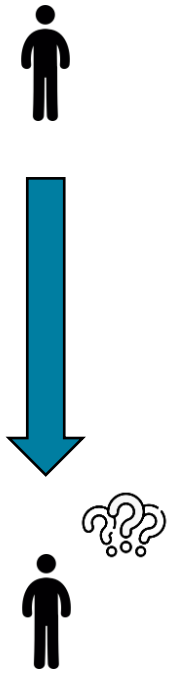
Naming, structuring and organising data

Why?



You :

- Easier to locate a file
- Find similar files together
- Moving files becomes much easier
- Easy to identify which files you want to back up
- Keep organized in the long-run
- Increases productivity
- Helps you to keep and maintain a record of the project

Someone else (& their machines) : Projects can easily be understood

Sources : - https://datamanagement.hms.harvard.edu/plan-design/directory-structures

- Data Management: File Organization by Data Management Services.Copyright © 2022-04-12MASSACHUSETTS INSTITUTE OF TECHNOLOGY is licensed under a Creative Commons Attribution 4.0 International License except where otherwise noted. [https://creativecommons.org/licenses/by/4.0/]. Access at http://bit.ly/fileOrgSlides.

Naming Convention – General Guidelines

what not to do	what to do
no long names	Limite file names to 32 characters (or less)
no special character ($^&()+=?\/!@*%{}[]<>#)	use numbers and letters
no space, hyphen or dash (-) between two words	use underscore (_)
Do not vary	same information in the same order same abbreviations and acronyms same date format = naming convention

Naming Convention - Advices

Folder	File
Begin with a number	Indicate the version number
Example : [number]_[foldername]	Example : [project]_[filesubject]_[subsubject]_[date]_[version].[extension]

Naming Convention - Tools

– ExifToolGUI (Windows) : https://exiftool.org/gui/

– pyExifToolGUI (Linux/Mac/Windows) : https://github.com/hvdwolf/jExifToolGUI/releases

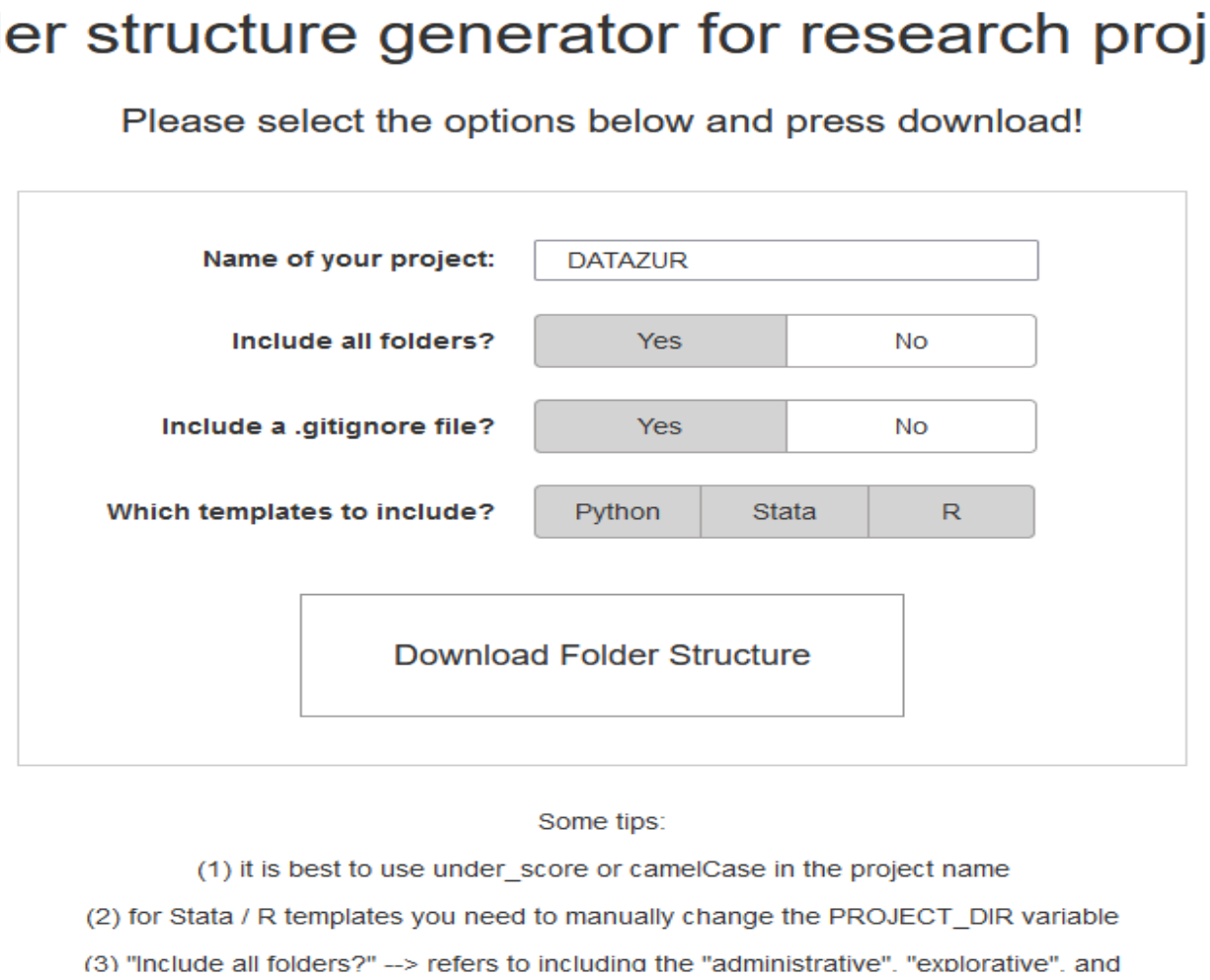
Folder structure : General Guidelines

- Use folders to group files with common properties
- Keep raw data separate
- Include source code in processed data folder
- Not too deep a tree structure (max. 5 levels)
- Separate consent forms of different types

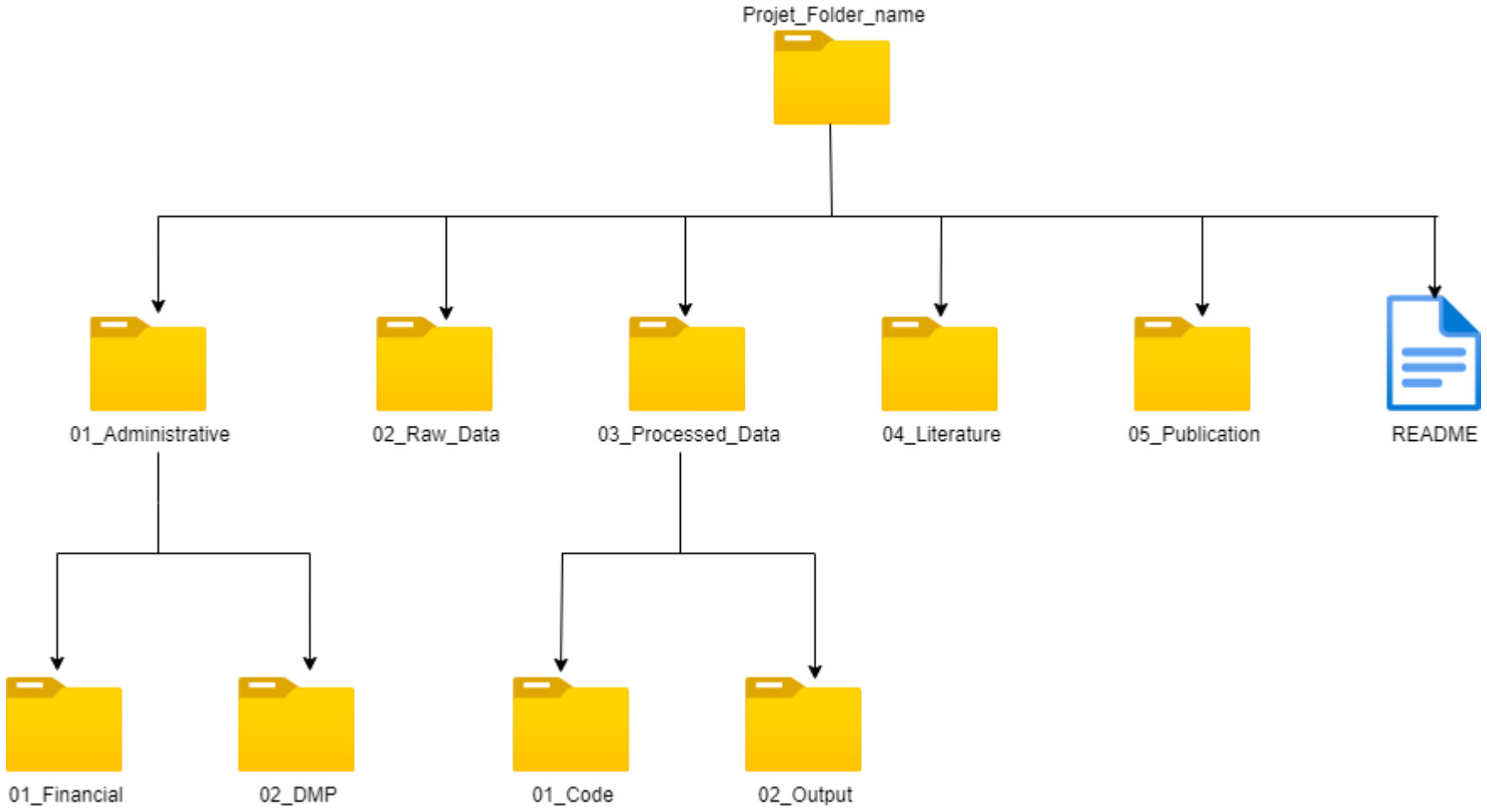
Source : https://library.bath.ac.uk/research-data/working-with-data/organising-data

Folder structure : Tool

Source : https://www.tiesdekok./folder-structure-generator/



Folder structure : an example



Files – Advice & Tool

– Organisation (spreadsheet) : 1 cell = 1 information

– Tool - File extension : FACILE (CINES) - https://facile.cines.fr/

AAC AAC	Advanced Audio Codings	[fmt/199]		Format Mpeg-4 contenant uniquement un flux audio au format AAC.	✔
AIFF PCM	Audio Interchange File Format	[fmt/414]	[audio/x-aif, audio/x-aiff]	Format audio contenant uniquement un flux PCM.	✔
APNG	Animated Portable Network Graphics	[fmt/935]	[image/vnd.mozilla.apng, image/apng]	L'APNG est une extension du format PNG permettant de réaliser des animations graphiques.	✖
DAE UTF-8 1.4.1	Collada		[application/xml]	Format permettant de stocker des données géométriques sous forme de scènes (plusieurs objets combinés dans le même référentiel), et d'y ajouter des informations supplémentaires pour décrire la scène et les objets (matériaux, environement lumineux, animations, ...) ou pour ajouter des notions sémantiques (relations entre les objets, découpage d'un objet en plusieurs éléments fonctionnels, etc...).	✖
FLAC FLAC 1.2.1	Free Lossless Audio Codec	[fmt/279]	[audio/ogg, audio/x-flac]	Format audio compressé sans perte.	✔
GIF 87a	Graphics Interchange Format	[fmt/3]	[image/gif]	Format image pouvant contenir également des animations.	✔
GIF 89a	Graphics Interchange Format	[fmt/4]	[image/gif]	Format image pouvant contenir également des animations.	✔
GeoTIFF	Geographic Tagged Image File Format	[fmt/155]	[image/tiff]	Format dérivé du TIFF contenant des informations de géoréférencement et de géolocalisation.	✔
HDF5 1.0	Hierarchical Data Format	[fmt/286]		Format de données à caractère scientifique.	✖
HDF5 2.0	Hierarchical Data Format	[fmt/287]		Format de données à caractère scientifique.	✖
JPEG RAW	Joint Photographic Experts Group - Raw JPEG Stream	[fmt/41]	[image/jpeg]	Format de représentation compressée d'une image fixe.	✖
JPEG2000	JPEG 2000	[fmt/151, x-fmt/392]	[image/jp2]	Extension du format JPEG.	✔
JPEG 1.00	Joint Photographic Experts	[fmt/42]	[image/jpeg]	Format de représentation compressée d'une image fixe.	✔

README - A File to list them all, a File to find them, a File to bring them all together and link them (in the in the structuring?)

– File .txt or .md

– Informations :

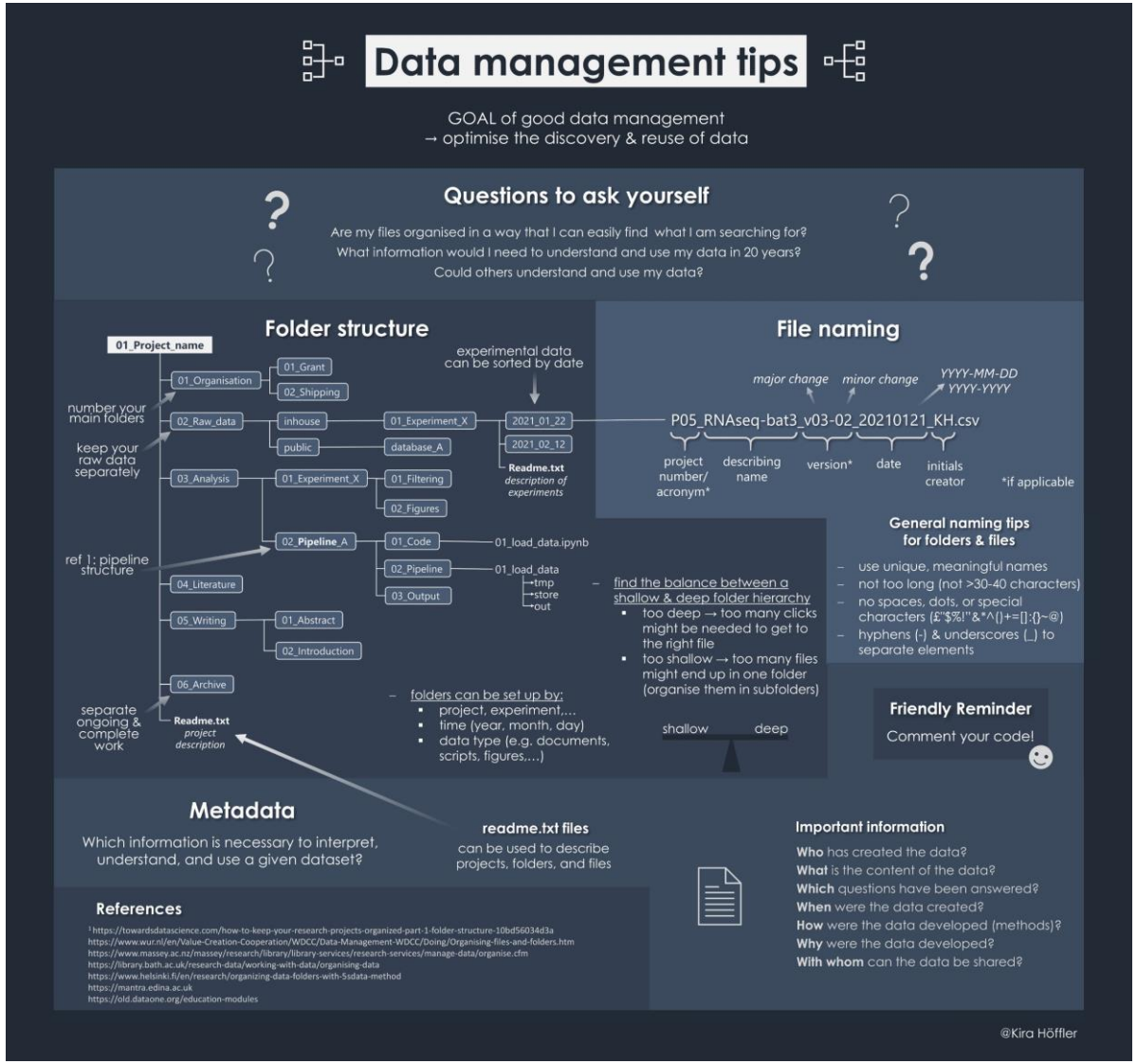
- Dataset title
- Author(s)
- Contact
- Date of last update
- Naming Convention (date format, abbreviations and acronyms…)
- Folder Structure
- Content of the dataset
- Methodological information (context of creation, reuse…)
- Any information relevant to the description of the dataset

README - A File to list them all, a File to find them, a File to bring them all together and link them (in the in the structuring?)

- Examples :
  - Entrepôt Recherche Data Gouv : https://recherche.data.gouv.fr/fr/categorie/33/guide/modele-de-readme
  - MIT : https://www.dropbox.com/sh/ys6f4wi5vtcxpg4/AAA1fWWejy5-Ayp4-yXbf1tQa?dl=0

- Mandatory file for deposit in a distribution warehouse (like https://entrepot.recherche.data.gouv.fr/dataverse/root)

Conclusion



Source: https://twitter.com/KiraHoeffler/status/1367804034413920259/photo/1

Thanks!