PhD Library Camp

Module 1

Managing research data as a junior scientist

Vincent.Huber@unige.ch
Audrey.Bellier@unige.ch

Winter 2020

AGENDA

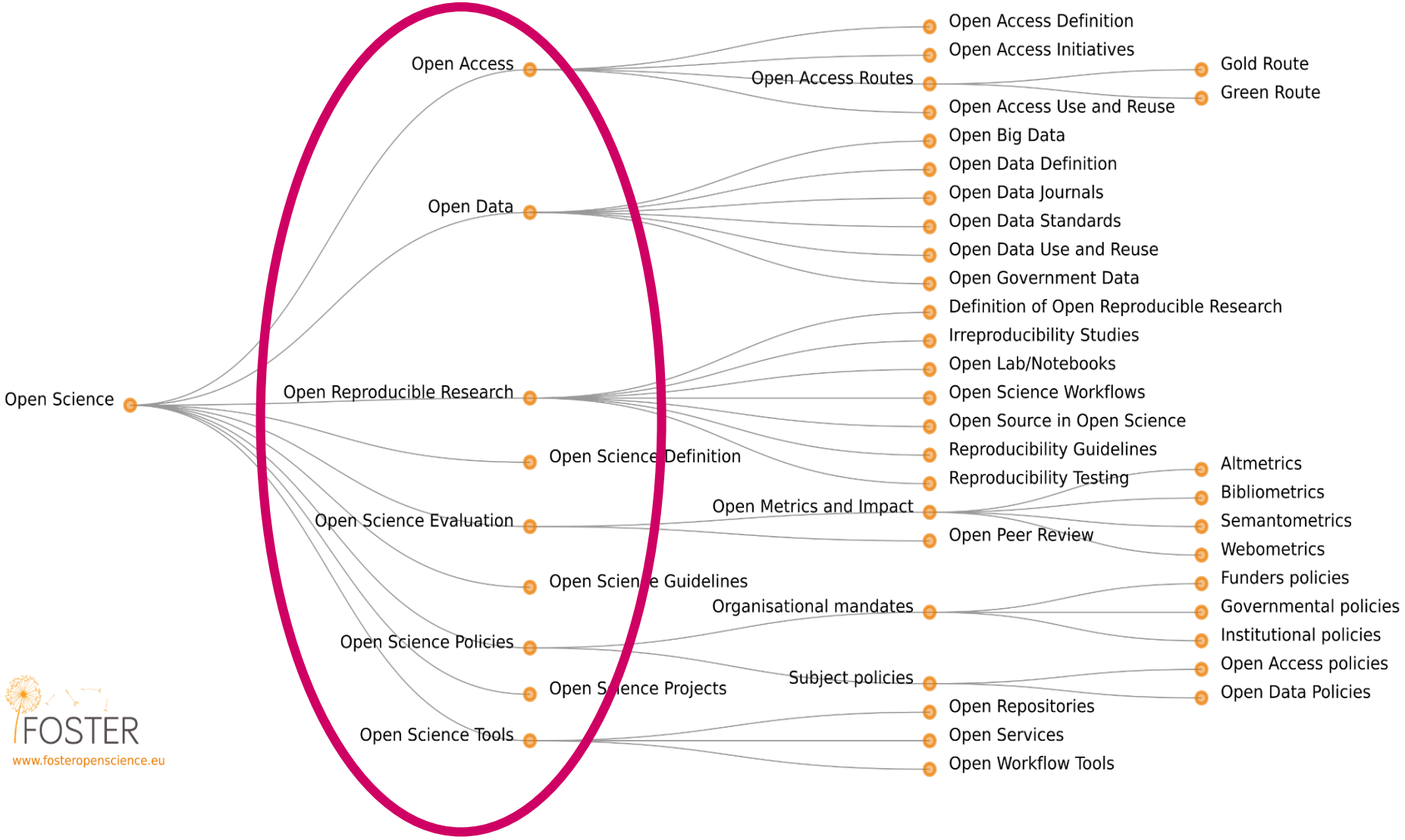
Content
Welcome and Presentation of the day
General introduction on Open Science
Research data – definitions, organization, description, formats, tools
Data Repositories, Copyrights and Licenses
Data Management Plan (DMP)
Escape the Lab
General discussion, Feedback

https://insightsresources.seek.com.au/the-top-icebreaker-interview-questions-you-should-be-asking

What is Open Science?

Open science is the movement to make scientific research (including publications, data, physical samples, and software) and its dissemination accessible to all levels of an inquiring society, amateur or professional.

*Woelfle, M.; Olliaro, P.; Todd, M. H. (2011). "Open science is a research accelerator". Nature Chemistry. **3**: 745–748. doi:10.1038/nchem.1149*



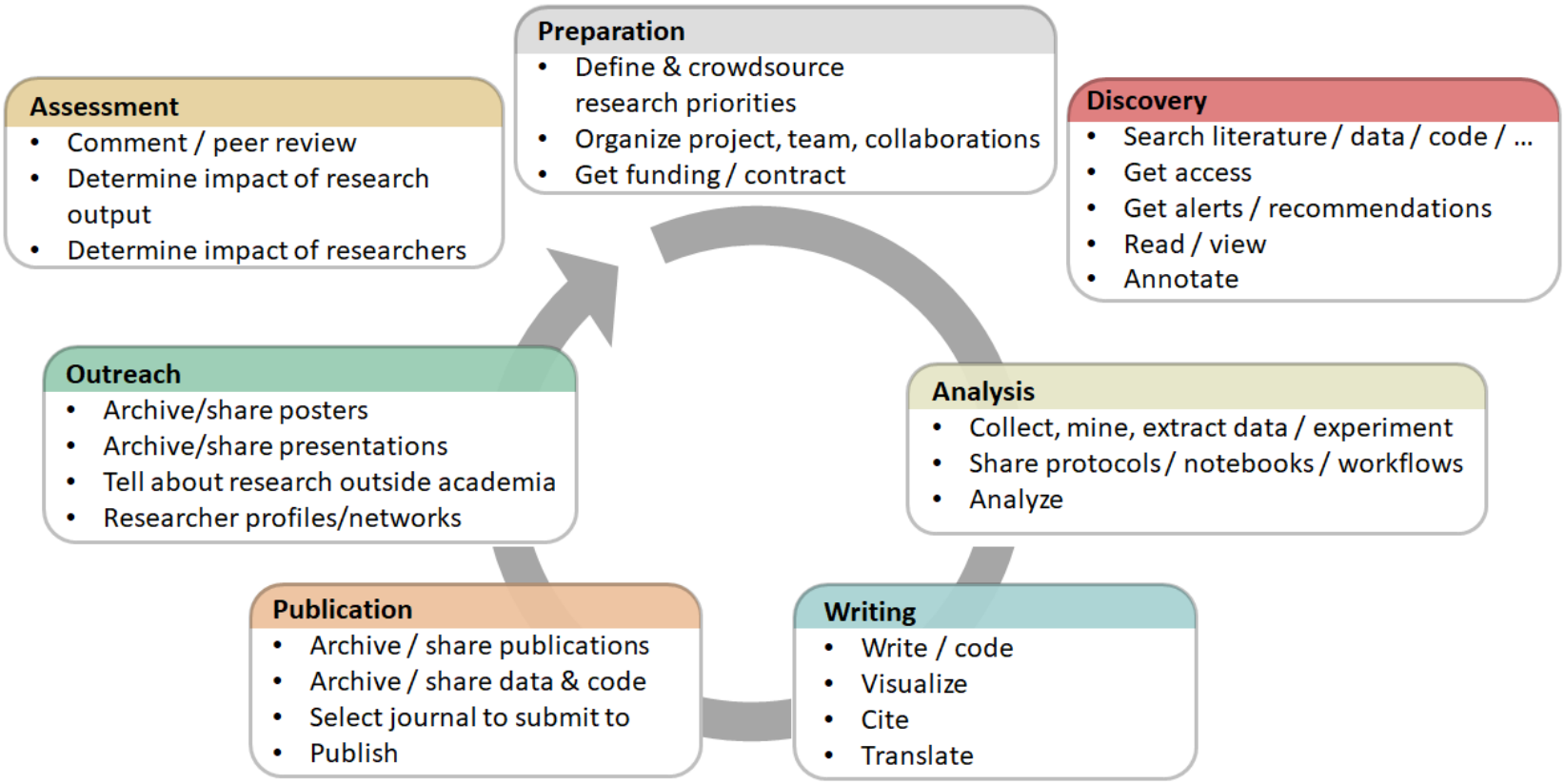
FOSTER
www.fosteropenscience.eu

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.

Dan Gezelter, 2009. http://openscience.org/what-exactly-is-open-science/

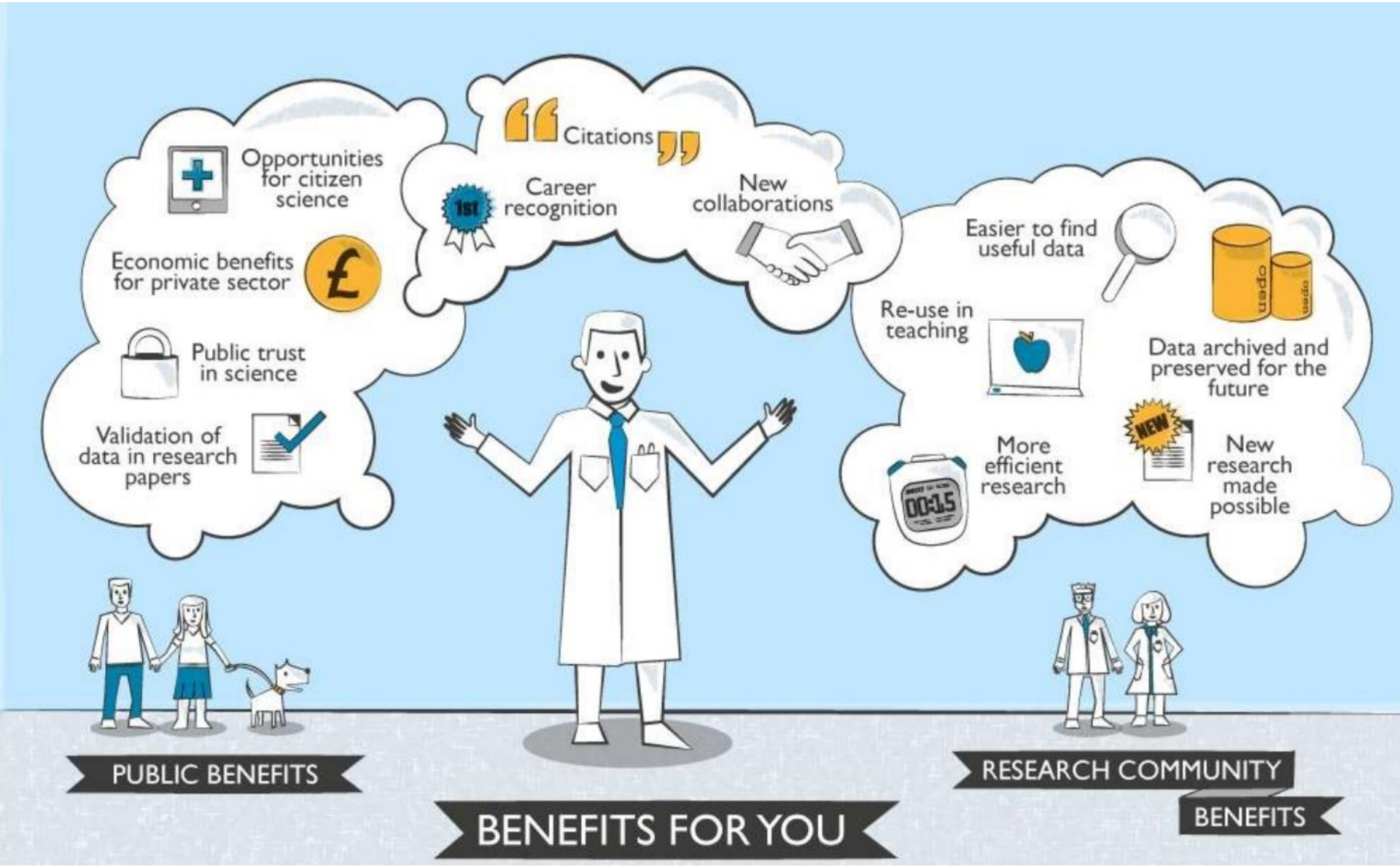
RESEARCH CYCLE



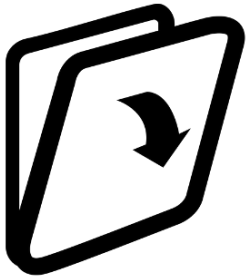
https://open-science-training-handbook.gitbook.io/book

Pros and cons of Open Science

- Broader dissemination of scientific results
  - Paywalls, licenses, poor formating, proprietary formats… empede the sharing of knowledge
- Reproducibility crisis
  - Too many researches cannot be reproduced
- Good use of public money
- In favor of scoopers and predators



Hole, Brian. The Journal of Open Archaeology Data and PRIME: Incentivising Open Data Archiving [Internet]. 18 Jan 2013 [cited 28 Nov 2016]. Available: http://www.slideshare.net/brianhole/the-journal-of-open-archaeology-data-and-prime-incentivising-open-data-archiving



MANAGE YOUR RESEARCH DATA

WHAT IS RESEARCH DATA?

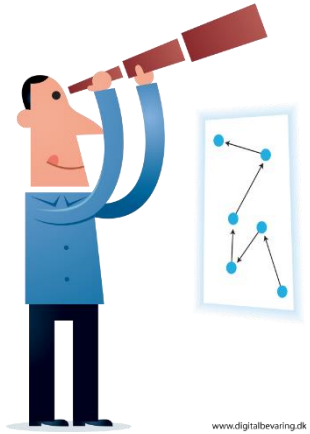
Factual records (numbers, text, images and sounds), which are used as primary sources for scientific research and are generally recognized by the scientific community as necessary to validate research results.

Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics, 2007, p. 18

http://www.oecd.org/fr/sti/sci-tech/38500823.pdf

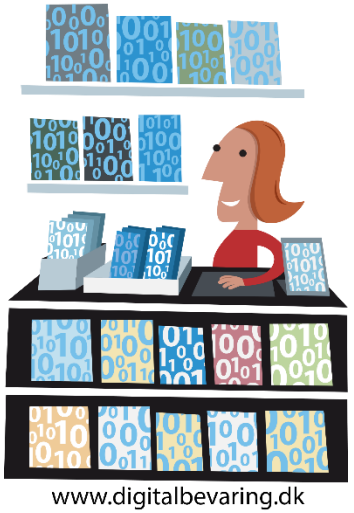
CLASSIFICATION OF RESEARCH DATA

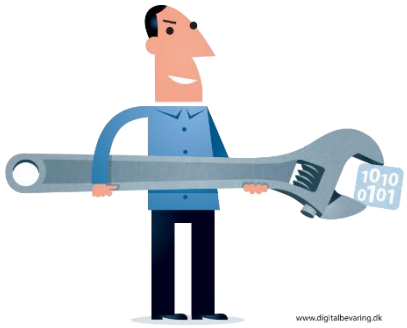
observational



simulation / models

derived / compiled



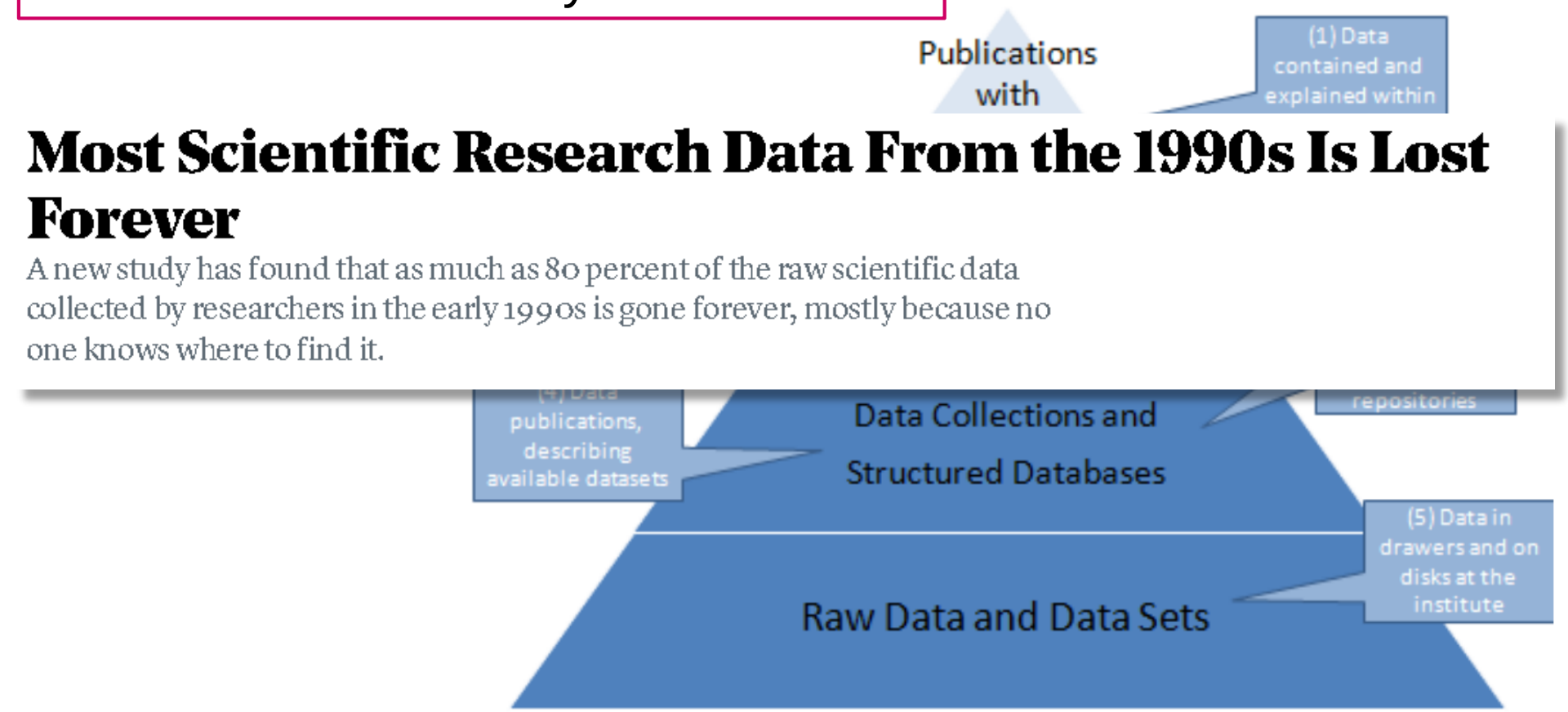


experimental

reference

RESEARCH DATA AND PUBLICATIONS

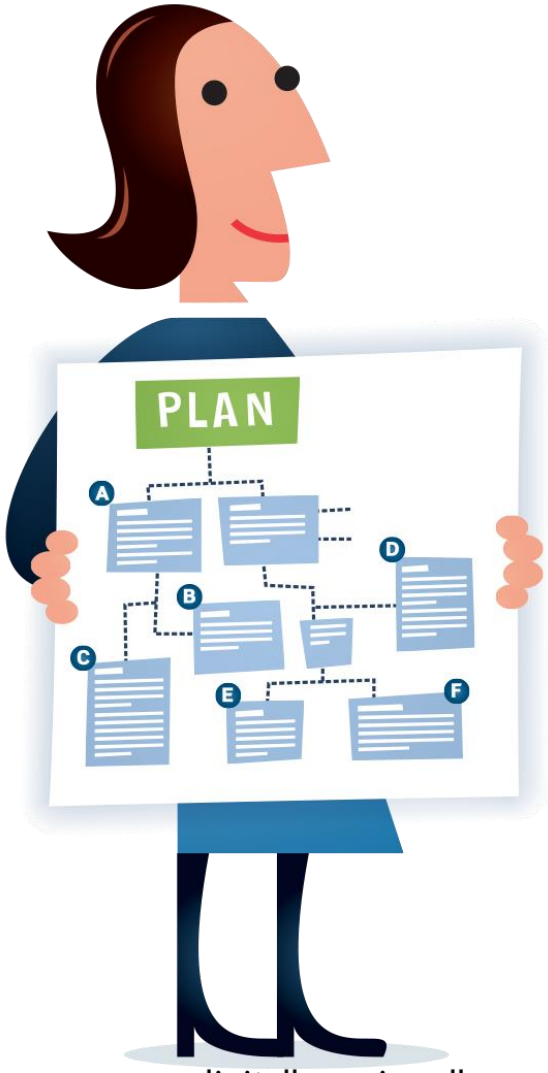
Most of the data are ignored or lost for the scientific community



https://www.theatlantic.com/national/archive/2013/12/scientific-data-lost-forever/356422/

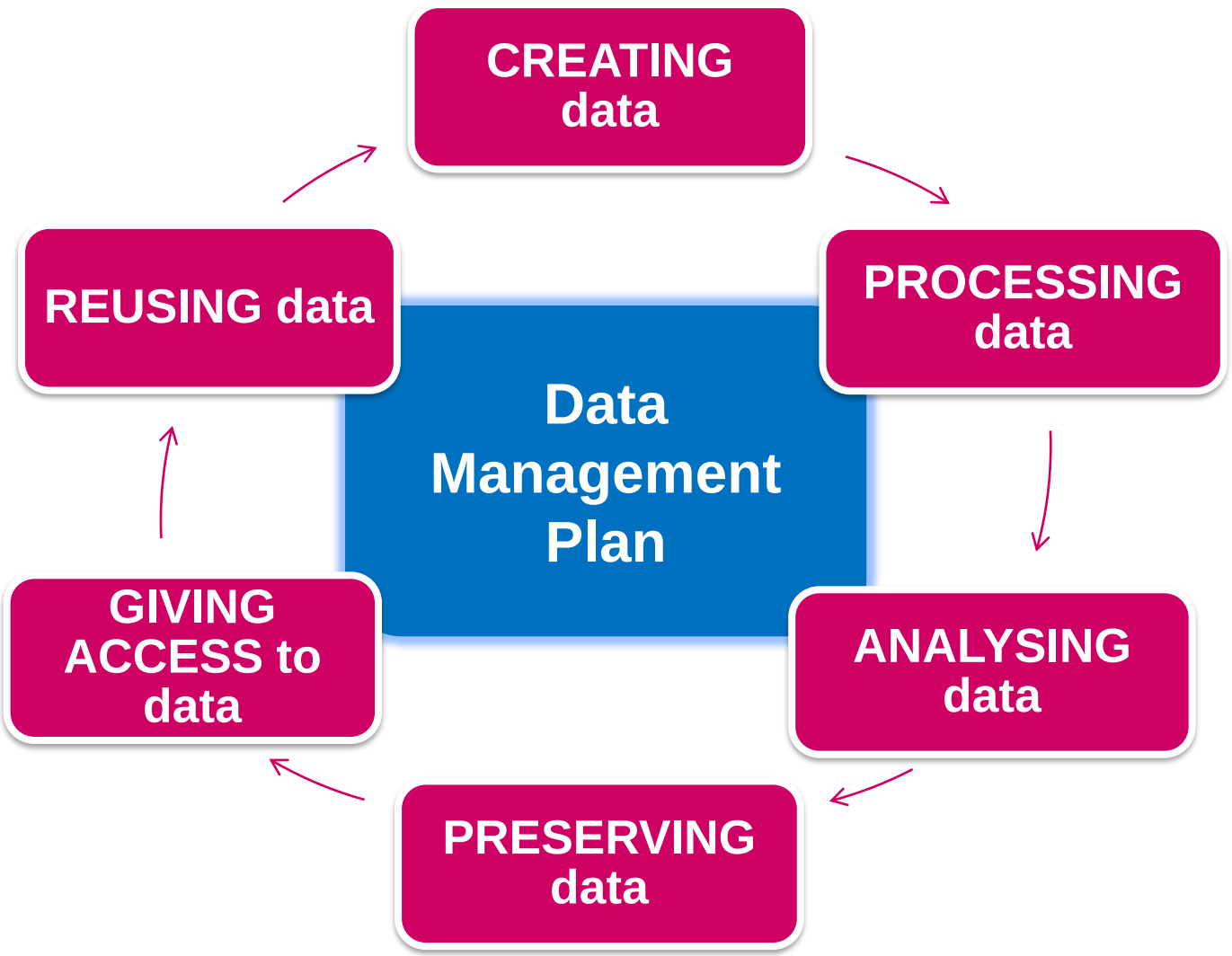
Report on integration of data and publications, S. Reilly, 2011

What to do?



Set up good practices to manage research data

DATA LIFE CYCLE AND THE DMP



MAKE YOUR DATA FAIR

http://www.datafairport.org/

- ✓ persistent identifier
- ✓ enriched metadata
- ✓ Searchable and findable online

- ✓ retrievable using standard communication protocols
- ✓ possibilily to define access rights

→ in a repository that ensures long-term preservation

- ✓ standard formats
- ✓ controlled vocabulary to describe data

→ data will be compatible and combinable with others

- ✓ well-described & documented (e.g. in a README file)
- ✓ clear conditions to cite and reuse data (e.g. CC licenses)

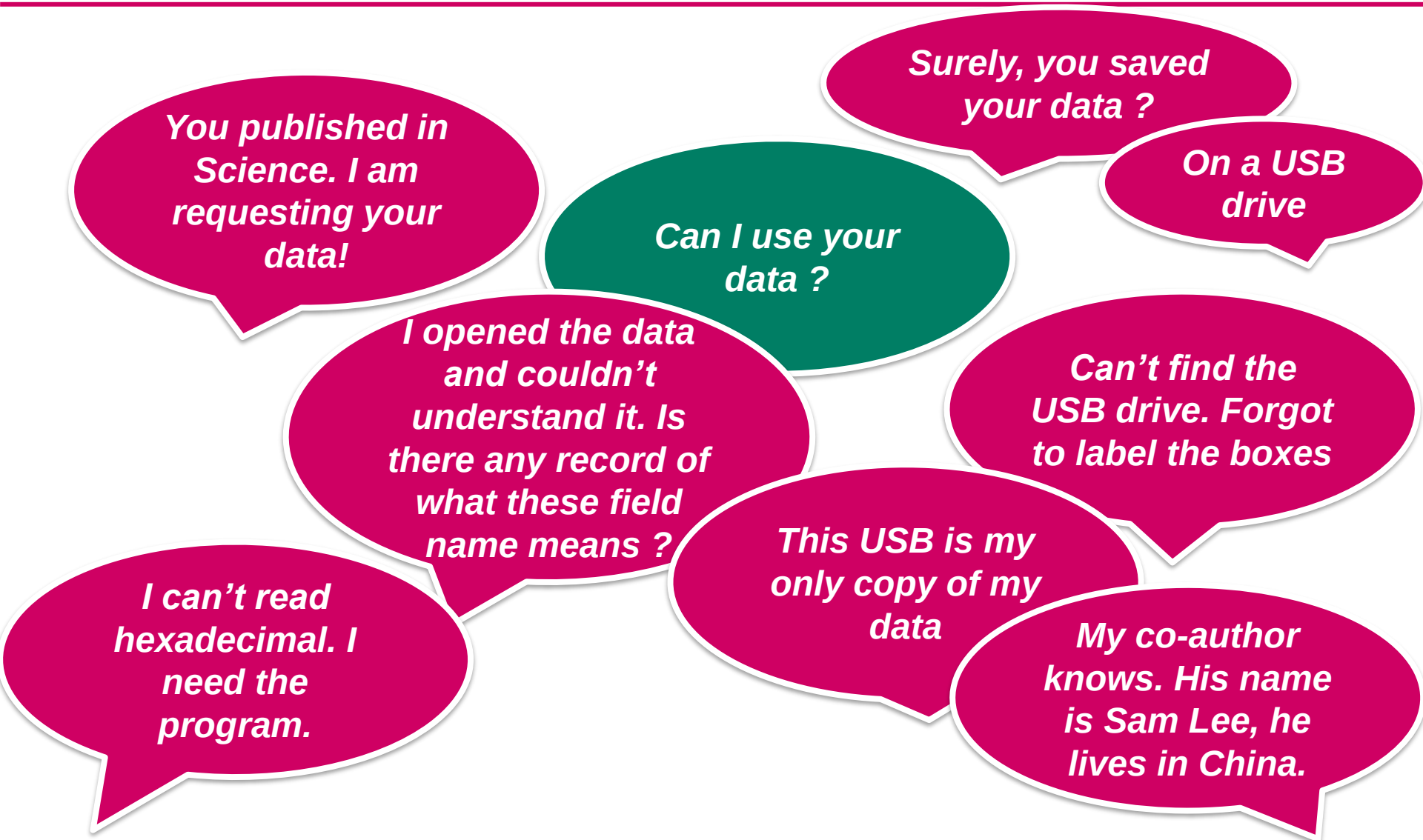
→ allow data to be correctly interpreted and reused

IN A NUTSHELL…

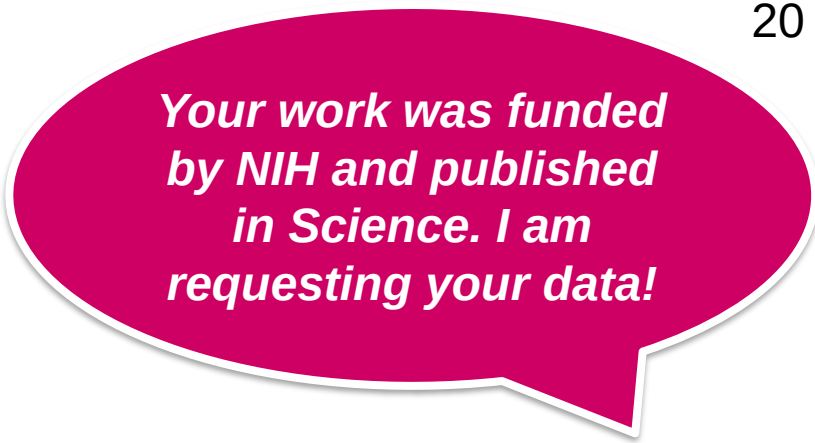
https://www.youtube.com/watch?v=N2zK3sAtr-4

WHAT STRUCK YOU IN THE VIDEO ?

MMMH. A FEW PROBLEMS…

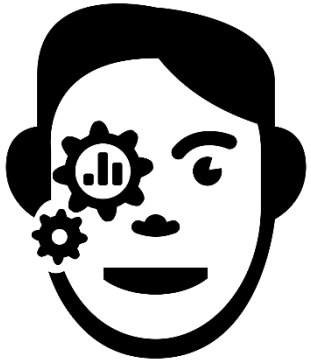


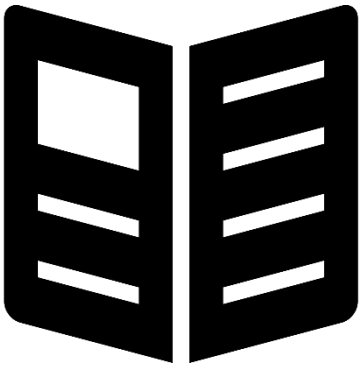
Requirements for data sharing



Funders

Institution





Journals

FUNDER REQUIREMENTS

Most of funding bodies now require DMP (Data Management Plans) and healthy data management practices :

FNSNF
FONDS NATIONAL SUISSE
DE LA RECHERCHE SCIENTIFIQUE

HORIZON 2020

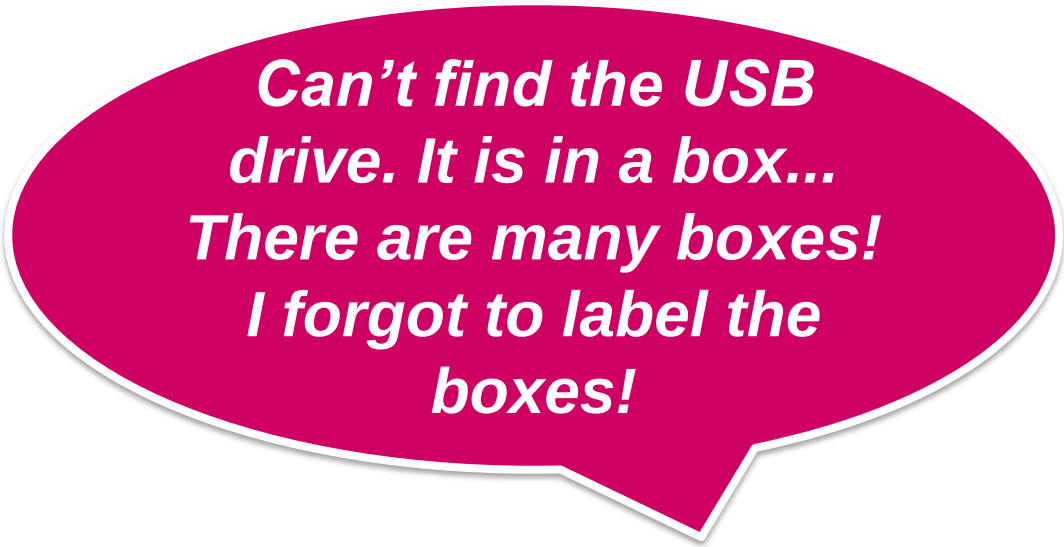
erc European Research Council

DFG Deutsche Forschungsgemeinschaft

NHS National Institute for Health Research

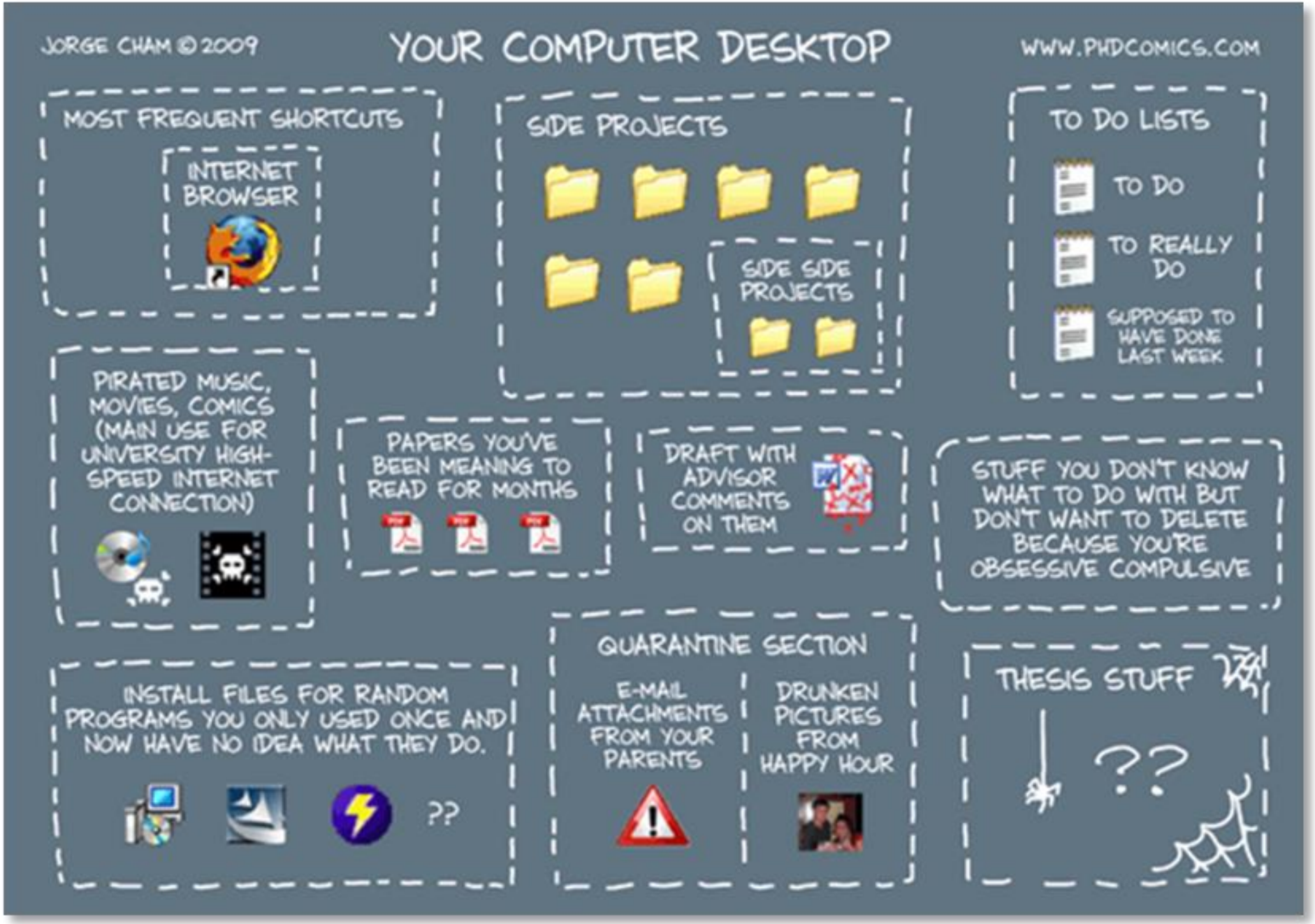
BILL & MELINDA GATES foundation

https://www.unige.ch/researchdata/en/footer/policy/other-policies/



Let’s talk about

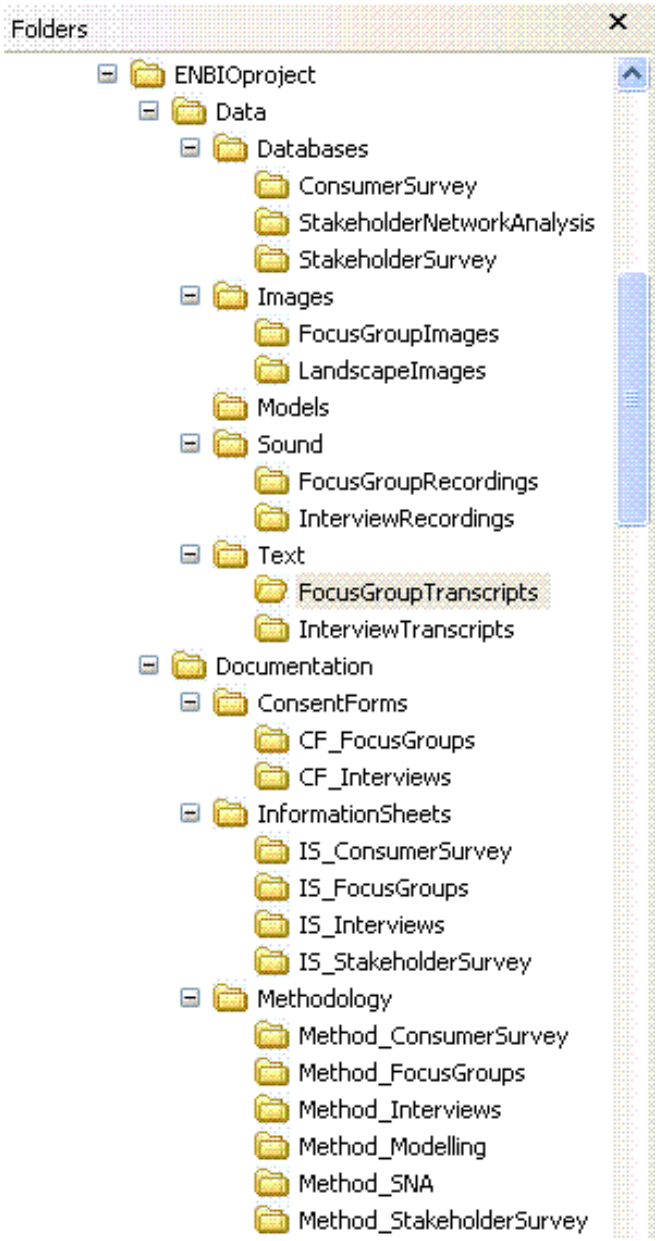
- *Folder architecture and naming*
- *Naming conventions*



Jorge Cham, 2009. http://www.phdcomics.com/comics/archive.php?comicid=1168 [Accessed January 9 2017]

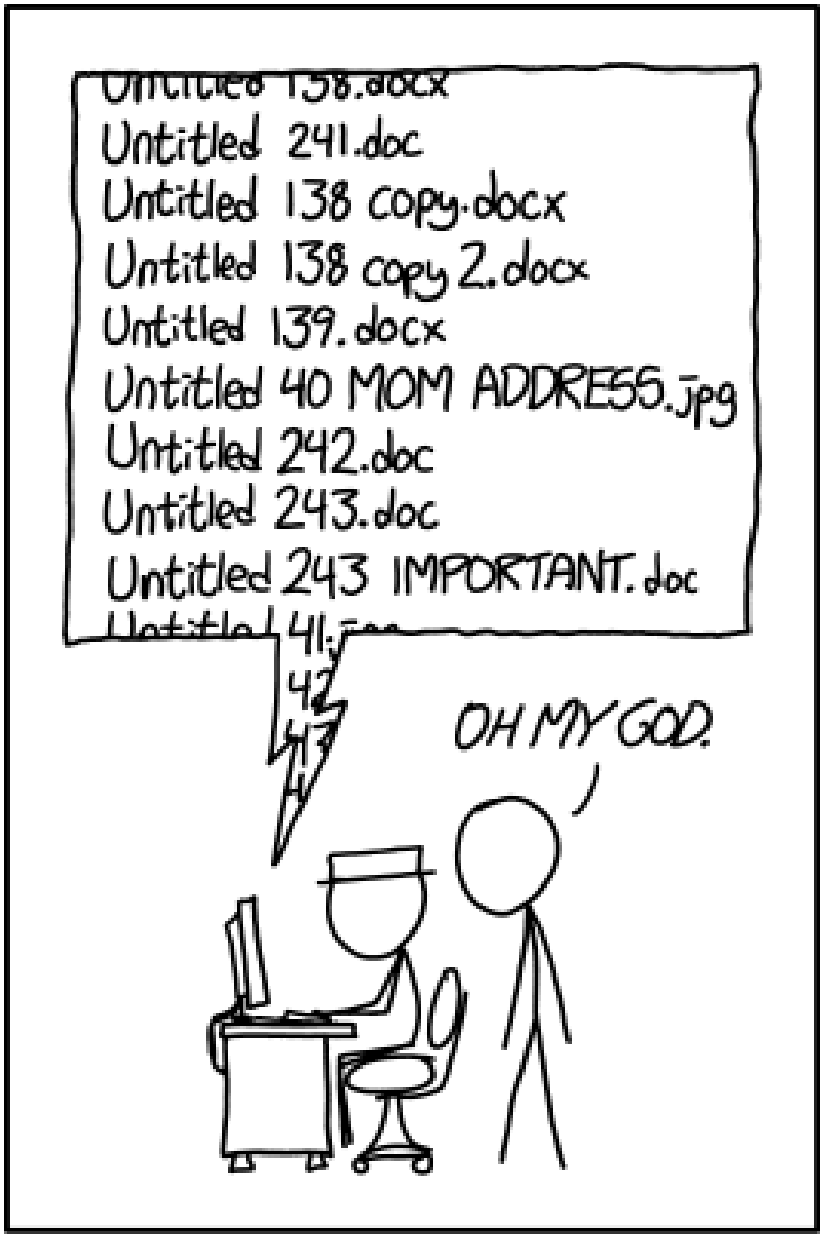
FILE STRUCTURES

Activity: Organize your files



- ✓ Avoid overlapping categories
- ✓ Don’t let folders get too big
- ✓ Don’t let structures get too deep

File naming



https://xkcd.com/1459/

Exercice: File naming

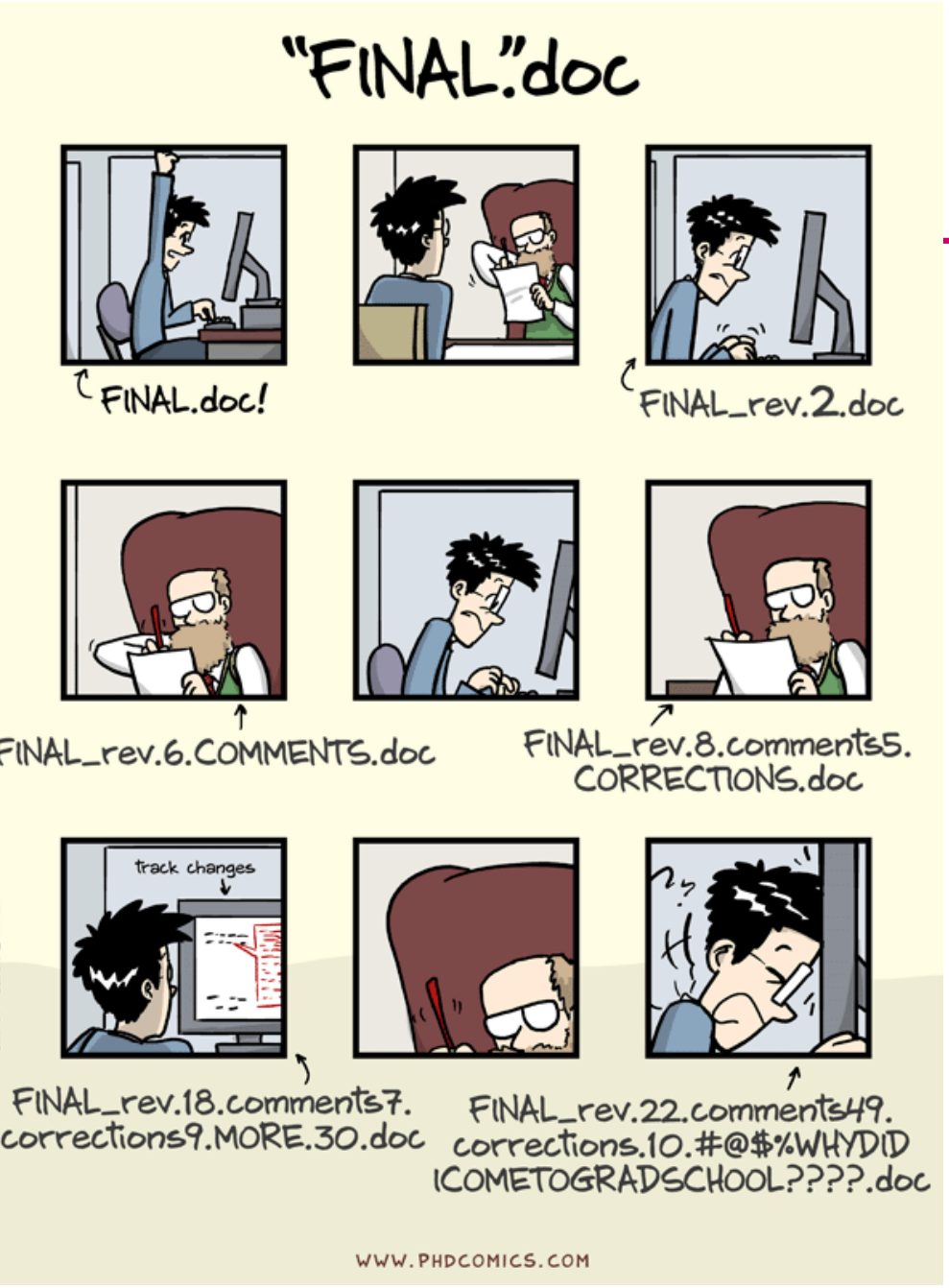
My passwords.doc	My data.xls
IMPORTANT.doc	My study.doc
My Thesis final final.doc	Doc.1.doc
Ma thèse version 12.doc	New doc.doc
Data 01/08/2016.xls	Int 1 (2).doc
Data 10 jan. 2016.xls	Interview 1.doc

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 pratices

Best Practice	Example
Order dates beginning with the **year** to enable sorting by date (kept 4 digit for year)	YYYY-MM-DD or YYYYMMDD
Limit the file name to **32 characters**	32CharactersLooksExactlyLikeThis.csv
For sequential numbering, use **leading zeros** to allow for multi-digit versions For a sequence of 1-10: 01-10 For a sequence of 1-100: 001-010-100	**NO :** ProjID_2.csv ProjID_12.csv **YES :** ProjID_02.csv ProjID_12.csv
Don't use **special characters** or **spaces** & , * % # ; * ( ) ! @$ ^ ~ ' { } [ ] ? < > -	**NO :** name&date@location.doc **YES :** name_date_location.doc
Use only **one period** (for the file extension)	**NO :** name.date.doc **YES :** name_date.doc
Avoid using **generic data file names** that may be ambiguous when moved	**NO :** MyData.csv **YES :** ProjID_date.csv

Adapted from MIT Library, DM workshop, 2016



FILE VERSIONING

1. Avoid imprecise “final” labels

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

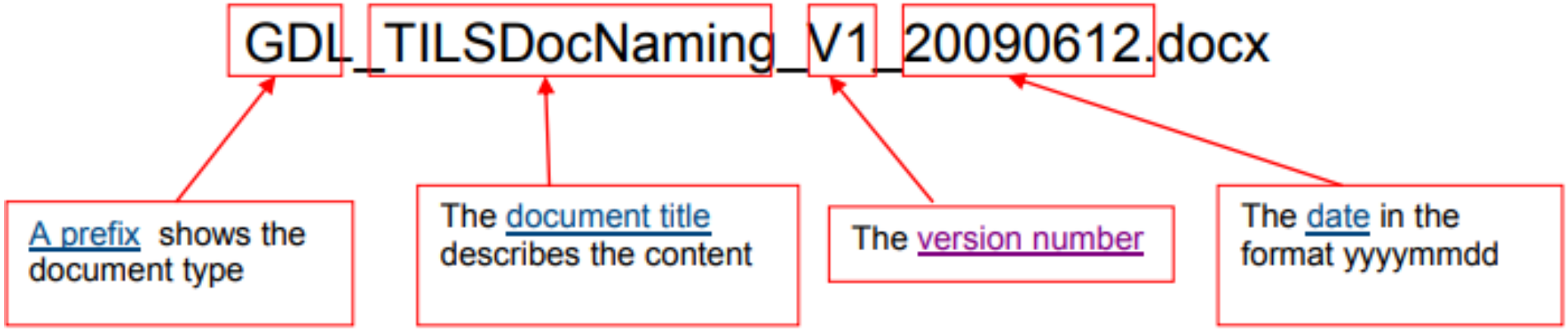
3. Document your convention

4. Consider your version control needs

EXAMPLE OF NAMING CONVENTION

TILS Document Naming Convention

Document naming for the TILS Division should follow this convention:



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.

http://www.data.cam.ac.uk/files/gdl_tilsdocnaming_v1_20090612.pdf

KEY PRINCIPLES OF FILE ORGANIZATION

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?

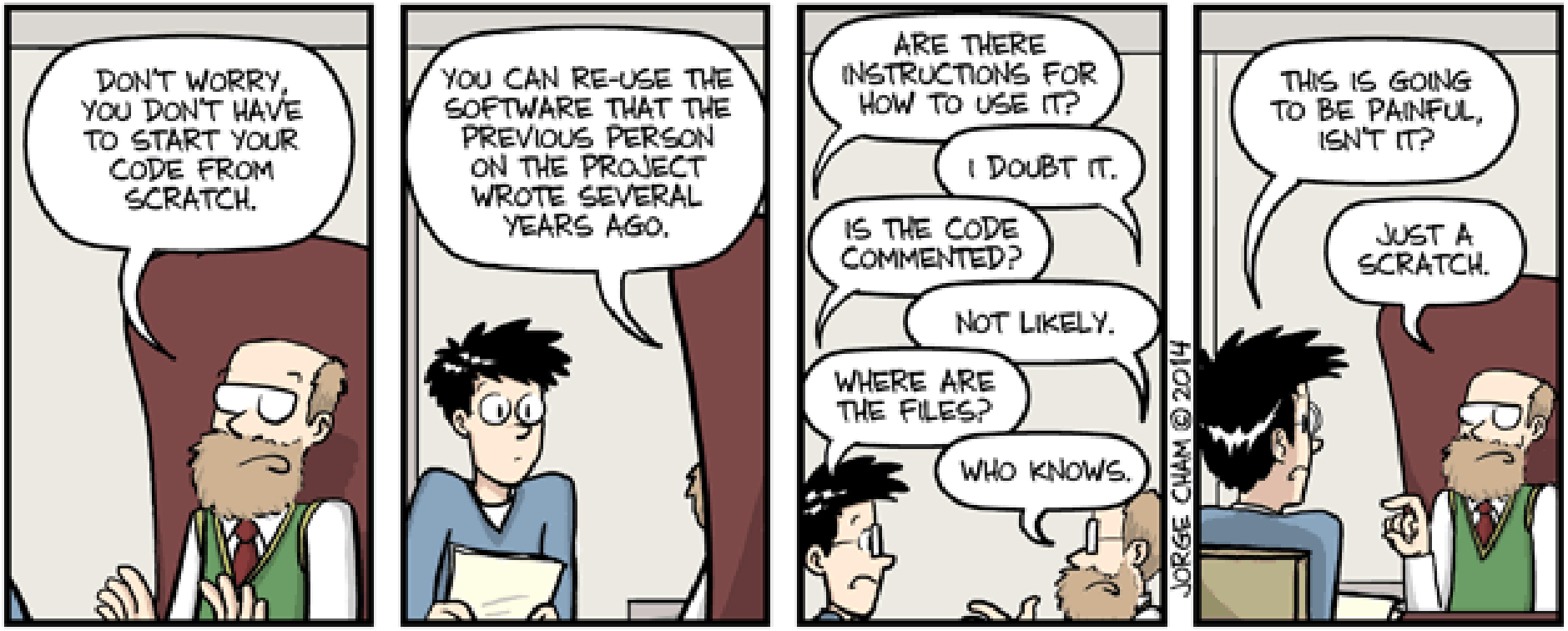
***Make* a system. *Share* the system. *Follow* the system.**

(MIT libraries)

Let's talk about

- *Metadata & documentation: Documenting the format, tools, fields, etc.*
- *Readme files*

Piled Higher and Deeper *by Jorge Cham*



title: "Scratch" - originally published 3/12/2014 WWW.PHDCOMICS.COM

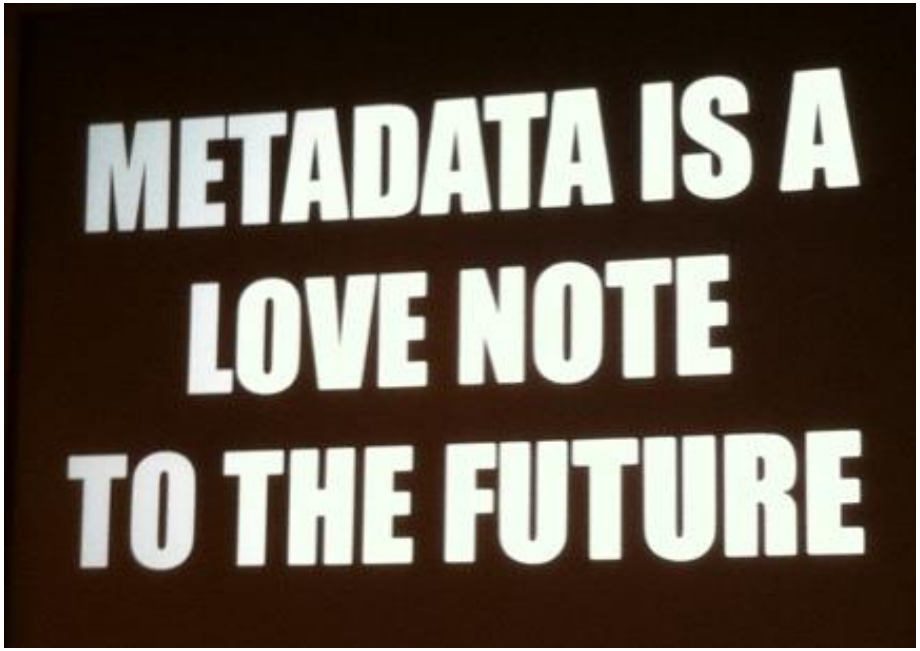
METADATA

- Research data need to be documented at various levels:
  - project level
  - file or database level
  - and variable or item level.

- Aim is to make them:
  - ✓ Understandable
  - ✓ Reusable
  - ✓ Findable

Cf. FAIR data ☺

-> Avoid ambiguity & misinterpretations



3 Main Types of Metadata

- **Descriptive metadata** enables discovery, identification, and selection of resources.
  - Title/unique ID, author, abstract, keywords, funder, date,..
- **Administrative metadata** facilities the management of resources.
  - Rights statement, timestamp, file format, version, preservation,..
- **Structural metadata** describes how different components of a set of data relate to one another
  - Data dictionary, variable list, database schema, taxonomy, abbreviations…

Examples of Data Documentation

- laboratory notebooks & experimental protocols
- questionnaires, codebooks, data dictionaries
- software syntax and output files
- information about equipment settings & instrument calibration
- database schema
- methodology reports
- provenance information about sources of derived or digitised data
- …

README.TXT

Readme.txt : File with metadata accompagnying 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

AUTHOR_DATASET_ReadmeTemplate.txt - Bloc-notes

Fichier Edition Format Affichage ?

```
This DATSETNAMEreadme.txt file was gene

------------------
GENERAL INFORMATION
------------------

Title of Dataset:

Author Information (Name, Institution,

        Principal Investigator:
        Associate or Co-investigator:
        Alternate Contact(s):

Date of data collection (single date, r

Geographic location of data collection:

Information about funding sources or sp


-------------------------
SHARING/ACCESS INFORMATION
-------------------------

Licenses/restrictions placed on the dat

Recommended citation for the data:

Citation for and links to publications

Links to other publicly accessible loca

Links/relationships to ancillary or rel


--------------------
DATA & FILE OVERVIEW
-----

File

Relat
```

```
--------------------
DATA & FILE OVERVIEW
--------------------

File list (filenames, directory structure (for zipped files) and brief description of all data files):

Relationship between files, if important for context:

Additional related data collected that was not included in the current data package:

If data was derived from another source, list source:

If there are there multiple versions of the dataset, list the file updated, when and why update was made:


-------------------------
METHODOLOGICAL INFORMATION
-------------------------

Description of methods used for collection/generation of data: <Include links or references to publicatio
or other documentation containing experimental design or protocols used in data collection>

Methods for processing the data: <describe how the submitted data were generated from the raw or collecte
data>

Software- or Instrument-specific information needed to interpret the data, including software and hardwar
version numbers:

Standards and calibration information, if appropriate:

Environmental/experimental conditions:

Describe any quality-assurance procedures performed on the data:

People involved with sample collection, processing, analysis and/or submission:


-------------------------
DATA-SPECIFIC INFORMATION <Create sections for each datafile or set, as appropriate>
-------------------------

Number of variables:

Number of cases/rows:
```

https://data.research.cornell.edu/content/readme -> download a template

```
========== HEADER ==========
Readme.txt for ____[add name/title here]___ dataset
Documentation written on ____[add date as YYYYMMDD here]___
By ____[add Last name, First name here]___
Updated <YYYYMMDD>, by ____[add Last name, First name here]___


========== DATA DESCRIPTION ==========
ACKNOWLEDGEMENTS
Project title:
Funding agency/agencies:
Award Number:
Award Period:

Investigator Name:
Investigator Institution:
Investigator Address:
Investigator Email:
Investigator Role (related to this dataset): [e.g., data collection, data processing/cleaning, data analysis, data visualization, lab coordinator, site manager, etc.]
Investigator ID (if applicable): [e.g. ORCID]

Investigator Name:
Investigator Institution:
Investigator Address:
Investigator Email:
Investigator Role (related to this dataset): [e.g., data collection, data processing/cleaning, data analysis, data visualization, lab coordinator, site manager, etc.]
Investigator ID (if applicable): [e.g. ORCID]

...(repeat as needed)


DESCRIPTION
What this project or dataset is about, in terms of topical, geographic, or temporal coverage:

Other details about the content, formats, and internal relationships the dataset:


CITATION
Author(s)/Creator(s):
Title:
Year of Publication[when dataset was published/released, not data collection or coverage date]:
Publisher: [data center or repository]:
Identifier (DOI or any applicable identifier, including edition or version):
Availability and Access (URL or other location information for data): [e.g. https://www.datacite.org/services/cite-your-data.html]


DATE(S) of DATA COLLECTION
[single date, range, approximate date]
<suggested format YYYYMMDD>

GEOGRAPHIC LOCATION(S) of DATA COLLECTION
[where data was collected]
<suggested format: city, state, zip code, country, GPS>

DIRECTORY/FILE NAMING CONVENTIONS
[explain any metadata embedded in the directory or file names]
e.g. YYYY-MM-DD-INSTRUMENT-NAME, YYYY-MM-DD-ALGORITHM-NAME, YYYY-MM-DD-HH:MM-ARTIST-NAME
...[repeat as needed]

CHANGELOG
[note any changes to, additions to, or replacement of files stored in the Scholarly Data Archive]
```

https://legacy.lib.utexas.edu/d7/sites/default/files/utl_readme.txt another example

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
  - to well describe the data, its contents, structure/organization
- Write your readme document as **a plain text file** (or PDF/A)
- Use a template (here/there) 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.

IMPORTANCE OF METADATA

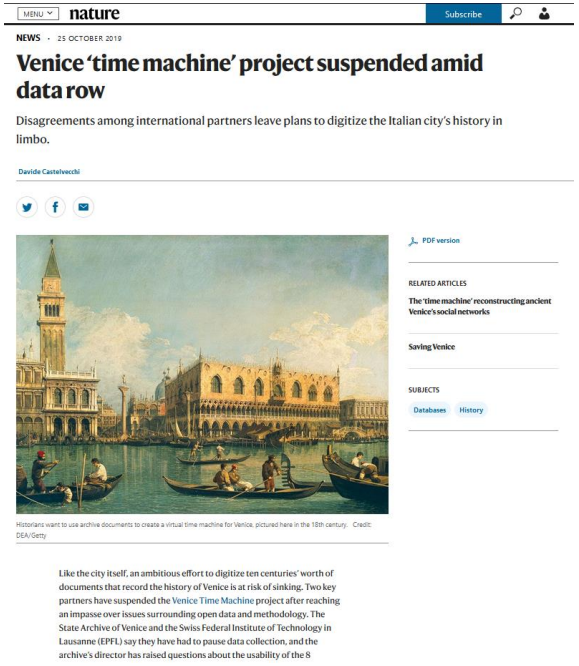
2012 – Project of officially **launched**:
Venice's State Archive + Ca' Foscari Univ. + EPFL (DHLAB)

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!**



MENU nature Subscribe

NEWS · 25 OCTOBER 2019

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

PDF version

RELATED ARTICLES

The 'time machine' reconstructing ancient Venice's social networks

Saving Venice

SUBJECTS

Databases History

Historians want to use archive documents to create a virtual time machine for Venice, pictured here in the 18th century. Credit: DEA/Getty

Like the city itself, an ambitious effort to digitize ten centuries' worth of documents that record the history of Venice is at risk of sinking. Two key partners have suspended the Venice Time Machine project after reaching an impasse over issues surrounding open data and methodology. The State Archive of Venice and the Swiss Federal Institute of Technology in Lausanne (EPFL) say they have had to pause data collection, and the archive's director has raised questions about the usability of the 8

DOI: 10.1038/d41586-019-03240-w

I opened the data and couldn't understand it. (…) I noticed that you called your data fields "Sam"... Is that an abbreviation?

Is there any record of what these field name means ?

Let's talk about

- *Documenting your variables, any abreviations, etc… in a readme file.*
- *Using standards*

DESCR

in Codebook or Rea

- Including references to st
- Including unitsof measurr

```
********************************** FRUBASE

                                VERSION 4

THE FRUBASE PACKAGE ACCOMPANIES:

Jordano, P. 1995. Angiosperm fleshy fru
analysis of adaptation and constraints
Naturalist 145: 163-191.

It contains a copy of the main data file
well as other accompanying files (see be

Taxonomic arrangement follows:
Cronquist, A. (1981). An integrated syst
plants. Columbia University Press.

Nomenclature follows Stevens, P. F. (20
Website. Version 8, June 2007. http://w
This scheme follows: A.P.G. [= Angiosper
of the Angiosperm Phylogeny Group class
of flowering plants: APG II. Bot. J. Li
```

```
CL          Class
SCL         SubClass
ORD         Order
FAM         Family
GEN         Genus
SP          Species
REF         Reference number - This is my maintenance code for updates.
NEWREF      New Reference number - These are the refs numbers in the
                                   files REFS and SUMMARY.
FAMLAB      Family Label -  An 8-character label for family.
GENLAB      Genus Label -   An 8-character label for genus.
SPLAB       Species Label - An 8-character label for species.
COD         Species code -  A 5-character code for the species.
DISPCAT     Disperser type category - BIRDS, MIXED, MAMMALS.
DISP        Disperser type - Finer categorization. Not yet completed.
                             Needs revision.
MEGAFAUNA Whether the fruit species is associated with dispersal
            of seeds by megafauna.
AREA        Geographic area - Major geographic areas of the data
              sources.
             MEurope: Mediterranean Europe (also includes Israel and
                      Morocco.
             NEurope: Temperate and Northern Europe.
             NAmerica: North America, excl. Southern Mexico.
             NTAmerica: Neotropical America, incl. Southern Mexico.
             Africa: Africa, south of Sahara.
             Australasia: Australia, Malaysia, New Guinea, and whole
                          SouthEast Asia.
HABIT       Growth habit - Tree, Shrub, Liana, Herb.
FRUIT       Fruit type - Type of fruit. Not completed yet.
                         Needs revision.
COLOR       Fruit color -
             Black, Blue, Red, Orange, Yellow, Green, Brown, White
COMPCOLOR Whether the fruit is monocolored, bicolored, or multicolored.
LENG        Fruit length (mm)
DIAM        Fruit diameter (mm)
FRFM        Fruit fresh mass (g)
PDM         Dry mass of pulp per fruit (g)
SDM         Dry mass of seed(s) per fruit (g)
SEEDS       Number of seeds per fruit
SEEDM       Seed dry mass (g)
RY          Relative yield of pulp (PDM/FRFM*100)
KJG         Specific energy content of pulp (kJ/g)
KJFR        Total energy content per fruit (kJ)
PCW         Percent water content of fruit
LIP         Proportion of lipids (per g dry pulp)
PRO         Proportion of protein (per g dry pulp)
NSC         Proportion of non-structural carbohydrates (per g dry pulp)
ASH         Proportion of minerals (per g dry pulp)
FIB         Proportion of acid-detergent fiber (per g dry pulp)
```

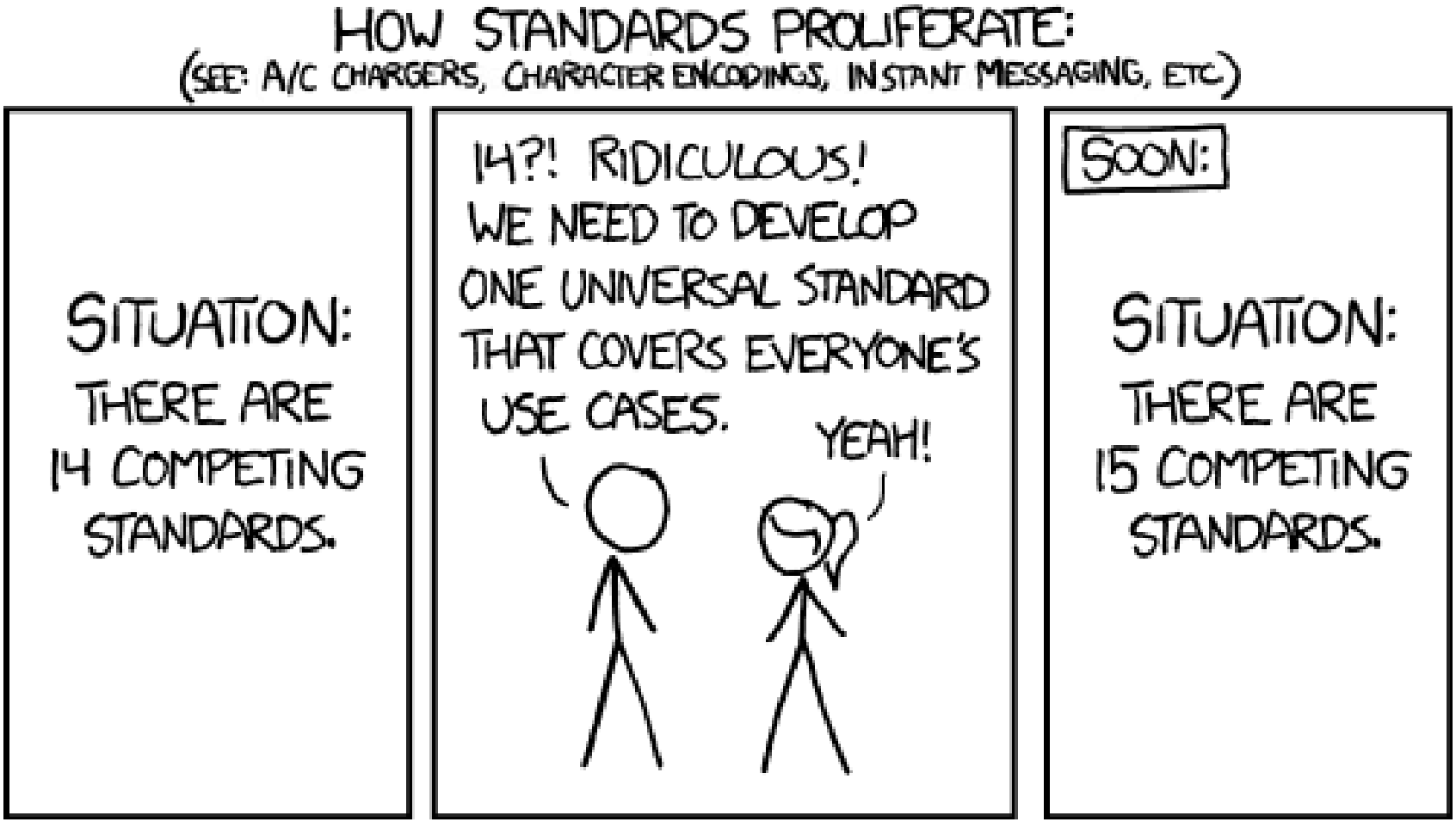
STRUCTURE YOUR DATA

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

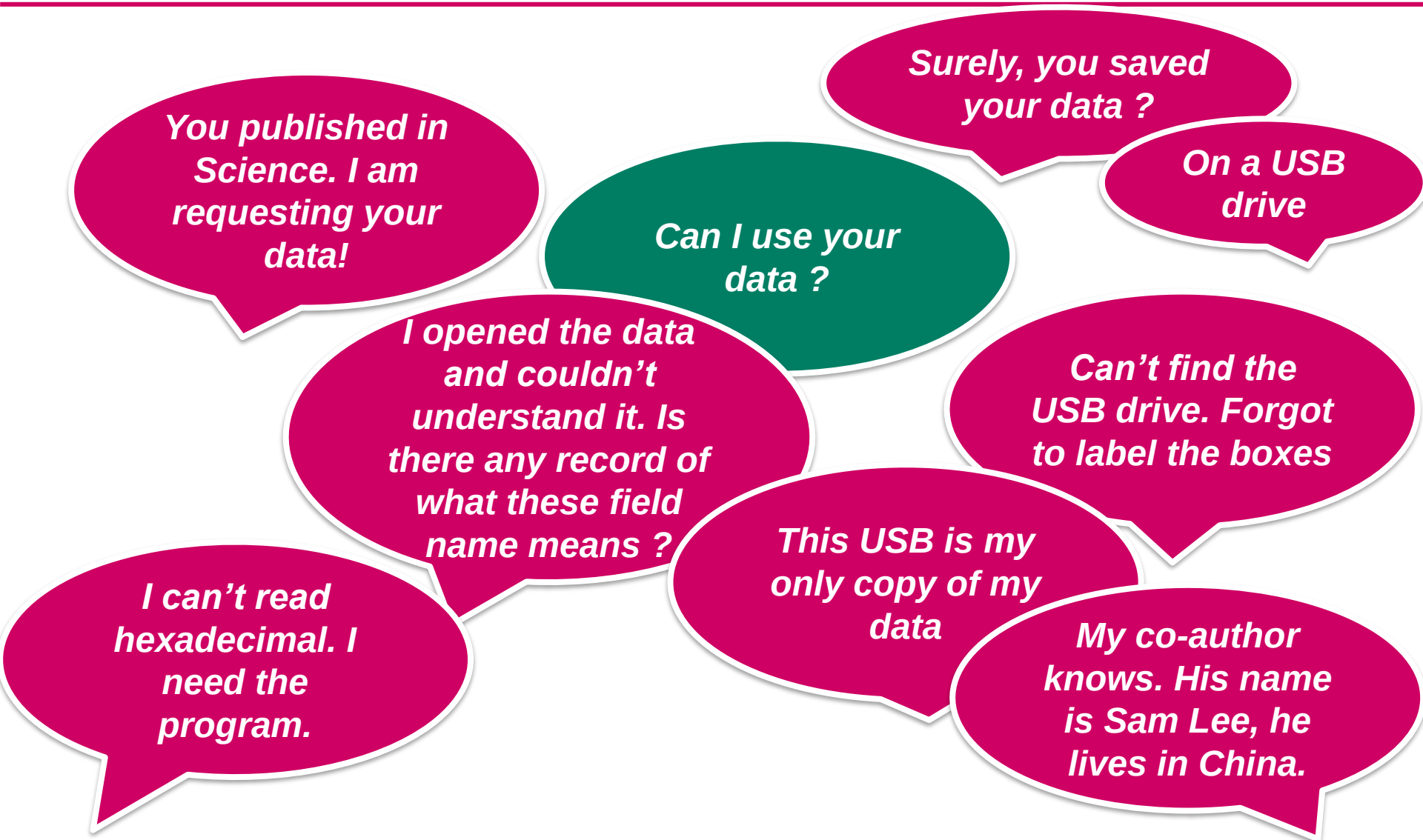
Each standard is clearly defined and documented → Each dataset can use the terms / be organized consistently → Easy to share data across institutions, applications and disciplines

To find standards: https://fairsharing.org/standards/

FAIRsharing.org
standards, databases, policies



MMMH. A FEW PROBLEMS…



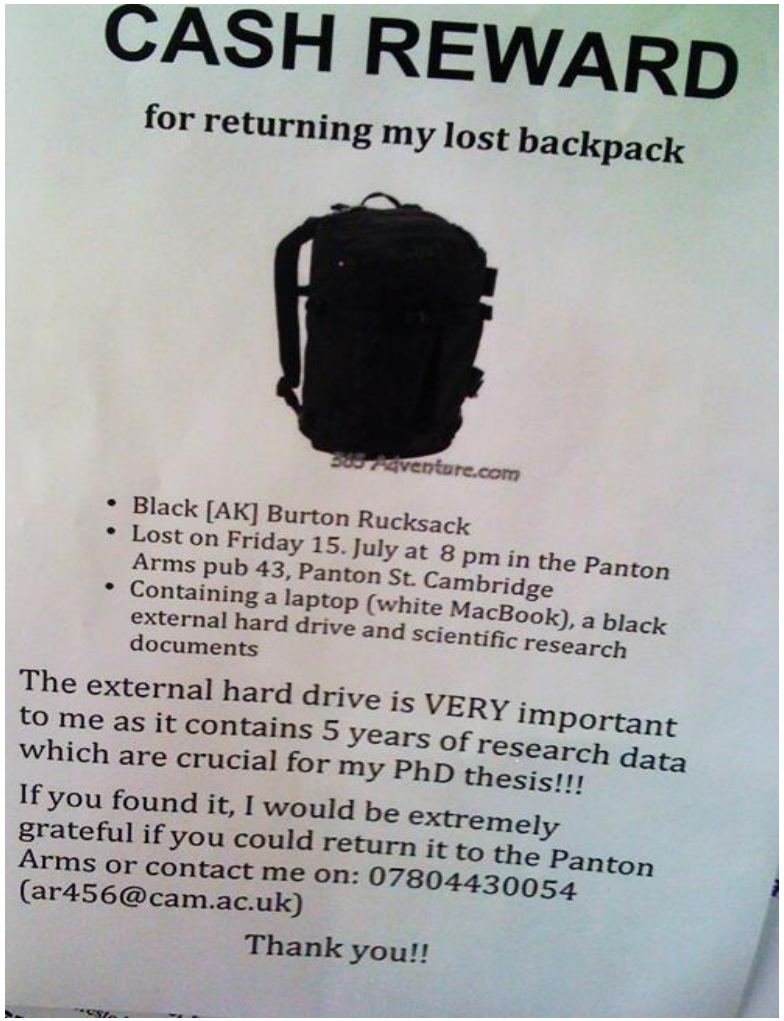
But surely, you saved your data ?

- I did, I saved it on a USB drive

- *Let's talk about*
  - *Data storage*
  - *Back-ups*

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

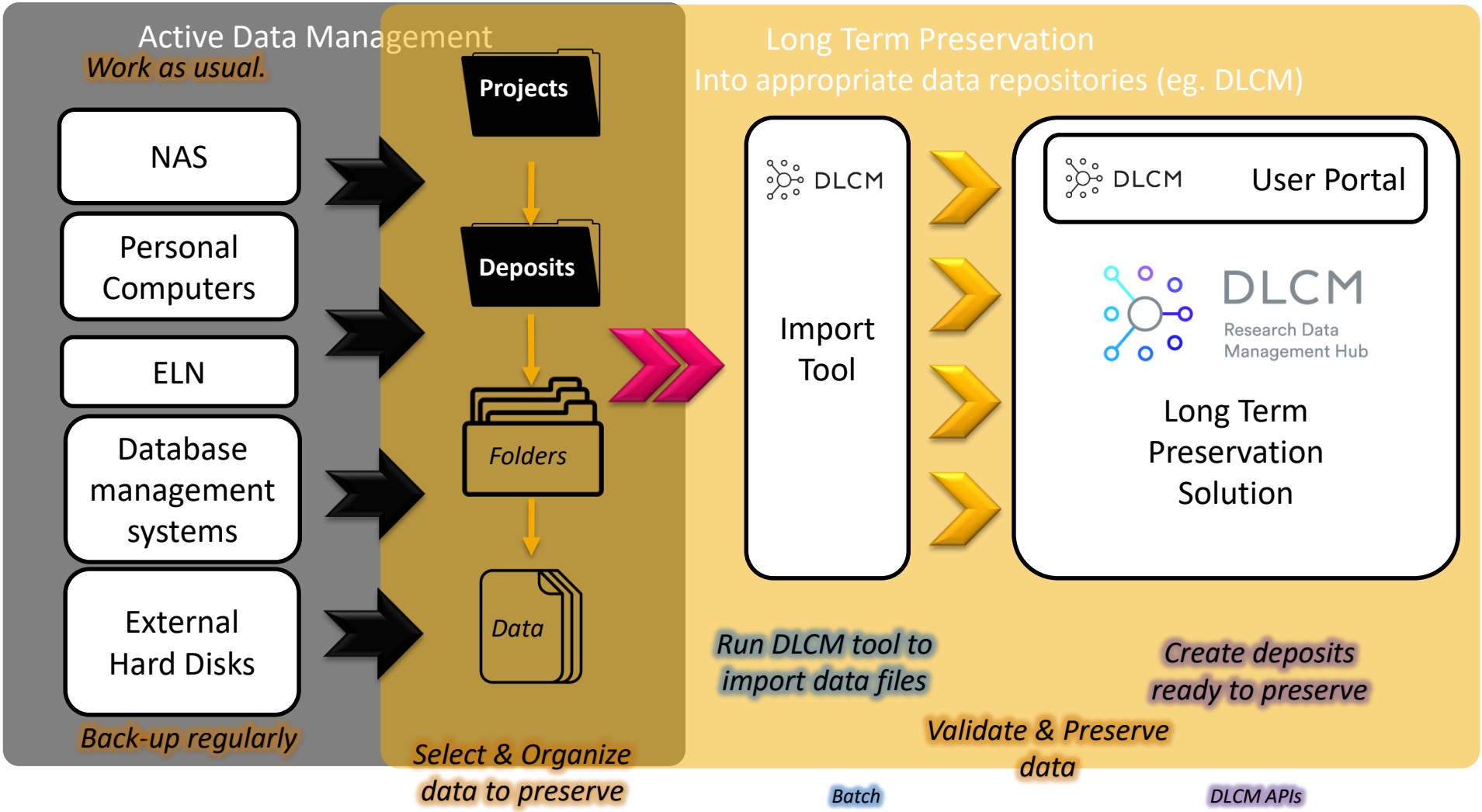
LOST HARDWARE



Your laptop AND your external hard drive got stolen!

Credit: Peter Murray-Rust, http://blogs.ch.cam.ac.uk/pmr/2011/08/01/why-you-need-a-data-management-plan/, August 2011, CC-BY

THE WHOLE PICTURE



Reproduced with permission of T. Mangelle, adapted

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 keep 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).

Back-up VS Archive

Backup

- Recoverable
- Copy
- Changing
- Short-term

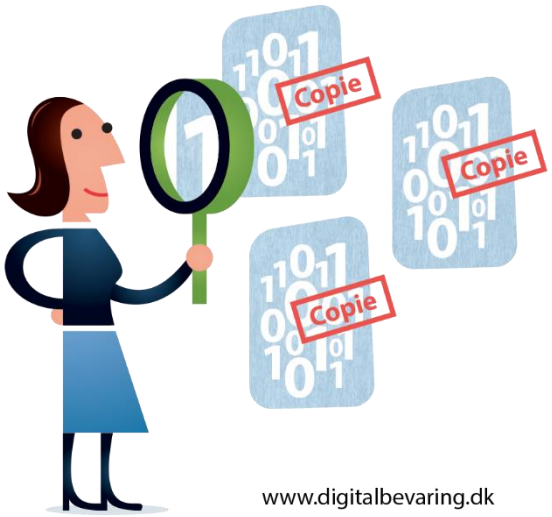
Archive

- Accessible
- Moved
- Inactive
- Long-term

https://www.youtube.com/watch?v=34zmiNectAY

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

SWITCHdrive

Store, synchronise, share and edit files quickly and reliably - all in the secure SWITCH cloud.

Facts at a glance

- Standard 50GB online storage per user
- Easy collaboration and protected access thanks AAI-Login
- Synchronize files & folders across multiple devices
- Access via desktop client, browser or mobile application

Your data at SWITCH

- All data is stored on SWITCH servers in Switzerland.
- Full compliance with Swiss data protection regulations
- No data and metadata exchange with other Office companies - not even for diagnoses or future extensions.

New collaboration functions thanks to OnlyOffice integration

- Real-time collaboration on shared documents - work with others on a document and track all changes live
- No local software installation required
- Access via desktop client, browser or mobile applicatic

SWITCHdrive

	SWITCH	for Business	
Space included	50 GB (can be upped to 250 GB)	Unlimited	15GB
Price (per annum)	Free	CHF15/mo/user	Free (2 TB is CHF10/mo)
File history	Yes	30 days	Yes
Where are files stored?	Within Switzerland	Within EU	Anywhere

SENDING FILES

SWITCH Services Stories About us

Services → Send 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.

Data Security

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

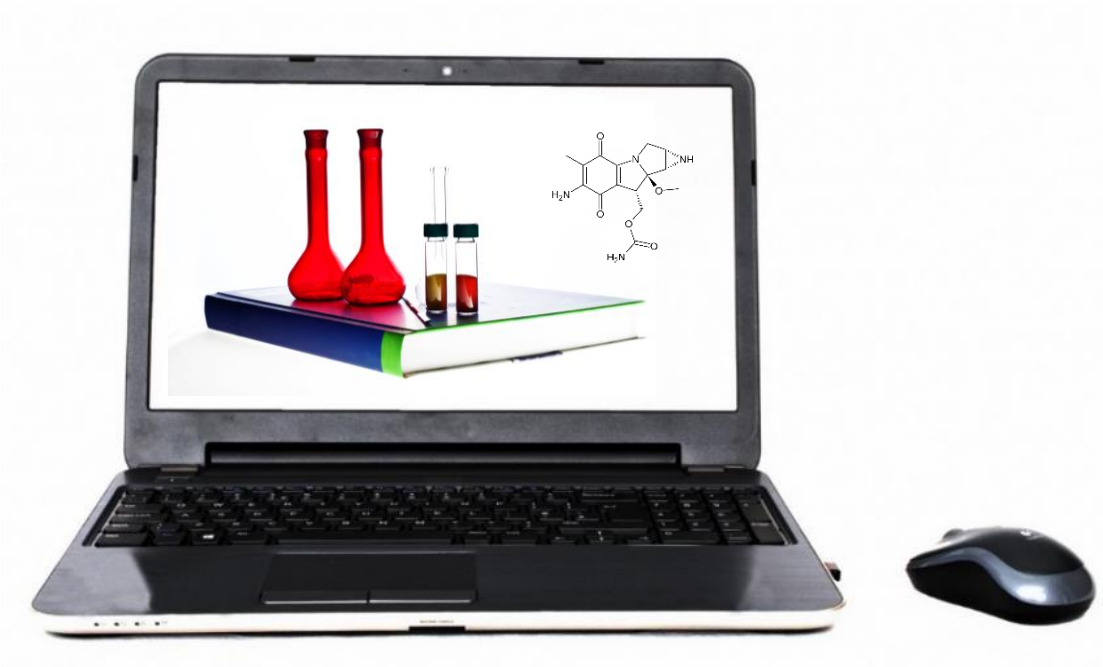
- Physical Access
- Digital access

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

NAS académique
Educloud (filr)
SWITCHdrive
SWITCHfilesender
Locally hosted ELN

Dropbox
Google Drive
Wetransfer
…

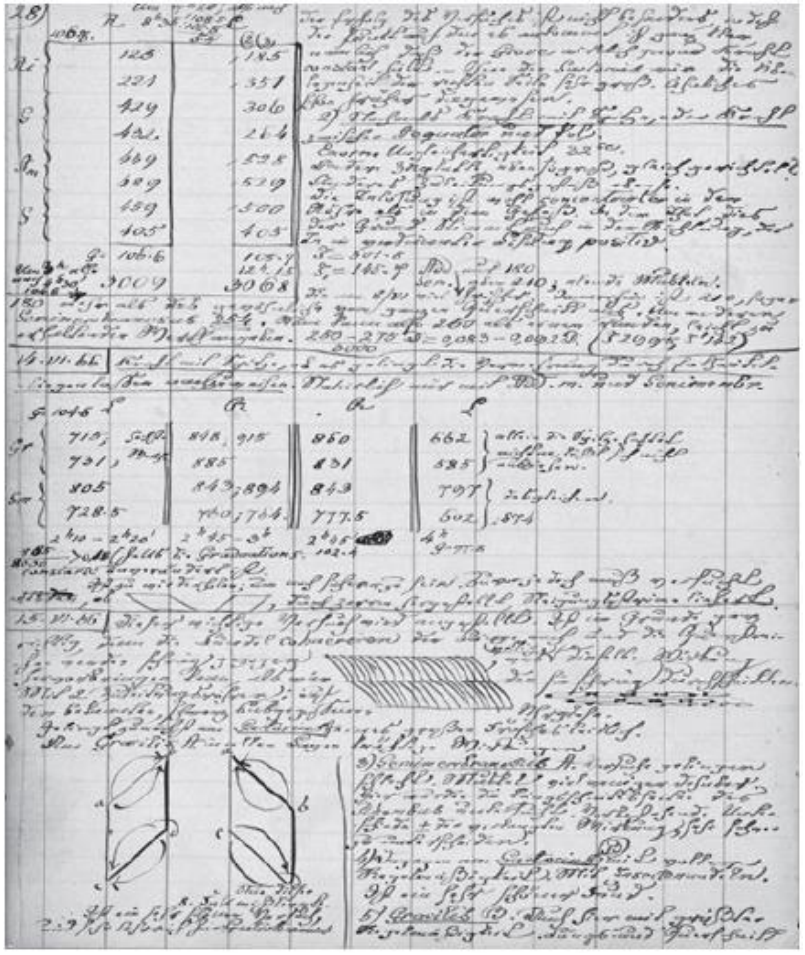
ELECTRONIC LABORATORY NOTEBOOKS



Another tool to save & share data & information within your lab

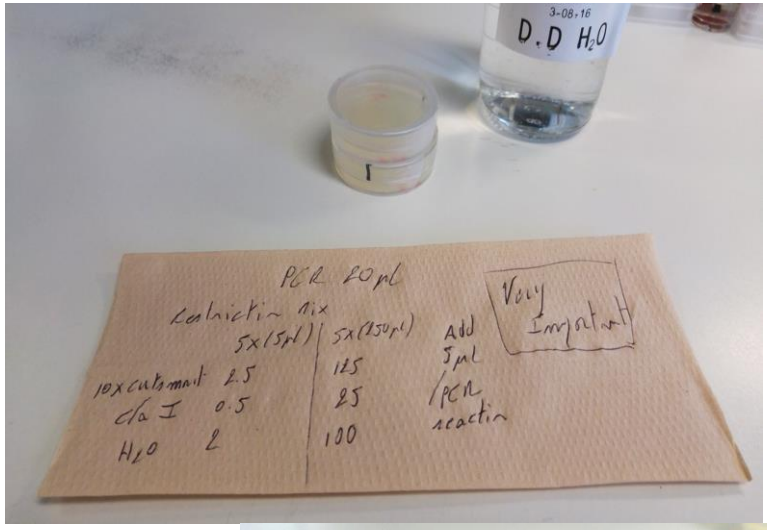
58

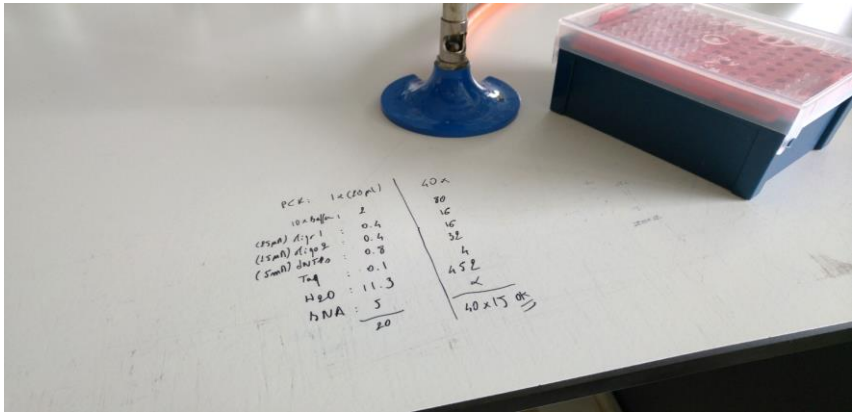
LABORATORY NOTEBOOKS



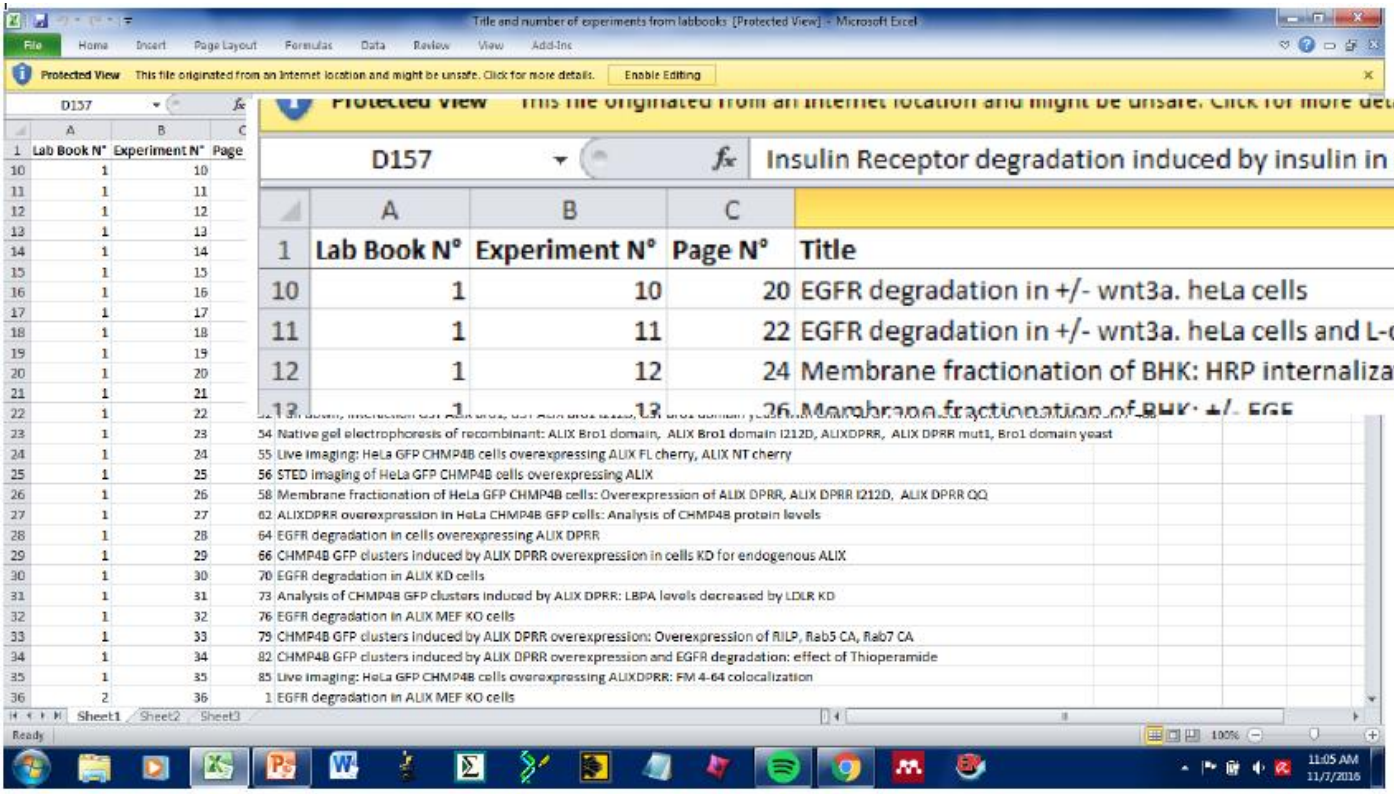
Dirnagl U, Przesdzing I. A pocket guide to electronic laboratory notebooks in the academic life sciences. F1000Res. 2016;5. doi:10.12688/f1000research.7628.1

ALTERNATIVES OF LAB NOTEBOOKS…





ALREADY TRANSITIONNING?



Title and number of experiments from labbooks [Protected View] - Microsoft Excel

Protected View This file originated from an Internet location and might be unsafe. Click for more details. Enable Editing

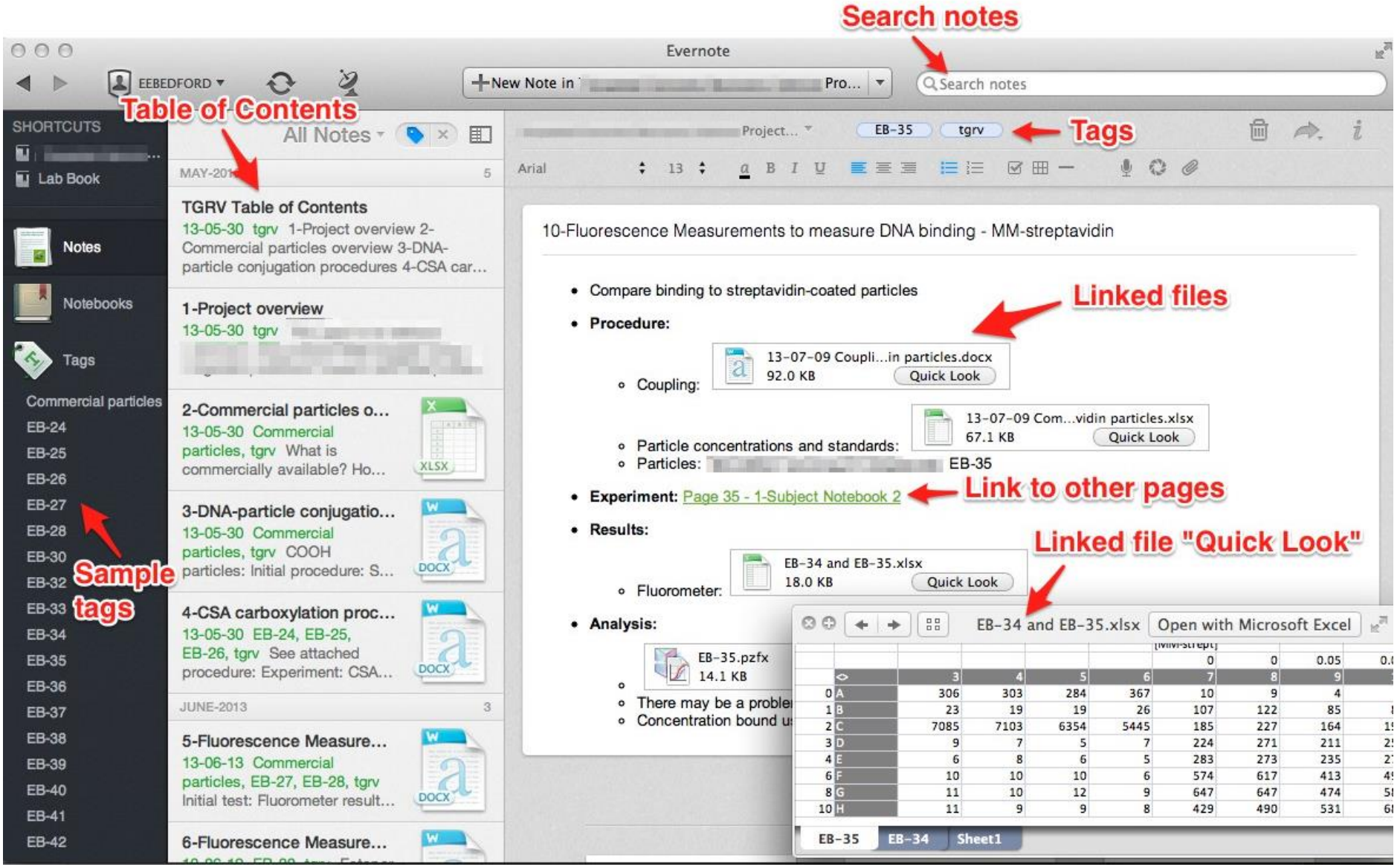
D157 — Insulin Receptor degradation induced by insulin in HeLa MZ. Analysis by WB. With and witho

	A	B	C	D
1	Lab Book N°	Experiment N°	Page N°	Title
10	1	10	20	EGFR degradation in +/- wnt3a. heLa cells
11	1	11	22	EGFR degradation in +/- wnt3a. heLa cells and L-cells
12	1	12	24	Membrane fractionation of BHK: HRP internalization
13	1	13	26	Membrane fractionation of BHK: +/- EGF

	A	B	C	
1	Lab Book N°	Experiment N°	Page	
10	1	10		
11	1	11		
12	1	12		
13	1	13		
14	1	14		
15	1	15		
16	1	16		
17	1	17		
18	1	18		
19	1	19		
20	1	20		
21	1	21		
22	1	22	54	Native gel electrophoresis of recombinant: ALIX Bro1 domain, ALIX Bro1 domain I212D, ALIXDPRR, ALIX DPRR mut1, Bro1 domain yeast
23	1	23	55	Live imaging: HeLa GFP CHMP4B cells overexpressing ALIX FL cherry, ALIX NT cherry
24	1	24	56	STED imaging of HeLa GFP CHMP4B cells overexpressing ALIX
25	1	25	58	Membrane fractionation of HeLa GFP CHMP4B cells: Overexpression of ALIX DPRR, ALIX DPRR I212D, ALIX DPRR QQ
26	1	26	62	ALIXDPRR overexpression in HeLa CHMP4B GFP cells: Analysis of CHMP4B protein levels
27	1	27	64	EGFR degradation in cells overexpressing ALIX DPRR
28	1	28	66	CHMP4B GFP clusters induced by ALIX DPRR overexpression in cells KD for endogenous ALIX
29	1	29	70	EGFR degradation in ALIX KD cells
30	1	30	73	Analysis of CHMP4B GFP clusters induced by ALIX DPRR: LBPA levels decreased by LDLR KD
31	1	31	76	EGFR degradation in ALIX MEF KO cells
32	1	32	79	CHMP4B GFP clusters induced by ALIX DPRR overexpression: Overexpression of RILP, Rab5 CA, Rab7 CA
33	1	33	82	CHMP4B GFP clusters induced by ALIX DPRR overexpression and EGFR degradation: effect of Thioperamide
34	1	34	85	Live imaging: HeLa GFP CHMP4B cells overexpressing ALIXDPRR: FM 4-64 colocalization
35	1	35	1	EGFR degradation in ALIX MEF KO cells
36	2	36		

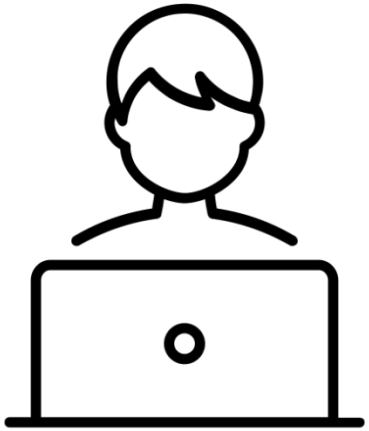
Sheet1 Sheet2 Sheet3 — Ready — 100% — 11:05 AM 11/7/2016

Screen shot from O. Schaad



Electronic Lab Notebook Unique Electronic Lab Notebooks Gradhacker Gallery. In: IslFamily.com [Internet]. 17 Aug 2009 [cited 1 Oct 2018]. Available: https://www.islfamily.com/electronic-lab-notebook.html/electronic-lab-notebook-unique-electronic-lab-notebooks-gradhacker-gallery

ADVANTAGES FOR YOU



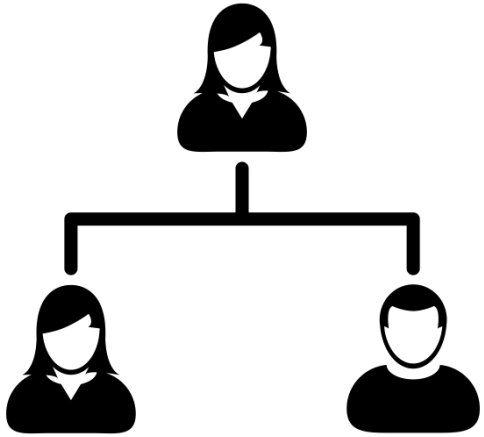
- Reduce data loss
- Insertion of all types of data and media

(images, movies, protein sequences…)

- Text-based searching
- Directly collect primary data from lab instruments
- Customized templates to facilitate data entry
- Revision history with digital timestamps → Improve data traceability
- Improve reproducibility of results

Workplace by kiddo from the Noun Project, https://thenounproject.com/

ADVANTAGES FOR YOUR COLLEAGUES



- Integrate metadata (data *about* the experiment, eg. date, temperatures…)
→ facilitates retrieval, analysis and re-use
- Keep in one place all the data for one project
- Archives kept in the lab after people are leaving the lab
- 24/24 access
- Provide a database for material discussed at group meetings
- Facilitate data management

Hazards of using an ELN

- Definition of access rights and restriction
- Copyrights
- Perceived complexity compared to using a paper notebook
- Too much standardization
- Long-term data storage
- Cost

Electronic Laboratory Notebooks

labarchives
Chance Favors the Organized Lab

hivebench

Benchling

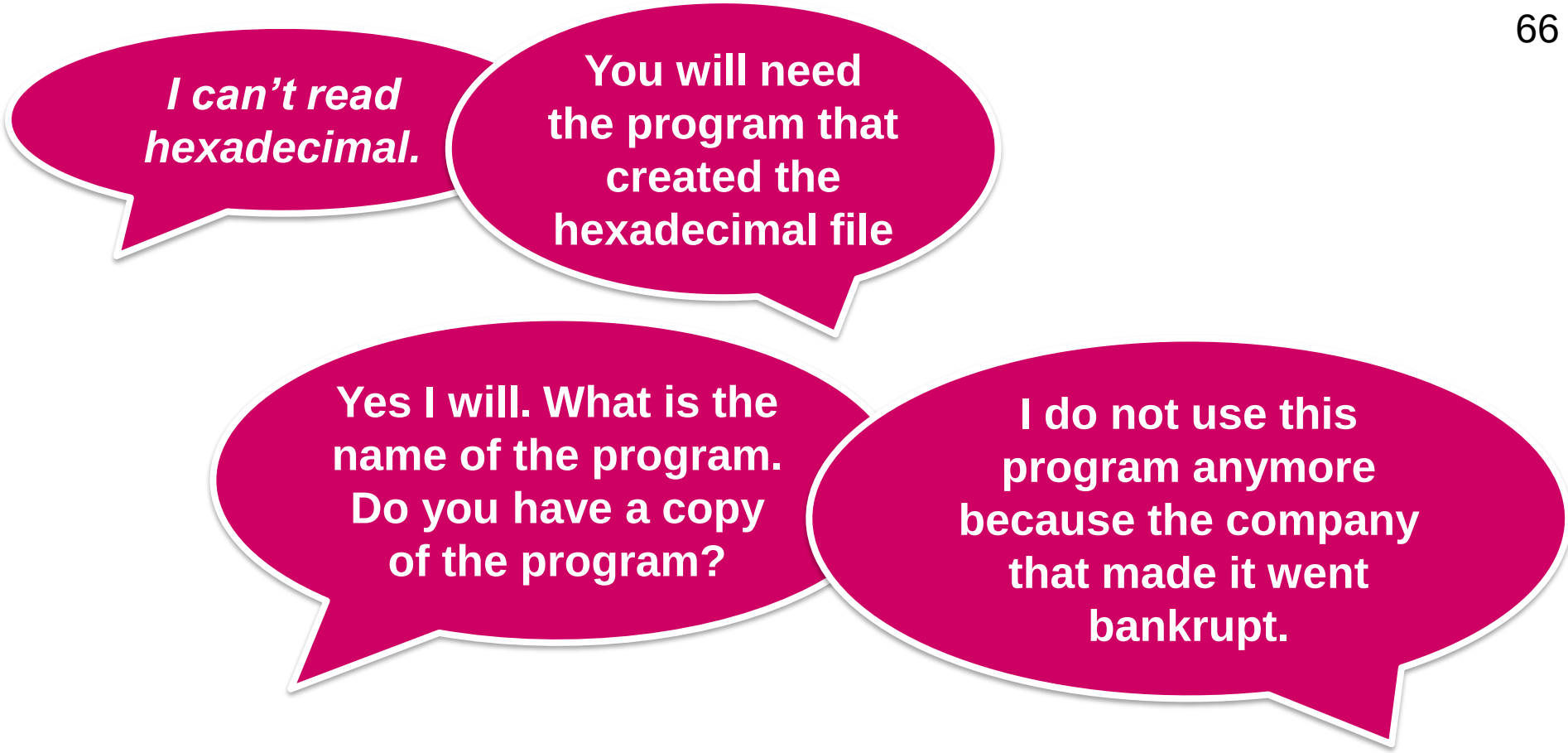
RSpace

SCINOTE

LabTrove

eLabFTW

WordPress



Let's talk about

- *Data conversion: open format for preservation*
- *Preserving the software alongside the data ?*

KEEP RAW DATA RAW

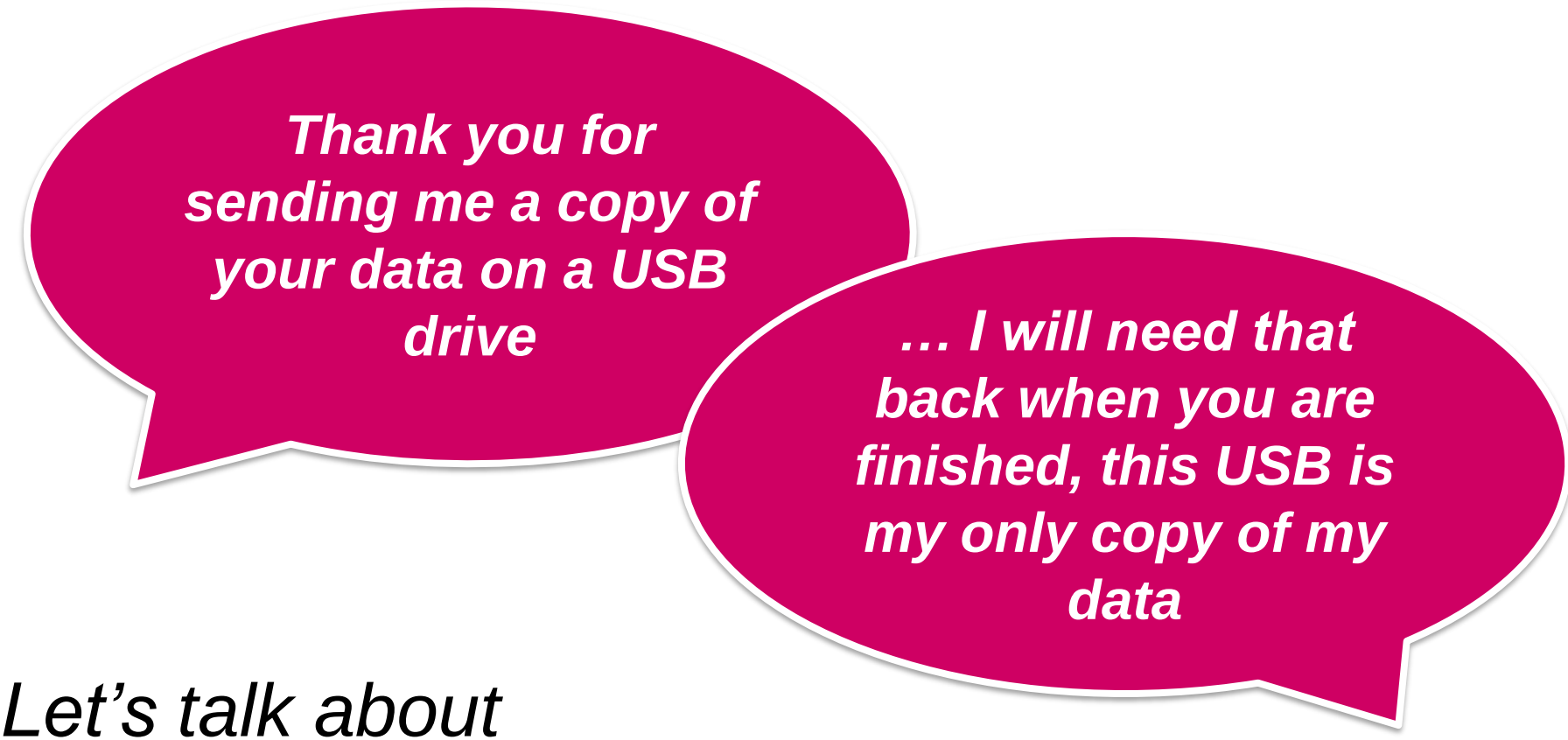
- Allows **new analysis** with old data and **transparency**

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

BUT PRESERVE DATA IN 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:
  .odt rather than .docx
  .csv rather than .xlsx
  .svg rather than .ai
  PDF/A rather than PDF

See Guidance from ETH, UK data service, Harvard



Let's talk about

- *Archiving & sharing*
  - *Caveat about special data*
- *Repositories & Licenses*

BE CAUTIOUS WITH 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;"

  (**Swiss Federal Act on Data Protection**)

- Confidential / «critical» data

Processing of personal data ➡ Cantonal Data Protection Law

Federal Act on Data Protection (FADP)

Swiss Federal Act on Data Protection https://www.admin.ch/opc/en/classified-compilation/19920153/index.html

ETHICS AND LEGAL POLICIES

Research with human subjects

Federal Act on Research involving Human Beings (HRA)

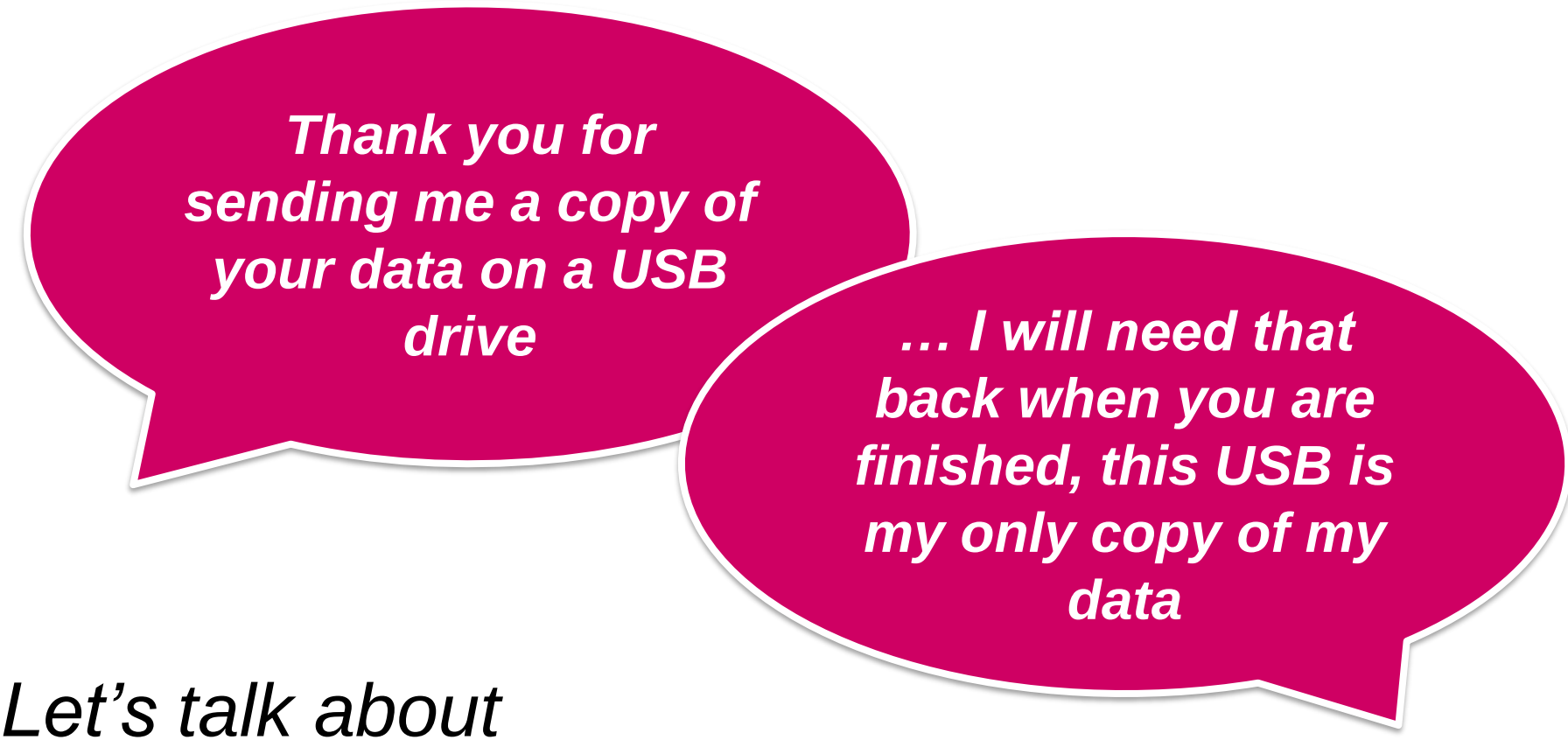
Applies to research on human diseases and on the structure and functioning of the human body, carried out:

a. on people;
b. on deceased people;
c. on embryos and fetuses;
d. on biological material;
e. on personal data related to health.

→ **Cantonal Ethics Commission (CCER)**
www.ge.ch/commission-cantonale-ethique-recherche-ccer

Does not apply to the research carried out:

a. on *in vitro* embryos (within the meaning of the Federal Act of 19 December 2003 on embryonic stem cell research)
b. on **anonymous** biological material;
c. on health-related data that have been collected **anonymously or anonymized**.



Let's talk about

- *Archiving & sharing*
  - *Caveat about special data*
- *Repositories & Licenses*

UNIVERSITÉ DE GENÈVE

UNIVERSITÉ FACULTÉS ÉTUDIANTS COLLABORATEURS SERVICES ALUMNI

Home Presentation [redacted] People Publications Data

Open Access Data

Texte intégral

Article (Published version) (665 Kb) - Accès libre
Autre version: https://academic.oup.com/hmg/article/27/4/732/4710066
Jeu de données: https://www.ebi.ac.uk/ega/studies/EGAS00001000805
Jeu de données: https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1866/

Texte intégral

Article (Published version) (3 MB) - Accès restreint UNIGE
Jeu de données: https://doi.org/10.5281/zenodo.1474531

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The dataset is available from https://doi.org/10.5281/zenodo.1474531.

Time-locked inhibition of OFC
Plasticity at OFC–striatum synapses
Bidirectional shift in synaptic strength
Discussion
Methods
Data availability
References
Acknowledgements

Abstract
Introduction
Results
Discussion
Materials and methods
Supporting information
Acknowledgments
References

Reader Comments (0)
Media Coverage (0)
Figures

Supporting information

pone.0208371.s003.xlsx

	A	B	C	D	
1	Sample	Reads	Clipped_reads	Tophat_mapped	Ma
2	C57_D300_F1_1	62370005	48072079	41581189	0.1
3	C57_D300_F1_2	61289415	50599210	44084120	0.1
4	C57_D300_F1_3	64705541	52738583	45347087	0.1
5	C57_D300_F1_4	73423174	58363679	48878978	0.1
6	C57_D300_F1_5	60720003	52309334	45704724	0.1
7	C57_CTL_F1_1	43268415	36994852	30080705	0.1
8	C57_CTL_F1_2	68710604	59640671	50884558	0.1
9	C57_CTL_F1_3	65719341	57936661	48286489	0.1

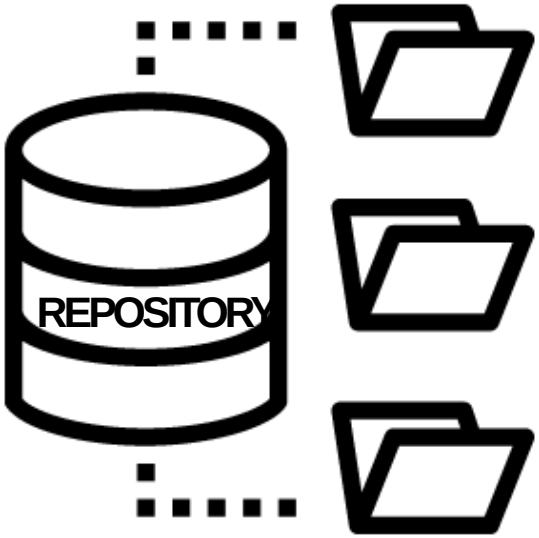
Feuil1 Feuil2 Feuil3

figshare 3 / 5 Download

(XLSX)

S1 Fig. Read lengths distribution across samples.
A Peaks specific for sperm RNAs were identified. The peak at 22 nucleotides is specific to the size of mature microRNAs. The peak at 32 nucleotides is specific to the size of tRNA-derived small RNAs. The peak at 50 bp involves mainly coding RNAs.
https://doi.org/10.1371/journal.pone.0208371.s001
(TIF)

ARCHIVING AND DISSEMINATION



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

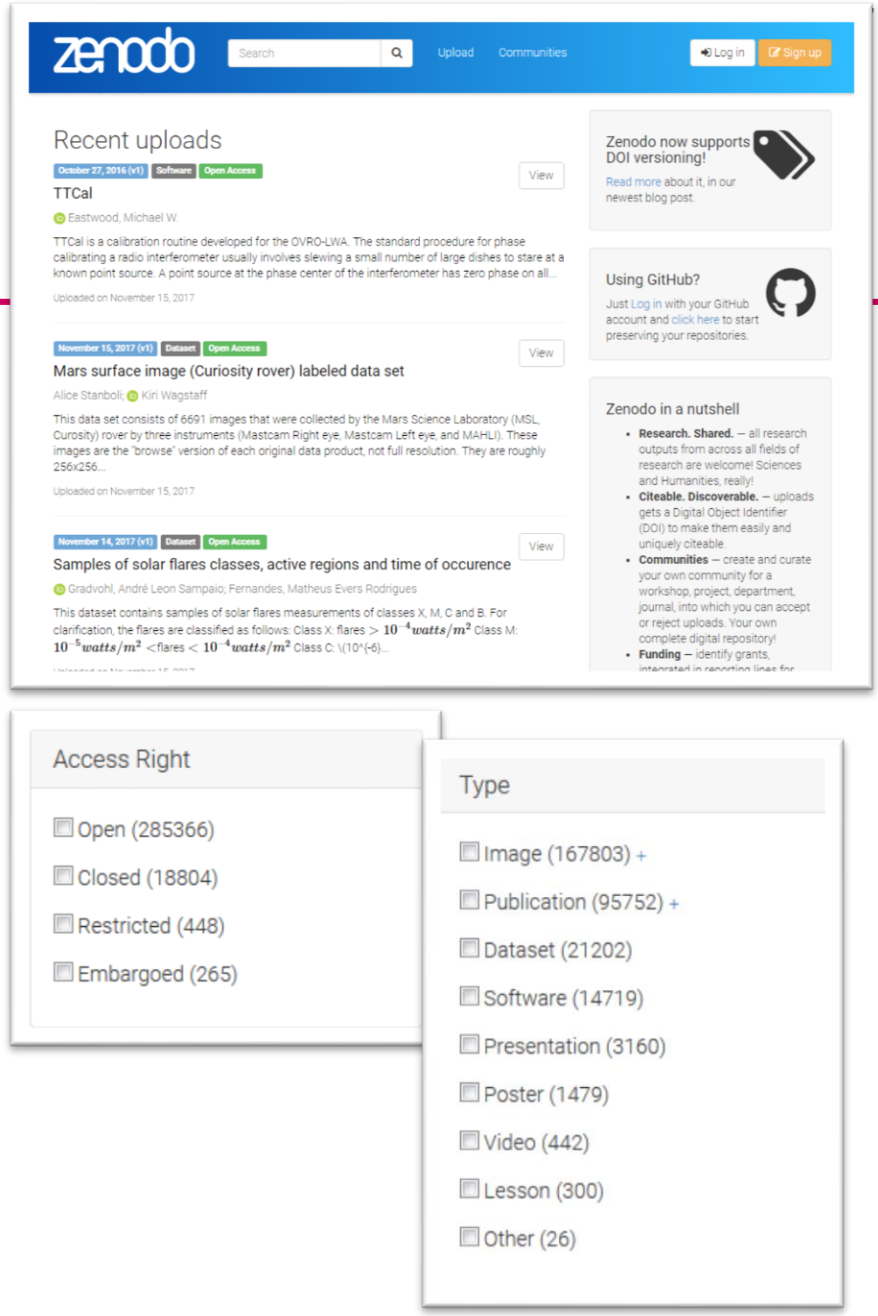
- Journal supplementary material service
- Departmental, project or personal web page

A LOCAL REPOSITORY : YARETA

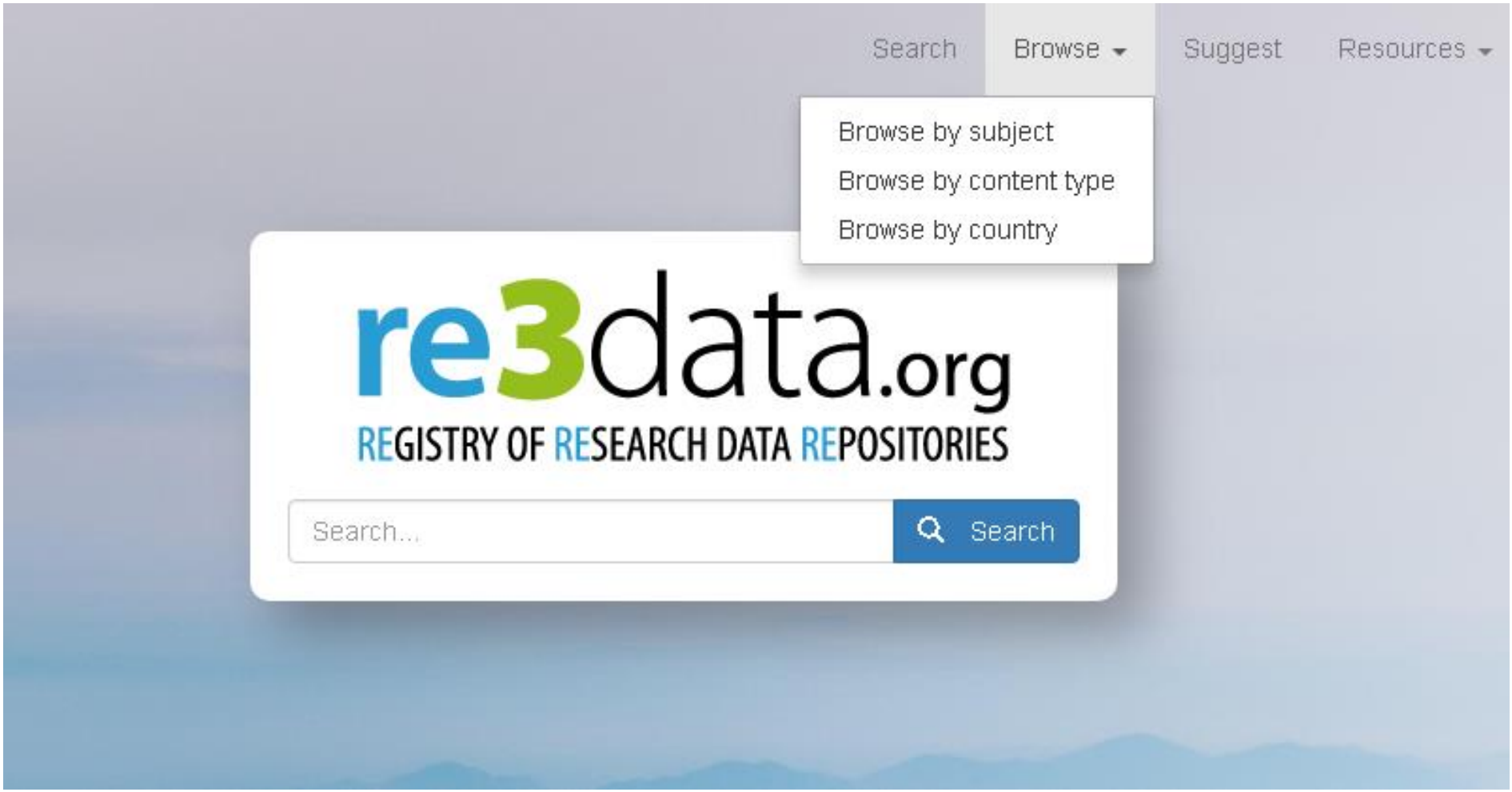
- At Geneva (canton): https://yareta.unige.ch
- Local solution (output from the www.dlcm.ch project)
- Servers (main and copy) in Switzerland
- Compliant with FAIR principles
- Provides DOI to datasets (Digital Object Identifier)
- Long-term preservation solution
  → You define the duration of retention of the data
- "Open" by default, restricted if required
- Fair pricing
  - Free up to 50 Go, then: **100 CHF / Tb / Year**

ZENODO

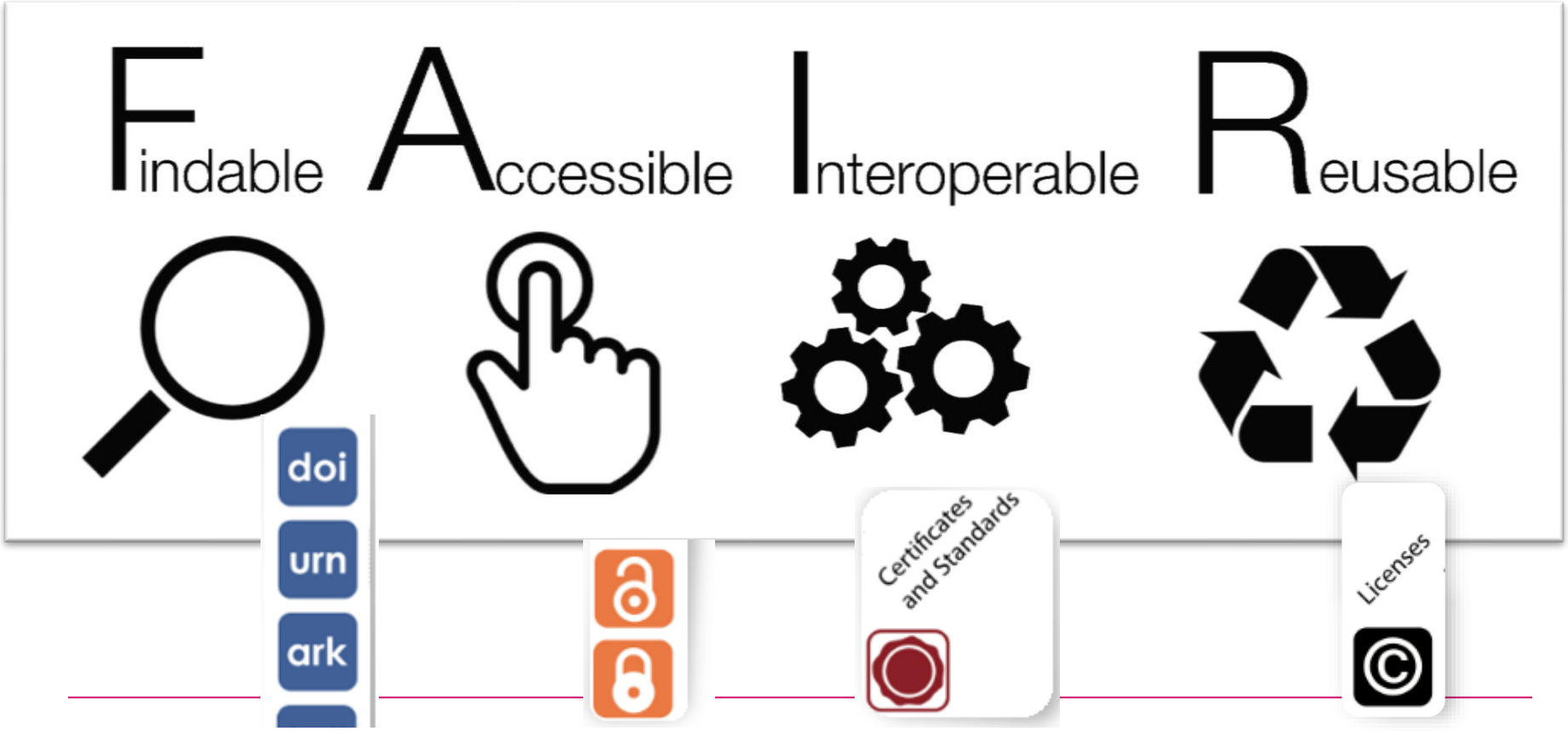
- https://www.zenodo.org/
- The OpenAIRE project was commissioned by the EC to support a catch-all repository for EC funded research
- CERN, partner of OpenAIRE, provides this tool
- **Zenodo** launched in 2013
- Free access and deposit



A REGISTRY OF REPOSITORIES



IS MY REPOSITORY FAIR ?



At least one blue icon

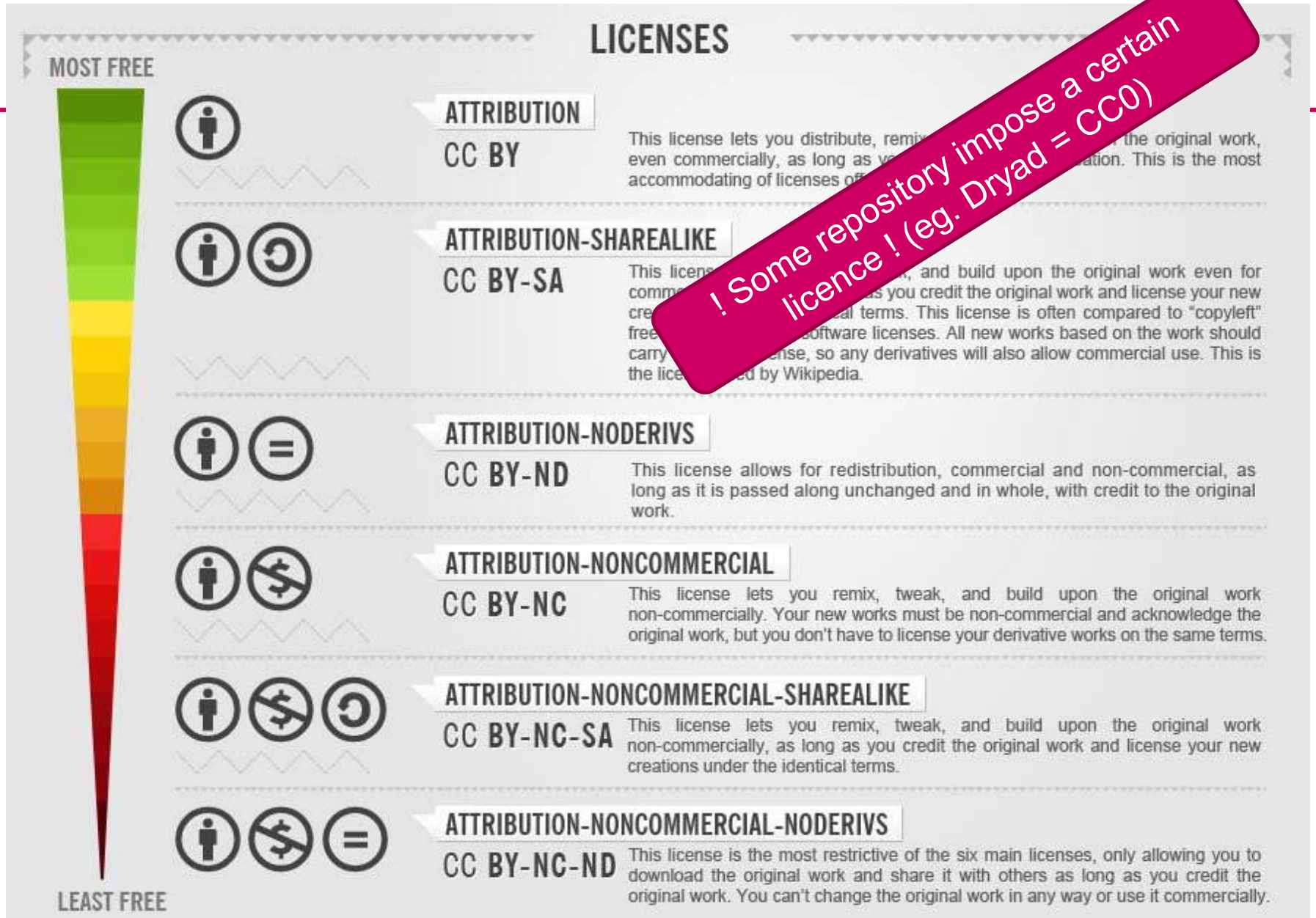
An orange icon with the lock

The «Standards» icon is a bonus

A «Licence» icon

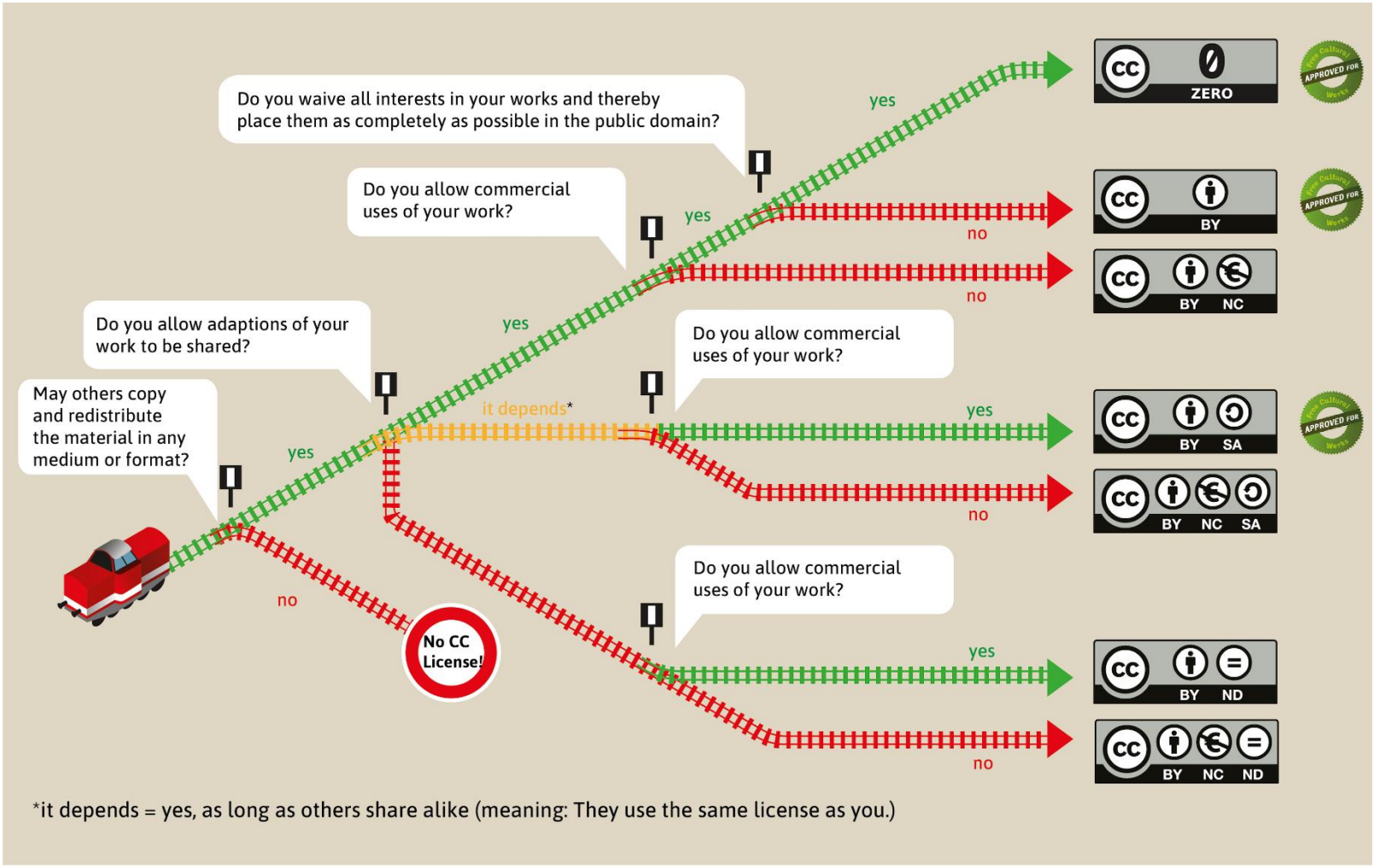
http://www.datafairport.org/

Creative Commons Licenses

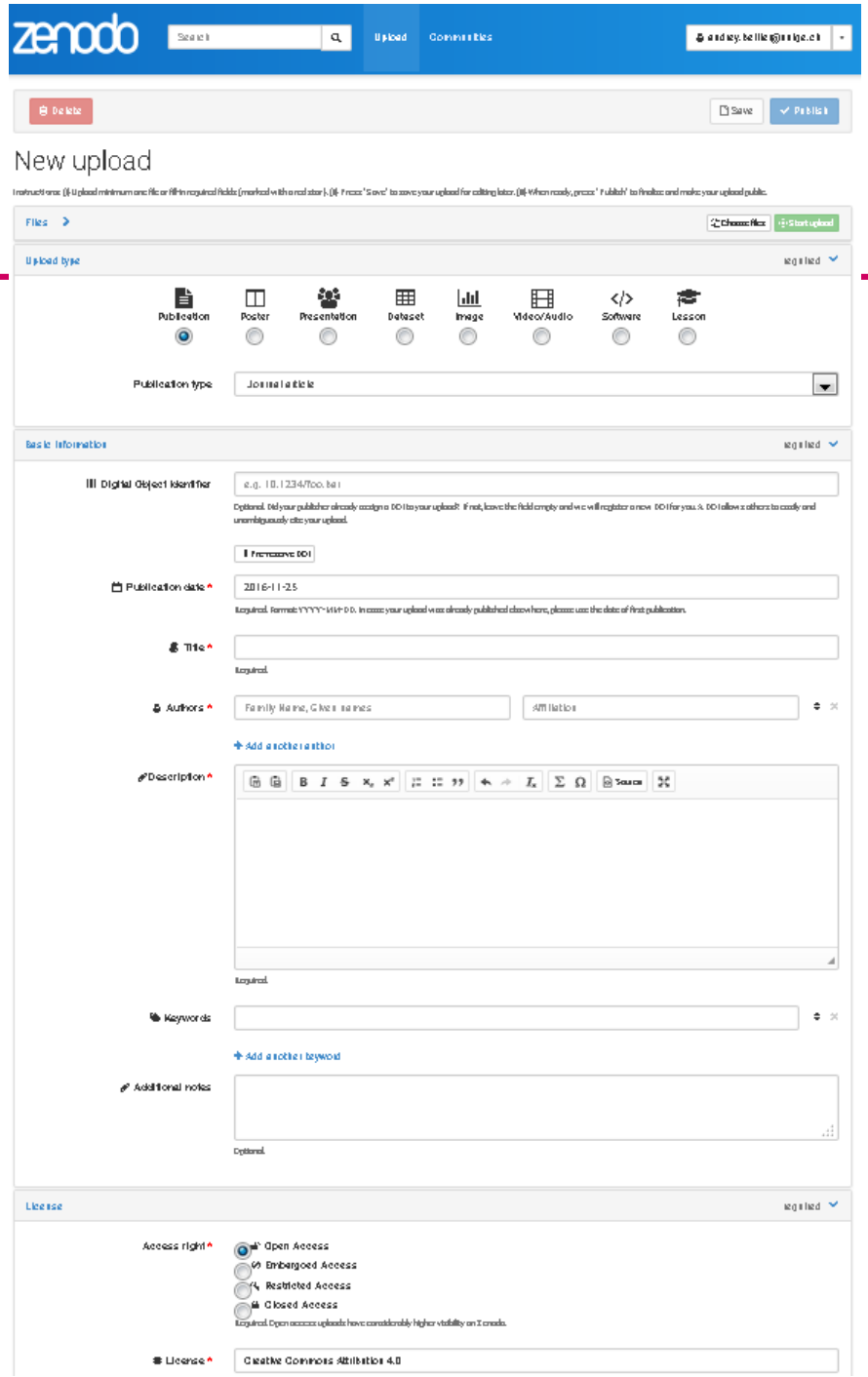


! Some repository impose a certain licence ! (eg. Dryad = CC0)

CREATIVE COMMONS



ZENODO SANDBOX VERSION



https://sandbox.zenodo.org/

If you want to try making a dumb deposit

Can I use your data ?

✓ ***Context***: to be able to understand it

- *Cf.* ***FAIR*** *data*

✓ ***Permission/License***

- ***permission*** *from data subject & ethics committee*
- ***license*** *given by the data owner*

? To whom does research data belong ?

✓ ***Access***

- *Data available on request VS Repositories*

TAKE HOME MESSAGE

- Check for **legal** and **regulatory** requirements
- **Anticipate** your data needs (volume, sensitivity, formats,…) and select apropriate options accordingly
- Ensure your data can be **reused** (open formats, proper naming, add metadata…)
- UNIGE solutions: **NAS**, **SWITCH, Yareta**

Give access to your data

- ✓ Open Access journals (Module 2)
- ✓ Preprints servers (Module 2)
- ✓ Research data repositories
- ✓ Data papers

What will happen to your shared data?

Adapted from Closed Data… Excuses, Excuses. In: UC3 Data Pub Blog [Internet]. 24 Apr 2013 [cited 28 Nov 2016]. Available: https://datapub.cdlib.org/2013/04/24/closed-data-excuses-excuses

DEFINITION

A data paper is a journal publication whose primary purpose is to describe data, rather than to report a research investigation.

Chavan, Vishwas, et Lyubomir Penev. 2011. « The Data Paper: A Mechanism to Incentivize Data Publishing in Biodiversity Science ». *BMC Bioinformatics* 12 Suppl 15: S2. doi:10.1186/1471-2105-12-S15-S2.

TERMINOLOGY

Data article

Data descriptor

Data in brief

Data note

Data original article

Data paper

Database article

Database paper

Dataset paper

IDENTIFY DATA PAPERS' KEY CONTENTS

- ✓ How different is it from a traditional journal article ?
- ✓ What strikes you as specific to data papers ?
- ✓ Can you easily access the data sets, having no further info than this article?

KEY COMPONENTS

General structure

- Titre
- Authors, affiliations
- Abstract
- Keywords
- Context (spatial coverage, temporal coverage)
- Methods
  - Steps, sampling strategy, quality control, contraints, ethical considerations
- Dataset description
  - Object names, data type, format names & versions, creators, creation dates, language, license, location (DOI), publication date
- Reuse potential
- Acknowledgements
- References

Data Paper

Data citation

Article citation

- **Upload in the journal website or in a repository**
- **In open access or with temporary restriction**

Dataset

Adapted from Open Science : gestion et partage des données de la recherche / Esther Dzale & Dominique l'Hostis

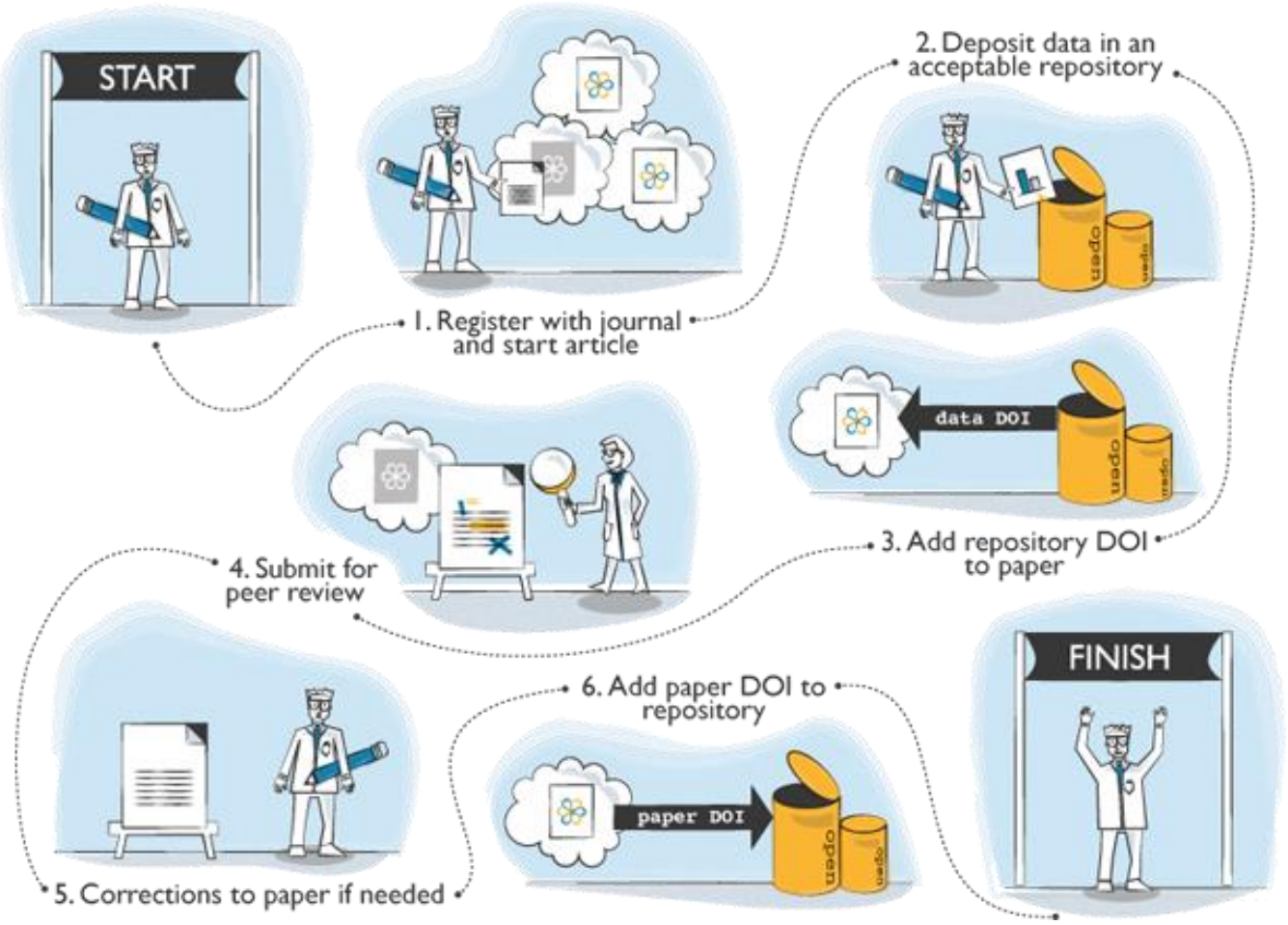
BENEFITS

- ✓ Provide a **citable journal publication**
  - ✓ Academic credit for your data
  - ✓ Paper included in PubMed, Web of Science,…
  - ✓ Ability to trace usage and citations of your data
- ✓ Describe your **data in a human-readable form**
  - ✓ To ensure it's well understood and reused appropriately
- ✓ Increase the **visibility** of the data you deposit in a repository
  - ✓ Opportunities for collaboration
  - ✓ Opportunity for reuse in other research context/fields
  - ✓ Avoid duplication of experiences/measurements (reduce costs)
- ✓ Get your data **peer reviewed**
- ✓ **Interlink** your research outputs to accelerate their **dissemination**
- ✓ Long term **archiving**

Drawbacks

- ✗ Time-consuming
- ✗ Costs
- ✗ Ethical concerns (anonymization, consent,…)
- ✗ Select appropriate license and embargo

How to submit a data paper

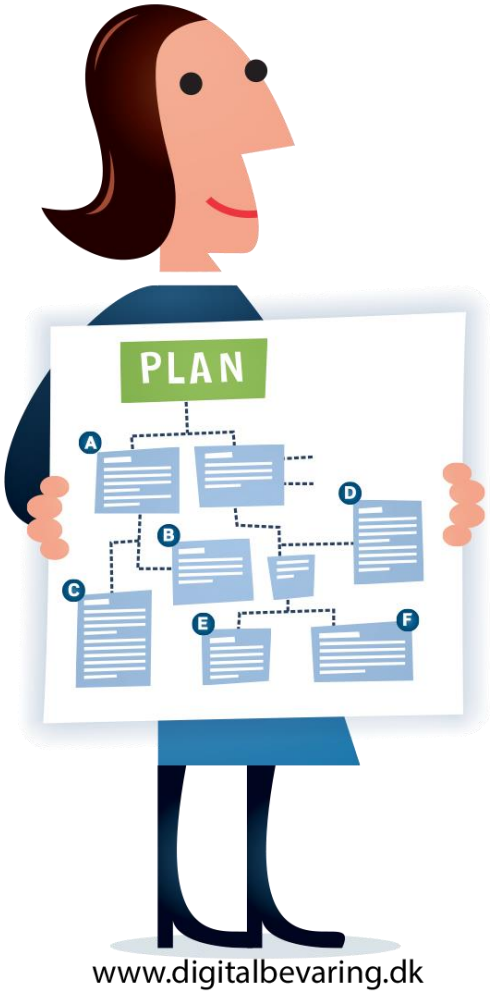


http://openhealthdata.metajnl.com/about/submissions/

DATA MANAGEMENT PLAN (DMP)

98

Data Management Plan



- Saves time
- Increases the impact of your research
- Participates in open science
- Facilitates new discoveries

EXAMPLES OF FUNDERS WHICH REQUIRE DMPS OR EQUIVALENT

REQUIREMENTS FROM SNF

mySNF

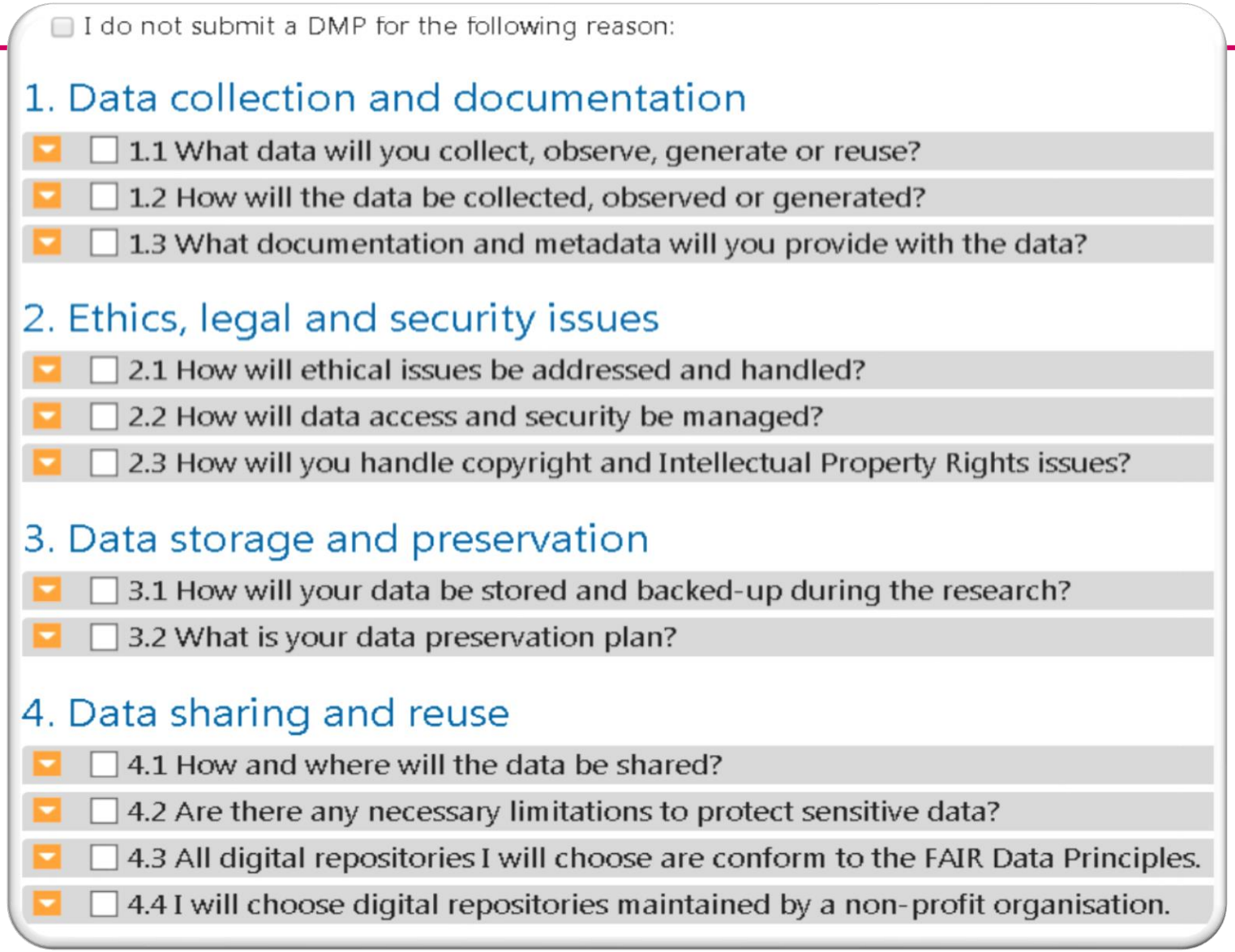
2. Application data

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

.. Données relatives à la requête

#	No/titre
2.1	Données de base I
2.2	Données de base II
2.3	Projet orienté vers l'application
2.4	Resoumission
2.5	Relation avec d'autres projets FNS
2.6	Moyens disponibles ou requis
2.7	Haute école
2.8	Collaboration (nationale et internationale)
2.9	Moyens additionnels
2.10	Besoin financier
2.11	Plan de gestion des données (DMP)
2.12	Recherche exigeant des autorisations ou des annonces

THE DMP ON MYSNF



1. Data collection and Documentation

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

→ *Type*

→ *Format*

→ *Volume*

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

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 in spreadsheets will be stored as .csv Data in 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.

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

Qualitative (e.g. phenotyping data for cell lines), quantitative (e.g. cell counts), mass spectrometry and image data will be generated. Raw data will be analysed and expressed as graphs, tables and annotated images.

Data generated will be in various formats and sizes of datasets. They include:

1. Cell images e.g. phase and fluorescence, and electron micrographs (~5,000 images over project). Software used includes OpenLab, Softworx and IN Cell Investigator, with data saved as software-specific files e.g liff and lg3 files, as well as generic formats such as jpeg, tiff etc.
2. mass spectrometry spectra (from <50 samples). MS data will be analysed using Bruker Data Analysis or Thermo Excalibur software (generating xml and raw files) and proteins will be matched to the *T. brucei* genome dataset using the Matrix Science Mascot search engine. Each LC-MS data file is between 1-2GB.
3. Cell line phenotyping data including growth curves and DAPI counts (Excel and GraphPadPrism files) and flow cytometry data (FlowJo and jpeg/tiff files) (~200 data sets).
4. …

(source)

1. Data collection and Documentation

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

→ *Standards*

→ *Methods*

→ *Files naming and versioning*

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

Giemsa-stained 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

☐ 1.3 What documentation and metadata will you provide with the data?

→ *Information required for future users*

→ *Annotations*

1.3 What documentation and metadata will you provide with the data?

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?

→ *Data protection*

→ *Confidentiality agreement*

→ *Personal and sensitive data management*

2.1 How will ethical issues be addressed and handled?

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, VitalIT)

Lettres | STM 1 | SSH | STM 2 | Suite

UNIGE GUIDELINES

General ethical charter

c) Each individual engaged in research is responsible for ensuring integrity in his or her quest for knowledge and in the interpretation of results.

Research integrity directive

Le responsable de projet doit veiller à ce que les données de base soient conservées en sécurité pendant cinq ans au moins après le terme de la recherche.

Federal Act on Research involving Human Beings (HRA)

Commission cantonale d'éthique de recherche (CCER)

Commission universitaire d'éthique (CUREG)

Faculty Committees:

Psychologie et Sciences de l'éducation
Traduction et Interprétation
Sciences de la Société

2. Ethics, legal and security issues

2.2 How will data access and security be managed?

→ *Data security*

→ *Data access rights and permissions*

2.2 How will data access and security be managed?

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.

2.2 How will data access and security be managed?

The main risks to data security are loss or damage to laboratory notebooks and loss or corruption of electronic data. Data will be safeguarded by the following measures:

1. Data in lab notebooks will, as described above, also be recorded in electronic form that is backed up daily to secure against loss or damage of the notebook.
2. Access to electronic data (prior to publication as described above) will be limited to the members of the research group and relevant collaborators via limiting access to shared drives on the University server.
3. Access to laboratories and offices are controlled by card access to reduce the likelihood of malicious loss/damage; all computers used in this project will run Standard Staff Desktop, whereby firewalls and antivirus software are automatically upgraded and secure remote access to data is enabled; staff will lock their workstation whenever they are away from it.

(source of this eg.)

2. Ethics, legal and security issues

☐ 2.3 How will you handle copyright and Intellectual Property Rights issues?

→ *Who is the owner of the data?*

→ *Which licence to apply?*

2.3 How will you handle copyright and Intellectual Property Rights issues?

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.

2.3 How will you handle copyright and Intellectual Property Rights issues?

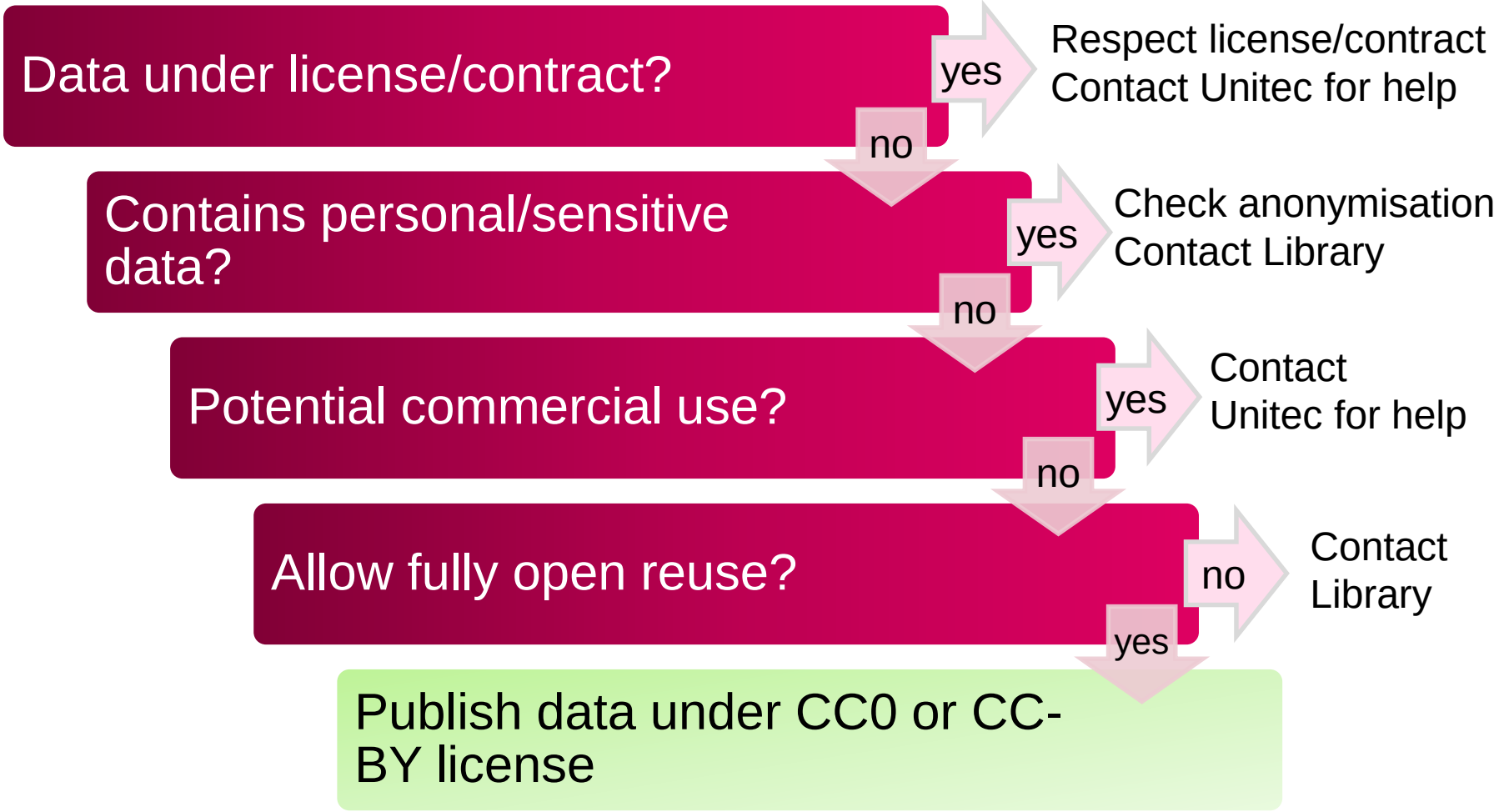
All data is the property of the sponsor institution (HUG), and is under the direct responsibility of the sponsor.
Intellectual Property Rights are defined as follows:

This project is being carried out in collaboration with an industrial partner. The intellectual property rights are set out in the collaboration agreement. The intellectual property generated from this project will be fully exploited with help from the institutional Technology Transfer Office. The aim is to patent the final procedure and then publish the work in a research journal and to publish the supporting {anonymised data / metadata} under an open Creative Commons Attribution (CC BY) license

(eg from DMP Model, HUG CRC)

Lettres | STM 1 | SSH | STM 2 | Suite

CHOOSE YOUR LICENSE



See also: https://creativecommons.org/choose/

3. Data storage and preservation

☐ 3.1 How will your data be stored and backed-up during the research?

→ Storage capacity

→ Storage facility

→ Back-up procedures

3.1 How will your data be stored and backed-up during the research?

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.

Lettres | STM 1 | SSH | STM 2 | Suite

3. Data storage and preservation

3.2 What is your data preservation plan?

→ Procedures to select data to be preserved

→ File formats for preservation

3.2 What is your data preservation plan?

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?

→ Which repository to select

→ How others will find out about your data?

4. Data sharing and reuse

☐ 4.2 Are there any necessary limitations to protect sensitive data?

→ Under which conditions will the data be made available?

→ Timing of data release

→ Delay if applicable

☐ 4.2 Are there any necessary limitations to protect sensitive data?

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)

Lettres | STM 1 | SSH | STM 2 | Suite

4. Data sharing and reuse

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

→ ☐ *[checkbox]*

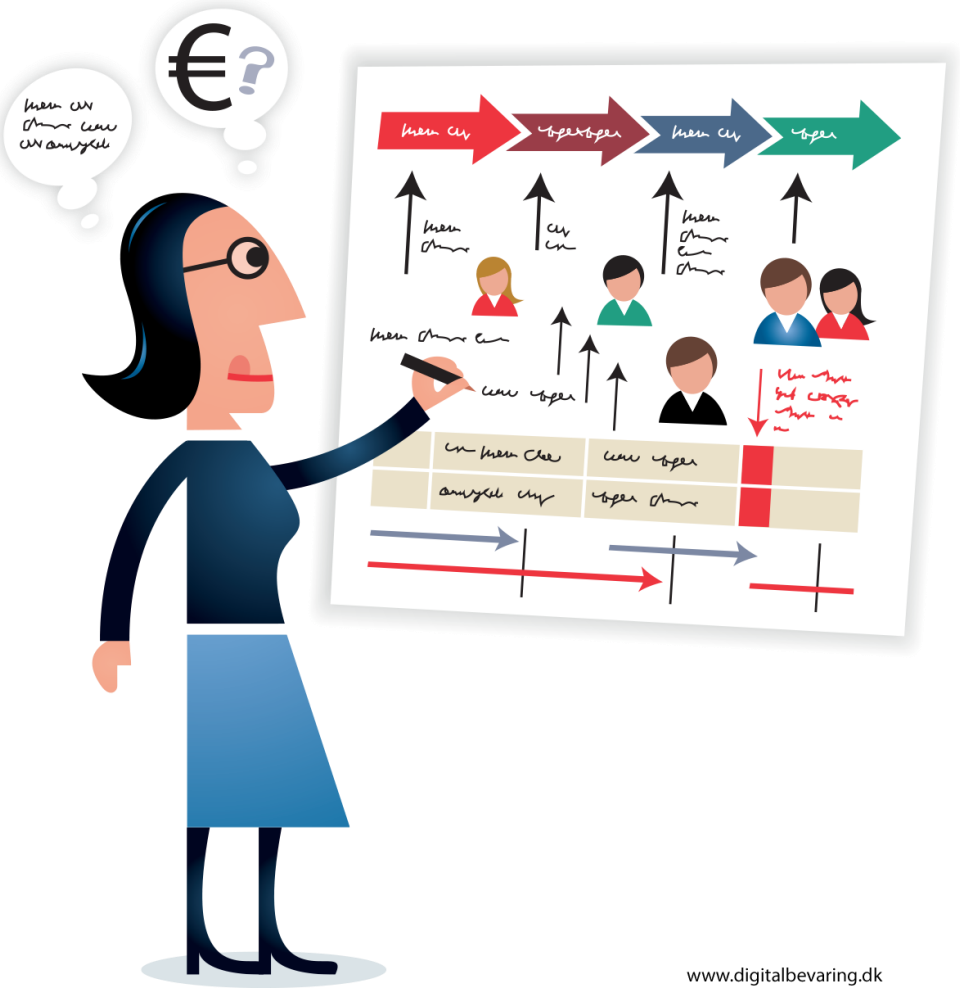
☐ 4.4 I will choose digital repositories maintained by a non-profit organisation.

→ *yes* / no *[radio button]*

TAKE HOME MESSAGE

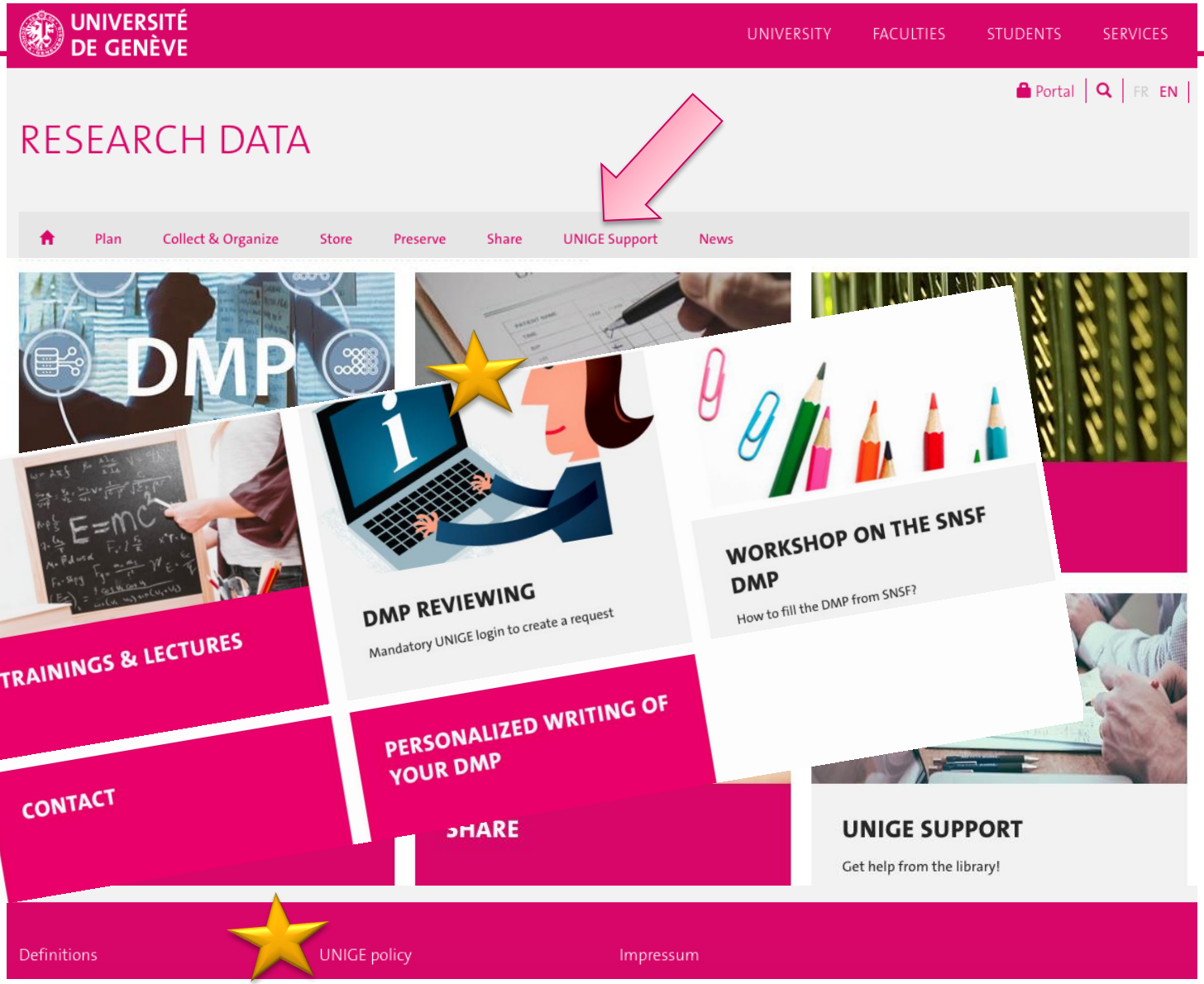
- ✓ A “plausible” DMP is a condition for the release of the funds
- ✓ DMPs are editable throughout the funding period (must be updated)
- ✓ Once SNSF funding has ended and the final scientific report has been approved, the DMP cannot be modified anymore
- ✓ The DMP is shared on P3 (SNSF’s public database) at the end of the project

WRITE YOUR OWN DMP



www.digitalbevaring.dk

WWW.UNIGE.CH/RESEARCHDATA

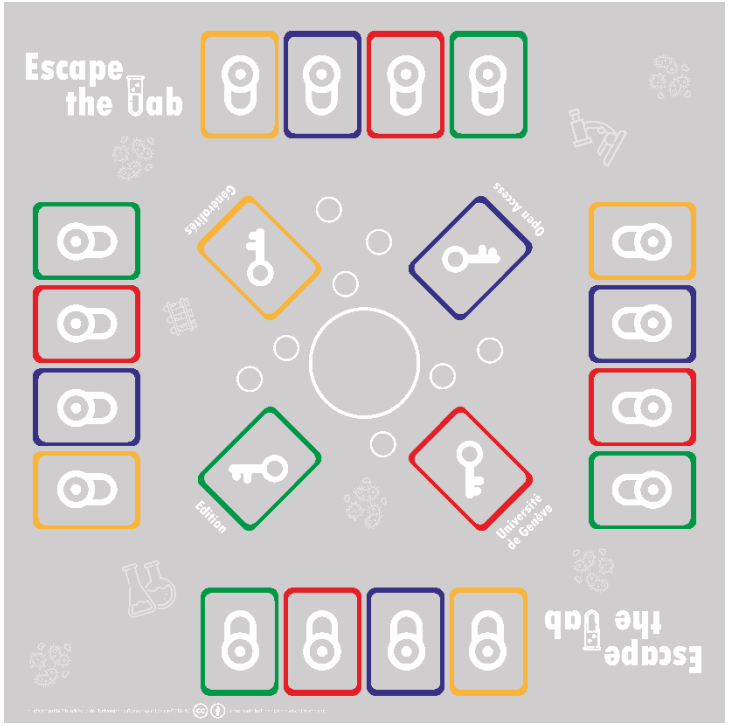


Escape the Lab

1. 4-5 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

Audrey.Bellier@unige.ch

Vincent.Huber@unige.ch

All images not cited here are in public domain