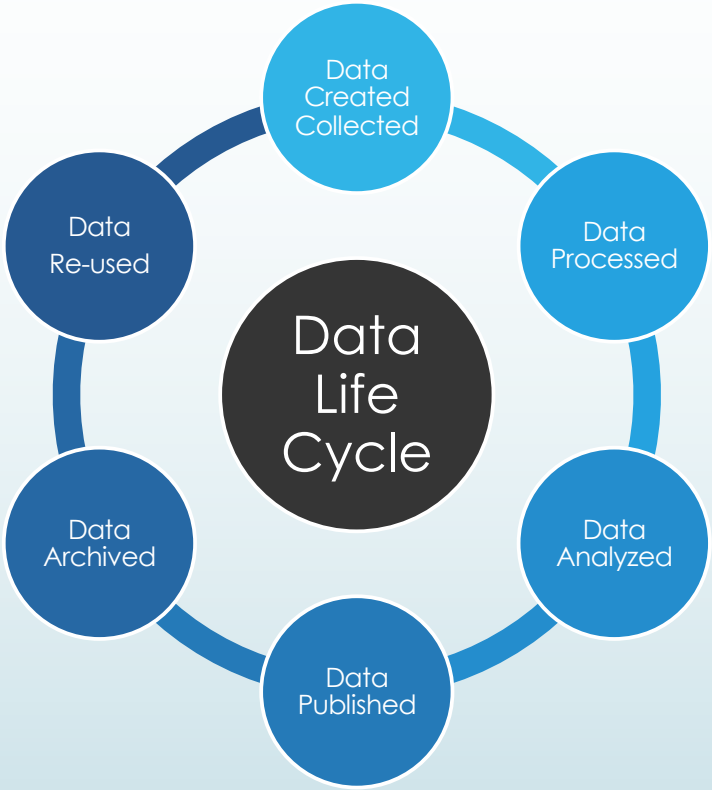
The Data Life Cycle and your role as a researcher?

A.M.Edwards, Ph.D., MLIS

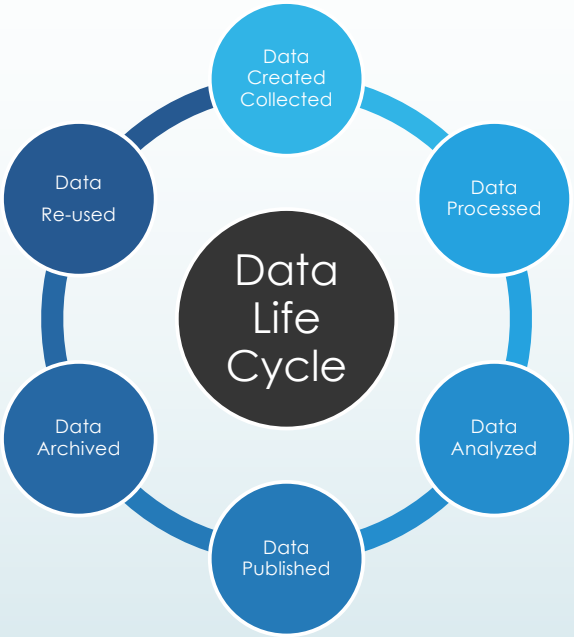
Director, Agri-food Data Strategy

ADC

Data Life Cycle



Where are you now with your research?



Data has a 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?

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

ADC

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?

ADC

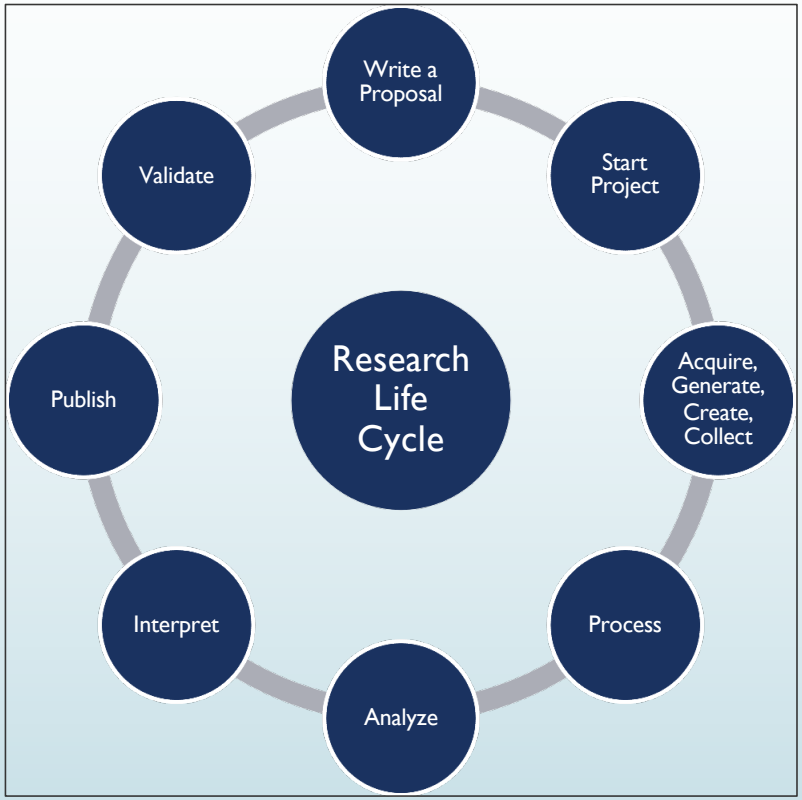
What happens when you DON'T?

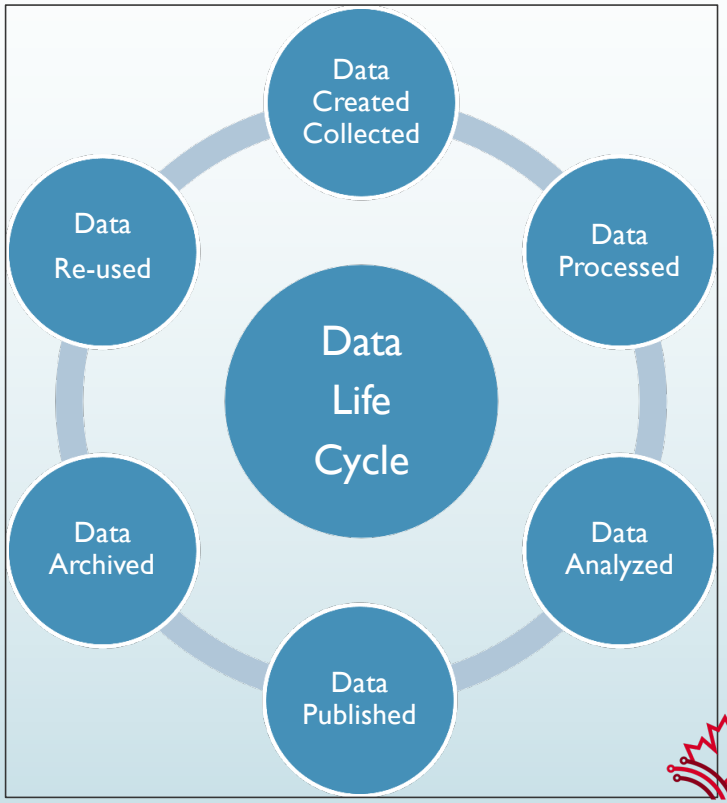
Worse case scenario

Brian Wansink – Cornell – 2018

(https://www.npr.org/sections/thesalt/2018/09/26/651849441/cornell-food-researchers-downfall-raises-larger-questions-for-science)

RESEARCH LIFE CYCLE | DATA LIFE CYCLE

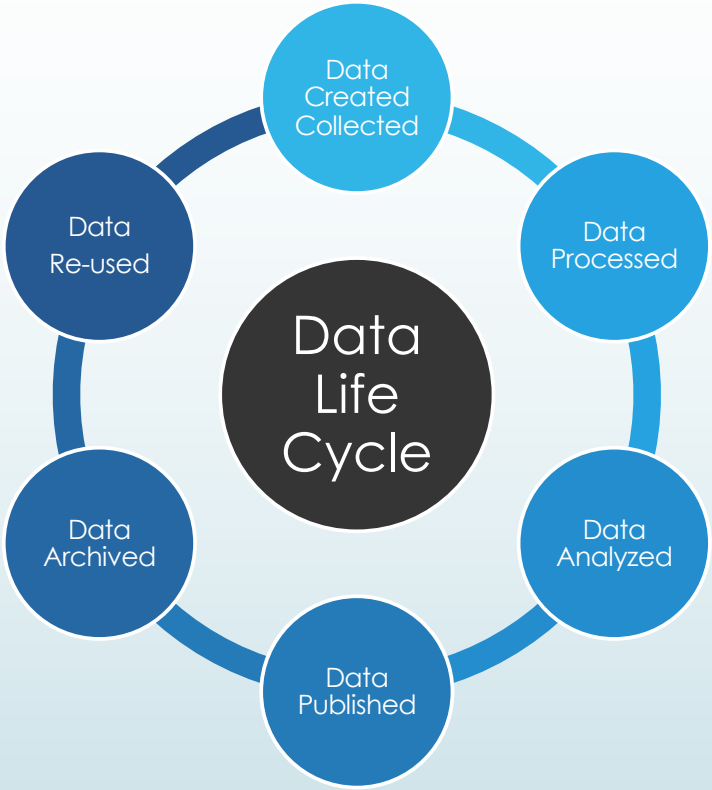




Research life cycle: Acquire, Generate, Create, Collect

- 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?

Data Life Cycle



Collecting the data

- How do you collect your data? What methods do you use?
- Who collects your data?
- Any challenges here? How do you mitigate these challenges?

Collecting the data

- How do enter your data into a file? Excel?
- 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

Organizing your project files

Does this look familiar? - **A**

- 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? - **B**

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

ADC

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 – Starting to Document

Title: Alpaca Fibre Study (AFS)

Short abstract or project statement

AFS_Budget = Budget information for the project

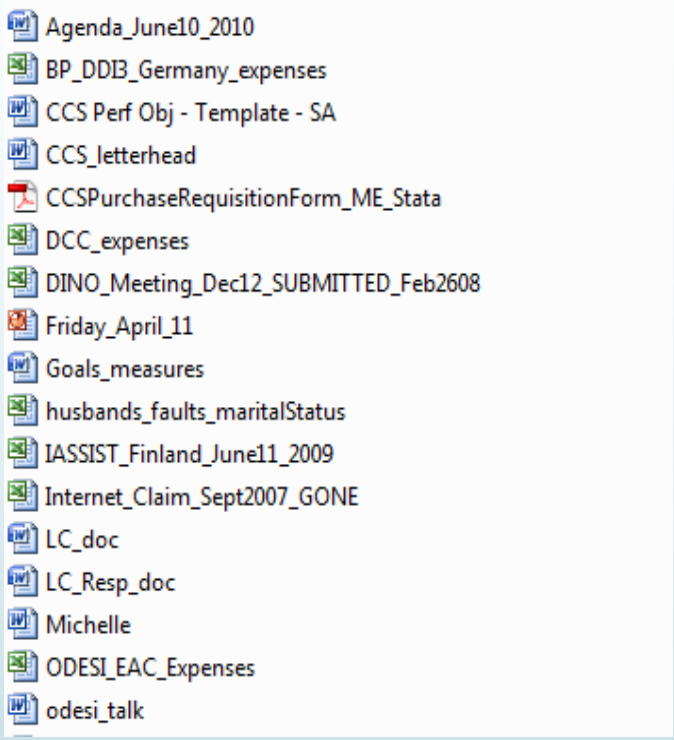
AFS_Data = Data collected

- 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

Filenames



➡ **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

ADC

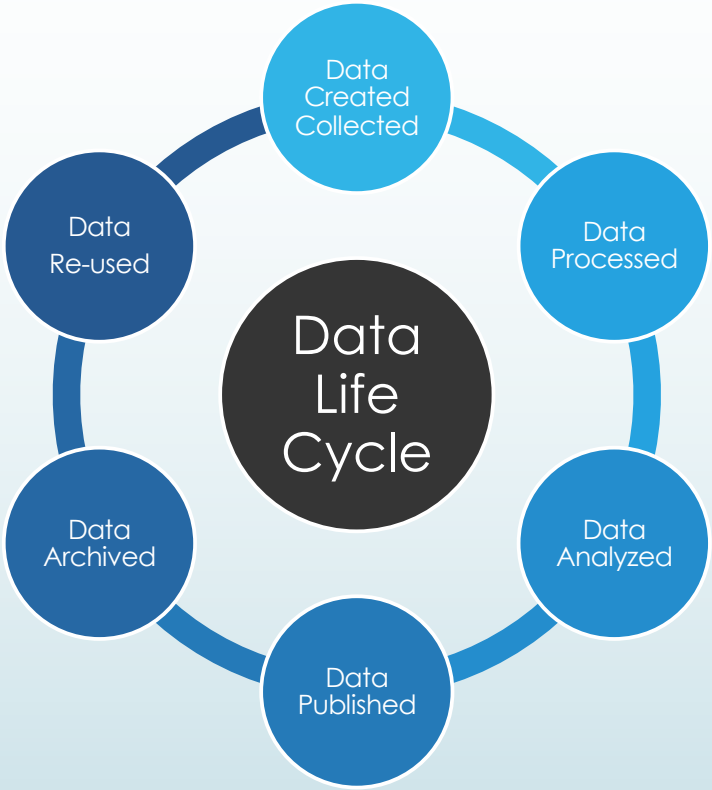
Plan now and Save time later!

- 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.

Discussion - Organization

- How are you currently organizing your files?
- Let's try to have a conversation

Data Life Cycle



Variable names: Limits and Restrictions

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 names: Limits and Restrictions

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

ADC

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

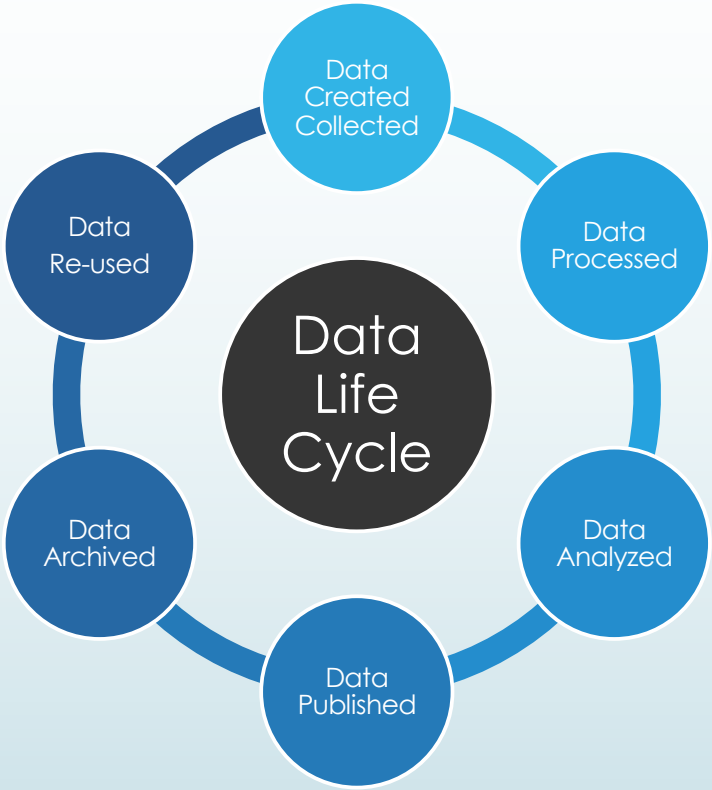
- Information or data that we are collecting:
  - diet_a
  - fibre_cm
  - location
  - price

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

Coffee Break

Data Life Cycle



Storing and Backing up data

What do you do???

Storing and Backing up data: 3 -2 -1 RULE

Keep at least **three** copies of your data

Store the copies on **two** different media

- (Department server, external hard drive, ~~USB,~~ etc.)

Keep **one** backup copy offsite

- Keep a '**master file**' – original untouched – for emergencies

Storing and backing up data

‘Master files’

- Secure location
- Separate folder with ‘master file’ in name
- Includes original documents, raw data files, final output files
- *NEVER* work on ‘master files’

‘Working files’

- Based on original master files
- May have multiple versions as you run tests, edit content, etc.
- If working file lost – create new copy from ‘master’ file

Storing and backing up data

- Backup all files on regular basis
  - Keep backup copy in separate location
  - Departmental servers offer regular backups
- Synchronize files on regular basis
  - Avoid multiple versions of same file (changed in one place but not updated in other copies)
- Encrypt sensitive data – see CCS services
  - Turns your information into a stream of what appears to be random symbols
  - You need a digital key to unlock – without the key the data is unusable
  - https://www.uoguelph.ca/ccs/encryption
    https://www.uoguelph.ca/ccs/infosec/encryptedusb

ADC

Storing and backing up data: File Versioning

AFS_Data_201806_Suri.xls

- If you edit the file …
  - Do you change the name? Do you include date of change?
    - V1.0, V1.1,
    - V1, V2, V3
- Be consistent
- Be clear
- Document changes made in a readme file or other note

Storing and backing up data: File Versioning

- Discusssion
  - What do you do right now?
  - Have you lost data or files?
  - What will you consider doing after today?

Securing your data

- What kind of data are you collecting?
- Does it include sensitive data?
  - Direct identifiers – names, SIN, Registration #s, …
- Who can access your data? Who should have access to your data?
  - Can we limit access?
  - How?
  - See CCS Information Security Policies https://www.uoguelph.ca/ccs/infosec/policies_and_procedures

Research data classification system

Works on three principles:

1. Level of sensitivity of the data – *i)* public, *ii)* internal/private, *iii)* confidential/sensitive

2. Level of risk or damage – 'probability of harm' assessed against 'magnitude of harm' (minimal, moderate, substantial)

3. Security measures recommended to mitigate risks – Level 1, Level 2, Level 3

ADC

Securing your data: When things go wrong

1. Parks Canada bans wildlife photographers from using radio receivers to locate animals (http://www.cbc.ca/news/canada/calgary/vhf-telemetry-receiver-ban-banff-kootenay-yoho-1.3717595)
   - Photographers were disturbing the natural environment of bears, elk, and wolves
   - New fines up to $25,000
2. Scientific data used to track and protect animals is vulnerable to hacking (http://www.cbc.ca/radio/day6/episode-324-lgbt-iranian-refugees-porn-o-nomics-microdosing-lsd-hans-rosling-and-more-1.3972896/scientific-data-used-to-track-and-protect-animals-is-vulnerable-to-hacking-1.3972929)
   - Bison reintroduction project in Banff – 5 are equipped with GPS collars or VHF radio collars

ADC

Securing your data: When things go wrong

Discussion

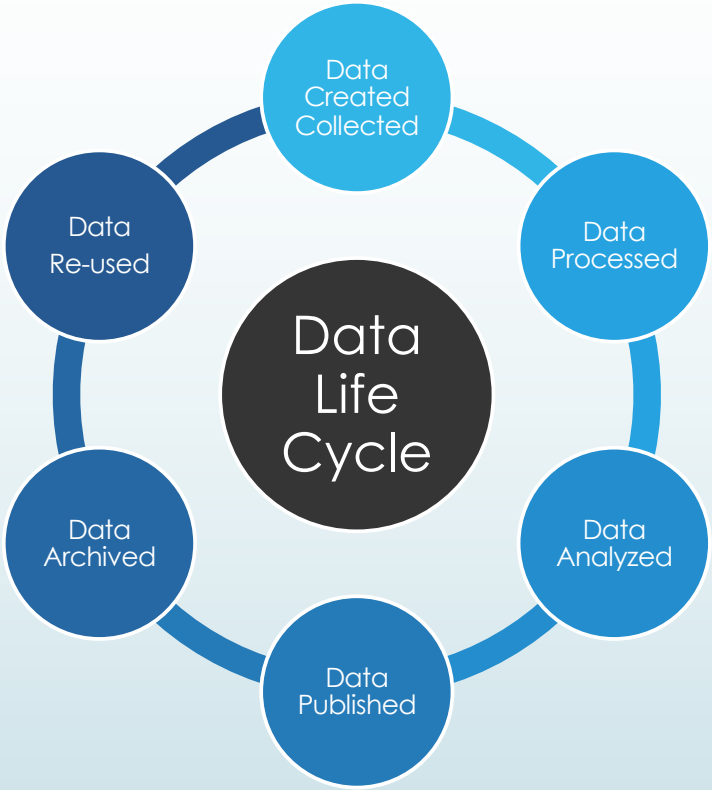
- Thinking about your data – should you be doing more to secure it?
- Does it matter?

Your project is complete!

- You've documented your data while working
- Filenames are fabulous!
- Readme files that describe everything to do with your data
- You've secured your data

Now what??

Data Life Cycle



Archiving vs Preserving your data

Data archiving:

- moving data that is no longer needed to a separate storage area for long-term storage. This data is still important and may be needed in the future.

Data preservation:

- ensuring that the data is still usable and accessible in the future. This requires a number of managed activities to ensure its use.

Data Preservation Challenges: Software

- Data collected in a variety of software…
  - VP Planner
  - Lotus 1-2-3
  - Quattro-Pro
  - first Mac version of Excel - 1985
  - first Windows version of Excel - 1987
  - Old excel version .xls

ADC

Data Preservation Challenges: Media

- Data stored on:
  - 1970s – magnetic tapes
  - 1980s – 5.25" diskettes
  - 1990s – 3.5" diskettes
  - 2000s – CD-ROMs
- Can we access this data today?
- If not, is this data irrelevant?

Data Preservation: Format

- At the end of your project save all data, statistical coding procedures, and outputs in a non-proprietary format.
- NO Excel, NO Word, NO SAS, NO SPSS, etc…

Data Preservation: Format

File type	Software used:	Save as:
Text	MS Word	pdf, pdf/A, txt, html
Images	Photoshop	Tiff (6.0 uncompressed)
Video	Quicktime	Mpeg4, Motion jpeg 2000
Data visualizations	Charts	Jpg, pdf
Mapping	ESRI	All files – shp, prj, sbx, sbn, dbf
Tabular data	MS Excel	csv
Numeric	Sas spss	Ascii, txt, csv
Database	MS Access	dbf

Check for errors after converting files

ADC

Data Preservation: Documentation

- When preparing your data and project for preservation you will need all of your files and

Documentation!

- Without documentation
  - your data is not meaningful and cannot be used in the appropriate manner
  - your study and analysis cannot be replicated!

ADC

Data Preservation: Documentation

- Documentation to include:
  - Codebook
    - explains variables – remember those labels?
    - shows what your values represent
  - Syntax files
    - all of your SAS codes
    - document what action each set of code is performing
  - Readme files
    - provide user with notes and description of directory structure
  - Any additional documentation necessary to understand your data

ADC

Sharing, Accessing, and Reusing your data

- Why would you want to share your data?
- What value does it add?
  - Increases impact of your research
  - Helps others replicate your research
  - Encourages further scientific enquiry
  - Reduces research costs by reducing duplication
  - Encourages transparency and accountability

ADC

Sharing, Accessing, and Reusing your data

- What data do you want to share?
  - Raw data files?
  - Data supporting publication?
- When do you want to share it?
  - Immediately upon publication?
  - After an embargo period to have time to expand on research?

Sharing, Accessing, and Reusing your data

- Before openly sharing your data you need to consider:
  1. Ethical and legal obligations and/or requirements
  2. Data anonymization
  3. Intellectual property rights

Ethical and Legal Obligations

1. Research ethics board
2. Funding agencies – Tri-Agency (NSERC, SSHRC, and CIHR); Polar Data Canada
3. Journal requirements
4. Partnership requirements

- Before sharing any data, review your ethical and legal obligations.
- In some instances you may not be ALLOWED to share or you may be REQUIRED to share

ADC

Data Anonymization

- Any shared data should be anonymized
  - no personal identifiers remaining in the dataset
  - users cannot recreate and identify individuals, units, etc..
- Methods of anonymization:
  - Aggregation
  - Pseudo anonymization

Data Anonymization

- How can you identify individuals in a dataset:
  - Direct identifiers
    - names
    - addresses
    - identification numbers – student ID, OHIP number….
  - Indirect identifiers
    - birth date,
    - detailed geographic areas
    - detailed information on income, place of birth, etc.
    - A combination of indirect identifiers could lead to the identification of an individual

ADC

Data Anonymization: Methods

- Aggregation
  - Example:
  - Income of the producer was collected
    - to anonymize this piece of information, aggregate income data
      (create larger grouping)

Producer X original data = $45,660/year anonymized data = $40,000 - $49,999

Age of Observer original data = 34 yrs anonymized data = 30-39 years

Data Anonymization: Methods

- Masking – pseudo anonymization
  - Data masking is a form of pseudo anonymization
  - Maintaining the integrity of the variable or information that was collected but replacing the real data with fictitious and masked information
    - **cannot** identify the individual unit
    - **can** still use the information to describe the sample or population

Data Anonymization: Pitfalls

- Tendency to replace all direct identifiers with pseudonyms or aggregate variables
- Avoid blanking out information
- Avoid over-anonymization – can lead to misleading conclusions
- Keep LOG of anonymization techniques!!
  - Secure data -> useable data

ADC

Intellectual Property Rights

- Who owns the data?
- Manage your intellectual rights to the data that you've collected through a license
- License will determine who can use the data, and what they can do with it.
- For more information please see Data licensing at the University of Guelph.. (https://www.lib.uoguelph.ca/get-assistance/maps-gis-data/research-data-management/preserving-sharing-data/licensing)

Sharing, Accessing, and Reusing your data

- We've finished our project
- Our data is clean and organized – and well documented!
- We CAN share our data
- So now what? Where and how do we share our data?
  - Pass out USB keys?
  - Provide the world with access to our Departmental Servers?

Preservation and Access options

- Institutional repository
- Publish data with results - journal
- Deposit in major data repository
- Deposit in discipline-specific data repository

University of Guelph based options

Atrium (institutional repository): e-theses, articles, reports, videos, etc.

https://atrium.lib.uoguelph.ca/

Agri-environmental Data Repository- research data - OAC data

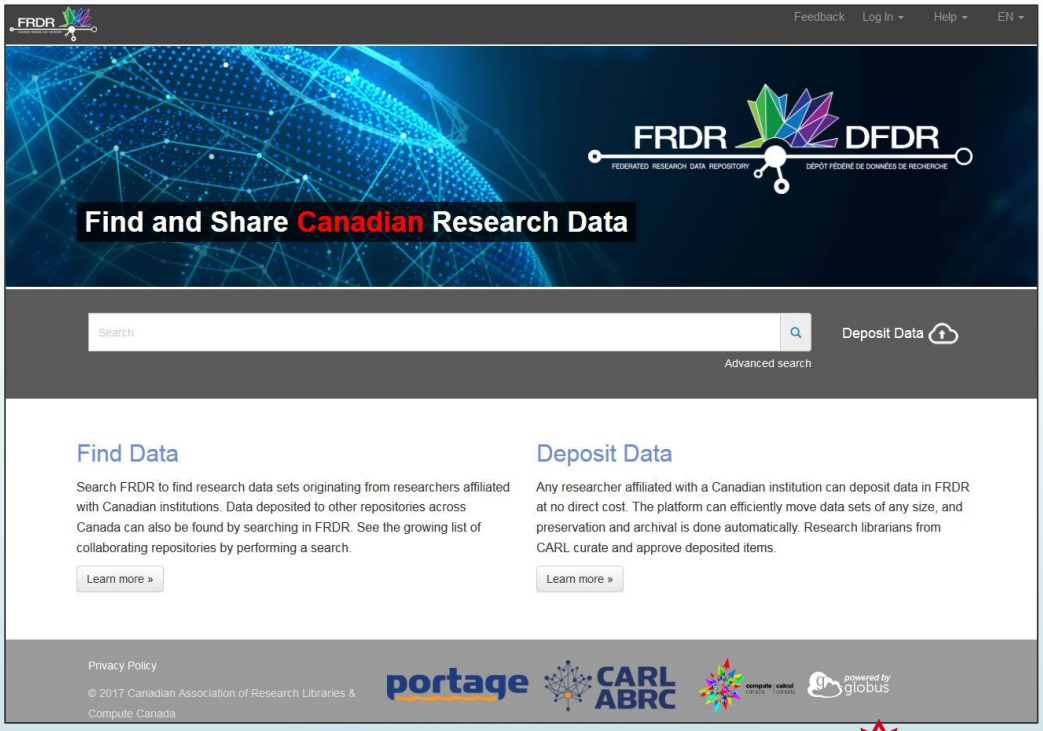
https://borealisdata.ca/dataverse/ugardr

University of Guelph Data Repository – research data – all disciplines

https://borealisdata.ca/dataverse/ugrdr

Federated Research Data Repository (FRDR) – now Lunaris

- https://www.frdr.ca
- https://www.lunaris.ca/en
- Canadian
- Data portal
- Big data repository



ADC

Discipline Specific Repositories

Re3data.org – global registry of research data repositories

www.re3data.org/

Choosing a Repository

- Is it reputable?
- Will it accept your data? Is it a good fit?
- Will your data be safe?
- Does it assign a persistent unique identifier?
- Does it provide analytics on data usage?
- Are there fees?
- What are your obligations under the service?
- What are the obligations of the service provider?

ADC

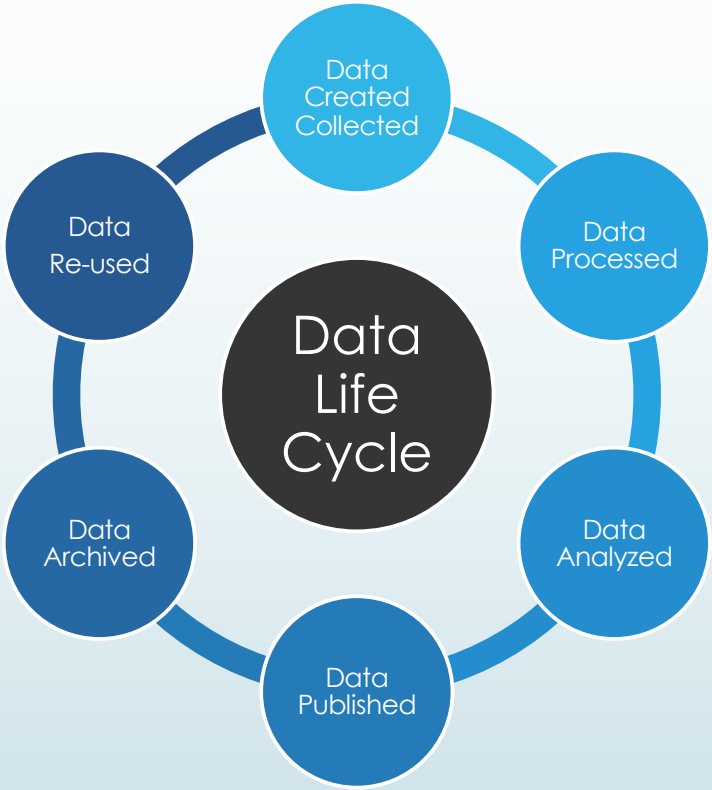
International initiatives: Data sharing and reuse

FAIR Principles: https://www.go-fair.org/fair-principles/

- Findable
- Accessible
- Interoperable
- Replicable

Research Data Alliance: https://www.rd-alliance.org/

Data Life Cycle



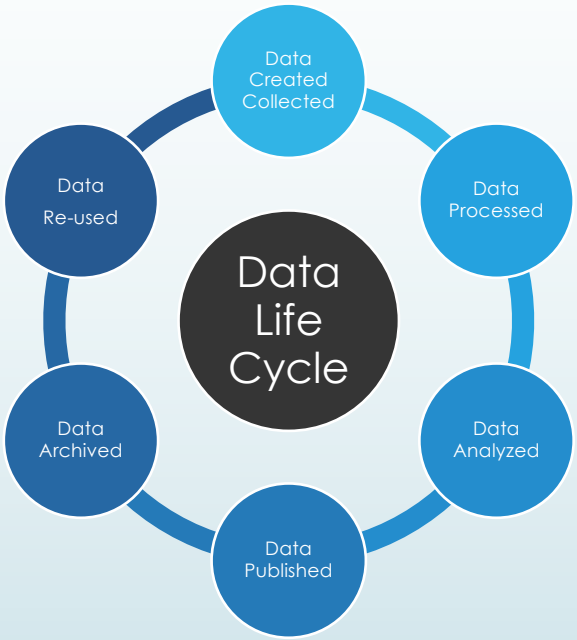
RDM – Reusing Data

- Re-run published analysis
- Add/incorporate into a new trial
- Meta-Analysis
- R-shiny apps

RDM – Reusing Data

- Other thoughts or ideas?

Data Life Cycle and your role as a researcher?
More than just collecting and analyzing the data



Contact

Michelle Edwards

edwardsm@uoguelph.ca

adc@uoguelph.ca

Next Agri-food Research Data Workshops

- Data Management Plans – April 19
- New ODRC Data Portal – April 26
- Documenting your Data! – May 3
- Introduction to R - May 10
- Documenting your data and processes with R Markdown** - May 17

Next Agri-food Research Data Workshops

- R Shiny** – May 31
- Introduction to Github – June 14
- Introduction to Linux – June 21

More topics? Please email adc@uoguelph.ca

ADC