CEPLAS
Cluster of Excellence on Plant Sciences

CEPLAS data management and DataPLANT

February 26th, 2020

Hajira Jabeen & Dominik Brilhaus

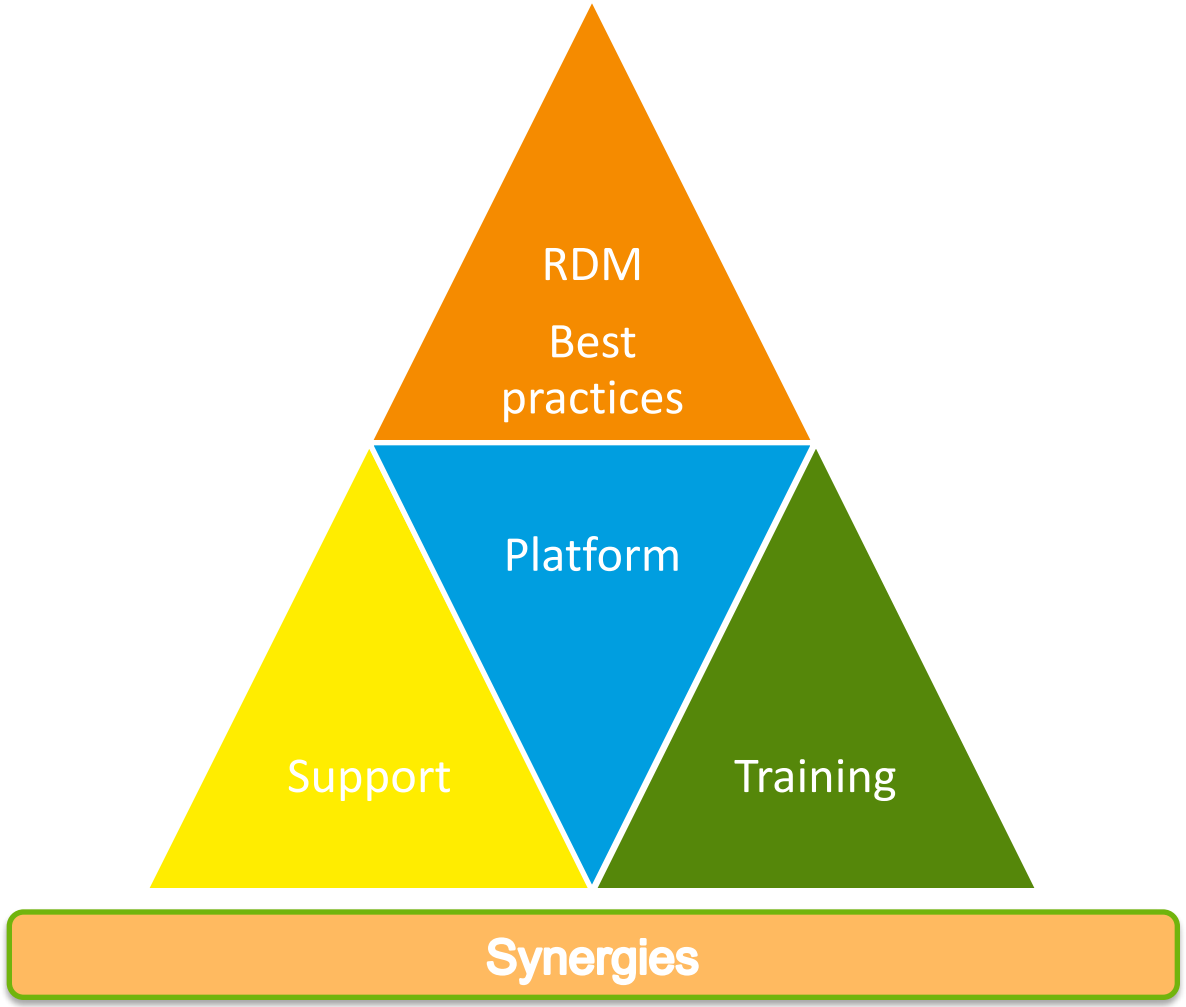
Our Vision

- Making CEPLAS research data FAIR
- Make Data Management @CEPLAS simple
  - organising data
  - making it reusable
  - find and use external data
  - publishing data in public repositories (required by journals & funders)
- Let collaborators find and work on your data
- Support to use internal resources for related tasks
  - e.g Computing, Backup, Sharing, etc . . .

CC BY NC SA

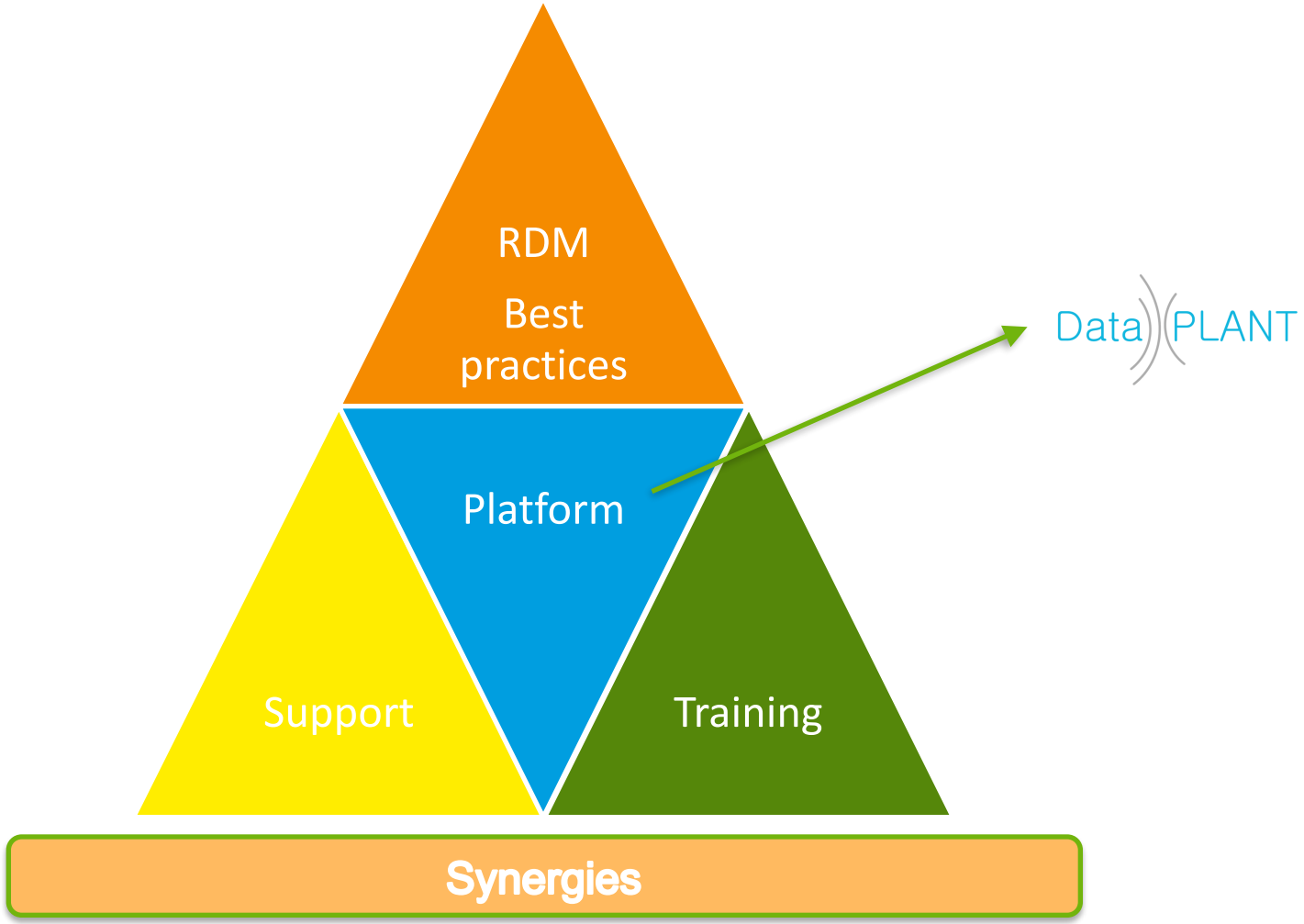
hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

Research Data Management @ CEPLAS



CC BY NC SA

Research Data Management @ CEPLAS



CC BY NC SA

ARC - Simply

A Directory + Template(Excel)

- A structured directory (following ISA model)
- An Excel file asking for necessary descriptions (MetaData)
- Useful to structure data of ANY
  - Type (transcriptomics, proteomics, metabolomics, microscopy, countable, code, workflows etc etc)
  - Size (a few MBs - - GBs)
  - Granularity (Assay, Analysis, Protocols . .)
  - Source (Wet Lab, Imaging facility . .)
    - Any data can be stored as into an ARC

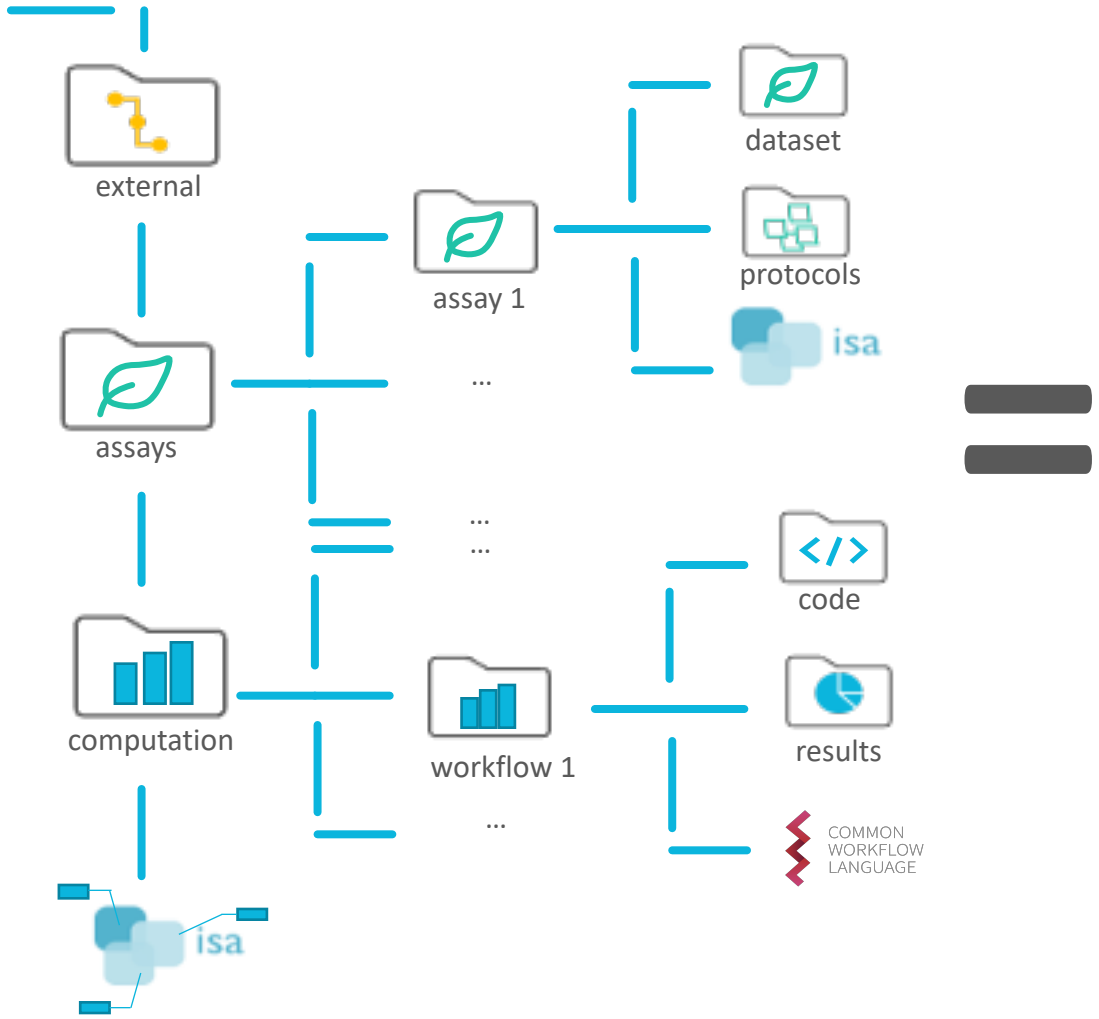
CC BY NC SA

ARC - Why

- A homogenised data packaging model to enable FAIR
- Easy to use
- Nominal additional effort
- Useful (inbuilt) features
  - Version control
  - Sharing
  - Backup
  - Easy publishing
  - Compatibility with platforms like Galaxy, Omero etc

CC BY NC SA

What does a real-world ARC look like?



=

CC BY NC SA

hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

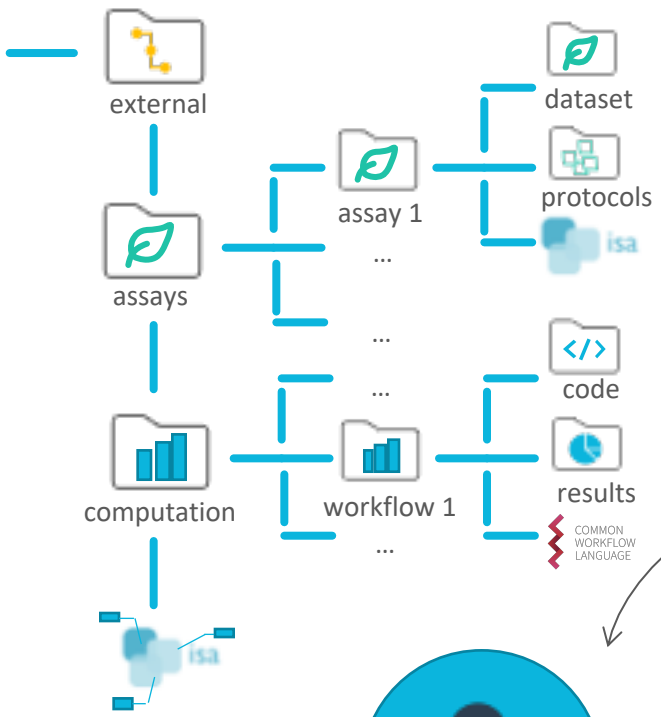
ARCs: Share yours and work together

- ARCs are the basis of the collaborative research platform
- ARCs can be shared between researchers seamlessly

CC BY NC SA

hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

Collaborative working with ARCs



versioning

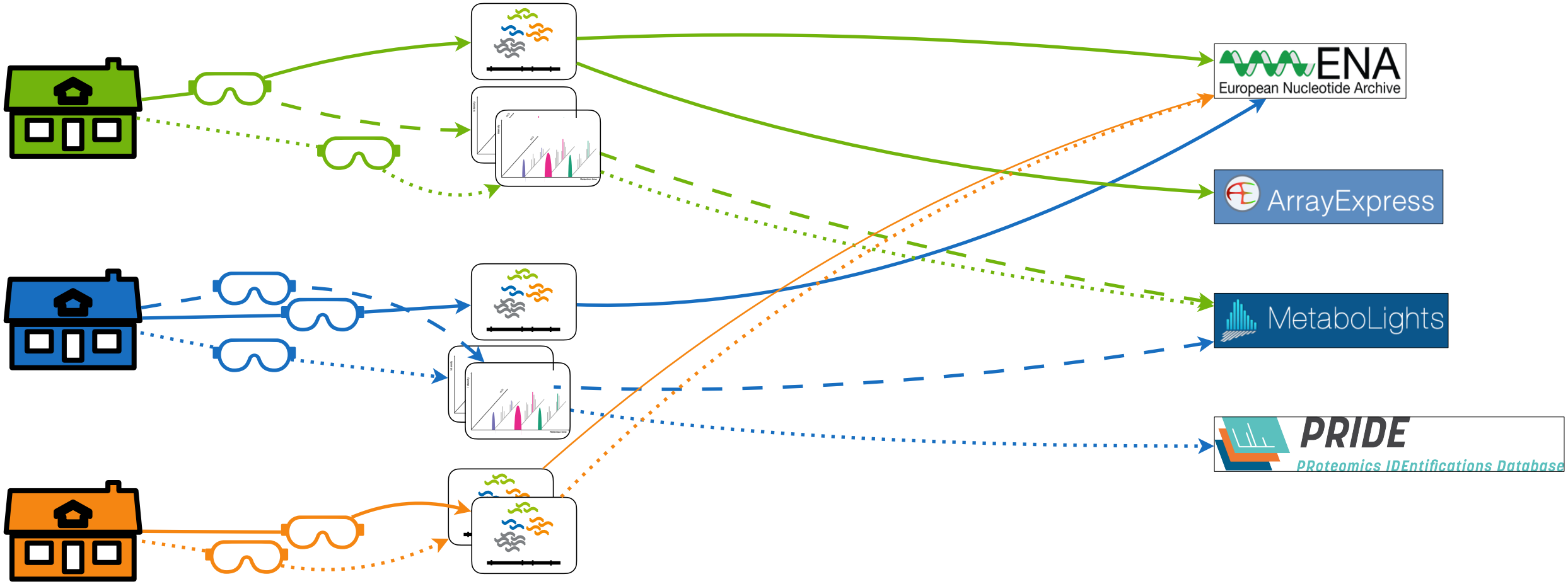
collaborative

- database
- object storage
- git repository

local + decentralized

hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de
CC BY NC SA

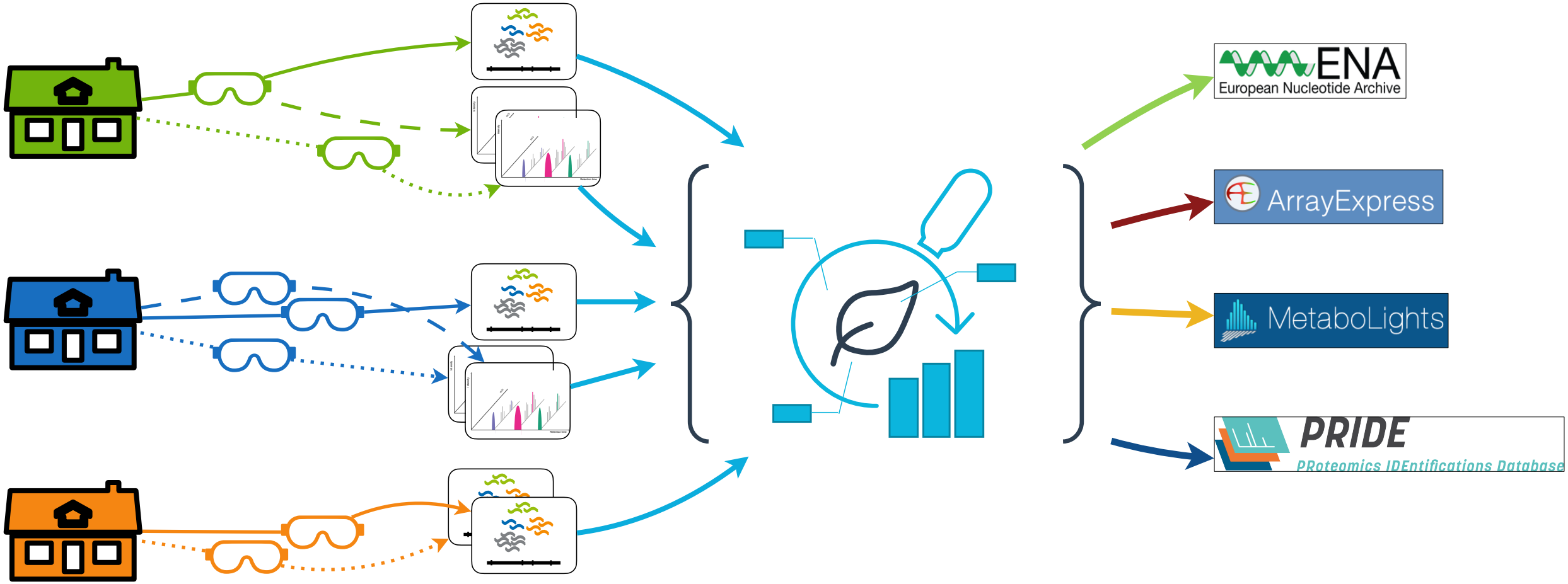
The guidelines for all repositories (challenge!)



CC BY NC SA

hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

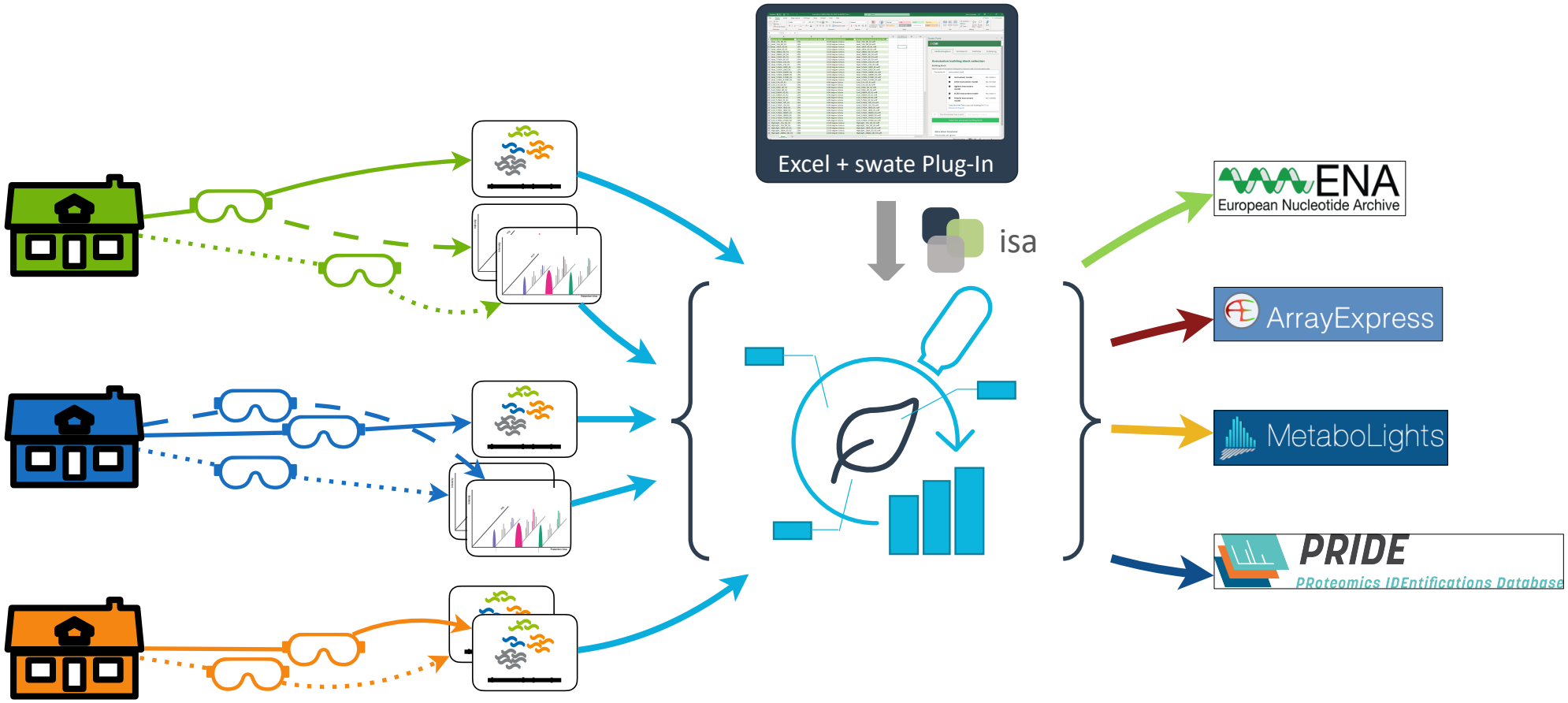
ARC to the rescue!



CC BY NC SA

hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

Reducing the overall workload



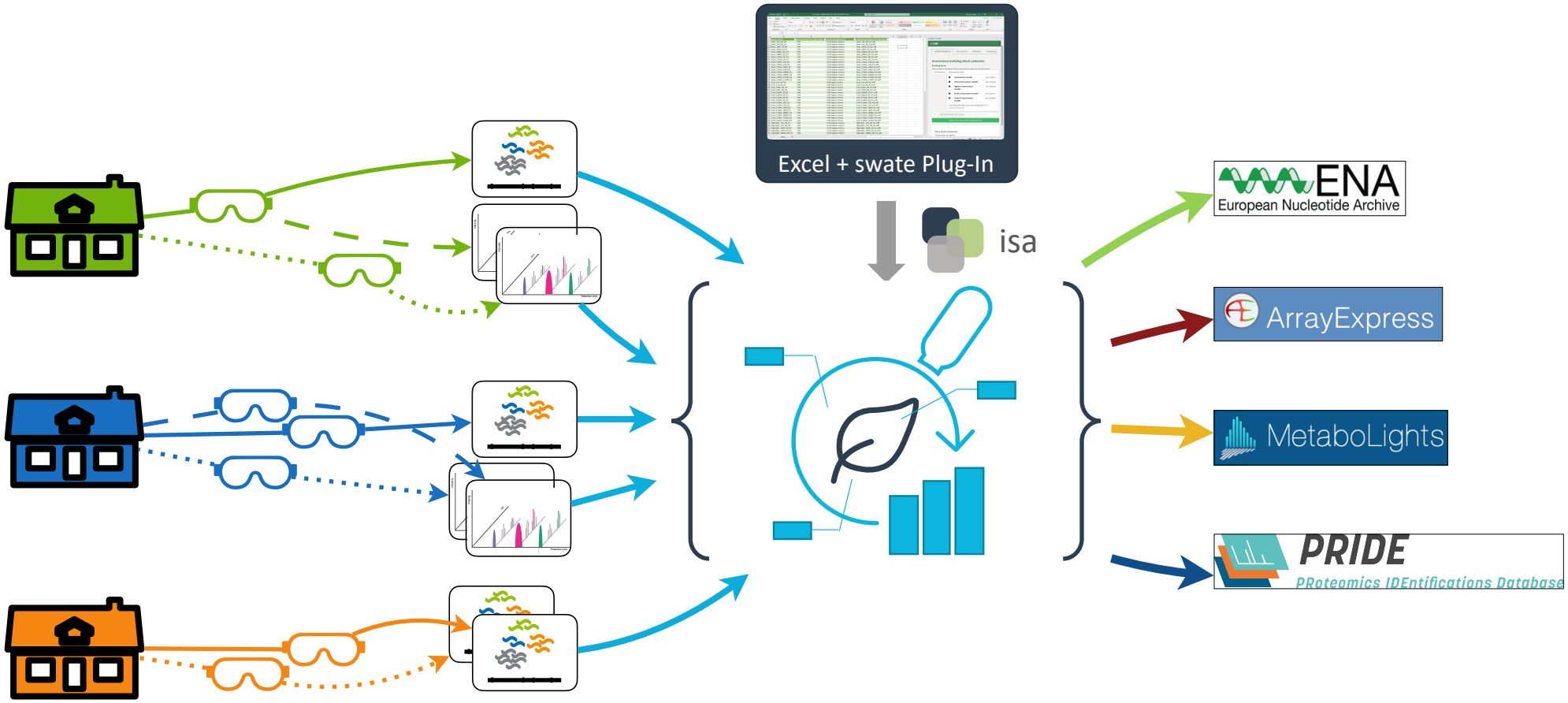
1

WE work on the ARC mechanisms

CC BY NC SA

hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

Reducing the overall workload



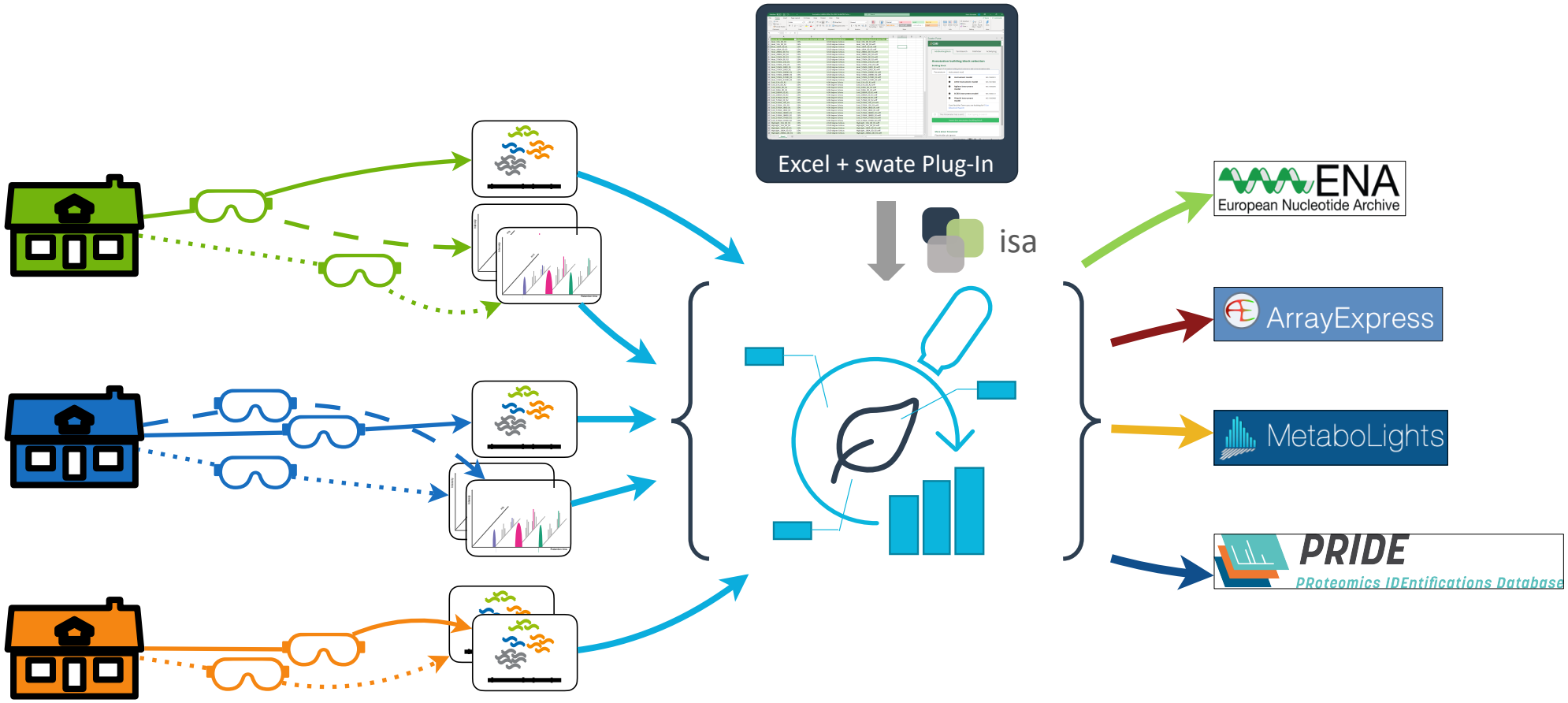
2 **WE** guide **YOU** how to use the ARC

1 **WE** work on the ARC mechanisms

CC BY NC SA

hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

Reducing the overall workload



3 So **YOU** can focus on your research

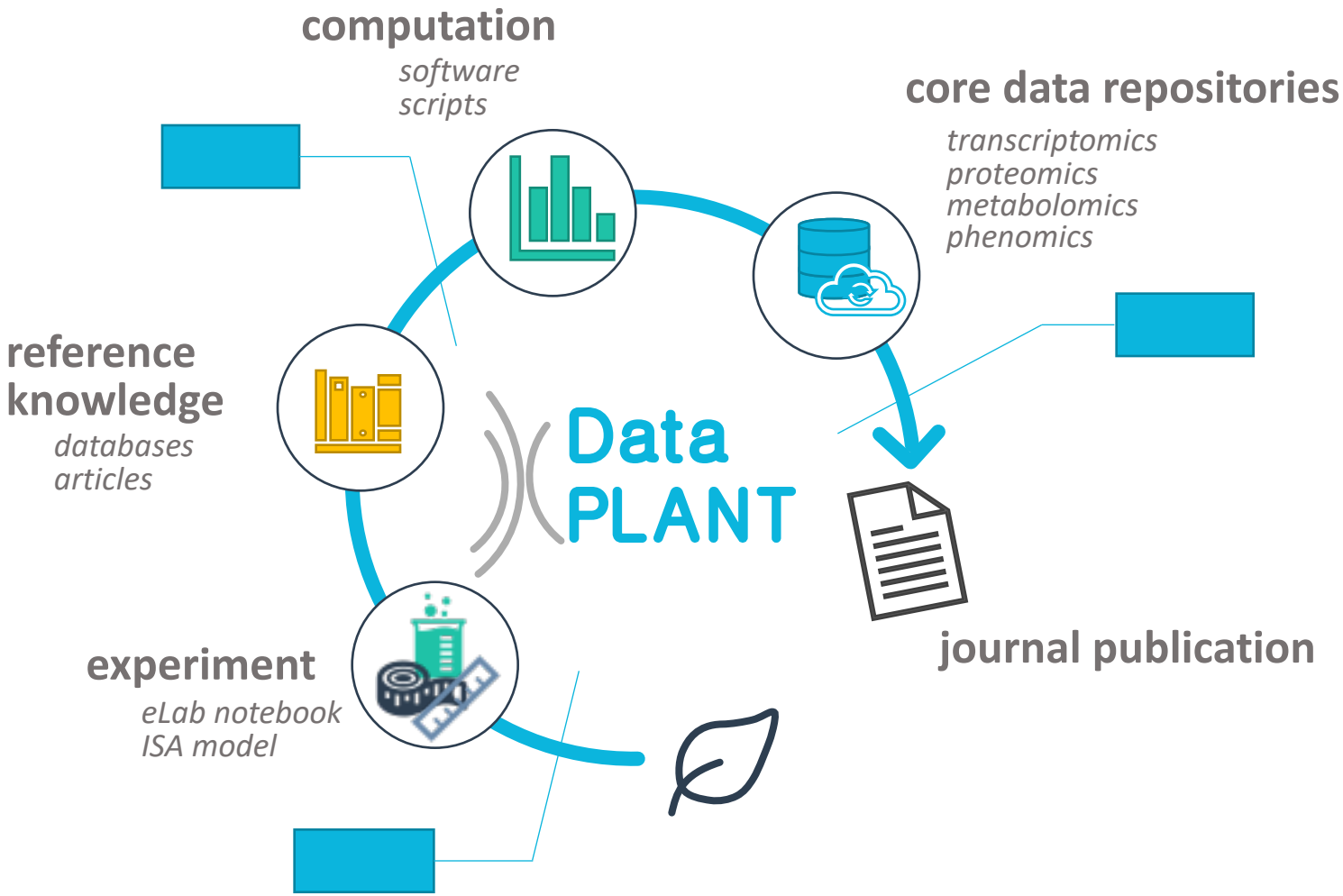
2 **WE** guide **YOU** how to use the ARC

1 **WE** work on the ARC mechanisms

CC BY NC SA

hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

ARC covers the complete research cycle

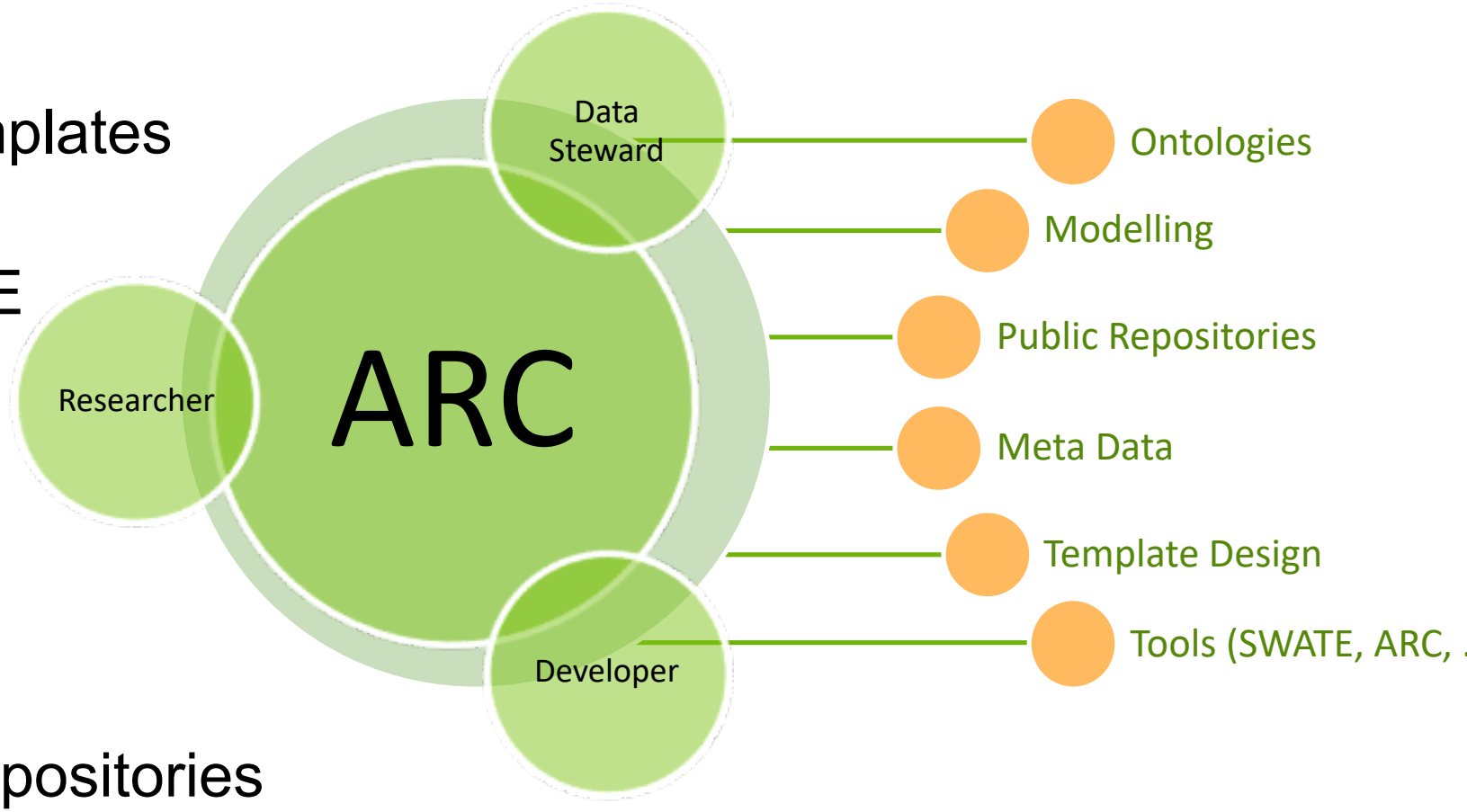


CC BY NC SA

hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

Behind the ARC

- A researcher only works with ARC
- ARC provides Excel Templates
  - Enriched using SWATE
  - Data type specific
  - Extended as needed
  - Exportable to public repositories



CC BY NC SA

Caution! Work in Progress ..

CEPLAS + DataPLANT

- Specialised templates
- Support of data types, and workflows
- Integration of ARC with analytic tools e.g.Galaxy
- DataHub with access controls
- Seamless version control

CC BY NC SA

hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

Summary

CEPLAS + DataPLANT

- An MoU is being developed with DataPLANT
  - CEPLAS uses DataPLANT hub as its “data management platform”
  - This is reflected in CEPLAS data policy
    - Taking care of IPR and technological aspects

CC BY NC SA

hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

Summary

CEPLAS + DataPLANT

- An MoU is being developed with DataPLANT
  - CEPLAS uses DataPLANT hub as its “data management platform”
  - This is reflected in CEPLAS data policy
    - Taking care of IPR and technological aspects

Thank you
Questions .. ?

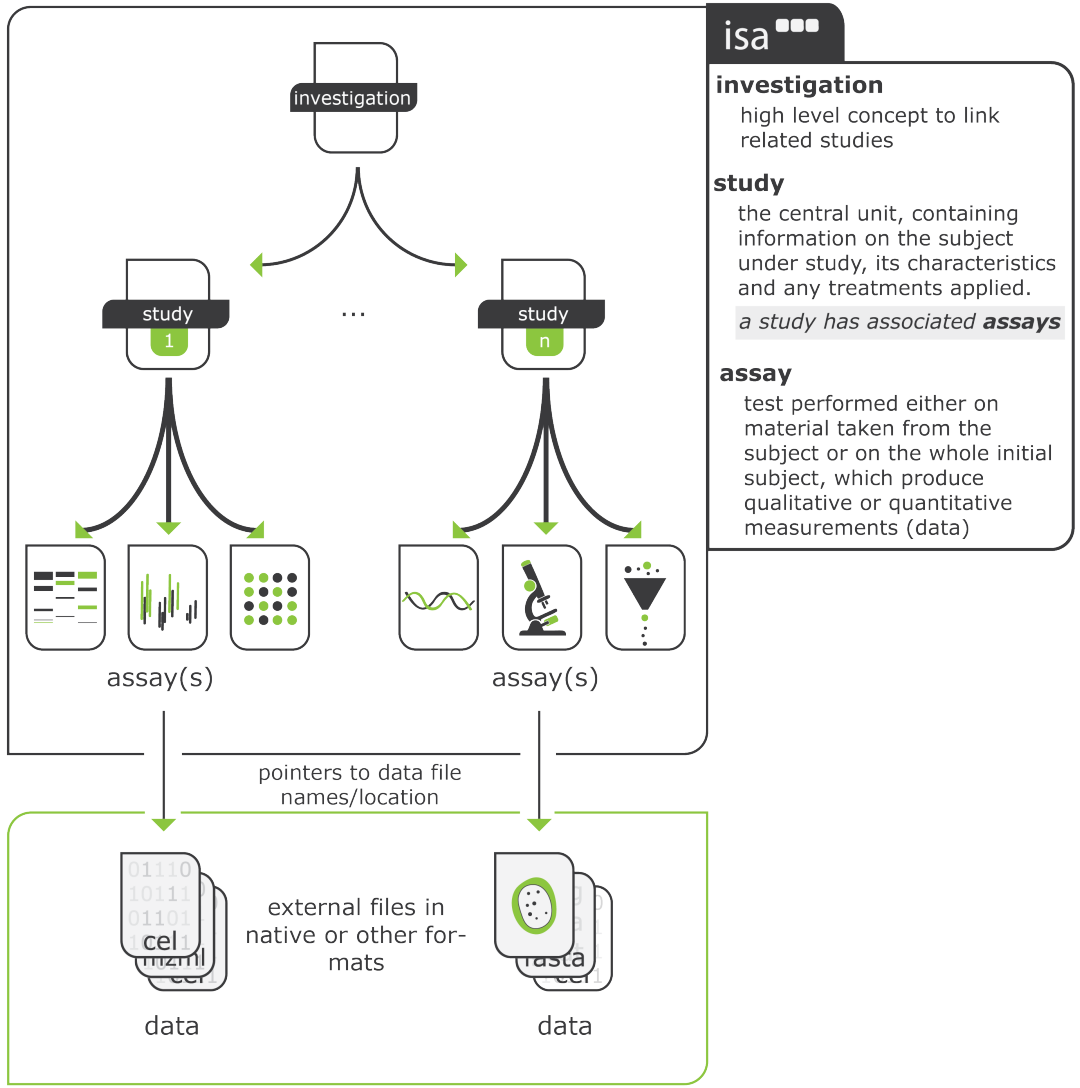
CC BY NC SA

hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de
CC BY NC SA

The ISA Model of (experimental) metadata

- **Investigation**
  - Overall goals
  - Scientific context
- **Study**
  - Experimental steps
- **Assay**
  - Leading to (raw) data



https://isa-tools.org/format/specification.html

hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

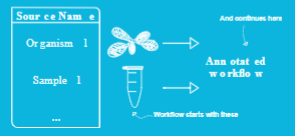
CC BY NC SA

Annotation principles

Source Name

The Source Name column defines the source of biological material used for your experiments. The name used must be a unique identifier. It can be an organism, a sample, or both.

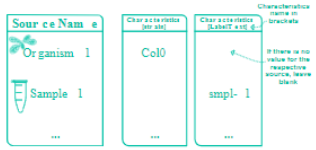
Every annotation table must start with the Source Name column



Characteristics

Use characteristics columns to annotate interesting properties of the source material.

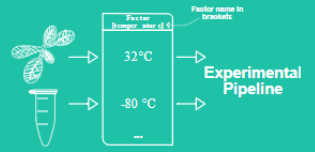
You can use any number of characteristics columns.



Factor

Use factor columns to track the experimental conditions that govern your study

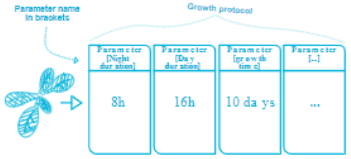
Most of the time, factors are the most important building blocks for downstream computational analysis.



Parameter

Use parameters to annotate your experimental workflow.

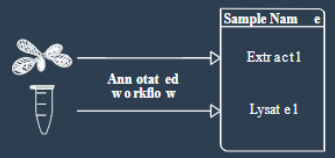
You can group parameters to create a protocol.



Sample name

The Sample Name column defines the resulting biological material of the annotated workflow. The name used must be a unique identifier.

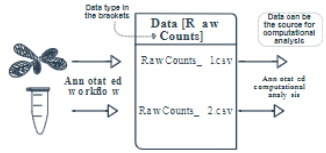
Samples can again be sources for further experimental workflows.



Data

The Data column describes data files that results from your experiments. Additionally to the type of data, the annotated files must have a unique name.

Data files can be sources for computational workflows.



Mühlhaus, von Suchodoletz, Krüger, Usadel. (2020, September). DataPLANT Kick-Off Meeting: 21st September 2020: First General Assembly. Zenodo. http://doi.org/10.5281/zenodo.4039749

CC BY NC SA

hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de