



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

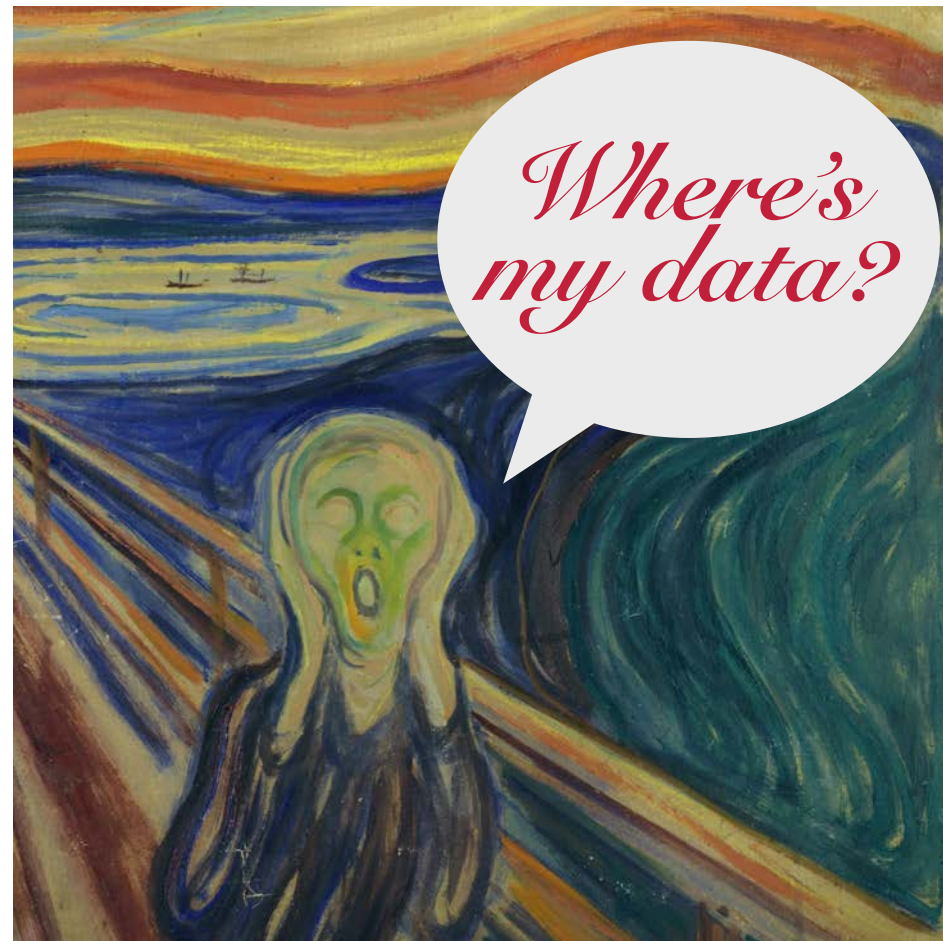
Data Management Workshop Series

— 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 2 nd , 2021
Session III	Types of Data	Aug. 6 th , 2021
Session IV	Reusability	Sept. 3 rd , 2021

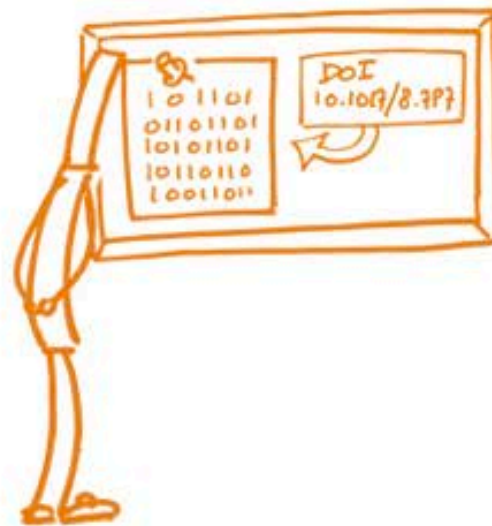
...

FAIR DATA PRINCIPLES



AH!

FINDABLE



ACCESSIBLE



HOW DO YOU
OPEN A .XZQ FILE?

INTEROPERABLE



HERE

REUSABLE

<https://book.fosteropenscience.eu/>



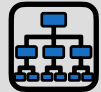
The benefits of FAIR data management



Findable



Accessible



Interoperable



Reusable

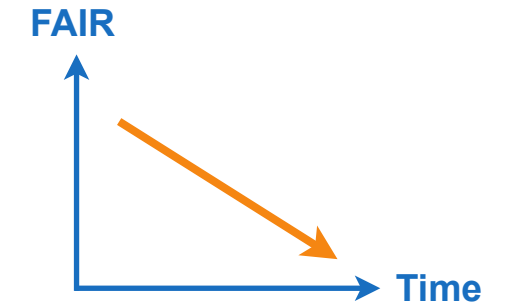
Increased findability and visibility



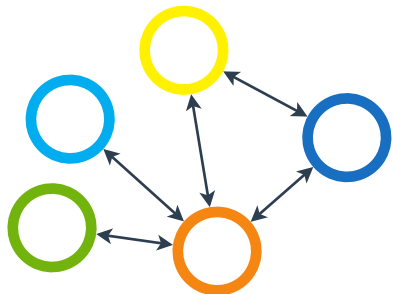
Reproducibility



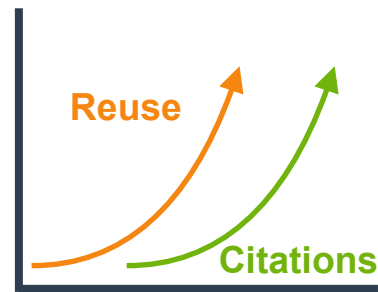
Saves time & workload



Easier collaboration & sharing



Receive due credit



Added-value to the research and community



Compliance with funding policies



Data Management (DM)

Your first paper or thesis

—

Can you find the data and
reproduce the results?

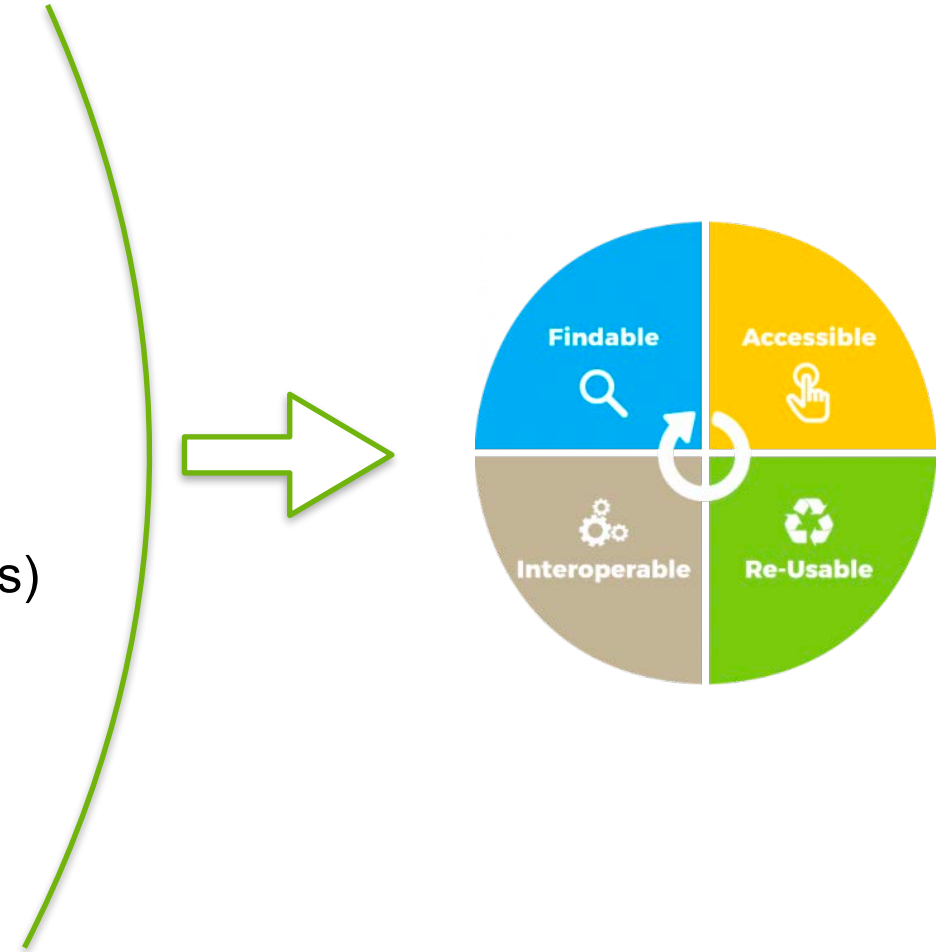
First Step: Planning the Data..

- 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..

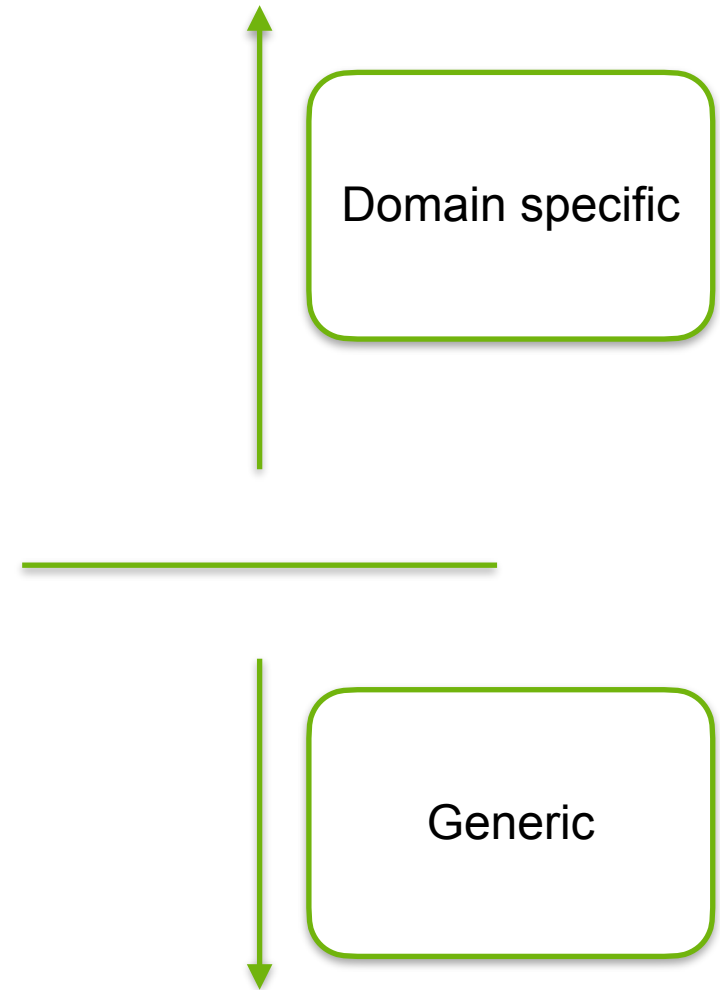
- Collection
 - software, hardware, staff, location, time
- Type and size
 - videos, text, images, omics
- Formats
 - file formats (csv, xls), data in files (columns, rows etc)
- Organisation
 - simple files, specialised database (e.g. Omero for images)
 - folder structure
- Storage
 - PC, laptop, cloud,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



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

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

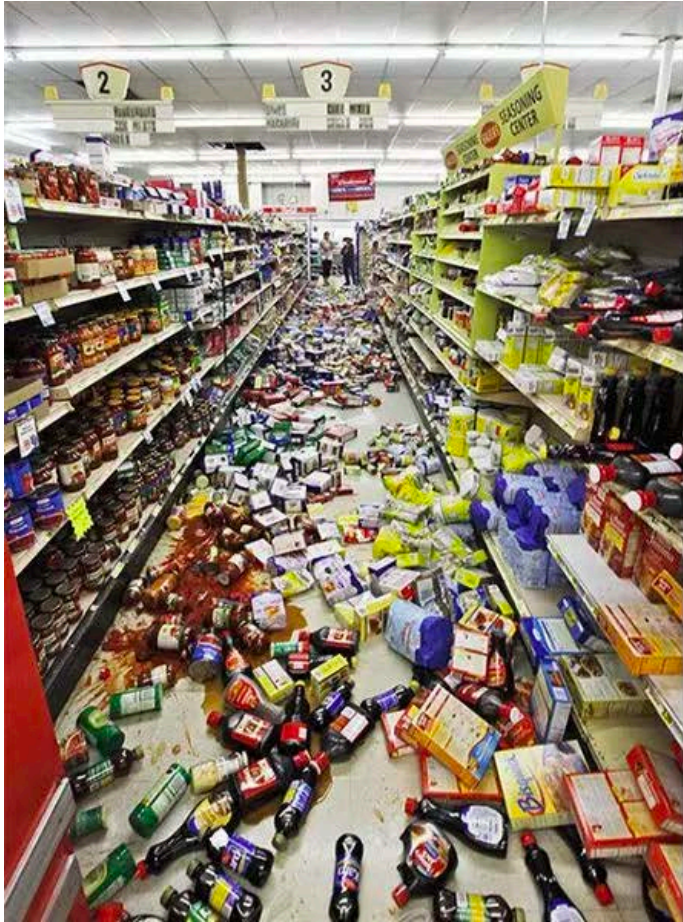


Directory structures





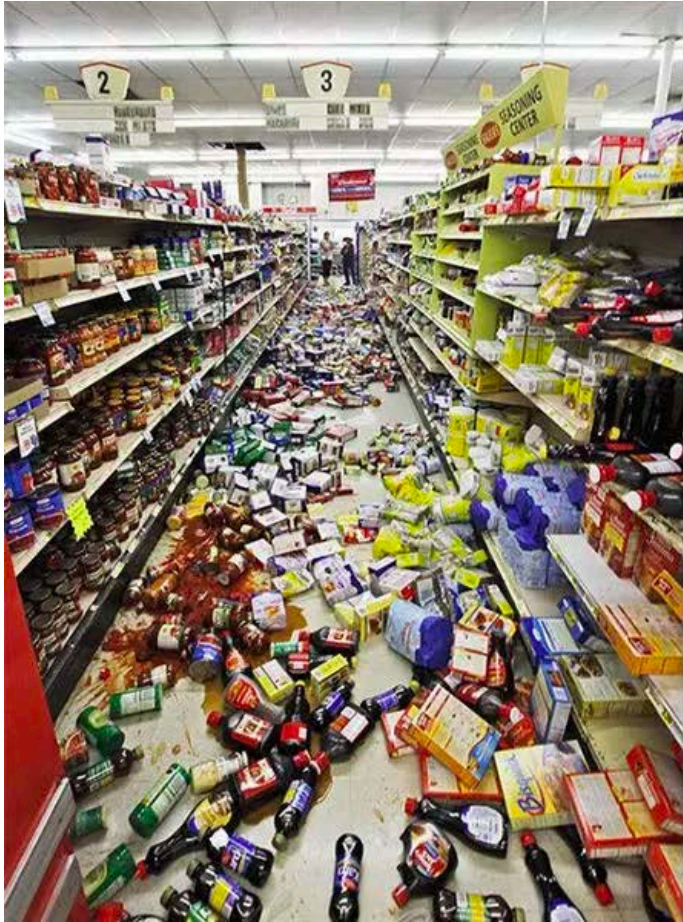
Why bother?



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



Why bother?

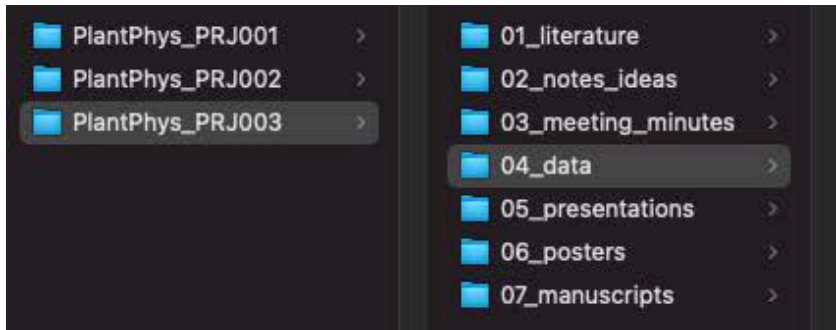


<https://840027.smushcdn.com/2225136/wp-content/uploads/2017/07/grocery-store-comparison-1024x600.jpg>



Directory Structure - Logical aspects

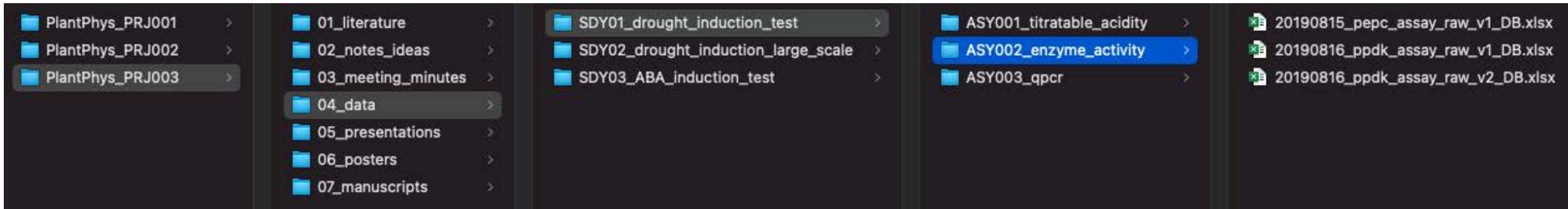
- Categorize and group files
 - Research projects
 - Time
 - Location
 - Methods





Directory Structure - Logical aspects

- Categorize and group files
 - Research projects
 - Time
 - Location
 - Methods
- Generic to specific
 - Understandable
 - Unambiguous
 - Easy to follow



Generic

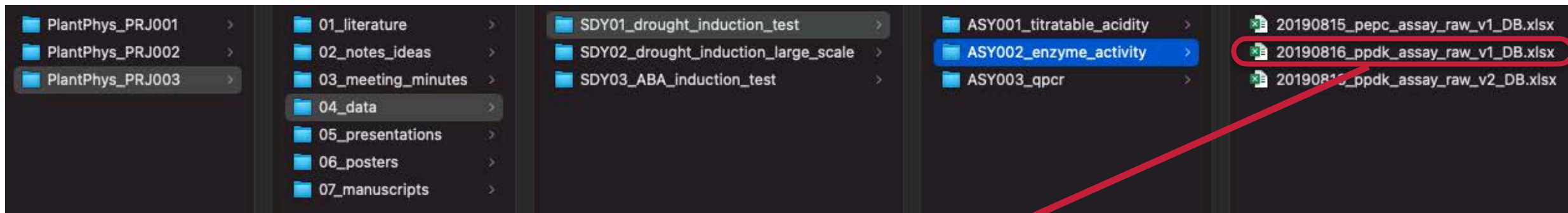
Specific





Directory Structure - Technical aspects

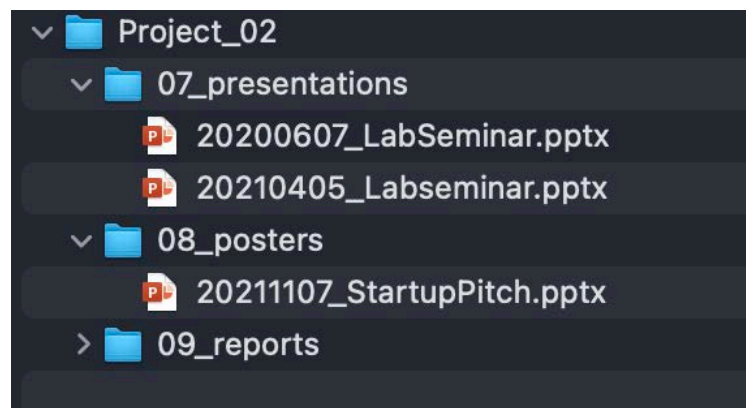
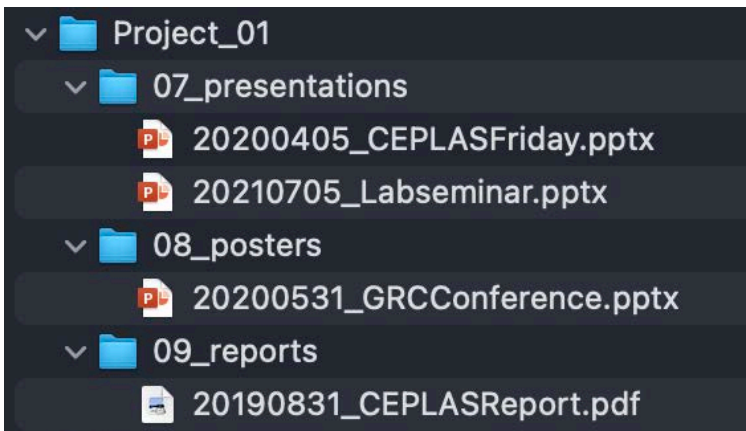
- Hierarchy (path) = high level descriptors for files
- Path + file name = unique file ID
- New folder vs. sub-folder?
- Different folder structures are usually suited to different data / needs / projects
- Avoid unnecessarily deep structures



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



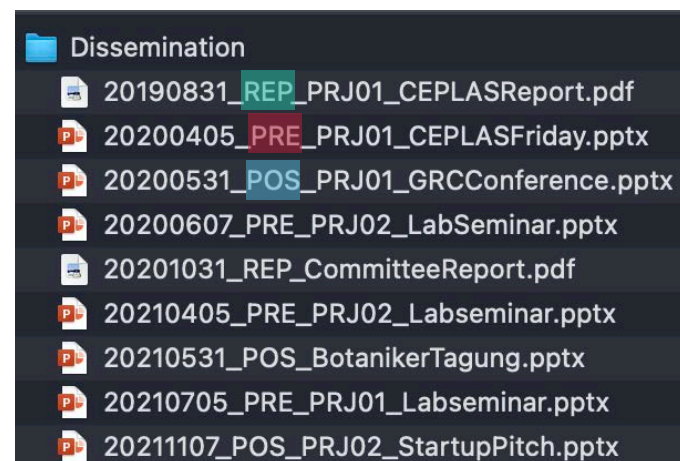
Avoiding deep structures



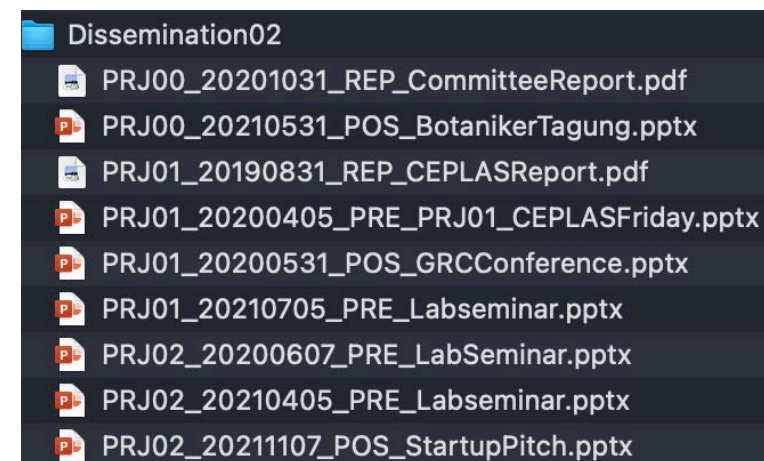
One folder for all files

Categorized by code

Chronological



By Project



Report

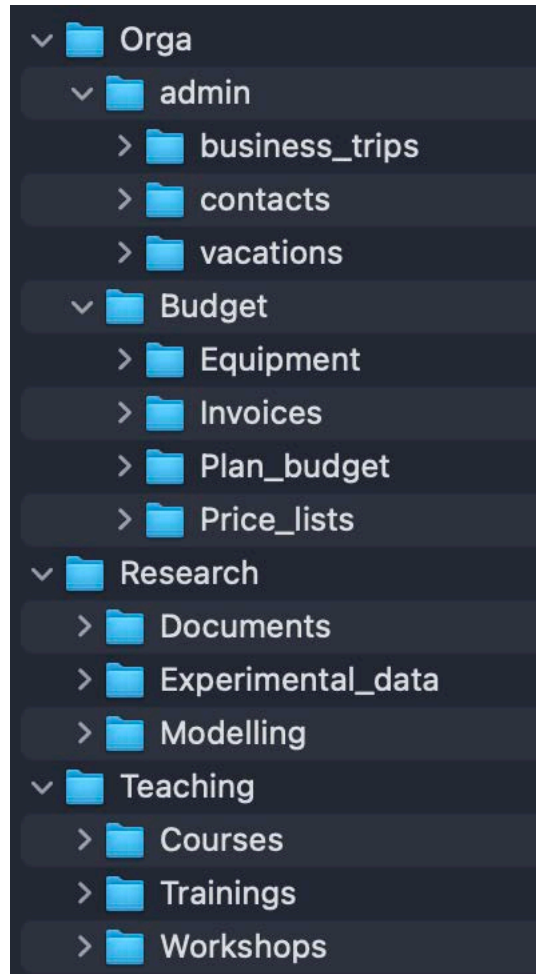
Presentation

Poster



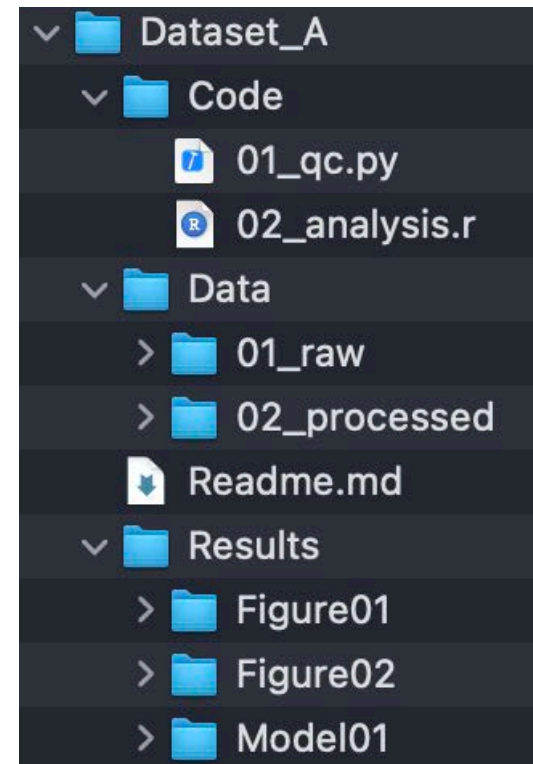
More examples..

PhD project directories

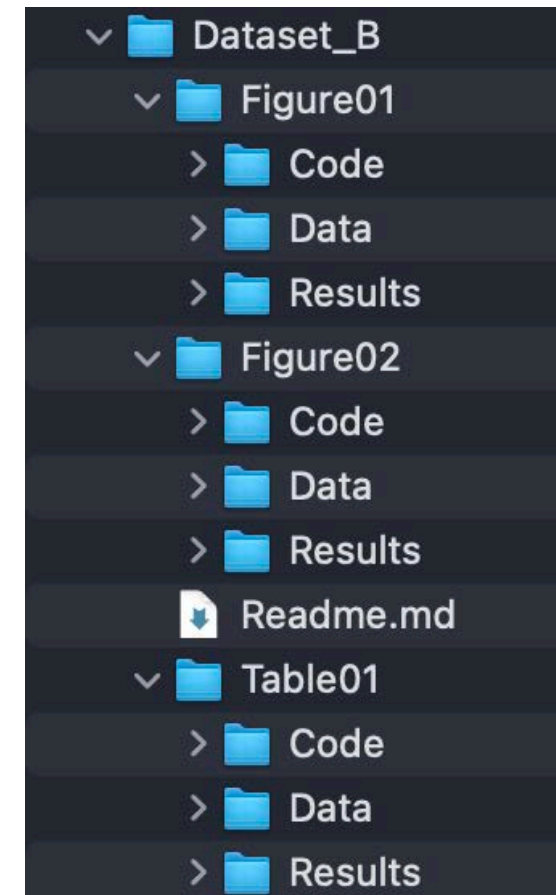


Research data directories

By research step

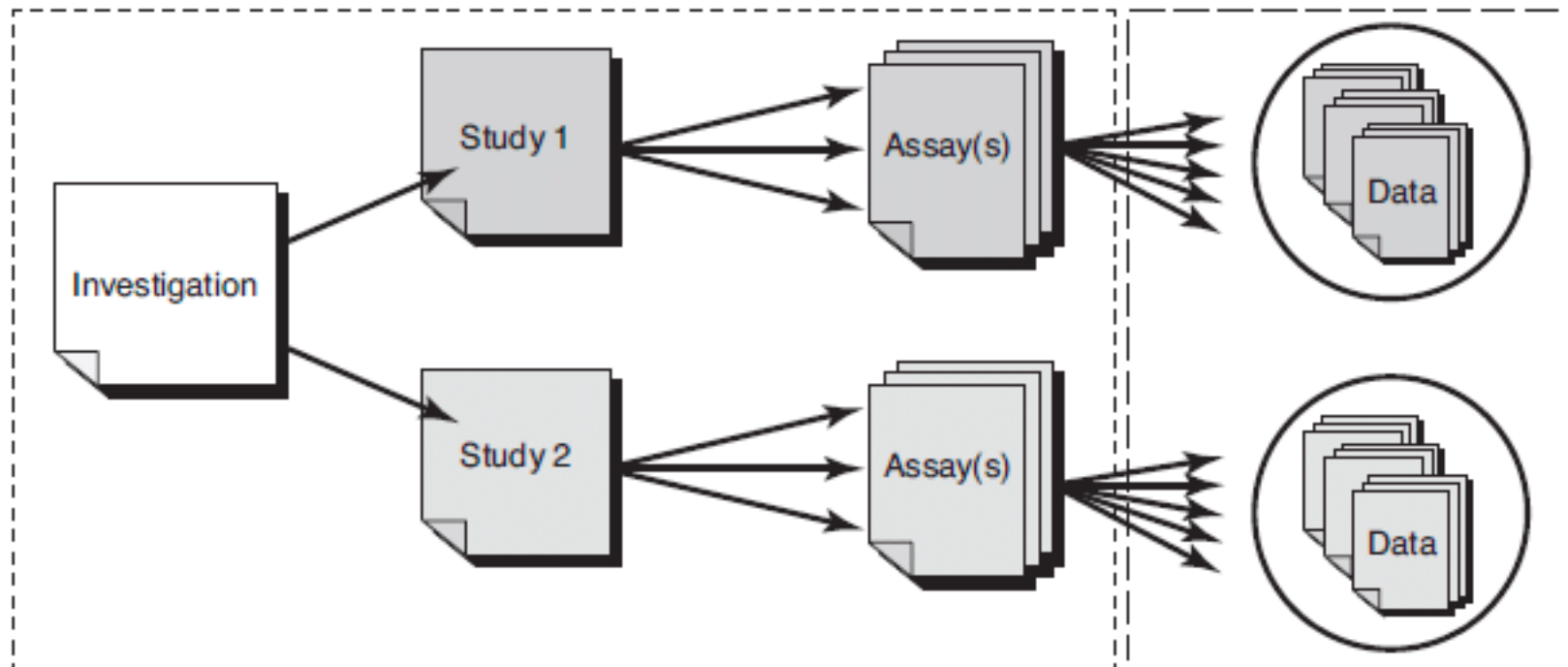


By output





The ISA Model



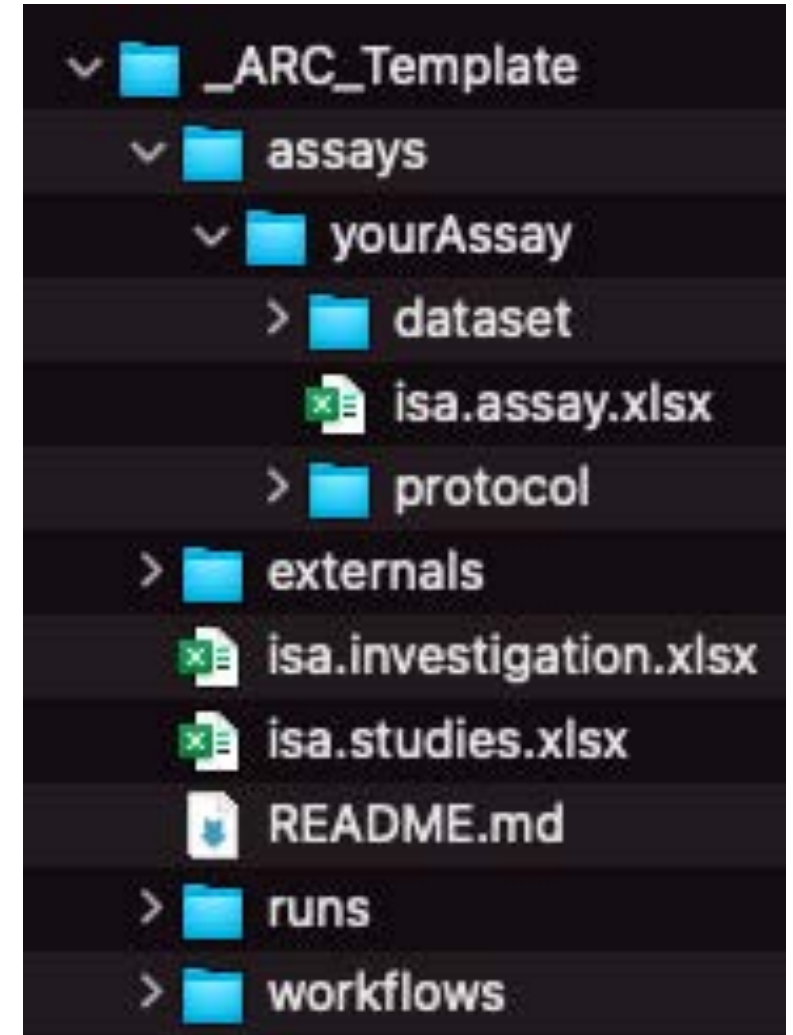
https://www.researchgate.net/publication/333163209_Design_and_Development_of_a_Phenotypic_Data_Model_PDM





Annotated Research Context (ARC)

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

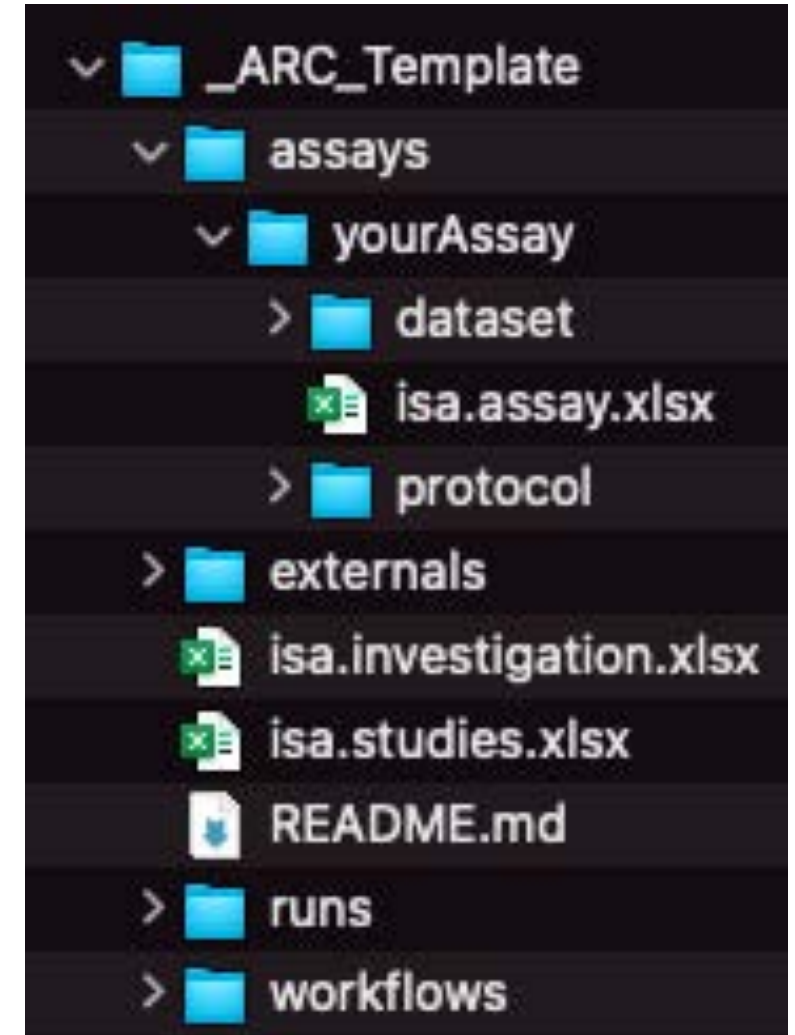




Annotated Research Context (ARC)

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

More on these in upcoming sessions ...



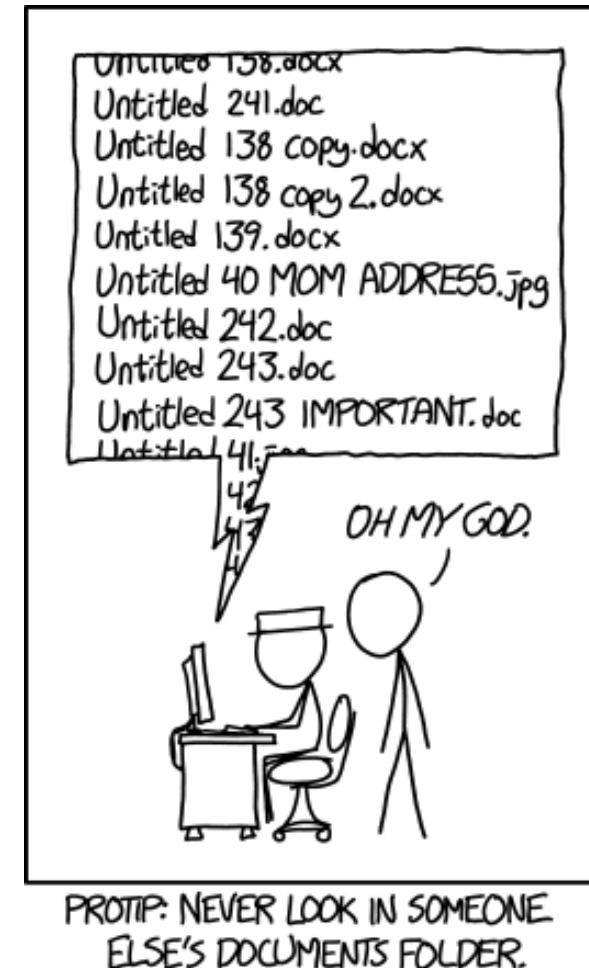
File Naming





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



<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 (lastnameef)
- Numbering (**001, 002, ... 010** - **NOT 1, 2, ... 10**)

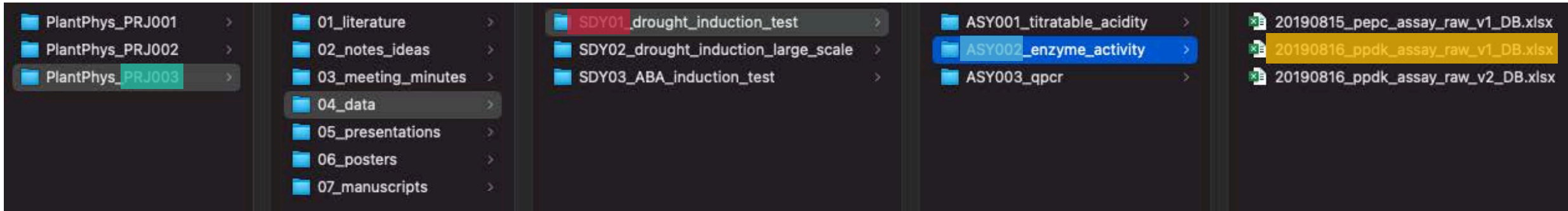




Directory structure or file name?

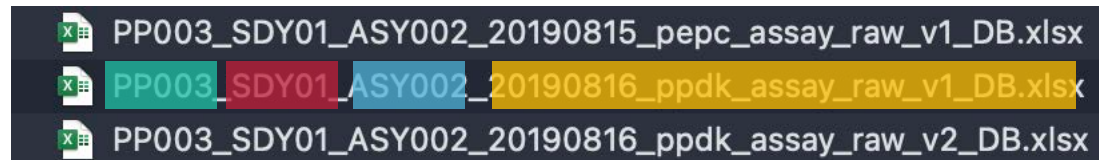
Find and understand data by

... location (path) + filename



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

...filename only



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”



- 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.