Managing Your Research Data

File Management

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Based on slides by Qi Wang & Jing Su

Outline

Data Management **Principles**

- Research Data Life-cycle
- Data Management Checklist

Techniques to help organize your research data

- File Organization
- File Naming
- Version Control
- Metadata (ReadMe)
- Running Low on Storage Space?

Outline

Data Management **Principles**

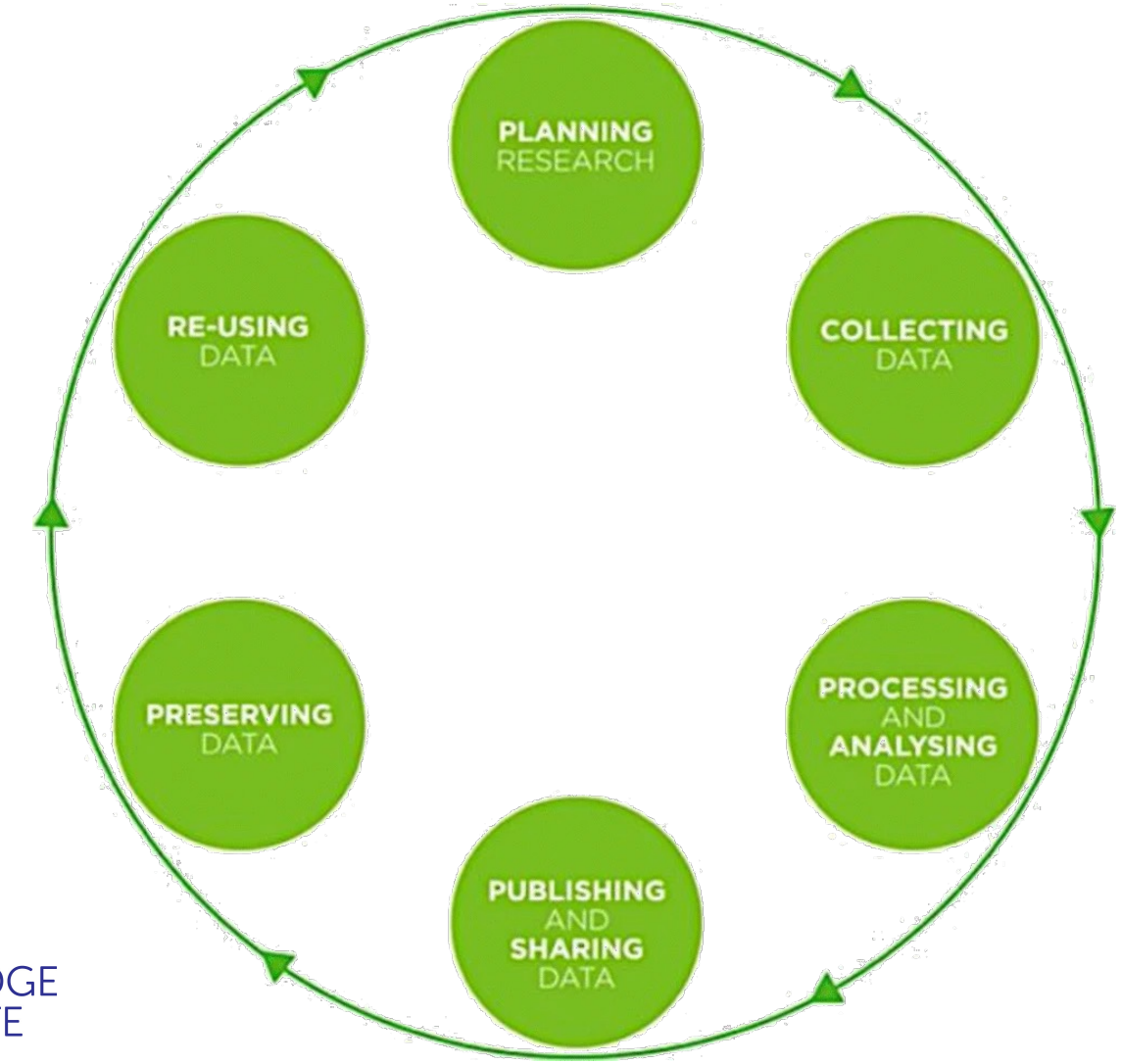
- Research Data Life-cycle
- Data Management Checklist

Techniques to help organize your research data

- File Organization
- File Naming
- Version Control
- Metadata (ReadMe)
- Running Low on Storage Space?

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Research Data Life Cycle



CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Data Management Checklist

- **What** types of data?
- **Who** will be responsible to collect and document the data?
- **How** to collect/document/store/back up/share data?

Data Management Checklist

- **What** types of data?
  - Experimental – raw data:
    - Tables of figures
    - Images
    - Sequencing data
    - Geographical
    - etc.
  - Processed data/Analysis results
  - Format and size for each

Data Types Recommended by UK Data Archive

Type of data	Recommended formats
Quantitative tabular data with extensive metadata. Variable labels, code labels, and defined missing values	Proprietary formats of statistical packages e.g. SPSS (.sav), Stata (.dta), .sas7bdat. Delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) containing metadata information. Some structured text or mark-up file containing metadata information, e.g. DDI XML file.
Quantitative tabular data with minimal metadata. A matrix of data with or without column	Comma-separated values (CSV) file (.csv). Tab-delimited file (.tab).
Geospatial data. Vector and raster data.	ESRI Shapefile (essential – .shp, .shx, .dbf, optional – .prj, .sbx, .sbn). Geo-referenced TIFF (.tif, .tfw). CAD data (.dwg). Tabular GIS attribute data.
Qualitative data. Textual.	eXtensible Mark-up Language (XML) text according to an appropriate Document Type Definition (DTD) (.xml). Rich Text Format (.rtf)/Plain text data, ASCII (.txt).
Digital image data.	TIFF version 6 uncompressed (.tif). Digital Imaging and Communications in Medicine (DICOM) (.dcm, .dcm30) – for CT/MRI data.
Digital audio data.	Free Lossless Audio Codec (FLAC) (.flac).
Digital video data.	MPEG-4 (.mp4). OGG video (.ogv, .ogg). motion JPEG 2000 (.mj2).
Documentation and scripts.	Rich Text Format (.rtf). PDF/A or PDF (.pdf). HTML (.html). OpenDocument Text (.odt). R Markdown files (.rmd) (with HTML version as well).

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

[Full table at the UK Data Archive website]

Data Types Recommended by UK Data Archive

Type of data	Recommended formats
Quantitative tabular data with extensive metadata. Variable labels, code labels, and defined missing values	Proprietary formats of statistical packages e.g. SPSS (.sav), Stata (.dta), .sas7bdat. Delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) containing metadata information. Some structured text or mark-up file containing metadata information, e.g. DDI XML file.
Quantitative tabular data with minimal metadata. A matrix of data with or without column	Comma-separated values (CSV) file (.csv). Tab-delimited file (.tab).
Geospatial data. Vector and raster data.	ESRI Shapefile (essential – .shp, .shx, .dbf, optional – .prj, .sbx, .sbn). Geo-referenced TIFF (.tif, .tfw). CAD data (.dwg). Tabular GIS attribute data.
Qualitative data. Textual.	eXtensible Mark-up Language (XML) text according to an appropriate Document Type Definition (DTD) (.xml). Rich Text Format (.rtf)/Plain text data, ASCII (.txt)
Digital image data.	TIFF version 6 uncompressed (.tif). Digital Imaging and Communications in Medicine (DICOM) (.dcm, .dcm30) – for CT/MRI data.
Digital audio data.	Free Lossless Audio Codec (FLAC) (.flac).
Digital video data.	MPEG-4 (.mp4). OGG video (.ogv, .ogg). motion JPEG 2000 (.mj2).
Documentation and scripts.	Rich Text Format (.rtf). PDF/A or PDF (.pdf). HTML (.html). OpenDocument Text (.odt). R Markdown files (.rmd) (with HTML version as well).

Open and non-proprietry

Information loss during conversion

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

[Full table at the UK Data Archive website]

Data Management Checklist

- **What** types of data?
  - Experimental – raw data:
    - Tables of figures
    - Images
    - Sequencing data
    - Geographical
    - etc.
  - Processed data/Analysis results
  - Format and **size for each**

Data Management Checklist

- **Who** will be responsible to collect and document the data?
  - Roles and responsibilities
  - Legal and ethical obligations and rights

Data Management Checklist

- **How** to collect/**document**/**store**/back up/share data?
  - Reproducibility & re-usability
  - Restrictions:
    - Ethical obligations
    - Legal obligations - privacy and data processing laws
    - Copyright/Intellectual property

Data Management Checklist

- **How** to collect/**document**/**store**/back up/share data?
  - Reproducibility & re-usability
  - Restrictions:
    - Ethical obligations
    - Legal obligations - privacy and data processing laws
    - Copyright/Intellectual property
  - Consider an Electronic Lab Notebook

Choosing an Electronic Lab Notebook

- Cost? One-time or subscription(monthly/yearly)?
- Access control. Other Users? Collaborators?
- Types of information to record and storage space required
- Any specialized functionality you require?
- Protection for sensitive data
- What happens if someone leaves the lab?
- What happens when you stop using this ELN?

Choosing an Electronic Lab Notebook

Further Reading:

- Kwok, Roberta. 2018. How to pick an electronic laboratory notebook. Nature 560 (7717): 269-270
- The Electronic Lab Notebook in 2023: A comprehensive guide
- Electronic Lab Notebook guide from Harvard Data Management

Data Management Checklist – 13 core questions

- **What** data will you collect or create?
- How will the data be **collected** or created?
- What **documentation** and **metadata** will accompany the data?
- How will you manage any **ethical issues**?
- How will you manage **copyright** and Intellectual Property Rights (IPR) issues?
- How will the data be **stored and backed up** during the research?
- How will you manage **access and security**?
- What is the long-term **preservation** plan for the dataset?
- Which data should be retained, **shared**, and/or preserved?
- How will you **share** the data?
- Are any **restrictions** on data sharing required?
- **Who** will be **responsible** for data management?
- **What resources will you require to deliver your plan? (people, time, hardware)**

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Full Checklist with guidance can be downloaded from the DCC website here

Outline

Data Management **Principles**

- Research Data Life-cycle
- Data Management Checklist

Techniques to help organize your research data

- File Organization
- File Naming
- Version Control
- Metadata (ReadMe)
- Running Low on Storage Space?


No one has
perfect data management habits,
but adopting even a few
goes a long way.

Outline

Data Management **Principles**

- Research Data Life-cycle
- Data Management Checklist

Techniques to help organize your research data

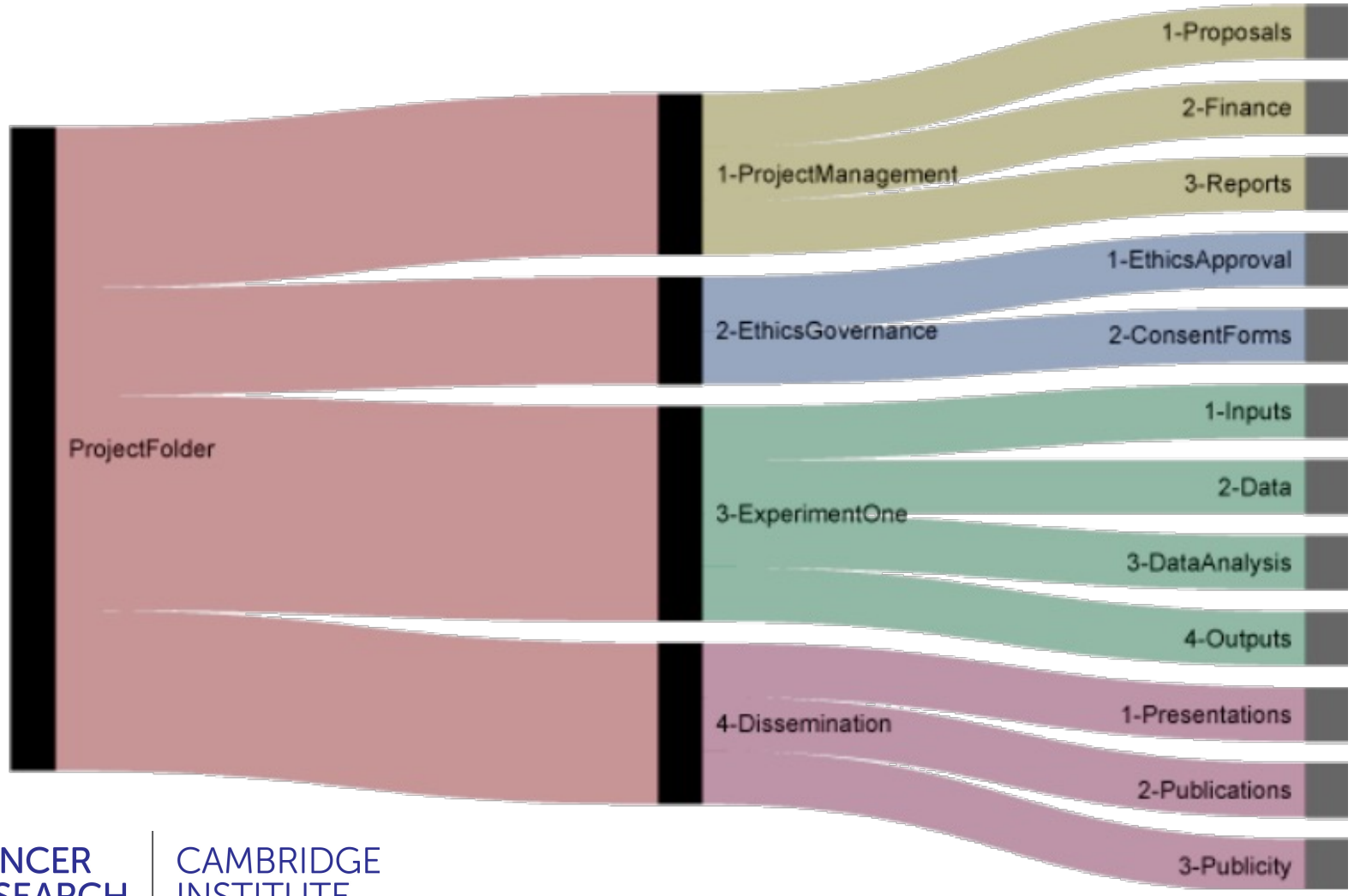
- **File Organization**
- File Naming
- Version Control
- Metadata (ReadMe)
- Running Low on Storage Space?

File Organization

Ways to organize electronic files

- Hierachical – files organized in folders and sub-folders
- Tag-based – each file is assigned one or more tags

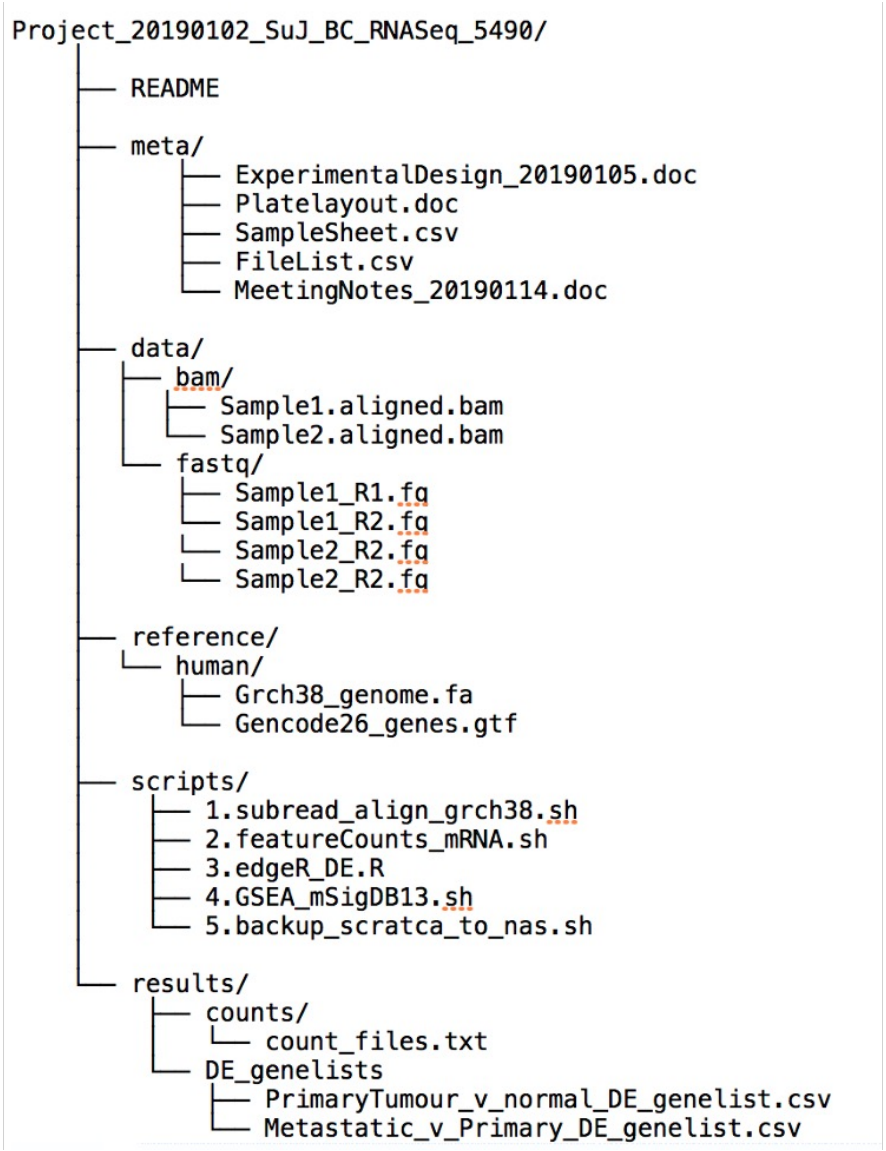
Hierachical folder structure



CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

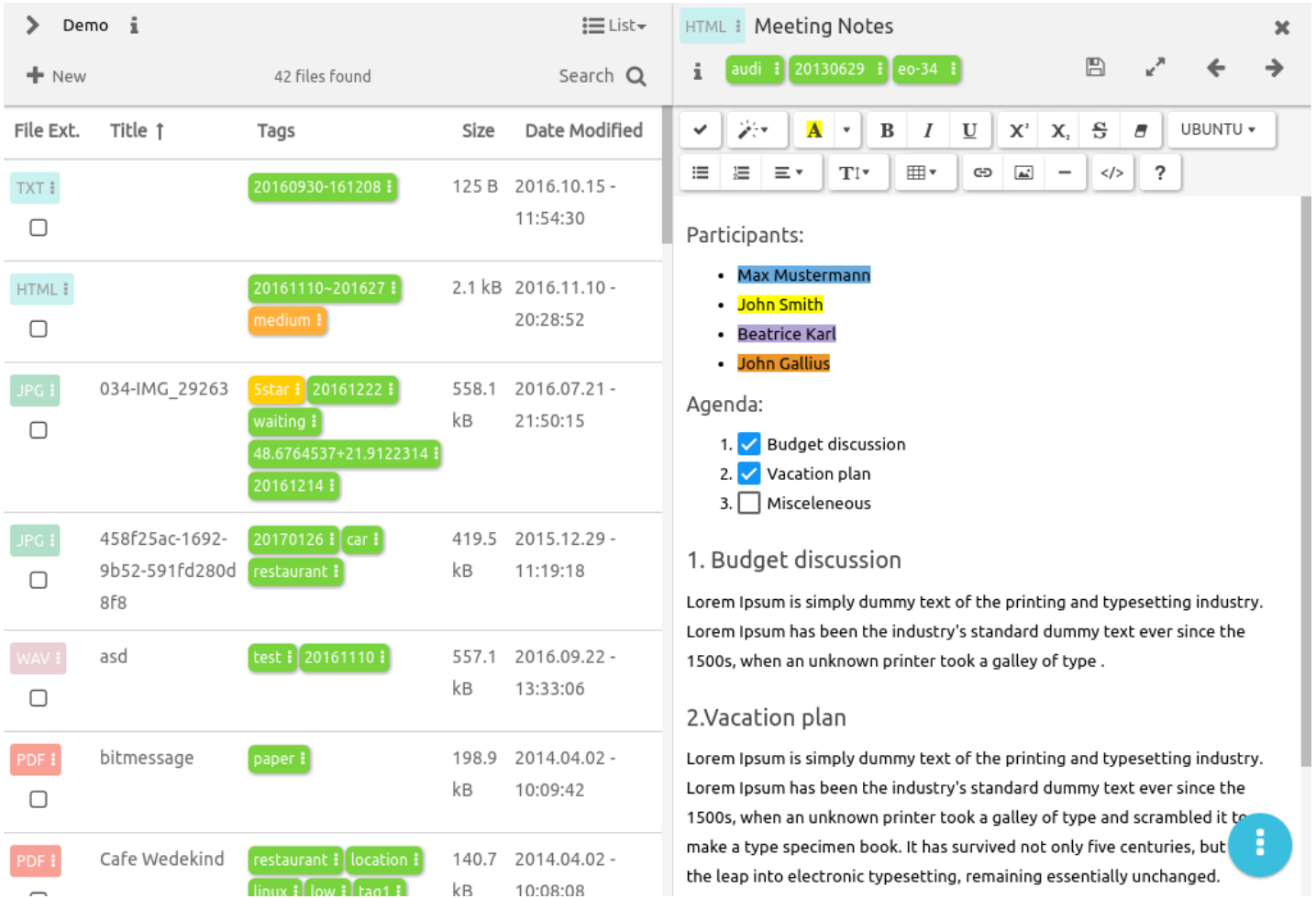
From: http://nikola.me/folder_structure.html

Hierachical folder structure



CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Tag based organization



CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

More about tagging: https://libguides.mit.edu/metadataTools

Outline

Data Management **Principles**

- Research Data Life-cycle
- Data Management Checklist

Techniques to help organize your research data

- File Organization
- **File Naming**
- Version Control
- Metadata (ReadMe)
- Running Low on Storage Space?

File Naming – the three Cs

Criteria: Can your collaborator (or you in 5 years time) identify the content of the file without opening it?

- **C**lear
  - Objective
    - ✗ my, current, latest, final
    - ✓JohnSmith, 20220418, version_1.0
  - Meaningful: “He”?
- **C**oncise
  - e.g. omit “the”, “and” etc.
- **C**onsistent
  - Have a defined naming convention e.g. [Date]_[Run]_[SampleType]_[SampleID]

File Naming – other tips

- Use underscores “_” to separate elements
  - Compare:
    - Averagetrendclustering20220814.png
    - Average_trend_clustering_20220814.png
- **Do not** use spaces or other special characters
- Use periods “.” only prior to the file extension
- Use leading zeros when using numbers in file names
  - Sample_01.png rather than Sample_1.png

File Naming – have a go...

This is a bad file name:

my Data @DryValley November 15 2010.v2.dat

How would you revise based on the principles discussed?

Type your proposal in the zoom chat window.

File Naming – have a go…

This is a bad file name:

my Data @DryValley November 15 2010.v2.dat

A better option:

DV_ICPOES_20101115_JDS_v02.dat

- DV: site code (Dry Valley)
- ICPOES: instrument name
- 20101115: date of data generation
- JDS: initials of the scientist
- v02: second version

Batching Renaming Tools

- Windows:
  - Ant Renamer
  - Bulk Rename Utility
- Mac:
  - Renamer6
  - Name Mangler
- Linux/Unix:
  - GNOME Commander
  - Use grep, sed and awk to search for and change file names

Outline

Data Management **Principles**

- Research Data Life-cycle
- Data Management Checklist

Techniques to help organize your research data

- File Organization
- File Naming
- **Version Control**
- Metadata (ReadMe)
- Running Low on Storage Space?

Version Control

Why?

- Track changes
- Enable reverting to earlier version

How?

- File naming (manually)
  - Date - Template_soil _testing_20120319.xlsx
  - Author's name - Template_soil _testing_by_AS.xlsx
  - Version number - Template_soil _testing_v03_02.xlsx
    - v01, v02 for major edit; v01_0, v01_1, v01_2 for minor edit
- Version control tools (automatic)
  - Wet lab
    - Electronic Lab Notebooks(ELN)
    - Laboratory Information Management System(LIMS)
  - Dry lab
    - Git (GitHub/GitLab)
    - Subversion

Version Control

VERSION CONTROL TABLE FOR A DATA FILE	
Title:	Vision screening tests in Essex nurseries
File Name:	VisionScreenResults_00_05
Description:	Results data of 120 Vision Screen Tests carried out in 5 nurseries in Essex during June 2007
Created By:	Chris Wilkinson
Maintained By:	Sally Watsley
Created:	04/07/ 2007
Last Modified:	25/11/ 2007
Based on:	VisionScreenDatabaseDesign_02_00

VERSION	RESPONSIBLE	NOTES	LAST AMENDED
00_05	Sally Watsley	Version 00_03 and 00_04 compared and merged by SW	25/11/2007
00_04	Vani Yussu	Entries checked by VY, independent from SK	17/10/2007
00_03	Steve Knight	Entries checked by SK	29/07/2007
00_02	Karin Mills	Test results 81-120 entered	05/07/2007
00_01	Karin Mills	Test results 1-80 entered	04/07/2007

Outline

Data Management **Principles**

- Research Data Life-cycle
- Data Management Checklist

Techniques to help organize your research data

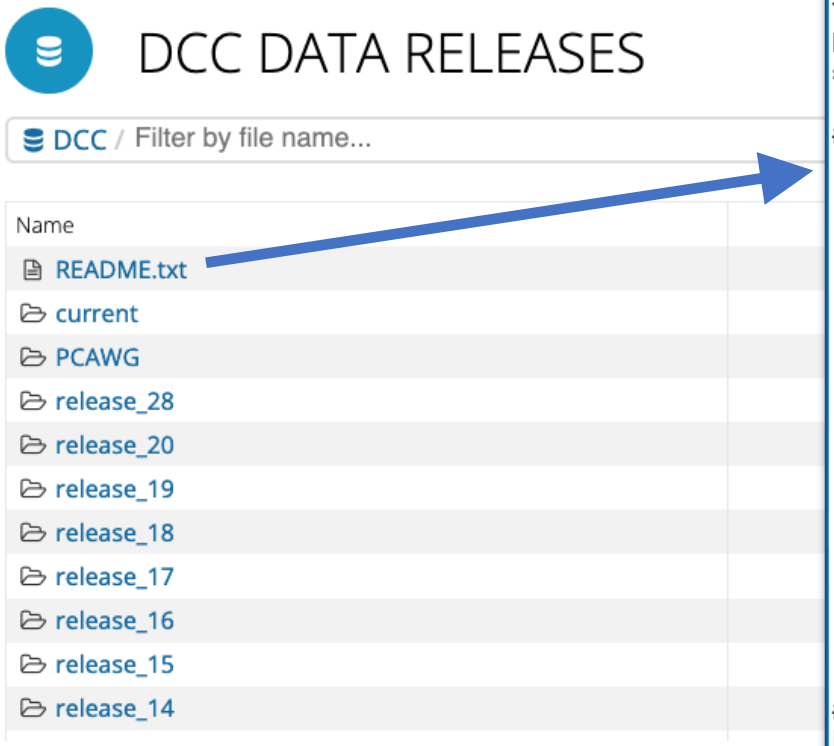
- File Organization
- File Naming
- Version Control
- **Metadata (ReadMe)**
- Running Low on Storage Space?

Metadata

- What is metadata?
  - Description that helps someone else understand the contents and organization of your files in your absence
- What should metadata include?
  - What?
  - Who?
  - Where & When?
  - How?

@ Project level @ Data level @ File level

Metadata - @Project level




```
# ICGC – DCC Data Releases
These are the DCC Data Releases of the International Cancer Genome Consortium (ICGC).
Release 28 also contains PCAWG mutation data. Please see below for more information on the
**PCAWG publication policy and embargo status.**

## Current DCC Data Releases
|Directory   |Contents              |Release Date|
|------------|----------------------|------------|
|Release_28  |DCC Data Release 28   |03/27/2019  |
|Release_27  |DCC Data Release 27   |04/30/2018  |
|Release_26  |DCC Data Release 26   |12/08/2017  |
|Release_25  |DCC Data Release 25   |06/08/2017  |
|Release_24  |DCC Data Release 24   |05/17/2017  |
|Release_23  |DCC Data Release 23   |12/07/2016  |
|Release_22  |DCC Data Release 22   |08/23/2016  |
|Release_21  |DCC Data Release 21   |05/16/2016  |
|Release_20  |DCC Data Release 20   |11/27/2015  |
|Release_19  |DCC Data Release 19   |06/16/2015  |
|Release_18  |DCC Data Release 18   |01/21/2015  |
|Release_17  |DCC Data Release 17   |09/12/2014  |
|Release_16  |DCC Data Release 16   |05/15/2014  |
|Release_15.1|DCC Data Release 15.1|02/12/2014  |
|Release_14  |DCC Data Release 14   |09/26/2013  |

## Legacy DCC Data Releases
For downloading data from previous releases before Release 14, please go to [Legacy DCC
Data Releases](https://dcc.icgc.org/releases/legacy_data_releases).

## ICGC Publication and Embargo Policy
If you plan to publish using data obtained from this portal please read the [ICGC
Publication Policy](https://daco.icgc.org/assets/site/files/
ICGC%20November%2015%202011%20Updates%20to%20Section%20E.3.pdf).
ICGC Publication guidelines and the current embargo status of each ICGC member project is
available at http://docs.icgc.org/portal/publication/#current-moratorium-status-for-icgc-
```

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

https://dcc.icgc.org/releases

Metadata - @Data level

```
<EXPERIMENT_SET>
   <EXPERIMENT alias="exp_mantis_religiosa">
      <TITLE>The 1KITE project: evolution of insects</TITLE>
      <STUDY_REF accession="SRP017801"/>
      <DESIGN>
         <DESIGN_DESCRIPTION/>
         <SAMPLE_DESCRIPTOR accession="SRS462875"/>
         <LIBRARY_DESCRIPTOR>
            <LIBRARY_NAME/>
            <LIBRARY_STRATEGY>RNA-Seq</LIBRARY_STRATEGY>
            <LIBRARY_SOURCE>TRANSCRIPTOMIC</LIBRARY_SOURCE>
            <LIBRARY_SELECTION>cDNA</LIBRARY_SELECTION>
            <LIBRARY_LAYOUT>
               <PAIRED NOMINAL_LENGTH="250" NOMINAL_SDEV="30"/>
            </LIBRARY_LAYOUT>
            <LIBRARY_CONSTRUCTION_PROTOCOL>Messenger RNA (mRNA) was isolated using the Dynabeads mRNA Purification Kit
            (Invitrogen, Carlsbad Ca. USA) and then sheared using divalent cations at 72*C. These cleaved RNA fragments
            were transcribed into first-strand cDNA using II Reverse Transcriptase (Invitrogen, Carlsbad Ca. USA) and N6
            primer (IDT). The second-strand cDNA was subsequently synthesized using RNase H (Invitrogen, Carlsbad Ca.
            USA) and DNA polymerase I (Invitrogen, Shanghai China). The double-stranded cDNA then underwent end-repair, a
            single `A? base addition, adapter ligati on, and size selection on anagarose gel (250 * 20 bp). At last, the
            product was indexed and PCR amplified to finalize the library prepration for the paired-end cDNA.</
            LIBRARY_CONSTRUCTION_PROTOCOL>
         </LIBRARY_DESCRIPTOR>
      </DESIGN>
      <PLATFORM>
         <ILLUMINA>
            <INSTRUMENT_MODEL>Illumina HiSeq 2000</INSTRUMENT_MODEL>
         </ILLUMINA>
      </PLATFORM>
      <EXPERIMENT_ATTRIBUTES>
         <EXPERIMENT_ATTRIBUTE>
            <TAG>library preparation date</TAG>
            <VALUE>2010-08</VALUE>
         </EXPERIMENT_ATTRIBUTE>
      </EXPERIMENT_ATTRIBUTES>
   </EXPERIMENT>
</EXPERIMENT_SET>
```


https://ena-docs.readthedocs.io/en/latest/submit/reads/programmatic.html

Metadata - @File level

File Descriptions

Open-access analyzed data:

clinical.[ICGC project code].tsv.gz: contains aggregated clinical donor, specimen and sample information
exp_array.[ICGC project code].tsv.gz: gene expression measured at the transcriptional level (mRNA) using array-based platforms
exp_seq.[ICGC project code].tsv.gz: gene expression measured at the transcriptional level (mRNA) using sequencing-based platforms

Details of the columns in Table S3:

1. sample_id
   tumor sample id (aliquot id)
2. ttype
   Tumor type name
3. chr
   Chromosome number
4. position
   Chromosome position
5. ref
   Reference allele
6. alt
   Alternate allele

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Metadata

Further reading on metadata and README files from Cornell University

Metadata – avoiding file corruption

- When copying, moving, downloading or uploading files, it is possible that the file may be corrupted or truncated.
- We can check this using an MD5 checksum

file name	md5sum
PCAWG16.consensus.virus.genus.normal.2out3.v3.icgc.controlled.tsv.gz	854b6a4dce3b46891c8cc4afc65a40d3
PCAWG16.consensus.virus.genus.normal.3out3.v3.icgc.controlled.tsv.gz	82f20aa61129522672fb8e1d7036cdfc
PCAWG16.consensus.virus.genus.tumour.2out3.v3.icgc.controlled.tsv.gz	1787e28e61651b19701cfbb9c108b908
PCAWG16.consensus.virus.genus.tumour.3out3.v3.icgc.controlled.tsv.gz	054200b756d059fc435c6f39ae9646b3
PCAWG16.consensus.virus.genus.normal.2out3.v3.tcga.controlled.tsv.gz	bba31c95dad98dc3b796c6937969a4e7
PCAWG16.consensus.virus.genus.normal.3out3.v3.tcga.controlled.tsv.gz	af0d91d2be2263f68c40e10a7780aced

- MD5 checksums are like “fingerprints” for files
- Any alterations to the file will cause the MD5 checksum to change

Outline

Data Management **Principles**

- Research Data Life-cycle
- Data Management Checklist

Techniques to help organize your research data

- File Organization
- File Naming
- Version Control
- Metadata (ReadMe)
- **Running Low on Storage Space?**

Running out of Space – Do we need all the files?

Five steps to decide what data to keep:

1. Identify purposes that the data could fulfil
2. Identify data that must be kept
3. Identify data that should be kept
4. Weigh up the costs £££
5. Complete the data appraisal

For more details:

https://www.dcc.ac.uk/guidance/how-guides/five-steps-decide-what-data-keep

Summary

- Principle: Can someone else (as well as yourself years from now) understand the contents and organization of your files in your absence.
- Data Management Checklist
  - What?
  - Who?
  - How?
- File Structure & File Name (3C)
  - Metadata (ReadMe) @Project-level @Data-level @File-level
- Keep track of changes with Version Control
- Avoid pitfalls in data transfer using md5sum check