CEPLAS

Cluster of Excellence on Plant Sciences

Friday FAIRday
Session I: Directory structures & file naming

CEPLAS FAIR Data Team

Dominik Brilhaus & Hajira Jabeen

Friday FAIRday

Data Management Workshop Series

— Save the dat(e/a) —

June 11th | 2 - 2:30 pm

Directory structures and file naming

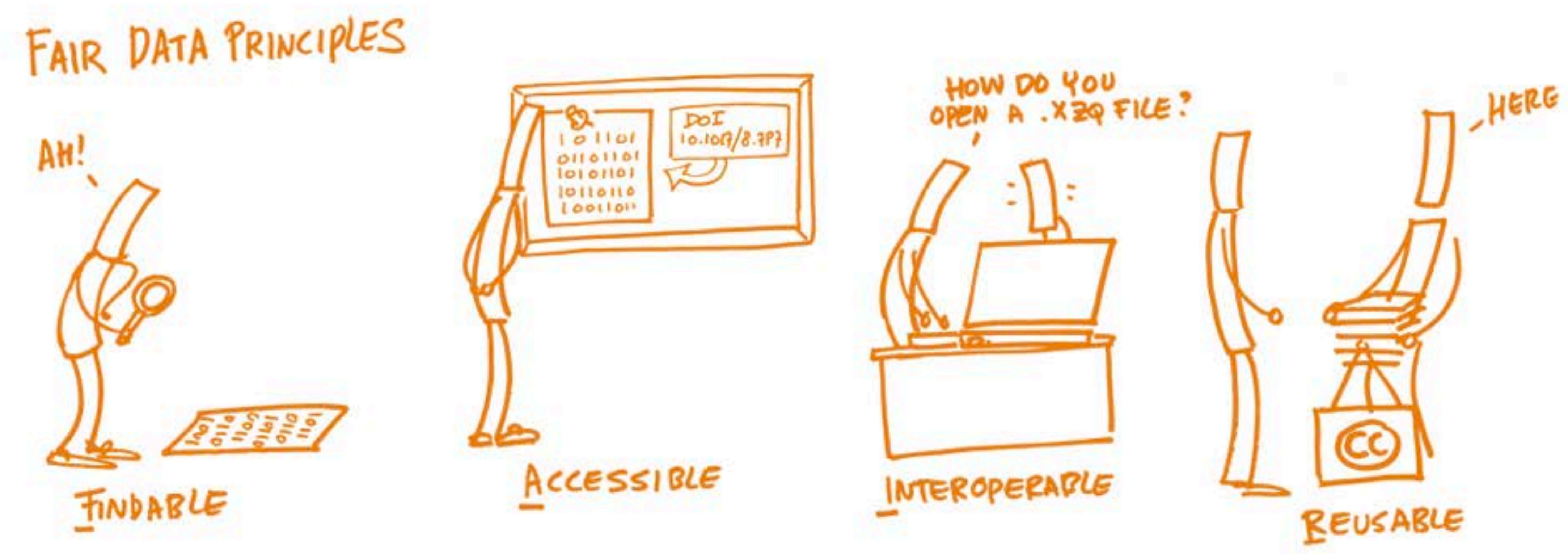
- Independent sessions
- Open to everyone
- Online via zoom (check ceplas.eu for details)

Manipulated from Skrik ("The Scream"), Edvard Munch, 1910. Photo: DPA

Friday FAIRday — Approximate schedule 2021

Session I	Directory structures & file naming	Today
Session II	Storage & Backup	July 2nd, 2021
Session III	Types of Data	Aug. 6th, 2021
Session IV	Reusability	Sept. 3rd, 2021
…		

CEPLAS-FAIRidise



https://book.fosteropenscience.eu/

The benefits of FAIR data management

Findable

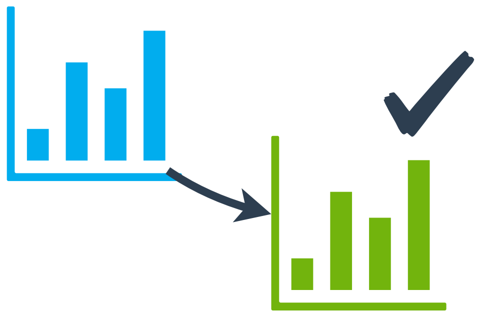
Accessible

Interoperable

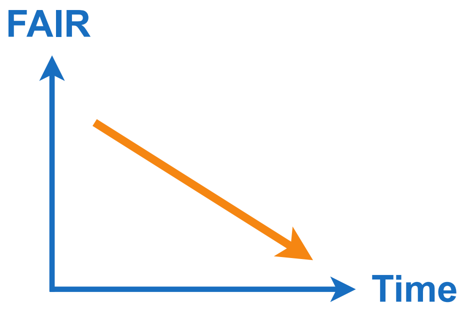
Reusable

Increased findability and visibility

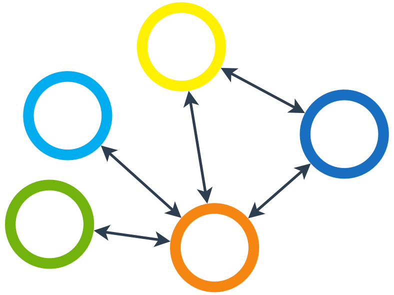
Reproducibility



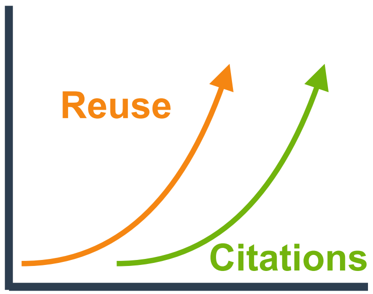
Saves time & workload



Easier collaboration & sharing



Receive due credit



Added-value to the research and community

EMBL-EBI

NCBI

Compliance with funding policies

DFG

Data Management (DM)

Your first paper or thesis

—

Can you find the data and reproduce the results?

Data Management (DM)

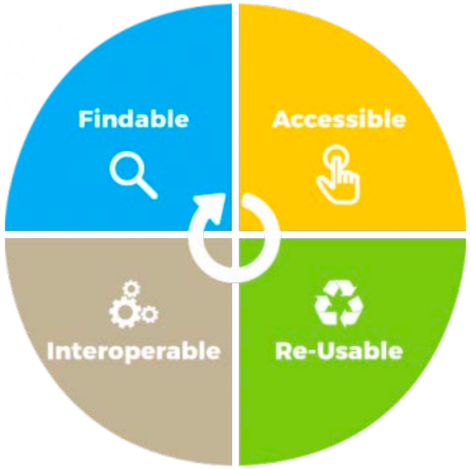
First Step: Planning the Data..

- Collection
  - software, hardware, staff, location, time
- Type and size
  - videos, text, images, omics
- Formats
  - file formats (csv, xls), data in files (columns, rows etc)
- Organisation
  - simple files, specialised database (e.g. Omero for images)
  - folder structure
- Storage
  - PC, laptop, cloud, institute, external disk
- Documentation
  - readme, ontologies, metadata

Data Management (DM)

First Step: Planning the Data..

- Collection
  - software, hardware, staff, location, time
- Type and size
  - videos, text, images, omics
- Formats
  - file formats (csv, xls), data in files (columns, rows etc)
- Organisation
  - simple files, specialised database (e.g. Omero for images)
  - folder structure
- Storage
  - PC, laptop, cloud, institute, external disk
- Documentation
  - readme, ontologies, metadata

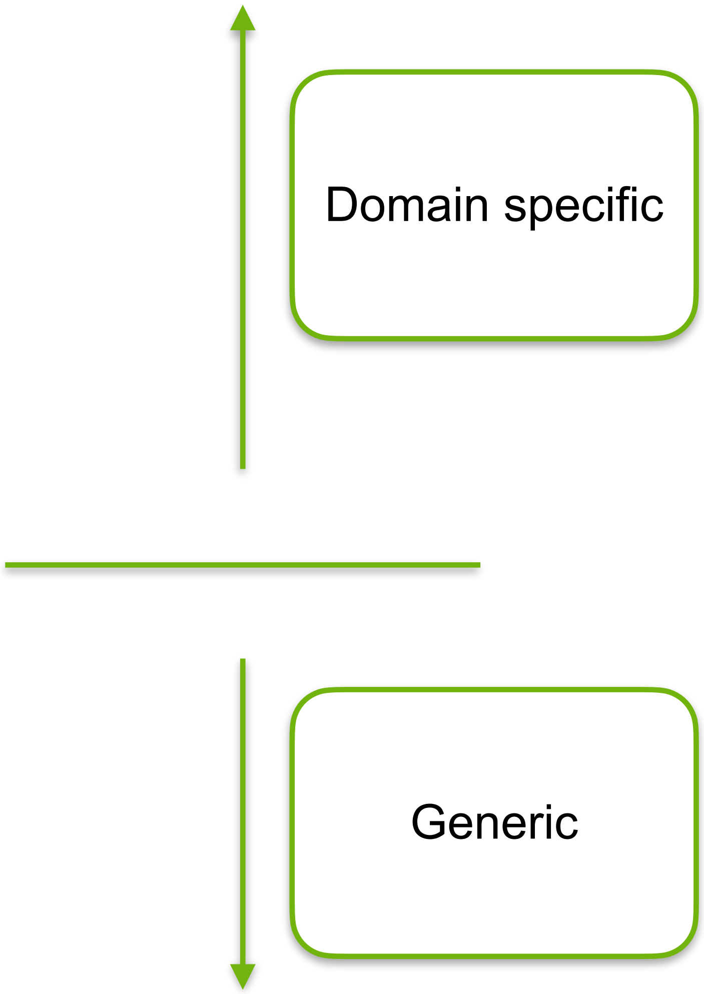


https://osimap.org/wp-content/uploads/2020/04/FAIR_EN-364x366-1.png

Data Management (DM)

First Step: Planning the Data..

- Collection
  - software, hardware, staff, location, time
- Type and size
  - videos, text, images, omics
- Formats
  - file formats (csv, xls), data in files (columns, rows etc)
- Organisation
  - simple files, specialised database (e.g. Omero for images)
  - folder structure
- Storage
  - PC, laptop, cloud, institute, external disk
- Documentation
  - readme, ontologies, metadata

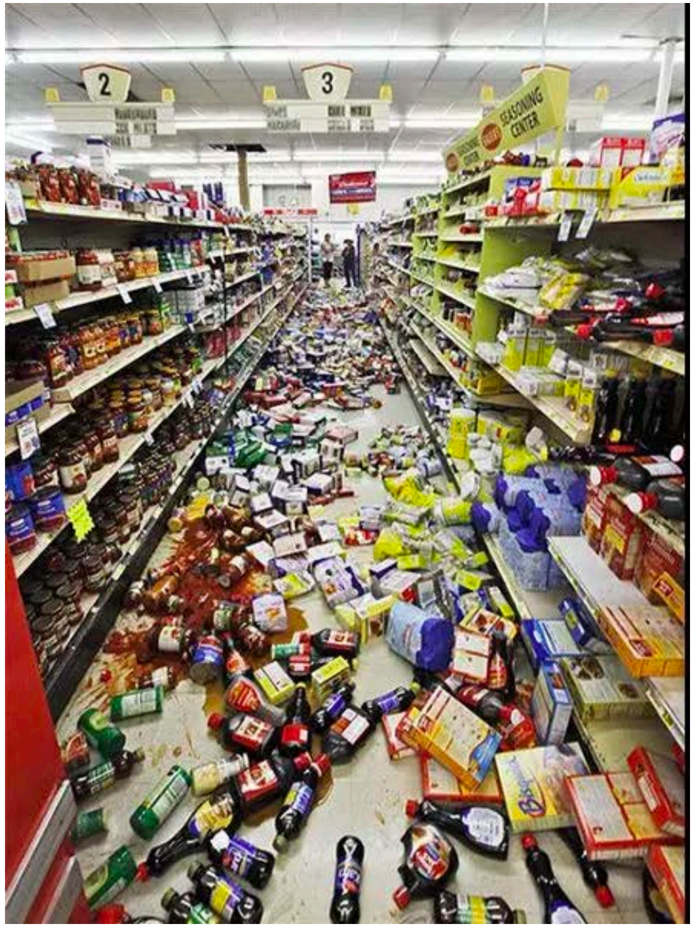


Domain specific

Generic

Directory structures

Why bother?



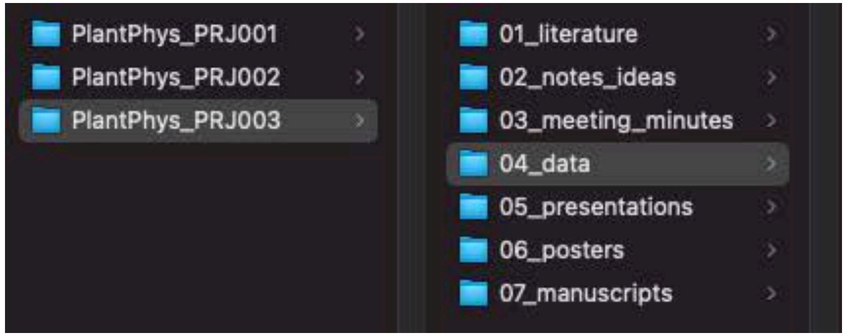
https://840027.smushcdn.com/2225136/wp-content/uploads/2017/07/grocery-store-comparison-1024x600.jpg

Why bother?

https://840027.smushcdn.com/2225136/wp-content/uploads/2017/07/grocery-store-comparison-1024x600.jpg

Directory Structure - Logical aspects

- Categorize and group files
  - Research projects
  - Time
  - Location
  - Methods



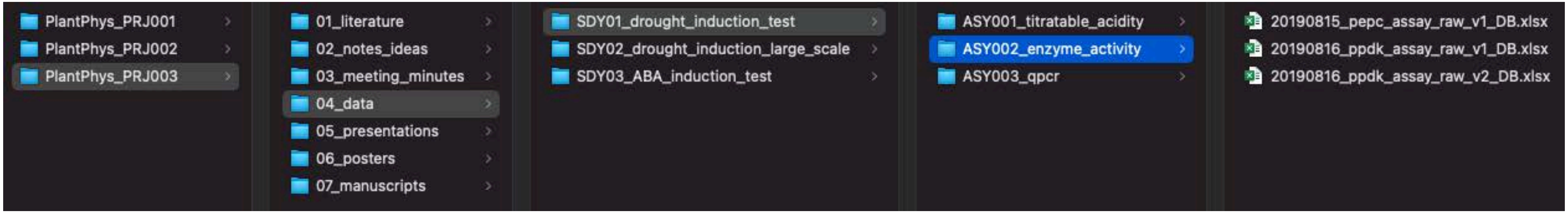
PlantPhys_PRJ001
PlantPhys_PRJ002
PlantPhys_PRJ003

01_literature
02_notes_ideas
03_meeting_minutes
04_data
05_presentations
06_posters
07_manuscripts

Directory Structure - Logical aspects

- Categorize and group files
  - Research projects
  - Time
  - Location
  - Methods

- Generic to specific
- Understandable
- Unambiguous
- Easy to follow



PlantPhys_PRJ001
PlantPhys_PRJ002
PlantPhys_PRJ003

01_literature
02_notes_ideas
03_meeting_minutes
04_data
05_presentations
06_posters
07_manuscripts

SDY01_drought_induction_test
SDY02_drought_induction_large_scale
SDY03_ABA_induction_test

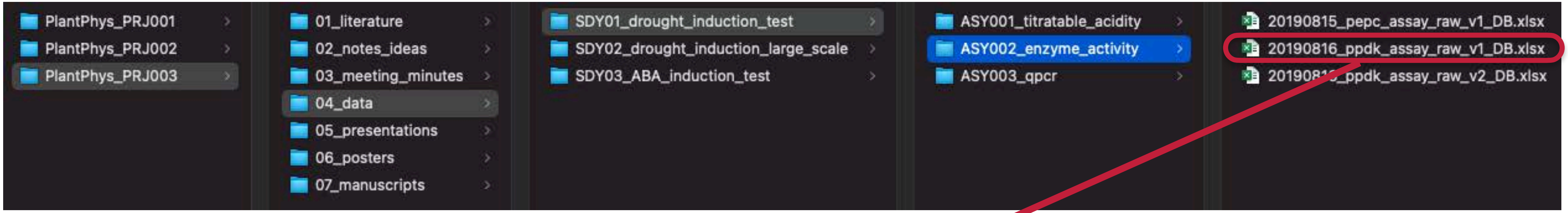
ASY001_titratable_acidity
ASY002_enzyme_activity
ASY003_qpcr

20190815_pepc_assay_raw_v1_DB.xlsx
20190816_ppdk_assay_raw_v1_DB.xlsx
20190816_ppdk_assay_raw_v2_DB.xlsx

Generic
Specific

Directory Structure - Technical aspects

- Hierarchy (path) = high level descriptors for files
- Path + file name = unique file ID
- New folder vs. sub-folder?
- Different folder structures are usually suited to different data / needs / projects
- Avoid unnecessarily deep structures



PlantPhys_PRJ001
PlantPhys_PRJ002
PlantPhys_PRJ003

01_literature
02_notes_ideas
03_meeting_minutes
04_data
05_presentations
06_posters
07_manuscripts

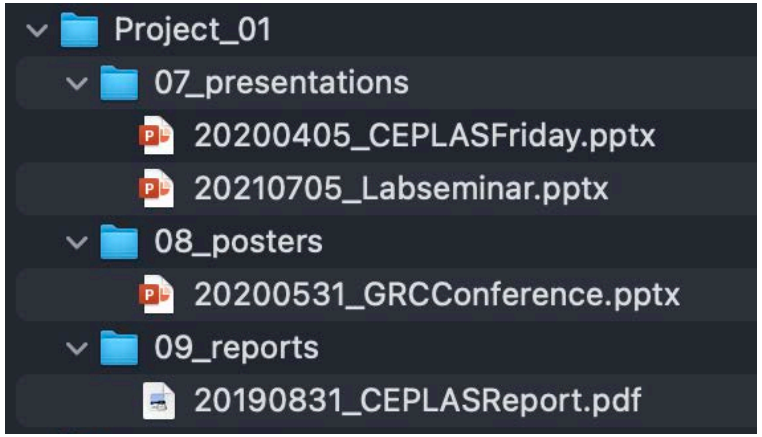
SDY01_drought_induction_test
SDY02_drought_induction_large_scale
SDY03_ABA_induction_test

ASY001_titratable_acidity
ASY002_enzyme_activity
ASY003_qpcr

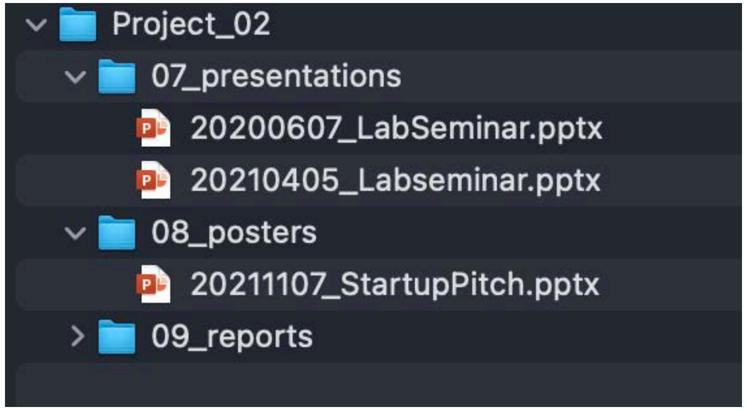
20190815_pepc_assay_raw_v1_DB.xlsx
20190816_ppdk_assay_raw_v1_DB.xlsx
20190816_ppdk_assay_raw_v2_DB.xlsx

Path+Filename: `~/PlantPhys_PRJ003/04_data/SDY01_drought_induction_test/ASY002_enzyme_activity/20190816_ppdk_assay_raw_v1_DB.xlsx`

Avoiding deep structures



- Project_01
  - 07_presentations
    - 20200405_CEPLASFriday.pptx
    - 20210705_Labseminar.pptx
  - 08_posters
    - 20200531_GRCConference.pptx
  - 09_reports
    - 20190831_CEPLASReport.pdf

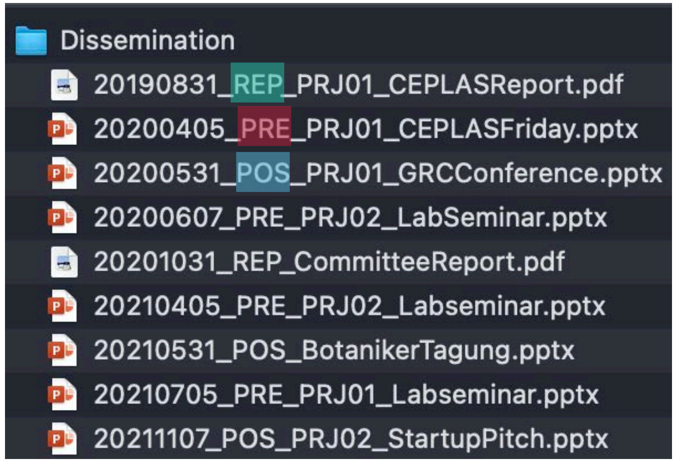


- Project_02
  - 07_presentations
    - 20200607_LabSeminar.pptx
    - 20210405_Labseminar.pptx
  - 08_posters
    - 20211107_StartupPitch.pptx
  - 09_reports

One folder for all files

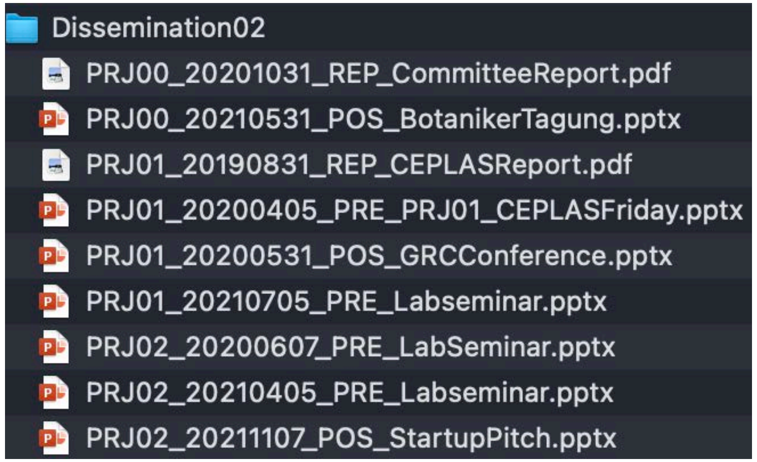
Categorized by code

Chronological



- Dissemination
  - 20190831_REP_PRJ01_CEPLASReport.pdf
  - 20200405_PRE_PRJ01_CEPLASFriday.pptx
  - 20200531_POS_PRJ01_GRCConference.pptx
  - 20200607_PRE_PRJ02_LabSeminar.pptx
  - 20201031_REP_CommitteeReport.pdf
  - 20210405_PRE_PRJ02_Labseminar.pptx
  - 20210531_POS_BotanikerTagung.pptx
  - 20210705_PRE_PRJ01_Labseminar.pptx
  - 20211107_POS_PRJ02_StartupPitch.pptx

By Project



- Dissemination02
  - PRJ00_20201031_REP_CommitteeReport.pdf
  - PRJ00_20210531_POS_BotanikerTagung.pptx
  - PRJ01_20190831_REP_CEPLASReport.pdf
  - PRJ01_20200405_PRE_PRJ01_CEPLASFriday.pptx
  - PRJ01_20200531_POS_GRCConference.pptx
  - PRJ01_20210705_PRE_Labseminar.pptx
  - PRJ02_20200607_PRE_LabSeminar.pptx
  - PRJ02_20210405_PRE_Labseminar.pptx
  - PRJ02_20211107_POS_StartupPitch.pptx

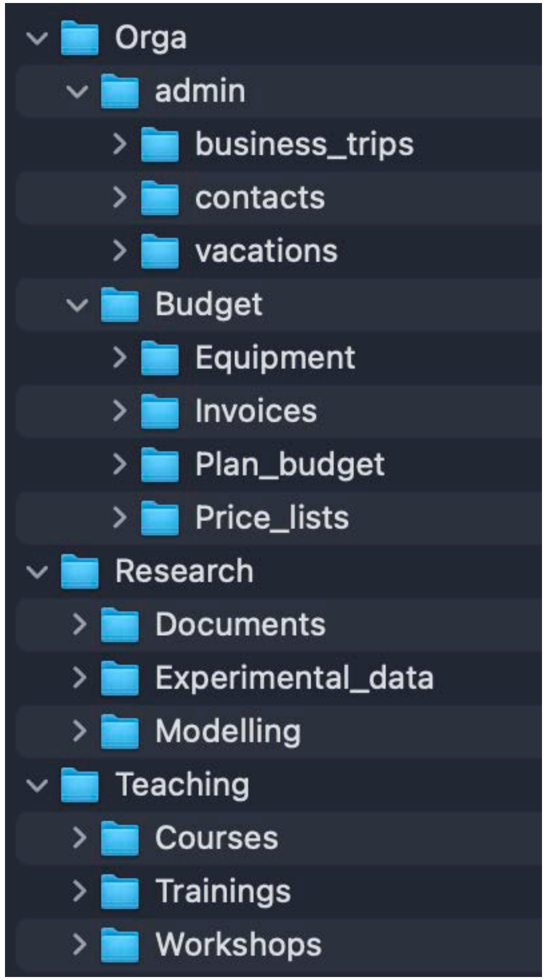
Report

Presentation

Poster

More examples..

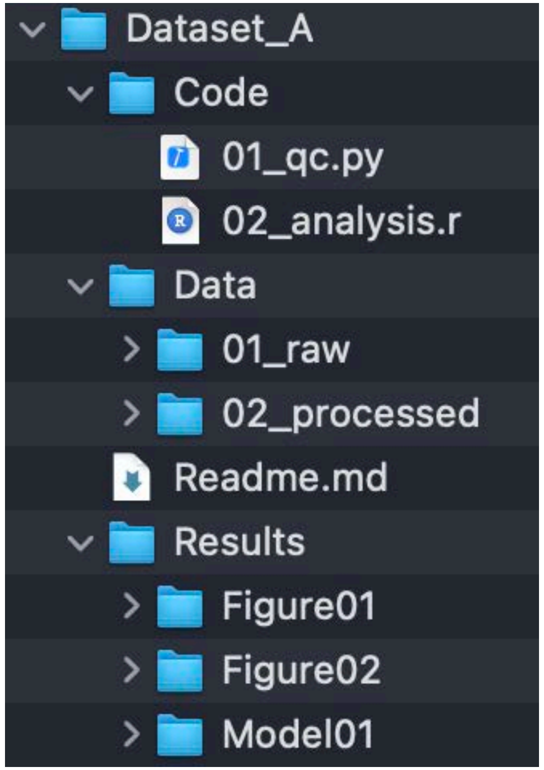
PhD project directories



```
v Orga
    v admin
        > business_trips
        > contacts
        > vacations
    v Budget
        > Equipment
        > Invoices
        > Plan_budget
        > Price_lists
v Research
    > Documents
    > Experimental_data
    > Modelling
v Teaching
    > Courses
    > Trainings
    > Workshops
```

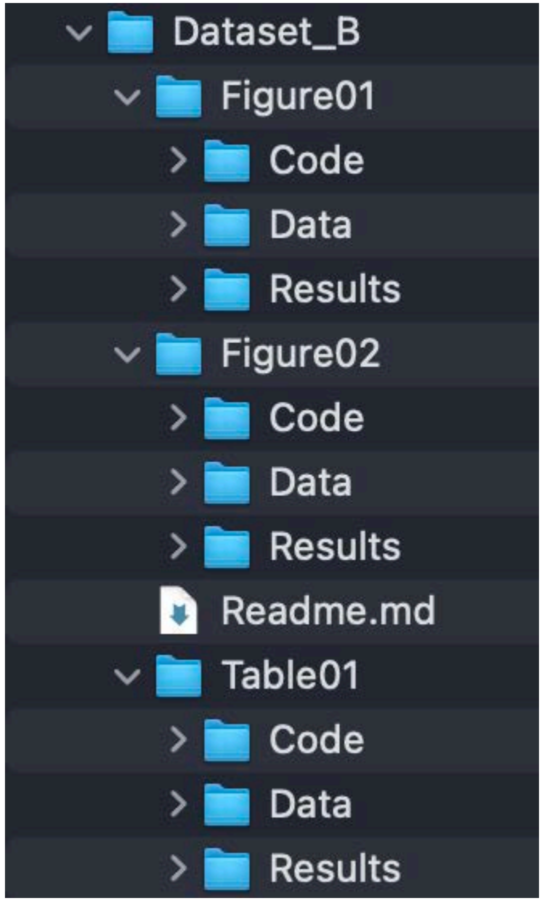
Research data directories

By research step



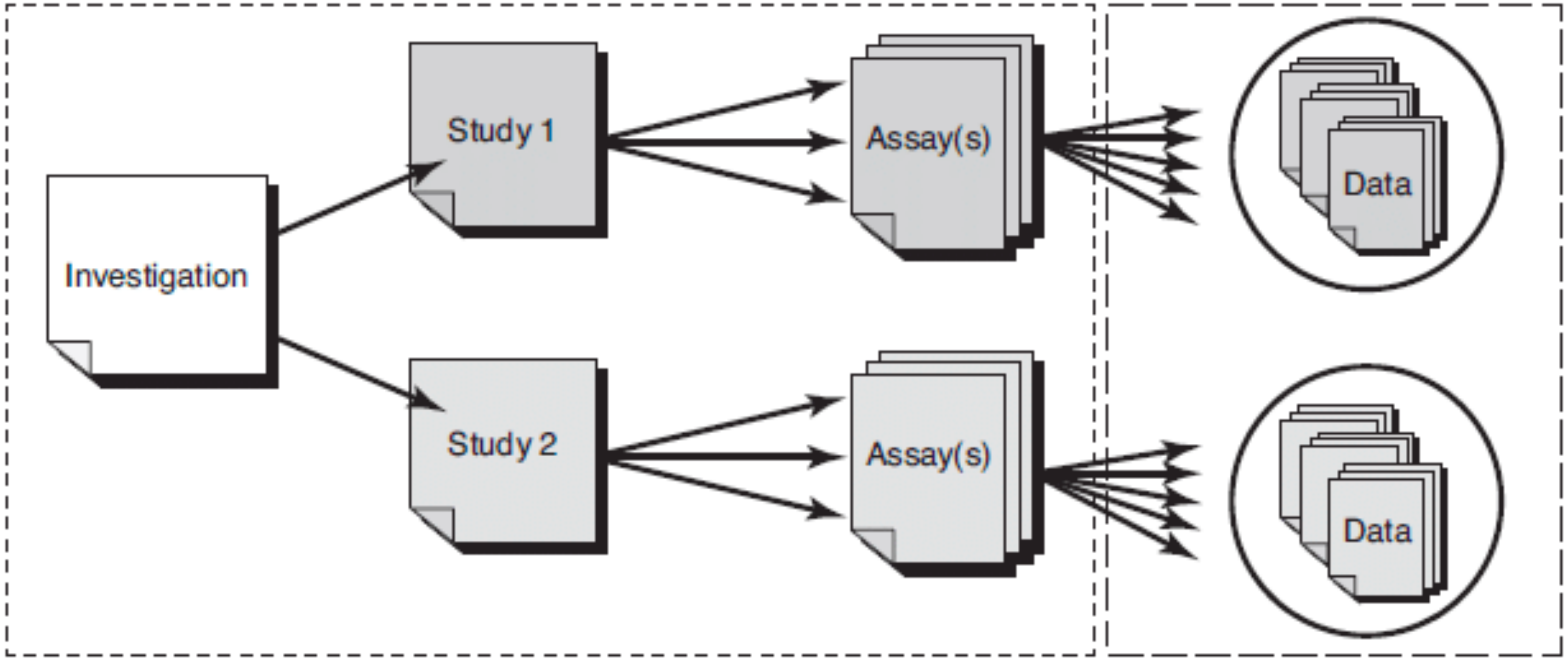
```
v Dataset_A
    v Code
        01_qc.py
        02_analysis.r
    v Data
        > 01_raw
        > 02_processed
    Readme.md
    v Results
        > Figure01
        > Figure02
        > Model01
```

By output



```
v Dataset_B
    v Figure01
        > Code
        > Data
        > Results
    v Figure02
        > Code
        > Data
        > Results
    Readme.md
    v Table01
        > Code
        > Data
        > Results
```

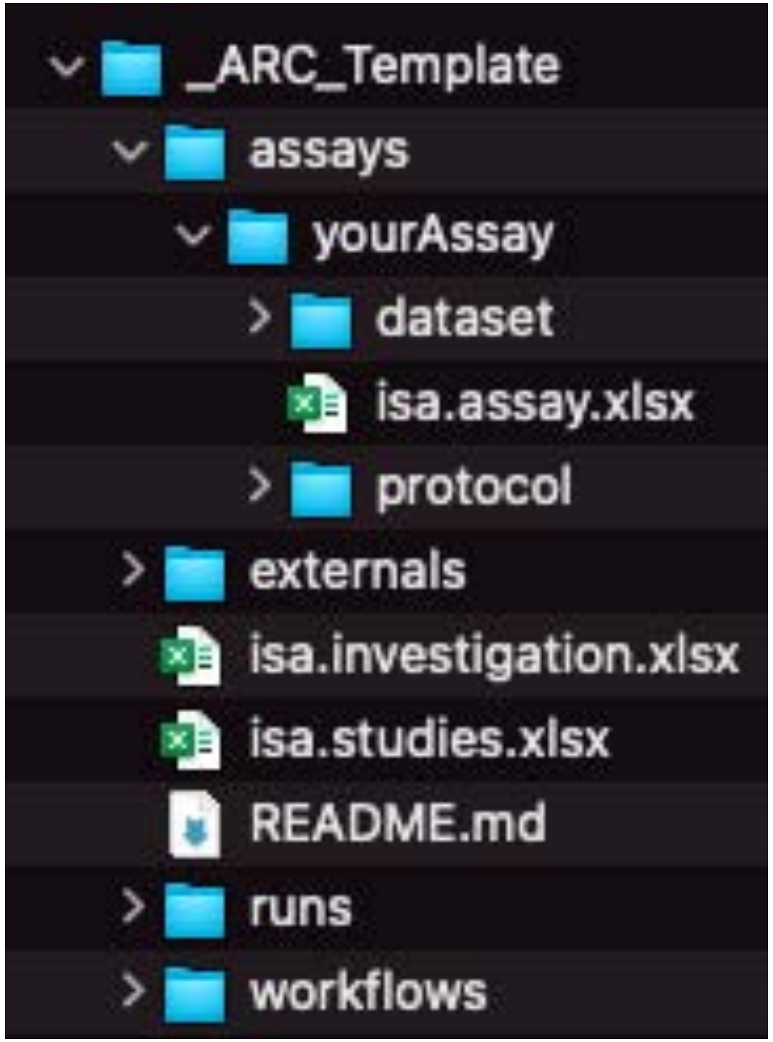
The ISA Model



https://www.researchgate.net/publication/333163209_Design_and_Development_of_a_Phenotypic_Data_Model_PDM

Annotated Research Context (ARC)

- **Directory Structure**
- Minimal amount of **naming convention**
- Raw data, processed data, metadata
- Version Control
- Sharing
- Backup
- Many more ..

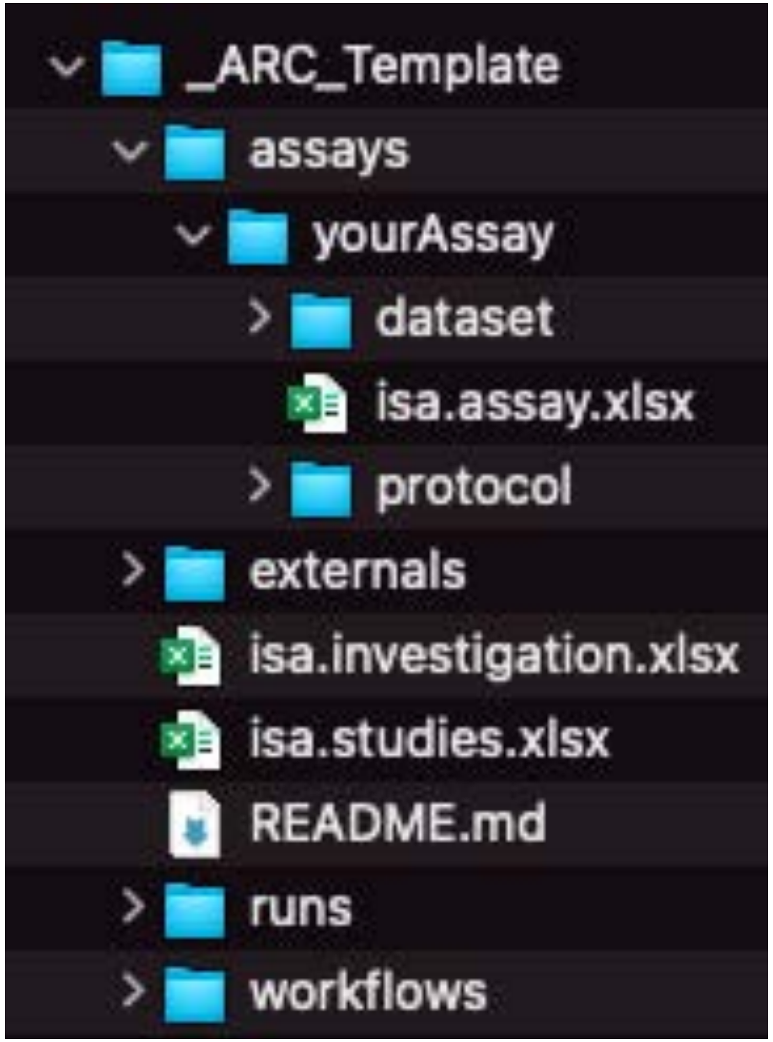


```
_ARC_Template
  assays
    yourAssay
      dataset
      isa.assay.xlsx
      protocol
  externals
  isa.investigation.xlsx
  isa.studies.xlsx
  README.md
  runs
  workflows
```

Annotated Research Context (ARC)

- **Directory Structure**
- Minimal amount of **naming convention**
- Raw data, processed data, metadata
- Version Control
- Sharing
- Backup
- Many more ..

More on these in upcoming sessions …

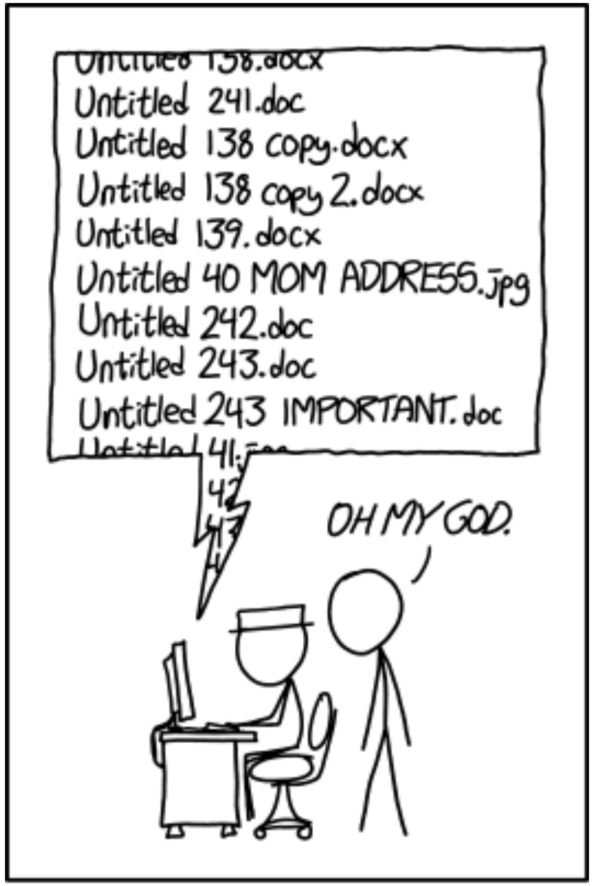


```
⌄ _ARC_Template
  ⌄ assays
    ⌄ yourAssay
      › dataset
        isa.assay.xlsx
      › protocol
  › externals
    isa.investigation.xlsx
    isa.studies.xlsx
    README.md
  › runs
  › workflows
```

File Naming

File Naming

- Primary identifier of a file
- Good and meaningful names
  - Hint towards contents of file
  - Help in discovery
    - Classification
    - Sorting
    - Versioning
- Consider (in future)
  - Searching
  - Sorting
  - Uniqueness



PROTIP: NEVER LOOK IN SOMEONE ELSE'S DOCUMENTS FOLDER.

https://xkcd.com/1459/

File Naming Checklist - Technical

- Avoid full-stops
- Avoid spaces
- Avoid special characters
- Use short, precise, relevant names
  - Less than 25 characters
  - Distinguishable (name+directory path)
  - Unique (search for filename should better not result in multiple results)

~ ! @ # $ % ^ &

* ( ) ` ; : < > ? . ,

[ ] { } ' " | ä ö ü ß

File Naming Checklist - Technical

- Example Cases
  - Kebab-case: The-quick-brown-fox-jumps-over-the-lazy-dog.txt
  - CamelCase: TheQuickBrownFoxJumpsOverTheLazyDog.txt
  - Snake_case: The_quick_brown_fox_jumps_over_the_lazy_dog.txt

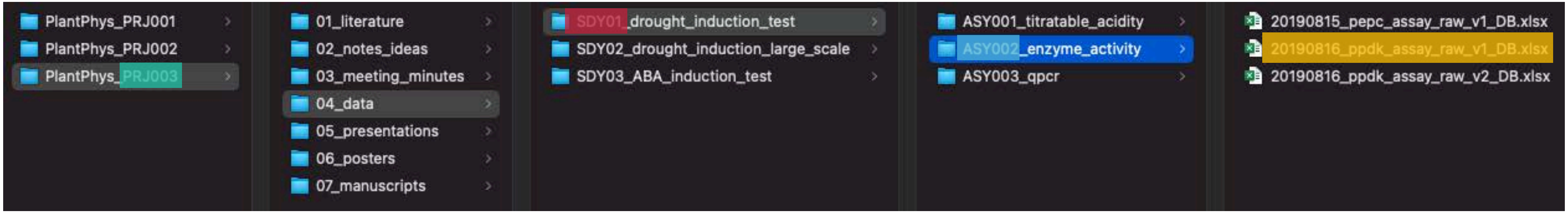
File Naming Checklist - Content

- Use descriptive names
- Abbreviations
    - project name - project number - deptname - team - location - version - date - sampletype - etc.)
- Reverse dates for recurring events (Timestamp : YYYYMMDD)
- For names (lastnamef)
- Numbering (001, 002, … 010 - NOT 1, 2, … 10)

Directory structure or file name?

Find and understand data by

… location (path) + filename



PlantPhys_PRJ001
PlantPhys_PRJ002
PlantPhys_PRJ003

01_literature
02_notes_ideas
03_meeting_minutes
04_data
05_presentations
06_posters
07_manuscripts

SDY01_drought_induction_test
SDY02_drought_induction_large_scale
SDY03_ABA_induction_test

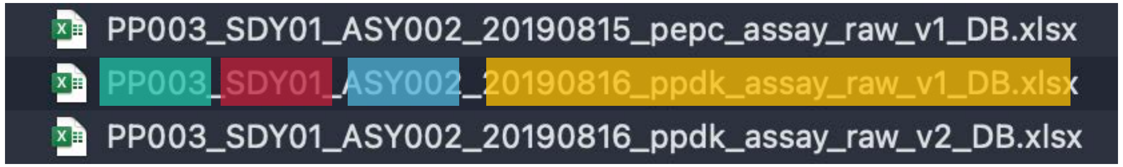
ASY001_titratable_acidity
ASY002_enzyme_activity
ASY003_qpcr

20190815_pepc_assay_raw_v1_DB.xlsx
20190816_ppdk_assay_raw_v1_DB.xlsx
20190816_ppdk_assay_raw_v2_DB.xlsx

Path+Filename:

```
~/PlantPhys_PRJ003/04_data/SDY01_drought_induction_test/ASY002_enzyme_activity/20190816_ppdk_assay_raw_v1_DB.xlsx
```

…filename only



PP003_SDY01_ASY002_20190815_pepc_assay_raw_v1_DB.xlsx
PP003_SDY01_ASY002_20190816_ppdk_assay_raw_v1_DB.xlsx
PP003_SDY01_ASY002_20190816_ppdk_assay_raw_v2_DB.xlsx

Generic **Specific**

Take action

Key takeaways

- Do not store your files on desktop !
- Do not be adhoc in creating folders / filenames
- Invest time in planning
- Limit folder creation
- Think strategically, adapt procedures to your requirements

Develop a system

(directory structure + naming scheme)

- Think easy adoption
- Document your system in a **README**
  - Create a shared folder hierarchy with the README for onboarding
- Be **consistent**
- Accept that there is no **perfect** method
- Be succinct
  - Often only 255 characters allowed for “filename AND path”

Resources

- Related Material
- File naming Checklist

- A readme schema
- An example Worksheet
  - Working on worksheet (Optional) ..

Disclaimer:

The practices we describe are neither binding, nor obligatory. Instead, we have tried to articulate useful principles for achieving a consistent and maintainable structure.