File and Data Management

Based on Slides by Jing Su

Introduction

- Why file management of your research data is important

- Specific techniques for organizing your research data,
  - File structures
  - File naming
  - Version control
  - Storage

- Focuses on research data, also applies to other types of files

Research Data Lifecycle

Data Management Checklist 1/2

- ☑ Are you using standardised and consistent procedures to collect, process, check, validate and verify data?
- ☑ Are your structured data self-explanatory in terms of variable names, codes and abbreviations used?
- ☑ Which descriptions and contextual documentation can explain what your data mean, how they were collected and the methods used to create them?
- ☑ How will you label and organise data, records and files?
- ☑ Will you apply consistency in how data are catalogued, transcribed and organised, e.g. standard templates or input forms?
- ☑ Which data formats will you use? Do formats and software enable sharing and long-term validity of data, such as non-proprietary software and software based on open standards?
- ☑ When converting data across formats, do you check that no data or internal metadata have been lost or changed?
- ☑ Are your digital and non-digital data, and any copies, held in a safe and secure location?
- ☑ Do you need to securely store personal or sensitive data?

Data Management Checklist 2/2

- ☑ If data are collected with mobile devices, how will you transfer and store the data?
- ☑ If data are held in various places, how will you keep track of versions?
- ☑ Are your files backed up sufficiently and regularly and are back-ups stored safely?
- ☑ Do you know what the master version of your data files is?
- ☑ Do your data contain confidential or sensitive information? If so, have you discussed data sharing with the respondents from whom you collected the data?
- ☑ Are you gaining (written) consent from respondents to share data beyond your research?
- ☑ Do you need to anonymise data, e.g. to remove identifying information or personal data, during research or in preparation for sharing?
- ☑ Have you established who owns the copyright of your data? Might there be joint copyright?
- ☑ Who has access to which data during and after research? Are various access regulations needed?
- ☑ Who is responsible for which part of data management?
- ☑ Do you need extra resources to manage data, such as people, time or hardware?

Data Management Checklist

- **What** types of data and for how long?

  Five steps to decide what data to keep

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

- **Who** will be responsible to collect and document the data?

  Roles and responsibilities. Legal and ethical obligations and right. Plan and consent to share.

- **How** to document different types of data?

  Study-level, Data-level, and Metadata

  Wet lab: Electronic Lab Notebook (ELN)

  Computational: large size sequencing data, consortium data (TCGA, ICGC)

Some current ELN products

	Suitability	Platform	Storage	Comments
Benchling	Individual, Group	Browser	Vendor cloud	Free, user-friendly, self-contained, Molecular Biology bias.
Biovia	Group, Department	Macintosh, Windows	Vendor cloud or local server	Basic but robust feature set and workflow, strong in compliance, deployed campus-wide at some institutions.
Docollab	Individual, Group	Browser	Vendor cloud	Basic feature set with simple, modern interface.
e-Notebook	Group, Department	Windows	Local server	Complex, Chemistry/Pharma bias.
e-Workbook	Group, Department	Browser	Vendor cloud	Strong inventory management, Chemistry/Pharma bias.
eLabFTW	Group, Department	Browser	Local server	Free, Open Source, requires local server (Docker containers recommended). Community-driven development, sponsored by Institut Curie.
eLABJournal	Individual, Group, Department	Browser	Vendor cloud, private cloud, or local server	Comprehensive product with strong inventory management integration
Findings	Individual	Macintosh, iOS	Local HD, Dropbox sync	Simple, attractive interface, good synchronisation with Apple devices.
Hivebench	Individual, Group, Department	Browser, Macintosh, iOS	Vendor cloud or local server	New product - pilot deployment in one Cambridge research group has been a positive experience.
LabArchives	Individual, Group	Browser, iOS	Vendor cloud	Comprehensive features including Graphpad Prism integration. [Trial comments (Cambridge only)]

Formats: Data type and sources

File formats currently recommended by UK Data Archive for long term preservation for research data

FILE FORMATS CURRENTLY RECOMMENDED BY THE UK DATA ARCHIVE FOR LONG-TERM PRESERVATION OF RESEARCH DATA

TYPE OF DATA	RECOMMENDED FILE FORMATS FOR SHARING, RE-USE AND PRESERVATION
Quantitative tabular data with extensive metadata a dataset with variable labels, code labels, and defined missing values, in addition to the matrix of data	SPSS portable format (.por) 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 headings or variable names, but no other metadata or labelling	comma-separated values (CSV) file (.csv) tab-delimited file (.tab) including delimited text of given character set with SQL data definition statements where appropriate
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) or schema (.xml) Rich Text Format (.rtf) plain text data, ASCII (.txt)
Digital image data	TIFF version 6 uncompressed (.tif)
Digital audio data	Free Lossless Audio Codec (FLAC) (.flac)
Digital video data	MPEG-4 (.mp4) motion JPEG 2000 (.jp2)
Documentation	Rich Text Format (.rtf) PDF/A or PDF (.pdf) OpenDocument Text (.odt)

File Naming Conventions

- Make file names unique
- Include most important identifying information of the project:
  - ✓ project name
  - ✓ acronym, or research data name
  - ✓ study title
  - ✓ location information
  - ✓ researcher initials
  - ✓ date (consistently formatted, e.g. YYYYMMDD)
  - ✓ version
- Use underscores to separate elements; avoid special characters, spaces and periods.
- Use leading zeros when incorporating numbers to enable sorting (a sequence of 1-100 should be numbered 001-100).
- File names should be short enough to be readable, while still conveying enough pertinent information

File Naming Conventions Examples

- **The Good: DryValleySoil_ICPOES_20101115_JDSv2.dat**
  - DryValleySoil, project name
  - ICPOES, instrument name
  - 20101115 date of sample created
  - JDS, initials of the scientist
  - V2, second version

- **The Bad: my Data @DryValley November 15** 2010.v2.dat

- **The Ugly:**

  Can you understand/use these data files? Would anyone 5 years from now?

  - SrvMthdDraft.doc
  - SrvMthdFinal.doc
  - SrvMthdLastOne.doc
  - SrvMthdRealVersion.doc

  Use content-or descriptive information

Batching Renaming Tools

- Windows:
  - Ant Renamer: http://www.antp.be/software/renamer
  - Bulk Rename Utility: http://www.bulkrenameutility.co.uk/
  - PSRenamer: http://www.powersurgepub.com/products/psrenamer.html
- Mac:
  - PSRenamer: http://www.powersurgepub.com/products/psrenamer.html
  - Renamer4Mac : http://renamer4mac.com/
  - Name Mangler: http://manytricks.com/namemangler/
- Linux/Unix:
  - GNOME Commander: http://www.nongnu.org/gcmd/
  - PSRenamer: http://www.powersurgepub.com/products/psrenamer.html
  - Use *grep, sed and awk* to search for and change

Version Control

Aim: Keep raw data untouched and reverse to earlier version

- Save an untouched copy of the raw data, work on save untouched copy
- Use a file naming convention (like v001, v002 or v1_0, v1_2, v2_0
- Use a directory structure naming convention that includes version information
- Date can be part of the file name, e.g.

  2012-02-27_Template_soil _testing.xlsx
- Append the author's name to the file name, e.g.

  Template_soil _testing_modified_by_AH.xlsx
- Add a version number after reach major edit, e.g.

  Template_soil _testing_v03.xlsx
- Directory top-level folders should include the project title, unique identifier, and date (year), but the files themselves should be well-described independent of the directory structure.
- Version control tools:
  - Wet lab: Electronic Lab Notebooks/Box/LIMS
  - Dry lab: Git (GitHub/GitLab), Subversion (SVN)

Version Control Example

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

Folder Structure

- Methods of organising electronic material
  - **Hierarchical**: Items organised in folders and sub-folders
  - **Tag-based**: Each item assigned one or more tags
  - **Hybrid** combination of hierarchical and tag-based

Folder Structure Examples – Tag-based

Demo

List

New

42 files found

Search

File Ext.	Title	Tags	Size	Date Modified
TXT		20160930-161208	125 B	2016.10.15 - 11:54:30
HTML		20161110~201627, medium	2.1 kB	2016.11.10 - 20:28:52
JPG	034-IMG_29263	5star, 20161222, waiting, 48.6764537+21.9122314, 20161214	558.1 kB	2016.07.21 - 21:50:15
JPG	458f25ac-1692-9b52-591fd280d8f8	20170126, car, restaurant	419.5 kB	2015.12.29 - 11:19:18
WAV	asd	test, 20161110	557.1 kB	2016.09.22 - 13:33:06
PDF	bitmessage	paper	198.9 kB	2014.04.02 - 10:09:42
PDF	Cafe Wedekind	restaurant, location, linux, low, tag1	140.7 kB	2014.04.02 - 10:08:08

HTML Meeting Notes

audi 20130629 eo-34

UBUNTU

Participants:

- Max Mustermann
- John Smith
- Beatrice Karl
- John Gallius

Agenda:

1. [x] Budget discussion
2. [x] Vacation plan
3. [ ] Misceleneous

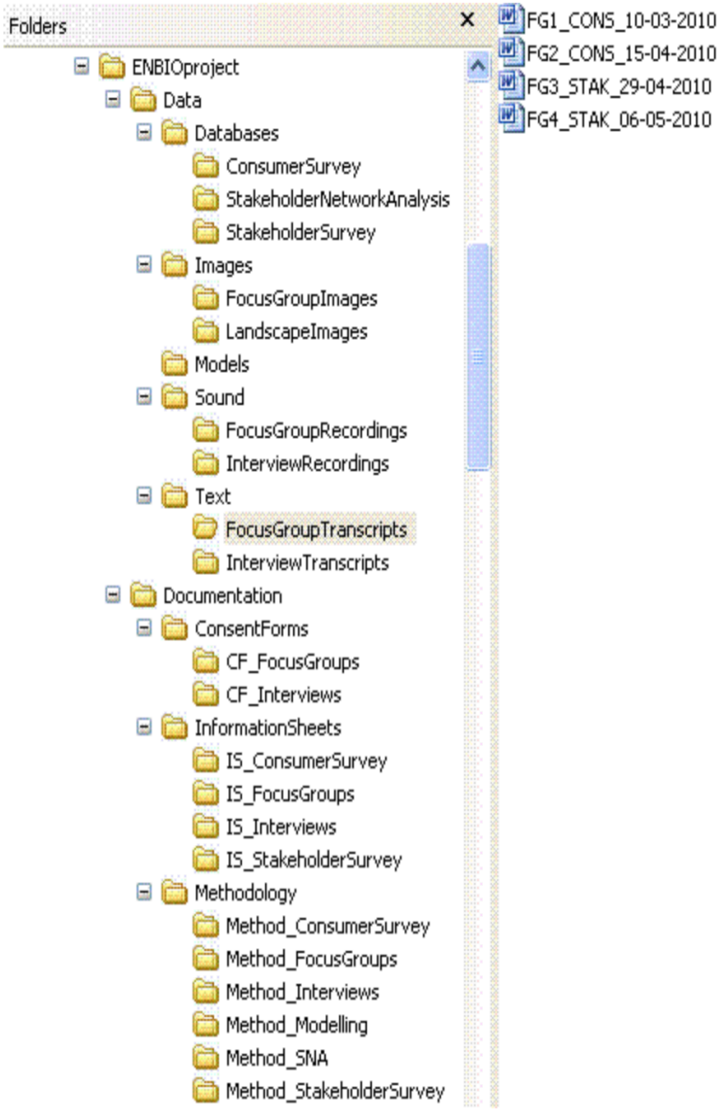
1. Budget discussion

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type .

2.Vacation plan

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but the leap into electronic typesetting, remaining essentially unchanged.

Folder Structure Examples – Hierarchical



```
Project_20190102_SuJ_BC_RNASeq_5490/
    │
    ├── README
    │
    ├── meta/
    │       ├── ExperimentalDesign_20190105.doc
    │       ├── Platelayout.doc
    │       ├── SampleSheet.csv
    │       ├── FileList.csv
    │       └── MeetingNotes_20190114.doc
    │
    ├── data/
    │   ├── bam/
    │   │   ├── Sample1.aligned.bam
    │   │   └── Sample2.aligned.bam
    │   └── fastq/
    │       ├── Sample1_R1.fq
    │       └── Sample1_R2.fq
    │       └── Sample2_R2.fq
    │       └── Sample2_R2.fq
    │
    ├── reference/
    │   └── human/
    │       ├── Grch38_genome.fa
    │       └── Gencode26_genes.gtf
    │
    ├── scripts/
    │     ├── 1.subread_align_grch38.sh
    │     ├── 2.featureCounts_mRNA.sh
    │     ├── 3.edgeR_DE.R
    │     ├── 4.GSEA_mSigDB13.sh
    │     └── 5.backup_scratca_to_nas.sh
    │
    └── results/
          ├── counts/
          │   └── count_files.txt
          └── DE_genelists
              ├── PrimaryTumour_v_normal_DE_genelist.csv
              └── Metastatic_v_Primary_DE_genelist.csv
```

Backups:
Project_20190212_SuJ_BC_RNASeq_5490_20190215
Project_20190212_SuJ_BC_RNASeq_5490_20190221

SpaceSniffer 1.3.0.1 - www.uderzo.it - [OSDisk (C:) (93.5GB Free Space) - Ready]

File Edit Windows Help

Filter Filter Ready

C:\ - 97.7GB

Users - 35.4GB

- 35.4GB

AppData 13.6GB

Documents 13.2GB

Desktop 3.8GB

Downloads 1.6GB

OneDrive - THE 1.3GB

Applications 1.3GB

Windows - 25.4GB

WinSxS - 11.6GB

System32 - 4.9GB

DriverStore 1.8GB

servicing - 1.7GB

LCU 1.5GB

Installer - 1.6GB

SysWOW64 - 1.1

MEMORY.DMP 1.3GB

Program Files (x86) - 13.4GB

Microsoft Office - 3.4GB

root 2.3GB

Updates 864.8MB

Office12

Esri - 2.1GB

Data Interope 2.1GB

ArcGIS - 2.9GB

Desktop10.5 2.9GB

Adobe - 1.7GB

Acrobat 10.0 1.3GB

Acrobat Reader DC

Common7 1.0GB

Google -

Chrome 437.8MB

Program Files - 11.6GB

ArcGIS - 6.7GB

Pro 3.8GB

Data Interoperability 2.8GB

QGIS 3.6 - 2.0GB

apps 1.2GB

bin 403.8MB

share 253.3MB

RStudio - 492.1MB

bin 319.3MB

Endpoin 198.2MB

AS OLI 230.9MB

R - 206.1

R-3.4.3

Audio 51.0MB

Common Files - 1.1GB

McAfee 696.2MB

microsoft sha 489.2MB

VLC

SynTP

Classif 5.2MB

pagefile.sys 5.2GB

hiberfil.sys 3.1GB

ProgramData - 1.

Package Cache -

Microsoft - 474.5

Windows Defen

Python27 - 1

Lib 579.0MB

ArcGIS10.5

Lib 335.8MB

QGIS-OSGeo4W 484.0MB

Master status: Scanning...