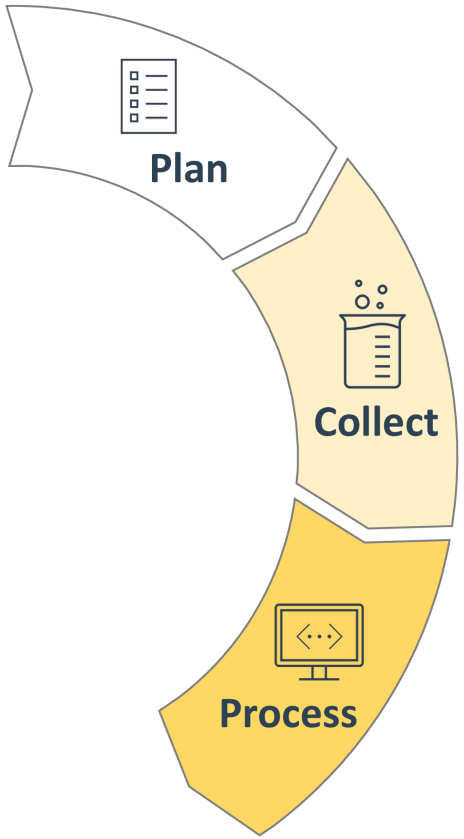
The Research Data Lifecycle



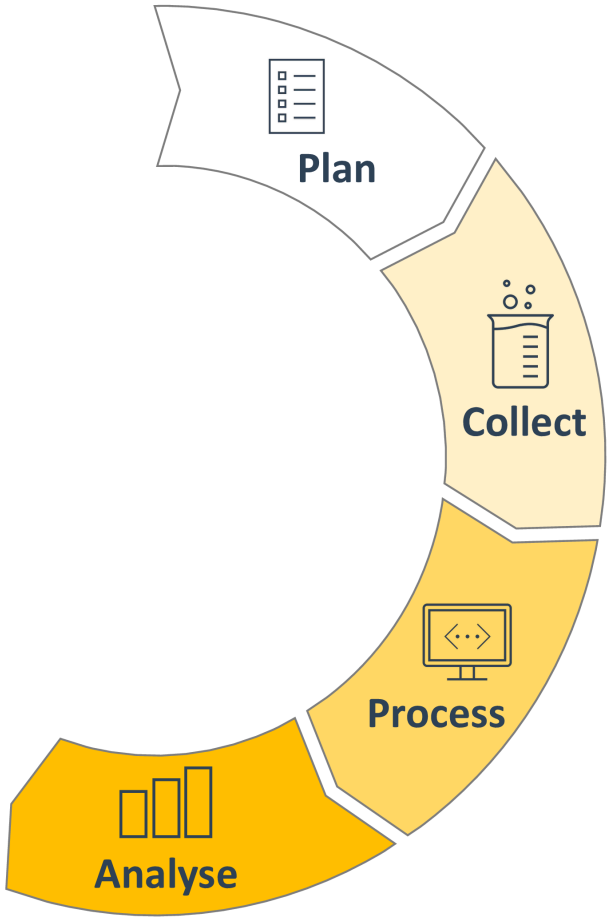
The Research Data Lifecycle



The Research Data Lifecycle



The Research Data Lifecycle

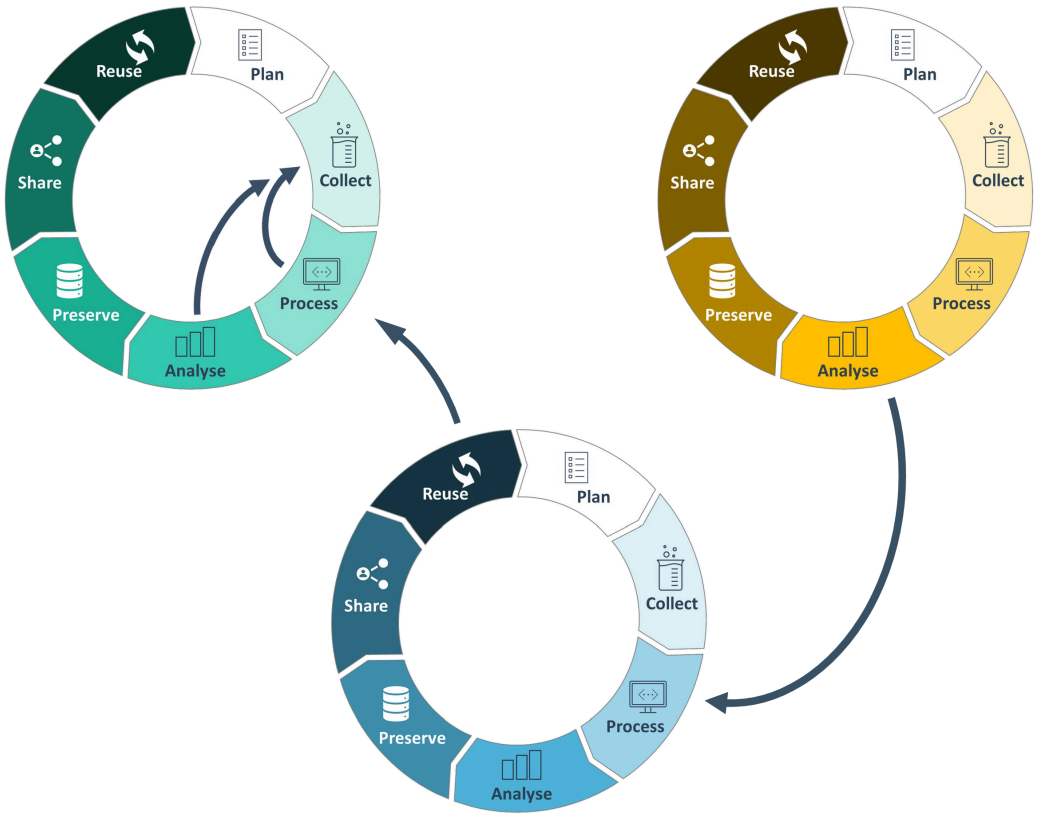


The Research Data Lifecycle

The Research Data Lifecycle

The Research Data Lifecycle

The Research Data Lifecycle *is mutable*



FAIR

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

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

nature > scientific data > comment > article

Open Access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

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

— Show fewer authors

Scientific Data **3**, Article number: 160018 (2016) | Cite this article

The FAIR principles

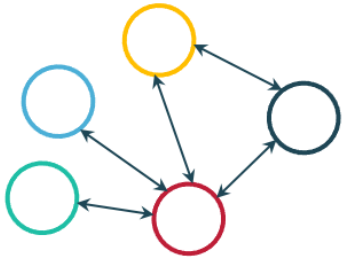
Findable

Accessible

Interoperable

Reusable

Easier collaboration & sharing



Increased findability and visibility

Reproducibility

Added-value to the research community

nfdi

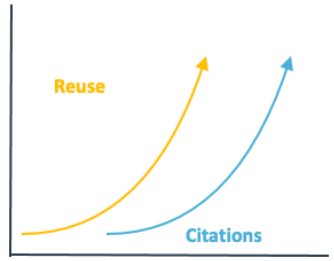
NCBI

EMBL-EBI

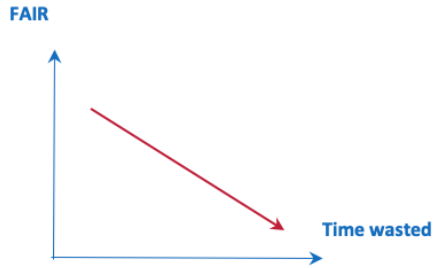
Compliance with funding policies

DFG

Receive due credit



Saves time & workload



Is your data FAIR?

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

- 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.

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

Accessible

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

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

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

Interoperable

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

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

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

Reusable

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

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

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

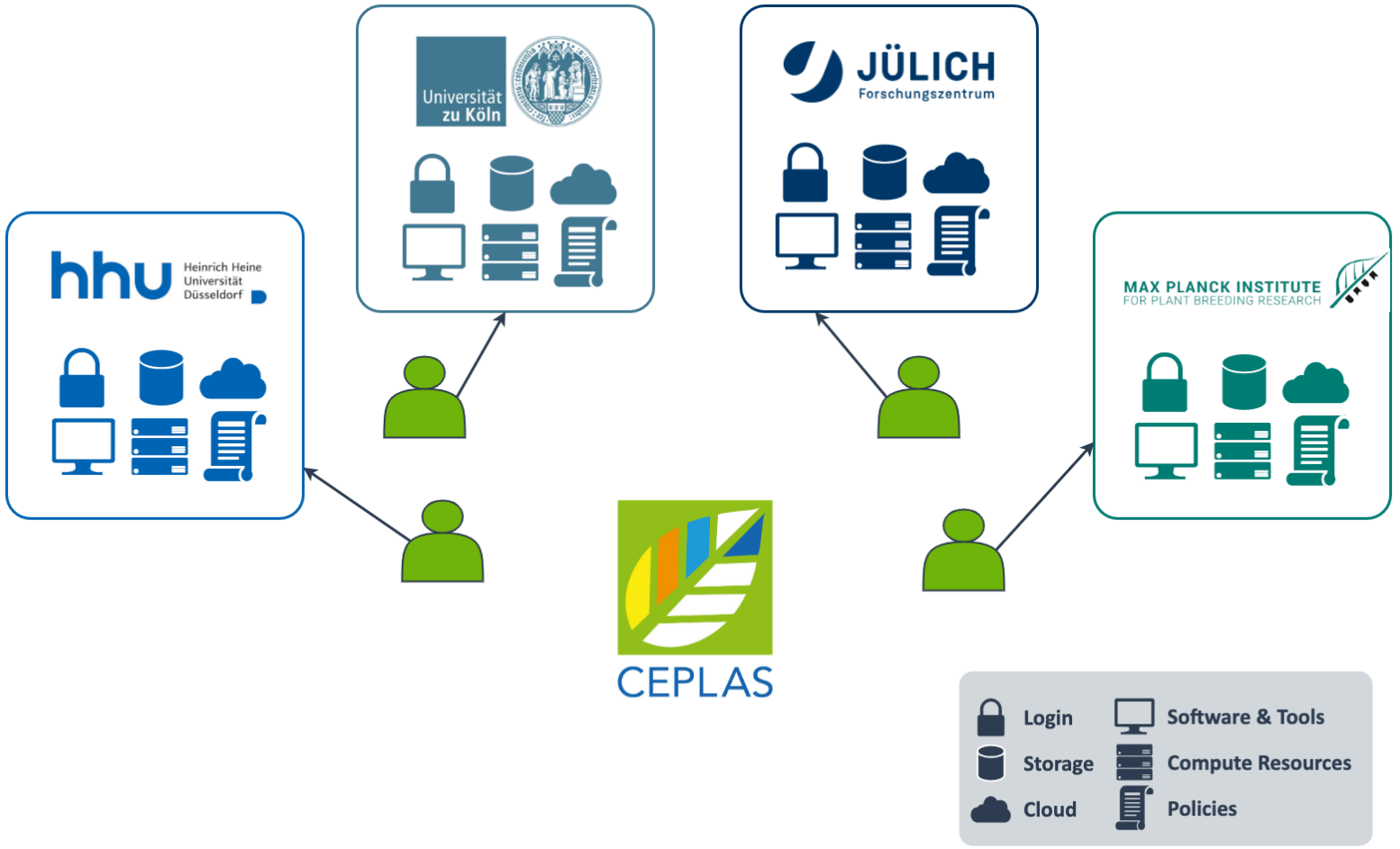
FAIR on multiple layers

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

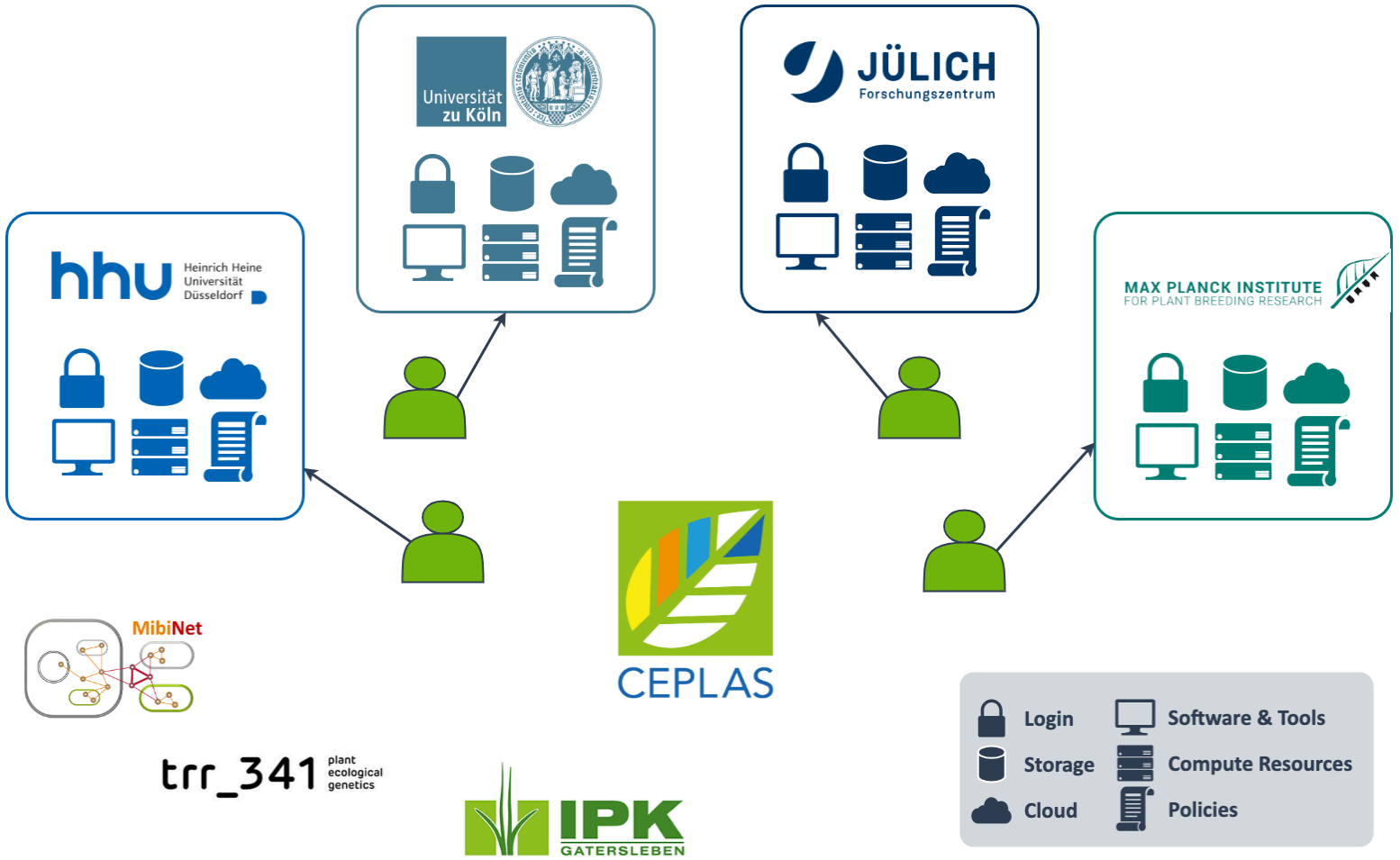
https://www.go-fair.org/fair-principles/

CEPLAS – One cluster, four locations

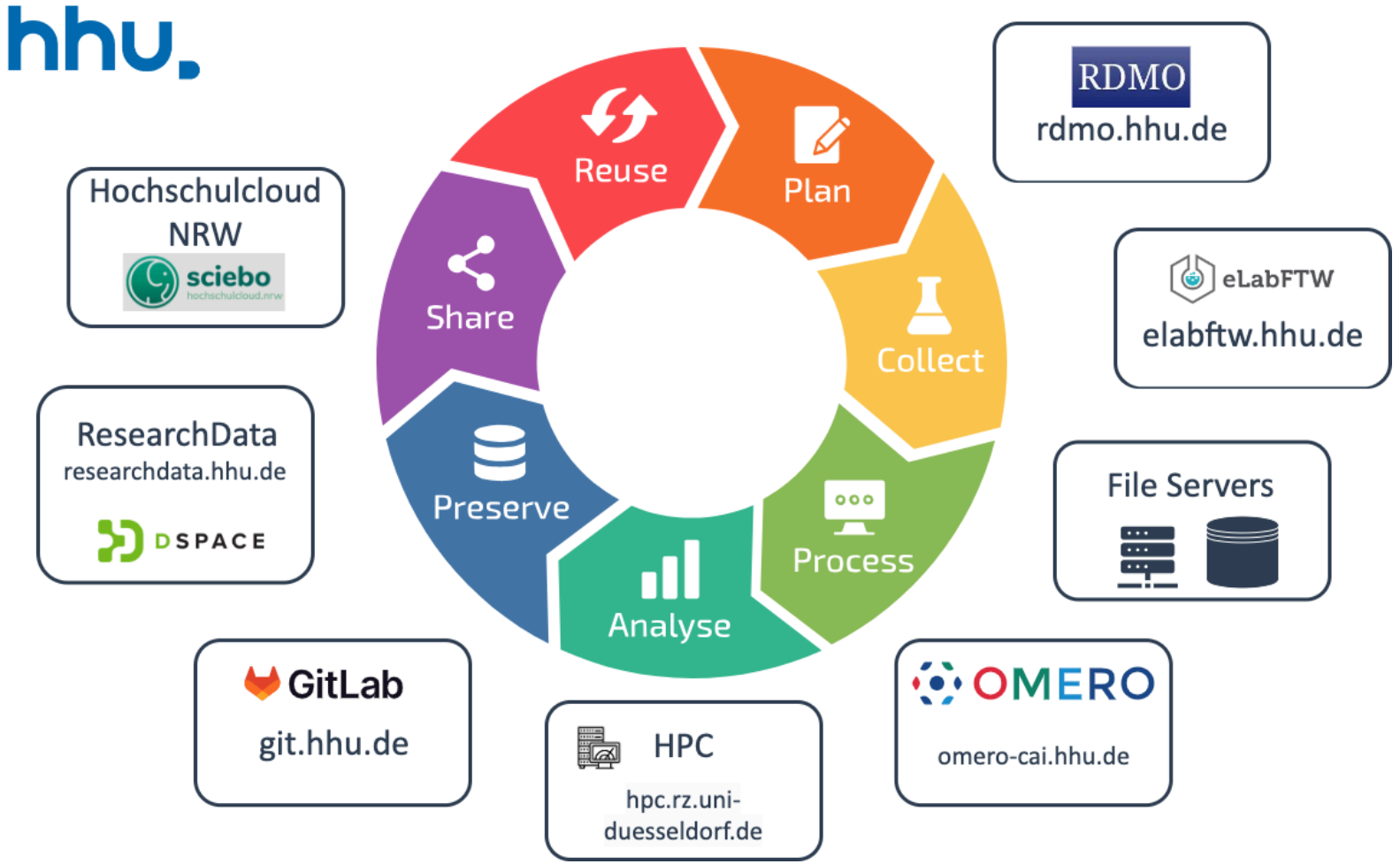
Data silos impede collaboration



Data silos impede collaboration



Missing interfaces impede collaboration



📝 Understand your colleague's project 📝

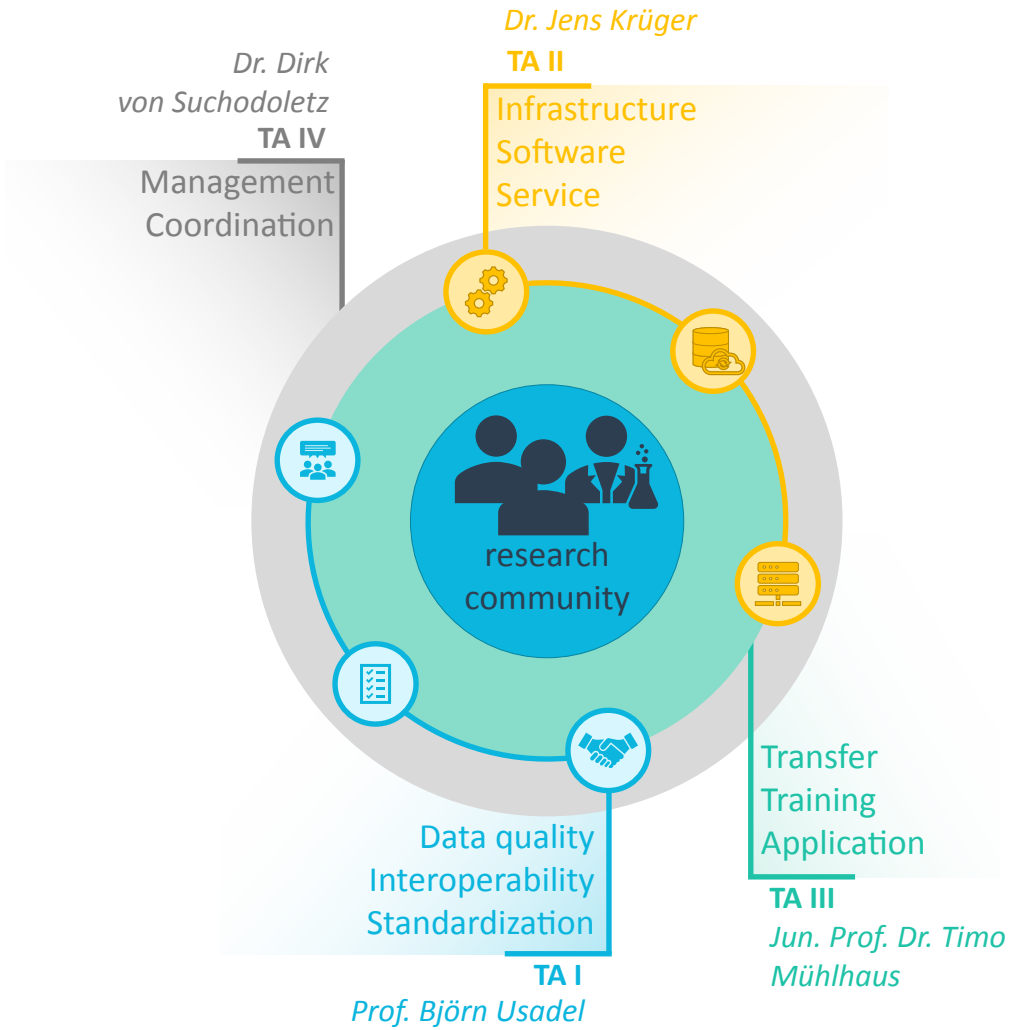
1. Go to the shared folder prepared before class
2. Try to identify one experiment that led to results (e.g. a figure in the thesis)
3. What are the samples (e.g. plants, bacteria)?
4. How were the samples prepared (~ materials)?
5. How was the experiment performed (~ methods)?
6. What is the raw data (~ 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

DataPLANT – The NFDI4Plants

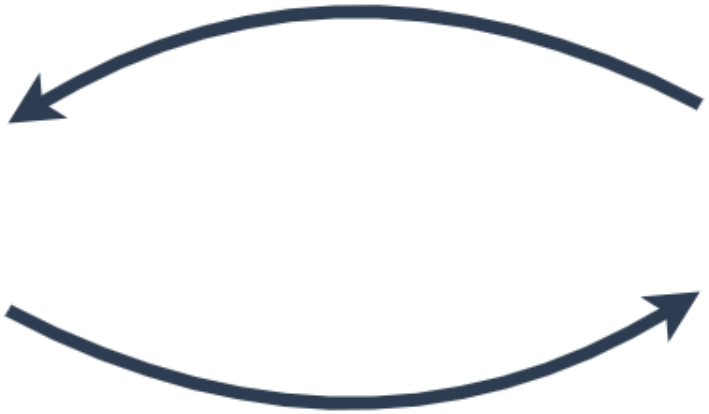
- NFDI: "Nationale Forschungsdaten Infrastruktur" – www.nfdi.de
- Funded since end of 2020



Data Stewardship between DataPLANT and the community

Community

nfdi4plants

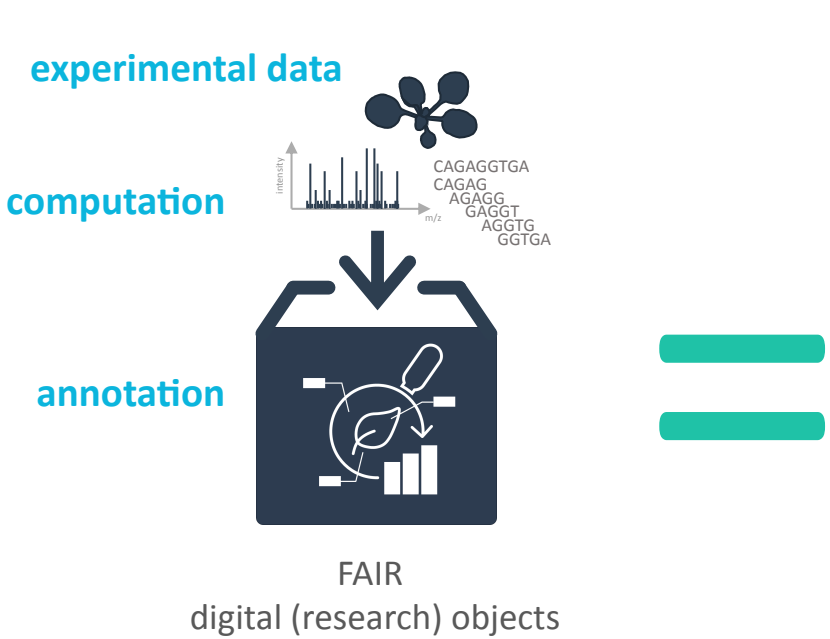


CEPLAS

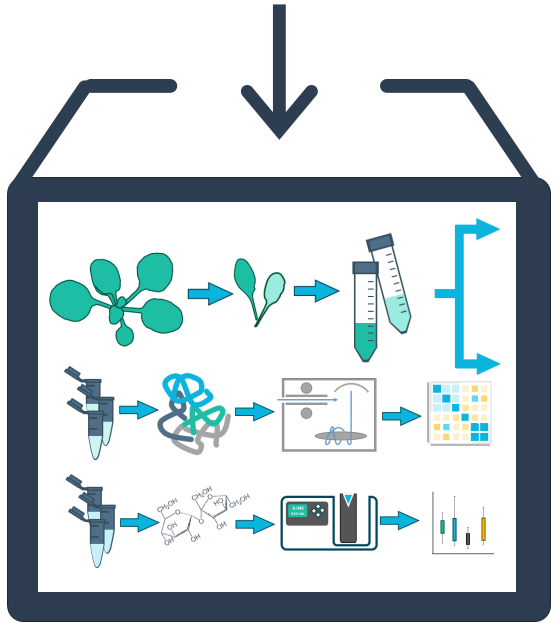
Domain experts
User experience
Training

Service provider
Developers
Tech experts

Annotated Research Context (ARC)



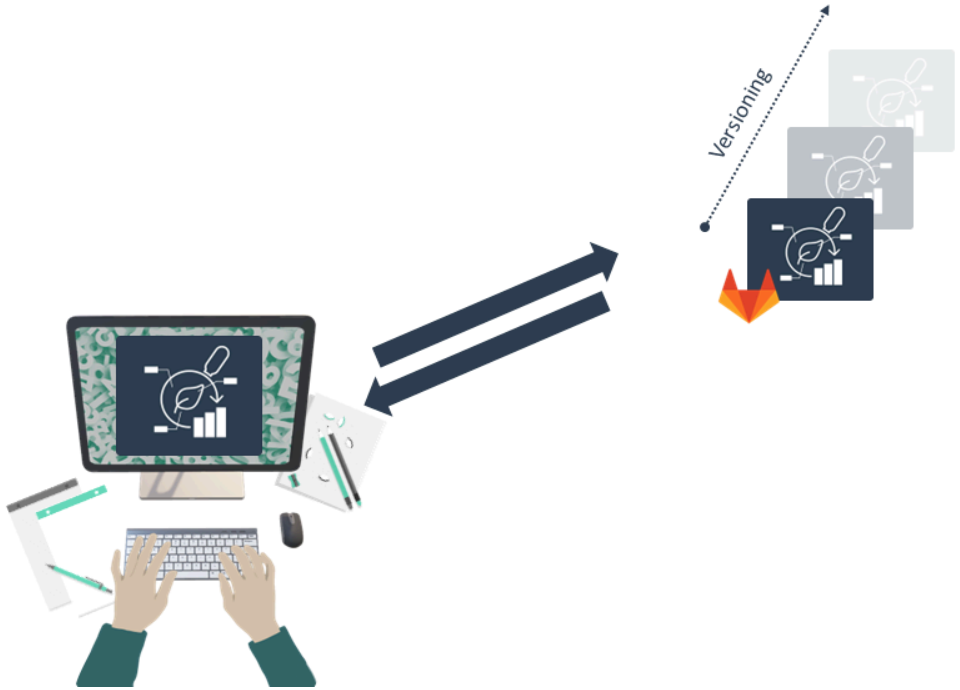
FAIR
digital (research) objects



Your entire investigation in a single unified bag

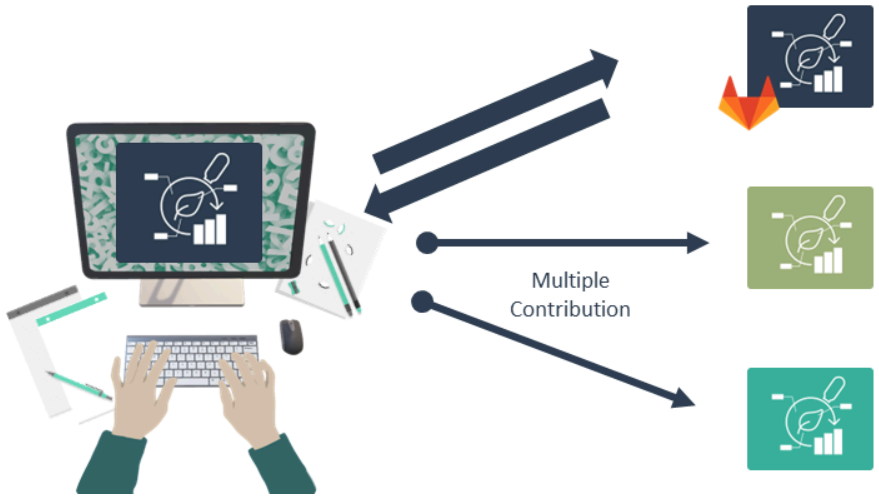
You can store your ARC in the DataHUB

ARCs are versioned

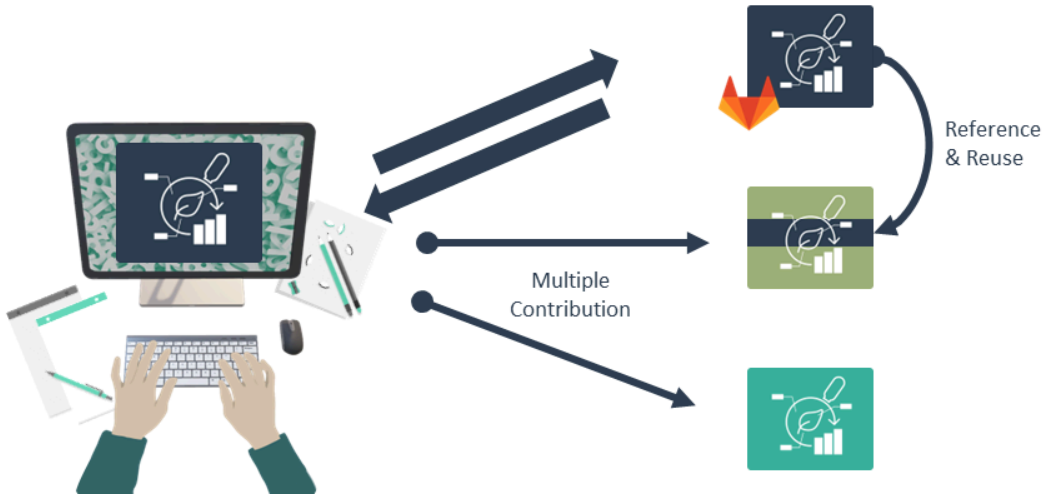


You can invite collaborators

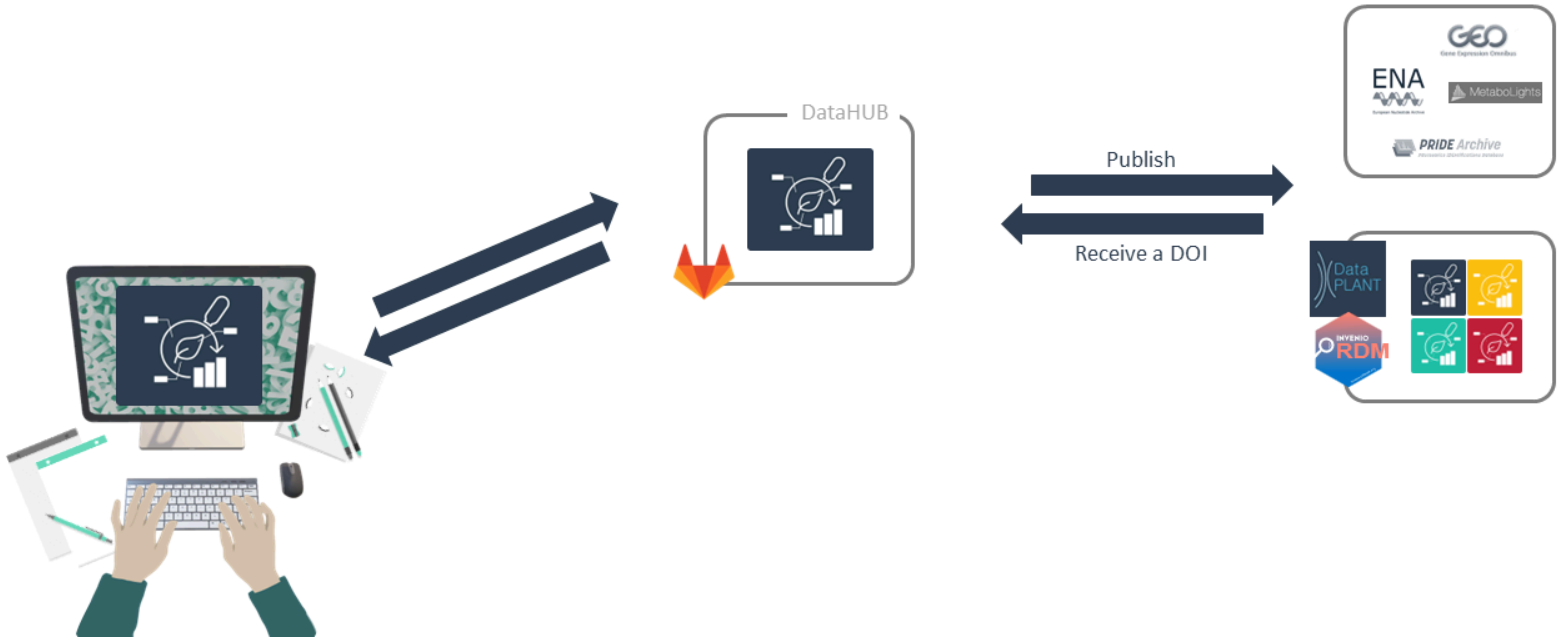
Collaborate and contribute



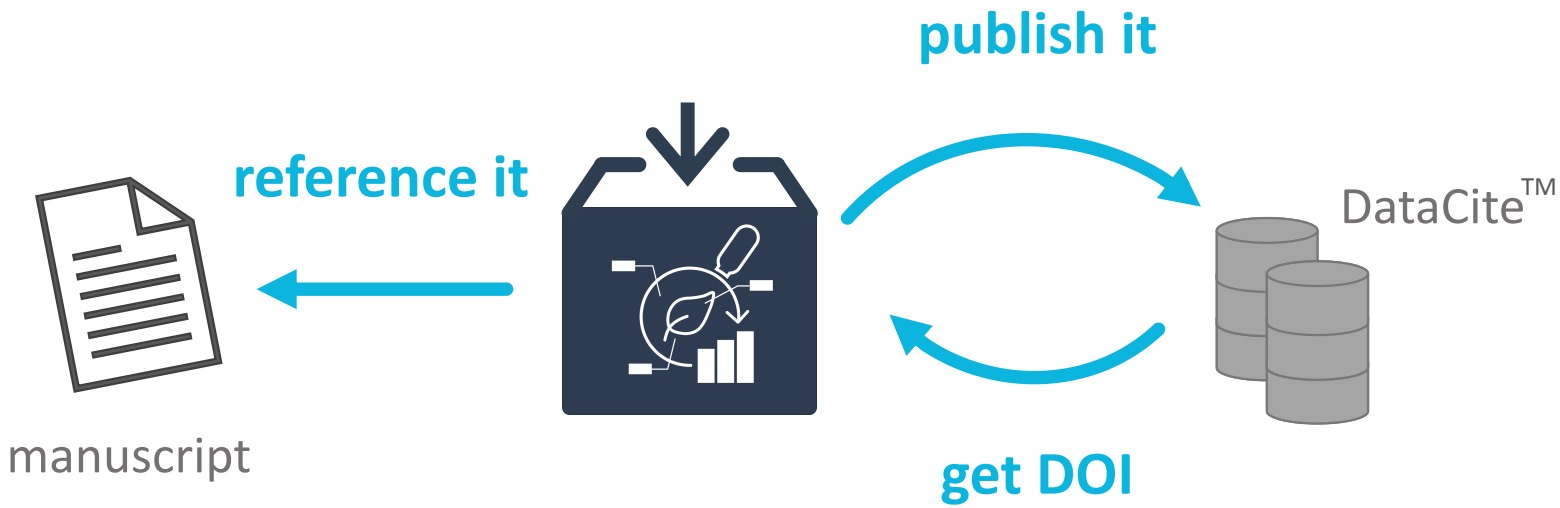
Reuse data in ARCs



Publish your ARC

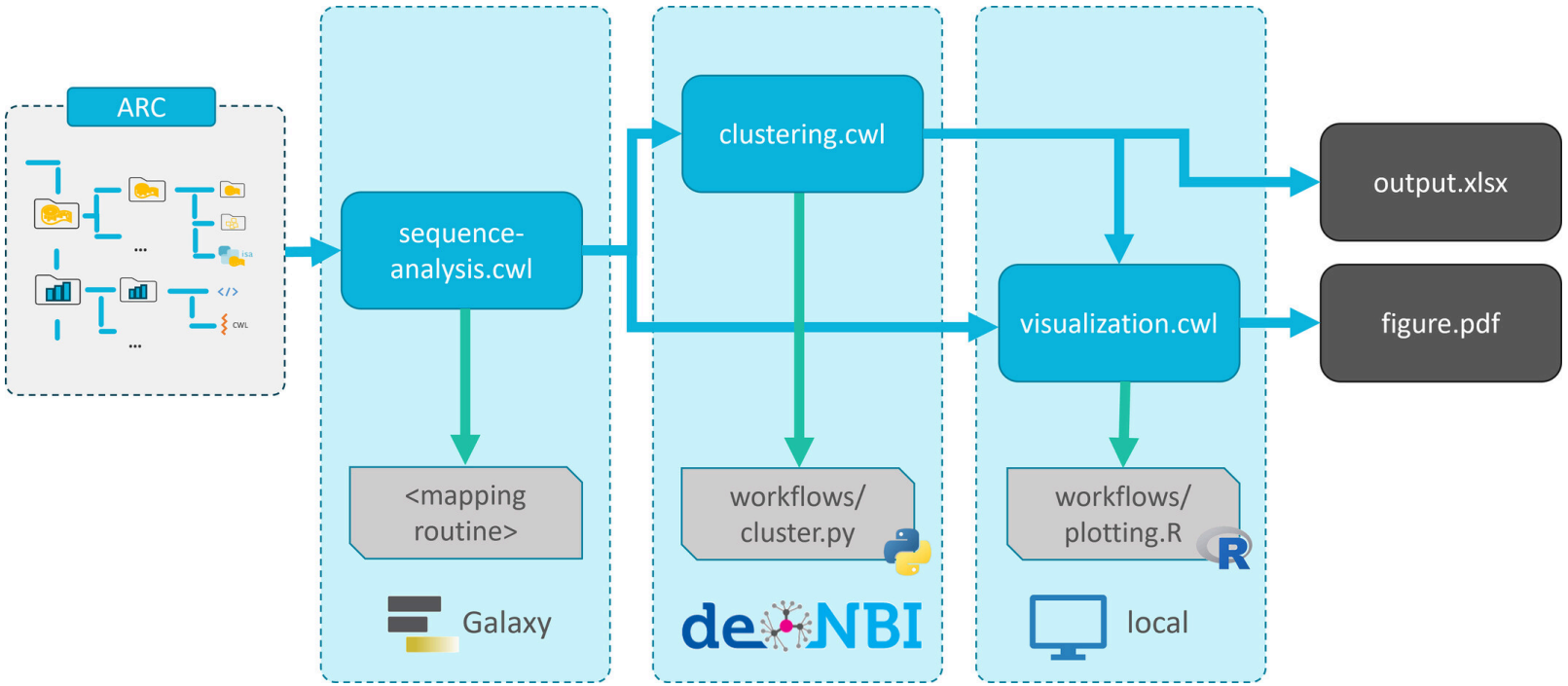


Publish your ARC, get a DOI



ARC as single-entry point

Data analysis and workflows



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

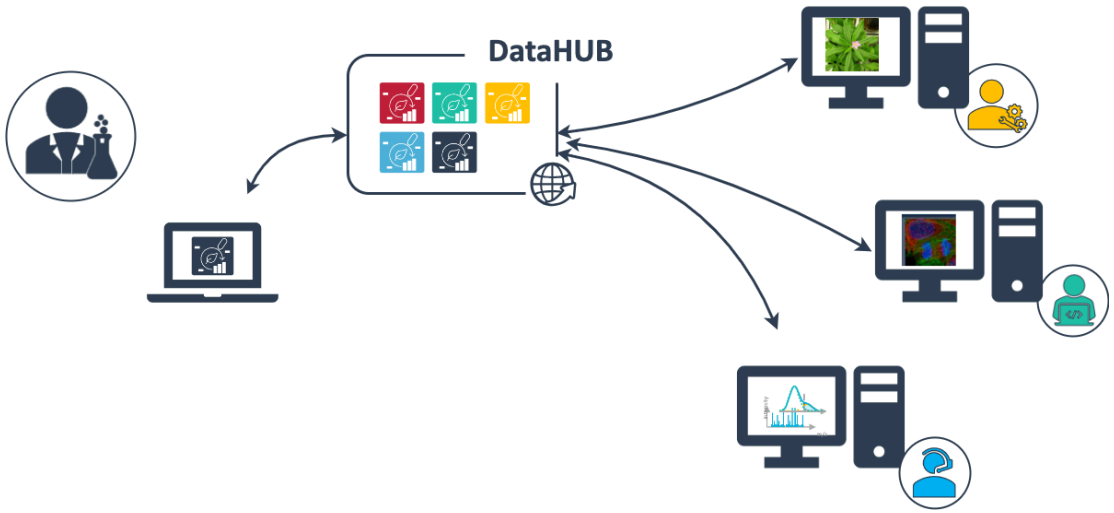
Galaxy integration: Extra value for plant research

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

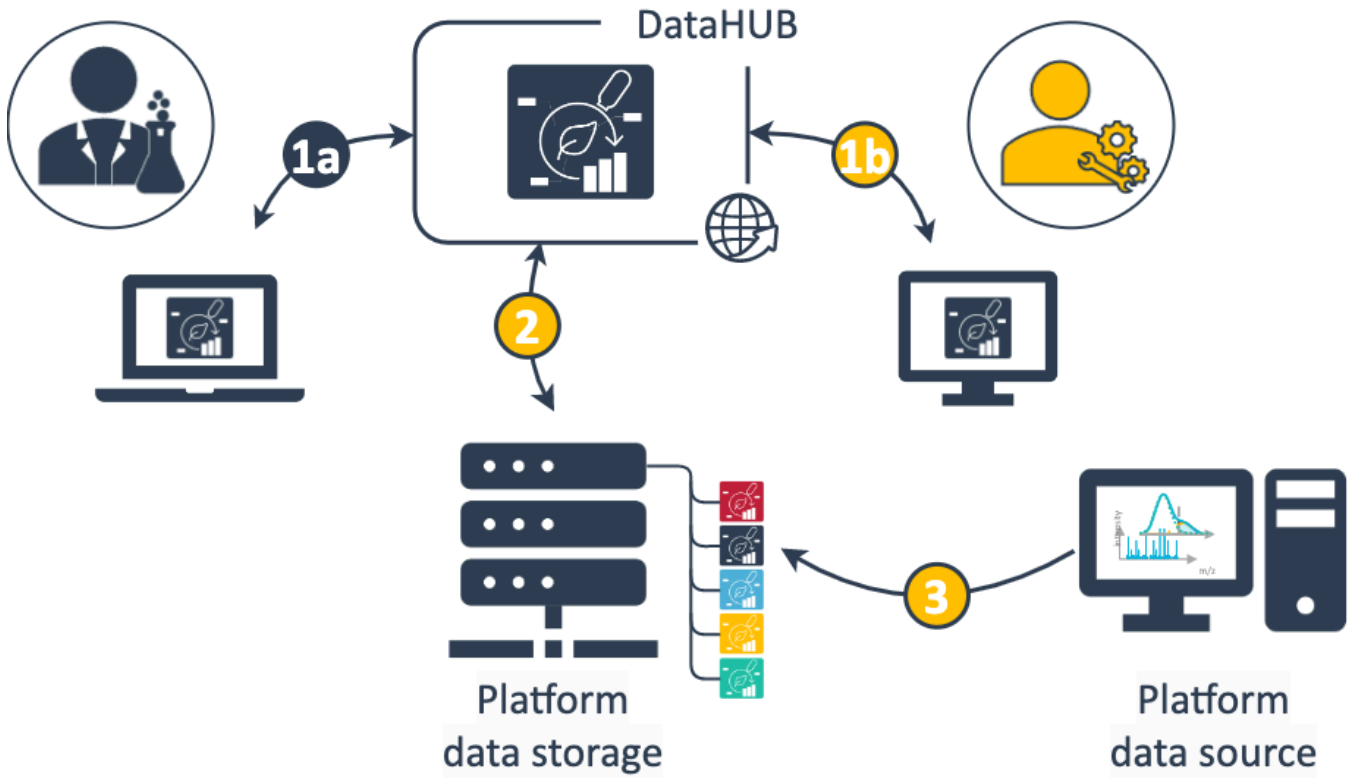
- Full ARC compatibility
- Automated metadata generation
- Specialized tools and workflows for ‘omics processing and analysis
- Public repository compatibility
- Galaxy teaching resource for data analysis

Enabling platforms

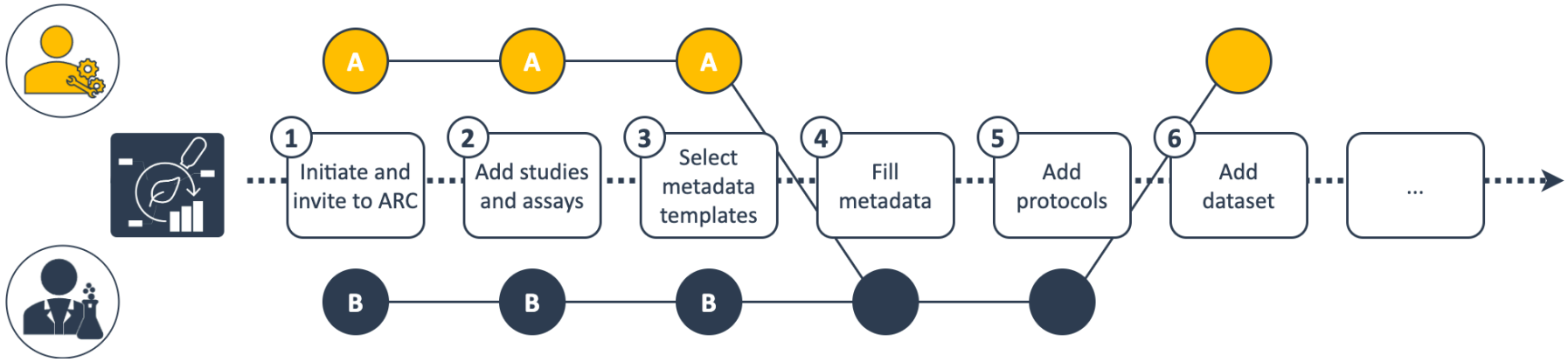
- Streamlined exchange of (meta)data
- Communication and project management



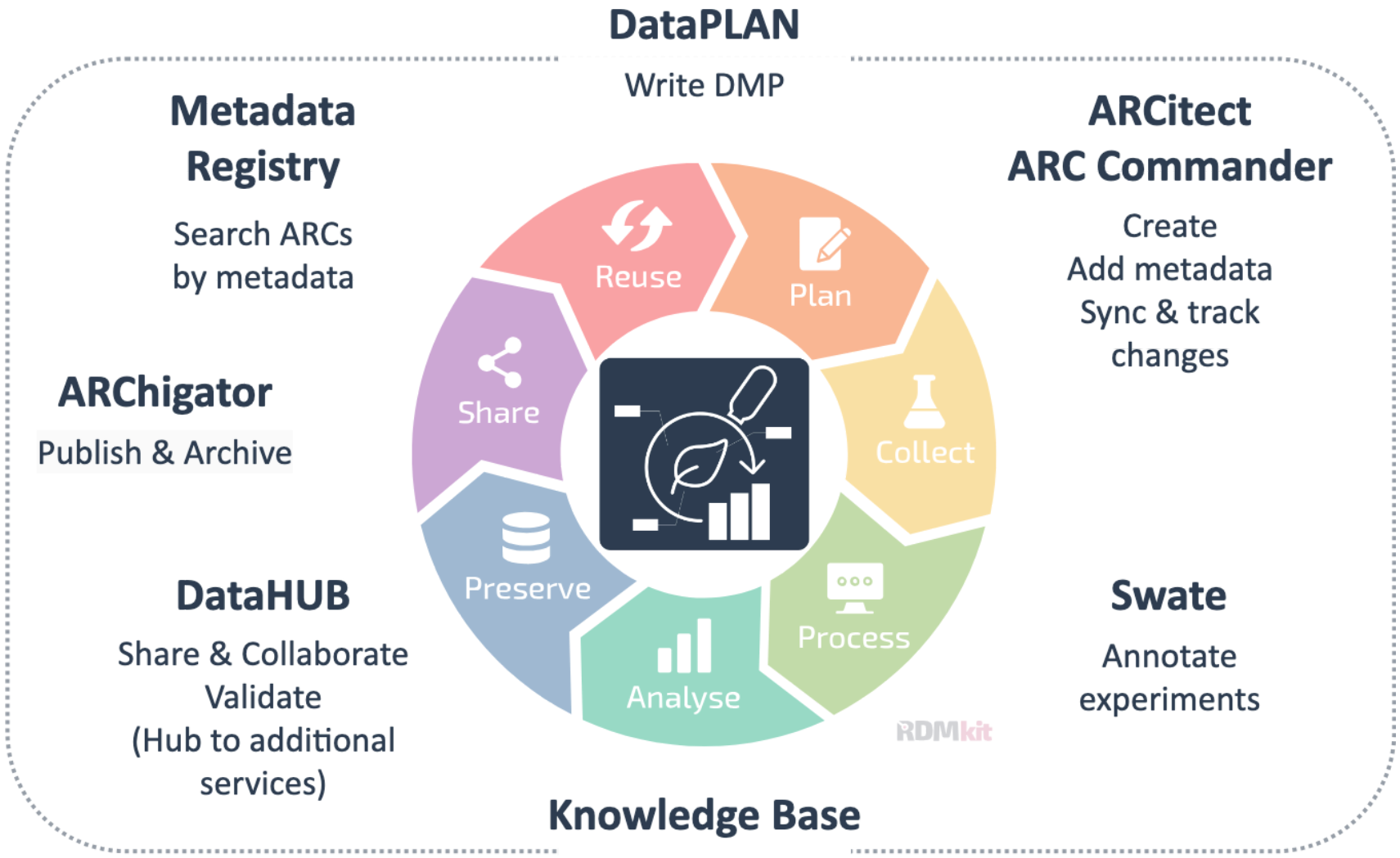
Streamlined data exchange



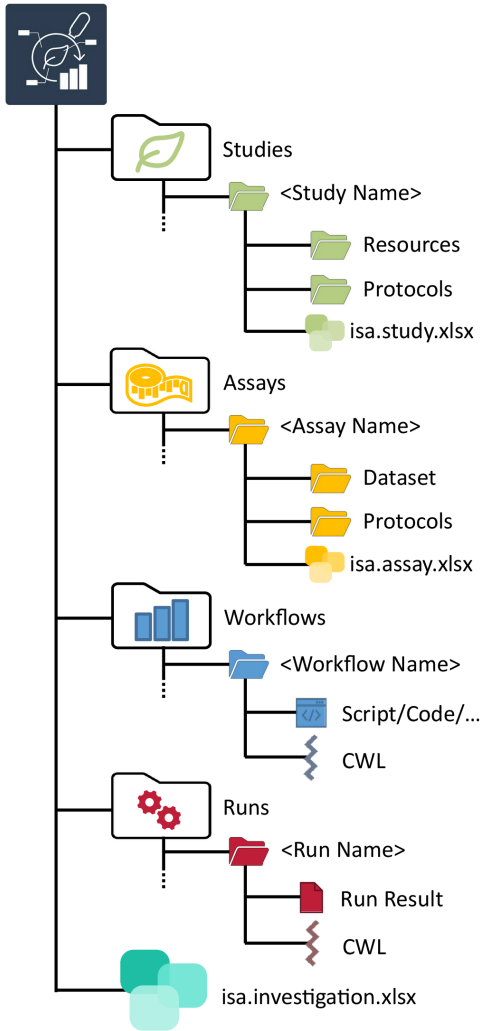
Meet your collaborators in an ARC



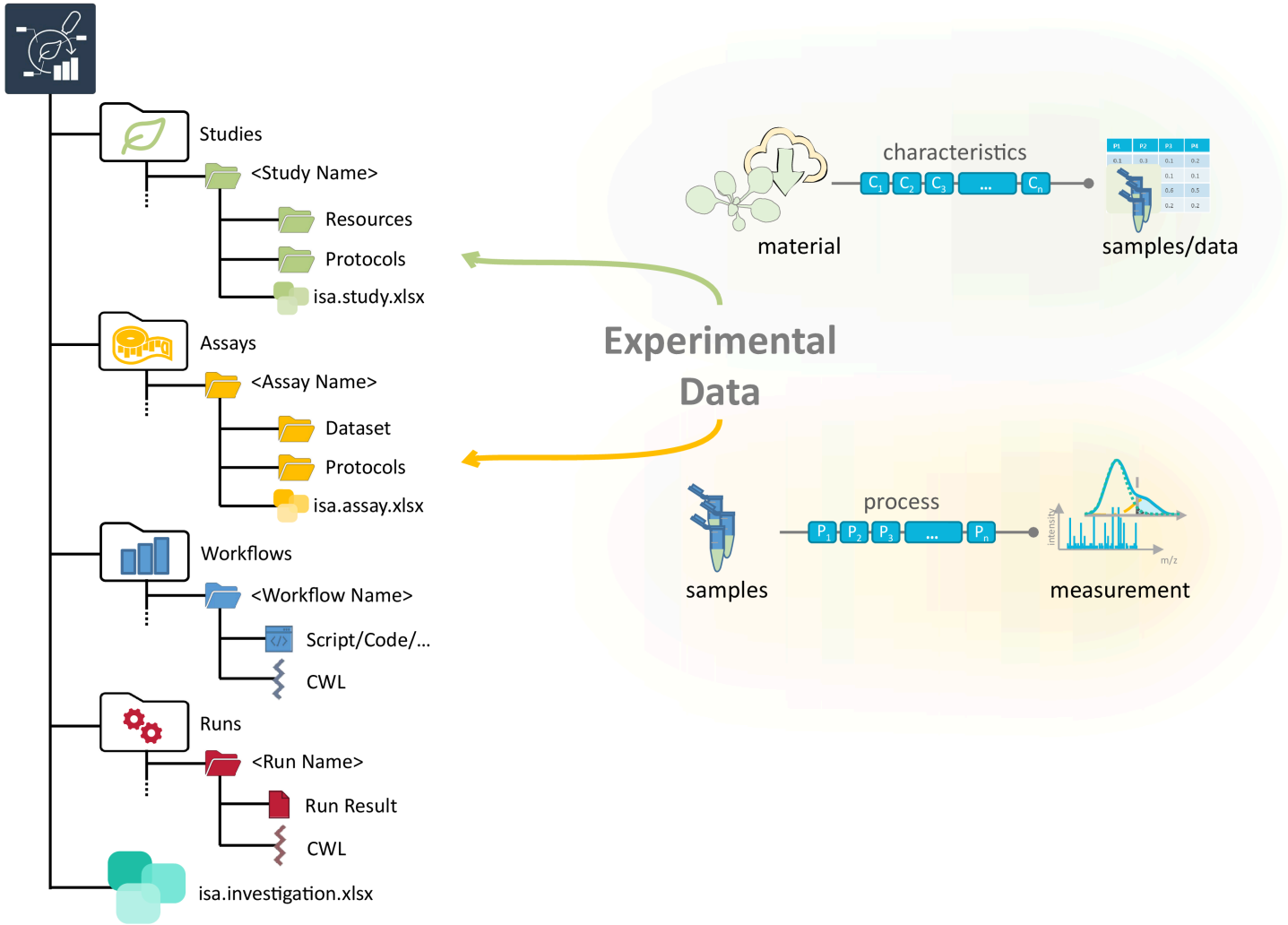
The ARC ecosystem



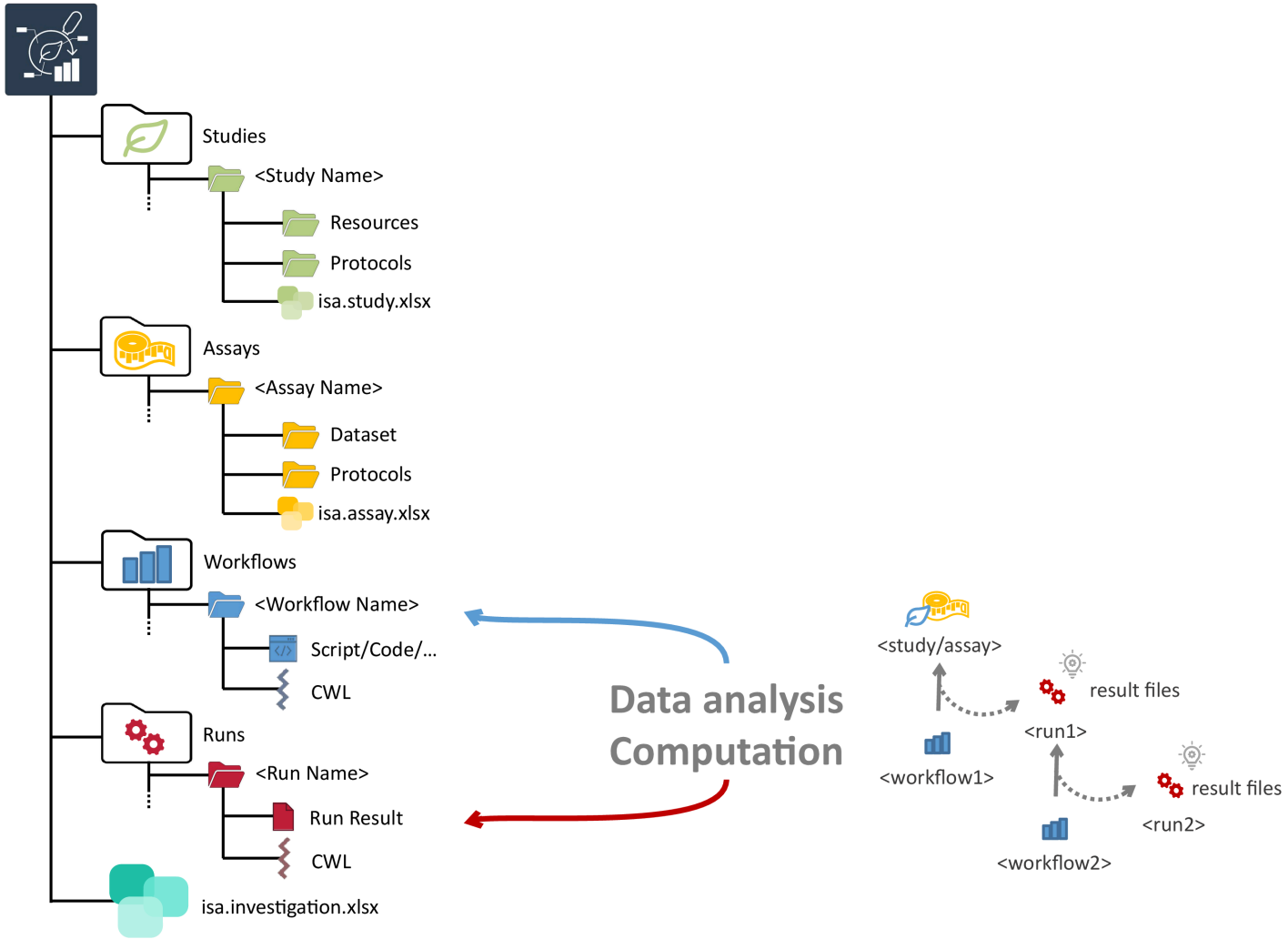
What does an ARC look like?



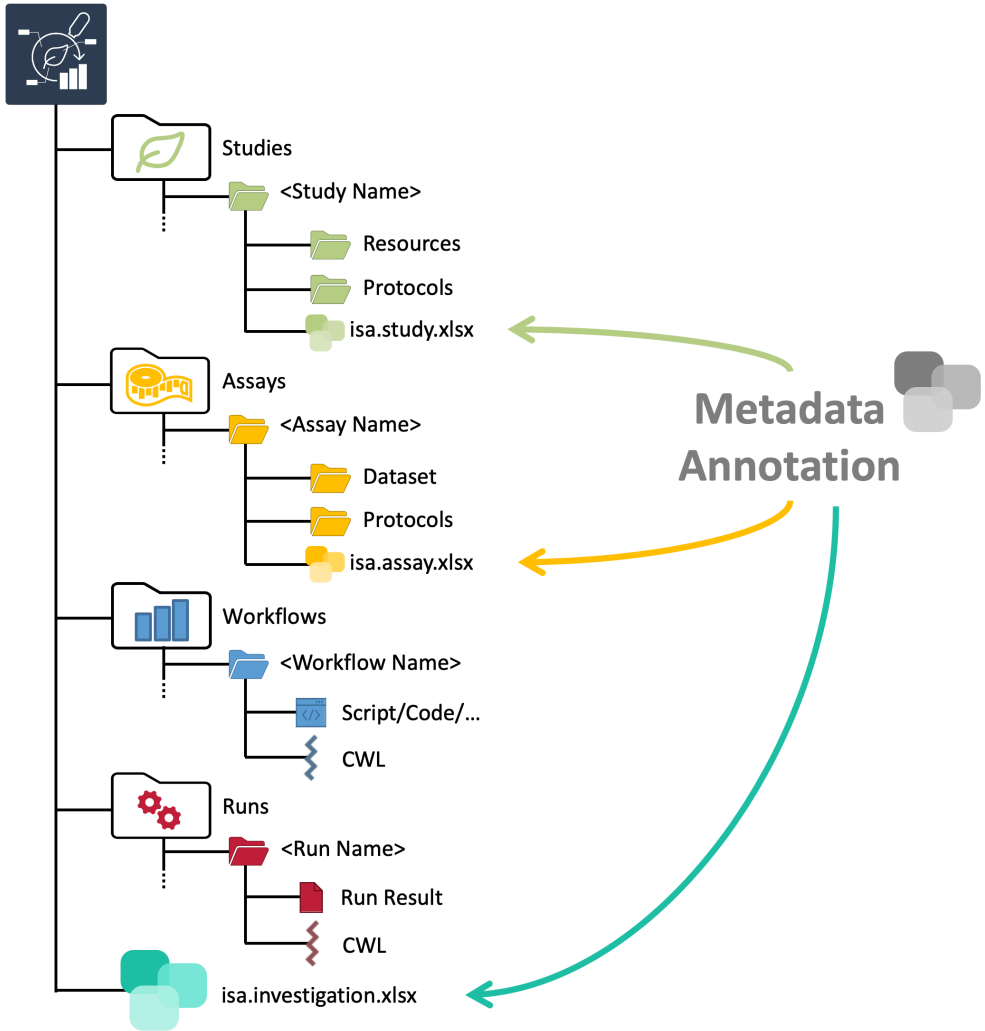
ARCs store experimental data



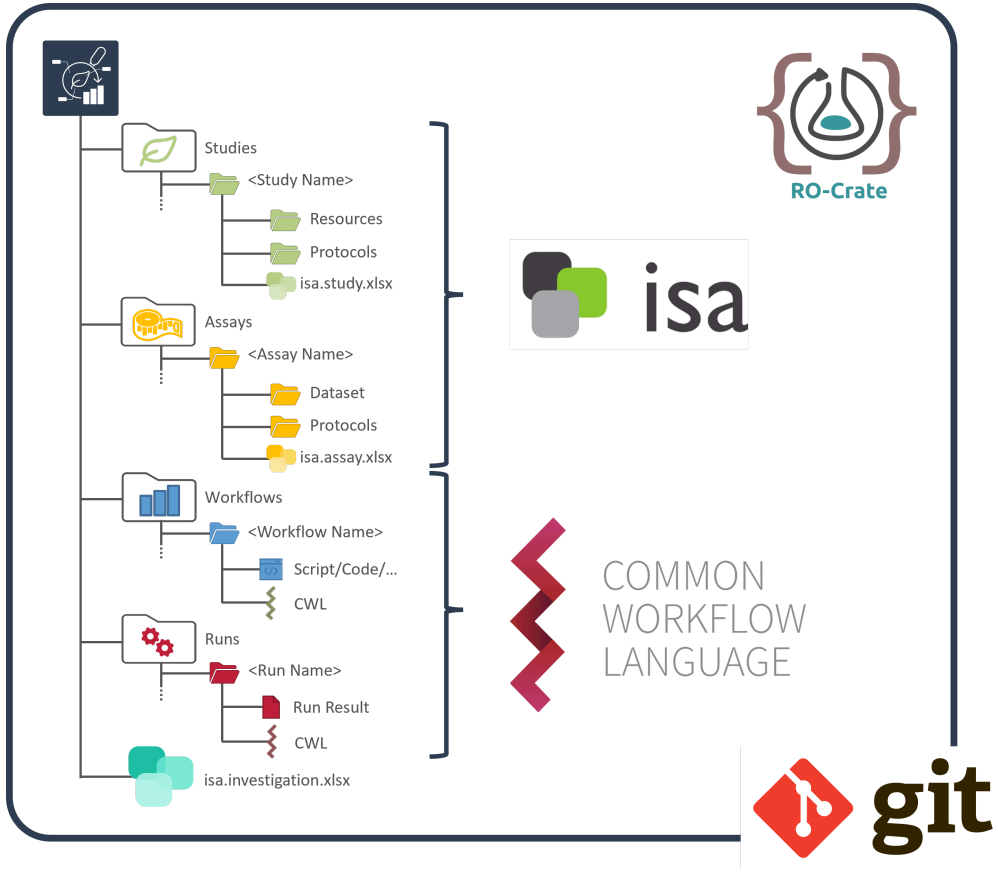
Computations can be run inside ARCs



ARCs come with comprehensive metadata



ARC builds on standards

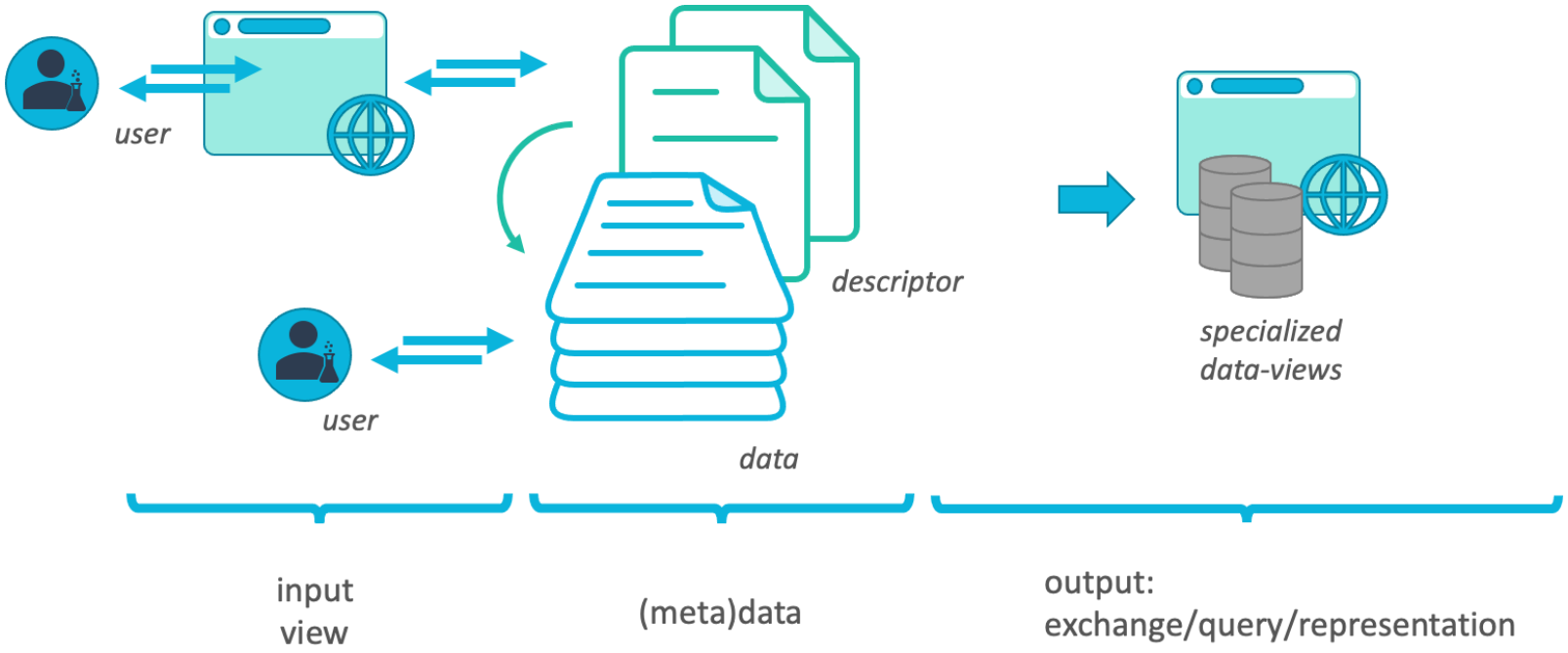


- RO-Crate: standardized exchange
- ISA: structured, machine-readable metadata
- CWL: reproducible, re-usable data analysis
- Git: version control

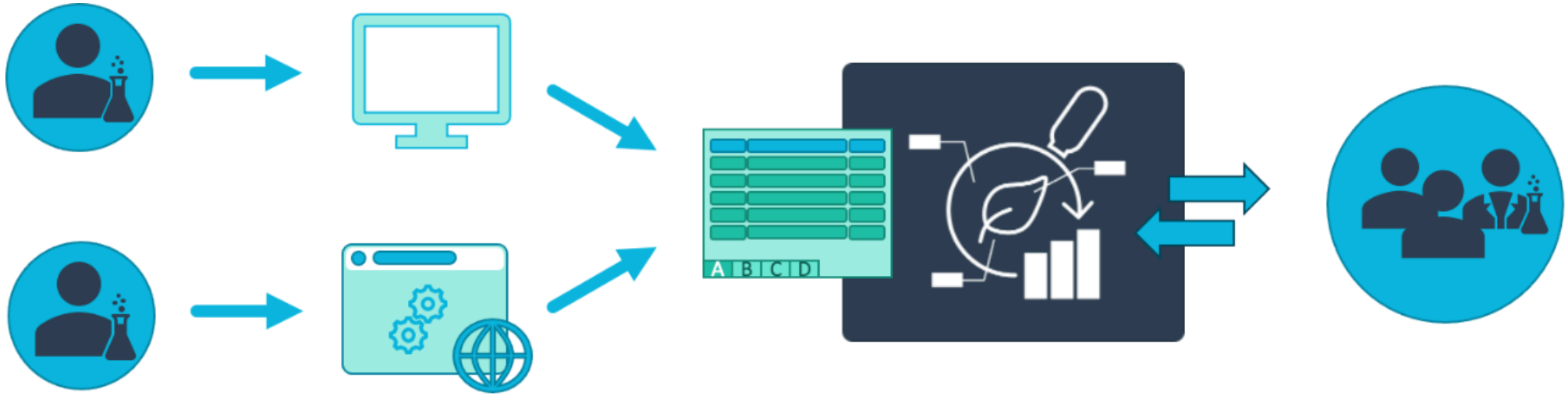
https://isa-tools.org/ | https://www.commonwl.org/
https://www.researchobject.org/ro-crate/ | https://git-scm.com

Everything is a file

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

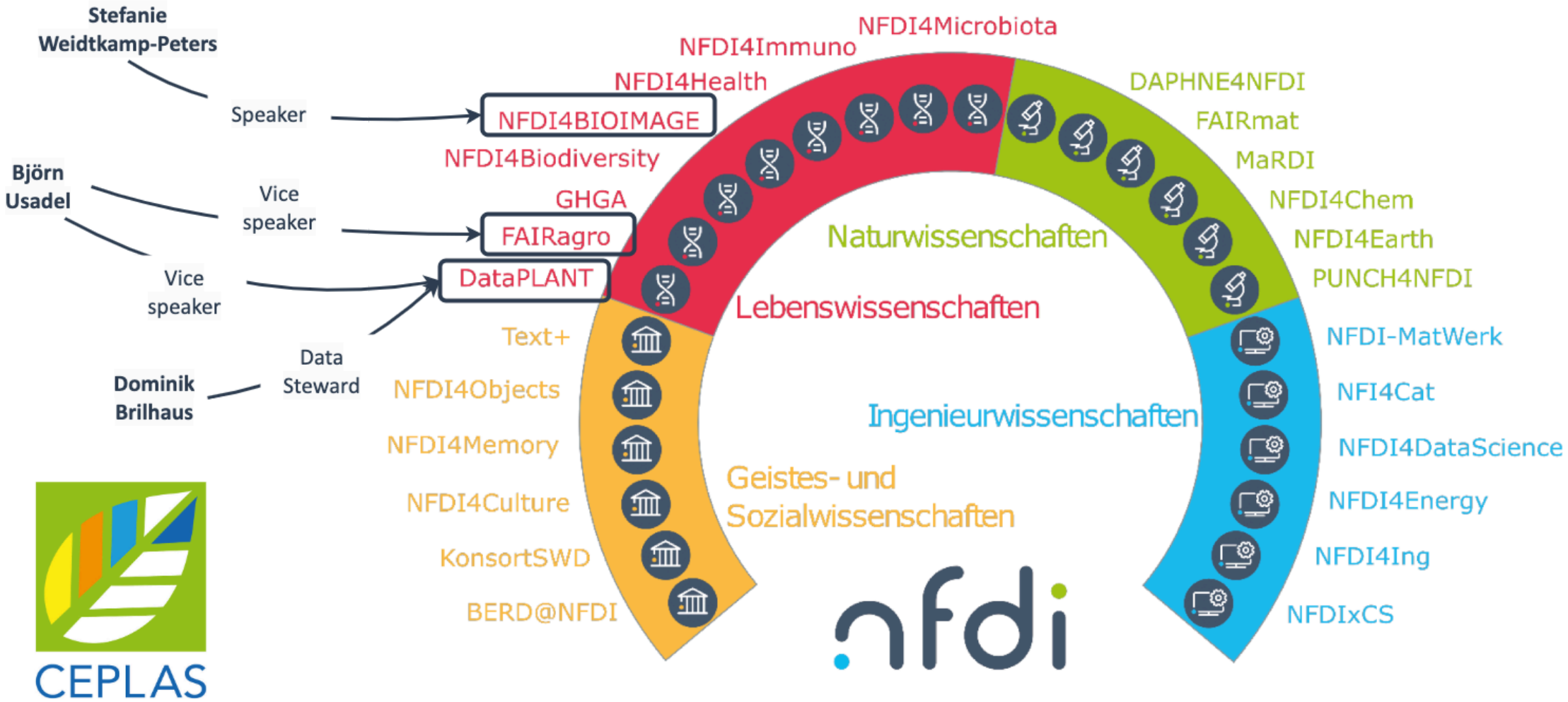


No technical lock-in



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

CEPLAS connection to the NFDI

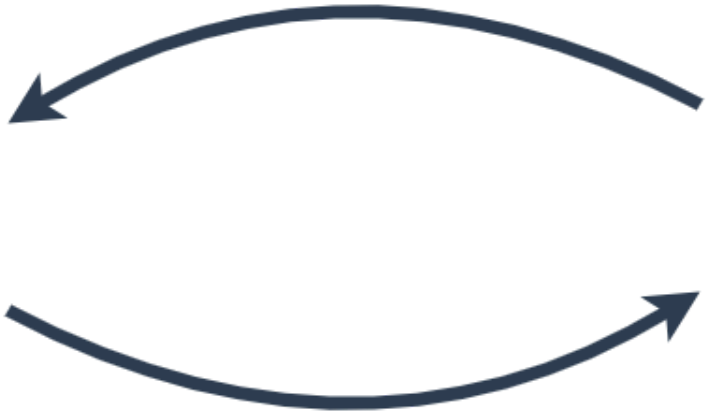


CEPLAS

Data Stewardship between DataPLANT and the community

Community

nfdi4plants

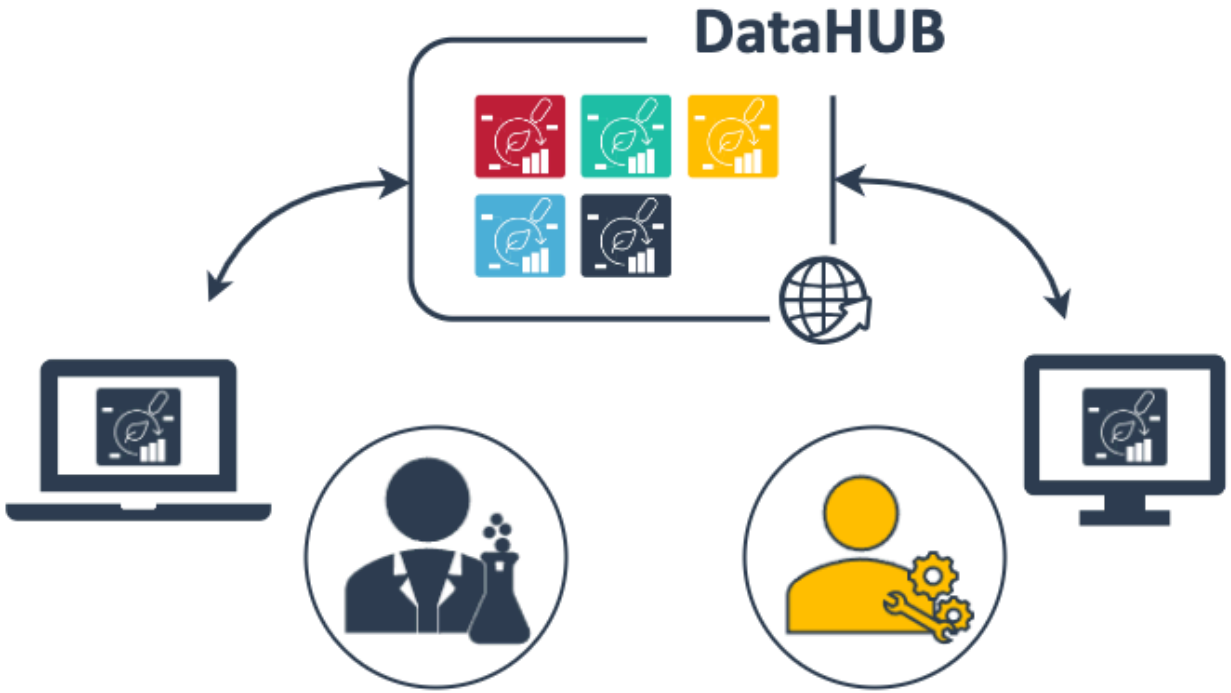


CEPLAS

Domain experts
User experience
Training

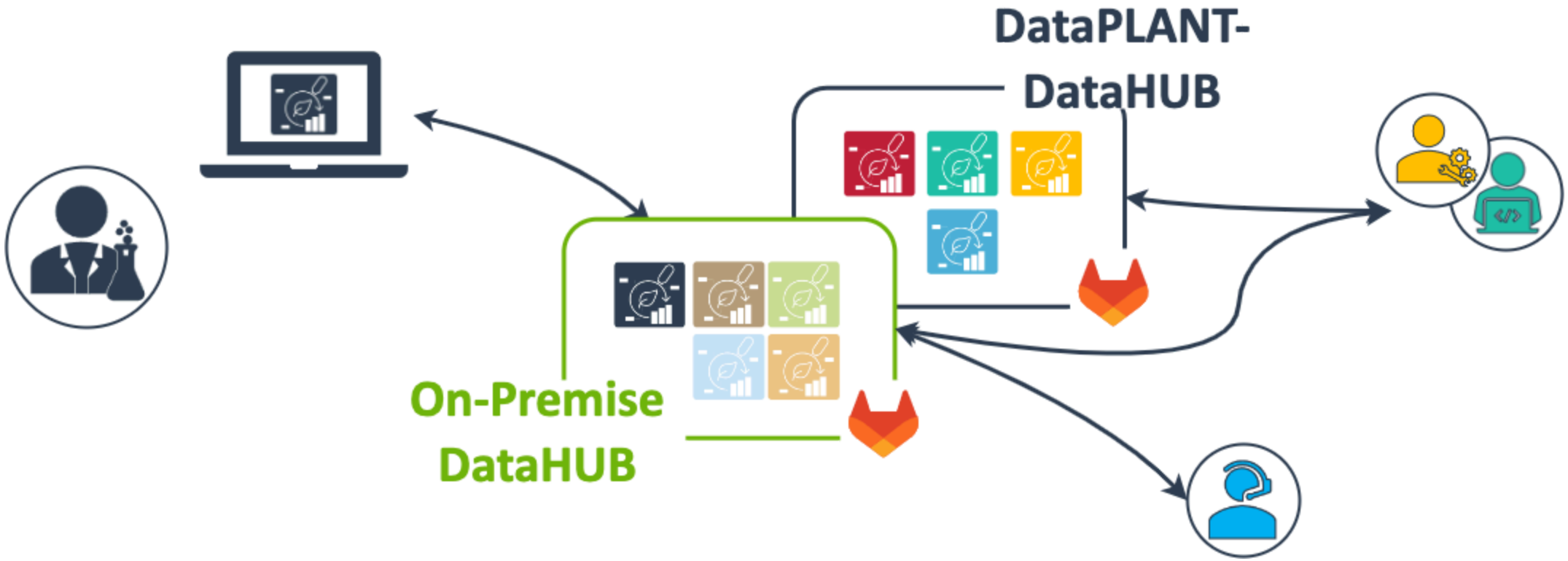
Service provider
Developers
Tech experts

The DataPLANT DataHUB – a GitLab *Plus*



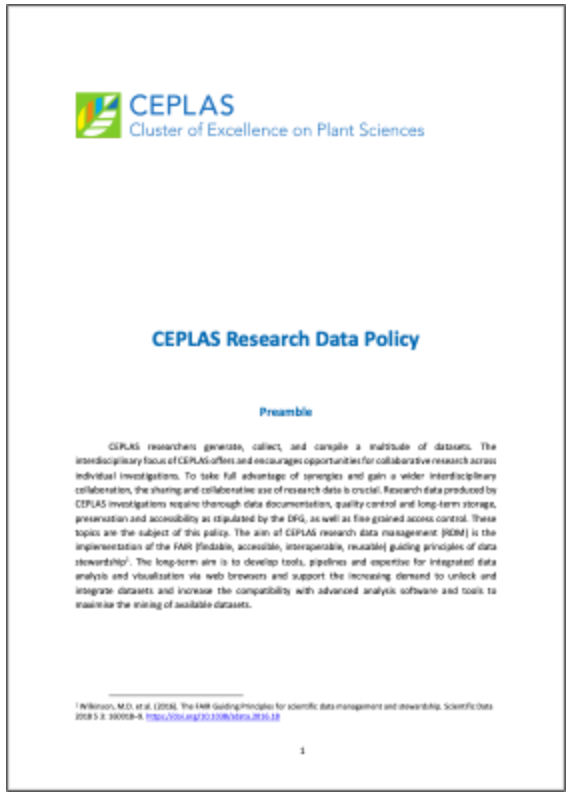
=

On-premise DataHUBs



ARC services are available as on-premise option

CEPLAS Research Data Policy


CEPLAS
Cluster of Excellence on Plant Sciences

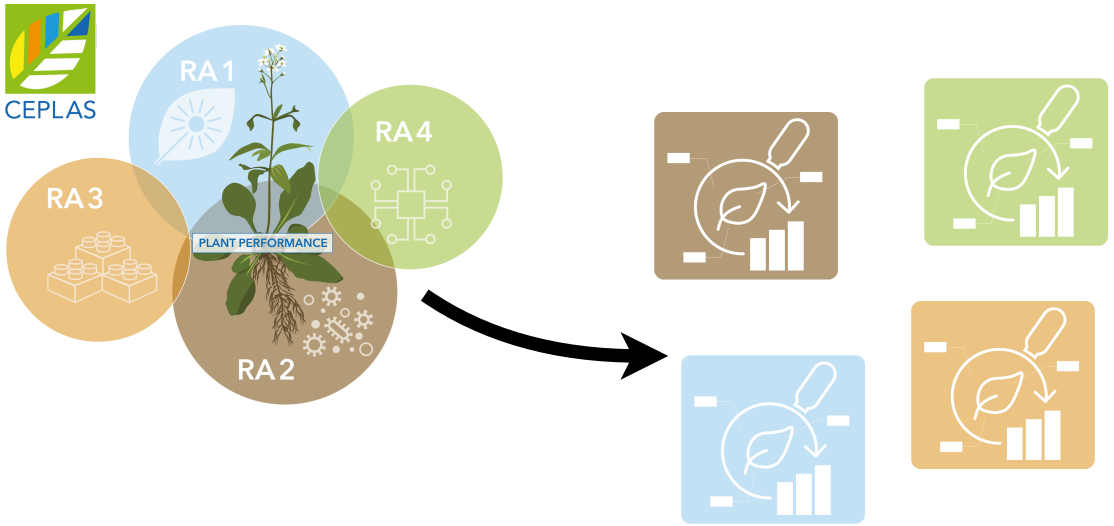
CEPLAS Research Data Policy

Preamble

CEPLAS researchers generate, collect, and compile a multitude of datasets. The interdisciplinary focus of CEPLAS offers and encourages opportunities for collaborative research across individual investigations. To take full advantage of synergies and gain a wider interdisciplinary collaboration, the sharing and collaborative use of research data is crucial. Research data produced by CEPLAS investigations require thorough data documentation, quality control and long-term storage, preservation and accessibility as stipulated by the DFG, as well as fine-grained access control. These topics are the subject of this policy. The aim of CEPLAS research data management (RDM) is the implementation of the FAIR (findable, accessible, interoperable, reusable) guiding principles of data stewardship[1]. The long-term aim is to develop tools, pipelines and expertise for integrated data analysis and visualization via web browsers and support the increasing demand to unlock and integrate datasets and increase the compatibility with advanced analysis software and tools to maximize the mining of available datasets.

[1] Wilkinson, M.D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 2018 3: 160018-9. https://doi.org/10.1038/sdata.2016.18

1



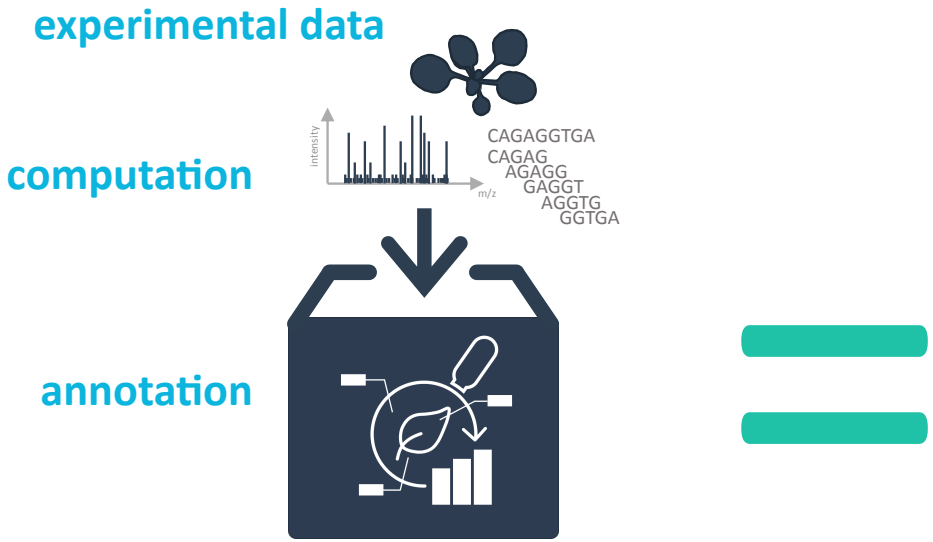
Hands-on: ARCitect part 1

Follow the **Start Here** guide in the DataPLANT knowledge base.

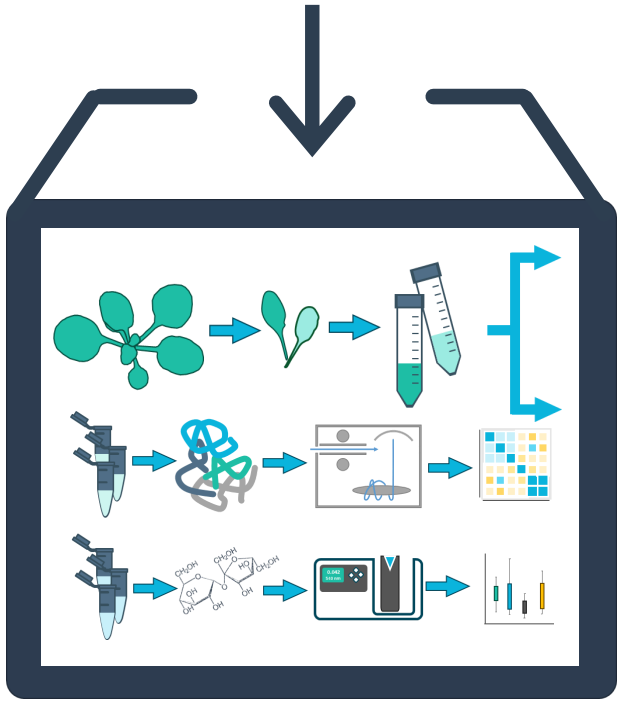
Until step "add a study"

- 01-create-arc-scaffold.mdx
- 02-investigation.mdx
- 03-study.mdx

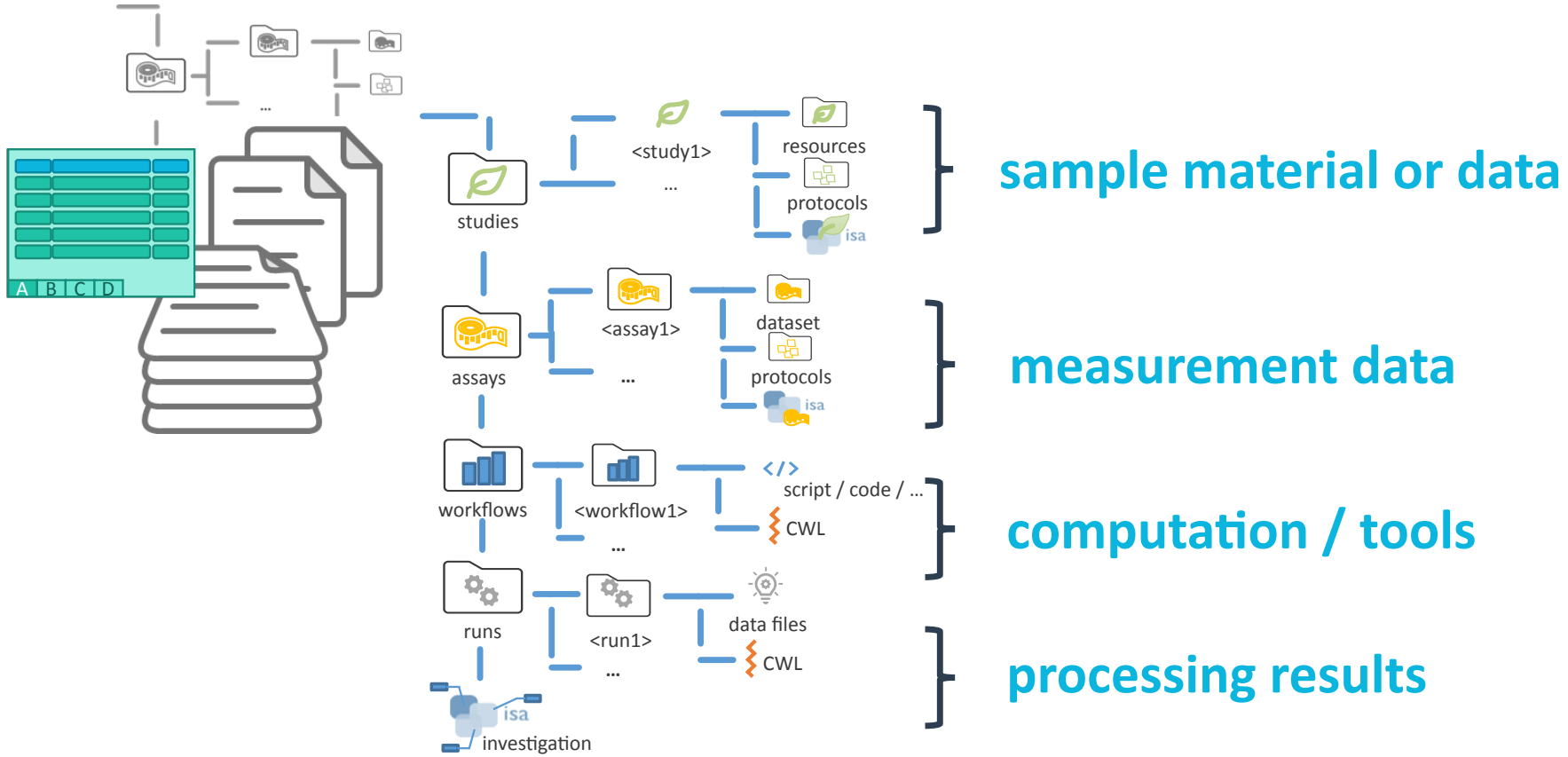
ARC: Annotated research context



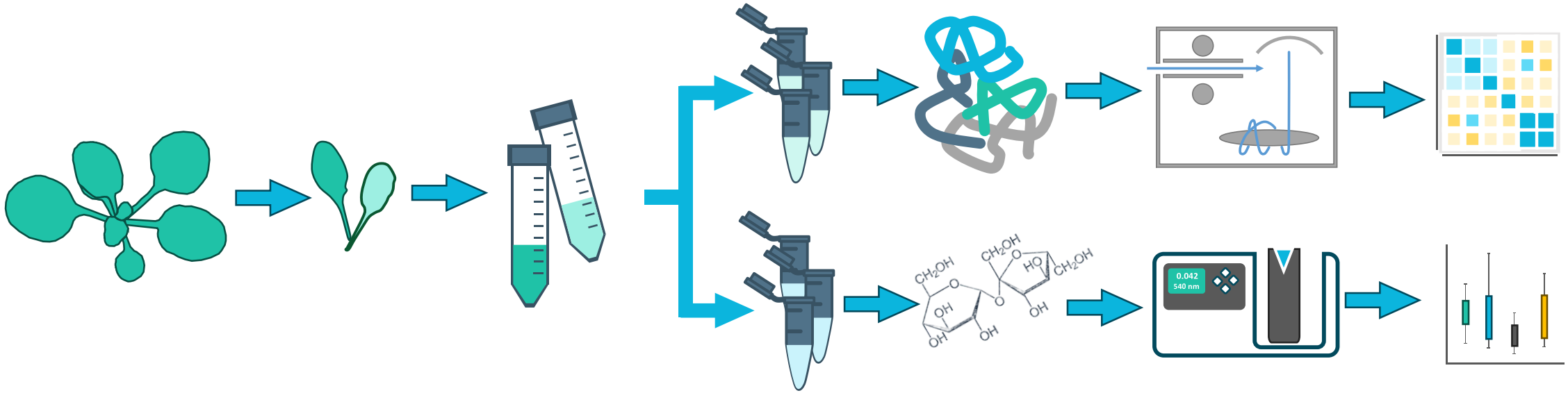
FAIR
digital (research) objects



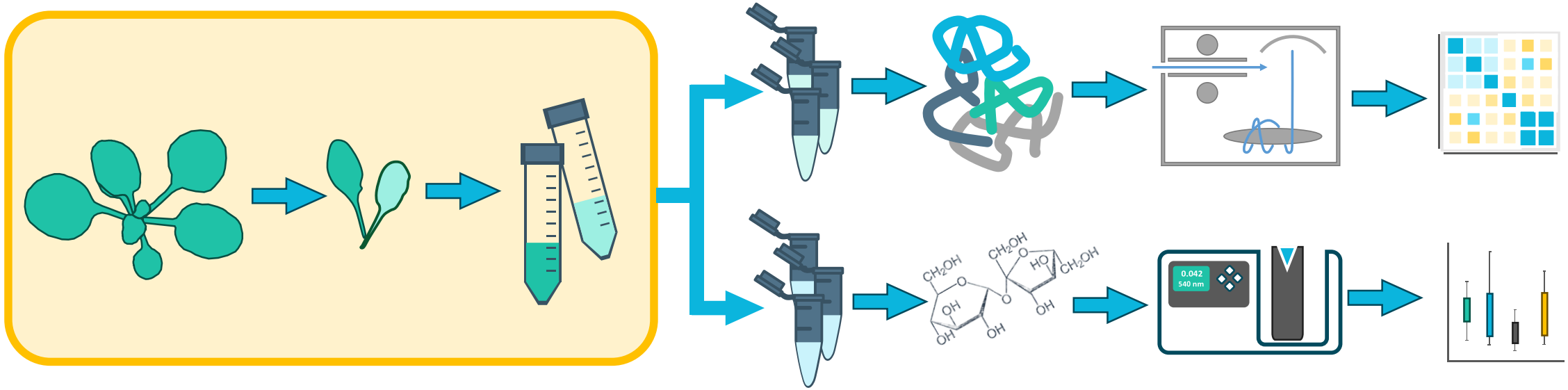
The ARC scaffold structure



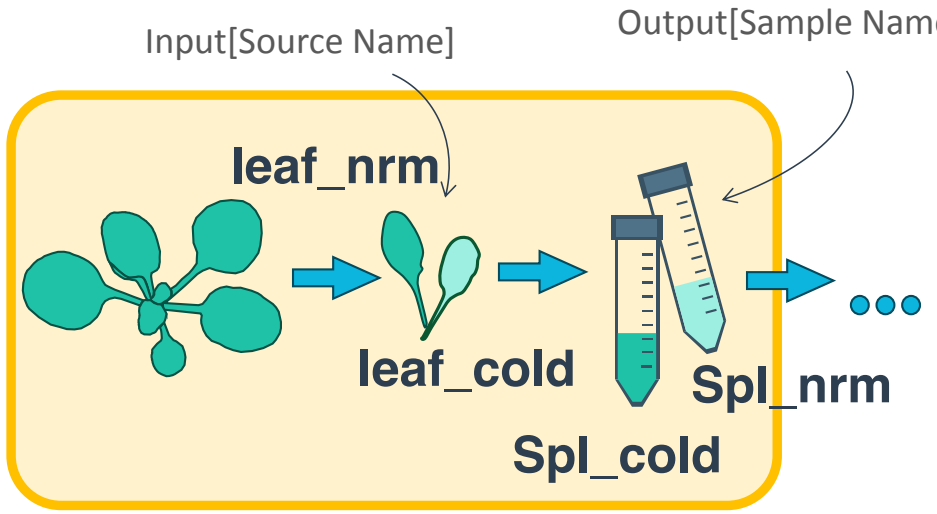
A small prototypic project

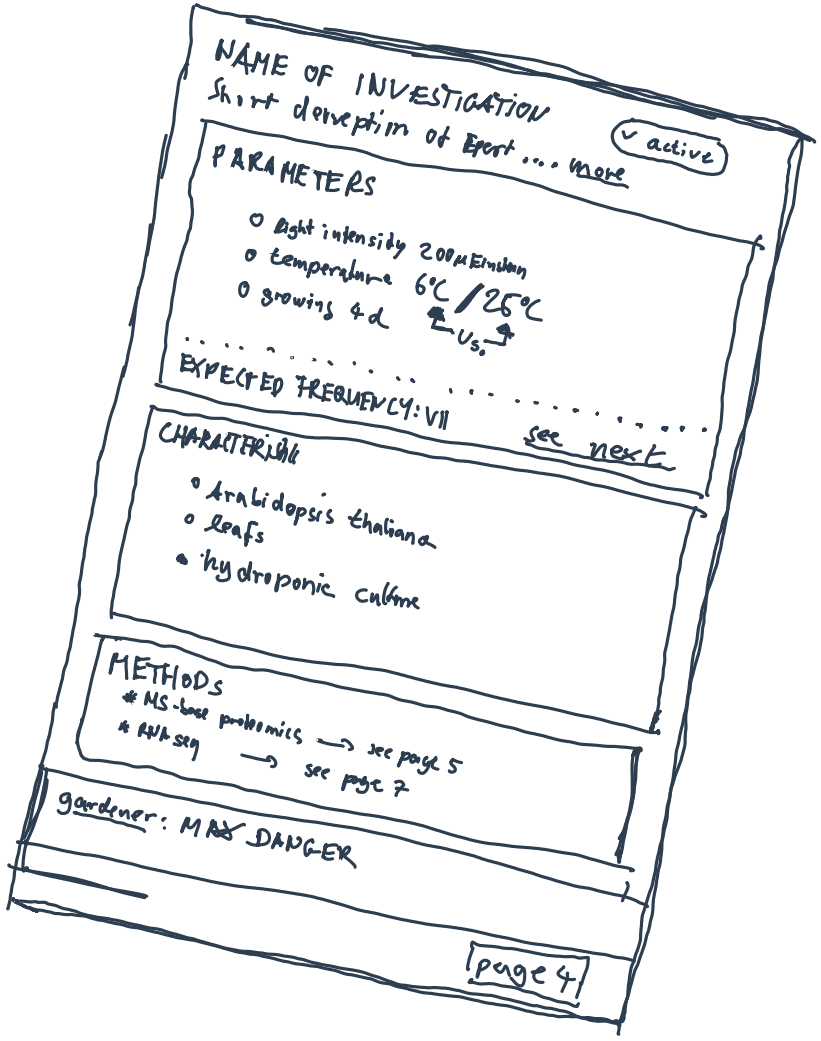


Divide and conquer for reproducibility



Identifying the ‘study’ part

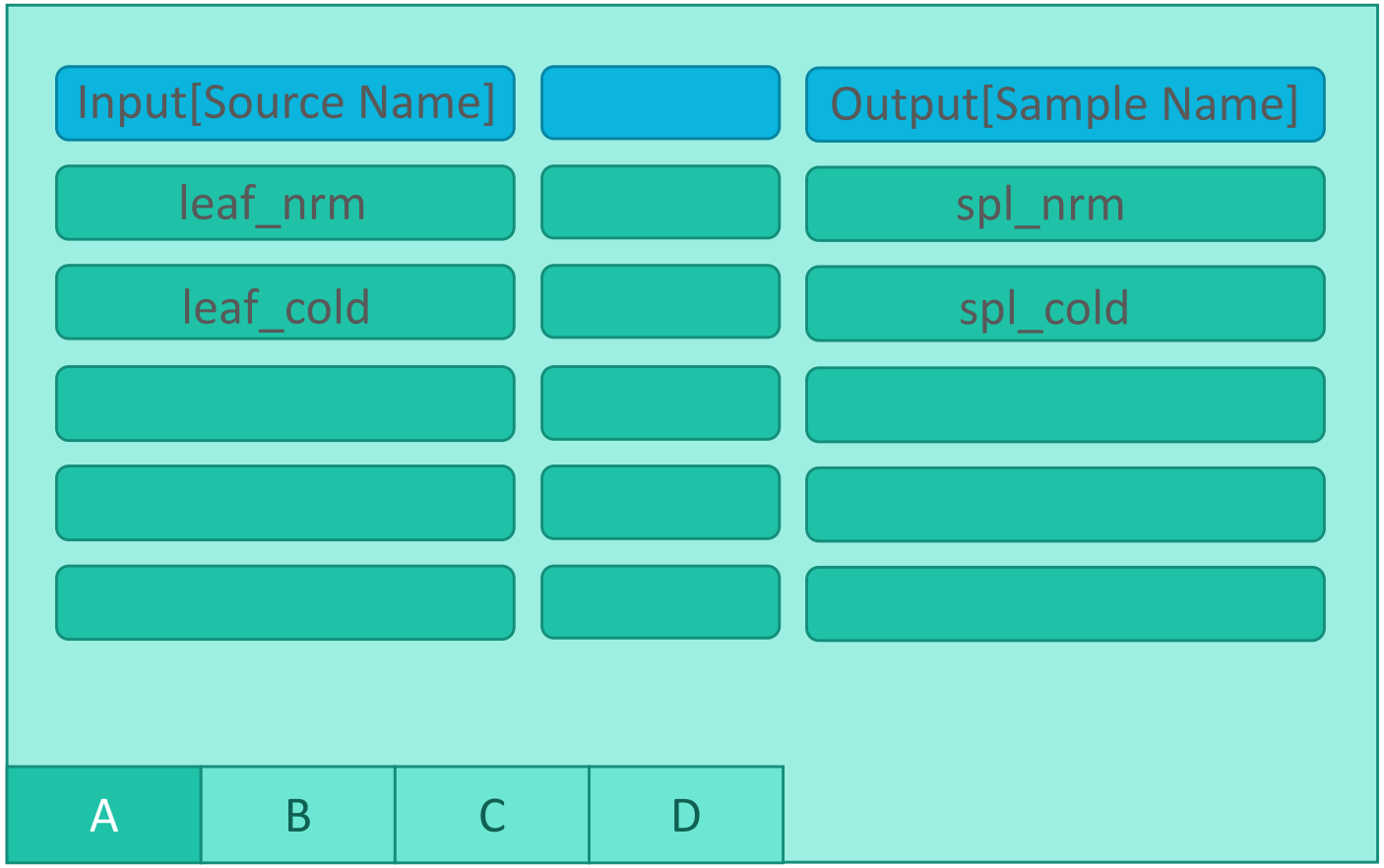




Protocol REF

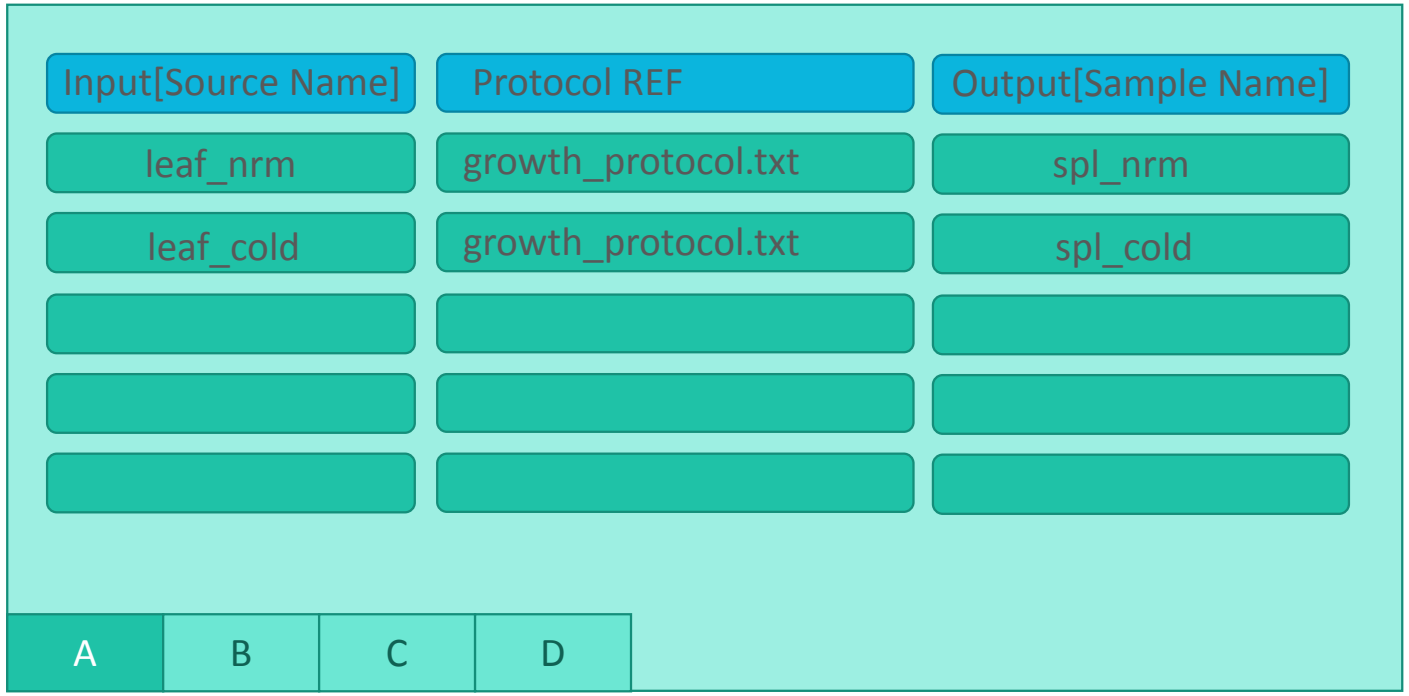
growth_protocol.txt

A table-based organization schema



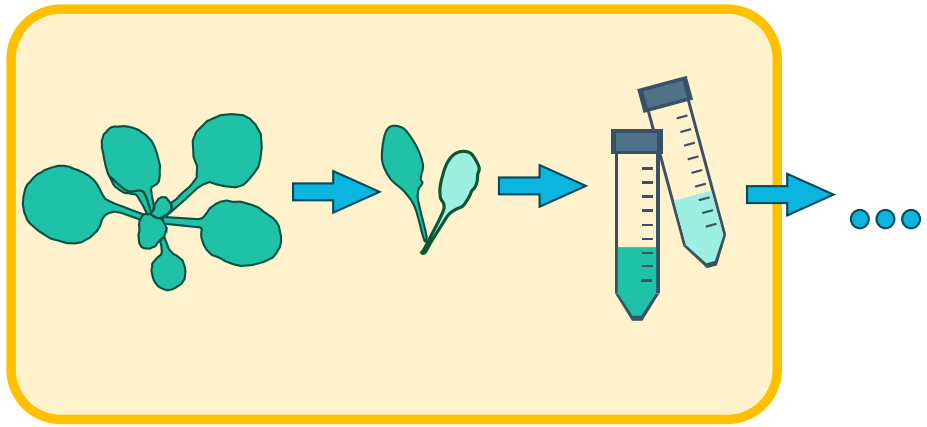
Referencing a protocol

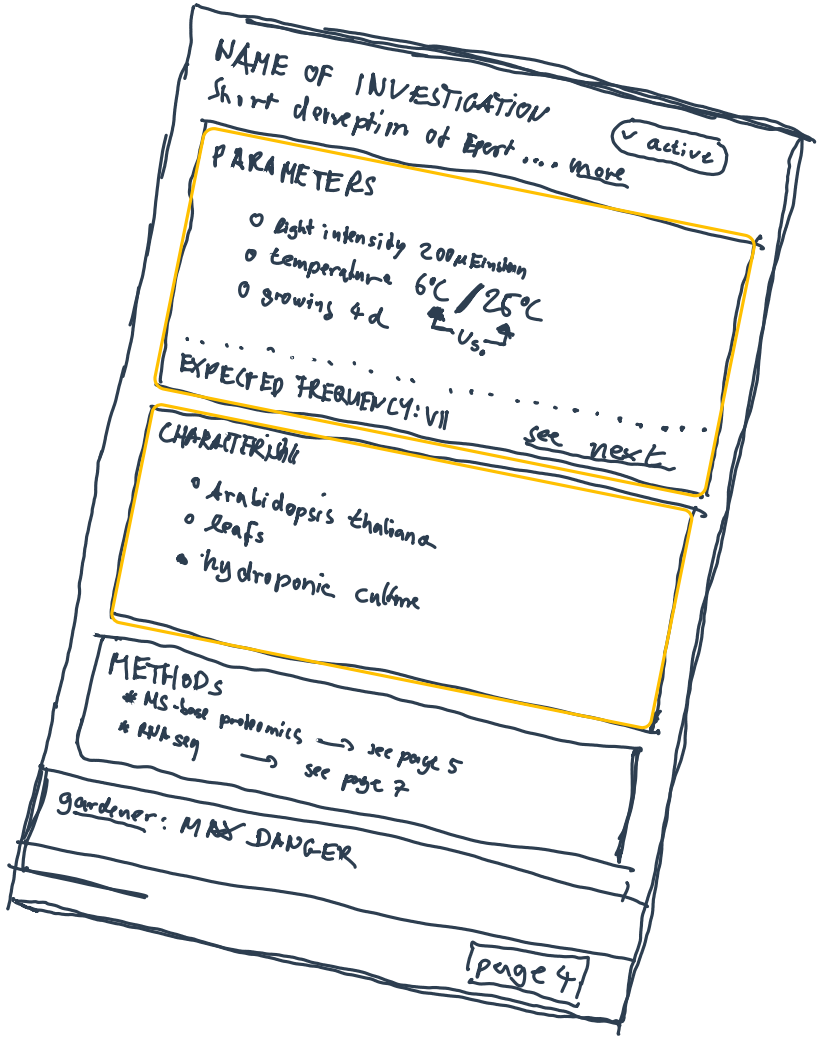
This allows you to reference the free-text, human-readable protocol.



- 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: `UNIPROTRDFS`

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`

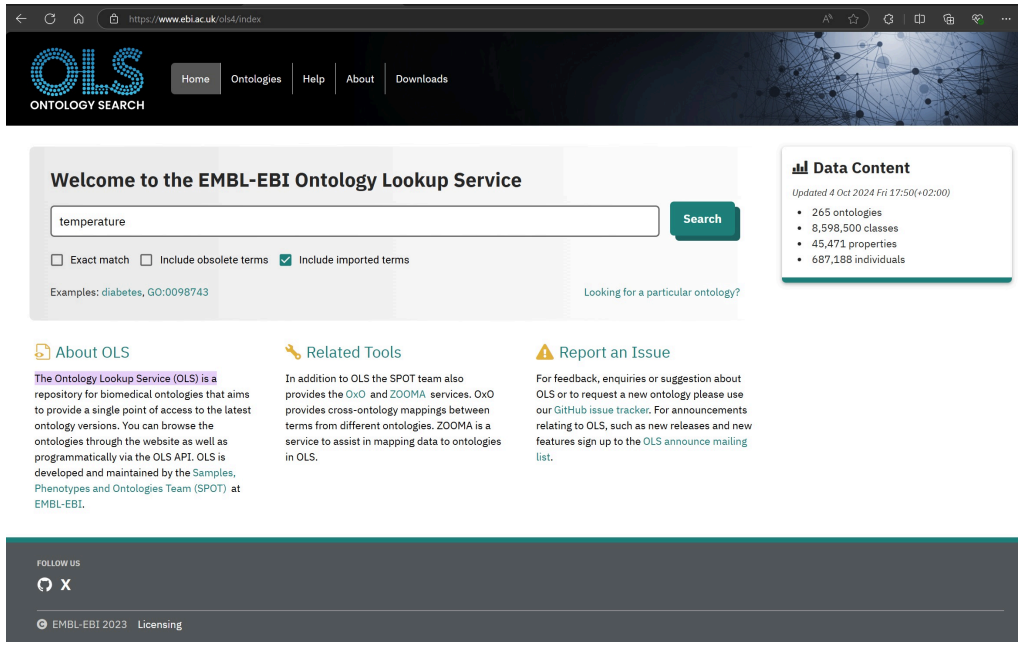
temperature `PATO:0000146`

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

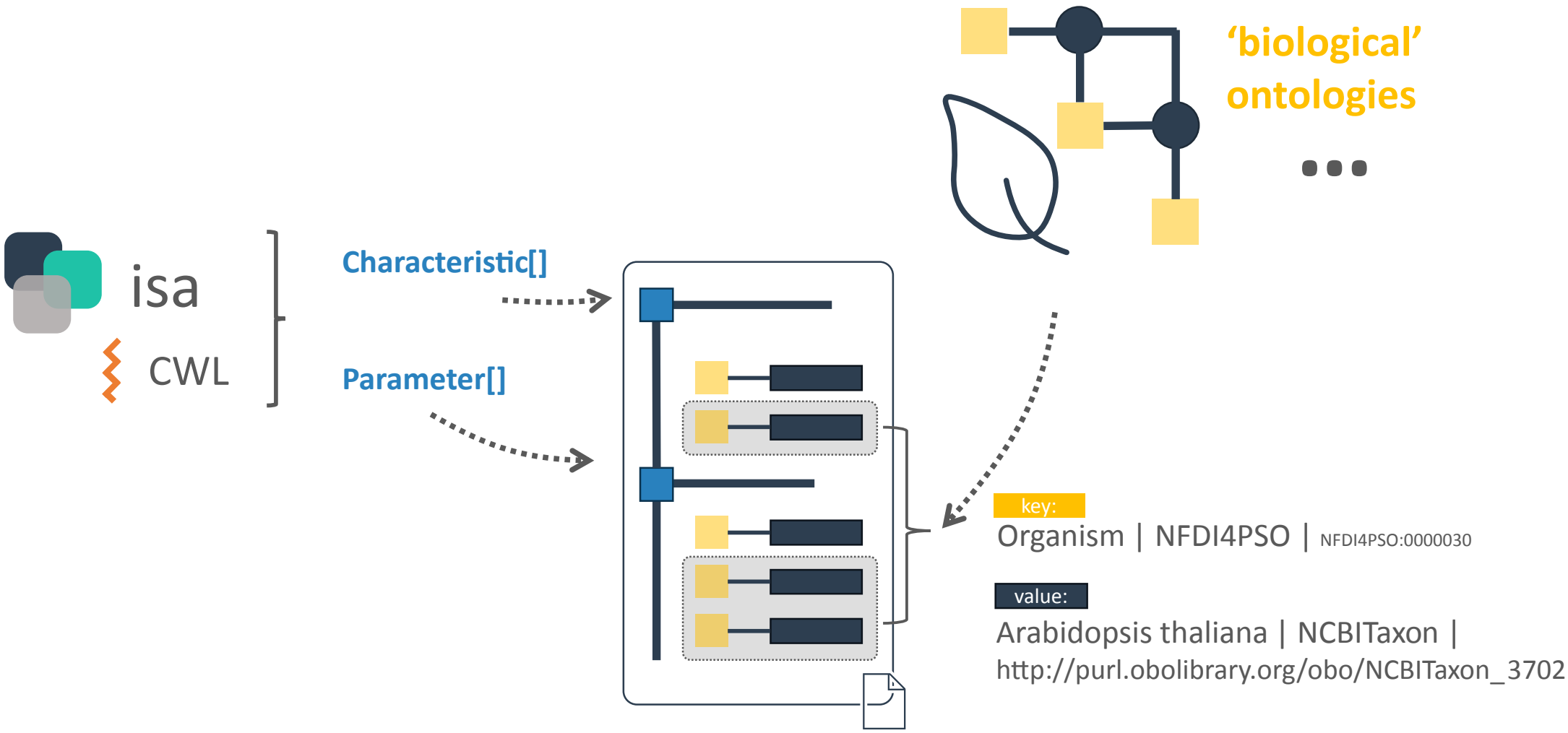
A physical quality of the thermal energy of a system.

Ontology: `PATO`

Also appears in: `NGBO` `HTN` `CAO` `ZP` `AGRO` `OMIABIS` `OBIB` `MONDO` `TXPO` `MCO` +



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

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, rovenance,
- licensing
- ethical aspects
- ...

technical

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

Metadata from a FAIR perspective

Findable

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

Accessible

- information about origin
- location of storage
- access rights

Interoperable

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

Reusable

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

Metadata "Standards"

Examples from Minimum Information for Biological and Biomedical Investigations (MIBBI):

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

💡 Check out https://fairsharing.org/ for more examples

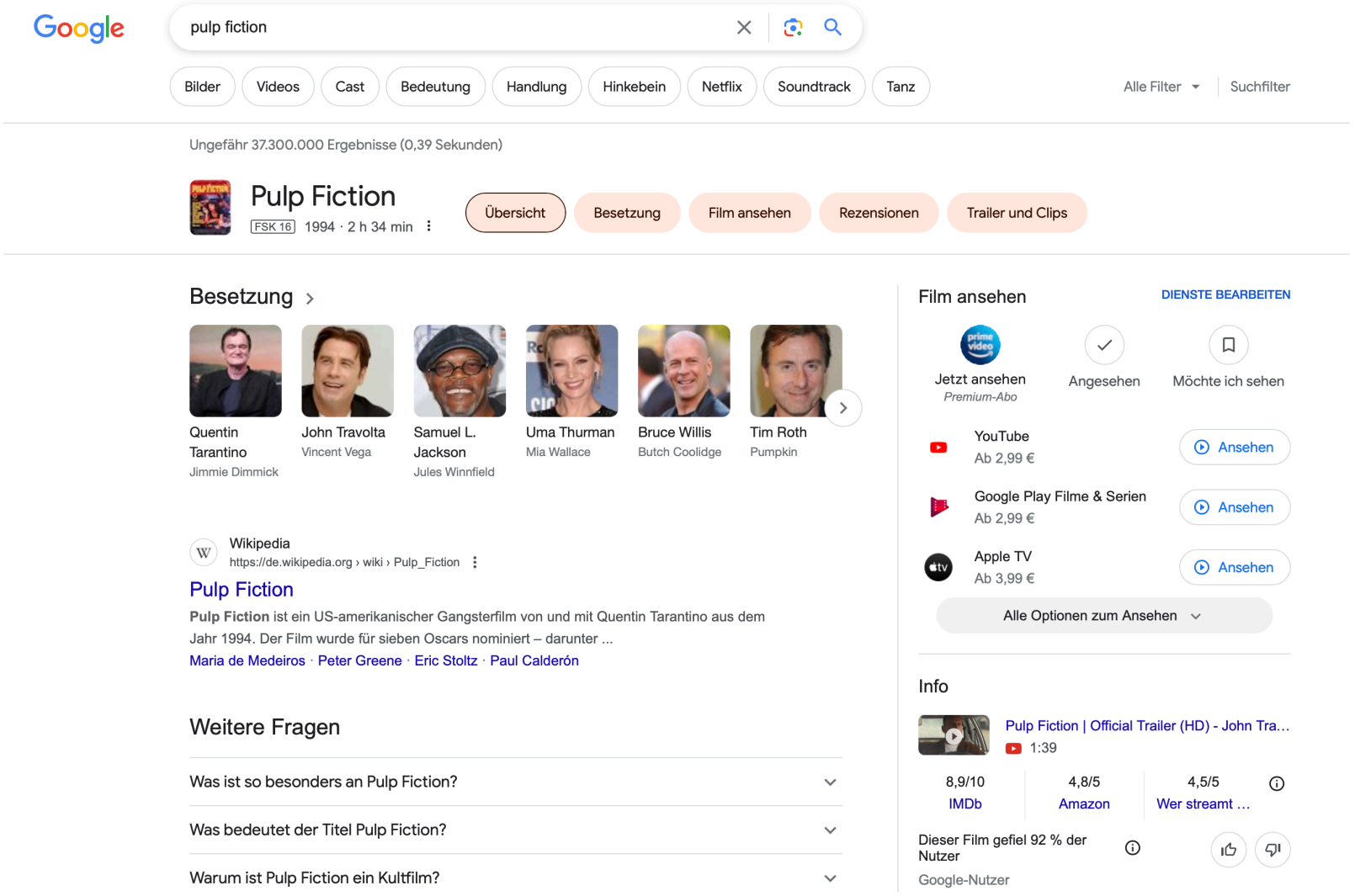
Metadata standards ≈ Checklists

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

A small Interactive detour

-> favorite Movie

How does google "know"?!



Schemas and machine-readability

Structured data and the internet

Schema.org

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

Structured data and the internet: Schema.org

https://schema.org/Person

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

JSON-LD

JSON-LD = JavaScript Object Notation for Linked Data

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

RDFa

RDFa = Resource Description Framework in Attributes

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

Standards

Dublin Core

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

DataCite Schema

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

DataCite Schema: Simple Example

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

Ontologies

Ontology

(Sometimes also referred to "semantic model")

An ontology combines features of

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

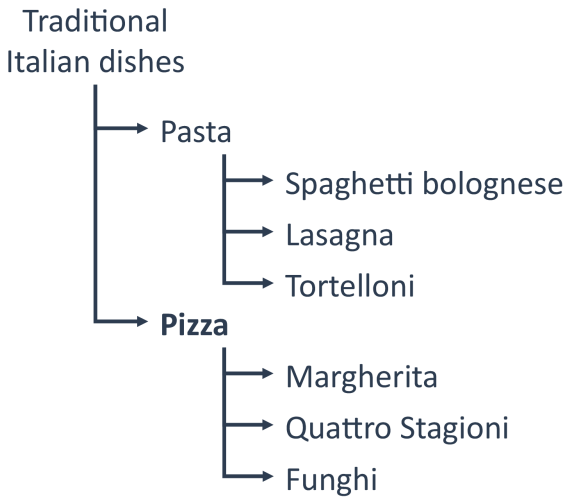
Dictionary

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

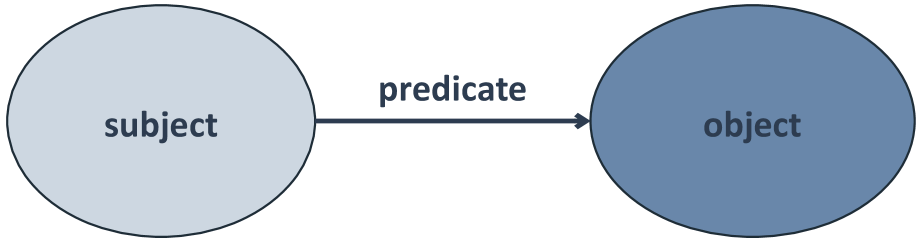
Dictionary of synonyms and relations

Pizza ≈ Lahmacun ≈ Focaccia ≈ Flammkuchen

Ontology

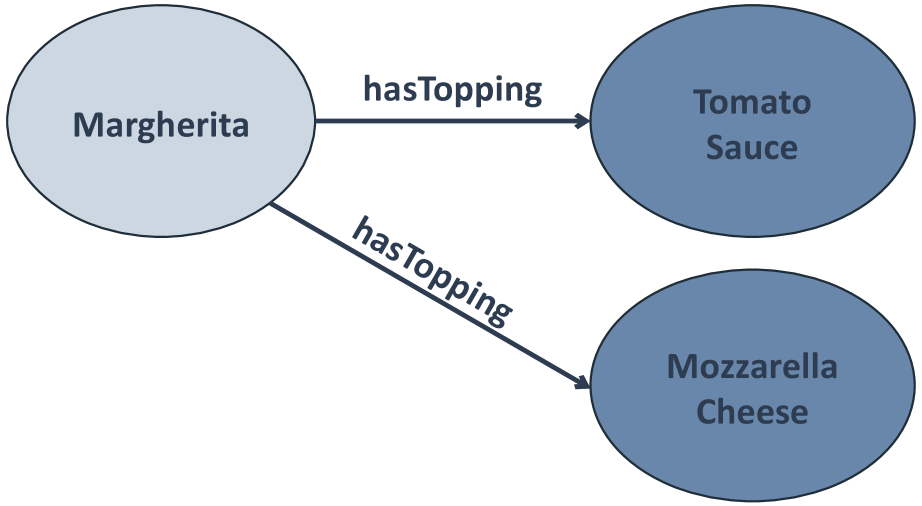
- Structures a set of **concepts** in a particular area and the relations between them in a **graph-like manner**
- Can be used in disambiguation, defining hierarchies, a standard to define terms
- Define a common vocabulary of concepts and their relationships to **model** a particular domain while making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

The semantic triple

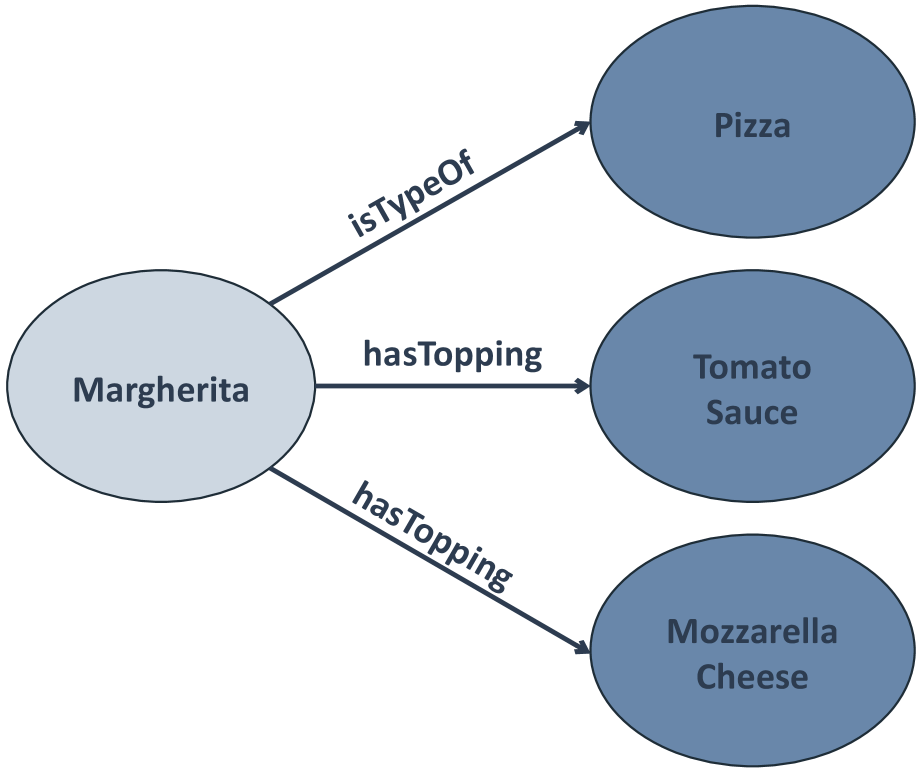


Modeling a pizza menu

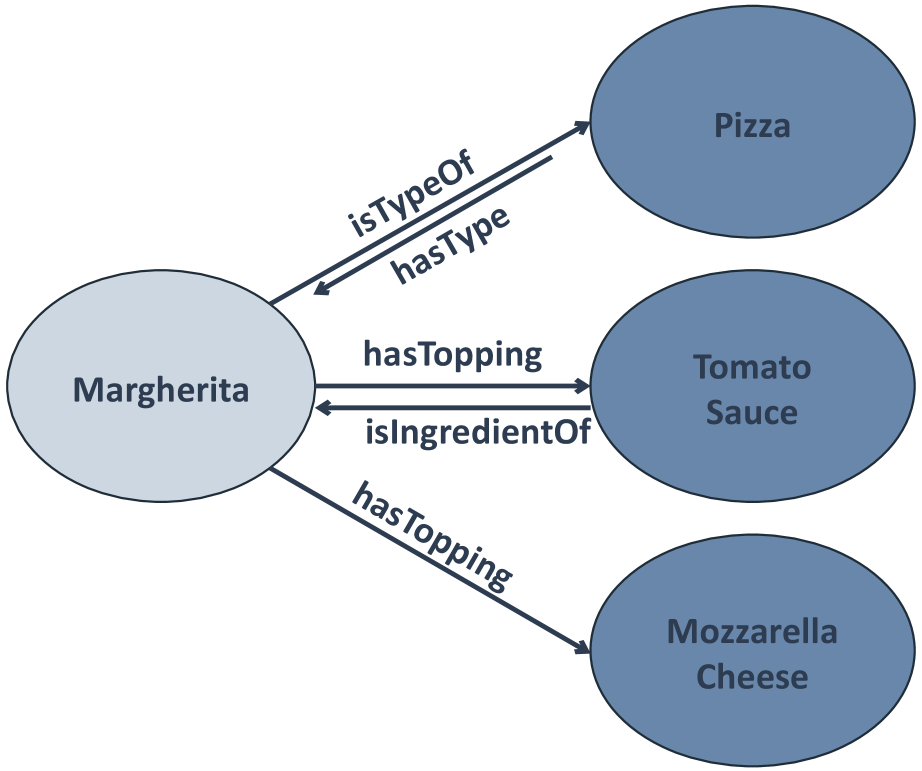
Modeling a pizza menu



Modeling a pizza menu

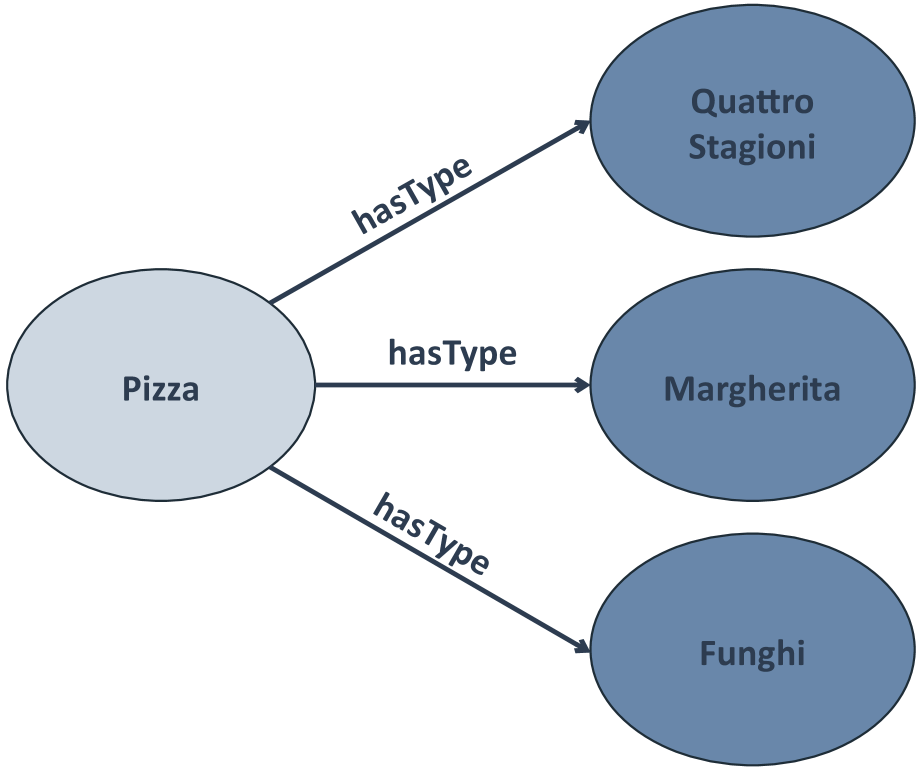


Predicates have two directions

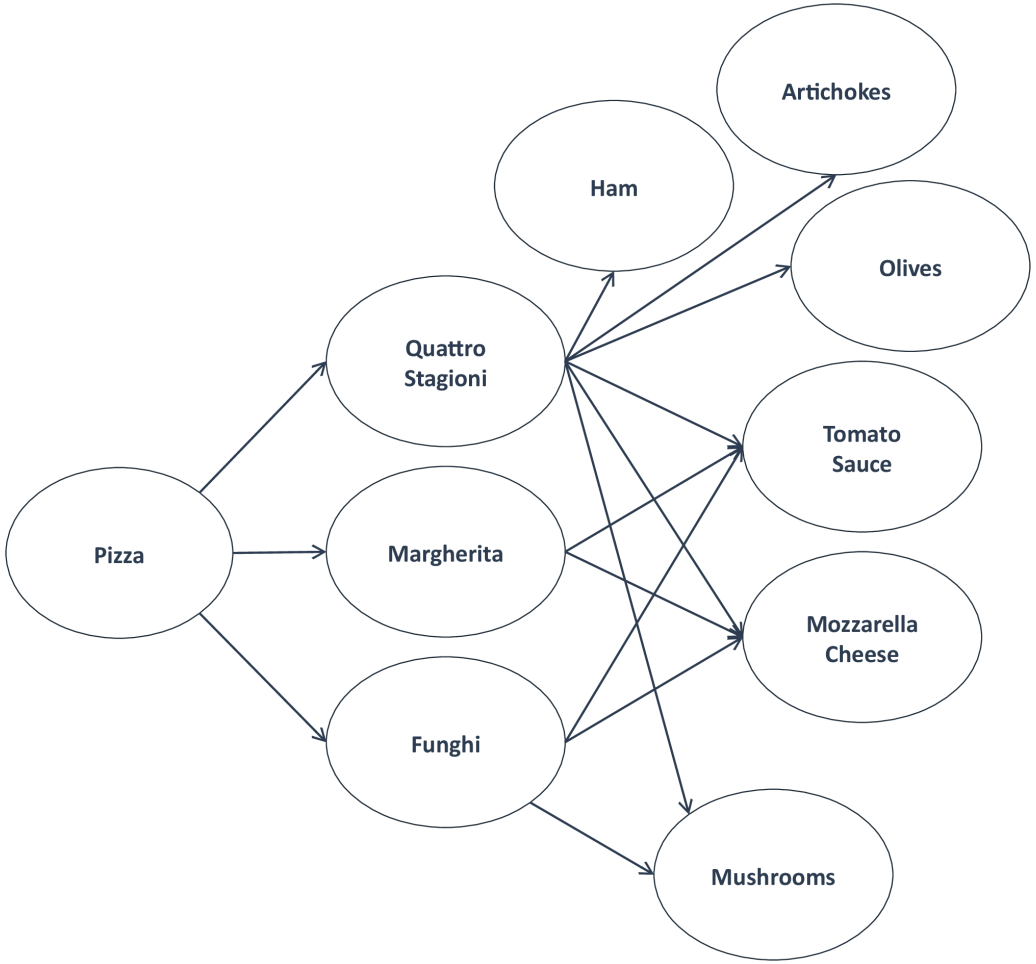


Looking at the menu from a different perspective

An object of one triplet can be the subject to another



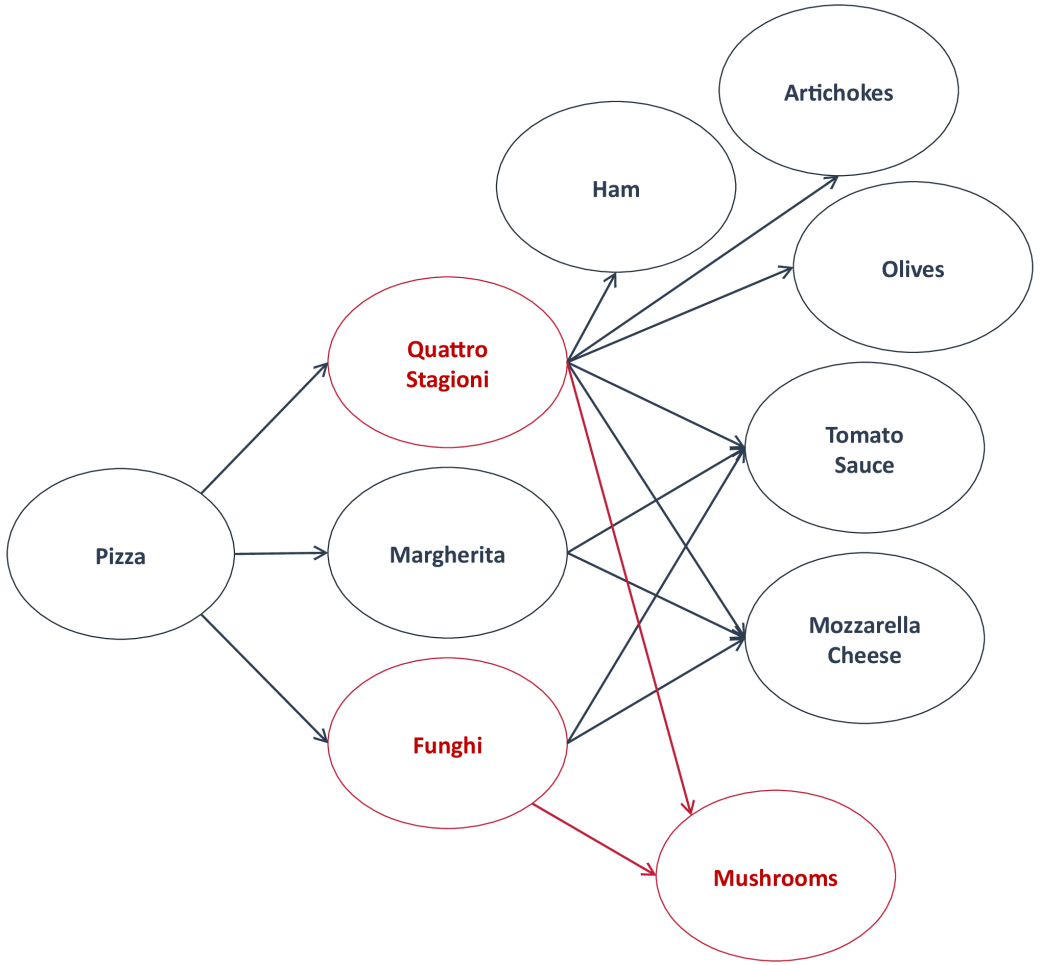
(Towards) a knowledge graph



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

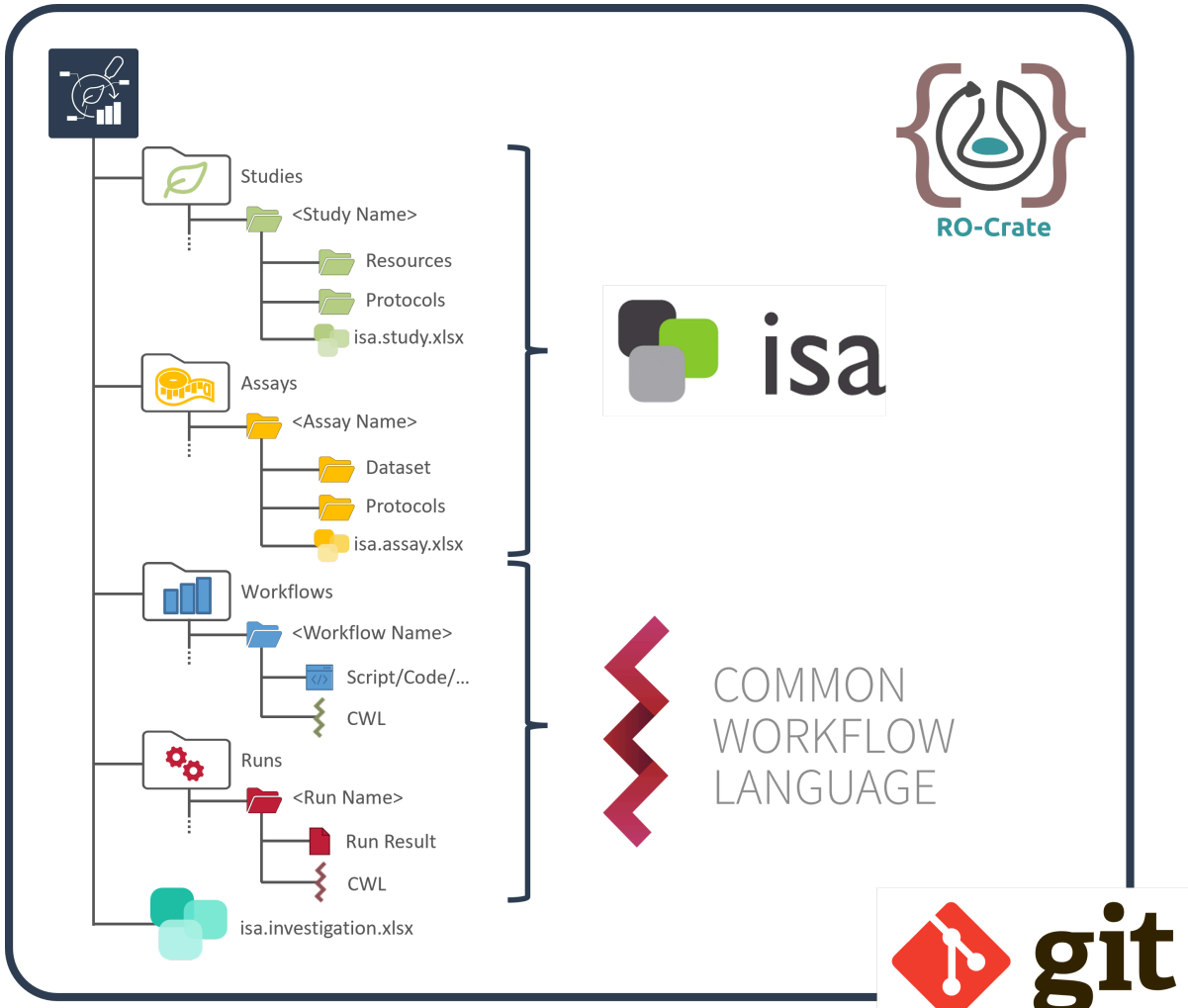
- Description: http://edamontology.org/page
- Browser: https://edamontology.github.io/edam-browser

PECO ontology

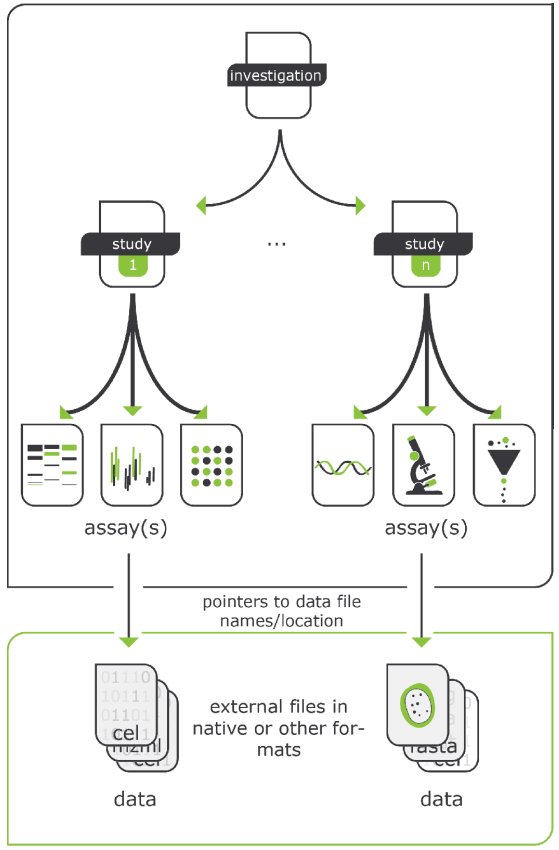
- Human-readble: https://www.ebi.ac.uk/ols/ontologies/peco
- Raw (OWL): http://purl.obolibrary.org/obo/peco.owl

> Explore more examples
>
> - https://www.ebi.ac.uk/ols/
> - https://bioportal.bioontology.org

ARC builds on standards



ARC builds on ISA



Investigation

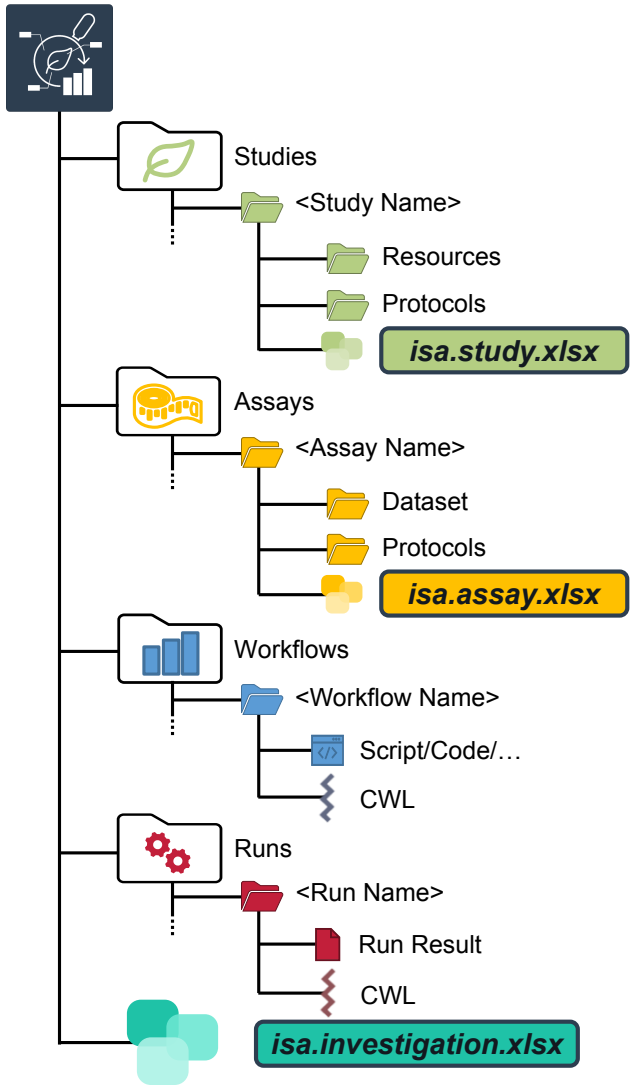
Overall goals
Scientific context

Study

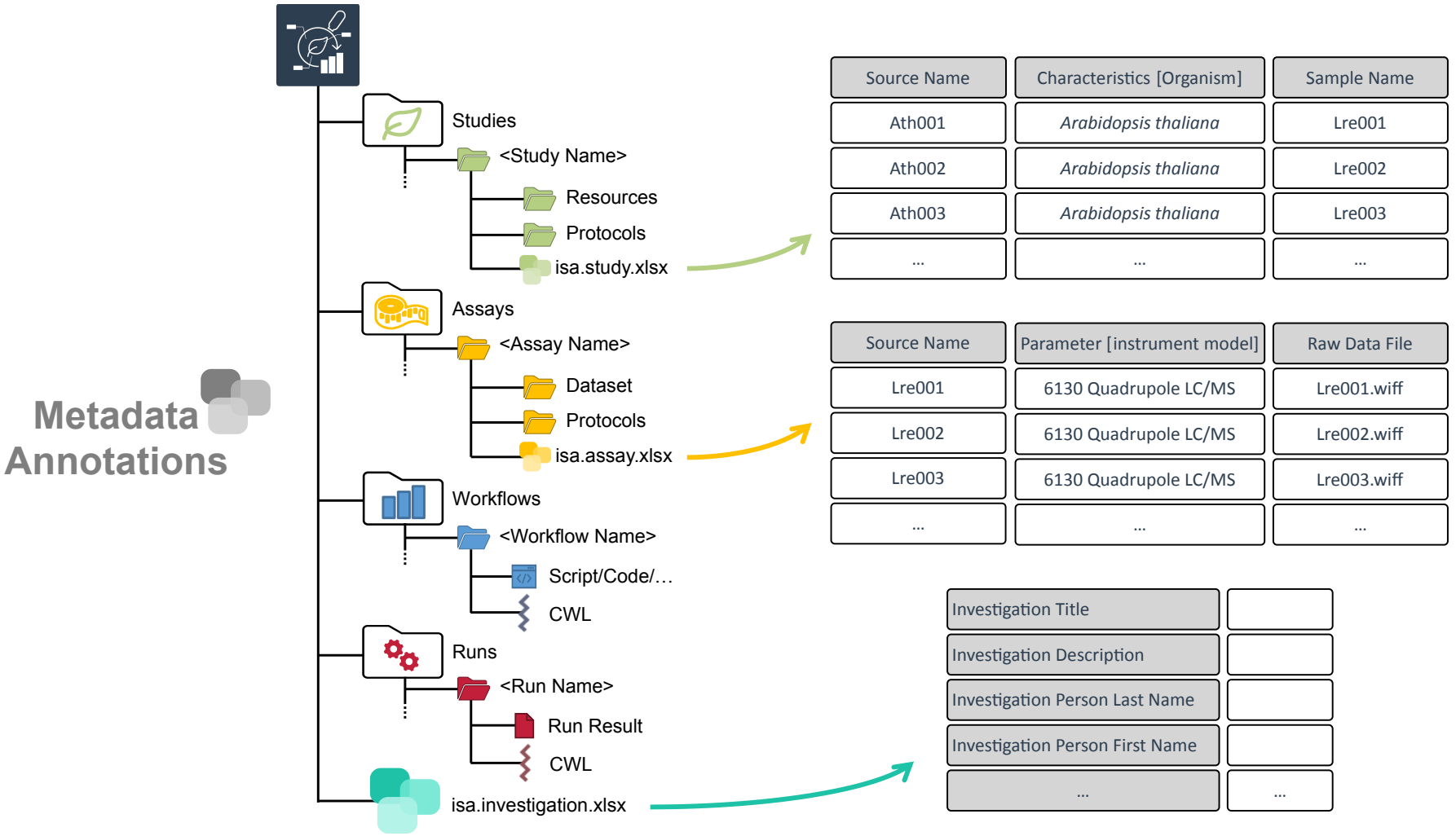
Experimental design

Assay

Leading to (raw) data



ARC builds on ISA

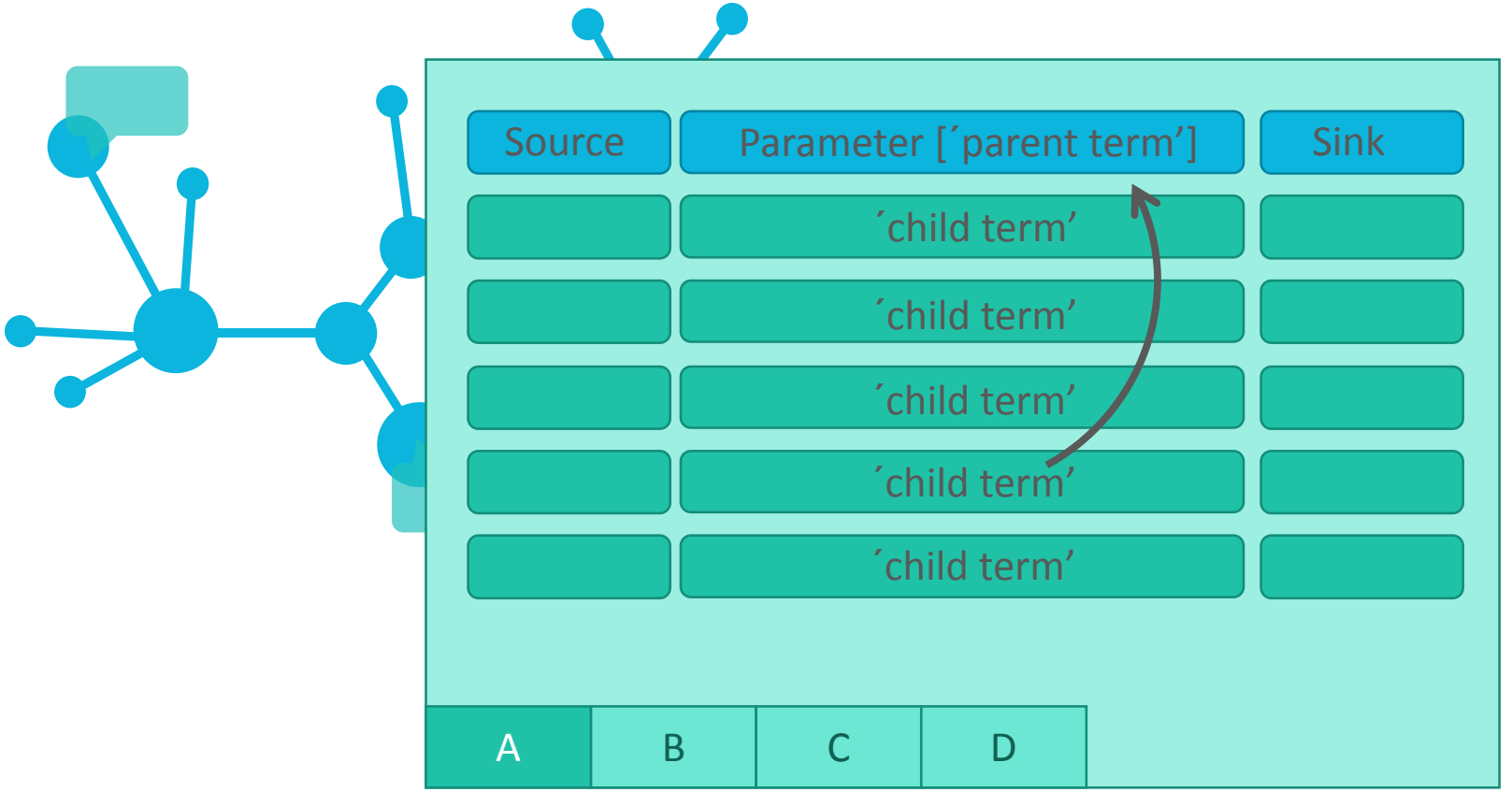


Source Name	Characteristics [Organism]	Sample Name
Ath001	*Arabidopsis thaliana*	Lre001
Ath002	*Arabidopsis thaliana*	Lre002
Ath003	*Arabidopsis thaliana*	Lre003
…	…	…

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

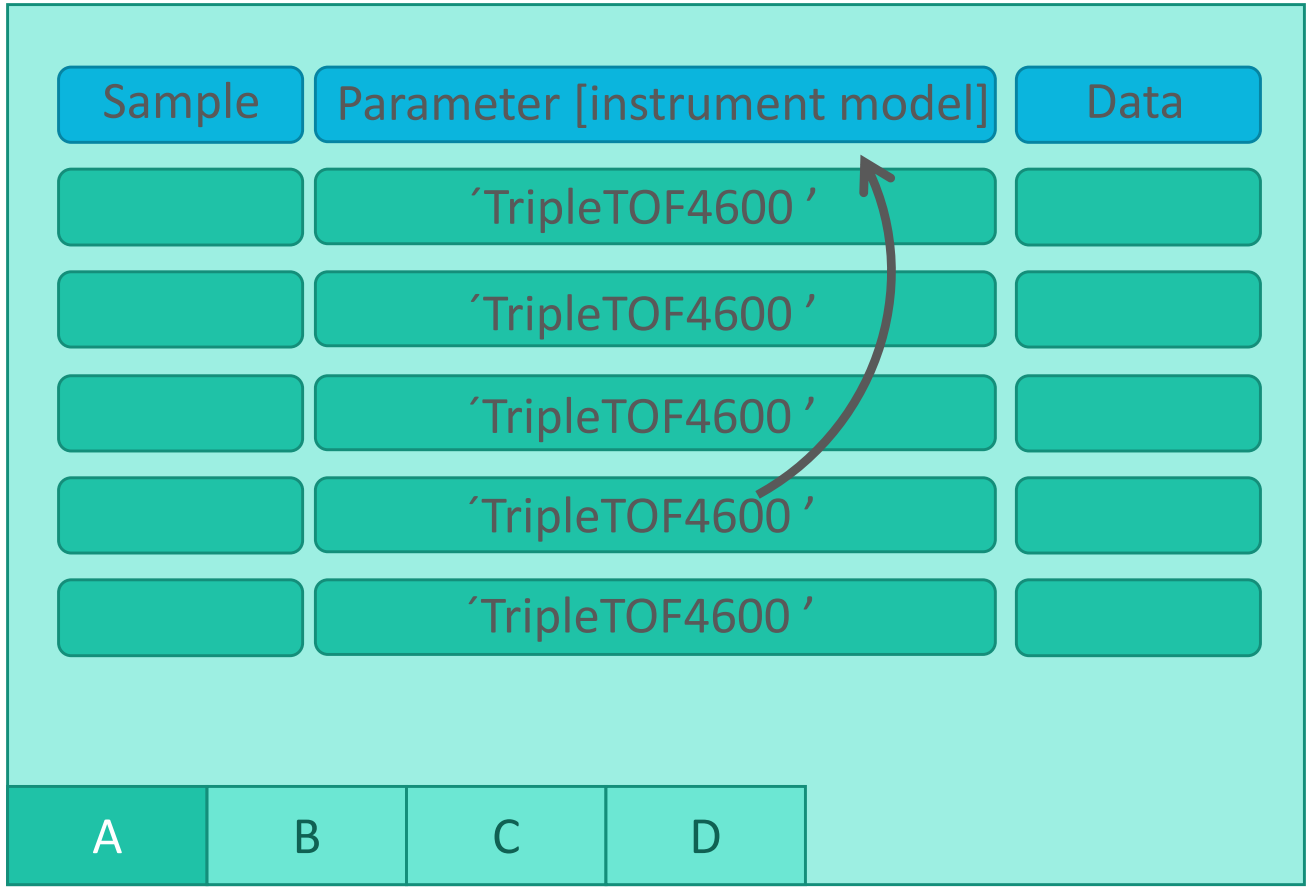
Investigation Title	
Investigation Description	
Investigation Person Last Name	
Investigation Person First Name	
…	…

Annotation by flattening the knowledge graph



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

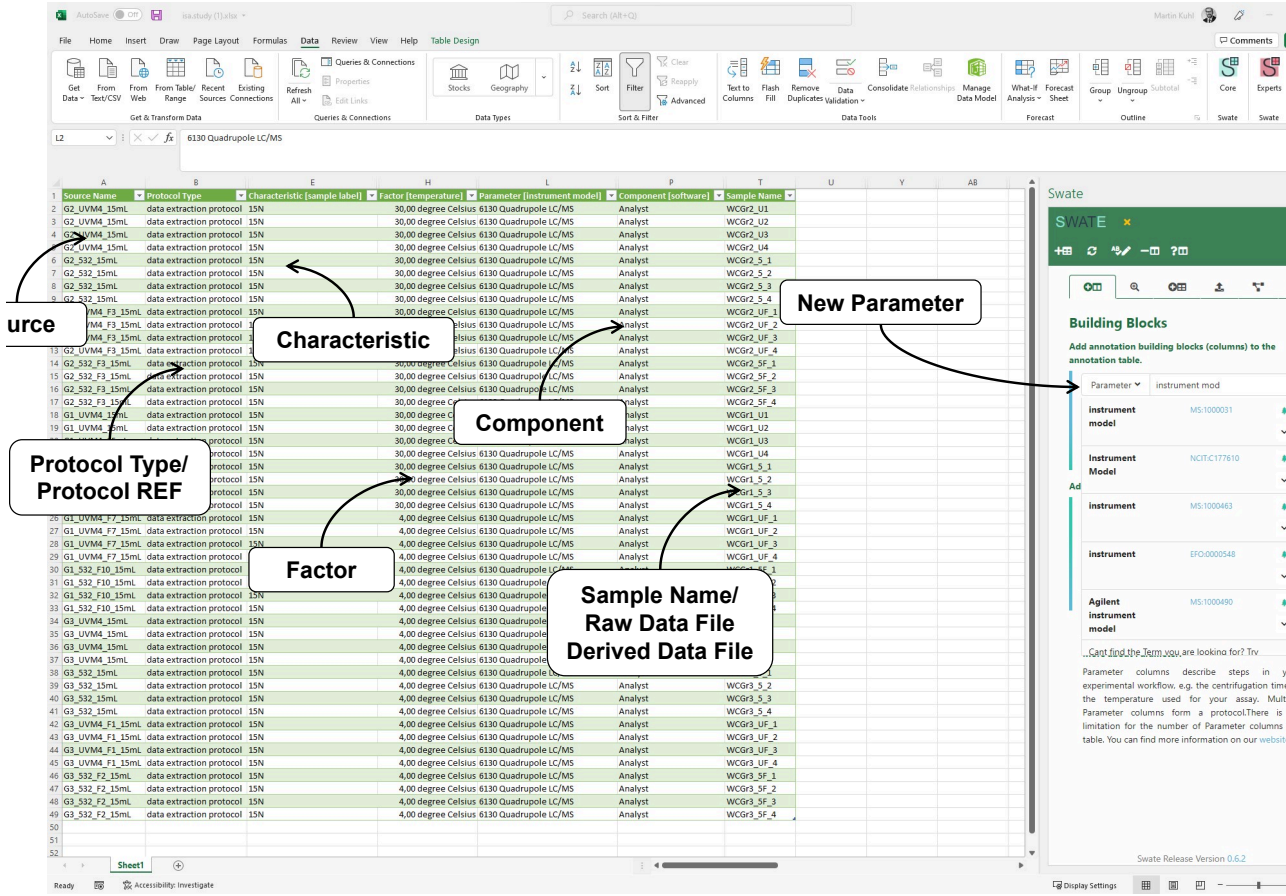
Annotation principle



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

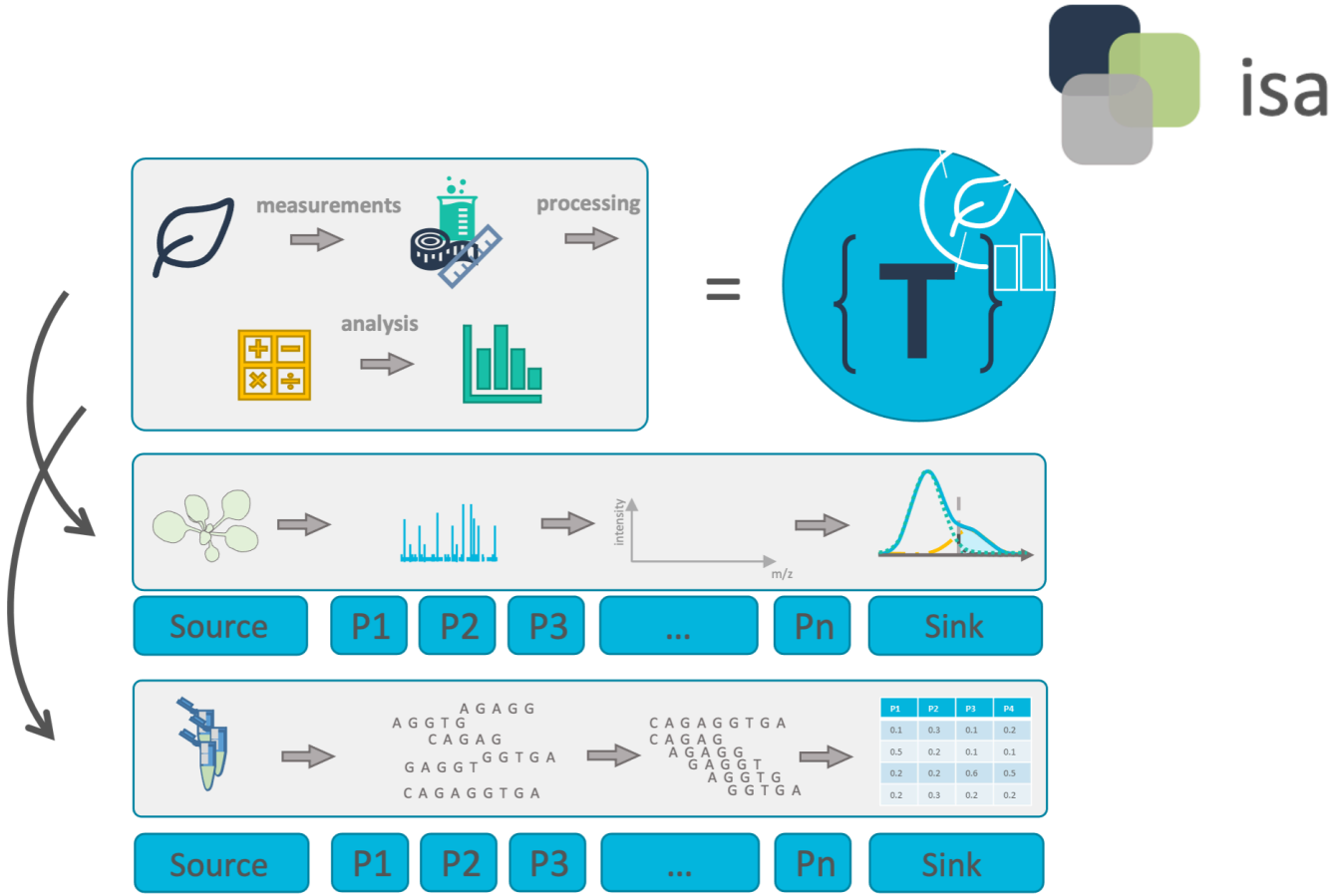
Annotation Building Block types

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



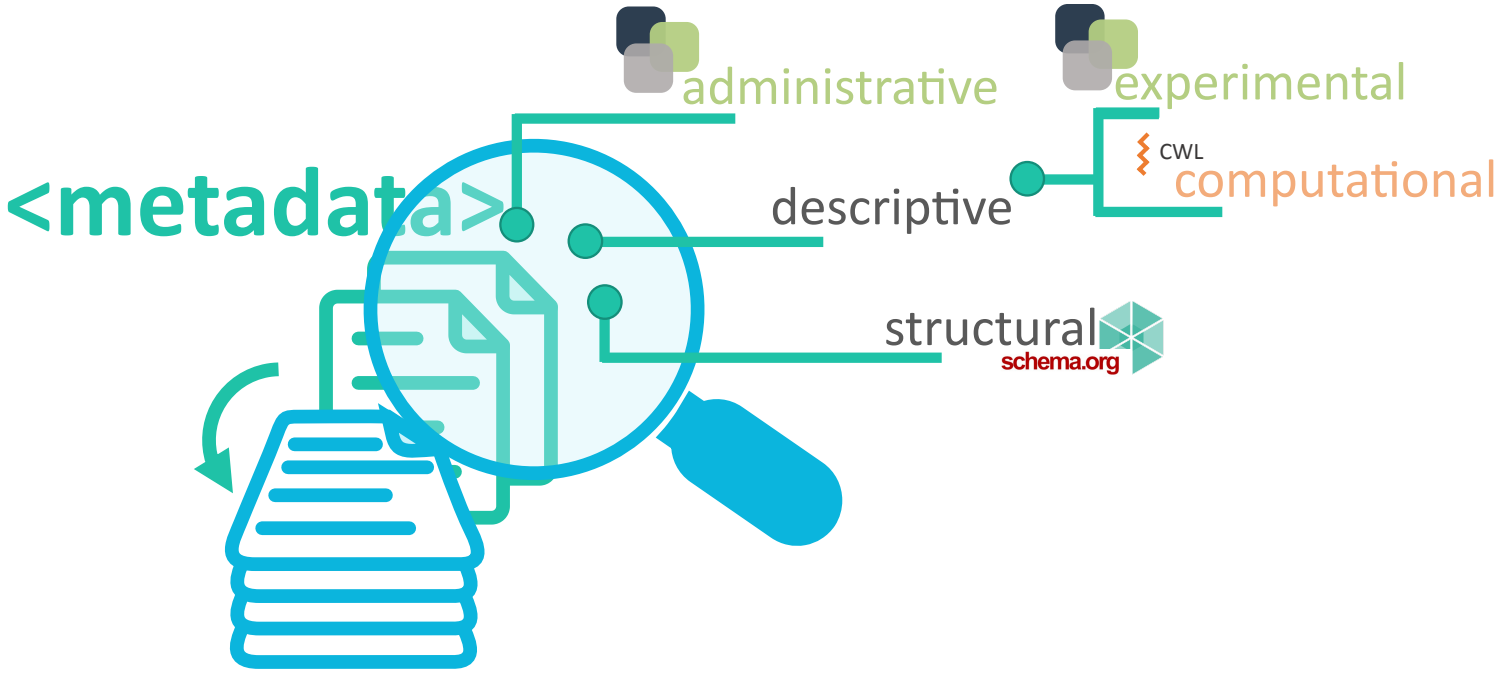
Let's take a detour on Annotation Principles | slides

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

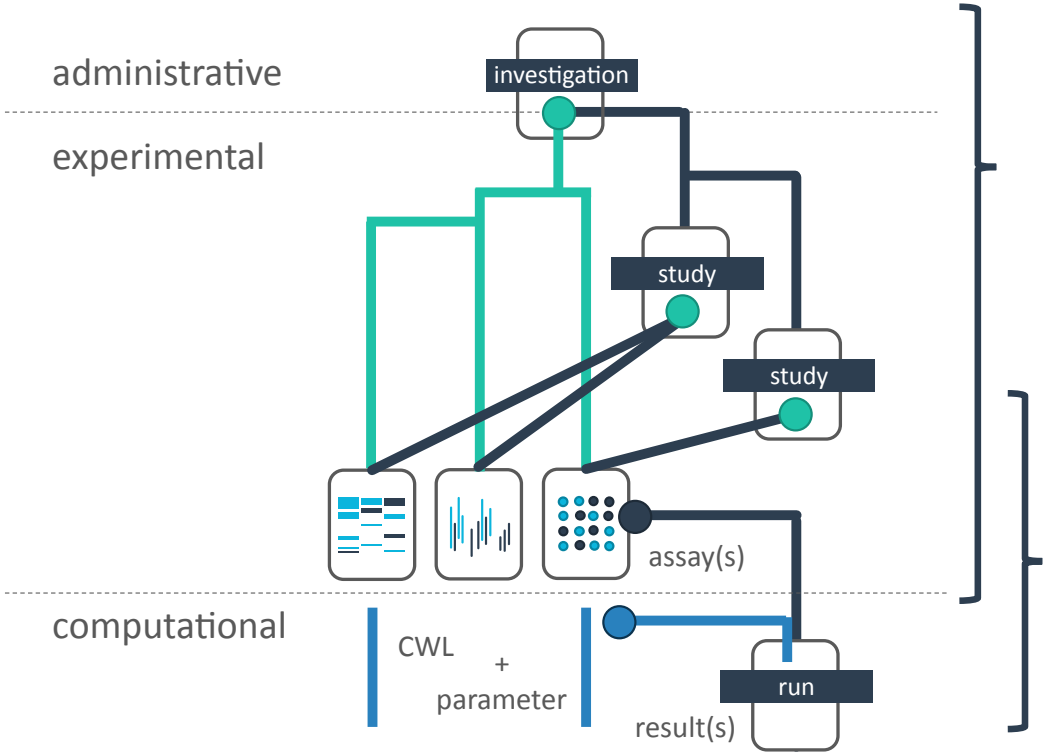
Study
descriptive (meta)data
information on the subject

- Characteristics
- Parameters
- Components
- Factors

Assay
descriptive (meta)data
information on the measurement

- Characteristics
- Parameters
- Components
- Factors

ISA and CWL – Connected by similarity



isa

COMMON WORKFLOW LANGUAGE

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

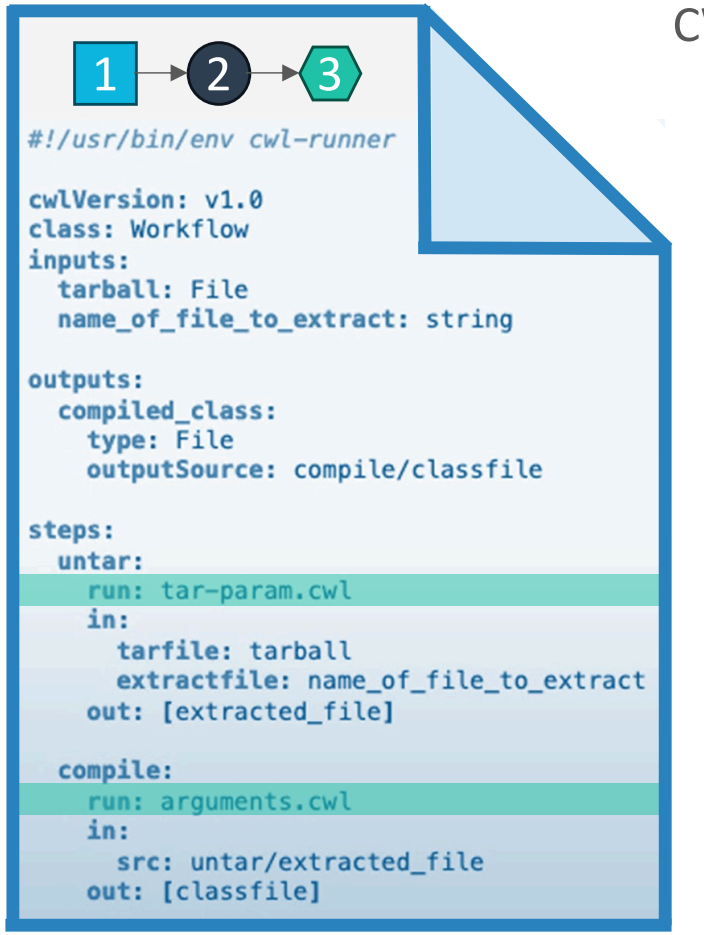
https://www.commonwl.org
https://www.nextflow.io
https://nf-co.re/
https://snakemake.github.io

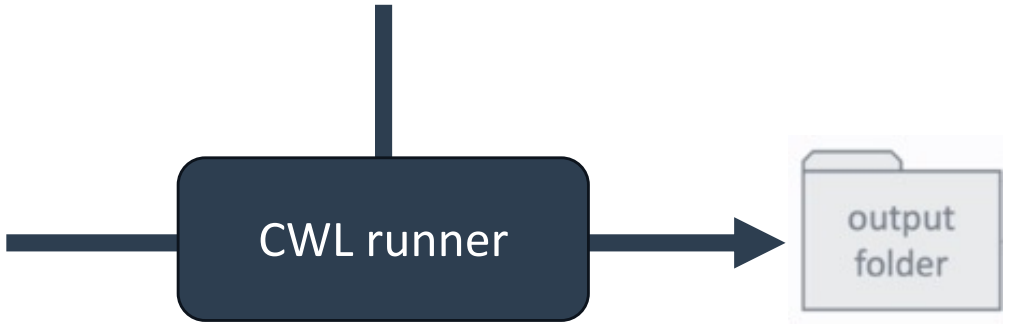
Common Workflow Language

CWL workflow document (*.cwl)

CWL job parameter (*.yaml)

CWL
tool descriptors (*.cwl)





Common Workflow Language

```yaml
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/

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

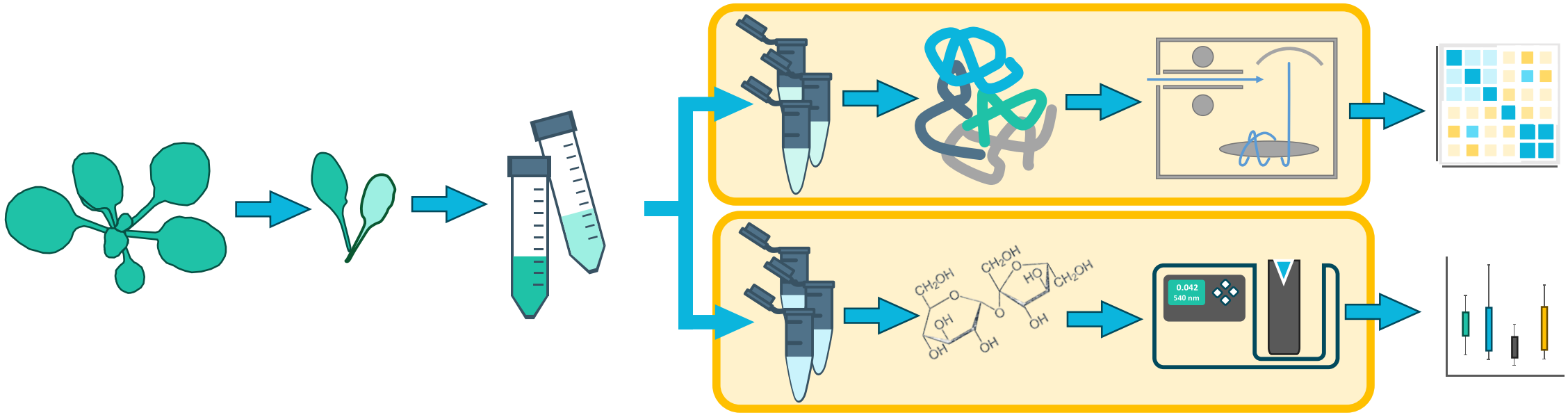
- https://www.bioconductor.org
- BioContainers: https://github.com/BioContainers/

Hands-on part 2: ARCitect (and Swate)

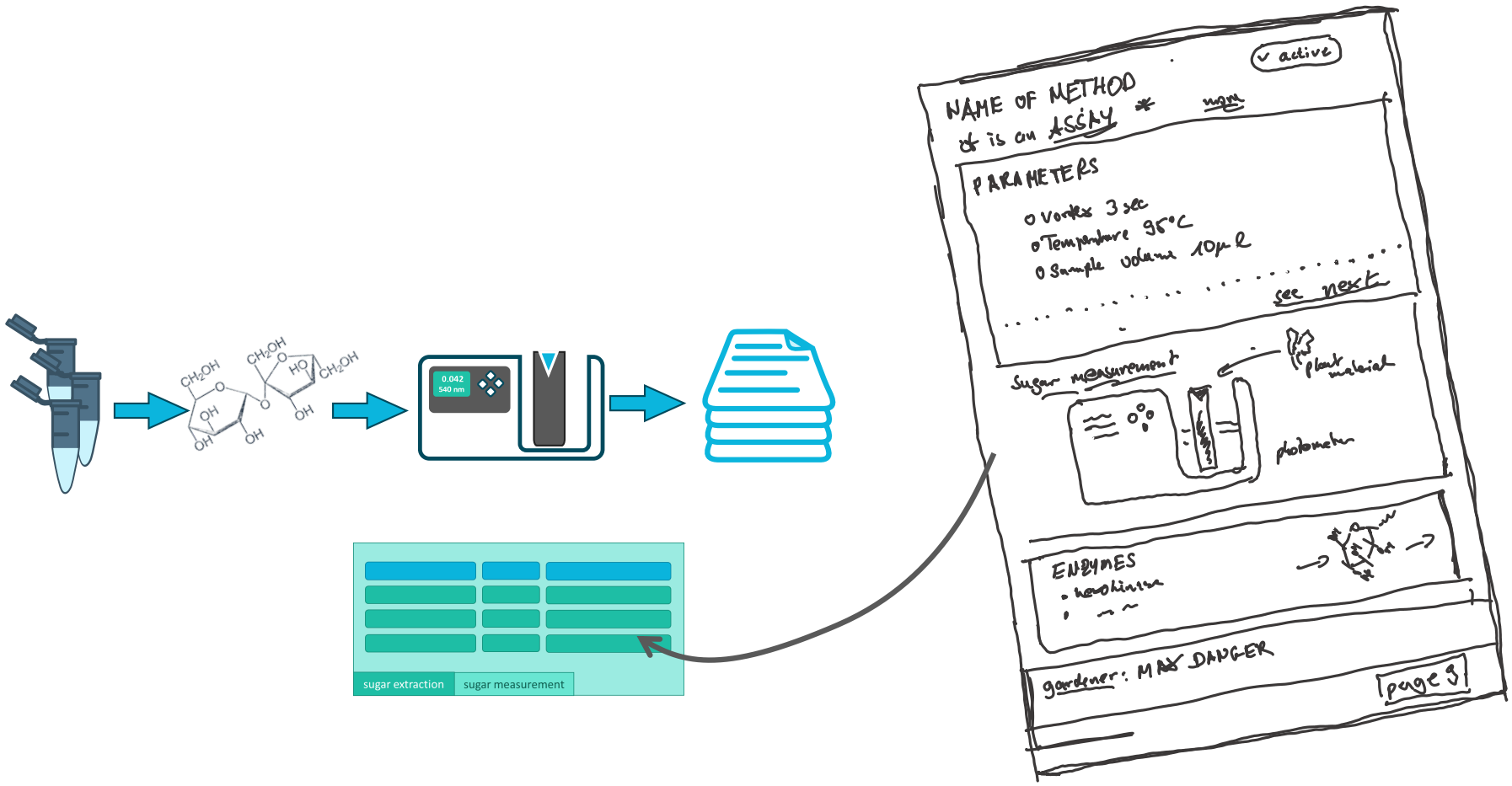
Follow the **Start Here** guide in the DataPLANT knowledge base.

Until "Data analysis"

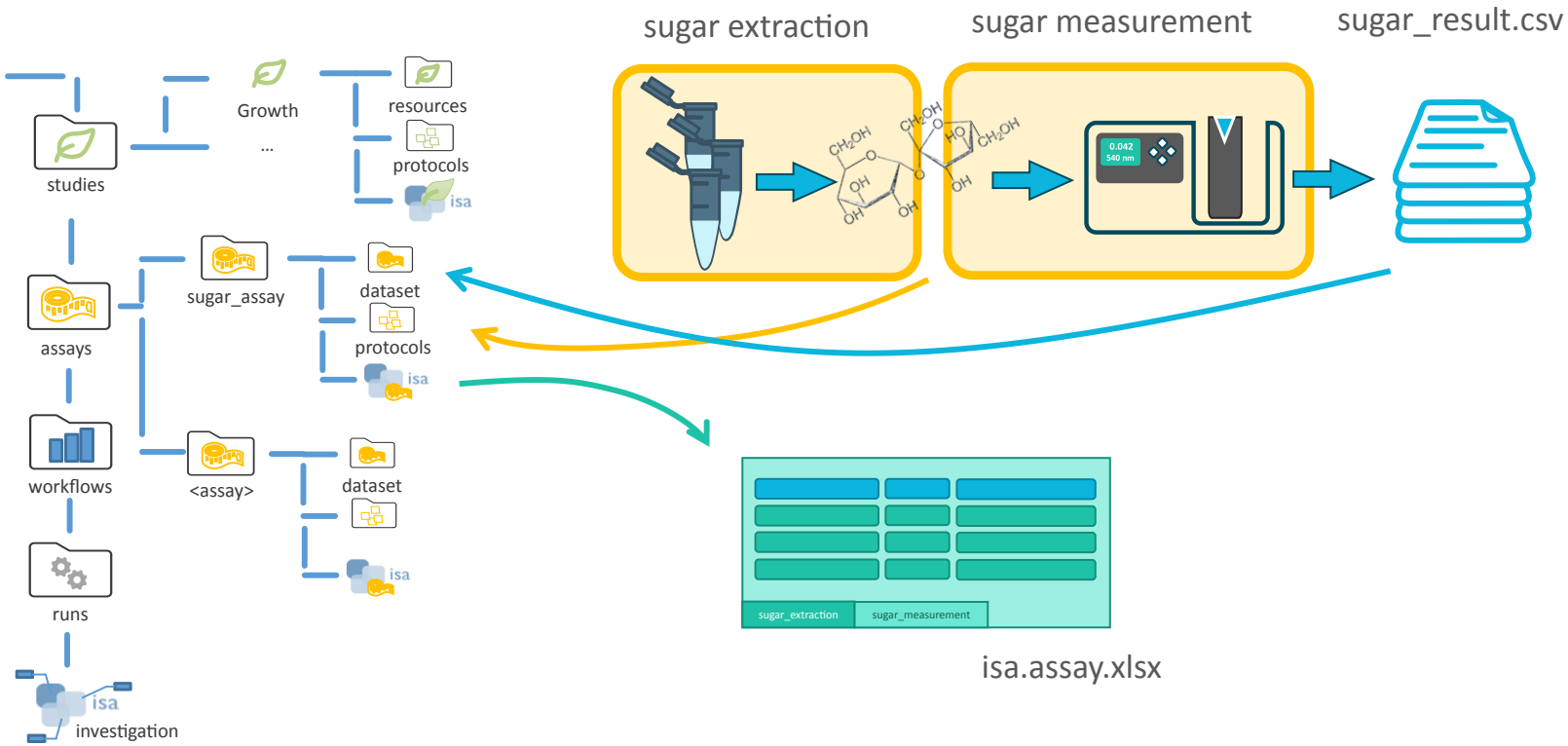
Identifying assays



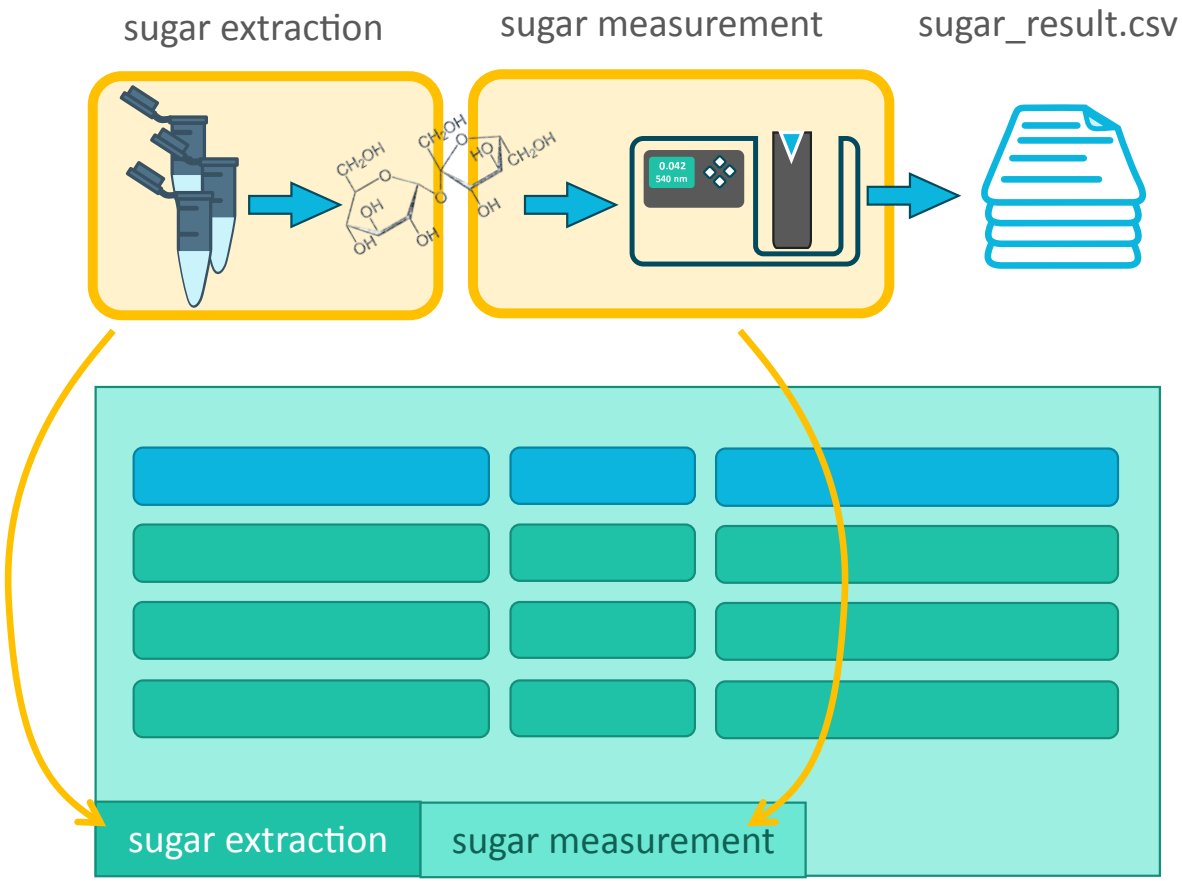
Assay for sugar measurement



Separating different assay elements

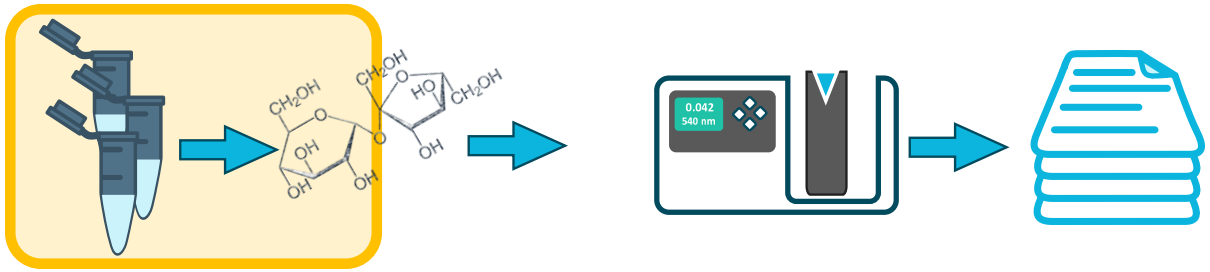


Isolating the lab processes in an assay



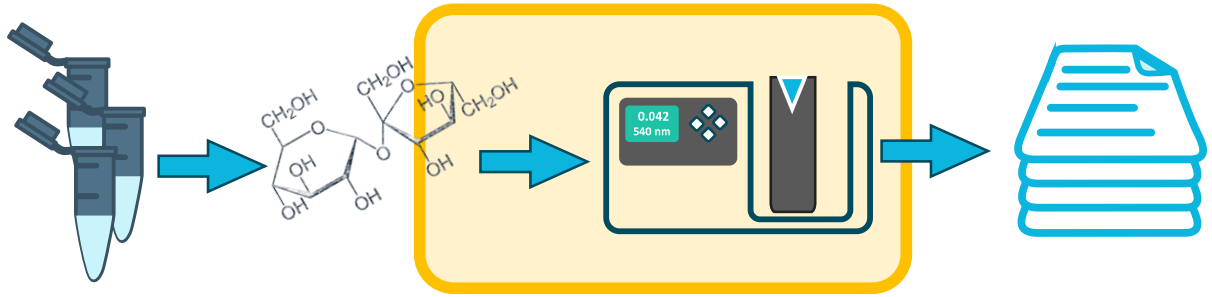
Parameteterization: sugar extraction

- 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



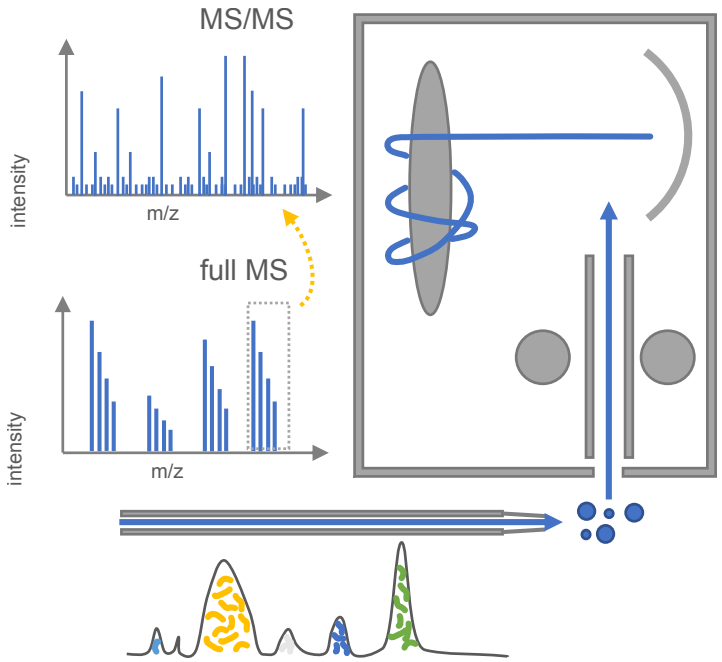
Save time using standard methods and SOPs

Parameter []

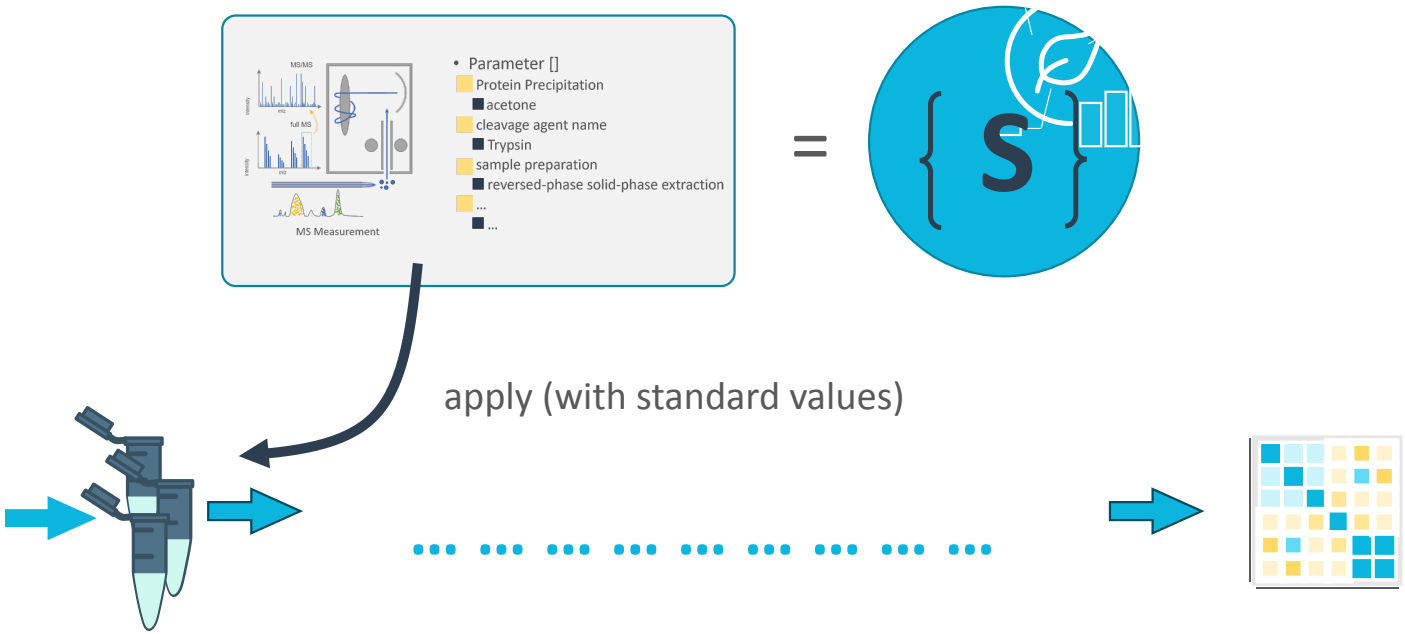
- Protein Precipitation
  - acetone
- cleavage agent name
  - Trypsin
- sample preparation
  - reversed-phase solid-phase extraction
- ...

Component []

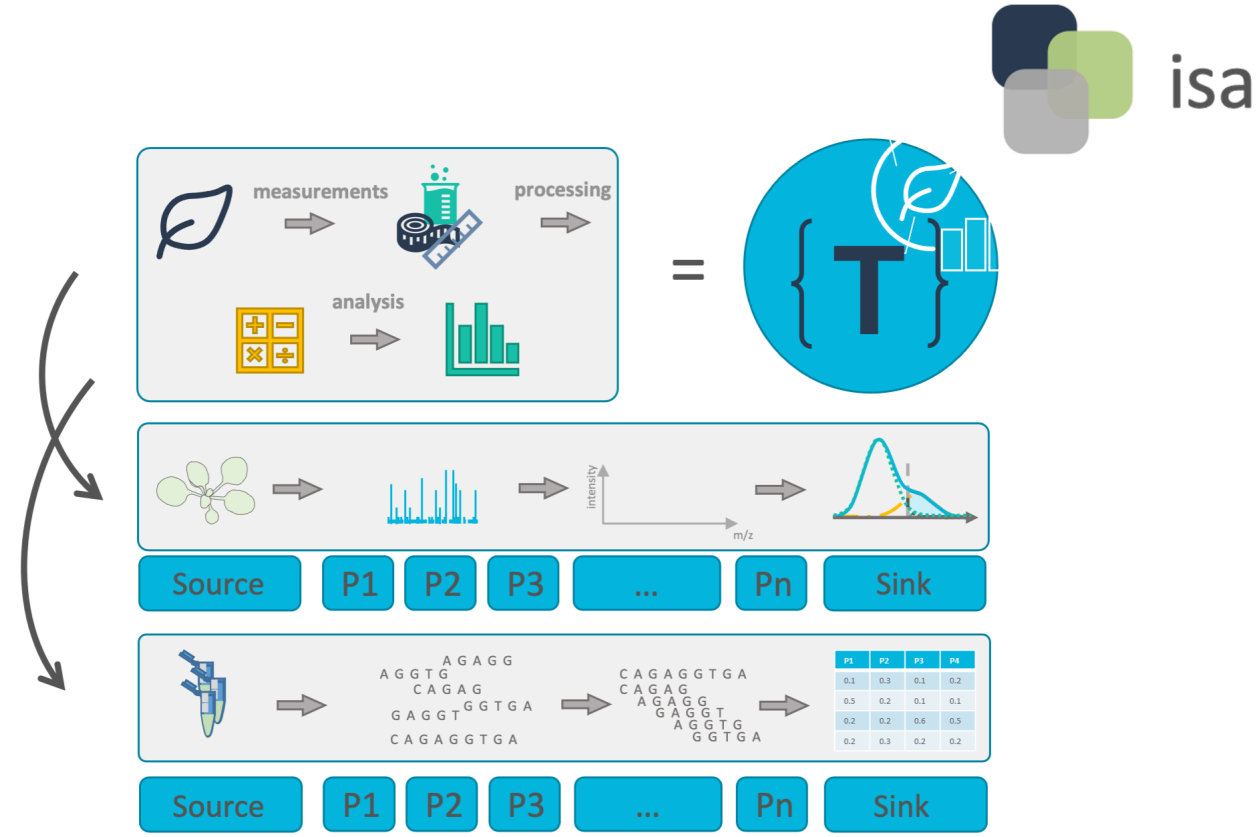
- chromatography instrument model
  - nanoElute2
- chromatography column model
  - PepSep C18 1.9u, 25cm x 75um
  - ...



Applying standard procedures to sample record

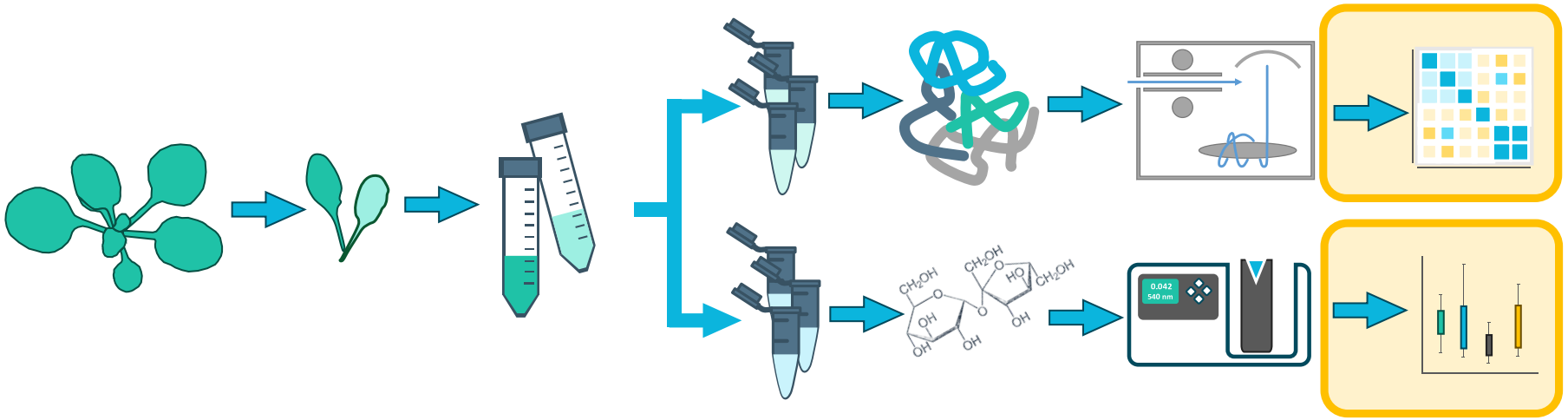


Realization of lab-specific metadata with templates

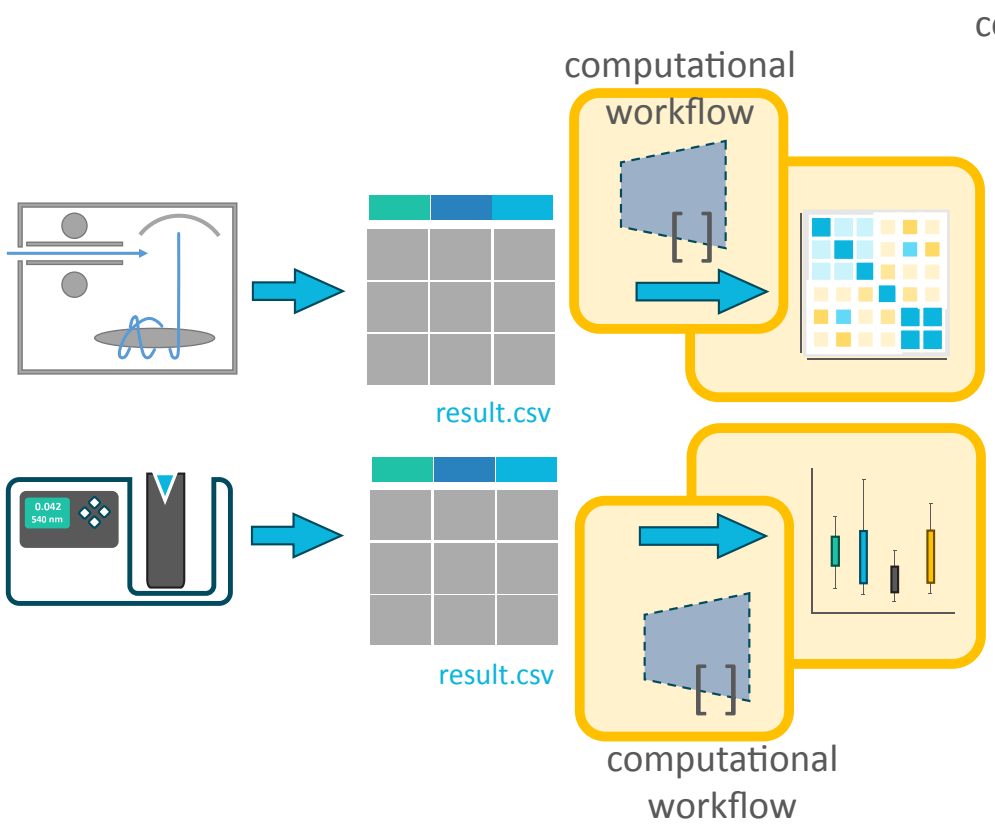


Facilities can define their most common workflows as templates

Options to annotate the data analysis



A computational workflow is like a protocol

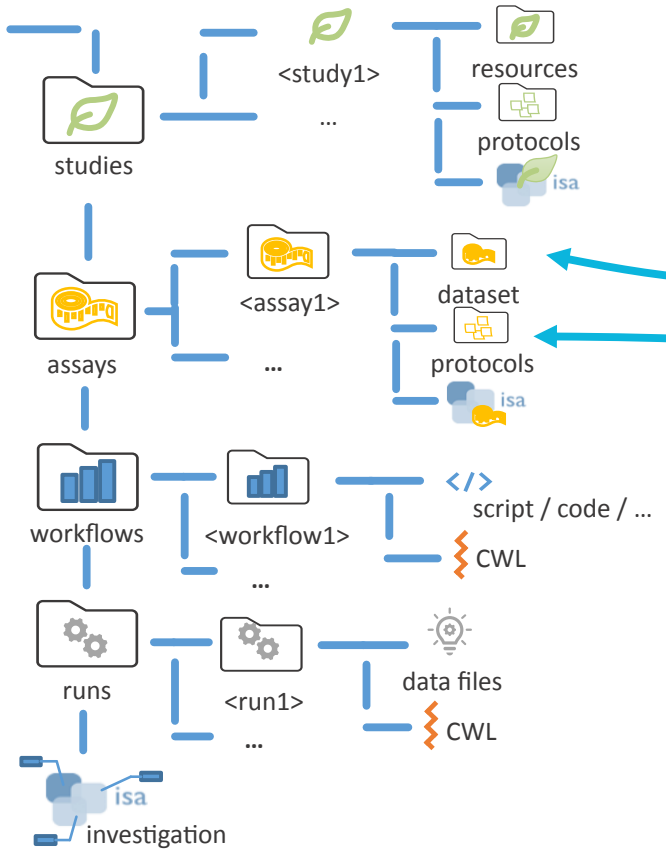


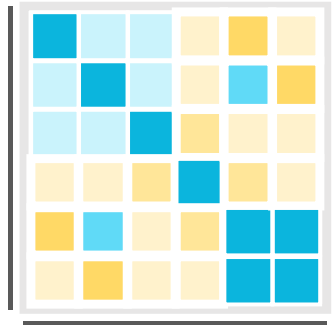
computational
workflow

```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()
```

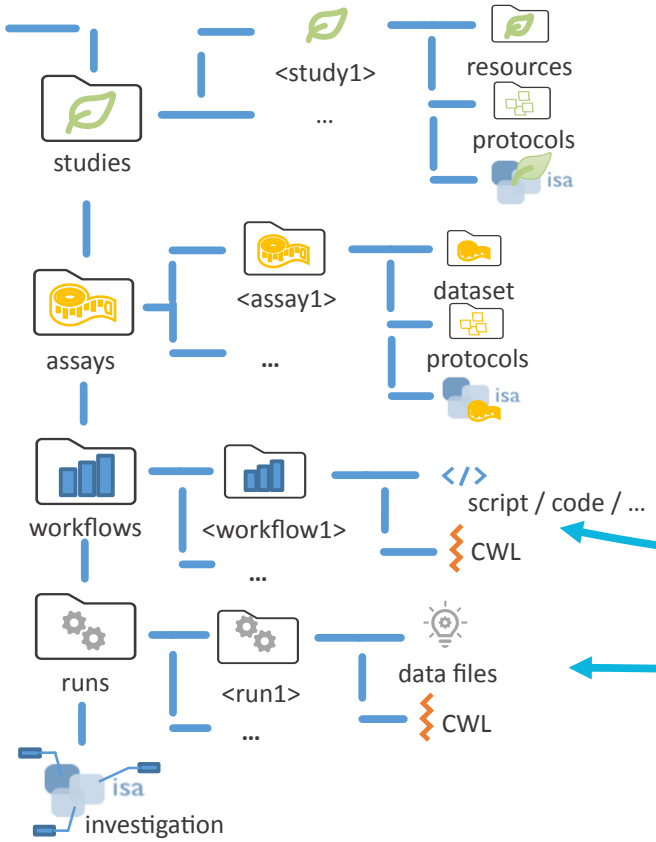
heatmap.py

Option I: Create a virtual assay



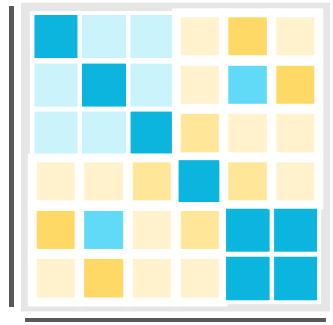


Option II: Create a workflow and run




```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()
```

heatmap.py

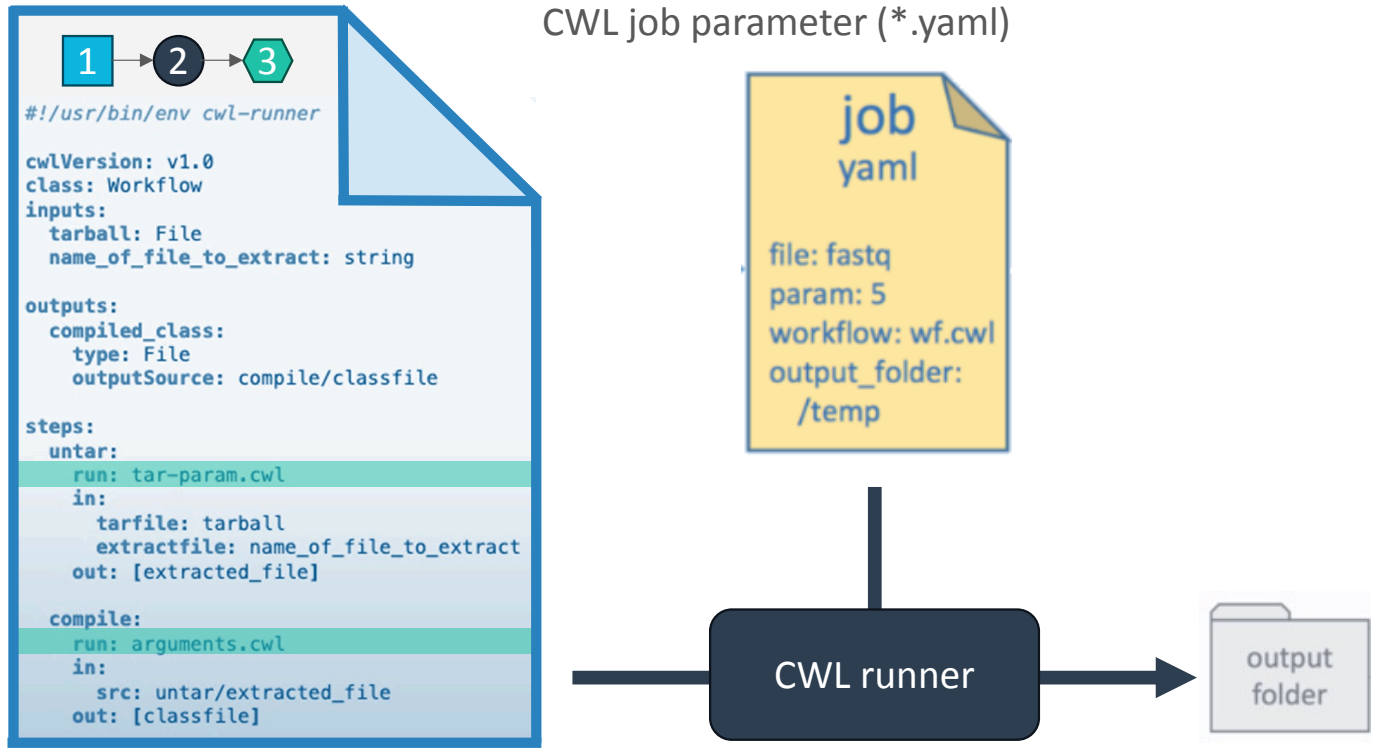


Use CWL to wrap your workflow

CWL workflow document (*.cwl)

CWL job parameter (*.yaml)

CWL
tool descriptors (*.cwl)



Q&A and Wrap-up Day1

Preparation for next day

- Please try to prepare your own ARC
- Please install SWATE

Resources

 **DataPLANT (nfdi4plants)**

Website: https://nfdi4plants.org/

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

DataHUB: https://git.nfdi4plants.org

GitHub: https://github.com/nfdi4plants

HelpDesk: https://helpdesk.nfdi4plants.org

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

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/0NdPKO5LQ5CHBRN2iuG91Q

Points to discuss from and since day 1

Data Storage and Versioning

Data stores

Institute server

University server

Local hard disks

Cloud services

sciebo
dieCampuscloud

Chat

Your data

Mail

Wiki, Project management

eLabFTW

labfolder

Electronic lab notebooks

Git repositories

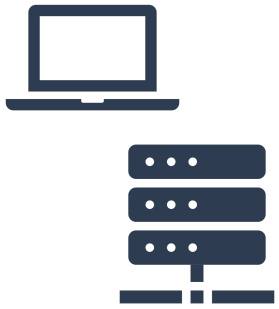
Backup vs. Archive

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

3-2-1 backup rule

3 copies of data

2 storage media



1 copy off-site

Version control and track changes

It’s good practice to document:

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

Types of Version Control

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

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

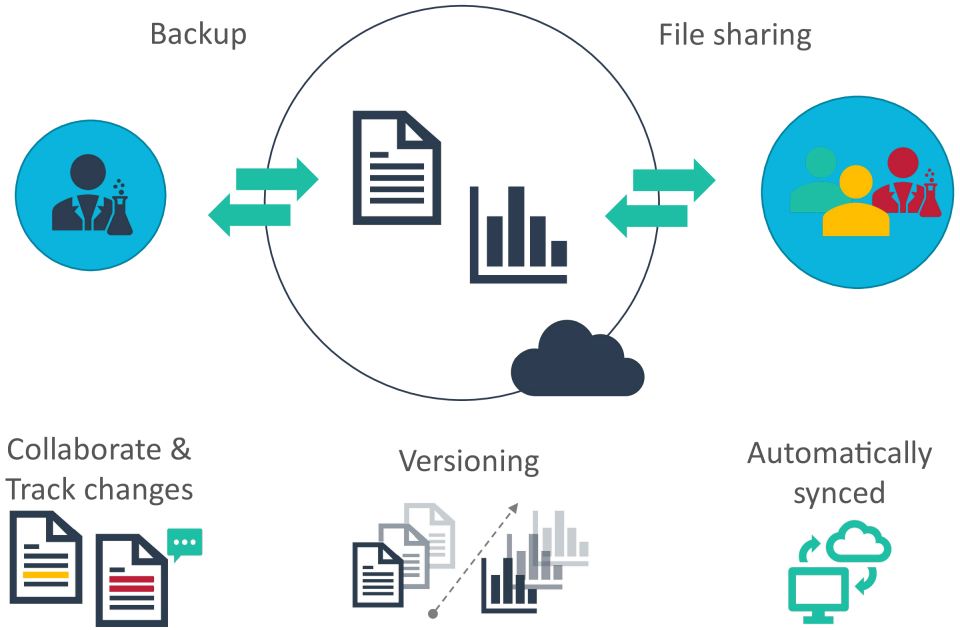
✓ Small data

✓ 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

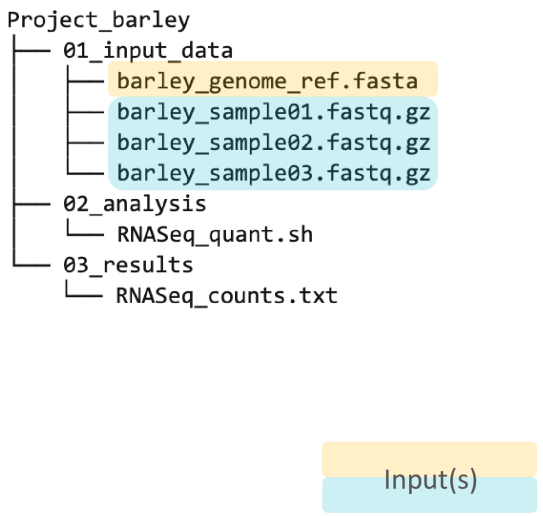
Why git? ≈> Why code?

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

A simple example: RNASeq project

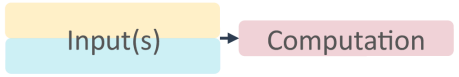
```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
├── 02_analysis
│   └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```

A simple example: RNASeq project



A simple example: RNASeq project

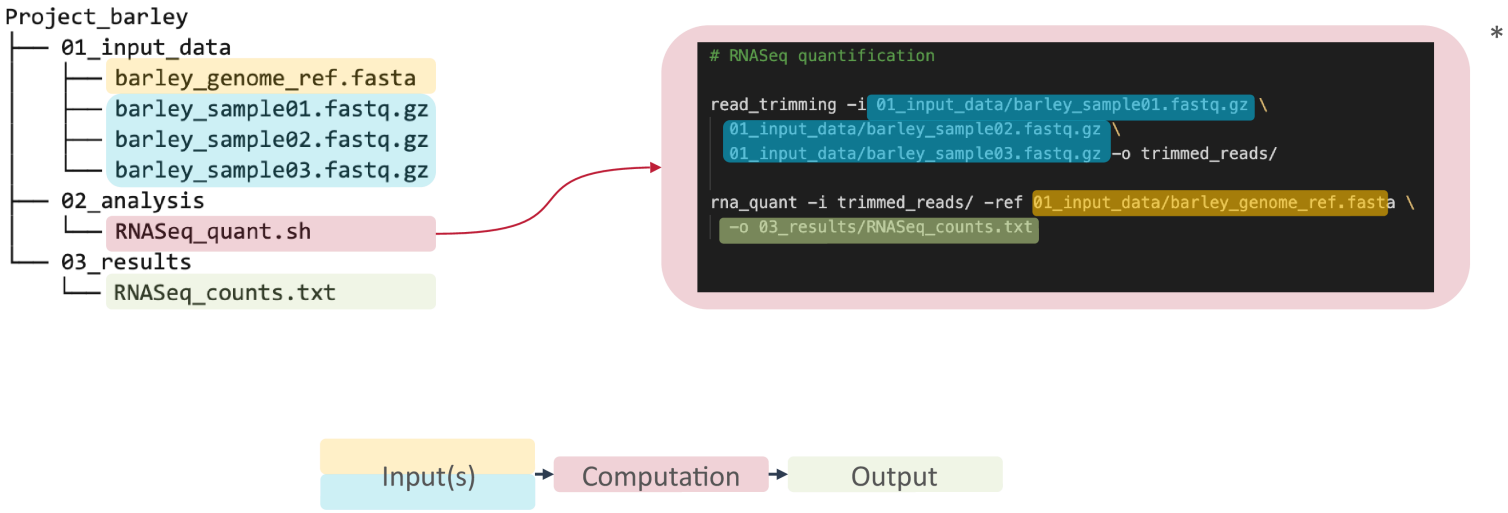
```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   └── barley_sample03.fastq.gz
├── 02_analysis
│   └── RNASeq_quant.sh
└── 03_results
    └── RNASeq_counts.txt
```



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
```

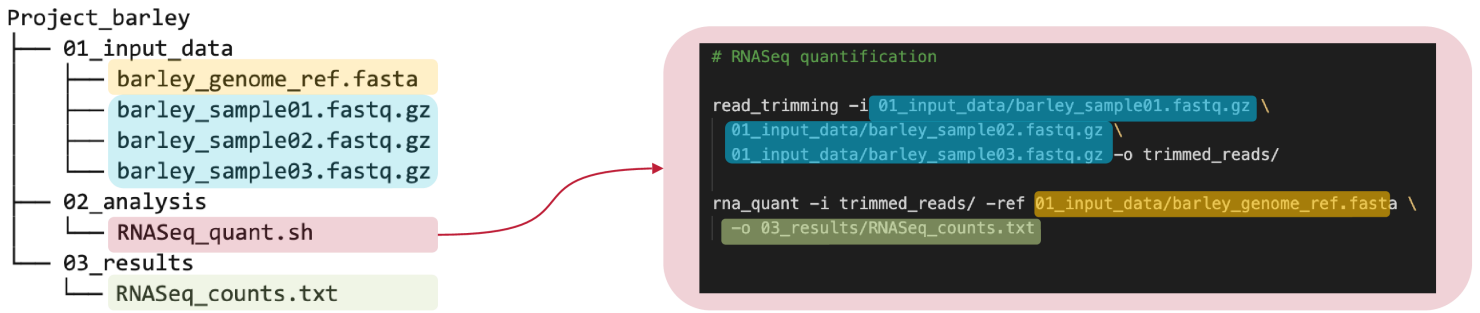
A simple example: RNASeq project



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

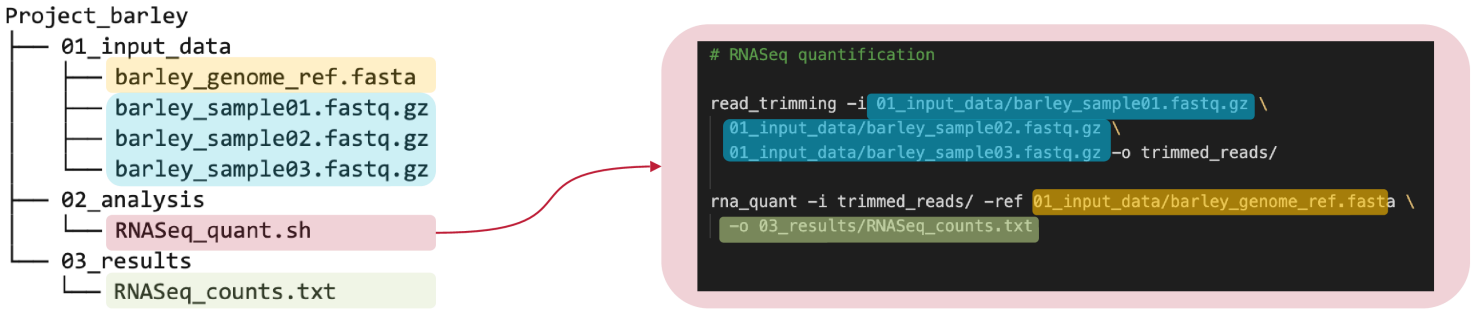
Take snapshots of your code work…

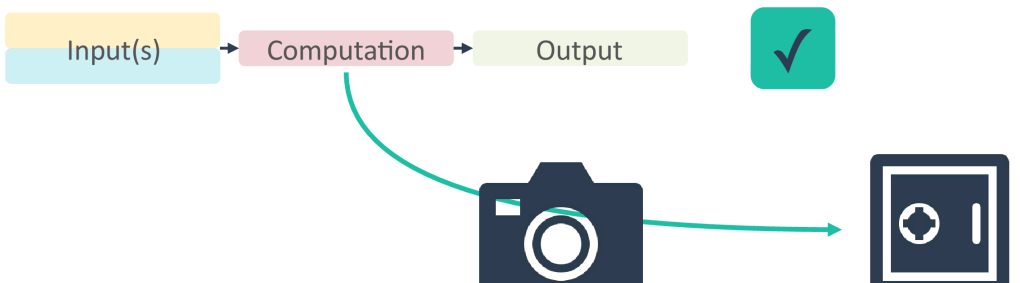
(… as long as it works)



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
```

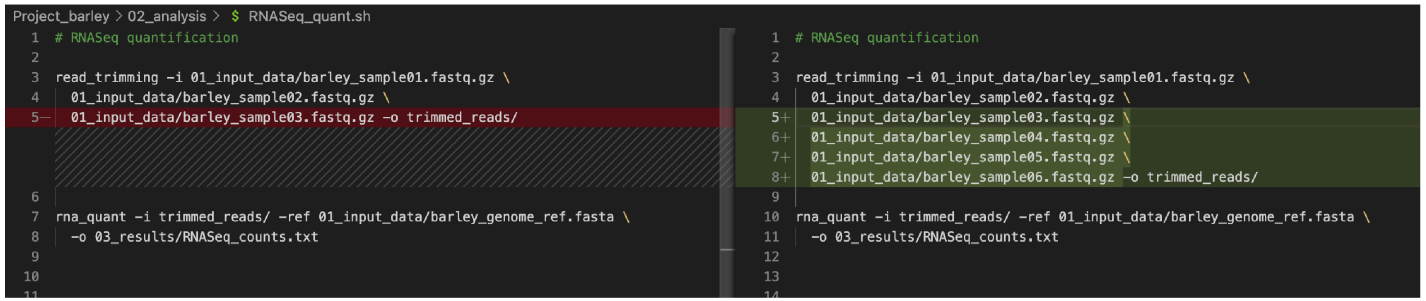
```
Project_barley
├── 01_input_data
│   ├── barley_genome_ref.fasta
│   ├── barley_sample01.fastq.gz
│   ├── barley_sample02.fastq.gz
│   ├── barley_sample03.fastq.gz
│   ├── barley_sample04.fastq.gz
│   ├── barley_sample05.fastq.gz
│   └── barley_sample06.fastq.gz
├── 02_analysis
│   ├── RNASeq_quant.sh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



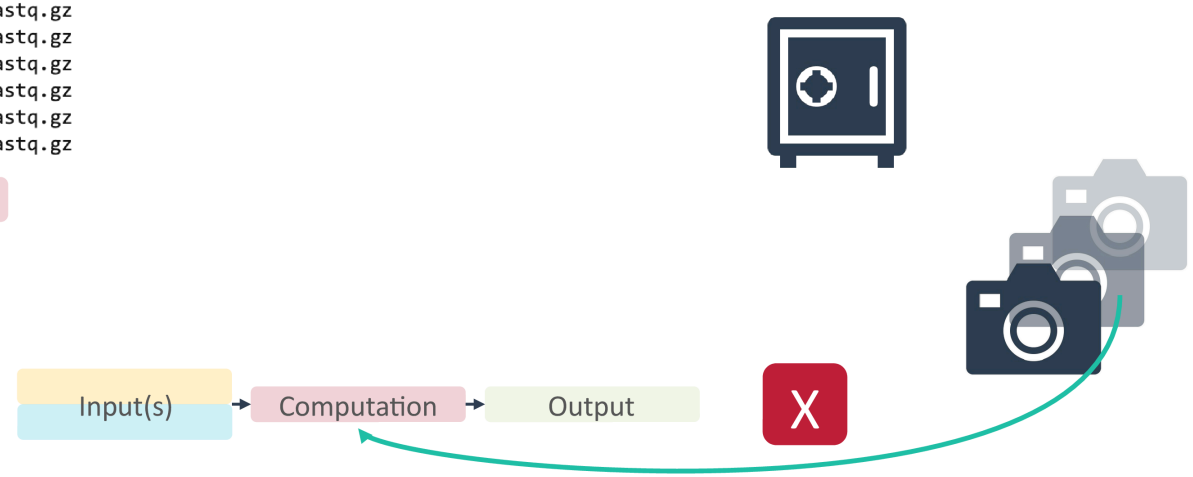
“version 1” “version 2”

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
```

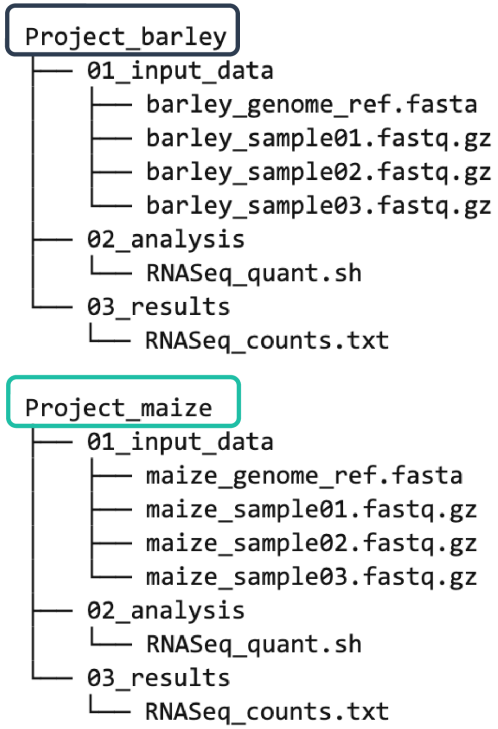
Revert to snapshot

```
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
```



“version xy”

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



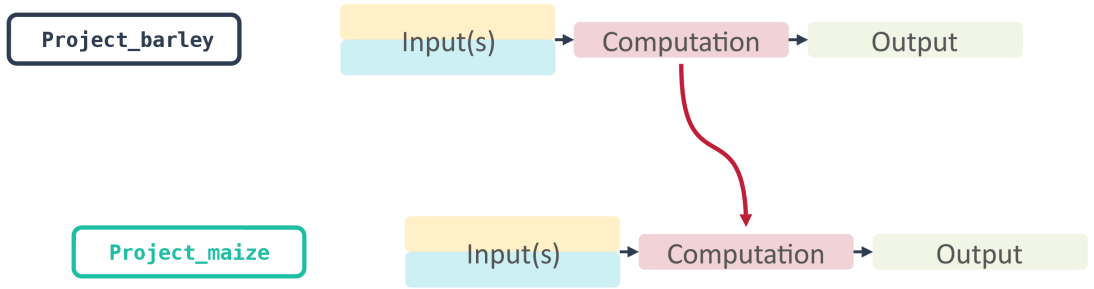
Scenario 3: New project, same type of data and analysis

```
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
```

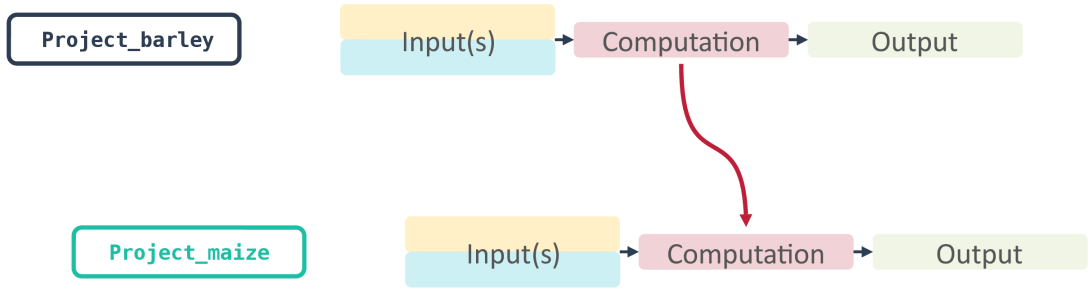
```
# RNASeq quantification

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


Project_barley
Input(s)
Computation
Output
Project_maize
Input(s)
Computation
Output

Re-use code



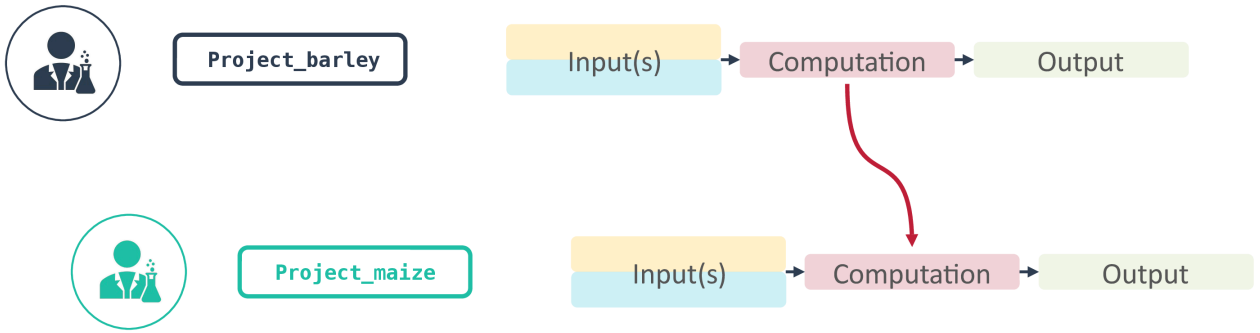

```
# RNASeq quantification

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
```

“version barley”

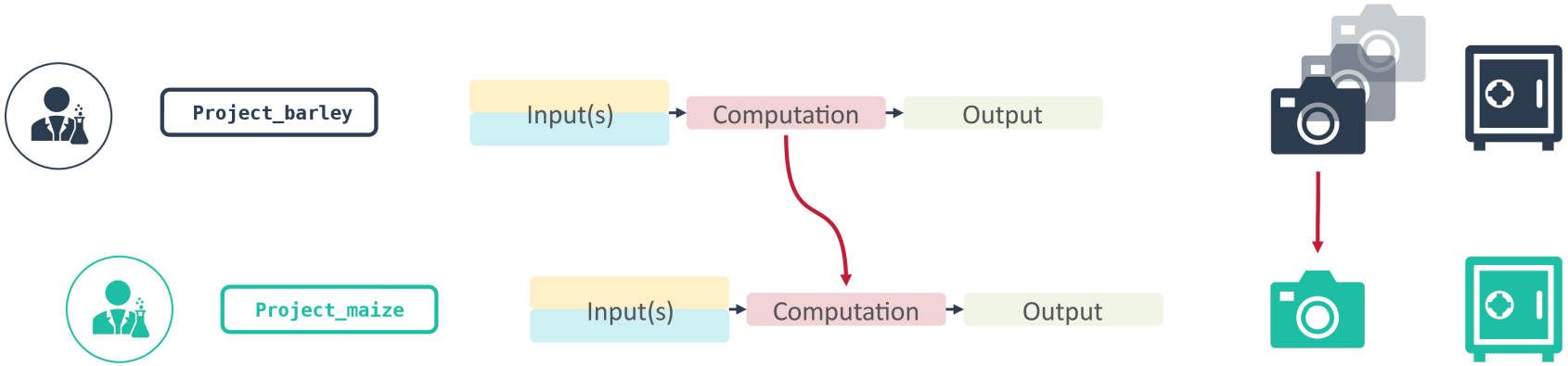
```
# RNASeq quantification

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
```

“version maize”

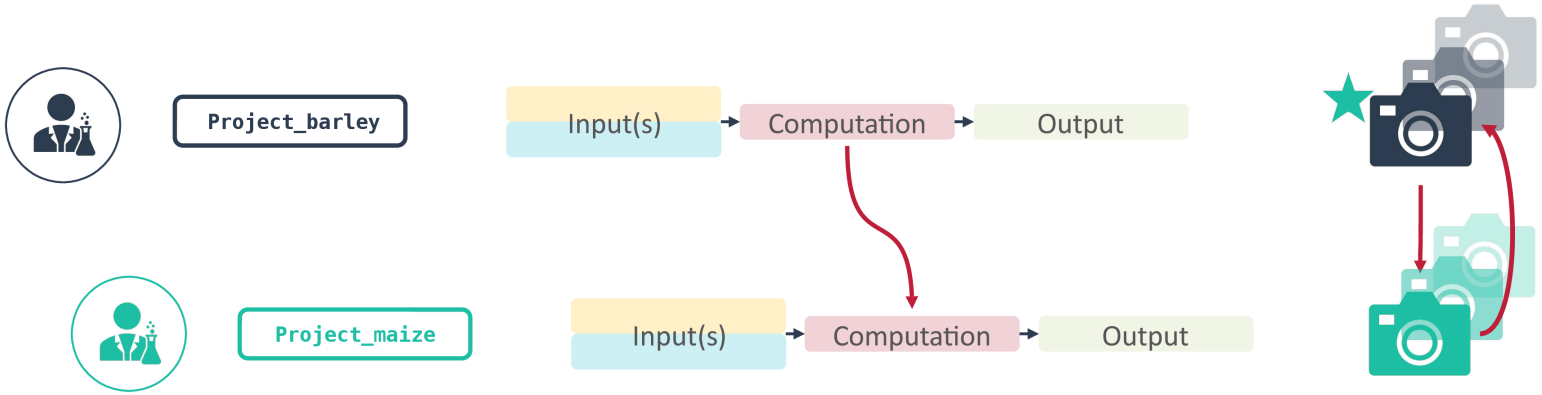
Re-use code – People have done this



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

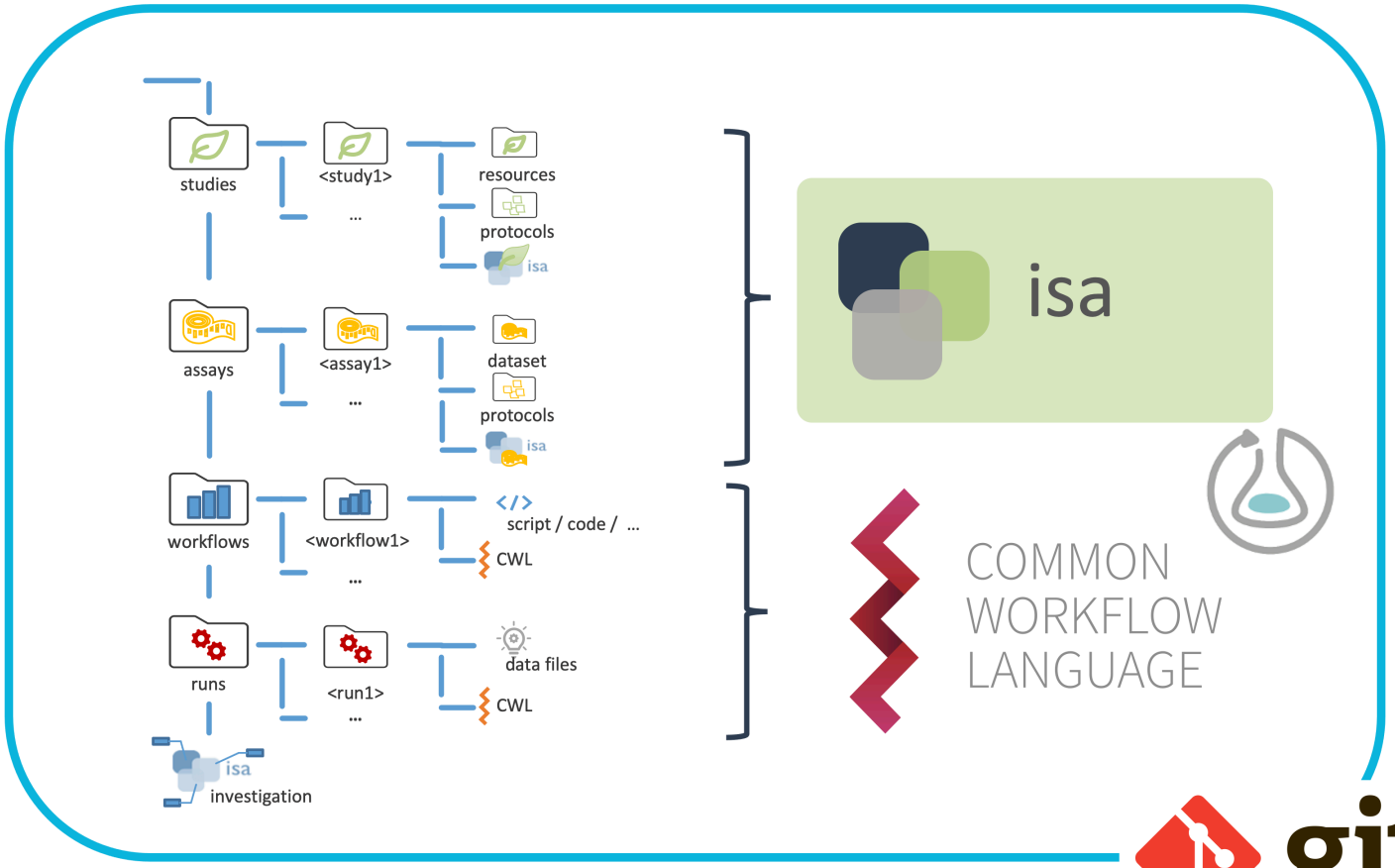
- Online service to host our projects
- Share code with other developers
- Others can download our projects, work on and contribute to them
- They can upload their changes and merge them with the main project

Cloud vs. Git

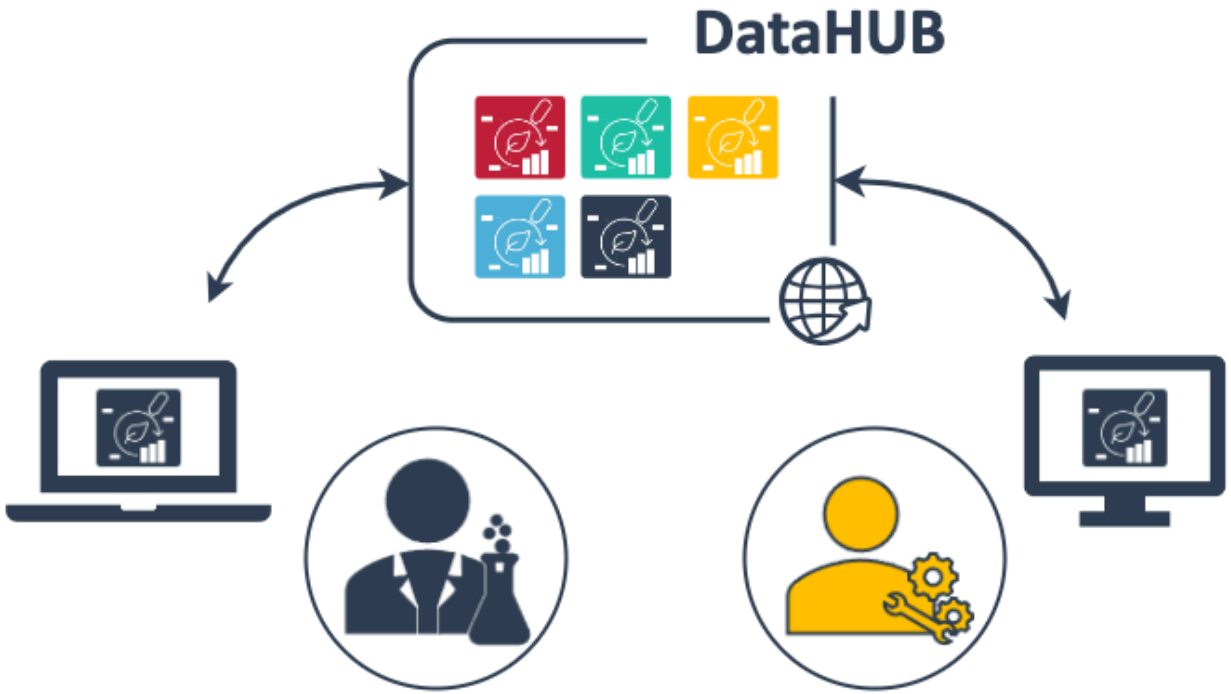
	Cloud services	Git / GitHub / GitLab
Track changes	✓ Documents ✓ Small data ✓ Presentations	✓ Code ✓ Data analytical projects
Collaboration		issue tracker, tracked contribution
Versioning	Automated	Well-documented (commit history)
Syncing	Automated	Active / controlled by user
Access	Oftentimes only within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

DataPLANT DataHUB

ARC builds on standards + Git

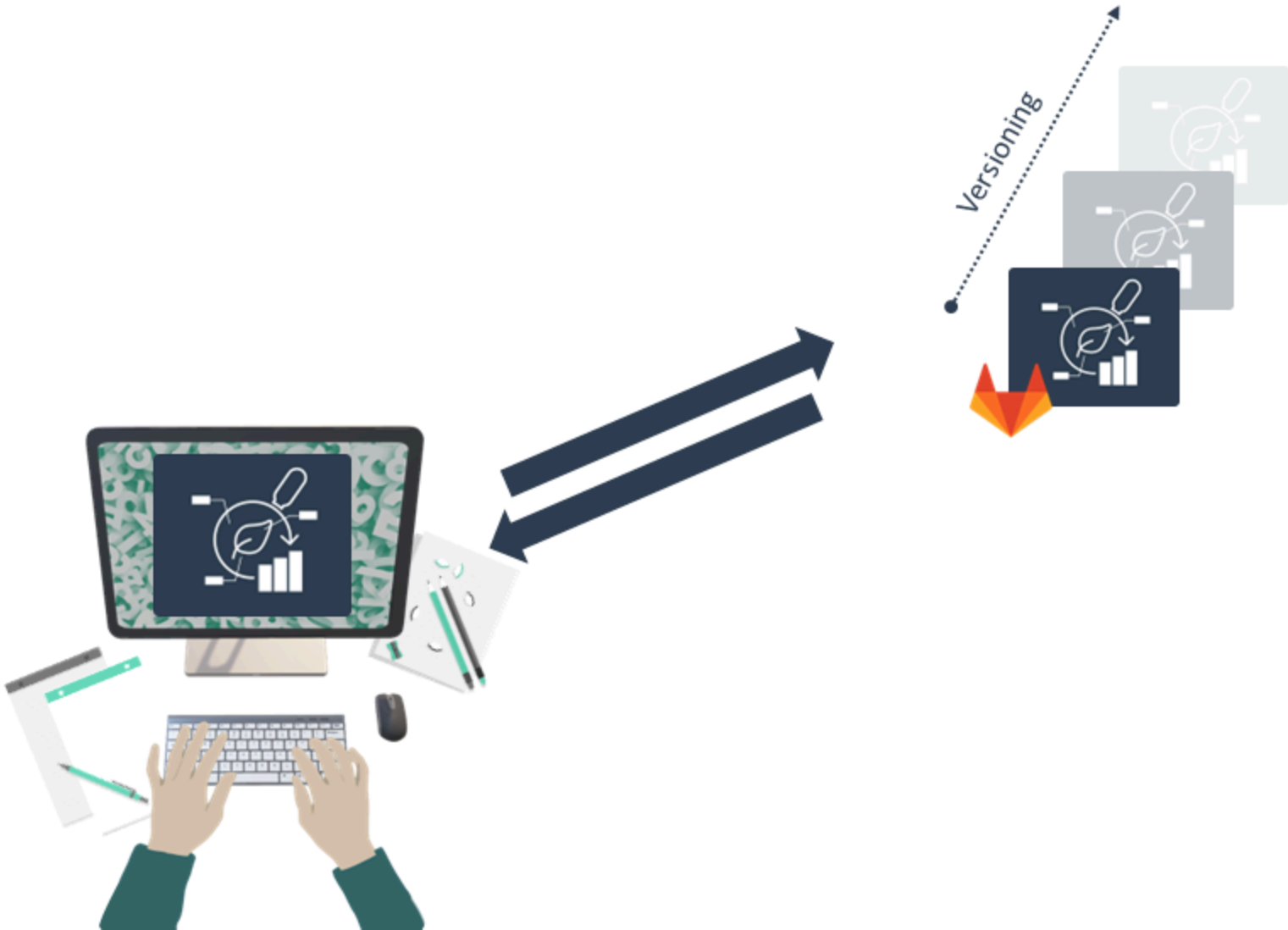


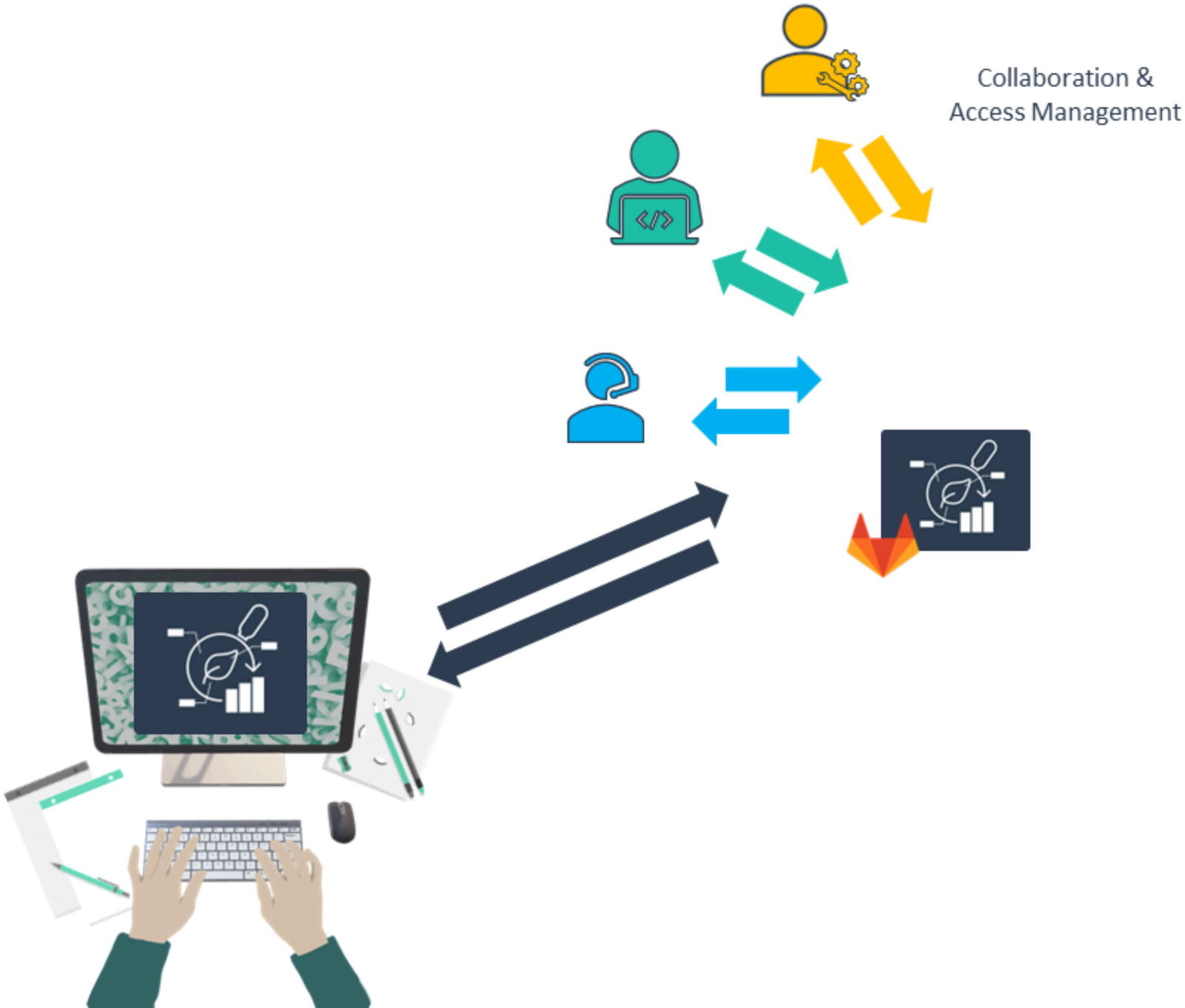
The DataPLANT DataHUB – a GitLab *Plus*

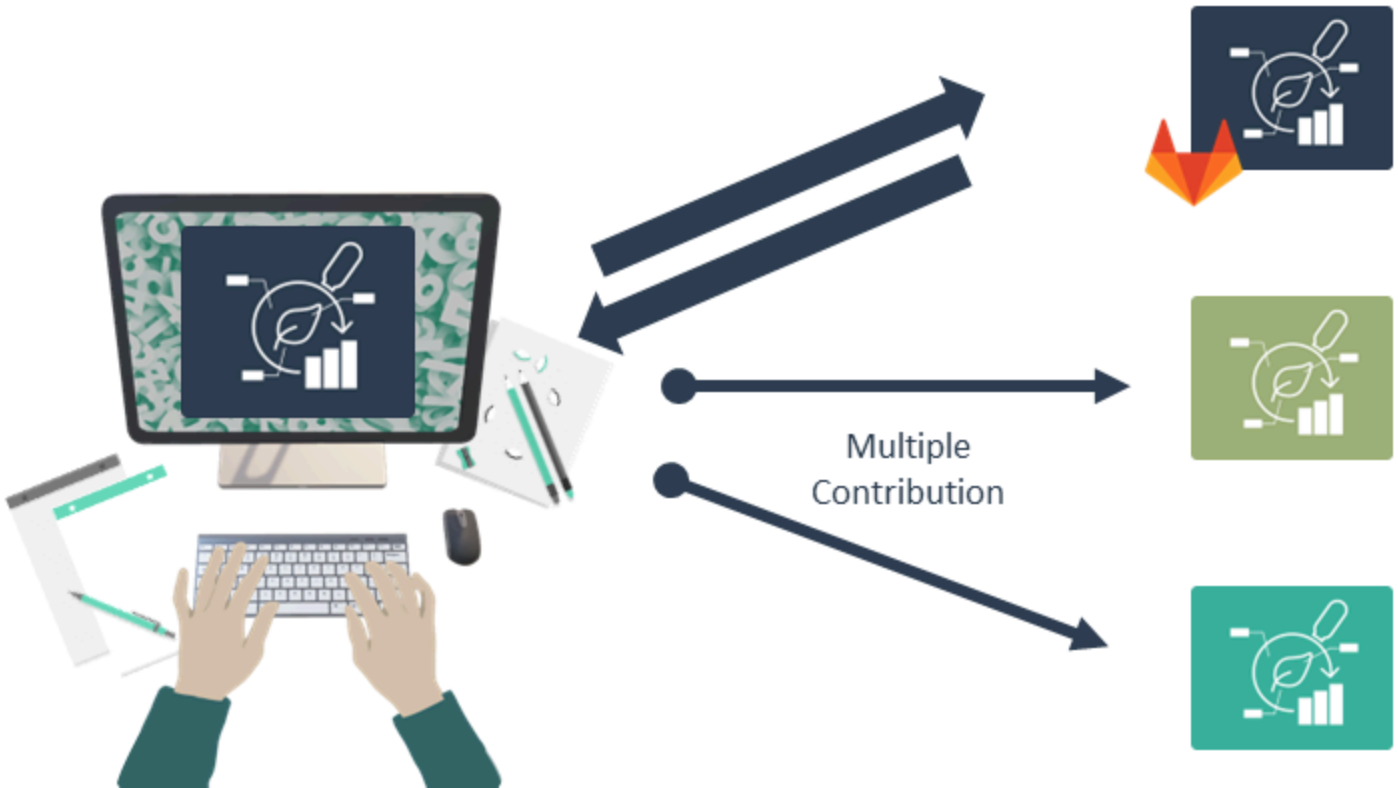


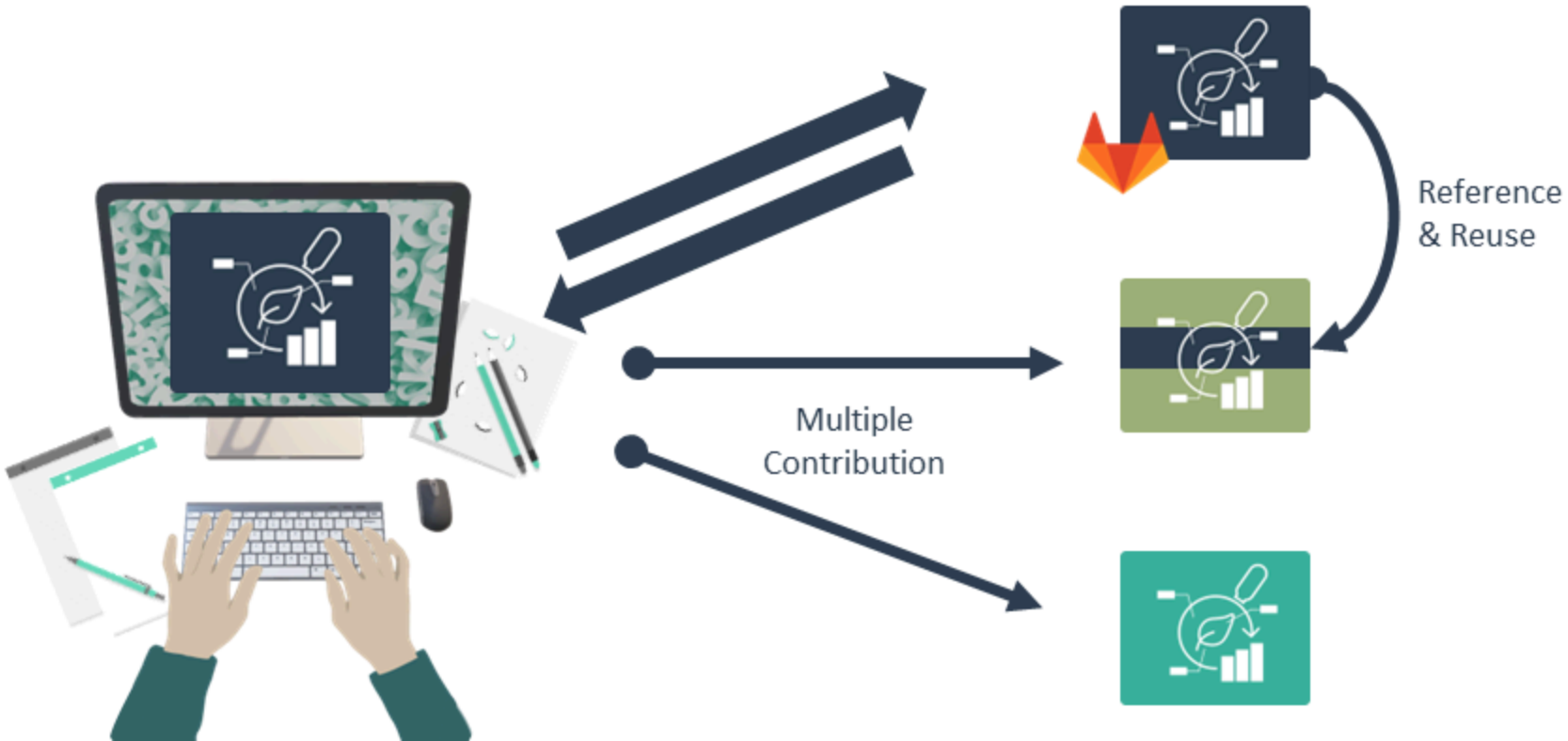
=

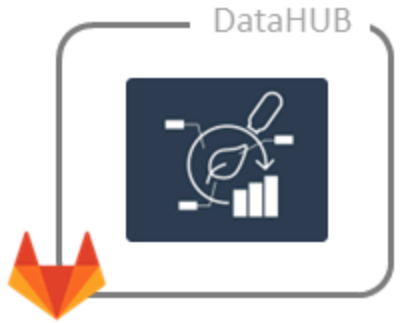

DataHUB
Storage & Backup


Versioning


Collaboration &
Access Management

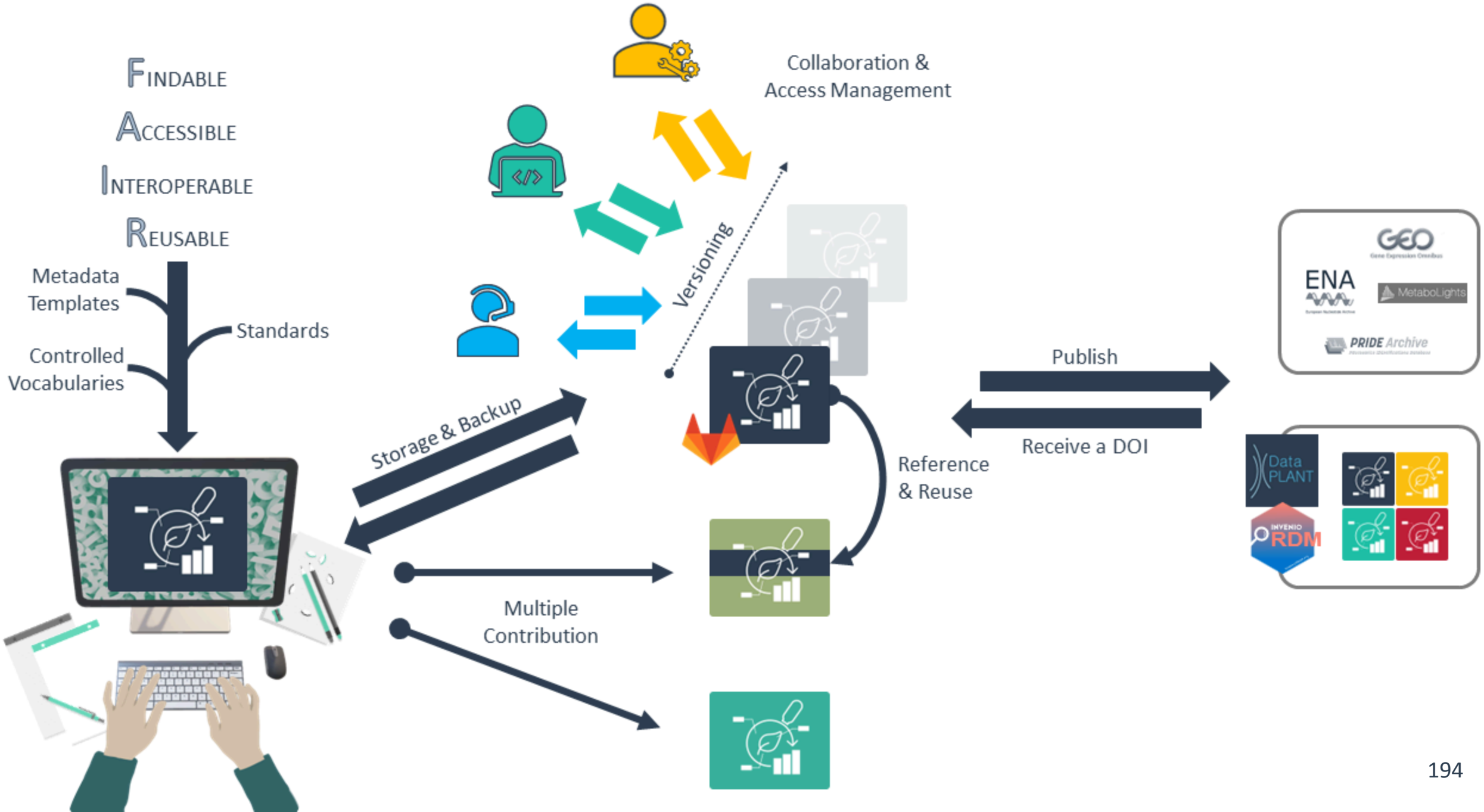

Multiple
Contribution


Reference
& Reuse
Multiple
Contribution

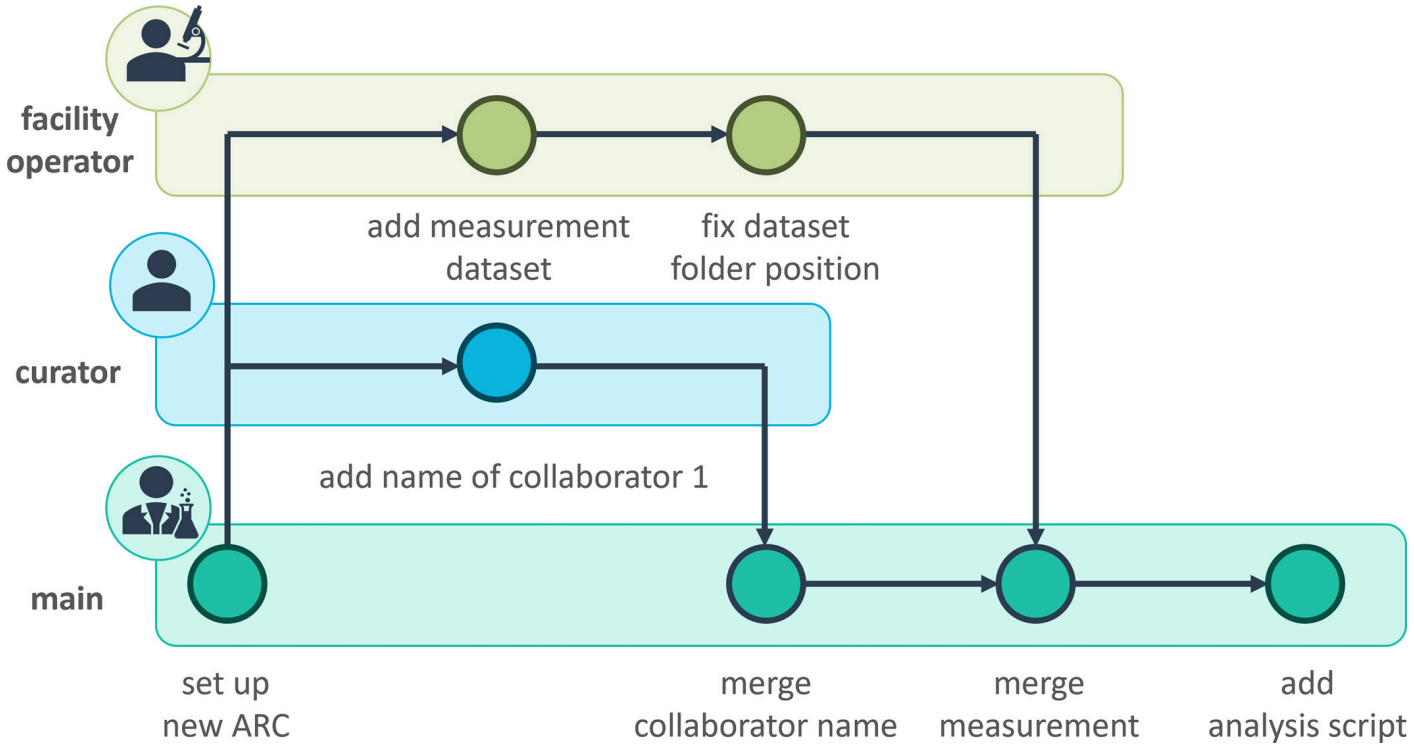

DataHUB


Publish
Receive a DOI


GEO
ENA
MetaboLights
PRIDE Archive
Data PLANT
INVENIO RDM

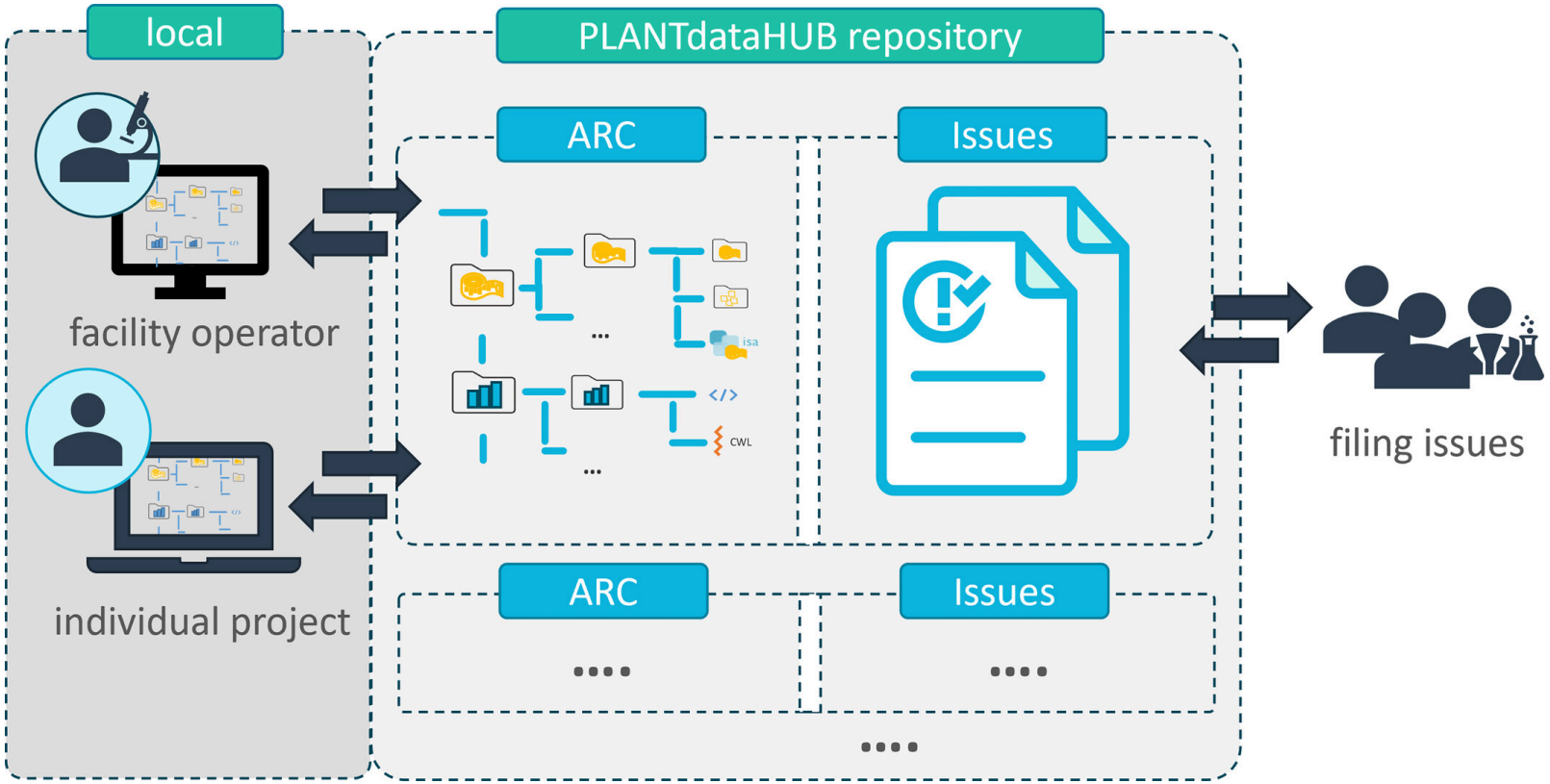

FINDABLE
ACCESSIBLE
INTEROPERABLE
REUSABLE
Metadata Templates
Controlled Vocabularies
Standards
Collaboration & Access Management
Versioning
Storage & Backup
Multiple Contribution
Reference & Reuse
Publish
Receive a DOI
GEO
ENA
MetaboLights
PRIDE Archive
Data PLANT
INVENIO RDM

Mutable data life cycle



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

Project management



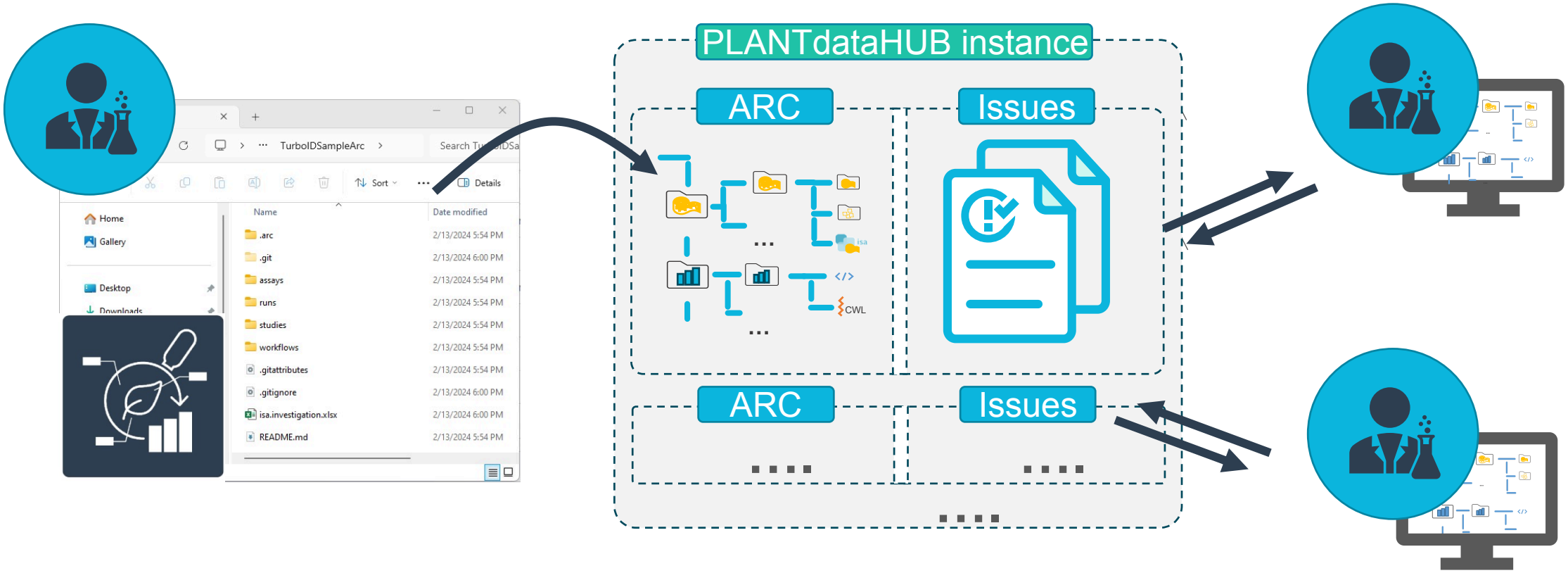
Weil, H.L., Schneider, K., et al. (2023), PLANTdataHUB: a collaborative platform for continuous FAIR data sharing in plant research. Plant J. https://doi.org/10.1111/tpj.16474

Hands-on part 3: DataHUB

Follow the **Start Here** guide in the DataPLANT knowledge base.

Until "Complete"

Using the DataHUB to collaborate



📝 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

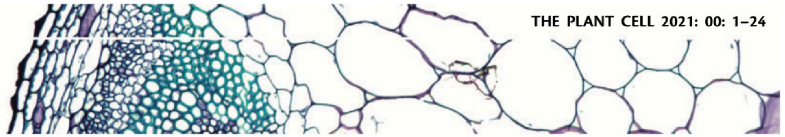
Peristent Identifiers (PIDs)

Spot the PIDs

doi:10.1093/plcell/koab243

THE PLANT CELL

THE PLANT CELL 2021: 00: 1–24



Research Article

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

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

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

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

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

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

Abstract

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

Globally unique, stable, persistent identifiers (PIDs)

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

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

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

Research
Resource Identifiers

https:// www.rrids.org

ePIC consortium

https ://www.pidconsortium. net

Research
Organization Registry

https://ror.org

Global
Research Identifier Database

https://grid.ac

Properties of a PID

Ideally, PIDs are

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

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

Additional resources

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

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-BioImage Archive	Stores and distributes biological images	imaging, microscopy
e!DAL-PGP	Plant Genomics & Phenomics Research Data Repository	phenome

Choosing a data repository

Domain-specific >> Generic >> Institutional

Find repositories at:

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

Generic data repositories

Good

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

Intermediate

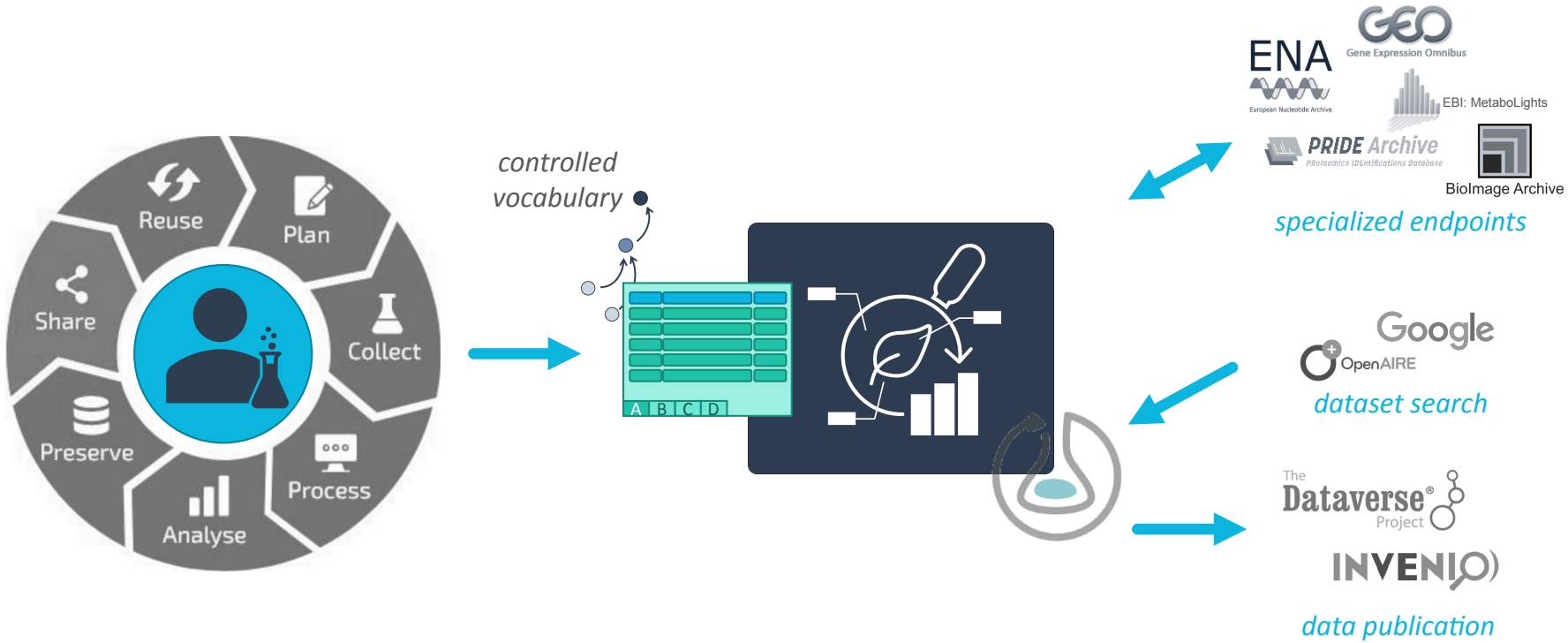
- Only generic / high-level metadata schema
- Limited reusability

https://zenodo.org

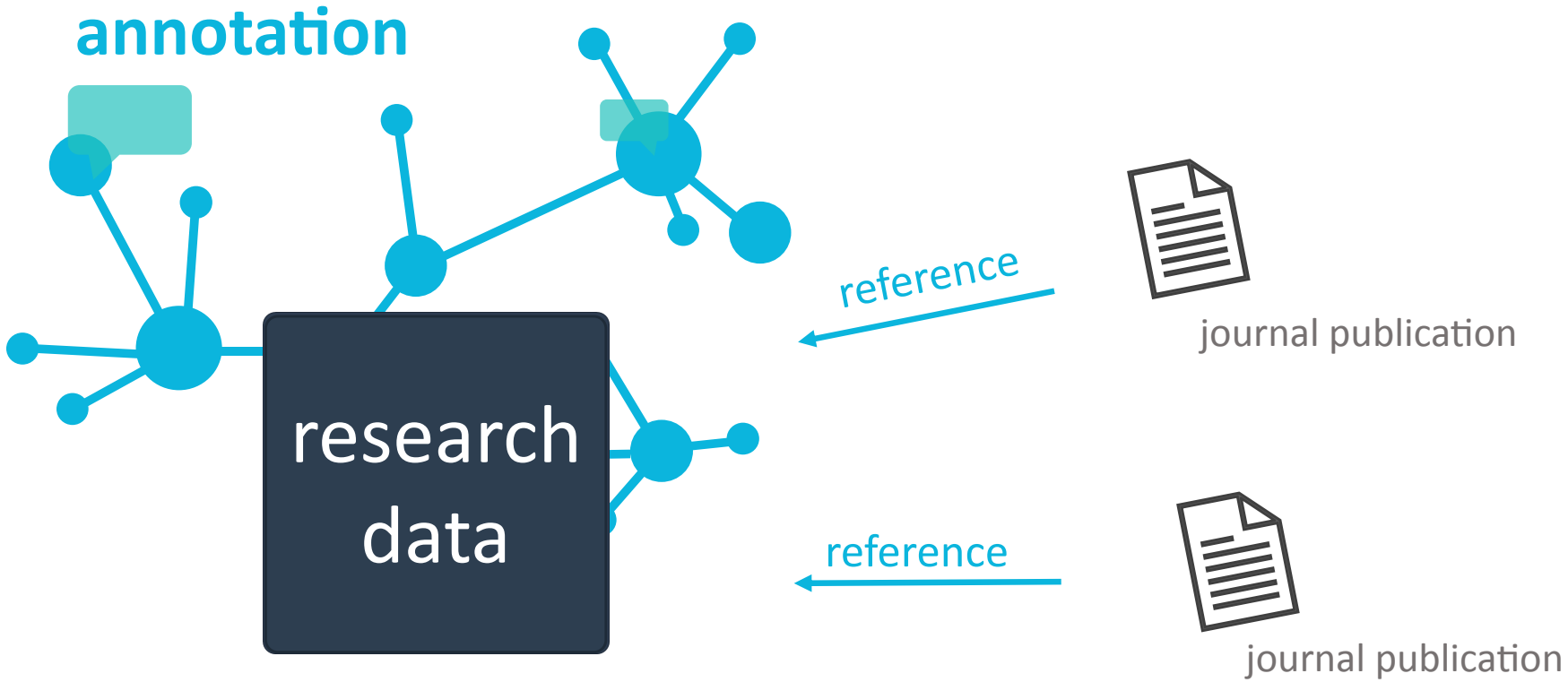
https://datadryad.org/

https://figshare.com

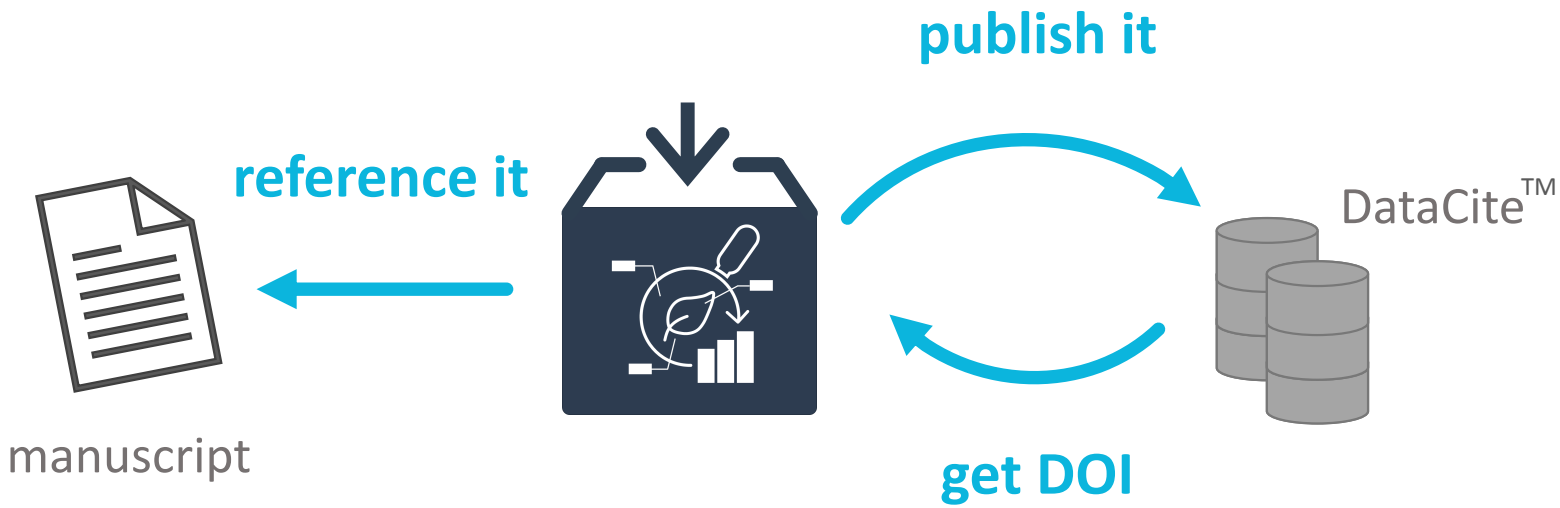
From ARC to repositories



Moving from paper to data publications

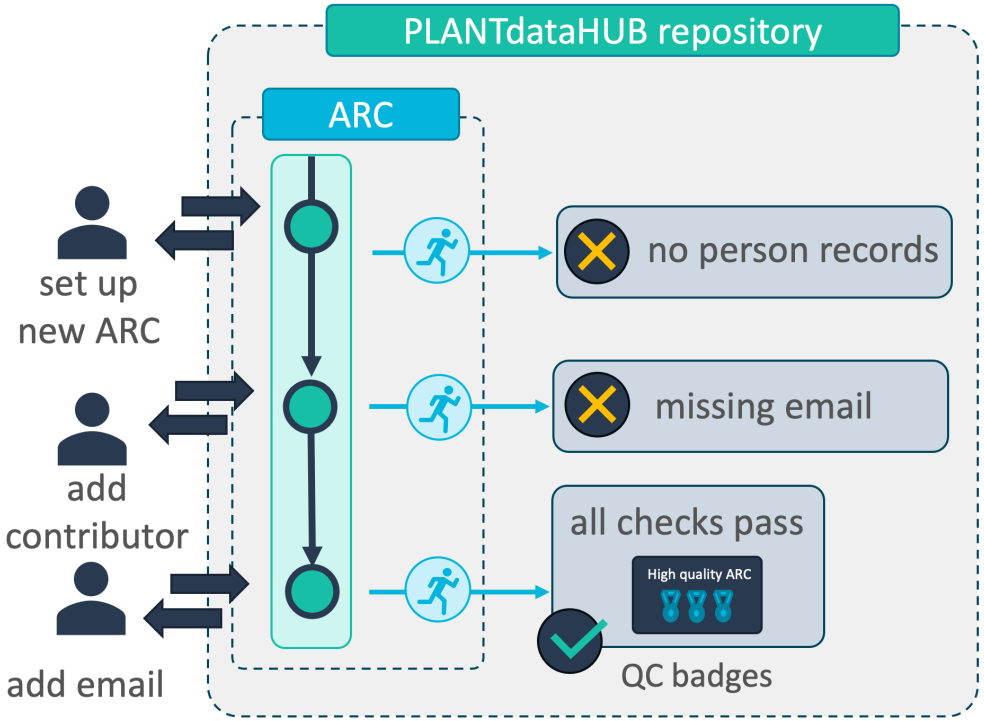


Publish your ARC, get a DOI



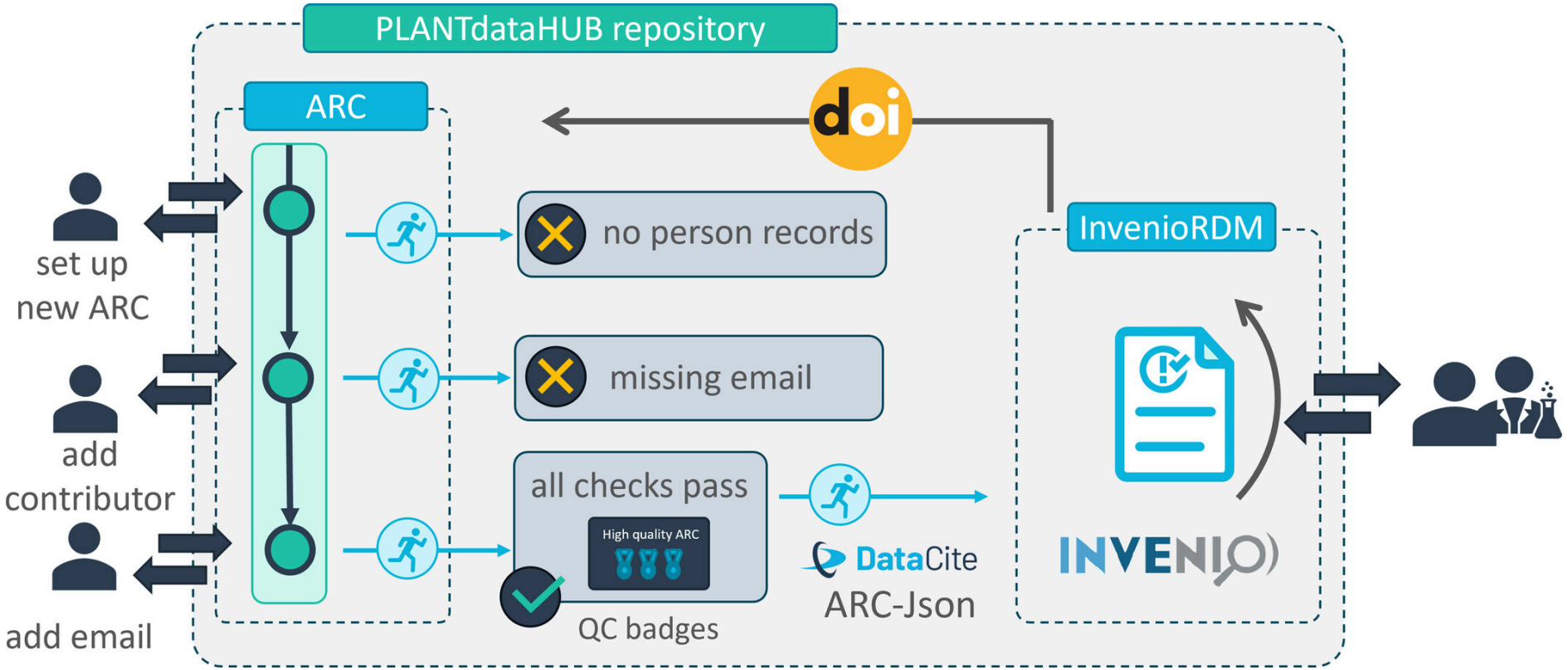
Publish your ARC with a few clicks

Validation



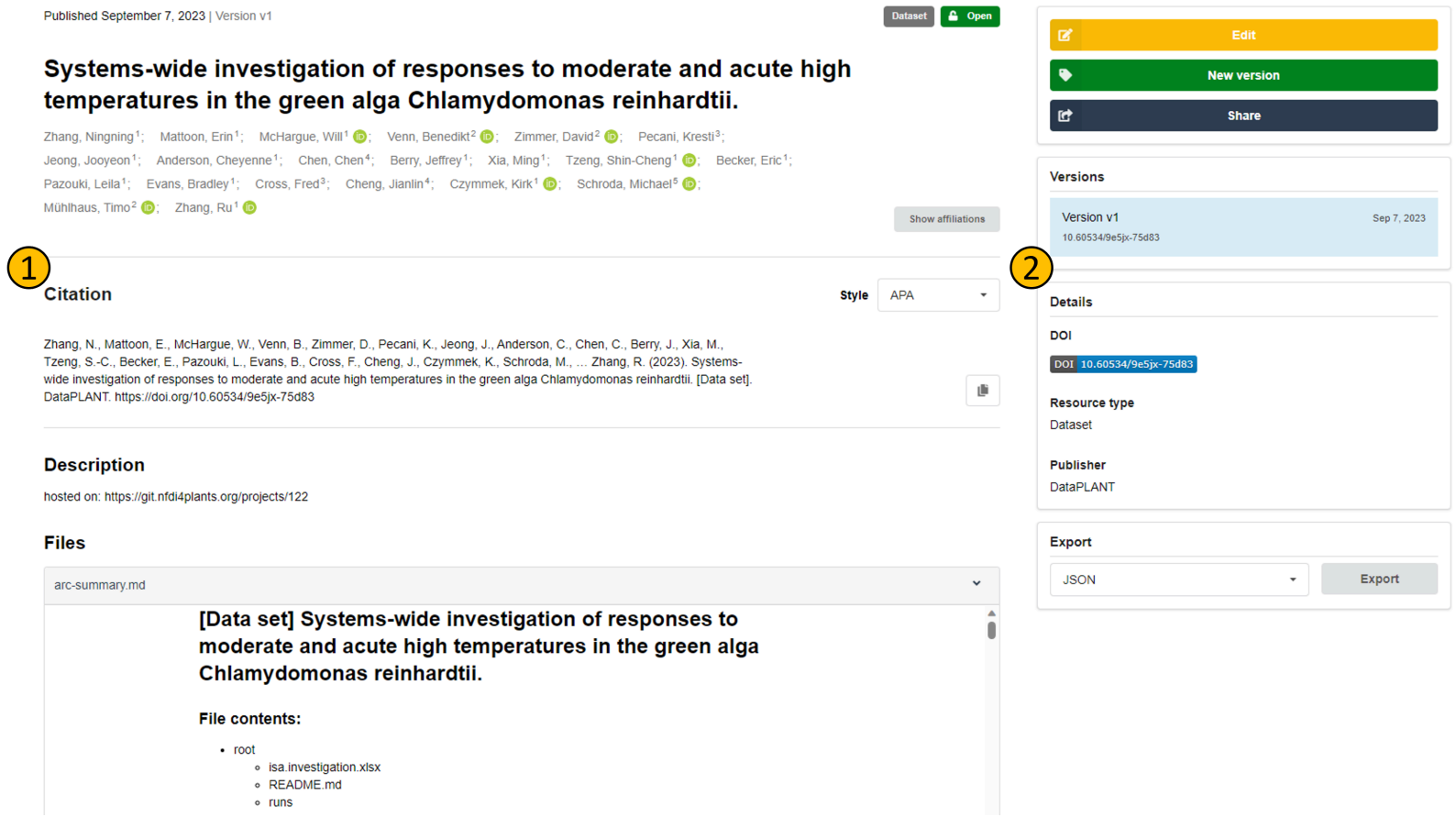
adapted from Weil, H.L., Schneider, K., et al. (2023), PLANTdataHUB: a collaborative platform for continuous FAIR data sharing in plant research. Plant J. https://doi.org/10.1111/tpj.16474

Validate & publish



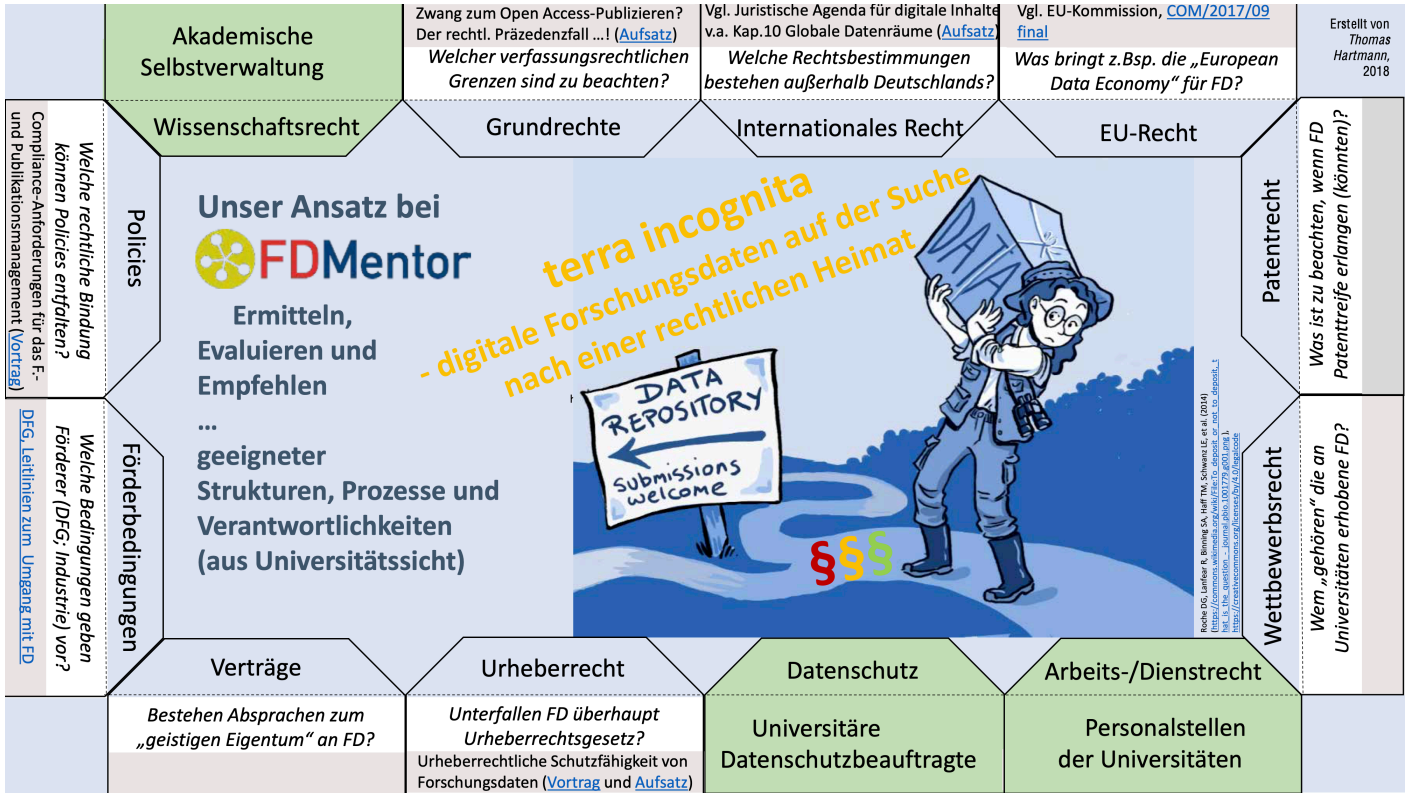
Weil, H.L., Schneider, K., et al. (2023), PLANTdataHUB: a collaborative platform for continuous FAIR data sharing in plant research. Plant J. https://doi.org/10.1111/tpj.16474

Receive a DOI



Legal aspects of RDM

Different laws touched by RDM



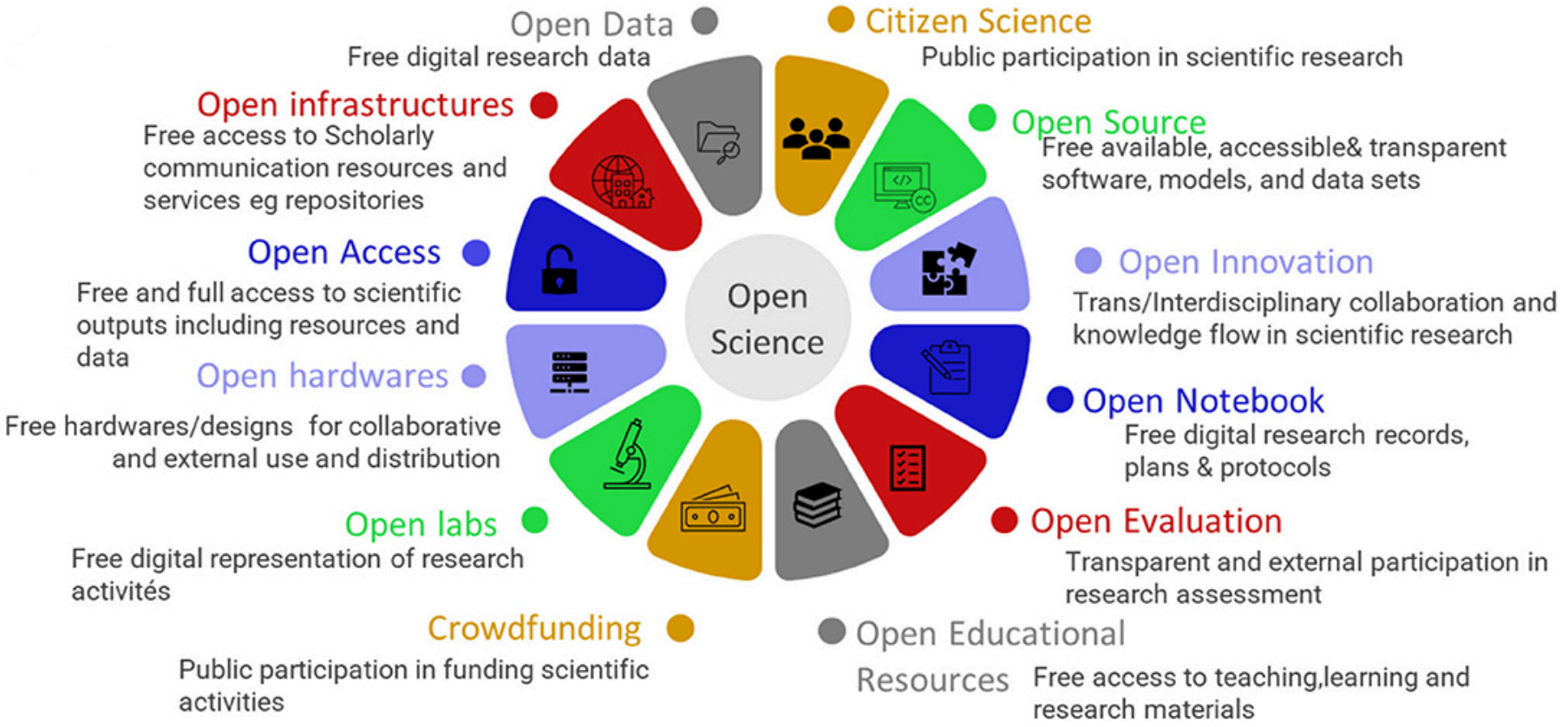
Hartmann, Thomas. (2019). Rechtsfragen: Institutioneller Rahmen und Handlungsoptionen für universitäres FDM. Zenodo. https://doi.org/10.5281/zenodo.2654306

Open Access (OA) categories

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

Piwowar H et al. (2018), PeerJ https://doi.org/10.7717/peerj.4375

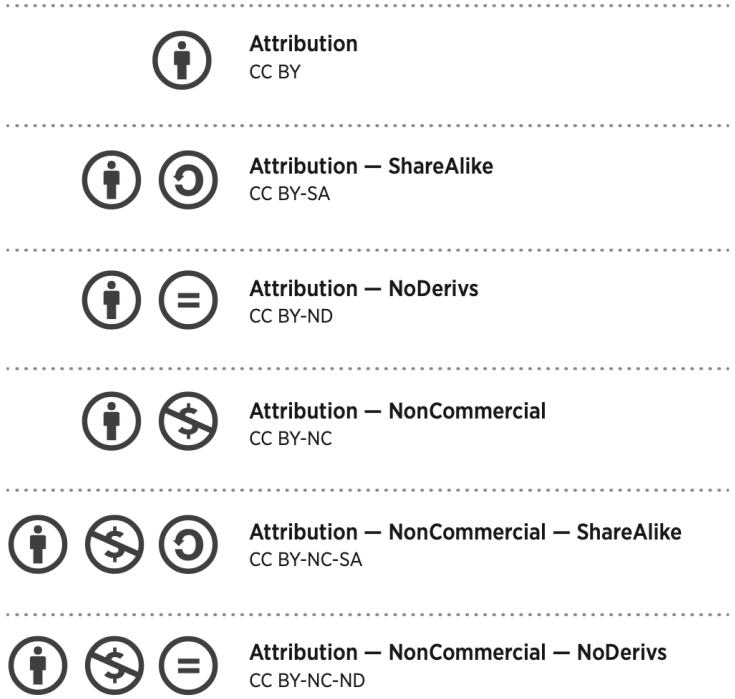
Open Science is more than Open Access



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

Creative commons

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



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

Data protection

GDPR: General Data Protection Regulation

DS-GVO (german): Datenschutz-Grundverordnung

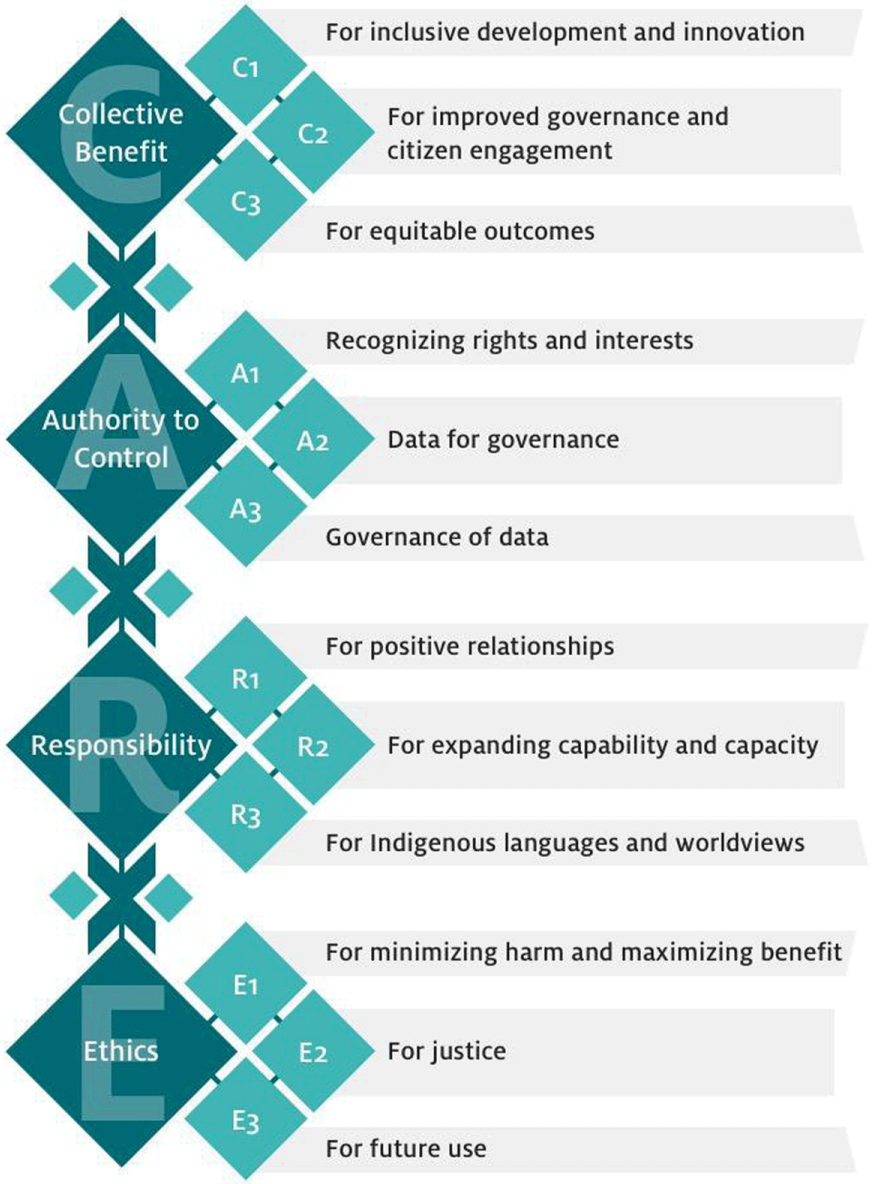
Use of biological materials

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

FAIR and CARE

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

CARE principles



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

Research Data policies

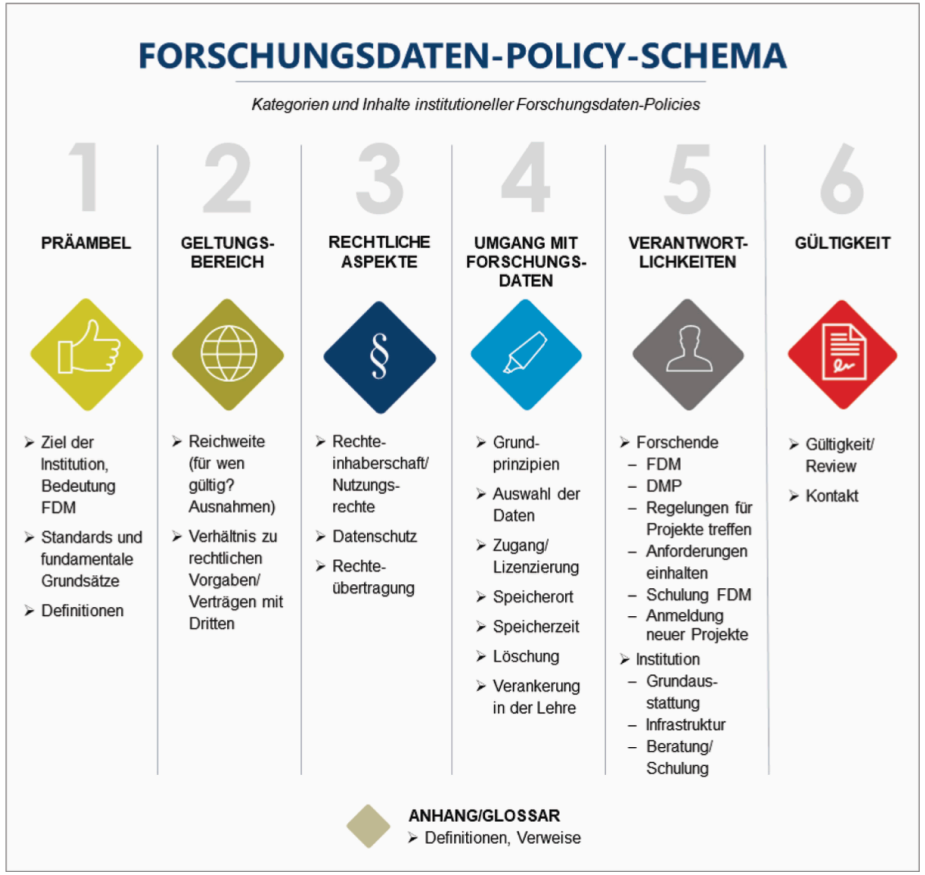


Abbildung 2: FD-Policy-Schema: Die sechs Kategorien einer FD-Policy mit ihren inhaltlichen Bestandteilen

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

CEPLAS relevant data handling guidelines & policies

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

The Data Management Plan (DMP)

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

DMP tools

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

> Check out the Elixir RDMkit for more

📝 Share your ARC 📝

1. Follow the next slides to learn how to share your ARC with a colleague
2. Go to your colleague's ARC and try to understand it

📝 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.

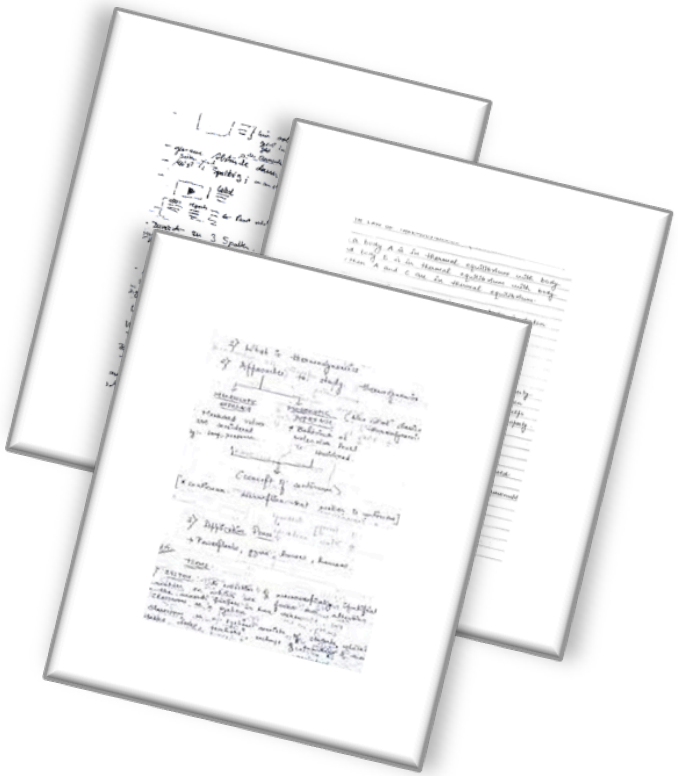
Assignment

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

Electronic Lab Notebooks (ELNs)

ELN ≈ Digital Protocol Editors

- Documenting daily lab routine
- Lab methods & protocols
- Lab inventory (biologicals, chemicals, instruments)
- Local sharing & collaboration
- Backup (locally)



✅ ELNs help to digitalize research documentation

Are ELNs FAIR?

FAIR indicator*	ELN
Findable	
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	
A1. (Meta)data are retrievable by their identifier using a standardised communications protocol	
A1.1 The protocol is open, free, and universally implementable	

*FAIR classified by https://www.go-fair.org/fair-principles/

Findable

FAIR indicator*	elabFTW
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.	🔴

*FAIR classified by https://www.go-fair.org/fair-principles/

Accessible

FAIR indicator*	elabFTW
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	❓

*FAIR classified by https://www.go-fair.org/fair-principles/

Interoperable

FAIR indicator*	elabFTW
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.	

*FAIR classified by https://www.go-fair.org/fair-principles/

Reusable

FAIR indicator*	elabFTW
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 classified by https://www.go-fair.org/fair-principles/

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
- name: Martin Kuhl