CEPLAS
Cluster of Excellence on Plant Sciences

Good Data Management-Practices

November, 2020

Hajira Jabeen & Dominik Brilhaus

Before we start today

- Please download and install git
  - https://git-scm.com/downloads

- Register at https://github.com/
- Register at https://elabftw.hhu.de/register.php
  - Team “CEPLAS Demo”
  - Use your @hhu.de or @uni-koeln.de email address

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

CEPLAS data science and management

Academic experts in data science

Hajira Jabeen

hajira.jabeen@uni-koeln.de

+49 221 470 89997

https://hajirajabeen.github.io/

Dominik Brilhaus

dominik.brilhaus@hhu.de

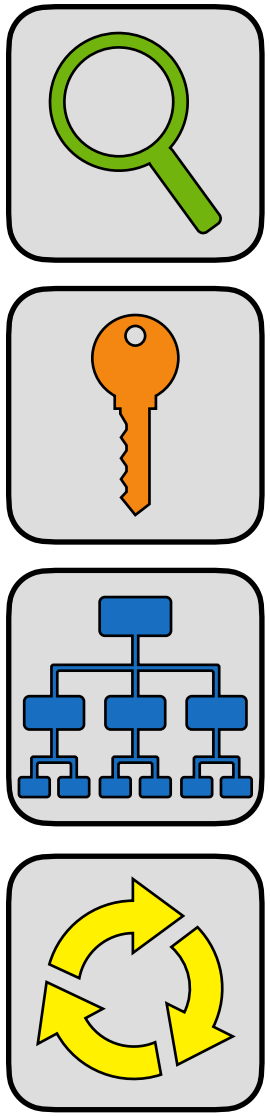
+49 211 81 15523

Data management officer

Björn Usadel

- Head of Bioinformatics Institute (IGB-4) @ Jülich
- Head of Institute for Biological Data Science @ HHU

FAIR principles of data stewardship



Findable

Accesible

Interoperable

Reusable

Wilkinson, M., *et al.* (2016)
The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018.
https://doi.org/10.1038/sdata.2016.18

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

Community-integrated plant research data

CEPLAS institutes

CEPLAS

hhu Heinrich Heine Universität Düsseldorf

Universität zu Köln

JÜLICH Forschungszentrum

Max-Planck-Institute for Plant Breeding Research

CEPLAS Data Science and Management

hajira.jabeen
@uni-koeln.de

dominik.brilhaus
@hhu.de

DataPLANT Community

DataPLANT

nfdi4plants.de

Local computing infrastructure

RRZK
fdm.uni-koeln.de
ZIM
fdm.hhu.de

F A I R

CEPLAS
Research data
platform

Public repositories

*Findable, Accessible, Interoperable, Reusable

Goals

- Appreciate FAIR principles
- Provide tools for FAIR Data Management
- Communication and terminology is key
- Collect and structure metadata with your data as early as possible

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

Introduction to research data

Research Data

- Any information;
  - Collected
  - Observed
  - Generated
  - Created
- . . to validate research findings
- Types of research data
  - Digital
  - Non-digital

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

Research Data Sources

- Observational data
  - Captured for research, often irreplaceable e.g. surveys, sample data, neuro-images etc
- Experimental data
  - Captured by lab equipment, reproducible(expensive), e.g. gene-sequences
- Reference/Canonical data
  - Collection of smaller, (published) datasets. e.g. data from gene sequence banks etc

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

Research Data Sources

- Simulation data
  - Generated by test models (model and metadata are more important than output) e.g. climate models
- Derived/Compiled data
  - Transformed from pre-existing data, reproducible if lost. e.g. data mining, compiled databases, 3D models etc

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

Research Data

What is it? Forms and formats

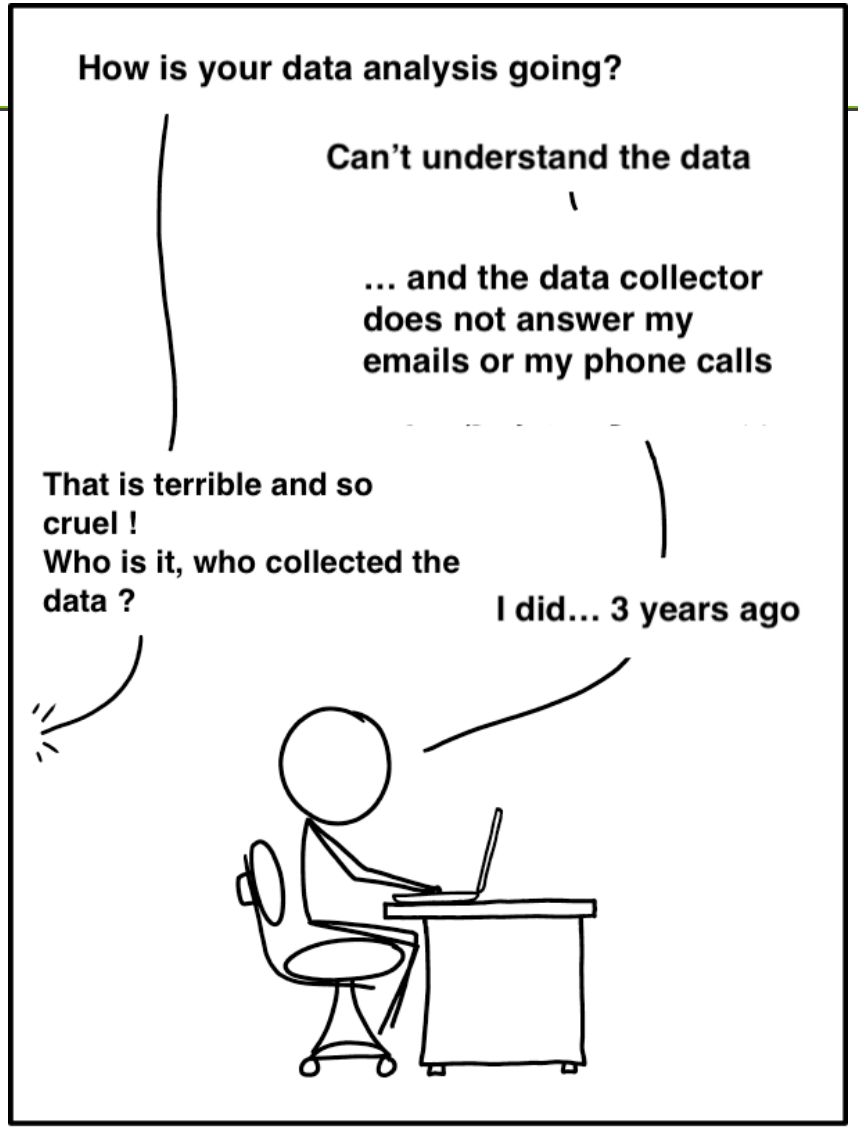
- Documents
- Spreadsheets
- Laboratory Notebooks
- Field Notebooks
- Diaries
- Questionnaires
- Transcripts
- Codebooks Audiotapes
- S O Ps
- Videotapes
- Photographs
- Films
- Test Responses
- Slides
- Artifacts
- Specimens
- Models
- Protocols
- Algorithms
- Scripts
- Contents Of An Application
- Input, Output,
- Analysis Software
- Simulation Software
- Schemas
- Methodologies
- Workflows

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

Research Data Management

Why ?

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



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

https://rdmpromotion.rbind.io/promotion/

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

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

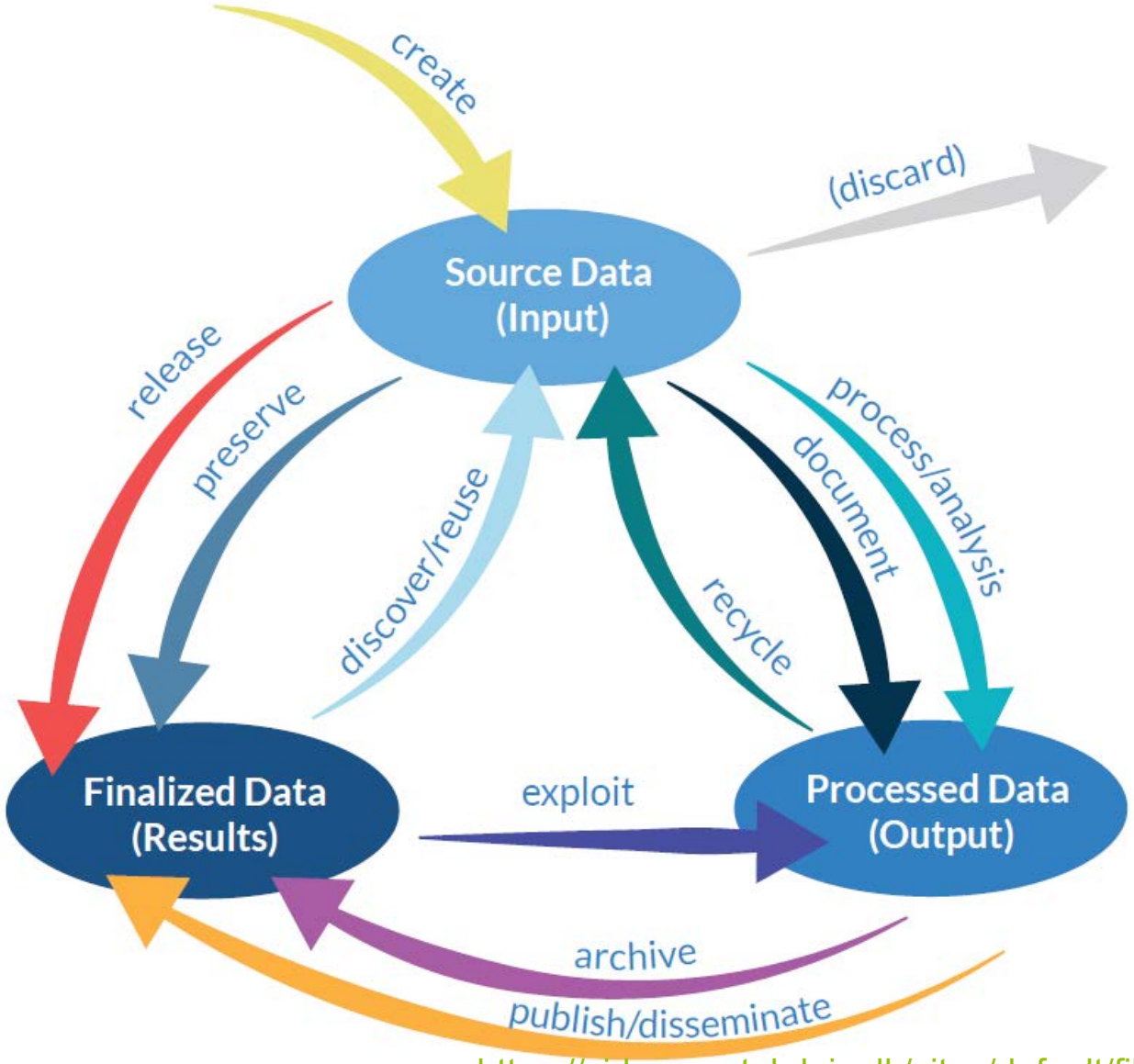
Research Data Management

What is it?

- Organization of Data
  - The process: from entry into research cycle, to dissemination and archiving of the results
  - Concerns
    - Data creation
    - Data organization
    - Data storage and backup
    - Data sharing and search

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

Research Data Lifecycle



https://vidensportal.deic.dk/sites/default/files/uploads/FAIR/Diagram.png

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

Research Data Lifecycle

Source Data

- Data creation
  - Raw data
  - File format
    - Used for analysis
    - Used for sharing
  - Choose a format that is flexible to enable easy reuse, standard format highly utilised in your domain
  - Non proprietary and unencrypted

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

Research Data Lifecycle

Processed Data (Analysis, methods and results)

- Use well known tools and compute environments
- Keep track of
  - tool versions
  - Parameters
  - Document everything
    - Keep raw data separate
    - Document changes to raw data
    - Create a version for analysis (more on this later..)

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

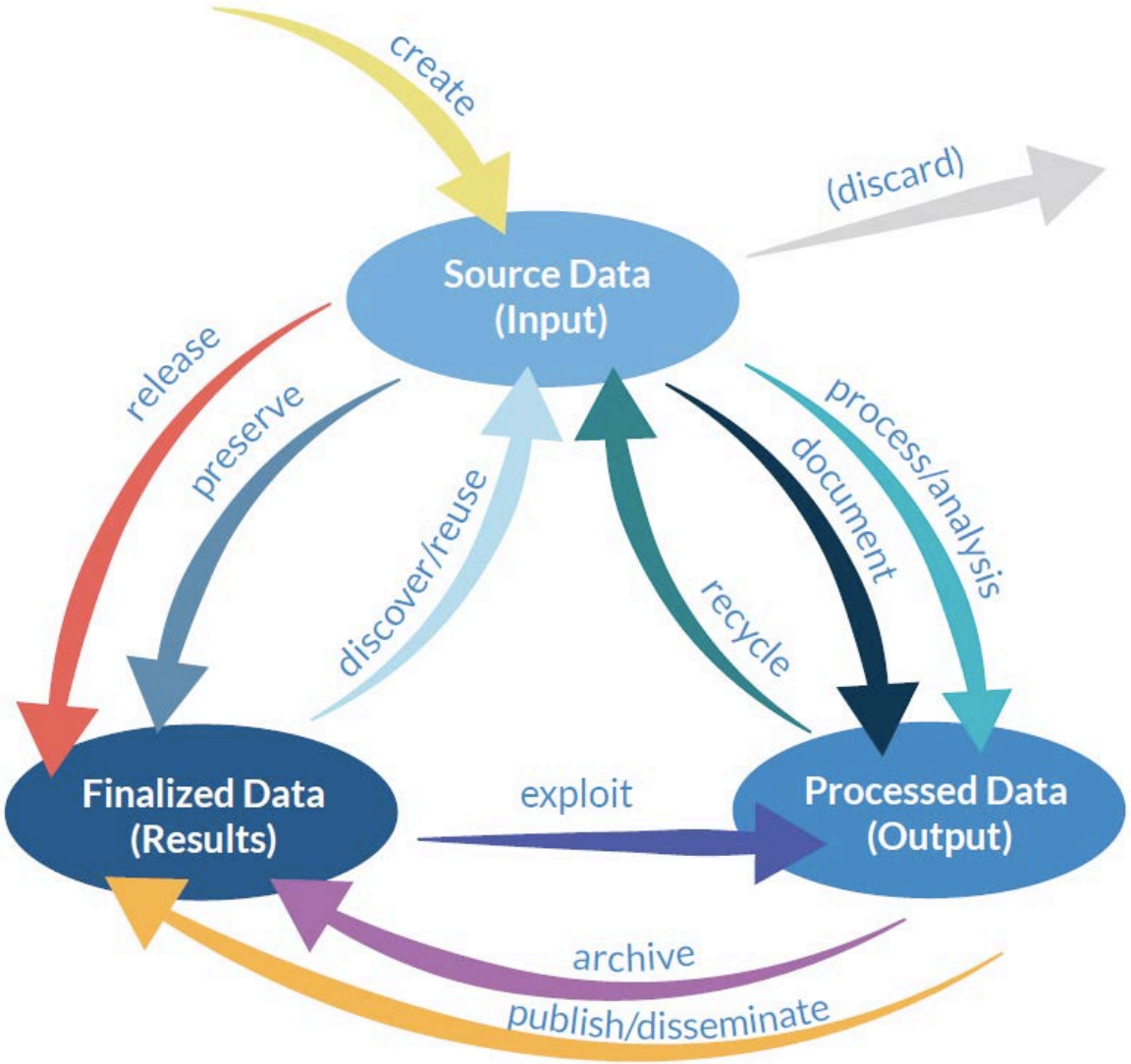
Research Data Lifecycle

Finalised data (results)

- If best practices are followed during the lifecycle. The resulting data is
  - Understandable
  - Easily reusable
  - Findable
  - Accessible
  - Results are reproducible
    - Transparency in research
    - Facilitates reuse and helps in citations

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

Source Data



https://vidensportal.deic.dk/sites/default/files/uploads/FAIR/Diagram.png

Raw data and file formats

- Why do we care for file formats
  - Proprietary formats are mostly associated with tools
    - e.g xls , docx .. prefer txt, csv, instead

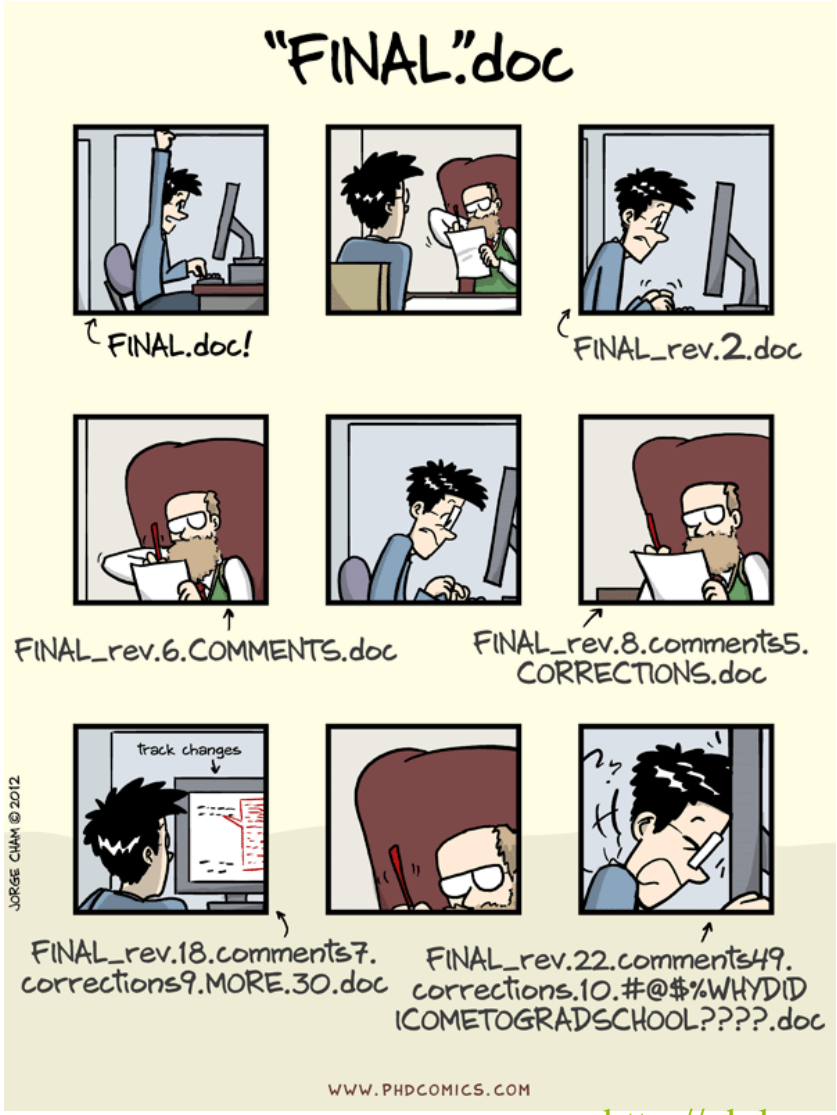
Data Stages	Examples
Raw Data	raw .txt file produced by an instrument
Processed Data	data with Z-scores calculated
Analyzed Data	rendered computational analysis
Finalized/Published Data	polished figures appear in Cell

https://twitter.com/cs3org/status/1309512729120698369/photo/1

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

File naming and directory structure

Proper file naming



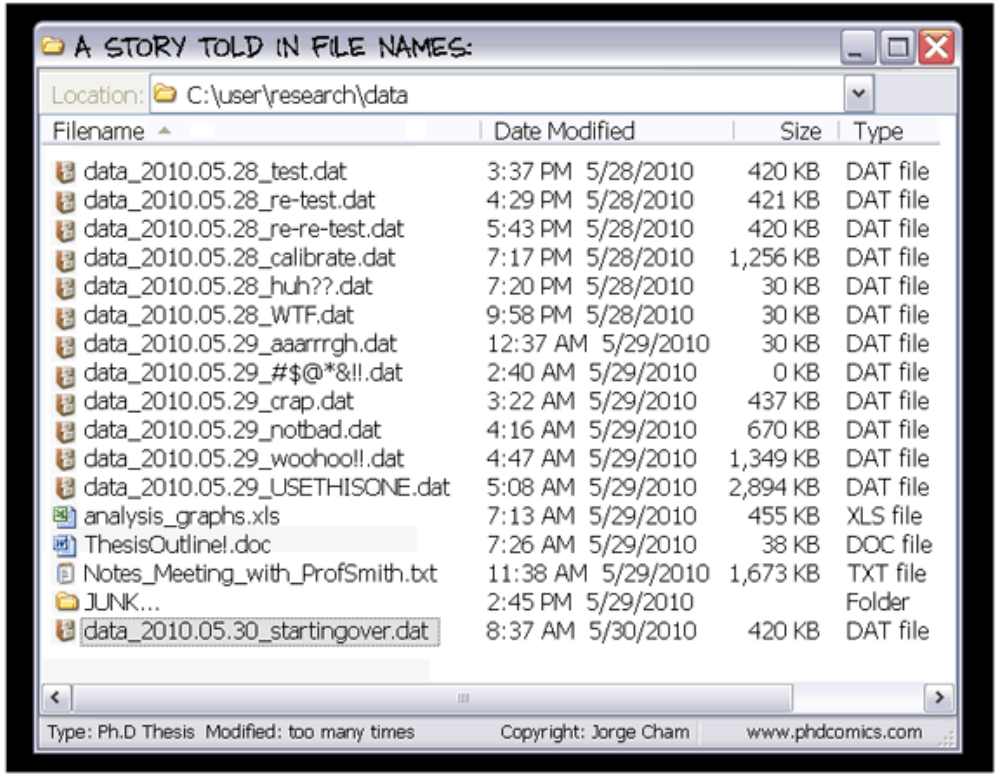
http://phdcomics.com/comics/archive_print.php?comicid=1531

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

File naming

Descriptive and consistent → document your system

- Use YYYYMMDD date format
- Sequential numbering (e.g. 001, 010)
- Project or experiment acronym
- Lab name/location
- Researcher name/initials
- Version number of file
- Avoid spaces and special characters:
  ~ ! @ # $ % ^ & * ( ) ` ; : < > ? . , [ ] { } ' " |



http://phdcomics.com/comics.php?f=1323

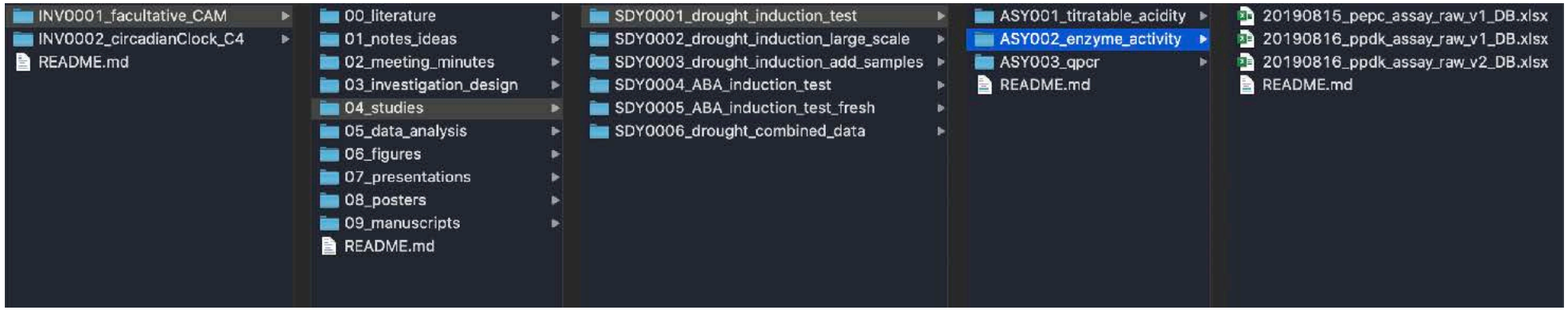
Files with naming conventions:
20161104_ProjectA_Ex1Test1_SmithE_v1.xlsx
20180204-ProjectA-Report-SmithE-v5-FINAL.docx

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

Directory structure

Any system is better than none → document your system

- Establish a system (e.g. in your lab) and use it consistently
- Example
  - One folder per project (investigation)
  - Subfolders for each study (experiment) and assay (measurement)
  - Separate folders for data or project stages
  - Date-based files folders (pairs well with lab notebook)



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

Metadata

What’s metadata and why does it matter?

The peril of missing metadata

a

Pathogen: clinical or host-associated sample from Severe acute respiratory syndrome coronavirus 2

Identifiers	BioSample: SAMN14751340; Sample name: WA-UW-6185; SRA: SRS6545124
Organism	Severe acute respiratory syndrome coronavirus 2 Viruses; Riboviria; Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Sarbecovirus; Severe acute respiratory syndrome-related coronavirus
Package	Pathogen: clinical or host-associated; version 1.0
Attributes	**isolate** WA-UW-6185 **collected by** University of Washington Virology Lab **collection date** missing **geographic location** USA **host** Homo sapiens **host disease** COVID-19 **isolation source** missing **latitude and longitude** missing
Submission	University of Washington, Pavitra Roychoudhury; 2020-04-27

b

Pathogen: clinical or host-associated sample from Severe acute respiratory syndrome coronavirus 2

Identifiers	BioSample: SAMN14656632; Sample name: hCoV-19/USA/WI-176/2020; SRA: SRS6514341
Organism	Severe acute respiratory syndrome coronavirus 2 Viruses; Riboviria; Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Sarbecovirus; Severe acute respiratory syndrome-related coronavirus
Package	Pathogen: clinical or host-associated; version 1.0
Attributes	**strain** hCoV-19/USA/WI-176/2020 **isolate** Homo sapien **collected by** Milwaukee Public Health Department **collection date** 2020-03-20 **geographic location** USA: Milwaukee, Wisconsin **host** Homo sapiens **host disease** COVID-19 **isolation source** nasal swab **latitude and longitude** 43.042180 N 87.908670 W **ARTIC barcode identifiers** NB23

Fig. 1 Lost opportunities for data reuse, SARS-CoV-2 (txid2697049[Organism:noexp]) BioSample records, where (**a**) **collection date** = "missing": 143; **latitude and longitude** = "missing": 1375; (**b**) SARS-CoV-2 BioSample record with complete metadata.

Schriml, L. M., Chuvochina, M., Davies, N., Eloe-Fadrosh, E. A., Finn, R. D., Hugenholtz, P., et al. (2020). COVID-19 pandemic reveals the peril of ignoring metadata standards. Scientific Data, 1–4. http://doi.org/10.1038/s41597-020-0524-5

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

Metadata

- "Data about data"
  - e.g. creator, project, PI, date
- Description of data
  - context, protocol
  - experimental & analysis information
  - Information to make the data
    - Understandable, reproducible, or reusable
- Would a new project member be able to step in and know how the data was created?

Documenting your metadata is key to reproducible science!!

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

Metadata

Who collects metadata and how?

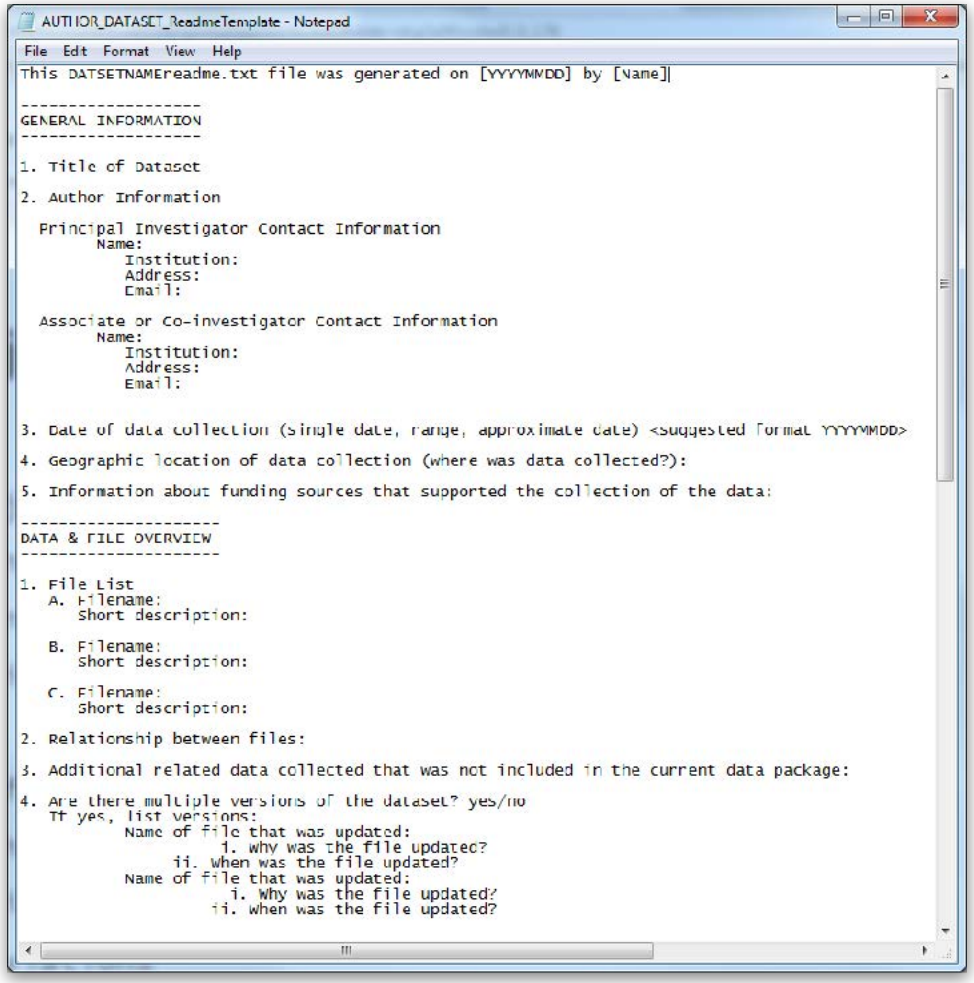
Example README File

Create a plain text file (README.txt) to document:

- Basic project information
- Title, Contributions, Grant Info
- Contact information
- All locations of where data live, including backups
- Useful information about the files and how they are organized
- Explain file naming conventions and abbreviations

Precede any comment about the data with “#”

Have a README file for each distinct dataset



Example Template: http://data.research.cornell.edu/content/readme

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

Metadata standards

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

- Examples (check https://fairsharing.org for more…)
  - MIAPPE | Minimum Information About a Plant Phenotyping Experiment
    https://www.miappe.org
  - MIAME | Minimum Information About a Microarray Experiment
    http://fged.org/projects/miame/
  - MIAPE | Minimum Information About a Proteomics Experiment
    http://www.psidev.info/miape
  - MINSEQE | Minimum Information about a high-throughput SEQuencing Experiment
    http://fged.org/projects/minseqe/

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

A closer look at MIAPPE

	MIAPPE				
line #	MIAPPE Check list	Definition	Example	Format	Cardinality
DM-1	**Investigation**	Investigations are research programmes with defined aims. They can exist at various scales (for example, they could encompass a grant-funded programme of work, the various components comprising a peer-reviewed publication, or a single experiment).			1 per MIAPPE submission
DM-2	**Investigation unique ID**	Identifier comprising the unique name of the institution/database hosting the submission of the investigation data, and the accession number of the investigation in that institution.	EBI:12345678	Unique identifier	0-1
DM-3	**Investigation title**	Human-readable string summarising the investigation.	Adaptation of Maize to Temperate Climates: Mid-Density Genome-Wide Association Genetics and Diversity Patterns Reveal Key Genomic Regions, with a Major Contribution of the Vgt2 (ZCN8) Locus.	Free text (short)	1
DM-4	**Investigation description**	Human-readable text describing the investigation in more detail.	The migration of maize from tropical to temperate climates was accompanied by a dramatic evolution in flowering time. To gain insight into the genetic architecture of this adaptive trait, we conducted a 50K SNP-based genome-wide association and diversity investigation on a panel of tropical and temperate American and European representatives.	Free text	0-1
DM-5	**Submission date**	Date of submission of the dataset presently being described to a host repository.	2012-12-17	Date/Time (ISO 8601, optional time zone)	0-1
DM-6	**Public release date**	Date of first public release of the dataset presently being described.	2013-02-25	Date/Time (ISO 8601, optional time zone)	0-1
DM-7	**License**	License for the reuse of the data associated with this investigation. The Creative Commons licenses cover most use cases and are recommended.	CC BY-SA 4.0, Unreported	Unique identifier	0-1
DM-8	**MIAPPE version**	The version of MIAPPE used.	1.1	Version number	1
DM-9	**Associated publication**	An identifier for a literature publication where the investigation is described. Use of DOIs is recommended.	doi:10.1371/journal.pone.0071377	DOI	0+
DM-10	**Study**	A study (or experiment) comprises a series of assays (or measurements) of one or more types, undertaken to answer a particular biological question.			1+ per investigation
DM-11	**Study unique ID**	Unique identifier comprising the name or identifier for the institution/database hosting the submission of the study data, and the identifier of the study in that institution.	EBI:12345678 http://phenome-fppn.fr/maugio/2013/t2351	Unique identifier	0-1
DM-12	**Study title**	Human-readable text summarising the study	2002 evaluation of flowering time for a panel of 375 maize lines at the experimental station of Maugio (France).	Free text (short)	1
DM-13	**Study description**	Human-readable text describing the study	2002 evaluation of male and female flowering time for a panel of 375 maize lines representing the worldwide genetic diversity at the experimental station of Maugio, France.	Free text	0-1
DM-14	**Start date of study**	Date and, if relevant, time when the experiment started	2002-04-04 2006-09-27T10:23:21+00:00	Date/Time (ISO 8601, optional time zone)	1
DM-15	**End date of study**	Date and, if relevant, time when the experiment ended	2002-11-27	Date/Time (ISO 8601, optional time zone)	0-1
DM-16	**Contact institution**	Name and address of the institution responsible for the study.	UMR de Génétique Végétale, INRA – Université Paris-Sud – CNRS, Gif-sur-Yvette, France	Free text (short)	1
DM-17	**Geographic location (country)**	The country where the experiment took place, either as a full name or preferably as a 2-letter code.	FR	Country name or 2-letter code (ISO 3166)	1
DM-18	**Experimental site name**	The name of the natural site, experimental field, greenhouse, phenotyping facility, etc. where the experiment took place.	INRA, UE Diascope - Chemin de Mezouls - Domaine expérimental de Melgueil - 34130 Mauguio - France	Free text (short)	1
DM-19	**Geographic location (latitude)**	Latitude of the experimental site in degrees, in decimal format.	+43.619264	Degrees in the decimal format (ISO 6709)	0-1 (1 if longitude is provided)
DM-20	**Geographic location (longitude)**	Longitude of the experimental site in degrees, in decimal format.	+3.967454	Degrees in the decimal format (ISO 6709)	0-1 (1 if latitude is provided)
DM-21	**Geographic location (altitude)**	Altitude of the experimental site, provided in metres (m).	100 m	Numeric + unit abbreviation	0-1
DM-22	**Description of the experimental design**	Short description of the experimental design, possibly including statistical design. In specific cases, e.g. legacy datasets or data computed from several studies, the experimental design can be "unknown"/"NA", "aggregated/reduced data", or simply 'none'.	Lines were repeated twice at each location using a complete block design. In order to limit competition effects, each block was organized into four sub-blocks corresponding to earliness groups based on a priori information.	Free text	1
DM-23	**Type of experimental design**	Type of experimental design of the study, in the form of an accession number from the Crop Ontology.	CO_715:0000145	Crop Ontology term (subclass of "CO_715:0000003")	0-1
DM-24	**Observation unit level hierarchy**	Hierarchy of the different levels of repetitions between each others	block>rep>plot	Formatted text (level>level)	0-1

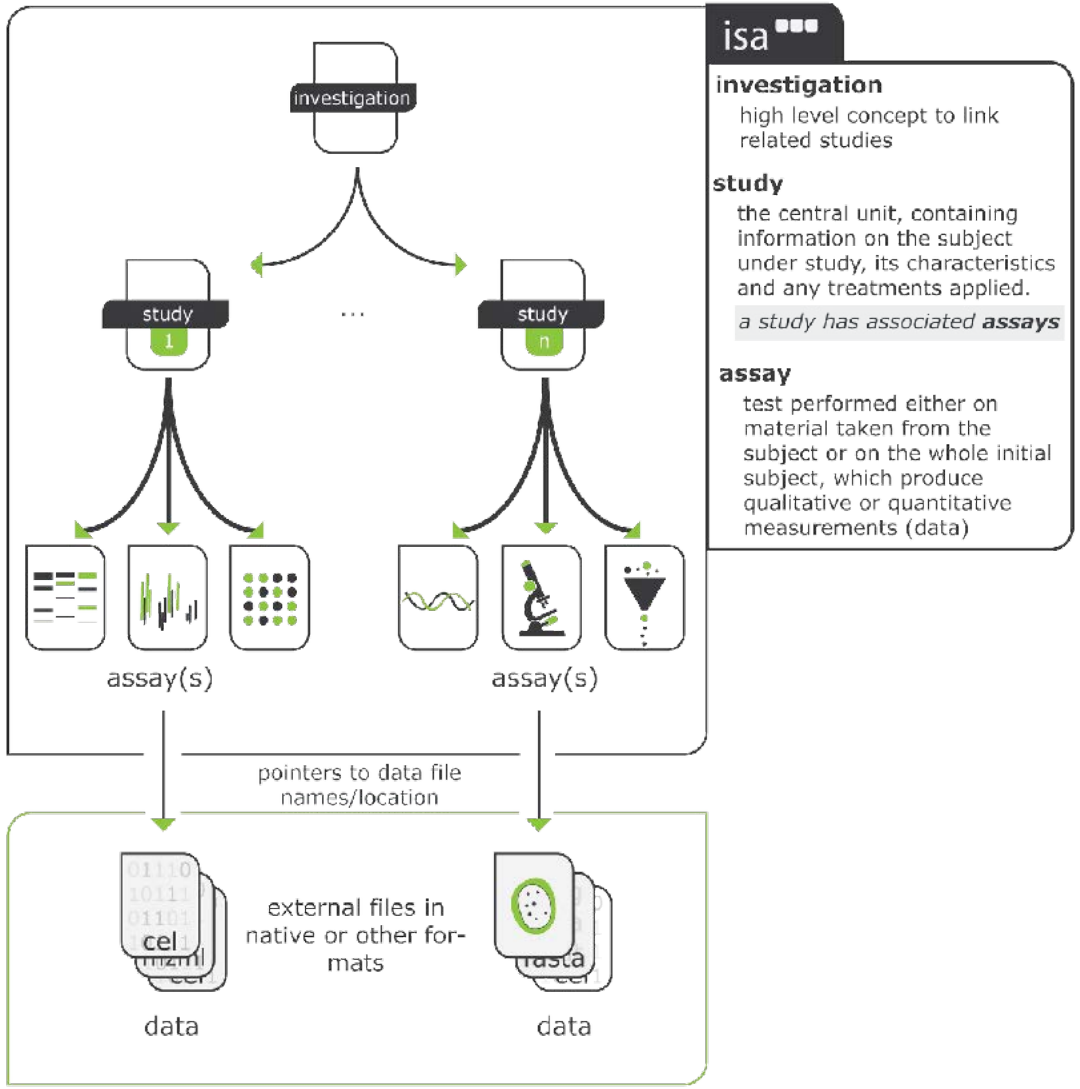
…this is just an excerpt…

https://github.com/MIAPPE/MIAPPE/blob/master/MIAPPE_Checklist-Data-Model-v1.1/MIAPPE_Checklist-Data-Model-v1.1.pdf

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

The ISA Model of (experimental) metadata

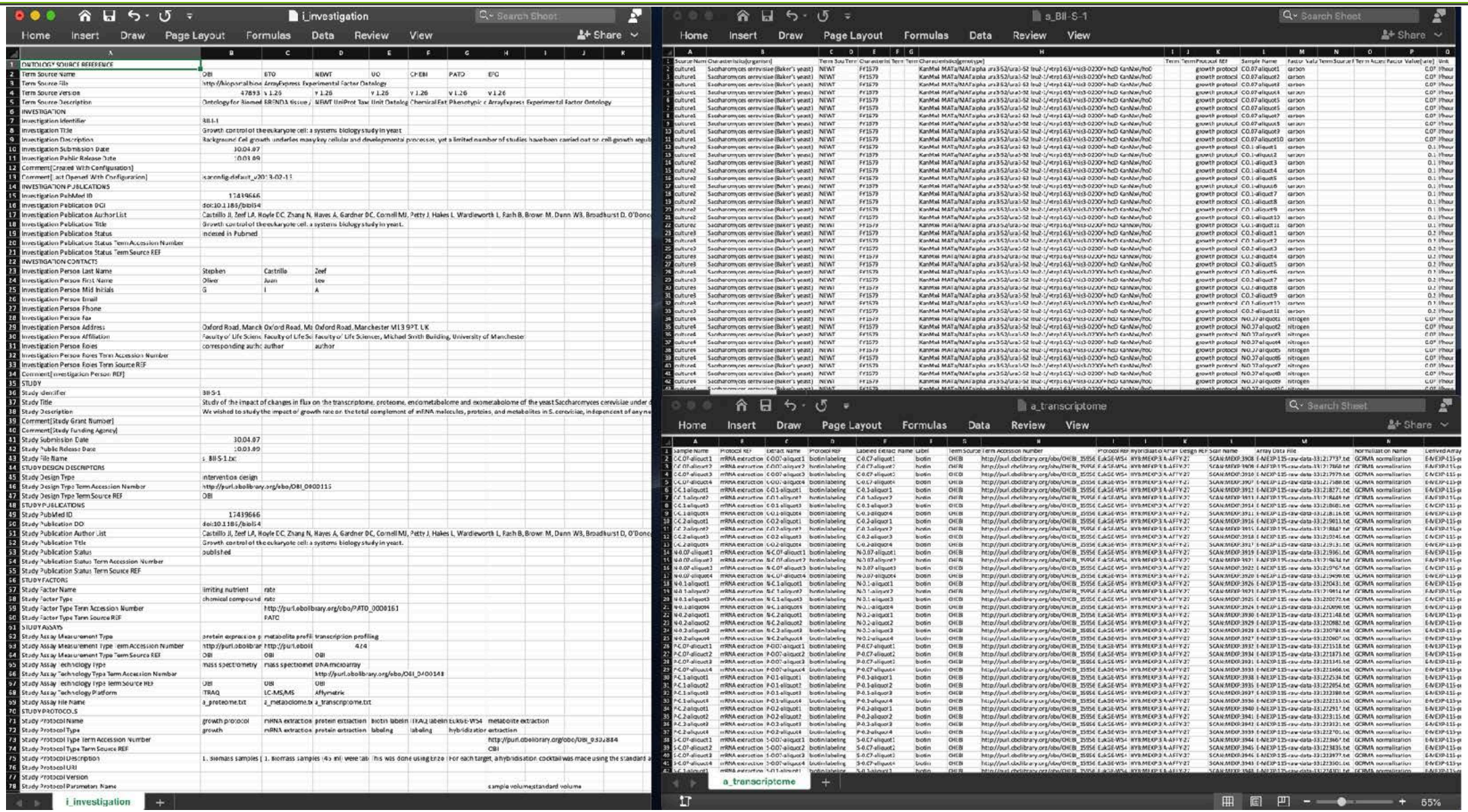
- **Investigation**
  - Overall goals
  - Scientific context
- **Study**
  - Experimental steps
- **Assay**
  - Leading to (raw) data



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

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

A closer look at ISA(-tab)



Example: https://github.com/ISA-tools/ISAdatasets/tree/master/tab/BII-I-1

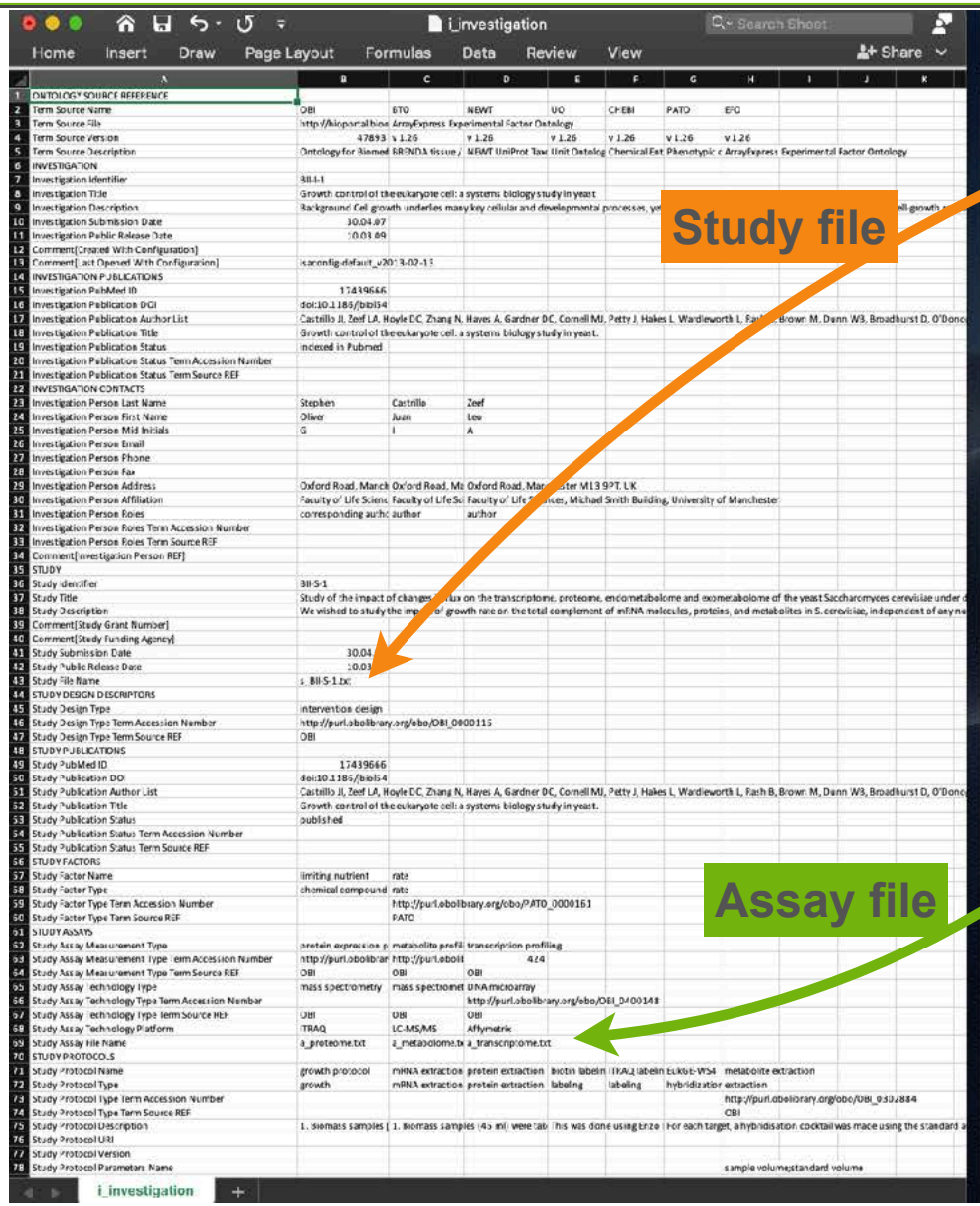
hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

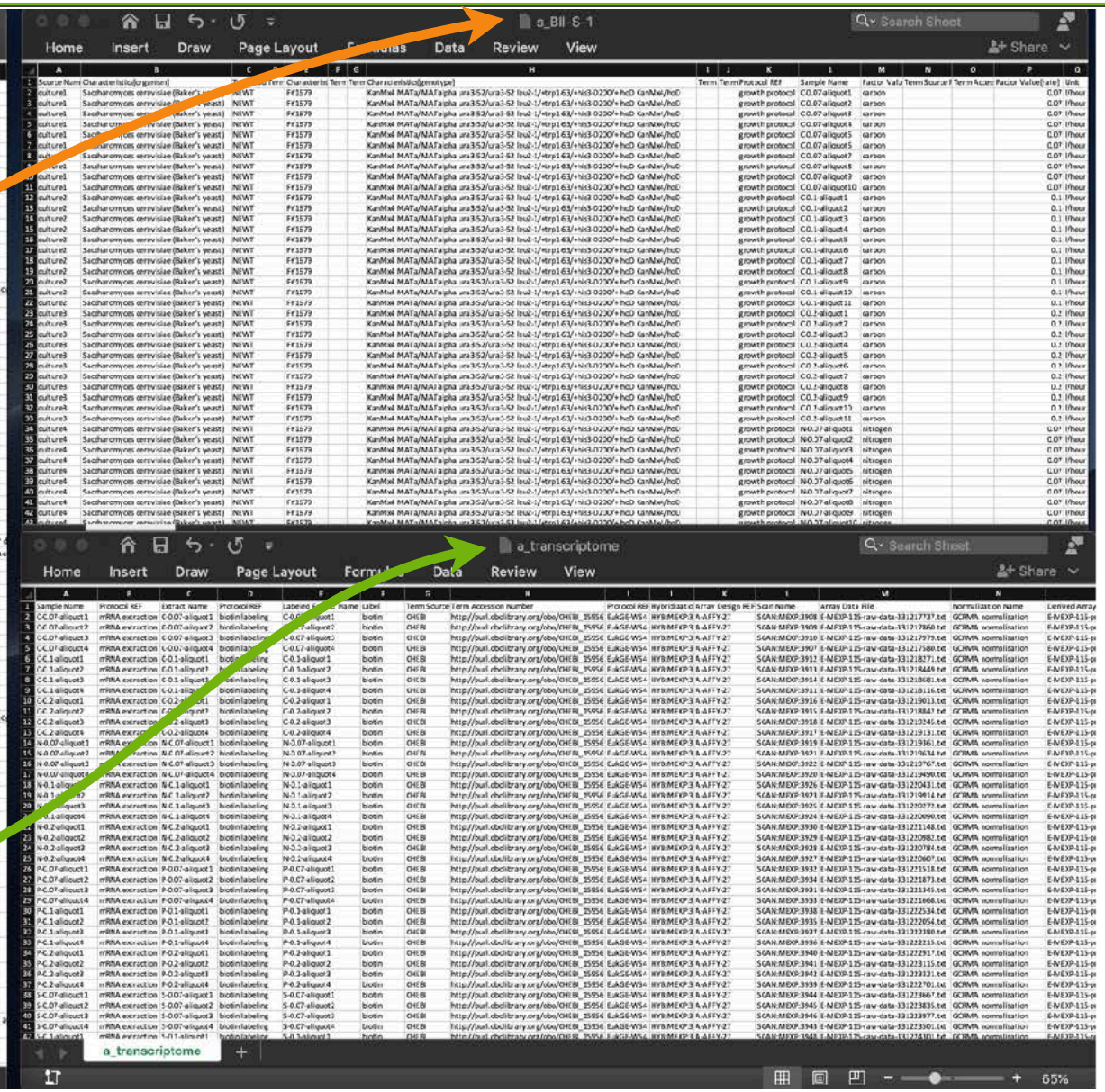
A closer look at ISA(-tab)

Investigation

Study file

Assay file





Study

Assay

Example: https://github.com/ISA-tools/ISAdatasets/tree/master/tab/BII-I-1

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

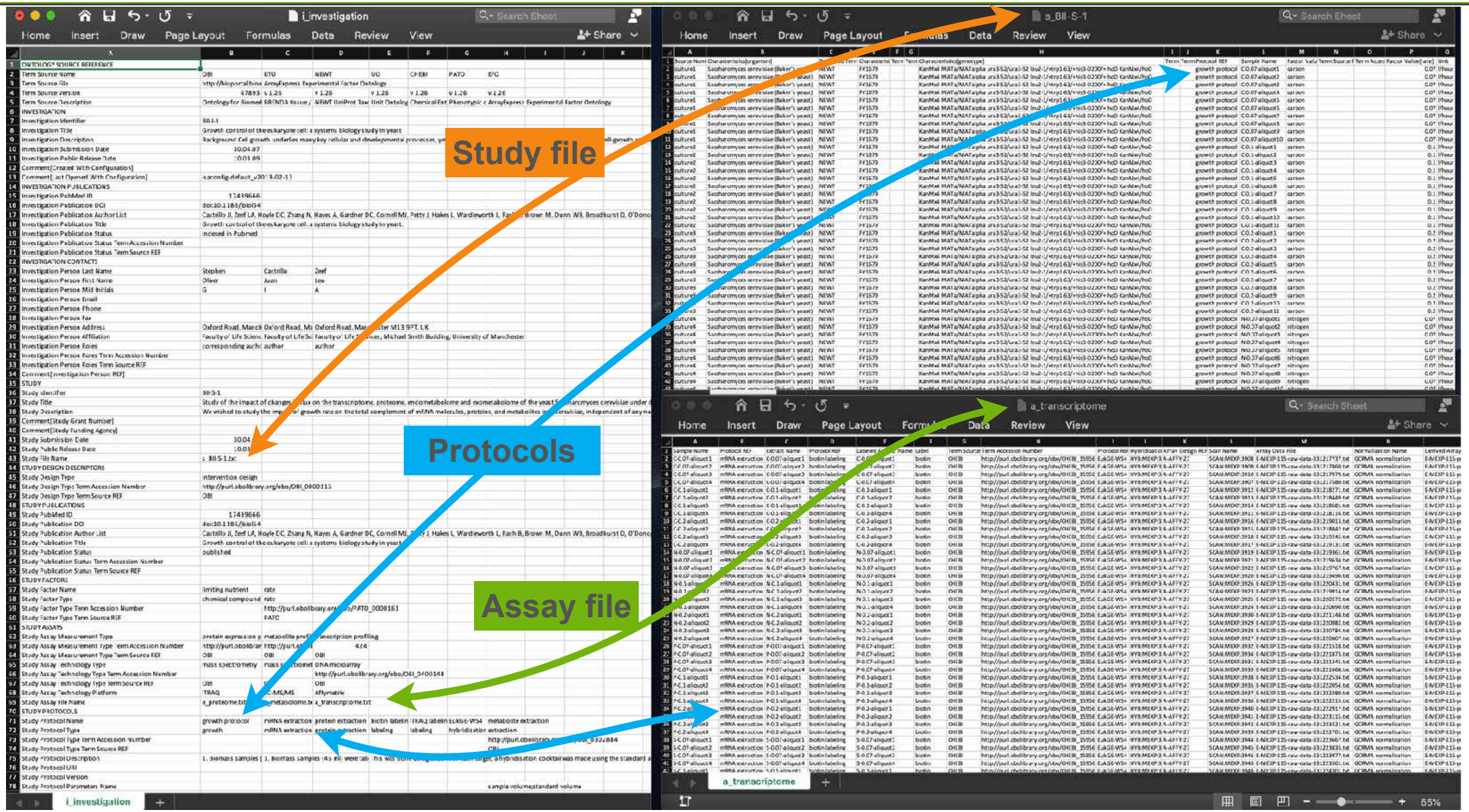
A closer look at ISA(-tab)

Investigation

Study

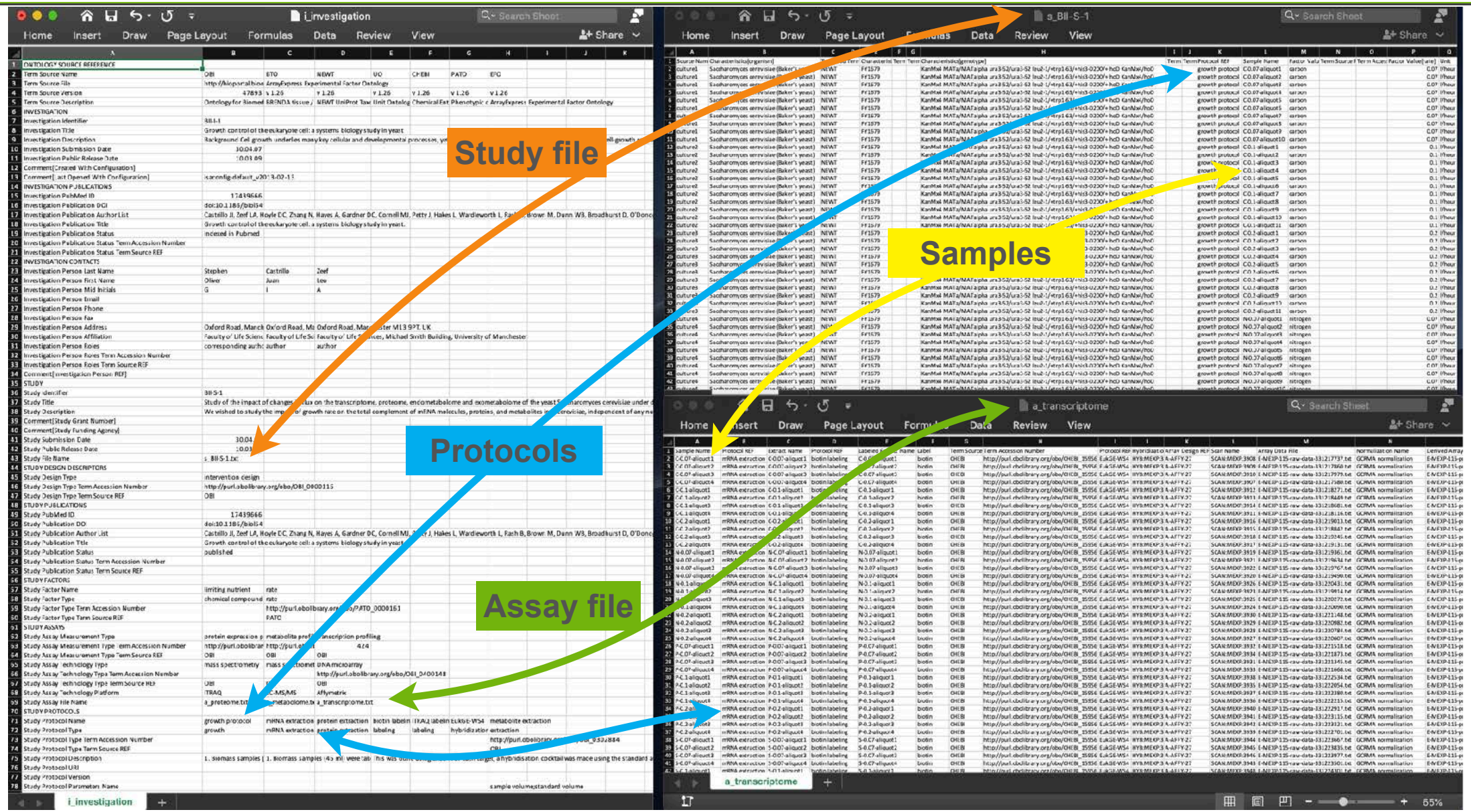
Assay

Example: https://github.com/ISA-tools/ISAdatasets/tree/master/tab/BII-I-1



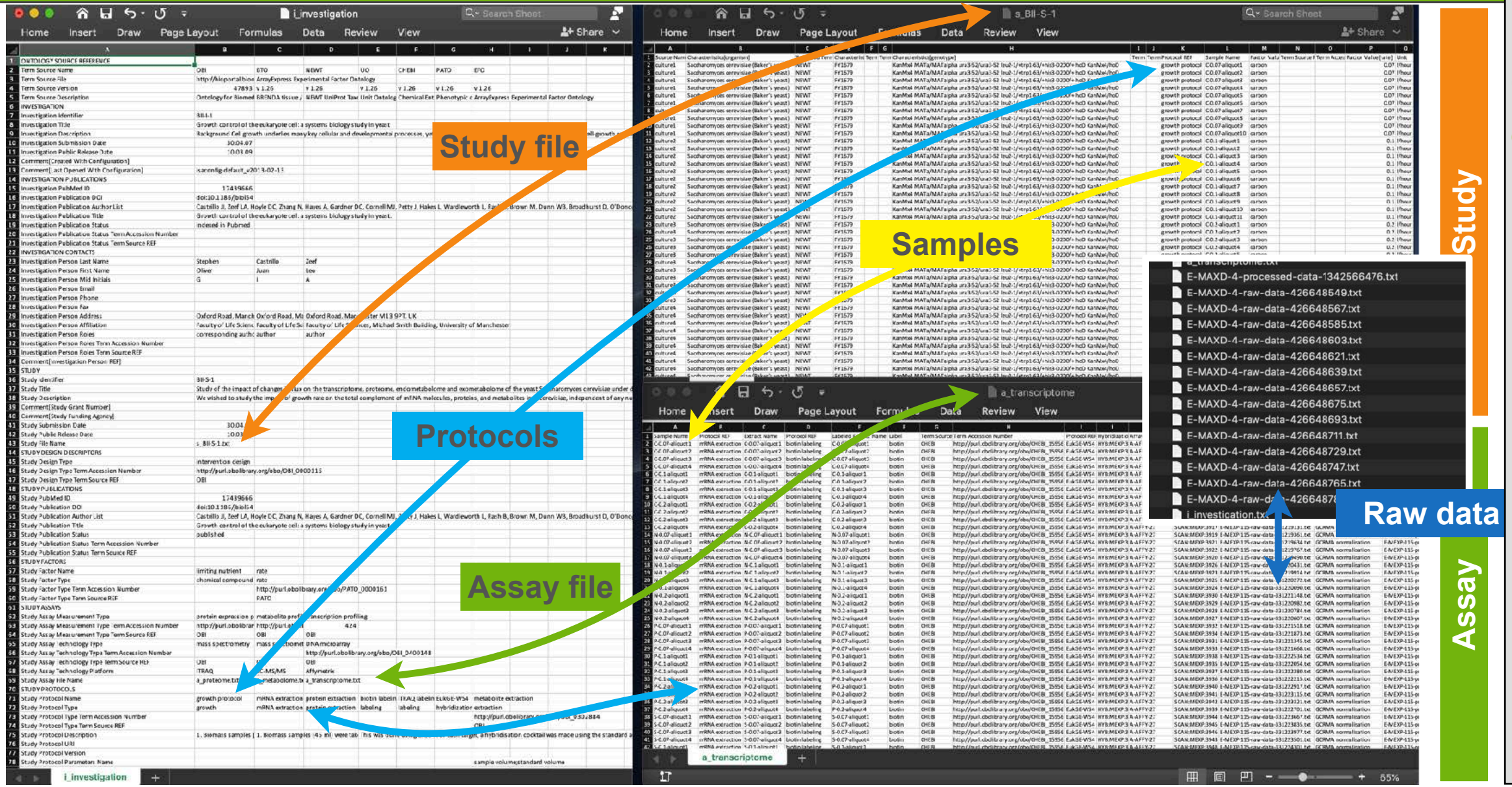
hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

A closer look at ISA(-tab)



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

A closer look at ISA(-tab)



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

Connecting (i.e. “mapping”) MIAPPE and ISA

Table 1 Mapping between MIAPPE and ISA-Tab sections.

MIAPPE section	ISA-Tab section	ISA-Tab section specification
Investigation	Investigation/investigation publications	
Study	Study/study design descriptors/study protocols	
Person	Investigation contacts/study contacts	
Data file	Study	With comment fields
Biological material	Source	
Environment	Study protocols	Growth type protocol
Experimental factor	Study Factors	
Event	Study protocols	Event type protocols and external Events file
Observation unit	Sample	
Sample	Extract/study protocols	Sampling type protocol
Observed variable	Observed variable	In external trait definition file

The table lists the MIAPPE sections with the ISA-Tab sections holding their fields. MIAPPE-exclusive fields have been added as comments in the corresponding sections. The detailed mapping can be found in Supporting Information Table S1, and in the MIAPPE repository (https://github.com/MIAPPE/MIAPPE/tree/master/MIAPPE_Checklist-Data-Model-v1.1/MIAPPE_mapping).

Papoutsoglou, E. A., Faria, D., Arend, D., Arnaud, E., Athanasiadis, I. N., Chaves, I., et al. (2020). Enabling reusability of plant phenomic datasets with MIAPPE 1.1. New Phytologist, 227(1), 260–273. http://doi.org/10.1111/nph.16544

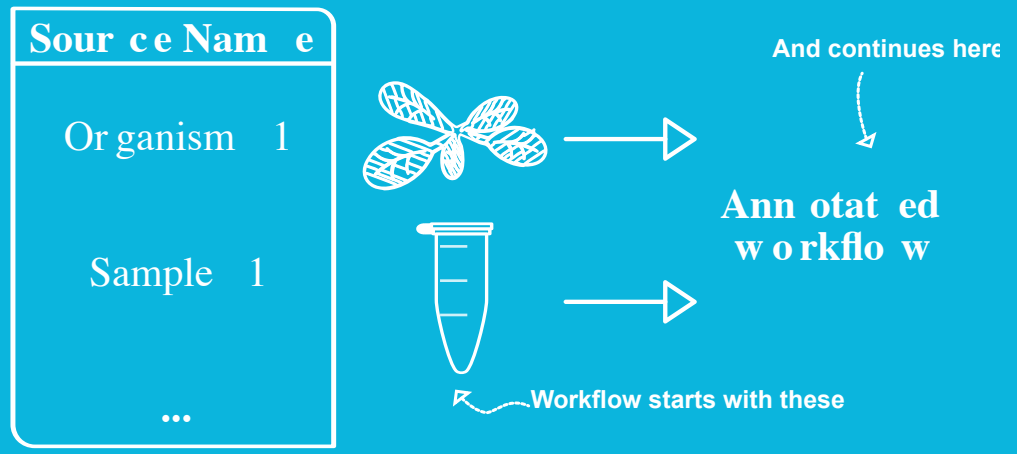
hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

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



https://nfdi4plants.github.io/AnnotationPrinciples/

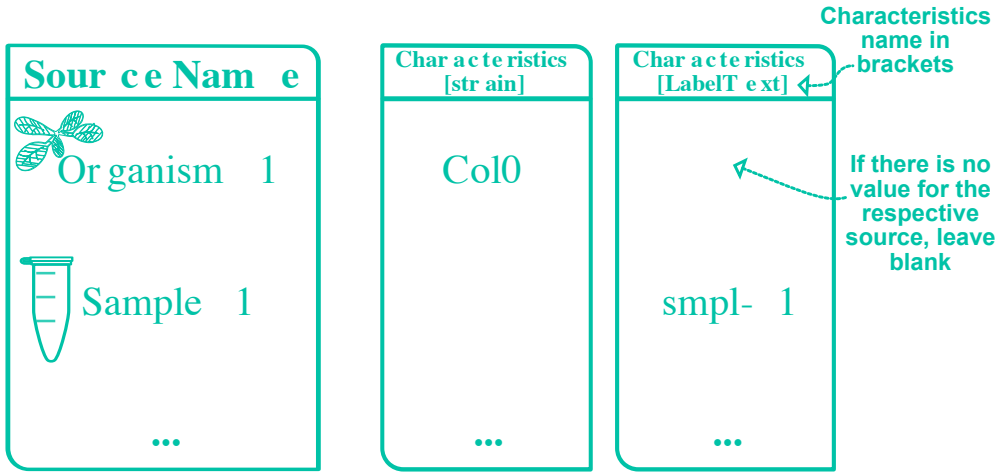
hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

Annotation principles

Characteristics

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

You can use any number of characteristics columns.



https://nfdi4plants.github.io/AnnotationPrinciples/

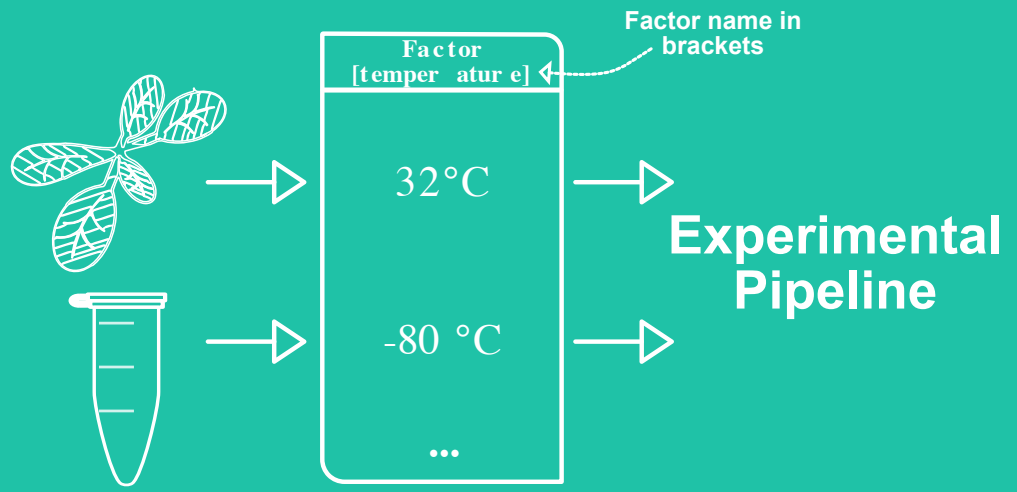
hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

Annotation principles

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.



https://nfdi4plants.github.io/AnnotationPrinciples/

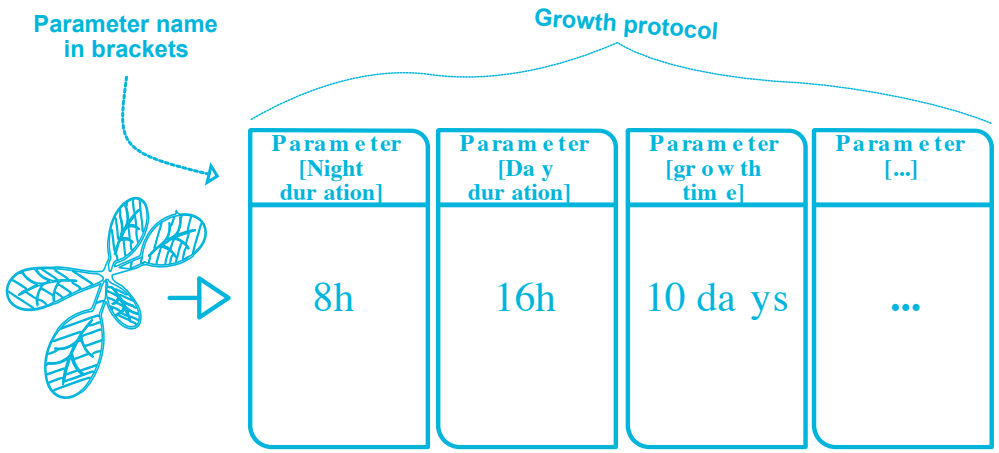
hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

Annotation principles

Parameter

Use parameters to annotate your experimental workflow.

You can group parameters to create a protocol.



https://nfdi4plants.github.io/AnnotationPrinciples/

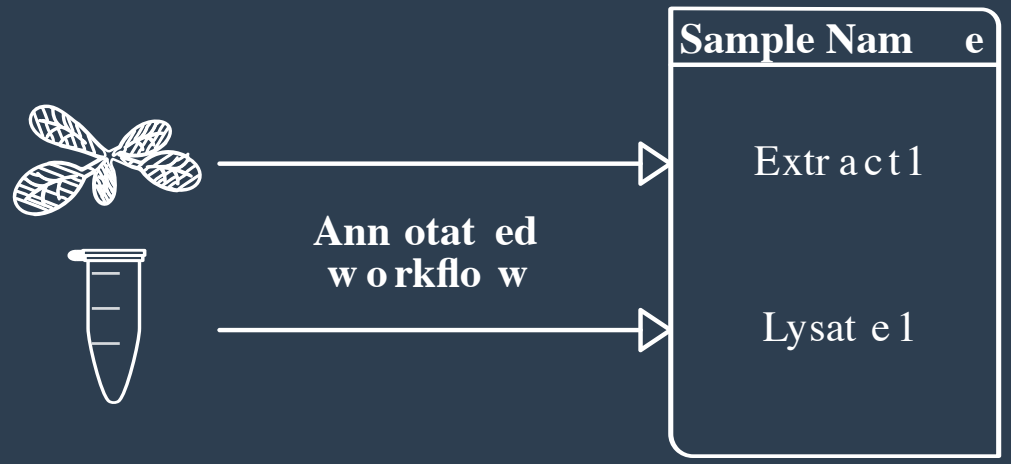
hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

Annotation principles

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.



https://nfdi4plants.github.io/AnnotationPrinciples/

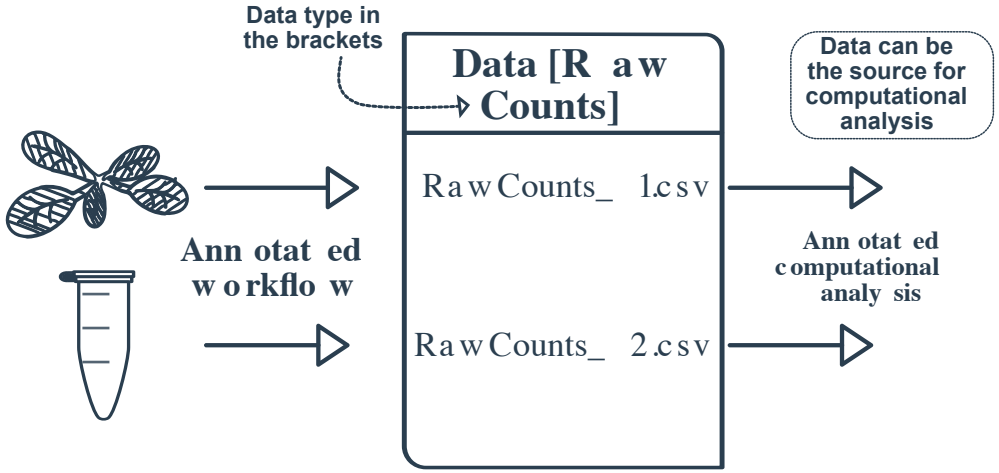
hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

Annotation principles

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.



https://nfdi4plants.github.io/AnnotationPrinciples/

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

A quick word on *words*

Kai joins a CEPLAS lab to pursue a PhD.
His supervisor Frida asks Kai to work on biomass conversion in algae.
Kai

(1) grows the algae on different ^{13}C-labelled carbon sources
(2) harvests some cells
(3) extracts metabolites
(4) submits the extracts to a metabolomics facility
(5) where they are measured
(6) and receives a nice dataset to be analyzed

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

A quick word on *words*

Kai joins a CEPLAS lab to pursue a PhD.
His supervisor Frida asks Kai to work on biomass conversion in algae.
Kai

(1) grows the algae on different ^{13}C-labelled carbon sources
(2) harvests some cells
(3) extracts metabolites
(4) submits the extracts to a metabolomics facility
(5) where they are measured
(6) and receives a nice dataset to be analyzed

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

Ontologies

Standard for defining terms

- 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
- Example: EDAM ontology
  - http://edamontology.org/page
  - https://ifb-elixirfr.github.io/edam-browser/
- Searching for desired terms at
  - https://bioportal.bioontology.org/

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

Ontologies for MIAPPE and ISA models

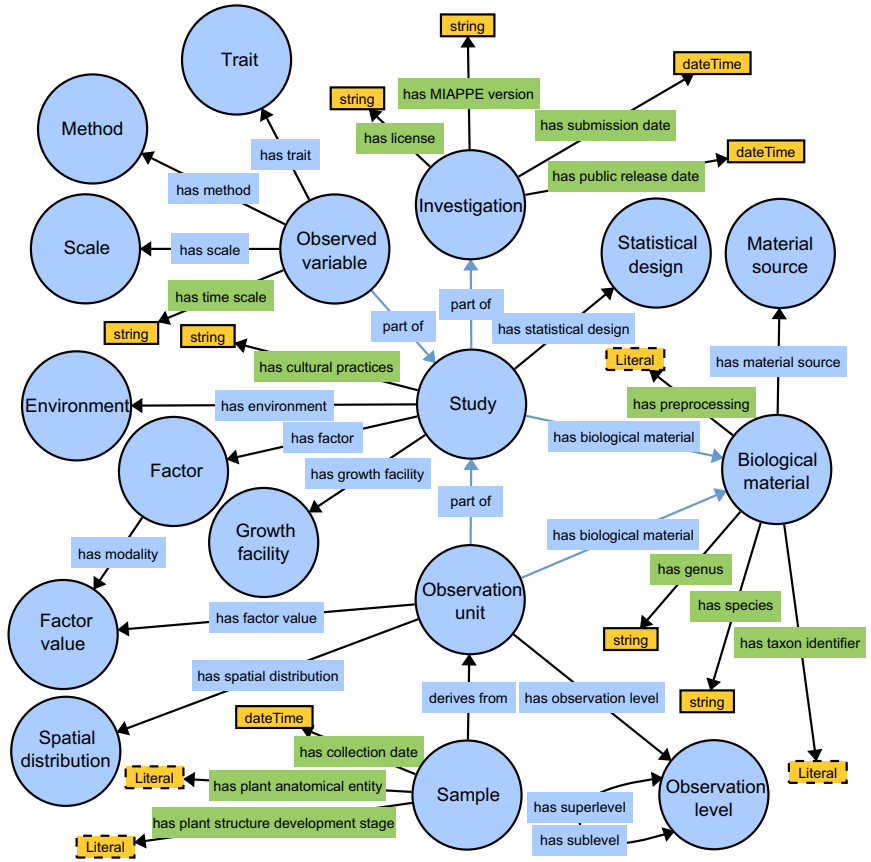
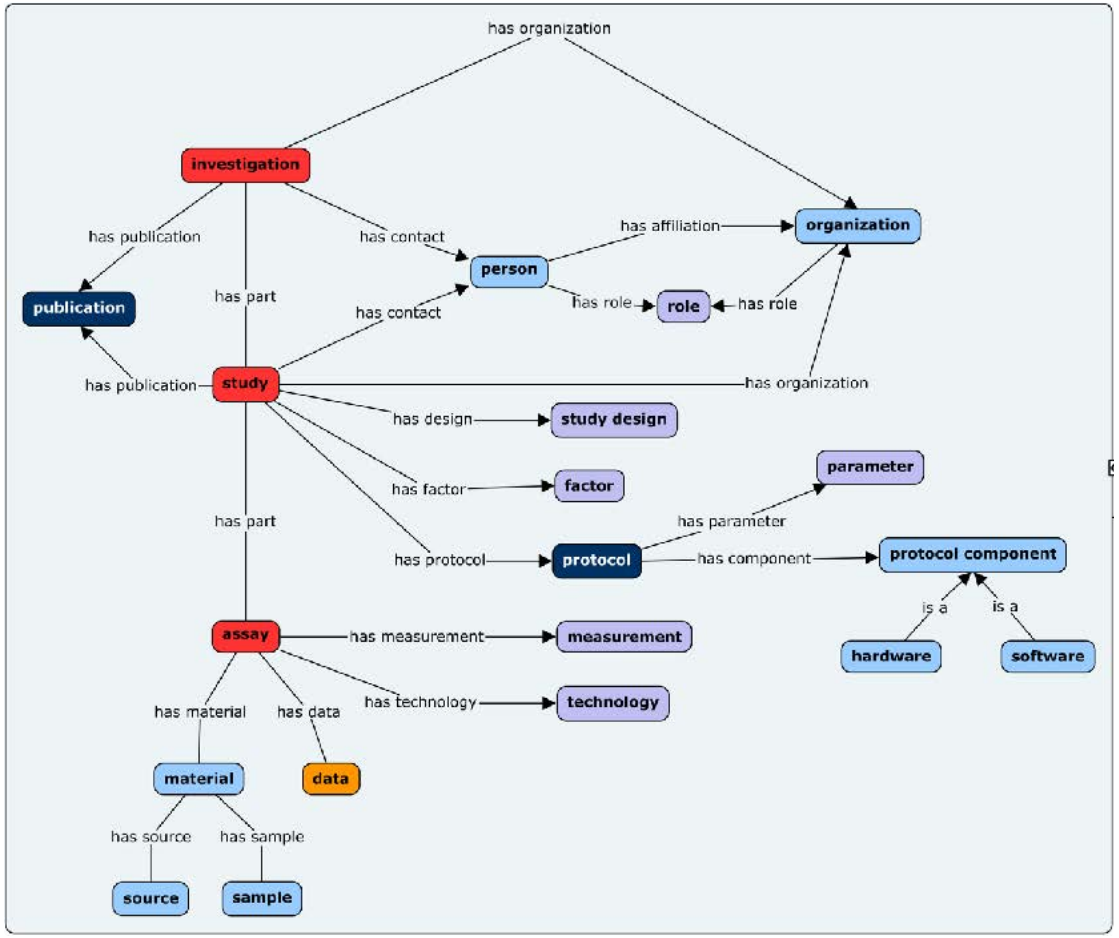


Fig. 1 Subset of the Plant Phenotyping Experiment Ontology representing the MIAPPE data model. Generated using WebVOWL (http://editor.visualdataweb.org/) and edited manually. Circles indicate classes. Object properties are shown in blue rectangles, and data properties are shown in green rectangles. Yellow rectangles represent literals.



Papoutsoglou, E. A., Faria, D., Arend, D., Arnaud, E., Athanasiadis, I. N., Chaves, I., et al. (2020). Enabling reusability of plant phenomic datasets with MIAPPE 1.1. New Phytologist, 227(1), 260–273. http://doi.org/10.1111/nph.16544

https://isa-specs.readthedocs.io/en/latest/isamodel.html#

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

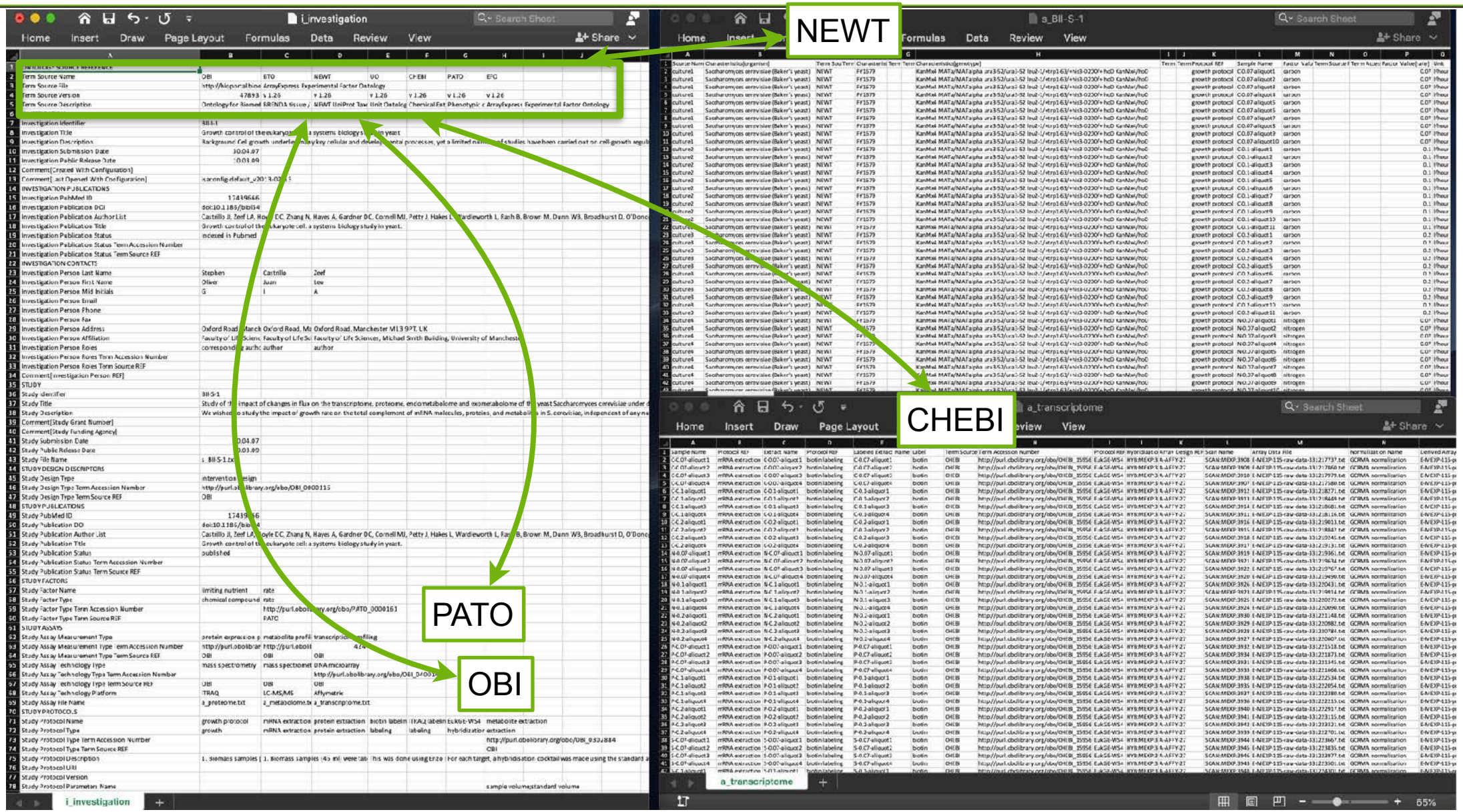
Use of ontologies in ISA

Investigation

Study

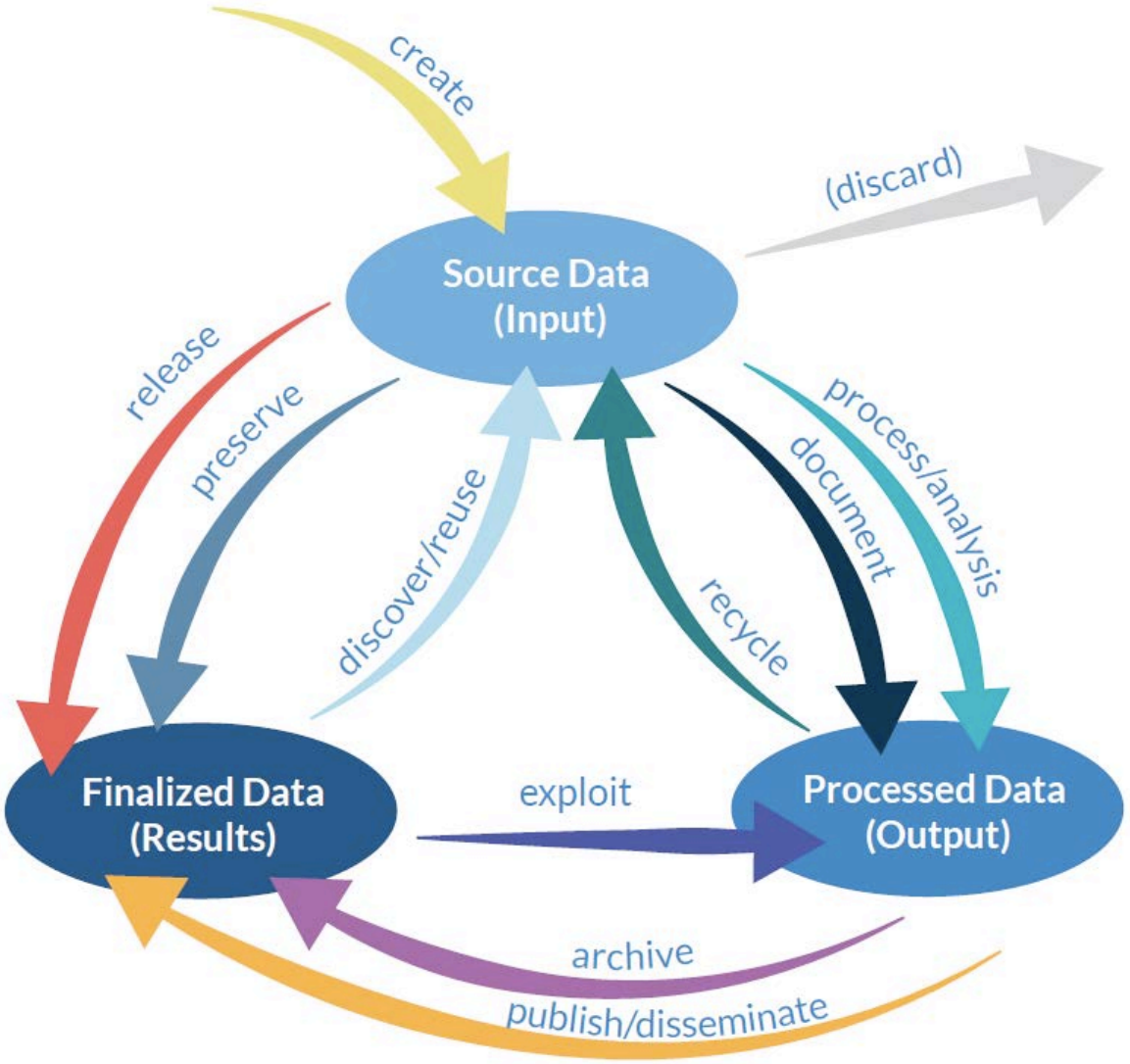
Assay

Example: https://github.com/ISA-tools/ISAdatasets/tree/master/tab/BII-I-1



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

Processed(ing) Data



https://vidensportal.deic.dk/sites/default/files/uploads/FAIR/Diagram.png

Versioning and version control

Versioning

- It’s good to document:
  - What was changed?
  - Who is responsible?
  - When did it happen?
  - Why the changes?

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

Git and Github hands-on

Why use git?

- Distributed version control system
  - Records changes to our files over time
  - Recall any version at any time
  - Allows easy collaboration on a project
    - Have own version of project files on their own computers

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

Version control

Revision control, source control

- Managing multiple versions of documents
  - Yourself
    - Time machine to go back to older versions
  - Team projects
    - Simplifies concurrent work & merging changes

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

Why use git?

- Store revisions as project history in one directory
- Rewind to any revision in the project
- Work on new features without messing with the main code
- Easily collaborate with others

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

Github

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

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

Installing git

Basic commands

- https://git-scm.com/downloads
- ubuntu : sudo apt install git-all
-

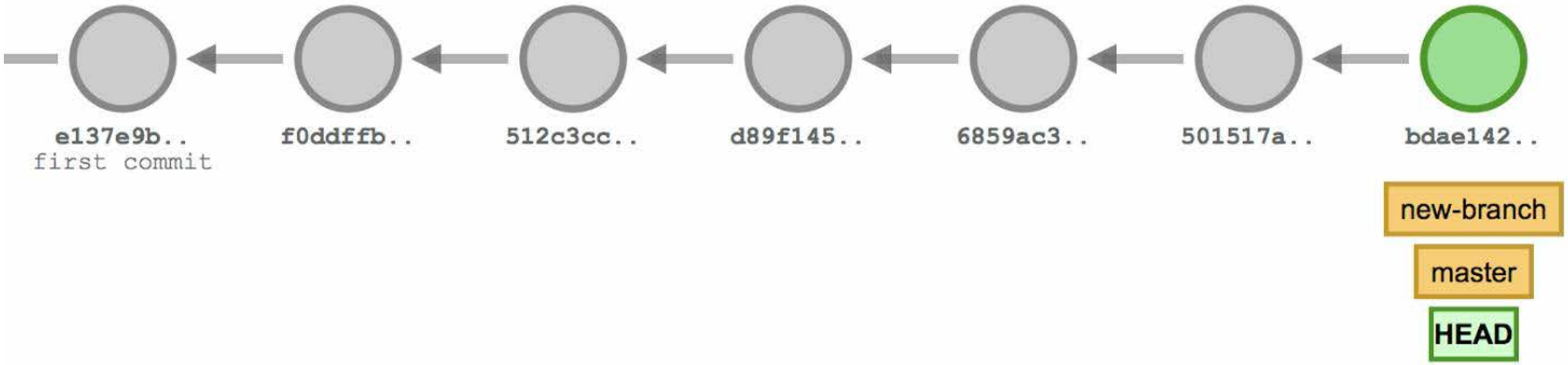
hajira.jabeen@uni-koeln.de, dominik.brilhaus@hhu.de

Repositories

- A repo is a container for a project that you want to track with Git
- One can have different repos for different projects on your computer
- Like project folders where git tracks the contents for us
- - - demo

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

Git commits and history

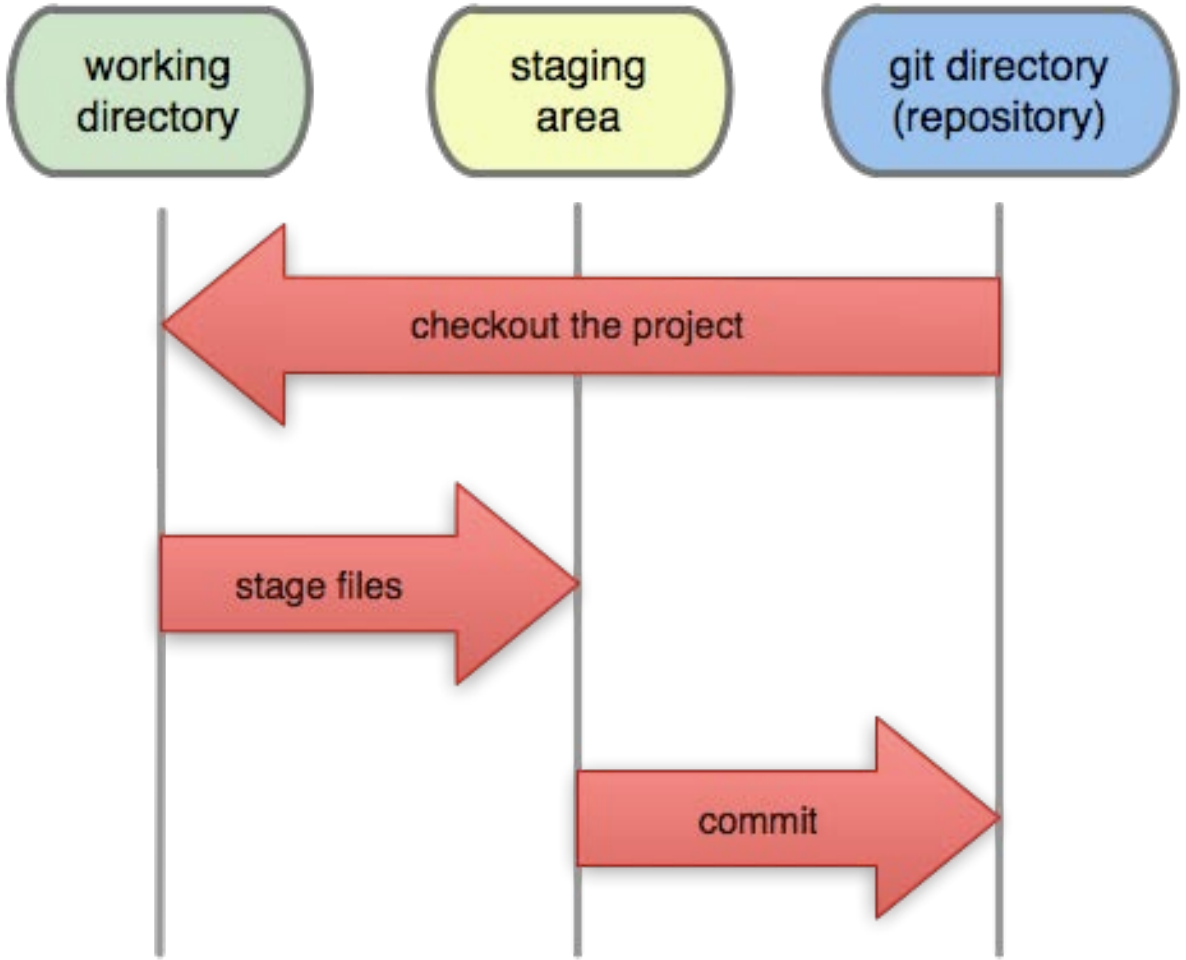


https://trailhead.salesforce.com/de/content/learn/modules/git-and-git-hub-basics/work-with-your-history-in-git?trail_id=set-up-your-workspace-and-install-developer-tools

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

Git workflow

- Modified
  - Changed files, not committed
- Staging
  - Add any changed files to staging that will be committed
- Committed
  - Files in the staging area are finally committed to the remote repo



https://i.stack.imgur.com/zLTpo.png

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

Create a Git Repo

- Checking the status
- Staging the changes
- Committing the changes
- Summary of commits
-

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

Commit history

- Checkout commit
- Revert commit
- Reset commit

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

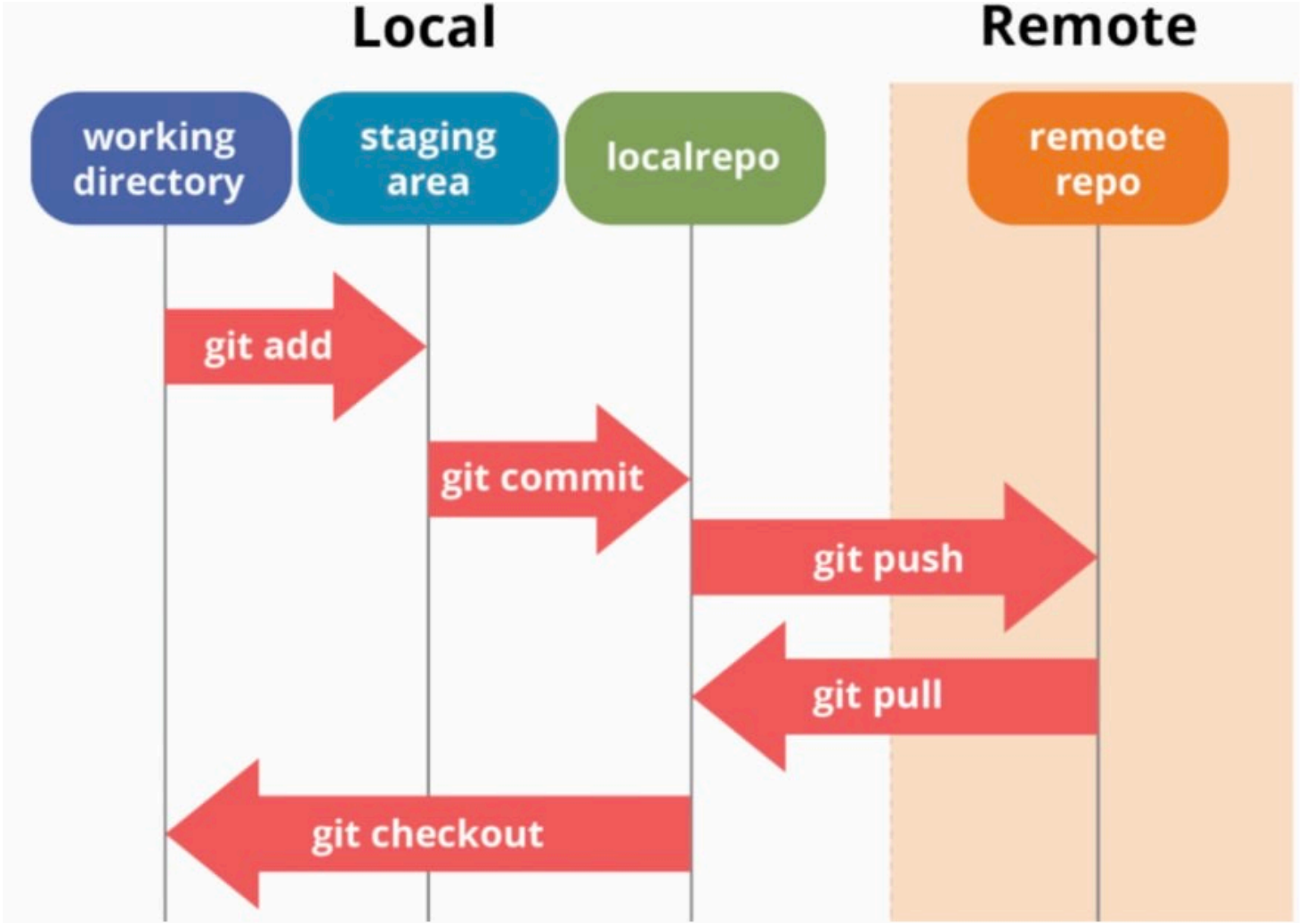
Working with branches

- Creating a branch
- Checking out a branch
- Deleting a branch
- Merging a branch with master
-

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

Git remote repo

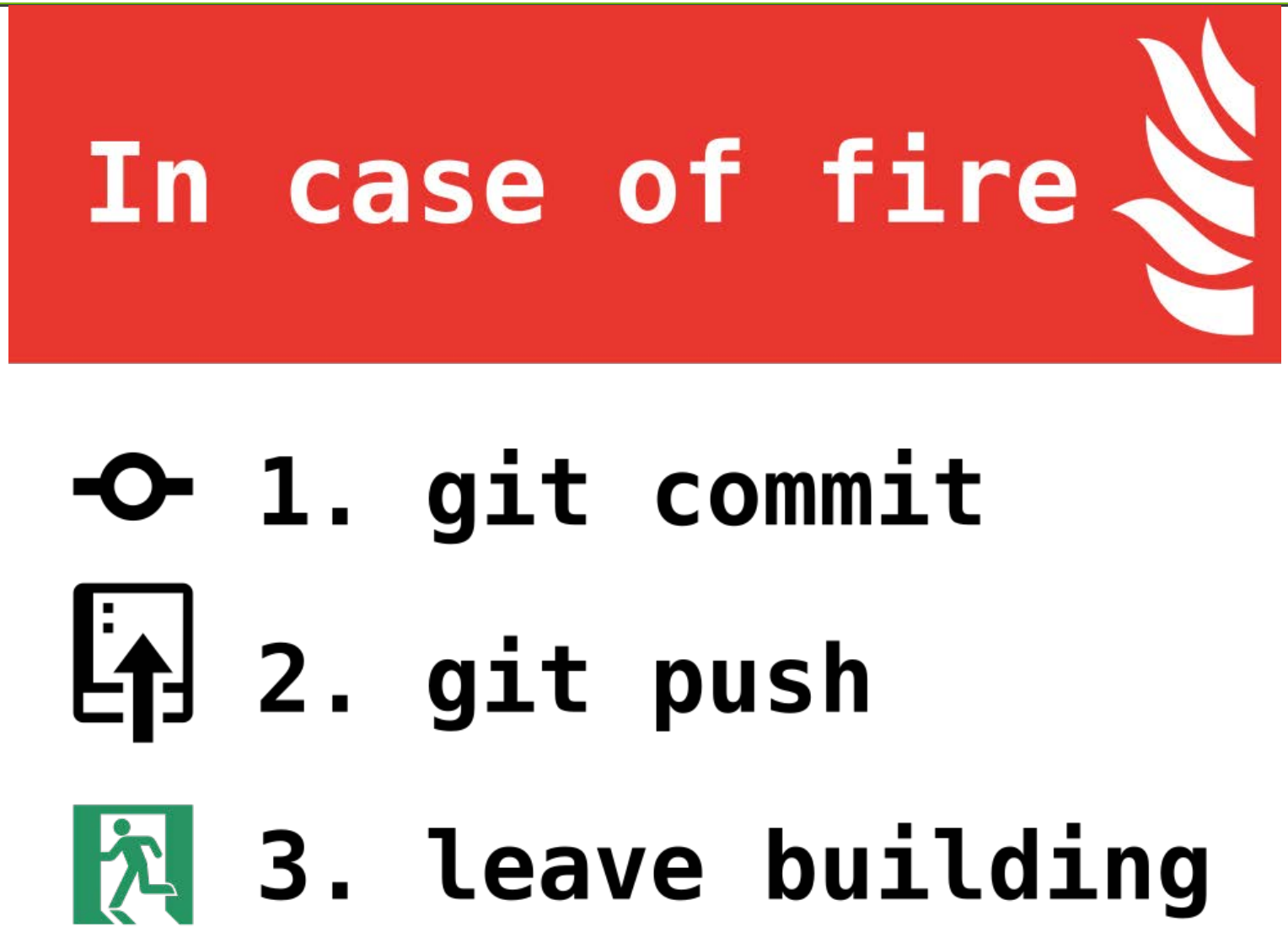
- Sign-in to github
- Create a remote repo
- Upload your changes
- Using remote repos



https://res.cloudinary.com/practicaldev/image/fetch/s--M_fHUEqA--/c_limit%2Cf_auto%2Cfl_progressive%2Cq_auto%2Cw_880/https://thepracticaldev.s3.amazonaws.com/i/128hsgntnsu9bww0y8sz.png

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

Save your work !



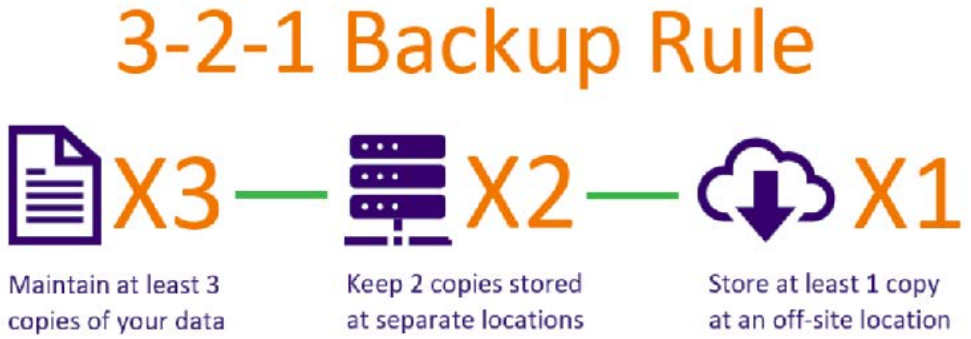
https://bids.github.io/dats/posts/2018-09-10-github-oss.html

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

Storage and backup

Storage and backup

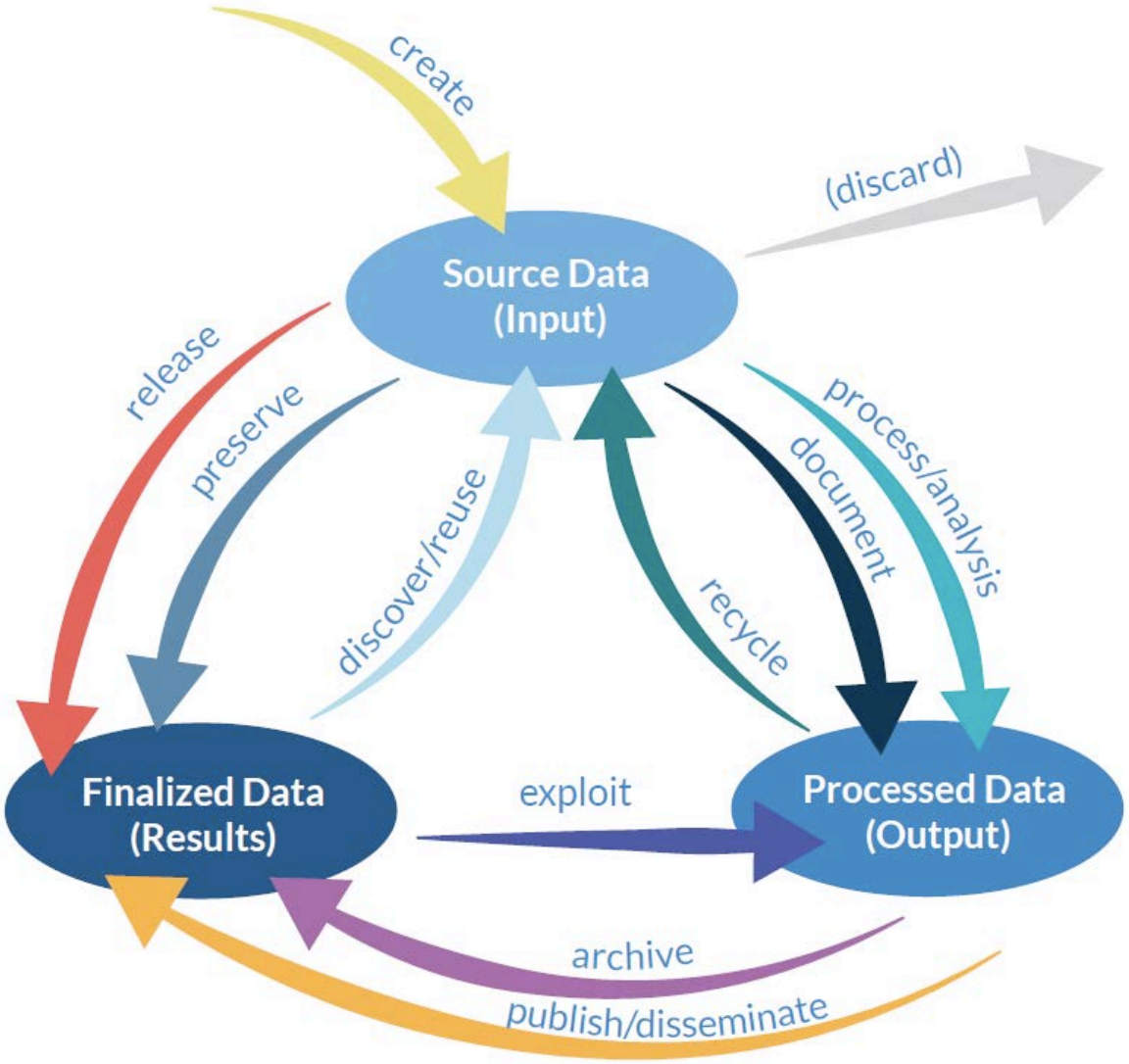
- Storage plan is integral to DM
  - How will data be stored and protected
  - Short-term vs. Long-term
  - Appropriate tiers, proper security
  - Store locally on servers or in the cloud
  - (Documented) plan to maintain system



https://www.thesslstore.com/blog/wp-content/uploads/2020/05/3-2-1-backup-rule.png

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

Process-ing/ed Data



https://vidensportal.deic.dk/sites/default/files/uploads/FAIR/Diagram.png

Digital Documentation

Digital documentation

*Of **anything** from green house to digital data object*

- Electronic lab notebook
  - Lab methods
  - Sample processing

- (Computational) processing steps
  - Data processing (excel sheets)
  - Analyses (excel, code, etc.)
  - Keep track of tools, versions and parameters used

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

Cataloging

To connect lab knowledge and resources

code	name
VEC	Vector
INS	Instrument
MTH	Method
PRM	Primer
INV	Investigation
EQP	Equipment
PLL	Plant line
ASY	Assay
SDY	Study

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

Intro to elabFTW

- HHU instance
  - https://elabftw.hhu.de/login.php
  - Register at https://elabftw.hhu.de/register.php
    - Team “CEPLAS Demo”
    - Use your @hhu.de or @uni-koeln.de email address

- More info: https://www.fdm.hhu.de/fdm-tools/elektronische-laborbuecher.html
- General demo: https://demo.elabftw.net/login.php

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

Word (.docx) alternatives (towards code)

For documents, articles, slides:

- Latex (https://www.latex-project.org)
  - See also: https://www.overleaf.com
- Markdown (https://daringfireball.net/projects/markdown/)
  - See also: https://pad.hhu.de or https://demo.hedgedoc.org/

For (interactive, commented) code:

- Rmarkdown (https://rmarkdown.rstudio.com/)
- Jupyter Notebooks (https://jupyter.org/)

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

Useful text editors

- Please do not open markdown (.md) files in MS word

- Recommended text editors with code highlighting
  - BBEdit https://www.barebones.com/products/bbedit/
  - Sublime https://www.sublimetext.com/
  - Visual Studio Code https://code.visualstudio.com/
  - Atom https://atom.io

- These highlight the structure / hierarchy of markdowns and make them a bit more comprehensible than pure text editors

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

Computational workflows

Standards and management tools

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

https://www.docker.com

nextflow

https://www.nextflow.io

nf-core

https://nf-co.re/

COMMON WORKFLOW LANGUAGE

https://www.commonwl.org

snakemake

A framework for reproducible data analysis

https://snakemake.github.io

Galaxy EUROPE

https://galaxyproject.eu/

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

Data Sharing

Sharing data with collaborators

- Data sharing agreement
  - what (meta)data or code
  - to whom
  - when
  - where (how)
  - for what purpose
- Regulate appropriate attribution (e.g. citation)

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

Registries and persistent IDs (PIDs)

Good URIs (Uniform Resource Identifiers)

- Globally unique
  - One URI should never refer to two different concepts at the same time
- Persistent
  - A URI should continue to resolve for the foreseeable future
- Stable
  - A URI should never be re-used […] even if the original is deleted
- Resolvable
  - A URI should redirect to a suitable document

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

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

Globally unique, stable, persistent identifiers (PIDs)

- Make data, digital objects, people, … uniquely identifiable
- See also
  - https://pidservices.org/
  - https://datacite.org
  - https://www.project-freya.eu/en

People

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

Digital objects

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

ePIC consortium (https://www.pidconsortium.net)

Resources

Research Resource Identifiers (https://www.rrids.org)

Institutions

Research Organization Registry (https://ror.org)

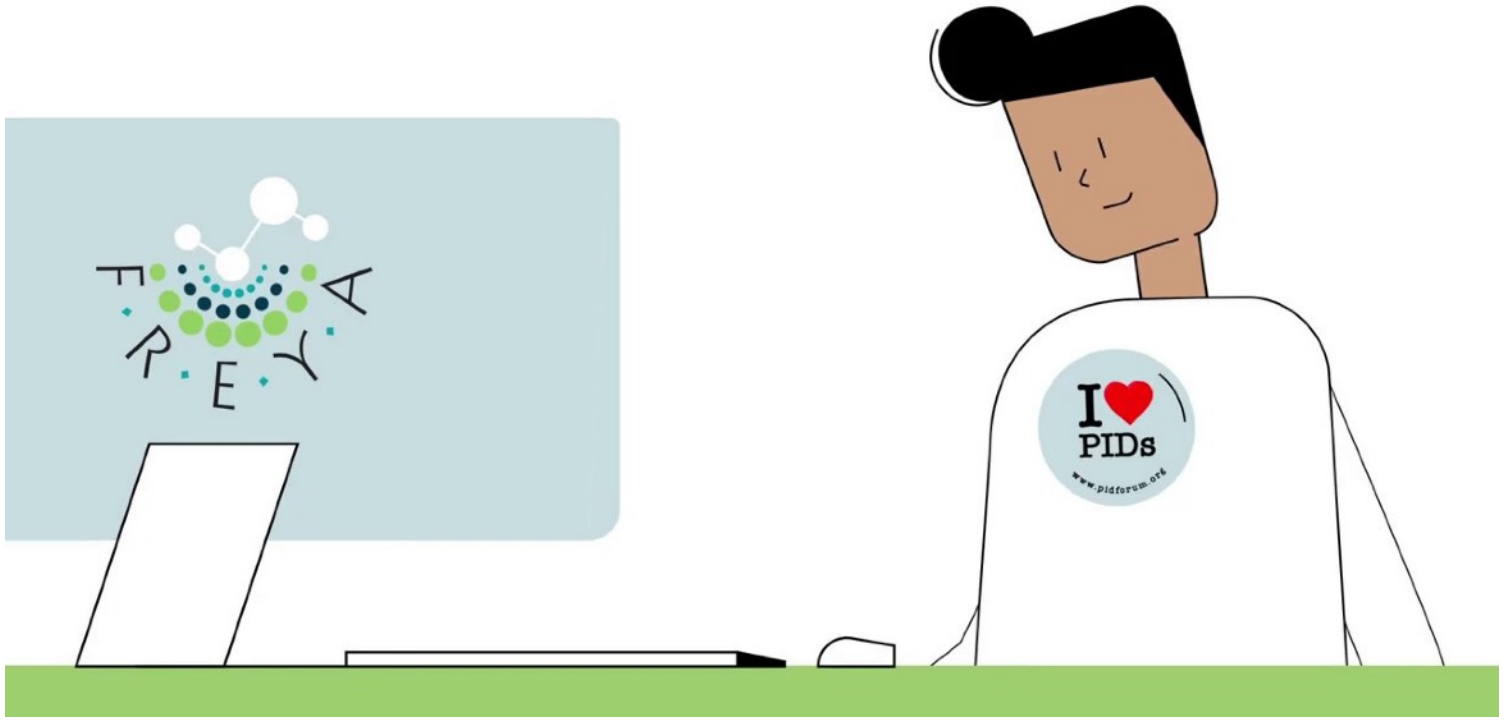
CEPLAS: https://ror.org/034waa237

Global Research Identifier Database (https://grid.ac)

CEPLAS: https://grid.ac/institutes/grid.503026.2

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

Freya Project on PIDs



https://www.project-freya.eu/en

There's also a PIDapalooza festival…
https://www.pidapalooza.org/upcoming-festival

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

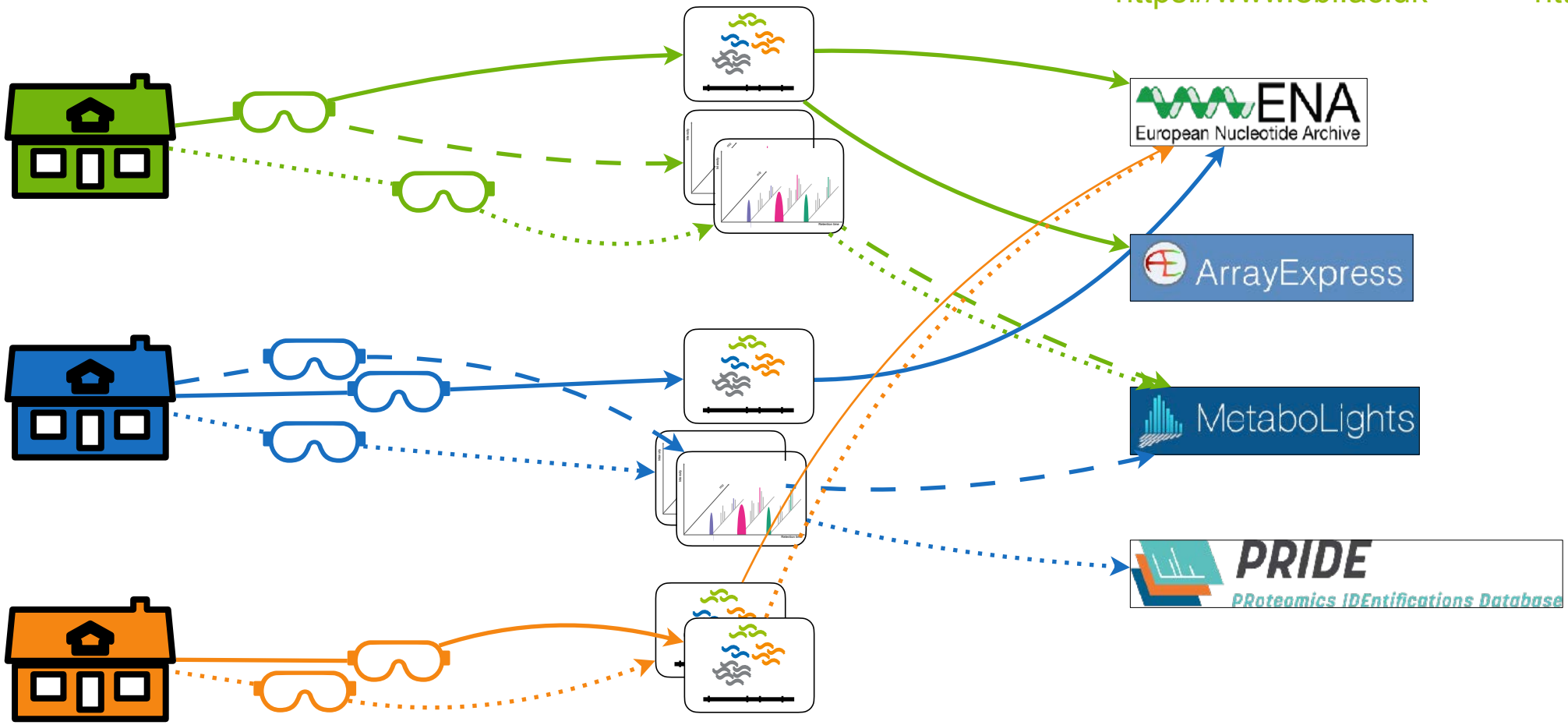
Public data repositories

Public domain-specific repositories

Examples for public end-point repositories of biological data hosted at

EMBL-EBI
https://www.ebi.ac.uk

NCBI
https://www.ncbi.nlm.nih.gov/



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

Public domain-specific repositories

	EBI	NCBI	Others
genome	EBI-ENA https://www.ebi.ac.uk/ena/	NCBI-GenBank https://www.ncbi.nlm.nih.gov/genbank/ NCBI-SRA https://www.ncbi.nlm.nih.gov/sra/	
imaging, microscopy	EBI-BioImage Archive https://www.ebi.ac.uk/bioimage-archive/ EBI-EMPIAR https://www.ebi.ac.uk/pdbe/emdb/empiar/		
meta - projects	EBI-BioStudies https://www.ebi.ac.uk/biostudies/	NCBI-BioProject https://www.ncbi.nlm.nih.gov/bioproject/	
meta - samples	EBI-BioSamples https://www.ebi.ac.uk/biosamples/	NCBI-BioSample https://www.ncbi.nlm.nih.gov/biosample	
metabolome	EBI-MetaboLights https://www.ebi.ac.uk/metabolights/		Metabolomics Workbench https://www.metabolomicsworkbench.org/
modeling	EBI-BioModels https://www.ebi.ac.uk/biomodels/		
phenome			e!DAL-PGP https://edal.ipk-gatersleben.de/index.html
proteome	EBI-PRIDE https://www.ebi.ac.uk/pride/		
transcriptome - quantification	EBI-ArrayExpress https://www.ebi.ac.uk/arrayexpress/	NCBI-GEO https://www.ncbi.nlm.nih.gov/geo/	
transcriptome - sequences	EBI-ENA https://www.ebi.ac.uk/ena/	NCBI-SRA https://www.ncbi.nlm.nih.gov/sra/	

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

Public generic repositories

- Zenodo https://zenodo.org
- DRYAD https://datadryad.org/
- FigShare https://figshare.com

- R3DATA Registry of Research Data Repositories https://www.re3data.org

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

Data Management Plan

The Data Management Plan (DMP)

- Helps you document and plan ***all*** your data management activities
- Covers the full research data lifecycle
- Frequently updated as your project develops

- Guidelines / templates:
  - http://www.dcc.ac.uk/resources/data-management-plans
  - https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
  - https://dmponline.dcc.ac.uk
  - https://ds-wizard.org

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

Conclusion / Take-home

(i.e. “Do I need to remember all of these”)

Disclaimer

Please be aware:

The following slides exemplarily show practices to get started with FAIR data management.

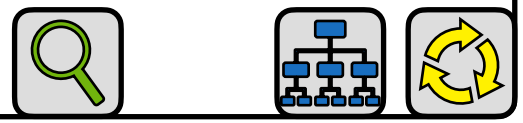
These examples are not the specific recommendations for data management within CEPLAS.

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

You can take action towards FAIR data **today!**

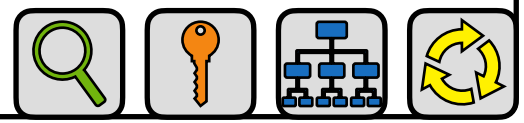
1

Rich metadata



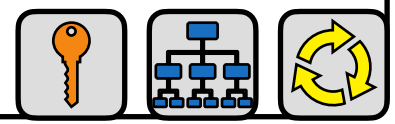
2

Conventions for
- folder structures -
- file naming -
- IDs -



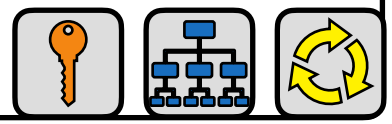
3

Digital documentation
- workflows -
- analyses -



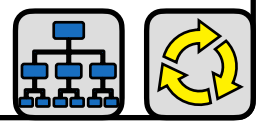
4

Versioning



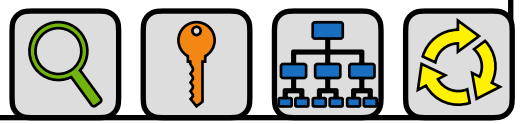
5

Storage and back-up plan



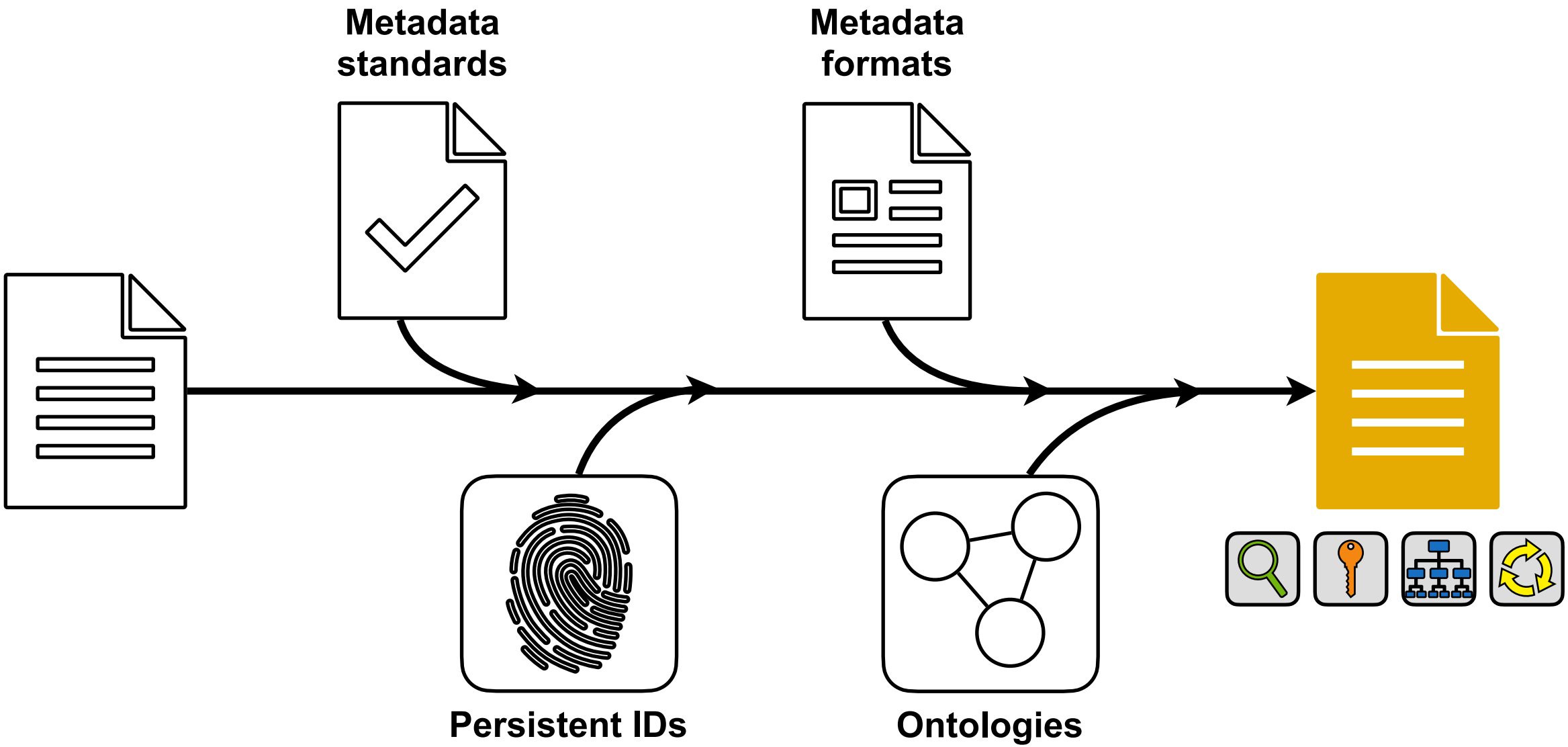
6

Data use agreements



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

From README to rich metadata



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

FAIRification protocol

- R1: Freeze-backup the "current state" (i.e. bulk of data as handed to you).
- R2: Separate and structure datasets according to "ISA model"
  - → (example) folder structure with ARC incorporated
- R3: Annotate with rich metadata
  - Add a "README" on Investigation / study levels into the folders, based on: https://github.com/nfdi4plants/ARC/blob/master/isa.investigation.xlsx
  - If you miss parameters add them in the first column using "Comment[*missing item*]"
- R4: Try to shape all data (experiment annotation, genes and functional annotation, raw data) into a single (data frame format) table
- R5: Check with PI, what data can be shared with whom
  - If your raw data is packed into little (ARC) packages (study, assay) together with rich metadata (incl. experiment conditions, contributors, protocols, etc.) it can be shared self-explanatory and referencable.
- R7: Document all of the above in a "FAIRification" document.

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