

File Management

Anne Pajon
Sergio Martínez Cuesta

Use descriptive and informative file names and directories



File names ... Best practices

Do not name all your files **data.xls** or **experiment.doc**

Include any information that will allow you to distinguish your files from one another

Project / experiment name / acronym

Type of data

Location / spatial coordinates

Conditions

Researcher name / initials

Version number

Date of experiment

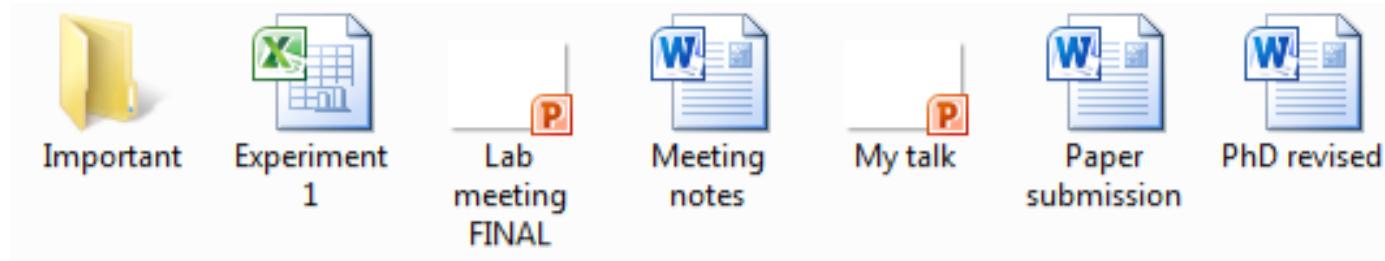
File names ... Best practices

Choose a consistent naming scheme and stick to it

Meaningful to you and your colleagues

Include in the directory a **README.txt** file that explains your naming format along with any abbreviations or codes you have used

Allow you to find files easily



File names ... Tips

Avoid **special characters** ~ ! @ # \$ % ^ & * () ` ; < > ? , [] { } ' " |

Use **short file names**

A good format for **dates** is **YYYY-MM-DD** or **YYMMDD**

All of your files will always stay in chronological order

Use **leading zeros** for clarity and to make sure files sort in sequential order

E.g. "001, 002 ... 010, 011 ..." instead of "1, 2, ...10, 11 ... "

File names ... Tips

Do not use spaces. Some softwares do not recognize file names with spaces.

e.g. data table.xls

Other options include:

Underscores, e.g. data_table.xls

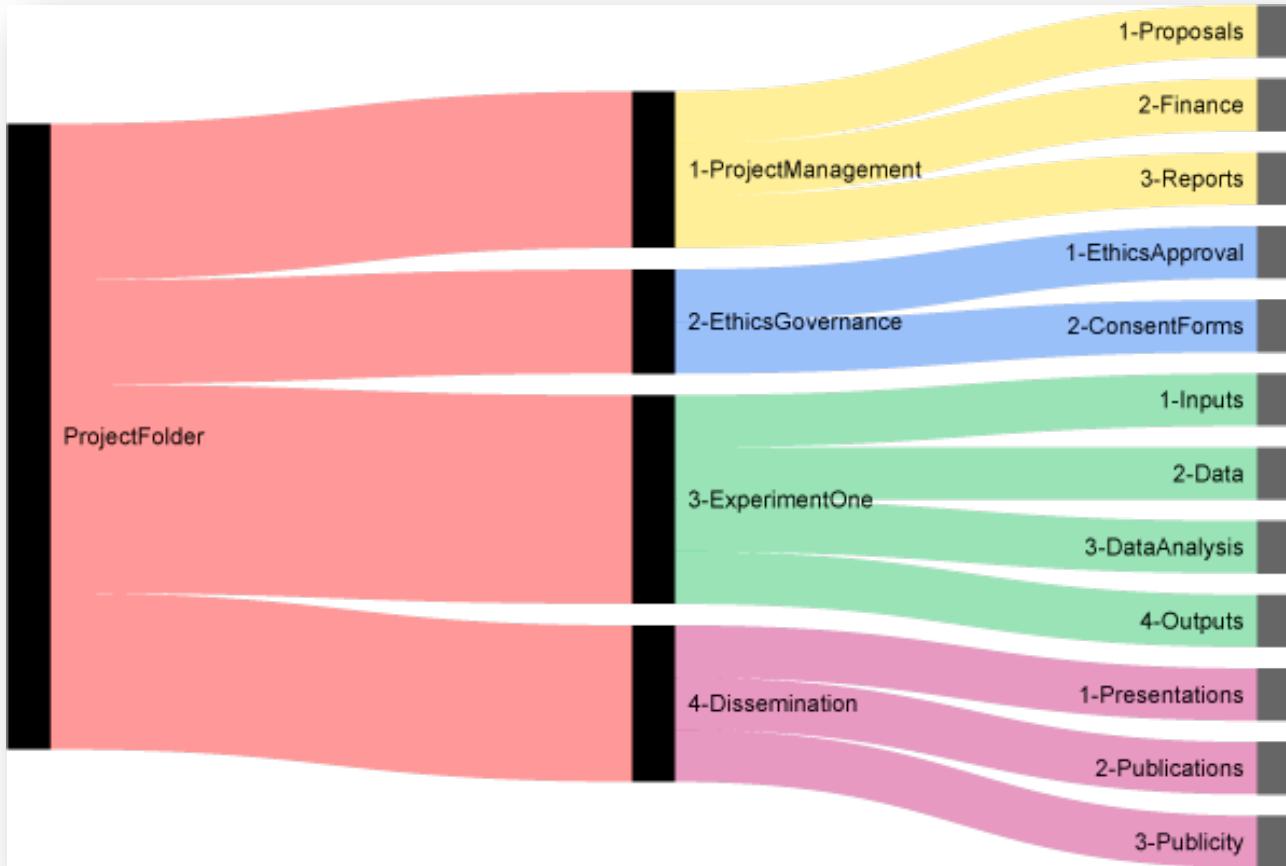
Dashes, e.g. data-table.xls

No separation, e.g. datatable.xls

Camel case, first letter of each section of text is capitalized, e.g. DataTable.xls

Keep an organised directory structure





Copyright: <http://www.vukovicnikola.info/folder-structure-for-research/>

**Choose file
formats that
will ensure long-
term access**



File formats ... Best practices

Save data in a **non-proprietary** (open) file format when possible

Usable on diverse platforms and by multiple applications

Export your data as tab separated file (.tsv)



Unencrypted and uncompressed

Common in your research community

Preferred formats

.tsv, comma separated file (.csv), .txt

Track different **versions** of your documents



File versioning

Versioning refers to saving new copies of your files when you make changes allowing you to reverse or roll back those changes or retrieve specific versions of your files later

Simple file versioning

Simple software options

Advanced software options

Simple file versioning

Manually save new versions when you make significant changes

Include a version number, e.g. "v01," "v02," or "v02.1" into file names

This works well if...

No need to keep lots of different versions

Only one person working on the files OR every collaborator knows what each version contains

Files are accessed from one location only

Simple software options: cloud services

Google Drive's word processing,
spreadsheet and presentation

Any time you edit files, new
versions are saved as you go

Version information includes who
was editing the file and when the
new version was created



OneDrive

Up to 1TB of available space for
University of Cambridge
members



Dropbox

Online and software to install
locally

Business option, £55-66 for
unlimited space



www.data.cam.ac.uk

Advanced software options: version control

Version control is the management of changes to documents, computer programs, and other collections of information.

Changes are usually identified by a number named the "revision number".

Each revision is associated with a **timestamp** and the **person** making the change.

Revisions can be compared, restored, and with some types of files, merged.

Systems like **Git** and **Subversion** can be used to do version control of files (e.g. computer code). Many people share projects on **GitHub**.



Course materials on GitHub

<https://github.com/>

Search for repository:

avoid-data-disaster

or

bioinformatics-core-shared-training

bioinformatics-core-shared-training / avoid-data-disaster

how to organise and keep your data tidy <http://bioinformatics-core-shared-training.github.io/avoid-data-disaster> — Edit

12 commits 2 branches 0 releases 1 contributor

Branch: master New pull request New file Find file HTTPS <https://github.com/bioinf/avoid-data-disaster> Download ZIP

markdunning add open refine project Latest commit 0d58eeb 9 minutes ago

File	Description	Age
images	add cruk banner	13 days ago
AvoidingDataDisastersFileMgmt.pdf	add file management talk	3 days ago
Presentation2.pptx	Add draft presentations	13 days ago
README.md	couple of tweaks	21 days ago
example1.google-refine.tar.gz	add open refine project	9 minutes ago
example1.xlsx	update example1 and add modified version	4 days ago
example1_modified.xlsx	update example1 and add modified version	4 days ago
example2.xlsx	Add draft presentations	13 days ago
principles.pdf	Add draft presentations	13 days ago
principles_nnotes.pdf	update Andy's slides	2 hours ago

README.md

Avoiding data disasters

Next GitHub course

<https://kirstiejane.github.io/friendly-github-intro>

www.bit.ly/GithubCam

Friday 13th January 2017 13:00 - 17:30

Spreadsheets and Databases

Anne Pajon
Sergio Martínez Cuesta

Spreadsheets

The good ...

Easy to **browse**, manually enter and edit data, and to **share** copies of files.

Fine control over **visual presentation**.

Very flexible structure.

Formulas make it a **living document**.

Built-in suite of helpers for charts, comments, spell checking ...

Relatively **easy to learn**.

The not so good ...

Lack data integrity. Data is not necessarily data.

Not good for **working with multiple datasets** and answering **detailed questions** about your data.

Do not scale. As spreadsheet size increases, performance suffers. Limits on cells (and spreadsheet) sizes.

Collaborating is hard. It is not easy to do version control.

Databases

System to store data (think of a huge library) and a **mechanism for searching** (think of a librarian).

The **Structured Query Language (SQL)** is a syntax for requesting things from the database (the language librarian speaks).

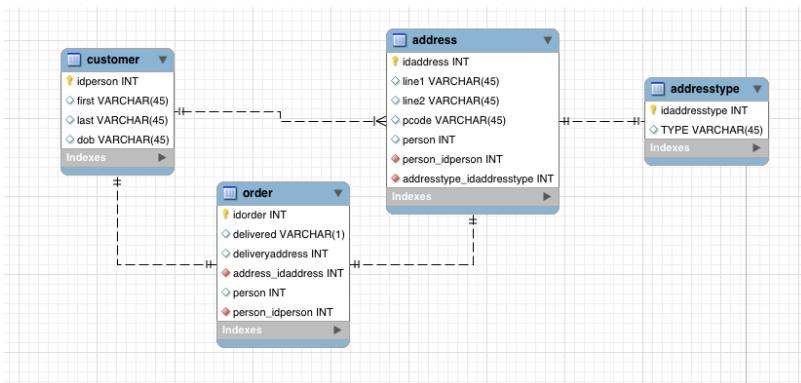
Relational databases consider **relationships between data**.



The database mantra

A database encourages/forces you to **store data logically**.

Every database consists of **tables and relationships** between them. Think of a table like a single spreadsheet. Just like in Excel, a table consists of **columns and rows**.



Columns define the structure of your data. Every column is given a **name** (like ‘Address’) and a defined **column type** (like ‘Integer,’ ‘Date’, ‘Date +Time’, or ‘Text’).

Rows contain the actual data in the table and have a value for every column. Once you establish the column structure, you can add in as many rows as you like.

The screenshot shows a database grid titled "Customers" with the following structure:

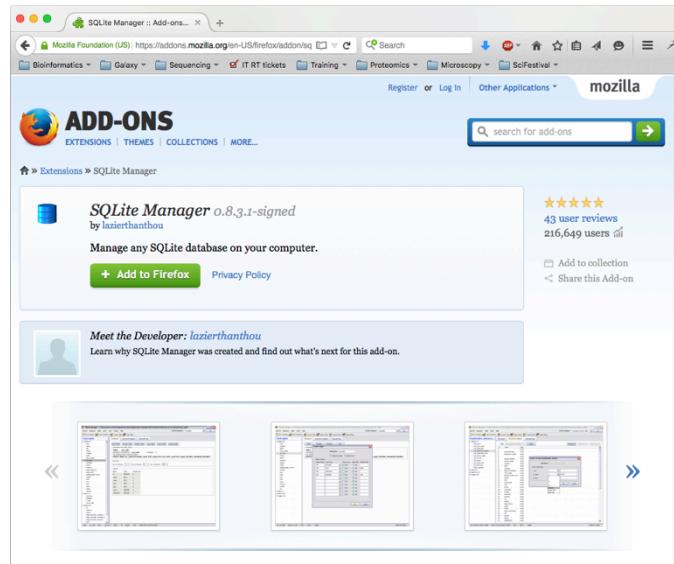
ID	First Name	Last Name	Street Address	City	State
1	Tracey	Ham	7 East Walker Dr.	Raleigh	NC
2	Lucinda	George	789 Brewer St.	Cary	NC
3	Jerrod	Smith	211 St. George Ave.	Raleigh	NC
4	Brett	Newkirk	47 Hillsborough St.	Raleigh	NC
5	Chloe	Jones	23 Solo Ln.	Raleigh	NC
6	Quinton	Boyd	4 Cypress Cr.	Durham	NC
7	Alex	Hinton	1011 Hodge Ln.	Cary	NC
8	Nisha	Hall	123 Huntingdon St.	Raleigh	NC
9	Hillary	Clayton	2516 Newman	Raleigh	NC
10	Kiara	Williams	9014 Miller Ln.	Durham	NC
11	Katy	Jones	456 Denver Rd.	Cary	NC
12	Beatrix	Joslin	85 North West St.	Raleigh	NC
13	Mariah	Allen	12 Jupe	Raleigh	NC
14	Jennifer	Hill	2100 Field Ave.	Raleigh	NC
15	Jaleel	Smith	123 Hill Top Drive	Garner	NC

Annotations highlight specific elements:

- An orange arrow points to the "Column" header in the table header.
- An orange box highlights the "Row" header in the table header.

Where to start?

SQLite is a good way to get started. You can install the “**SQLite Manager**” add-on for Firefox and start from within your browser.



University courses www.training.cam.ac.uk
search: relational database

Relational Database Design
<http://training.cam.ac.uk/event/1853176>
Monday 9th January 2017 9:00 – 13:00



Reference

Rosie Higman and Research data management team

www.data.cam.ac.uk

File management best practices



STANFORD UNIVERSITY LIBRARIES

<http://library.stanford.edu/research/data-management-services/data-best-practices>

Spreadsheets and Databases

<http://schoolofdata.org/2013/11/07/sql-databases-vs-excel/>

