GRADUATE CAMPUS
MODULE 1
Managing research data as a junior scientist

Jean-Blaise.Claivaz@unige.ch
Anouk.Santos@unige.ch

May 2022
FOREWORD

Slides from this workshop can be downloaded at:
https://www.unige.ch/researchdata/en/support/training/

Page:  Rendez-vous de l'info scientifique
Section: Other Workshops (at the very bottom of the page)
AGENDA

1. Welcome & presentation of the day
2. General introduction on Open Science
3. Research Data: definitions, organisation, description, formats & tools

------------------- Lunch break -------------------

4. Data repositories, copyrights and licences
5. Data Management Plan (DMP)
6. General discussion & feedback
7. Escape the Lab

Dice made by Dimi Kazak - https://www.flaticon.com
LET’S BREAK THE ICE!

• Fill out your Data scientist identity card

• Introduce yourself to your neighbor

• Introduce your neighbor
GOALS OF OPEN SCIENCE

1. **Transparency** in experimental methodology, observation, and collection of data.

2. **Public** availability and reusability of scientific data.

3. **Public** accessibility and transparency of scientific communication.

4. Using web-based tools to facilitate scientific collaboration.

**PROS AND CONS OF OPEN SCIENCE**

- **Reproducibility crisis**
  - Too many researches cannot be reproduced

- **Broader dissemination of scientific results**
  - Paywalls, licenses, poor formatting, proprietary formats... *empede* the sharing of knowledge

- **Good use of public money**

- **In favor of scoopers (?) and predators**

- **Bureaucracy**
“Misinformation is convenient, accessible, and ubiquitous and spreads to fill the vacuum left by credible information, which, conversely, is frequently inconvenient, inaccessible, and scarce.”

Thus

“Accessible open science can counter the infodemic.”

Childhood maltreatment and methylation of the glucocorticoid receptor gene NR3C1 in bipolar disorder

An investigation carried out at the request of the Dean of the Faculty of Medicine of the University of Geneva has concluded that one of the authors (Alain Malafosse) fabricated methylation data. …
OPEN SCIENCE TAXONOMY
MANAGE YOUR RESEARCH DATA
**WHAT IS RESEARCH DATA?**

Factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings.

ONE CONCEPT FOR A PLURALITY OF FORMS
CLASSIFICATION OF RESEARCH DATA

Observational

Simulation / models

Experimental

Derived / compiled

Reference
### What about your data?

<table>
<thead>
<tr>
<th></th>
<th>Observational</th>
<th>Experimental</th>
<th>Reference</th>
<th>Derived/compiled</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Digital</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Analog</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Set up good practices to manage research data
DATA LIFE CYCLE AND THE DMP

DATA LIFE CYCLE AND THE DMP

Creating data

Reusing data

Processing data

Analysing data

Preserving data

Giving access to data

Data Management Plan
IN A NUTSHELL...

https://www.youtube.com/watch?v=N2zK3sAtr-4
WHAT STRUCK YOU IN THE VIDEO?
MMMH. A FEW PROBLEMS…

You published in Science. I am requesting your data!

Surely, you saved your data?

Can I use your data?

I opened the data and couldn’t understand it. Is there any record of what these field names mean?

Can’t find the USB drive. Forgot to label the boxes.

I can’t read hexadecimal. I need the program.

This USB is my only copy of my data.

My co-author knows. His name is Sam Lee, he lives in China.

On a USB drive.
Let's have a "COFFEE BREAK"

1 hot beverage
1 pastry
1 fruit or a juice
Let’s talk about:

- Folder *architecture* and naming
- File naming
- File versioning

**Naming conventions**

Can’t find the USB drive. It is in a box... There are many boxes! I forgot to label the boxes!
YOUR COMPUTER DESKTOP

MOST FREQUENT SHORTCUTS

INTERNET BROWSER

SIDE PROJECTS

SIDE PROJECTS

TO DO LISTS

TO DO

TO REALLY DO

SUPPOSED TO HAVE DONE LAST WEEK

PIRATED MUSIC, MOVIES, COMICS
(MAIN USE FOR UNIVERSITY HIGH-SPEED INTERNET CONNECTION)

PAPERS YOU'VE BEEN MEANING TO READ FOR MONTHS

DRAFT WITH ADVISOR COMMENTS ON THEM

STUFF YOU DON'T KNOW WHAT TO DO WITH BUT DON'T WANT TO DELETE BECAUSE YOU'RE OBSESSIVE COMPULSIVE

INSTALL FILES FOR RANDOM PROGRAMS YOU ONLY USED ONCE AND NOW HAVE NO IDEA WHAT THEY DO.

E-MAIL ATTACHMENTS FROM YOUR PARENTS

DRUNKEN PICTURES FROM HAPPY HOUR

QUARANTINE SECTION

THESIS STUFF

??

??

??

Activity: Organize your files
FOLDERS AND FILES STRUCTURE

✓ Avoid overlapping categories

✓ Don’t let folders get too big

✓ Don’t let structures get too deep
PROTIP: NEVER LOOK IN SOMEONE ELSE’S DOCUMENTS FOLDER.
### File Naming: Exercice

<table>
<thead>
<tr>
<th>My passwords.doc</th>
<th>My data.xls</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMPORTANT.doc</td>
<td>My study.doc</td>
</tr>
<tr>
<td>My Thesis final final.doc</td>
<td>Doc.1.doc</td>
</tr>
<tr>
<td>Ma thèse version 12.doc</td>
<td>New doc.doc</td>
</tr>
<tr>
<td>Data 01/08/2016.xls</td>
<td>Int 1 (2).doc</td>
</tr>
<tr>
<td>Data 10 jan. 2016.xls</td>
<td>Interview 1.doc</td>
</tr>
</tbody>
</table>

1. Returning to your data in 1 year, will you recognize what these files contain?
2. What information needs to be in a file name to identify the content?
3. What would you change in these names?
# File Naming: Best Practices

<table>
<thead>
<tr>
<th>Best Practice</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order dates beginning with the <strong>year</strong> to enable sorting by date (kept 4 digit for year)</td>
<td>YYYY-MM-DD or YYYYMMDD</td>
</tr>
<tr>
<td>Limit the file name to <strong>32 characters</strong></td>
<td>32CharactersLooksExactlyLikeThis.csv</td>
</tr>
<tr>
<td>For sequential numbering, use <strong>leading zeros</strong> to allow for multi-digit versions</td>
<td><strong>NO</strong>: ProjID_2.csv ProjID_12.csv <strong>YES</strong>: ProjID_02.csv ProjID_12.csv</td>
</tr>
<tr>
<td>Don’t use <strong>special characters</strong> or spaces &amp; , * % # ; * ( ) ! @$ ^ ~ ' { } [ ] ? &lt; &gt; -</td>
<td><strong>NO</strong>: name&amp;<a href="mailto:date@location.doc">date@location.doc</a> <strong>YES</strong>: name_date_location.doc</td>
</tr>
<tr>
<td>Use only <strong>one period</strong> (for the file extension)</td>
<td><strong>NO</strong>: name.date.doc <strong>YES</strong>: name_date.doc</td>
</tr>
<tr>
<td>Avoid using <strong>generic data file names</strong> that may be ambiguous when moved</td>
<td><strong>NO</strong>: MyData.csv <strong>YES</strong>: ProjID_date.csv</td>
</tr>
</tbody>
</table>

Adapted from MIT Library, DM workshop, 2016
FILE VERSIONING: BEST PRACTICES

1. Avoid imprecise “final” labels

2. Save new versions using a consistent convention:
   • Major changes: v1; v2…
   • Minor changes: v1_1; v1_2

3. Document your convention

4. Consider your version control needs
EXAMPLE OF NAMING CONVENTION

TILS Document Naming Convention

Document naming for the TILS Division should follow this convention:

GDL_TILSDocNaming_V1_20090612.docx

A prefix shows the document type

The document title describes the content

The version number

The date in the format yyyymmdd

File names created from the TILS document naming convention are made up of four parts joined together with an underscore character ( _ ). There should not be any spaces in the file name.
1. Spending a little time *upfront*, can save a lot of time later on.

2. Be **realistic**: strike a balance between doing too much and too little.

3. There’s no single right way to do it; establish a **system that works for you**.

4. Think about **who your system needs to work for**: Just you? You and your lab group? Collaborators?

   “*Make a system. Share the system. Follow the system.*”

(MIT libraries)
"FINAL.doc"

FINAL.doc!

FINAL_rev.2.doc

FINAL_rev.6.COMMENTS.doc

FINAL_rev.8.comments5.
CORRECTIONS.doc

track changes

FINAL_rev.18.comments7.
corrections9.MORE.30.doc

FINAL_rev.22.comments49.
corrections.10.#@$%WHYDID
ICOMETOGRADECSCHOOL????.doc
Let’s talk about:

- **Metadata & documentation**: Documenting the format, tools, fields, etc.
- Documenting your variables, any abbreviations, etc… in a *readme* file
METADATA

Research data need to be documented at various levels:

- Project level
- File or database level
- Variable or item level

Aim is to make them:

- Understandable
- Reusable
- Findable

Cf. FAIR data 😊

→ Avoid ambiguity & misinterpretations

cea + from The Netherlands [CC BY 2.0]
3 MAIN TYPES OF METADATA

• **Descriptive metadata** enables discovery, identification, and selection of resources.
  o Title/unique ID, author, abstract, keywords, funder, date, ...

• **Administrative metadata** facilities the management of resources.
  o Rights statement, timestamp, file format, version, preservation, ...

• **Structural metadata** describes how different components of a set of data relate to one another
  o Data dictionary, variable list, database schema, taxonomy, abbreviations, …
Metadata enables connections to be made between published articles, researchers, datasets, computer programs, institutions, grants, funders, and more, eventually including things like shared facilities.

Phill Jones & Alice Meadows
Venice ‘time machine’ project suspended amid data row

Disagreements among international partners leave plans to digitize the Italian city's history in limbo.

Davide Castelvecchi

DOI: 10.1038/d41586-019-03240-w
ABOUT THE IMPORTANCE OF METADATA

2012 – Venice Time Machine project officially launched: Venice’s State Archive + Ca’ Foscari Univ. + EPFL (DHLAB) +…

2014 – Non-binding agreement signed. But … didn't specify the licensing that would regulate researchers’ use of the digitized data

2017 – At stake: 1,000 years of records in dynamic digital form: special high-speed scanners, thousands HD images per hour

2019 – Allegedly, the digitization of ~190,000 documents (8 TB) didn’t follow a common metadata policy: archival-science guidelines (require records of provenance for each document)

Now – … data collection has been paused, amid doubts on the usability of the data already collected!
File with metadata accompanying the dataset and providing all the necessary information about it

- For you, to stick with your standards
- For you, to understand your file names in 15 yrs (or 15 days!)
- For colleagues, to save files properly in your folders
- For other people, to use your files
**Readme.txt: Best Practices**

- Create one readme file **at the root** of the dataset
- **Have its title** reflect that the data user should read it (first)
- **Include anything you deem useful** to know for the potential users of your data
  - Descriptive, administrative and structural metadata
- Write your readme document as a **plain text file** (or PDF/A)
- Use a template to make sure you didn’t forget something important
- Additional readme files in subfolder may be created if judged useful
  - Format multiple readme files identically
  - Name the readme so that it is easily associated with the data file(s) it describes.
**DESCRIPTING VARIABLES**

In the Codebook or in the readme file

- Including references to standards
- Including units of measurement

---

<table>
<thead>
<tr>
<th>SP</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>REF</td>
<td>Reference number – This is my maintenance code for updates.</td>
</tr>
<tr>
<td>NEWREF</td>
<td>New Reference number – These are the refs numbers in the files REFS and SUMMARY.</td>
</tr>
<tr>
<td>FAMLAB</td>
<td>Family Label – An 8-character label for family.</td>
</tr>
<tr>
<td>GENLAB</td>
<td>Genus Label – An 8-character label for genus.</td>
</tr>
<tr>
<td>SPLAB</td>
<td>Species Label – An 8-character label for species.</td>
</tr>
<tr>
<td>COD</td>
<td>Species code – A 5-character code for the species.</td>
</tr>
<tr>
<td>DISPCAT</td>
<td>Disperser type category – BIRDS, MIXED, MAMMALS.</td>
</tr>
<tr>
<td>MEGAFANA</td>
<td>Whether the fruit species is associated with dispersal of seeds by megafauna.</td>
</tr>
<tr>
<td>AREA</td>
<td>Geographic area – Major geographic areas of the data sources.</td>
</tr>
<tr>
<td>HABIT</td>
<td>Growth habit – Tree, Shrub, Liana, Herb.</td>
</tr>
<tr>
<td>FRUIT</td>
<td>Fruit type – Type of fruit. Not completed yet. Needs revision.</td>
</tr>
<tr>
<td>COLOR</td>
<td>Fruit color – Black, Blue, Red, Orange, Yellow, Green, Brown, White</td>
</tr>
<tr>
<td>COMPCOLOR</td>
<td>Whether the fruit is monocolored, bicolored, or multicolored.</td>
</tr>
</tbody>
</table>

[https://doi.org/10.5061/dryad.9tb73/1](https://doi.org/10.5061/dryad.9tb73/1)
Structure your data

Use your discipline’s standards (if they exist!) to facilitate interoperability

- Each standard is clearly defined and documented
- Each dataset can use the terms / be organized consistently
- Easy to share data across institutions, applications and disciplines

To find standards: https://fairsharing.org/standards/
Piled Higher and Deeper by Jorge Cham

Don't worry, you don't have to start your code from scratch.

You can re-use the software that the previous person on the project wrote several years ago.

Are there instructions for how to use it?

I doubt it.

Is the code commented?

Not likely.

Where are the files?

Who knows.

This is going to be painful, isn't it?

Just a scratch.

Title: "Scratch" - originally published 3/12/2014
But surely, you saved your data?

- I did, I saved it on a USB drive

- Let’s talk about:
  - Data storage
  - Back-ups
  - Security and tools

... I will need that back when you are finished, this USB is my only copy of my data
Imagine, your laptop AND your external hard drive got stolen!

Credit: Peter Murray-Rust,
Storage & Backup

- Recoverable
- Copy
- Changing
- Short-term

Archive

- Accessible
- Moved
- Inactive
- Long-term
HAVE A SYSTEMATIC BACKUP SCHEME

• During the collection, data should be stored on local computers or institutional servers

• During analysis, a raw copy of the data should be kept and data should be backed-up to other locations:
  • Ideally: 2 on-site copies + 1 off-site copy

• After, data can be uploaded to a repository for long term preservation (and sharing, if possible)
GOOD PRACTICES FOR DATA STORAGE

→ Store 3 copies of your data:

1. The original
2. A copy kept on a local external device
3. A copy kept on an external device at a different location
NAS OF THE UNIGE

• Network-attached Storage connected to computer network
• Dedicated to research data
• Snapshots every 4 hours (can retrieve data from up to 6 wks)
• Backed up off-site (Campus Biotech)
• Free up to 50 GB
• CHF 75.- / TB / year

More info: https://catalogue-si.unige.ch/en/donnees
Cloud (EduCloud - Filr)

- Allows members of the UNIGE community with an ISIs account to access the NAS in cloud mode and to upload documents on it via the web, from most devices.
- You can share these documents with other members of the UNIGE community, as well as with people outside the university.
- You can synchronize the content of the NAS with a PC or a Mac outside the UNIGE (by installing the "Filr client" software).
- Storage volume = quota of your NAS.
- Recommended for research data.

https://educloud.unige.ch:8443/
CLOUD (OneDrive, SWITCHdrive)

OneDrive

- Access with your ISIs account
- Free up to 1 TB
- Data stored in Switzerland
- Allows to share files with external people managed by you
- Automatic synchronization with local copies

https://www.switch.ch/en/drive/

SWITCH

- Access with your SWITCH.edu account
- Free up to 50 GB
- Data stored in Switzerland
- Allows to share files/folders with people without an account via a public link (password optional)
- Automatic synchronization with local copies

https://www.switch.ch/en/drive/

https://plone.unige.ch/distic/pub/logiciels/onedrive/comment-installer-onedrive

Not recommended for research data storage! Nor personal/sensitive data
ALTERNATIVES OF LAB NOTEBOOKS...
Another **tool** to save & share data & information within your lab
SENDING FILES

SWITCHfilesender: send files via the SWITCH cloud

Some files are just too big to send in an e-mail. The best way round this is SWITCHfilesender. The service runs in the protected SWITCH cloud, and it’s free.

SWITCHfilesender is the best way to send large files up to 50 GB. The web-based service couldn’t be simpler:

1. Start SWITCHfilesender in your browser.
2. Log in (using your AAI or voucher login).
3. Upload your file to the protected SWITCH cloud.
4. Enter the recipient’s e-mail address.
5. Send the file.

The recipient then gets an e-mail containing the download link.

All university members can send external users a voucher for a single use of SWITCHfilesender. The voucher is valid for a maximum of 20 days and permits the holder to send one file.

Unlike comparable web-based services, SWITCHfilesender runs in the protected SWITCH cloud. Files are stored exclusively at SWITCH’s data centre in Switzerland and deleted after 20 days at the very latest. If you want to store files in the SWITCH Cloud with no time limit, SWITCHdrive is the service for you.
SPECIAL NOTE ABOUT PERSONAL DATA

• **Personal data:** “all information relating to an identified or identifiable person;”

• **Sensitive personal data:** “data on:
  1. religious, ideological, political or trade union-related views or activities,
  2. health, the intimate sphere or the racial origin,
  3. social security measures,
  4. administrative or criminal proceedings and sanctions;”

• **Confidential / «critical» data**

---

Processing of personal data

DATA SECURITY

During research, to ensure your data integrity, you need a good management of data access: who has access to what?

• Physical Access
• Digital access

Extra care and regulations when dealing personal data (eg. Keep it in Switzerland)

NAS académique
Educloud (filr)
SWITCHfilesender
Locally hosted ELN

Dropbox
Google Drive
Wetransfer
…
I can’t read hexadecimal.

You will need the program that created the hexadecimal file.

Yes I will. What is the name of the program. Do you have a copy of the program?

I do not use this program anymore because the company that made it went bankrupt.

Let’s talk about:

- Archiving
- Data conversion: open format for preservation
- Preserving the software alongside the data?
**BACK-UP VS ARCHIVE**

**Storage & Backup**
- Recoverable
- Changing
- Copy
- Short-term

**Archive**
- Accessible
- Inactive
- Moved
- Long-term
Tip 1

Use whenever open formats

• Day-to-day processing may use close formats

• But for **archival purposes**, maximize **accessibility** and **long-term value** by converting to open/portable formats that are not software dependent

• For instance, prefer:
  o `.csv` rather than `.xls`
  o `.svg` rather than `.ai`
  o PDF/A rather than PDF

• See Guidance from ETH, UK data service, Harvard
KEEP RAW DATA RAW

Tip 2

• Allows **new analysis** with old data and transparency

• When possible, use a **cryptographic hash** to ensure the dataset has not suffered any silent corruption and/or manipulation

• When appropriate, save the **software** along with the data
**Insurance policy against unforeseen system failures**
- Incremental
- Multiple snapshots
- Retained on short periods of time
- Not searchable

**Preserve information as required by regulations and institutional policies**
- Auditable
- Follows a life cycle
- Self-described
- Data integrity
Thank you for sending me a copy of your data on a USB drive

... I will need that back when you are finished, this USB is my only copy of my data

Let’s talk about:

- **Sharing**
  - Caveat about special data
- **Repositories & Licenses**
Requirements for Data Sharing

Your work was funded by NIH and published in Science. I am requesting your data!
ARCHIVING AND DISSEMINATION

- Discipline-specific (e.g. GenBank)
- Institutional (e.g. Yareta)
- Multi-disciplinary (e.g. Zenodo)

- Journal supplementary material service
- Departmental, project or personal web page
Find and Choose a Repository

Consult with colleagues in your discipline

Ask your librarians

Explore Re3data

Illustrations by Freepik Storyset
FAIR PRINCIPLES

- Findable
- Accessible
- Interoperable
- Reusable
Repository and FAIR data

Findable  Accessible  Interoperable  Reusable

Persistent  Access  Licenses
Identifier  Lock  Icon

doi  urn  ark

At least one blue icon

Presence of an icon license

Presence of a orange lock icon

Test the FAIR level of your data:
CREATIVE COMMONS LICENSES

Most open

This license is the "no copyright reserved" option in the Creative Commons toolkit - it effectively means relinquishing all copyright and similar rights that you hold in a work and dedicating those rights to the public domain.

This license lets you distribute, remix, tweak, and build upon the original work, even commercially, as long as you credit the original creation.

This license lets you remix, tweak, and build upon the original work, even commercially, as long as you credit the original creation and license your creations under the identical terms.

This licence allows for redistribution, commercial and non-commercial, as long as it is passed along unchanged and in while, with credit to the original work.

This license lets you remix, tweak and build upon the original work non-commercially. Your new works must be non-commercial and acknowledge the original work, but you don’t have to license your derivation work on the same terms.

This licence lets you remix, tweak and build upon the original work non-commercially as long as you credit the original work and license your new creations under the identical terms.

This licence is the most restrictive. It only allows to download the original work and share it with others as long as you credit the original work. You cannot change the original work in any way or use it commercially.

Least open

Some repositories impose a certain licence! (eg. Dryad = CC0)
INSTITUTIONAL REPOSITORY: YARETA

https://yareta.unige.ch

✓ Repository for all researchers of Geneva

✓ **Swiss-based** servers (original et copy)

✓ Compliant with **FAIR principles** & provides **DOI**

✓ Long-term preservation solution (you define the duration of retention of the data)

✓ **Free** up to 50 Go

  (above: 100 CHF / Tb / year of retention)
GENERIC REPOSITORY:

✓ Catch-all repository
✓ Operated by CERN

✓ Free access and deposit

• Also: https://sandbox.zenodo.org/
  o If you want to try making a dummy deposit
✓ Check for legal and regulatory requirements

✓ Anticipate your data needs (volume, sensitivity, formats,…) and select appropriate options accordingly

✓ Ensure your data can be reused (open formats, proper naming, add metadata…)

✓ UNIGE solutions: NAS, Yareta
DATA MANAGEMENT PLAN (DMP)
EXTENDED REQUIREMENT
## Requirements from SNF

### 2. Application data

<table>
<thead>
<tr>
<th>#</th>
<th>No./Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Basic data I</td>
</tr>
<tr>
<td>2.2</td>
<td>Basic data II</td>
</tr>
<tr>
<td>2.3</td>
<td>Use-inspired project</td>
</tr>
<tr>
<td>2.4</td>
<td>Re-submission</td>
</tr>
<tr>
<td>2.5</td>
<td>Link to other SNSF projects</td>
</tr>
<tr>
<td>2.6</td>
<td>Available or requested funds</td>
</tr>
<tr>
<td>2.7</td>
<td>University or research institution</td>
</tr>
<tr>
<td>2.8</td>
<td>Collaboration (national and international)</td>
</tr>
<tr>
<td>2.9</td>
<td>Information on additional funds</td>
</tr>
<tr>
<td>2.10</td>
<td>Requested funding</td>
</tr>
<tr>
<td>2.11</td>
<td>Data management plan (DMP)</td>
</tr>
<tr>
<td>2.12</td>
<td>Research requiring authorisation on own account</td>
</tr>
<tr>
<td>2.13</td>
<td>Children you are obliged to support</td>
</tr>
<tr>
<td>2.14</td>
<td>Diplomas / certificates / extension education</td>
</tr>
<tr>
<td>2.15</td>
<td>General remarks on the project</td>
</tr>
</tbody>
</table>

### Data collection and documentation

1. What data will you collect, observe, generate or reuse?
2. How will the data be collected, observed or generated?
3. What documentation and metadata will you provide with the data?

### Ethics, legal and security issues

1. How will ethical issues be addressed and handled?
2. How will data access and security be managed?
3. How will you handle copyright and Intellectual Property Rights issues?

### Data storage and preservation

1. How will your data be stored and backed-up during the research?
2. What is your data preservation plan?

### Data sharing and reuse

1. How and where will the data be shared?
2. Are there any necessary limitations to protect sensitive data?
3. All digital repositories I will choose are conform to the FAIR Data Principles.
4. I will choose digital repositories maintained by a non-profit organisation.
1. Data collection and Documentation

1.1 What data will you collect, observe, generate or reuse?

Type of data → recordings, images, texts...

Format → printed, pdf, xls, jpeg, etc.

Volume → Initial estimate
This project will generate three main types of raw data:

1. Images from transmitted-light microscopy
2. Images from confocal microscopy
3. Western blot data.

Measurements and quantification of the images will then be recorded in spreadsheets.

Images will be stored as .tif Data and spreadsheets will be stored as .csv Data and freetext documents will be stored as .txt.

Micrograph data is expected to total between 100GB and 1TB over the course of the project. Scanned images of western blots are expected to total around 1GB over the course of the project. Other derived data (measurements and quantifications) are not expected to exceed 10MB.
1. Data collection and Documentation

1.2 How will the data be collected, observed or generated?

- Data collection method
- Standards and best practices used
- Organization and naming of files and folders
Giemsastained squashed larval brains will be used for transmitted-light microscopy and of immunostained whole-mounted larval brains will be used for confocal microscopy.

All samples on which data are collected will be prepared according to published standard protocols in the field. All microscopes used for sample examination are serviced and recalibrated regularly. All Drosophila lines used in experiments are checked periodically for phenotypic markers. Drosophila are maintained in live culture according to standard methods in the field.

Files will be named according to a pre-agreed convention:

for images from transmitted-light microscopy (TLM) and from confocal microscopy (COM) and for quantifications (QUA), we will add the date (YYMMDD), the initials of the researcher (XY) and the version number V1, V2...:

TLM_YYMMDD_TLM_XY_V1.tif
COM_YYMMDD_TLM_XY_V1.tif
QUA_YYMMDD_TLM_XY_V1.csv
1. Data collection and Documentation

For the purpose of a future reuse:
- By humans / by machines
- By oneself / by others

Descriptive metadata

Readme.txt files
A README file will describe the directory hierarchy. The version of each software and library will be documented for each set of results in a text file e.g. INFO.txt.

The final datasets as deposited in the chosen data repositories will be accompanied by the relevant metadata documentation. Specifically:

- For our Microscopy image dataset, associated metadata will be compliant with OME-XML standard (https://docs.openmicroscopy.org/ome-model/5.5.7/ome-xml/index.html).
- For our Molecular Biology dataset, associated metadata will be compliant with DataCite standard (https://schema.datacite.org/).

This will ensure that our produced data can be analyzed together with other data types during the project and following project completion (eg from DMP Canva Generator, VitalIT).
2. Ethics, legal and security issues

2.1 How will ethical issues be addressed and handled?

Management of sensitive data

Data protection

Privacy

Permissions

www.digitalbevaring.dk
In our study there will be use of human samples.

Therefore, specific care will be taken to handle security and anonymization of sensitive data. Sensitive data transfers will be end-to-end encrypted and encryption keys will be managed only by authorized employees.

The human experimentation part of this grant has been approved by an ethics committee [please specify which].

All human data will be handled consistently with pre-signed formal consent agreements.

All personal data will be anonymized in such a way that it will be impossible to attribute data to specific persons.

(eg from [DMP Canva Generator](https://vitalit.com), VitalIT)
2. Ethics, legal and security issues

Risk Management

Data security

- Digital
- Physical

Access to data and premises
Our data is stored on the academic NAS managed by the UNIGE IT department (DiSTIC). Access to the data is limited to rights holders (central authentication). The head of the laboratory that owns this disk space manages access himself, with the possibility of registering additional users.
2. Ethics, legal and security issues

Legal issues:

- Who is the owner of the data?
- Which licence to apply?
- What restrictions apply to the reuse of third-party data?
Research data generated by UNIGE collaborators in the performance of their duties is the property of the institution. As the data is not subjected to a contract and will not be patented, it will be released as open data under Creative Commons CC0 license.
CHOOSE YOUR LICENSE

Data under license/contract?
- yes: Respect license/contract
  - Contact Unitec for help
- no

Contains personal/sensitive data?
- yes: Check anonymisation
  - before sharing!
- no

Potential commercial use?
- yes: Contact
  - Unitec for help
- no

Allow fully open reuse?
- yes
- no: Contact
  - Library

Publish data under **CC0** license

See also: [https://creativecommons.org/choose/](https://creativecommons.org/choose/)
3. Data storage and preservation

During the project

Storage capacity and location

Security backups
Storage and back up will be in three places: 1. On Laptop of [Name of Researcher] 2. On a portable storage device (hard drive) 3. On institutional collaborative storage.

[Name of Researcher] will be responsible for the storage and back up of data. This will be done weekly. Backups on the institutional infrastructure are automated.
After the project

Sorting of data to be preserved

File formats for preservation
Datasets from this work which underpin a publication will be deposited in Enlighten: Research Data, the University of Glasgow’s institutional data repository, and made public at the time of publication. Data in the repository will be stored in accordance with funder and University data policies.

The retention schedule for data in Enlighten: Research Data will be 10 years from date of deposition in the first instance, with extensions applied to datasets which are subsequently accessed. This complies with both University of Glasgow guidance and funder policies.
4. Data sharing and reuse

4.1 How and where will the data be shared?

Archiving and publishing

Which repository for sharing?

How will users find the data?
The project data will be shared with Yareta, the research data repository of Geneva's Higher Education Institutions, Along with the README file mentioned in 1.3.

The DOI issued to datasets in the repository will be included as part of a data citation in publications, allowing the datasets underpinning the publication to be easily identified and accessed.
4. Data sharing and reuse

4.2 Are there any necessary limitations to protect sensitive data?

**Conditions of data availability**

*Timing*

*Delay*

www.digitalbevaring.dk
Individual research subjects’ data cannot legally nor ethically be made available to non authorised people (HRA, *cf.* §2.1). Only the sponsor, the investigation team, reviewers, auditors and inspection authorities are entitled to access such data.

No personal data or data that may easily identify subjects will be provided, with respect to the Swiss law on human research (Federal Act on Research involving Human Beings (HRA)) and its applicable ordinance ClinO/KlinV/OClin/OSRUm.

(eg from [DMP Model](#), HUG CRC)
4. Data sharing and reuse

4.3 All digital repositories I will choose are conform to the FAIR Data Principles.

☐ [checkbox]

4.4 I will choose digital repositories maintained by a non-profit organisation.

→ yes / no  [radio button]
✓ A “plausible” DMP is often a condition for the release of the funds

✓ DMPs are editable throughout the funding period (must be updated)

✓ Once SNSF funding has ended and the final scientific report has been approved, the DMP cannot be modified anymore
SNSF’s DMP GUIDELINES

Key elements in English

SNSF’s Guidelines in English

SNSF’s Guidelines in French

Key elements in French
Let's have a Coffee Break!
UNIGE’s Data Ecosystem

https://datascience.unige.ch/ecosysteme-data-unige

- A multitude of data-related initiatives and services are available to the UNIGE community

- Working together, these initiatives constitute the UNIGE data ecosystem
1. 4 teams (4-8 people) per board game

2. Questions are about themes on Research Data
   - Yellow: Reuse RD
   - Red: Sharing RD + licences
   - Green: Storing RD + backups

3. Try to answer correctly to one of each card in order to escape your lab!
Any question left?
THANK YOU FOR YOUR ATTENTION

www.unige.ch/biblio

https://www.unige.ch/researchdata/en/

All images not cited here are in public domain