MANAGING YOUR RESEARCH DATA

Michelle Edwards & Carol Perry

Starting your Research on the Right Foot
Part 1

Used under CC BY-SA 2.0. Retrieved from: https://www.flickr.com/photos/australianshepherds/5501688419/

Objectives

- Understand the components of research data stewardship
- Apply concepts to your research project

Let‘s talk about you

- Where are you now with your research?

WHAT IS DATA?

- “factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation”

- “output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful”
  Merriam-Webster Dictionary - https://www.merriam-webster.com/dictionary/data

- “writings, notes, numbers, symbols, text, images, films, video, sound recordings, pictorial reproductions, drawings, designs or other graphical representations, procedural manuals, forms, diagrams, work flow charts, equipment descriptions, data files, data processing algorithms, or statistical records”
  CASRAI Dictionary http://dictionary.casrai.org/Data

RESEARCH DATA vs DATA

- Information collected during a research trial?
- Name some examples of research data

1.
2.
3.
4.
5.

DATA HASA STORY

- Different types of data – measurements, images, textual information
- Different sources of data – project, government, collaborative partners
- Use different parts of data for different analyses
- Use different parts of data for publication outputs – tables, plots, images

- Can we treat all of our data in the same way?

WHAT IS RESEARCH DATA MANAGEMENT?

- We collect data, save it on our computers, analyze it using some software, write-up our results, and hopefully publish
- What do we need to manage????

RESEARCH DATA MANAGEMENT (RDM)

- Allows us to ensure that the story about the data is captured and preserved
- The "story" of the researcher's data collection process
  - ensuring the processes are organized, understandable, and transparent
- By preserving data's story, we can reproduce data, analysis, outputs

WHY SHOULD WE CARE ABOUT RDM?

- Ethical and legal obligations
  - Research ethics board
  - Funding agencies – Tri-Agency: NSERC, SSHRC, and CIHR

- Publication requirements
  - Some journals require data to be included with paper
    - e.g. SpringerNature https://www.springernature.com/gp/authors/research-data-policy/data-policy-types/12327096

WHY SHOULD WE CARE ABOUT RDM?

- Reuse data later
  - Replication purposes
  - Sharing data

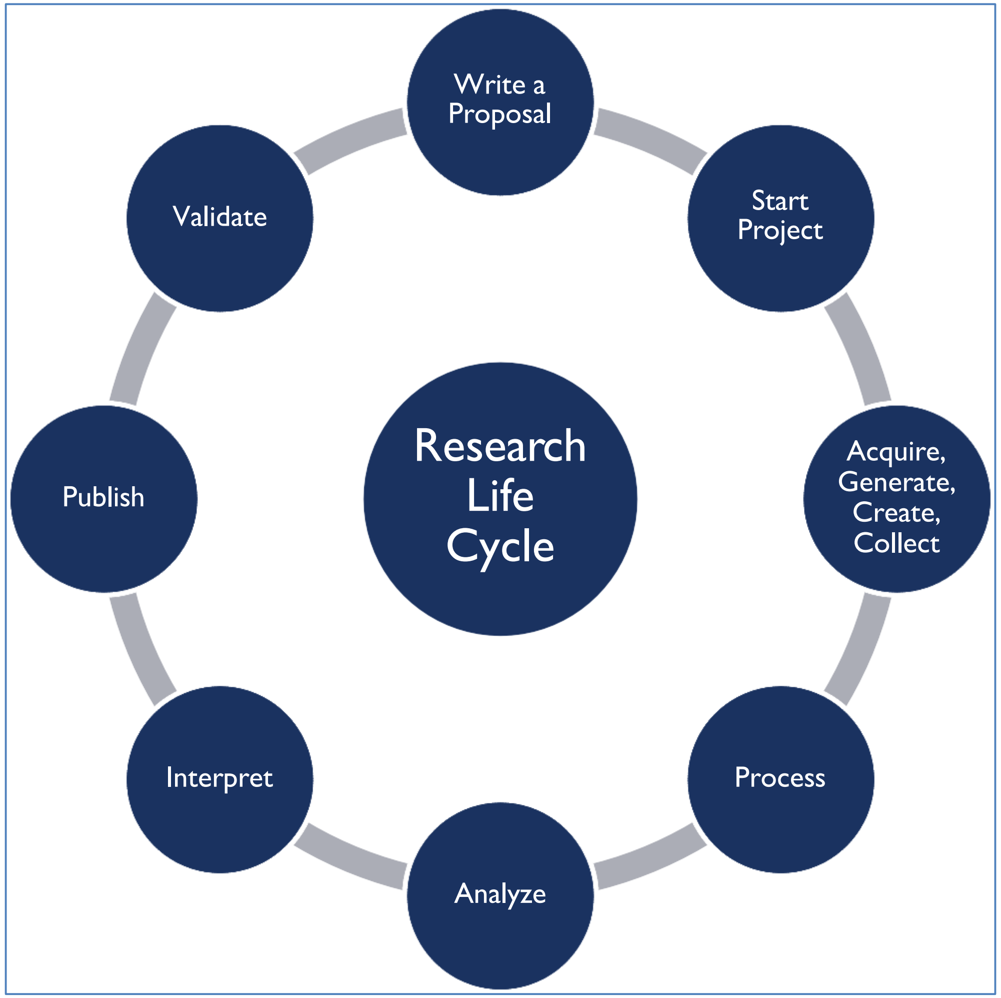
- Mitigate Risks
  - File corruption
  - Lost data
  - Hard drive failure
  - Old software
  - Human error
  - Unforeseen disasters

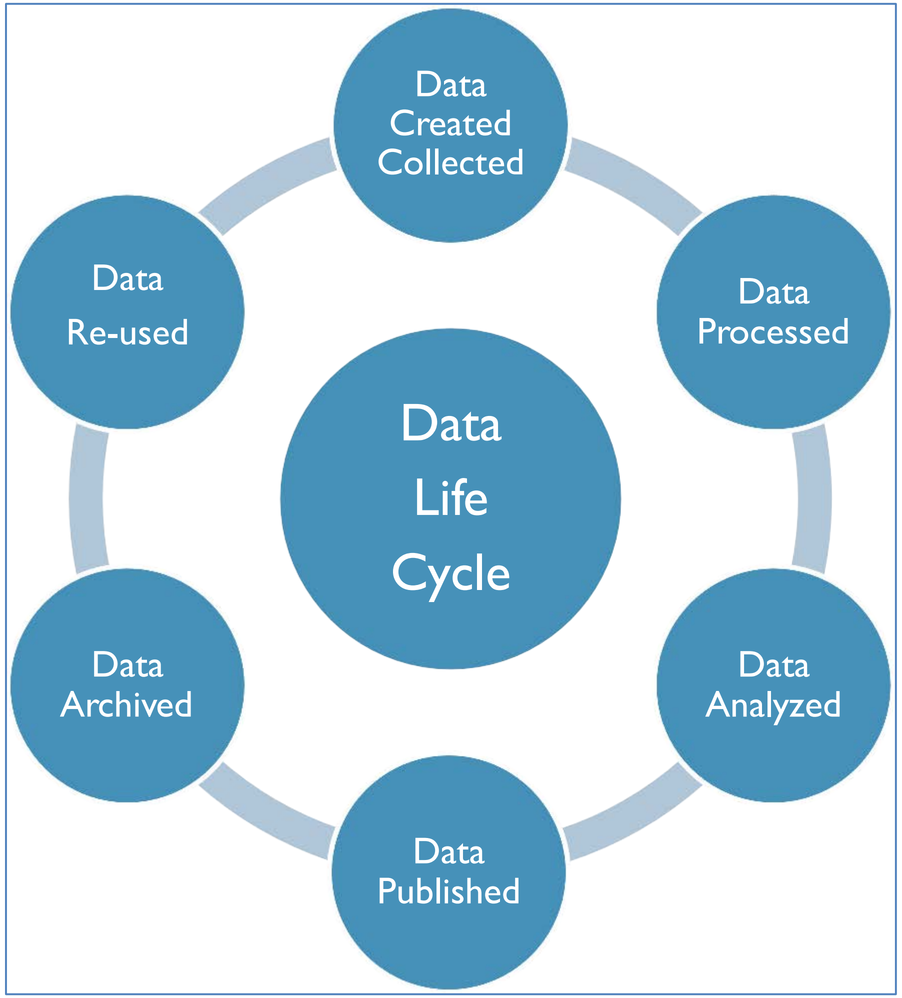
Sergio L.A. Used under CC BY NC licence Retrieved from: https://www.flickr.com/photos/sescobar/3291765330/sizes/m/

PERSONAL REASONS - WHY YOU SHOULD CARE

- Can you find your data?
- In 6 months – will you still understand your files?
- If you leave your data with your supervisor when you graduate – will they understand what you did?
- Do you need to provide your data to an agency or collaborator when you are finished?

RESEARCH LIFE CYCLE | DATA LIFE CYCLE





Research life cycle – Write a proposal

By Sizzlingbadger at English Wikipedia [Public domain], via Wikimedia Commons
https://upload.wikimedia.org/wikipedia/commons/c/c6/Suri-alpaca.jpg

- UpNorth Alpaca Research Team
  - Apply for an AAFC grant
  - Conduct a 2-year project
  - Assess the affect of diet on alpaca fleece weight and fibre quality
  - 2 breeds
    - Suri
    - Huacaya

By Sizzlingbadger at English Wikipedia [Public domain], via Wikimedia Commons
https://upload.wikimedia.org/wikipedia/commons/c/c6/Suri-alpaca.jpg

Research life cycle – Acquire, generate, create, collect

- How do you collect data?
- What format do you use when you collect data?
- How will you organize it?
- Where will you store it?
- Who will have access to it?

A Data Management Plan will answer ALL of these questions

DATA MANAGEMENT PLAN (DMP)

- Steps to developing a Data Management Plan (DMP)
  1. Organizing the data you've collected
  2. Documenting your work
  3. Managing your files – processing and analyzing your data
  4. Storing, securing, and backing up your files
  5. Preserving your data
  6. Accessing, sharing, and reusing your data

DMP Assistant

Canadian

Online

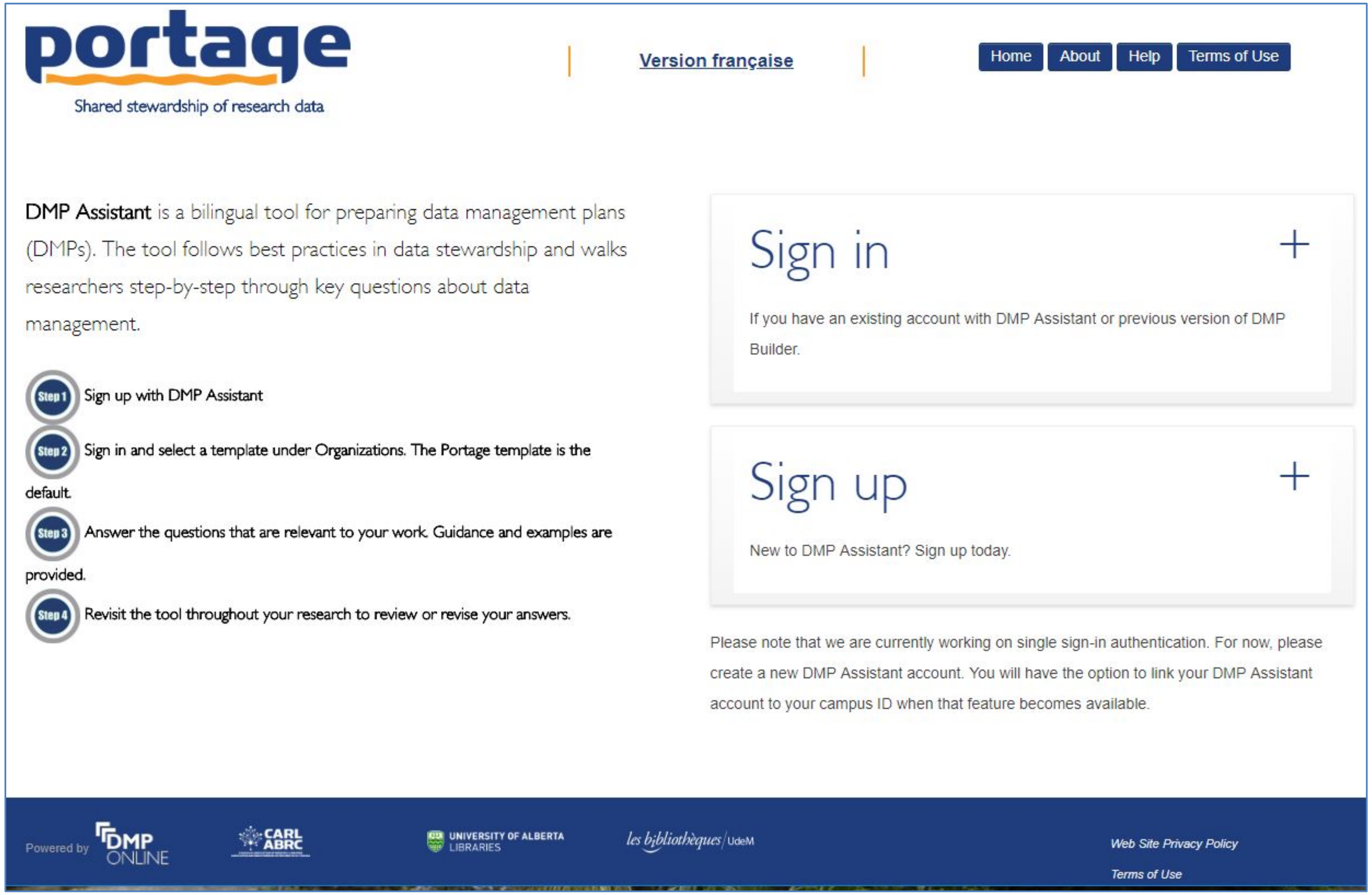
Bilingual

Data Management

Planning

Tool

https://assistant.portagenetwork.ca



DMP Assistant is a bilingual tool for preparing data management plans (DMPs). The tool follows best practices in data stewardship and walks researchers step-by-step through key questions about data management.

Step 1 Sign up with DMP Assistant

Step 2 Sign in and select a template under Organizations. The Portage template is the default.

Step 3 Answer the questions that are relevant to your work. Guidance and examples are provided.

Step 4 Revisit the tool throughout your research to review or revise your answers.

Sign in +

If you have an existing account with DMP Assistant or previous version of DMP Builder.

Sign up +

New to DMP Assistant? Sign up today.

Please note that we are currently working on single sign-in authentication. For now, please create a new DMP Assistant account. You will have the option to link your DMP Assistant account to your campus ID when that feature becomes available.

COLLECTING YOUR DATA

- How do you collect your data? What methods do you use?
  1.
  2.
  3.

- Who collects your data?
- Any challenges here? How do you mitigate these challenges?

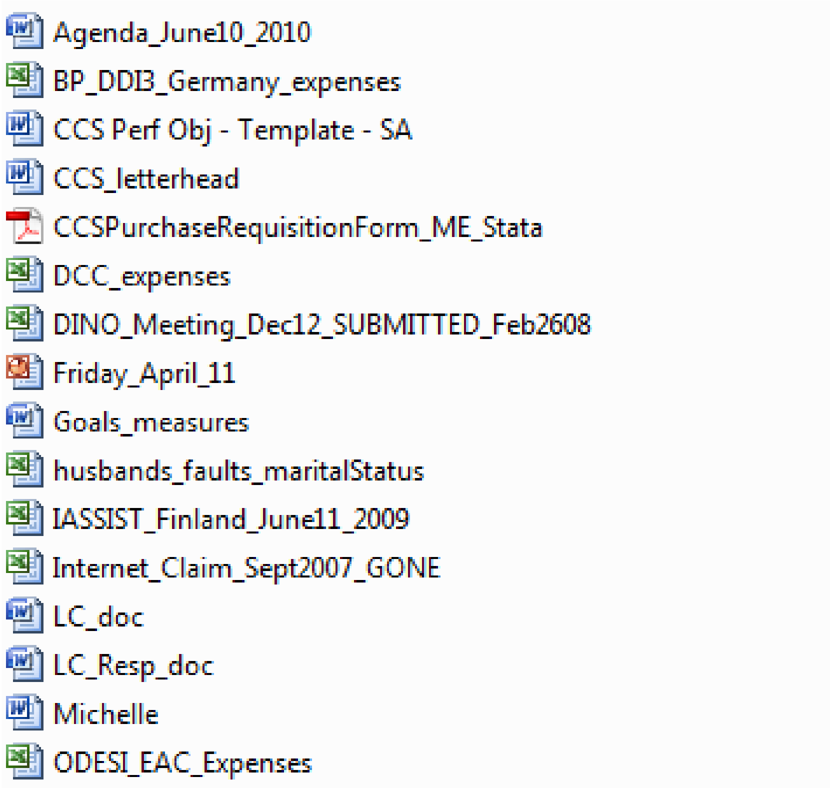
COLLECTING YOUR DATA

- How do enter your data into a file?   For instance an Excel file?
- May need to transcribe data from paper to Excel – who does this?
  - What happens when the transcriber cannot read the original paper?
- How/where is the process documented?
- Consider creating Standard Operating Procedures (SOP) for data collection and data entry https://www.uoguelph.ca/research/services-divisions/ethics/sops

Take a Break

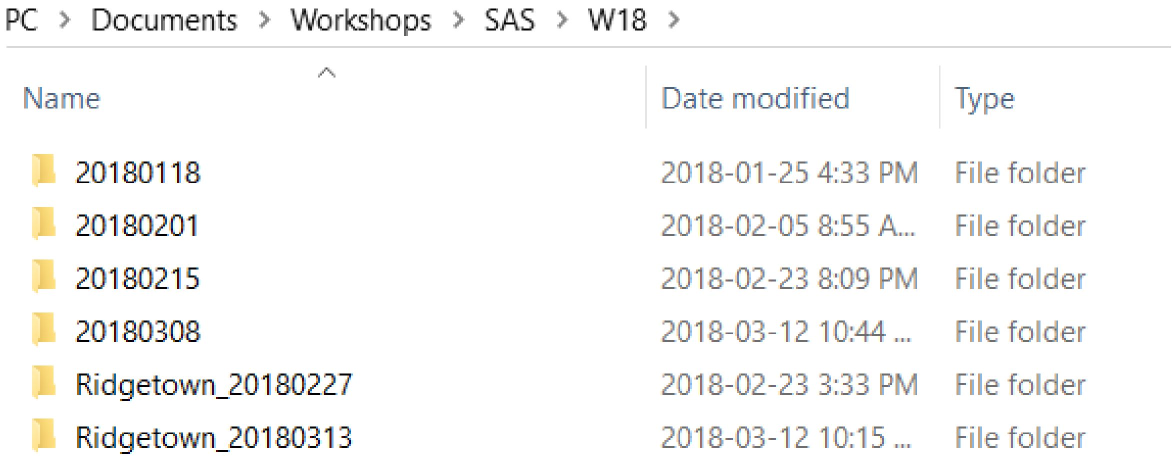
ORGANIZING YOUR PROJECT FILES

Does this look familiar?



- Agenda_June10_2010
- BP_DDI3_Germany_expenses
- CCS Perf Obj - Template - SA
- CCS_letterhead
- CCSPurchaseRequisitionForm_ME_Stata
- DCC_expenses
- DINO_Meeting_Dec12_SUBMITTED_Feb2608
- Friday_April_11
- Goals_measures
- husbands_faults_maritalStatus
- IASSIST_Finland_June11_2009
- Internet_Claim_Sept2007_GONE
- LC_doc
- LC_Resp_doc
- Michelle
- ODESI_EAC_Expenses

Or does this look familiar?



PC › Documents › Workshops › SAS › W18 ›

Name	Date modified	Type
20180118	2018-01-25 4:33 PM	File folder
20180201	2018-02-05 8:55 A...	File folder
20180215	2018-02-23 8:09 PM	File folder
20180308	2018-03-12 10:44 ...	File folder
Ridgetown_20180227	2018-02-23 3:33 PM	File folder
Ridgetown_20180313	2018-03-12 10:15 ...	File folder

Research life cycle - Process

- Set up Project Folder structure
  - Follow the structure of your project

- Assign an acronym to your project:
  - E.g. Alpaca Fibre Study = AFS

- All folders will start with this acronym – e.g. AFS_Budget
  - Keep your folder names short and clear to understand
  - NO spaces!!!! Use an underscore _

ORGANIZING YOUR PROJECT FOLDERS

- Sample Directory/ Folder Structure

- AFS
  - AFS_Budget
  - AFS_Data
    - AFS_Data_2018
    - AFS_Data_2019

- AFS
  - AFS_Budget
  - AFS_Data
    - AFS_Data_Huacaya
    - AFS_Data_Suri
  - AFS_SAS
  - AFS_Ouput

ORGANIZING YOUR PROJECT FOLDERS

AFS ← Top folder for project

AFS_Data ← Where all data will be saved for this project

AFS_Data_201806 ← Data collected in June 2018 is saved here

AFS_Data_201806_Suri.xlsx ← Data collected in June 2018 from Suri breeders

AFS_Data_201806_Huacaya.xlsx ← Data collected in June 2018 from Huacaya breeders

ORGANIZING YOUR PROJECT FOLDERS

- Create a README file – save in your top directory or main folder

  A text file that:

  - Defines your acronyms
  - Describes your project and the folder structure
  - Defines what files will be in each directory or folder
  - Think of this README file as an annotated Table of Contents to your project folder structure
  - AFS_README.txt

README FILE – STARTINGTO DOCUMENT!

Title:Alpaca Fibre Study (AFS)

Short abstract or project statement

AFS_Budget = Budget information for the project

AFS_Data = Data collected

- AFS_Data_201806; AFS_Data_201807;AFS_Data_201808

AFS_SAS =All SAS programs

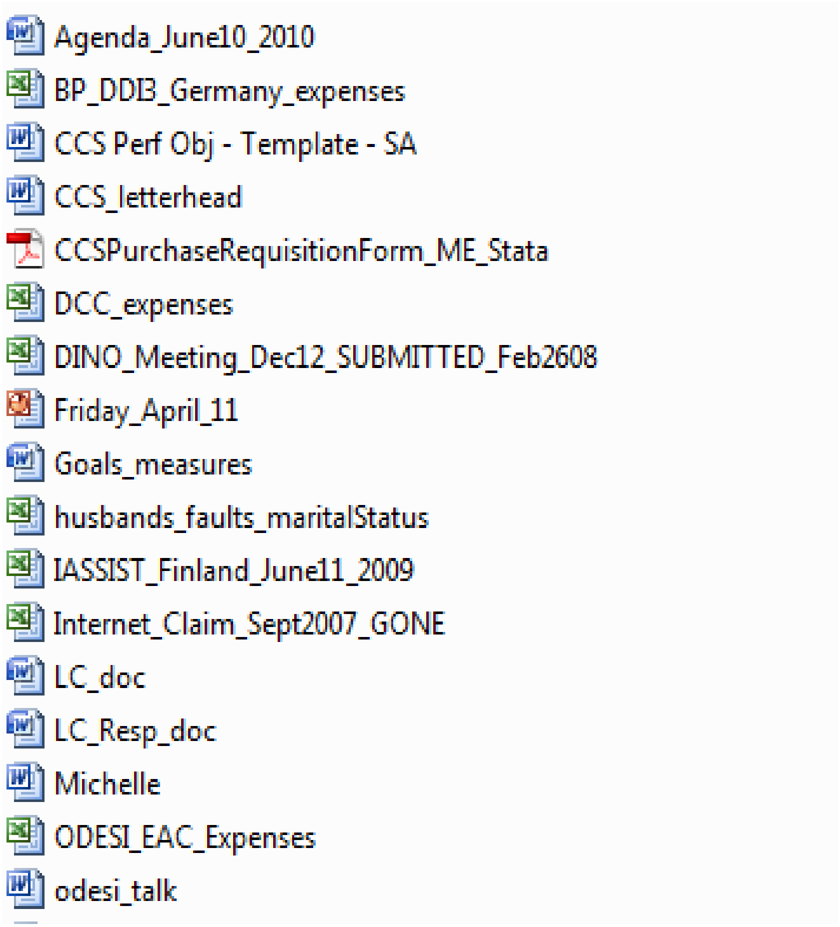
AFS_Output =All SAS outputs

Data collected from 2018-06-01 to 2019-06-01

Price data collected in dollars per pound

NOTE: 2018-06-15 Rain caused data collection to be delayed until 2018-06-20

FILE NAMES



Agenda_June10_2010

BP_DDI3_Germany_expenses

CCS Perf Obj - Template - SA

CCS_letterhead

CCSPurchaseRequisitionForm_ME_Stata

DCC_expenses

DINO_Meeting_Dec12_SUBMITTED_Feb2608

Friday_April_11

Goals_measures

husbands_faults_maritalStatus

IASSIST_Finland_June11_2009

Internet_Claim_Sept2007_GONE

LC_doc

LC_Resp_doc

Michelle

ODESI_EAC_Expenses

odesi_talk

Can you guess what is in these files?

1. Agenda_June10_2010
   - Agenda for what?
2. husbands_faults_maritalStatus.xlsx
3. Michelle.docx

FILE NAMING

Be descriptive

- Less than 25 characters - preferred
- Names independent of location (create project id or acronym)

Be consistent

- Version identification
- Use standard date formats eg. yyyymmdd
- Avoid unusual characters !@#$%^&*()+
- Use underscores between words or capitalize first letter of each word

- Example:
  - afs_codebook_2018_02_13.pdf
  - afsCodebook20180213.pdf

Interpretation = project id_description of file_ISO date format.file format

FILE NAMES - EXAMPLES

AFS_Budget

  - AFS_Budget_2018_Expenses.xslx
  - AFS_Budget_2018_Revenues.docx

- AFS_SAS
  - AFS_SAS_20180630_DescStats.sas
  - AFS_SAS_20180630_Model.sas

- Provide detailed explanation of contents in the README file!

MANAGING YOUR FILES

√ We have a project directory structure

√ We have consistent file name conventions

How are we going to manage our files?

- By month of data collection?
- By breed?
- What about comparisons over the years?
- Changes in an outcome variable from the beginning of the trial to the end?

MANAGING YOUR FILES

What is practical for data collection?

- One file with all the data? Or
- A separate file each time you take measurements?

1. Measure feed consumption every 30 days OR
2. Measure animal weights every 2 weeks.

- Same trial is repeated in 2018 and 2019 – one file or more than one file?

PLAN NOW SAVE TIME LATER!

- Sounds like a lot of work to plan out your directories, file names, and document it!
- It will save you a lot of time later! Especially when you go back after being away for a bit.

Exercise - Organization

VARIABLE NAME RESTRICTIONS AND LIMITS

Length of Variable Name

- SAS: 32 characters long
- Stata: 32 characters long
- Matlab: 32 characters long
- SPSS: 64 bytes long
  - 64 characters in English
  - 32 characters in Chinese
- R: 10,000 characters long

1st Character of Variable Name

- SAS: MUST be a letter or an underscore
- STAT: MUST be a letter or an underscore
- Matlab: MUST be a letter
- SPSS: MUST be a letter, an underscore or @,#,$
- R: No restrictions found

VARIABLE NAME RESTRICTIONS AND LIMITS

Special Characters in Variable Names

- SAS: NONE
- Stata: NONE
- Matlab: No restrictions found
- SPSS: NONE except Period, @
- R: NONE except Period

Case in Variable Names

- SAS: Mixed case –Presentation only
- Stata: Mixed case – Presentation only
- Matlab: Case sensitive
- SPSS: Mixed case – Presentation only
- R: Mixed case – Presentation only

NO BLANKS (SPACES) allowed in any of the Statistical Packages
Beware of Function names in all Statistical Packages – these cannot be used as Variable Names

BEST PRACTICES FOR VARIABLE NAMES

1. Set Maximum length to 32 characters
2. ALWAYS start variable names with a letter
3. Numbers can be used anywhere in the variable name AFTER the first character
4. ONLY use underscores "_" in a variable name
5. Do NOT use blanks or spaces
6. Use lowercase

VARIABLE NAMES INSIDE MY FILES

- Information or data that we are collecting:
  - Diet A⟶ diet_a
  - Fibre length in centimetres ⟶ fibre_cm
  - Location of farm ⟶ location
  - Price paid for fleece⟶ price

Exercise – Variables

DMP CHECKLIST

√ Organizing the data you’ve collected

√ Documenting your work

√ Managing your files – processing and analyzing your data

Storing, backing up, and securing your files

Preserving your data

Accessing, sharing, and reusing your data

DMP Assistant

- Can you fill in any of the sections of the DMP Assistant?
- Let’s review – remember want to use this as a Plan – stay ahead of the work

Contact

- Michelle Edwards
- edwardsm@uoguelph.ca

- Carol Perry
- carolp@uoguelph.ca
- lib.research@uoguelph.ca

By Funkipickle Retrieved from: https://www.flickr.com/photos/funkipickle/6137260892/
Used under CC BY-NC-ND 2.0