Research Data Management Training for vmITET

Educational Material

Author(s):
Petrus, Ana; Töwe, Matthias

Publication date:
2017-05-17

Permanent link:
https://doi.org/10.3929/ethz-b-000296500

ETH BIBLIOTHEK

Research Data Management Training for vmITET

Ana Sesartic & Matthias Töwe
Digital Curation Office

vmitet career event
17. May 2017

What is data?

"A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing."

Essence of RDM

«…tracking back to what you did 7 years ago
and recovering it (…) immediately
in a reusable manner.»

Henry Rzepa, Professor of Computational Chemistry, Imperial College London

Why spend time and effort on this?

© Seppo Leinonen: "Sticks and carrots" http://www.seppo.net/cartoons/displayimage.php?album=6&pid=1188 (4.9.2018)

- **Meet** funders' and institutional **requirements**
  - SNSF asks for data management plans as of October 2017
  - EU Horizon 2020 asking for data management plans
- **Good scientific practice**, transparency and validity
- **Avoid reputation risks**

- **Preserve data that cannot be replicated** (e.g. observational data)
- **Avoid redundancy** in data creation/collection
- **Enable** data **re-use and sharing** – even for yourself
- **Raise** your **impact**: your data can be cited
- **Facilitate collaboration** in your group and globally

ETH regulations, intellectual property, privacy and access rights

Recent Overview

https://itsecurity.ethz.ch/en/#/manage_your_data

Guidelines for Research Integrity

- At the ETH Zurich research is founded on **intellectual honesty**. Researchers […] are committed to **scientific integrity** and **truthfulness** in research and peer review.

- https://www.ethz.ch/content/dam/ethz/main/research/pdf/forschungsethik/Broschure.pdf

Article 11. Collection, documentation and storage of primary data

- All **steps** in the treatment of primary data (statistical analyses, reorganizations, etc.) must be **documented in a form appropriate to the discipline** in question (e.g. laboratory logs, other data carriers) in such a way as to **ensure that the results** obtained from the primary data **can be reproduced** completely.

- The **project management is responsible for data management** (data collection, storage, data access, compliance with data protection requirements, etc.). In particular, it should ensure that, following completion of the project, the **data and materials are retained for the period prescribed in the discipline**, and are duly destroyed within the period prescribed by law, if appropriate.

From: https://www.ethz.ch/content/dam/ethz/main/research/pdf/forschungsethik/Broschure.pdf

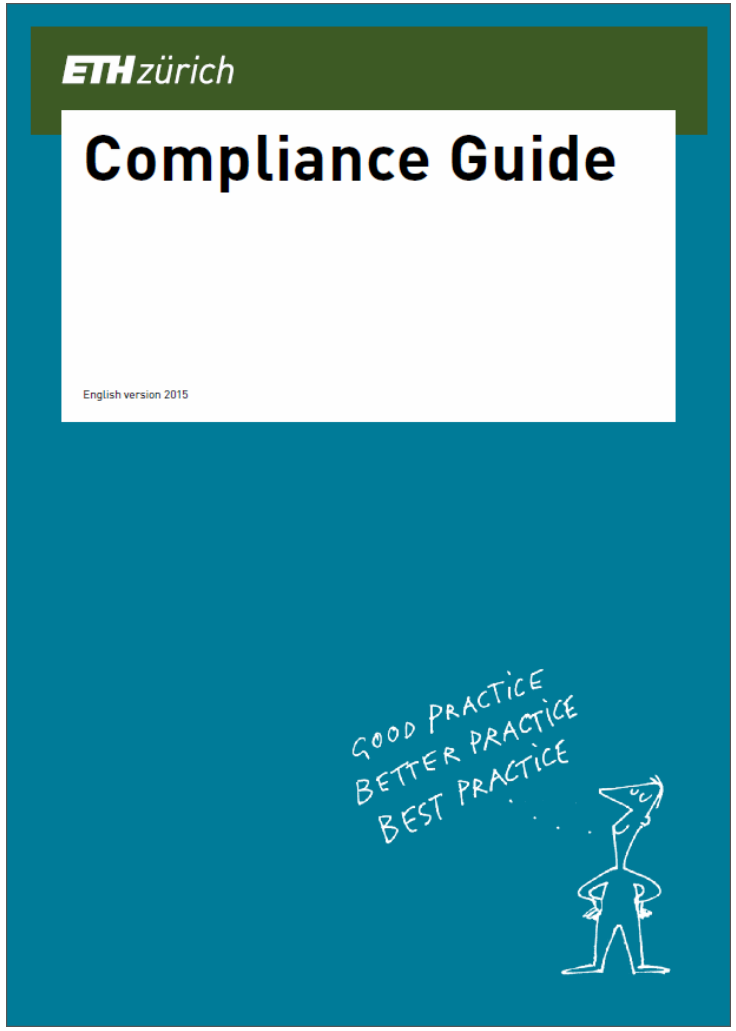
Roles and Responsibilities

- Project **Members**:
  - adhere to the principles of good scientific practice and the guidelines for Research Integrity at ETH.
  - All steps of treatment of primary data must be **documented** and results must be **reproducible**.

- Project **Manager**:
  - responsible for execution of a scientific project and **data management** (data collection, storage, data access, compliance with data protection requirements...).
  - Ensures that all research project participants are aware of the guidelines.
  - Determines together with the professor, which departed project members should retain access to the primary data or materials.

From: https://www.ethz.ch/content/dam/ethz/main/research/pdf/forschungsethik/Broschure.pdf

Compliance Guide



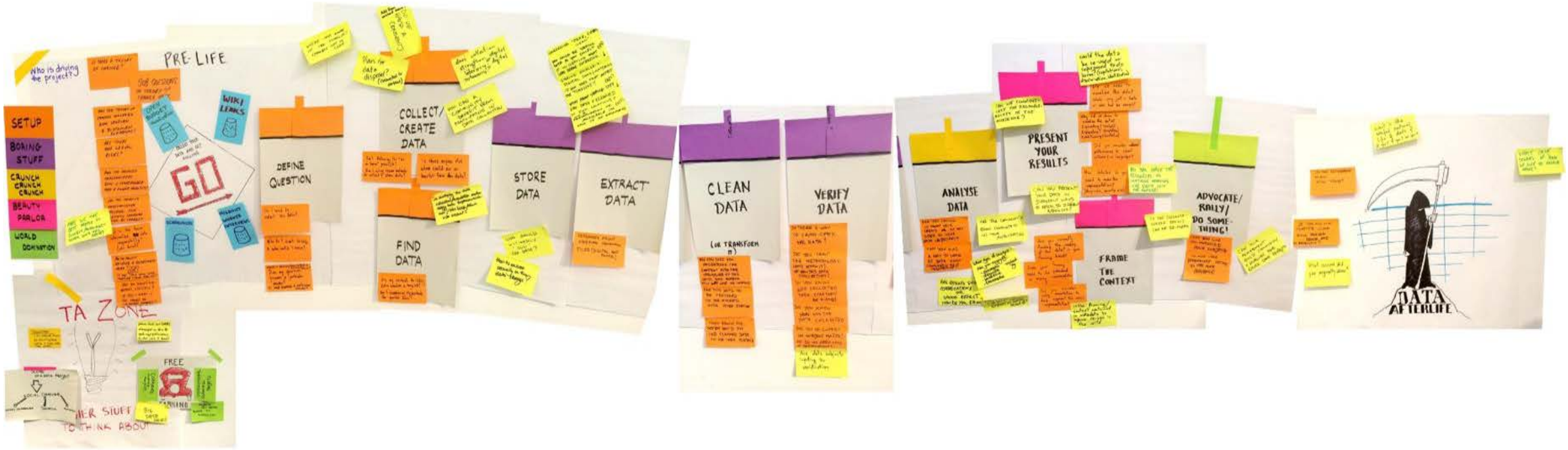
- […] all ETH members […] are required to integrate the general conditions and internal directives into the work process.
- In the research context, the **project manager plays an active role in guiding and monitoring junior scientists**. In particular, he or she is responsible for making sure that everyone involved in the project is aware of the research integrity guidelines.
- Junior scientists are given appropriate guidance.
- Primary data is carefully archived.

From: https://rechtssammlung.sp.ethz.ch/Dokumente/133_en.pdf

- Research must be **documented** and **reproducible**
- Existing **regulations must be complied with**
- The project **manager** is **responsible** for data management

How you ensure those points are observed is up to you

TL;DR … manage your data!



Data Management Planning

What is a Data Management Plan (DMP)?

A brief plan written at the start of a project and updated during its course **to define:**

- **What data** will be collected or created?
- How will the **data be documented and described**?
- **Where will** the data be **stored**?
- **Who will be responsible** for data security and backup?
- Which data will be **shared and/or preserved**?
- **How will the data be shared** and with whom?

DMPs are demanded by:

SNSF from October 2017 on
http://www.snf.ch/de/derSnf/forschungspolitische_positionen/open_research_data/Seiten/default.aspx

Horizon2020 EU funding programme
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

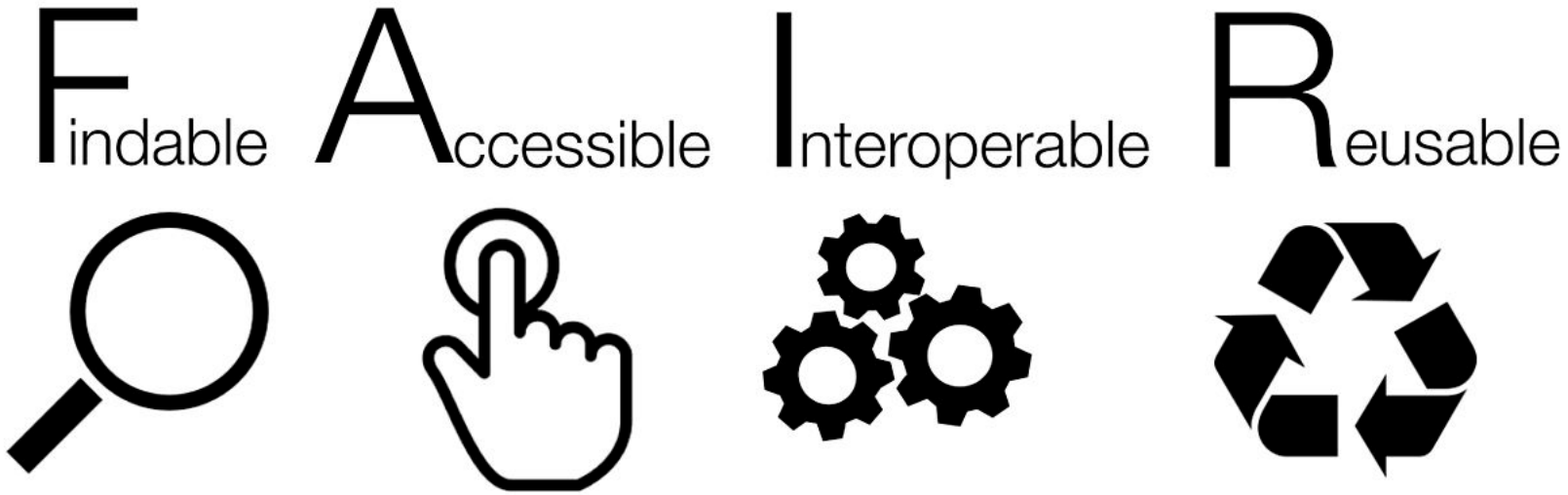
What to do?

- **Data Management Checklist by ETH and EPFL**
- Supports you in the creation of a DMP or in discussing data management in general, even if you don't need to do it to comply with funders
- http://bit.ly/rdmchecklist

- **DMPOnline**
- A tool by the UK Digital Curation Centre that helps you create Horizon 2020 compliant data management plans, by answering a questionnaire
- https://dmponline.dcc.ac.uk

Collection of DMP examples:
http://www.dcc.ac.uk/resources/data-management-plans/guidance-examples

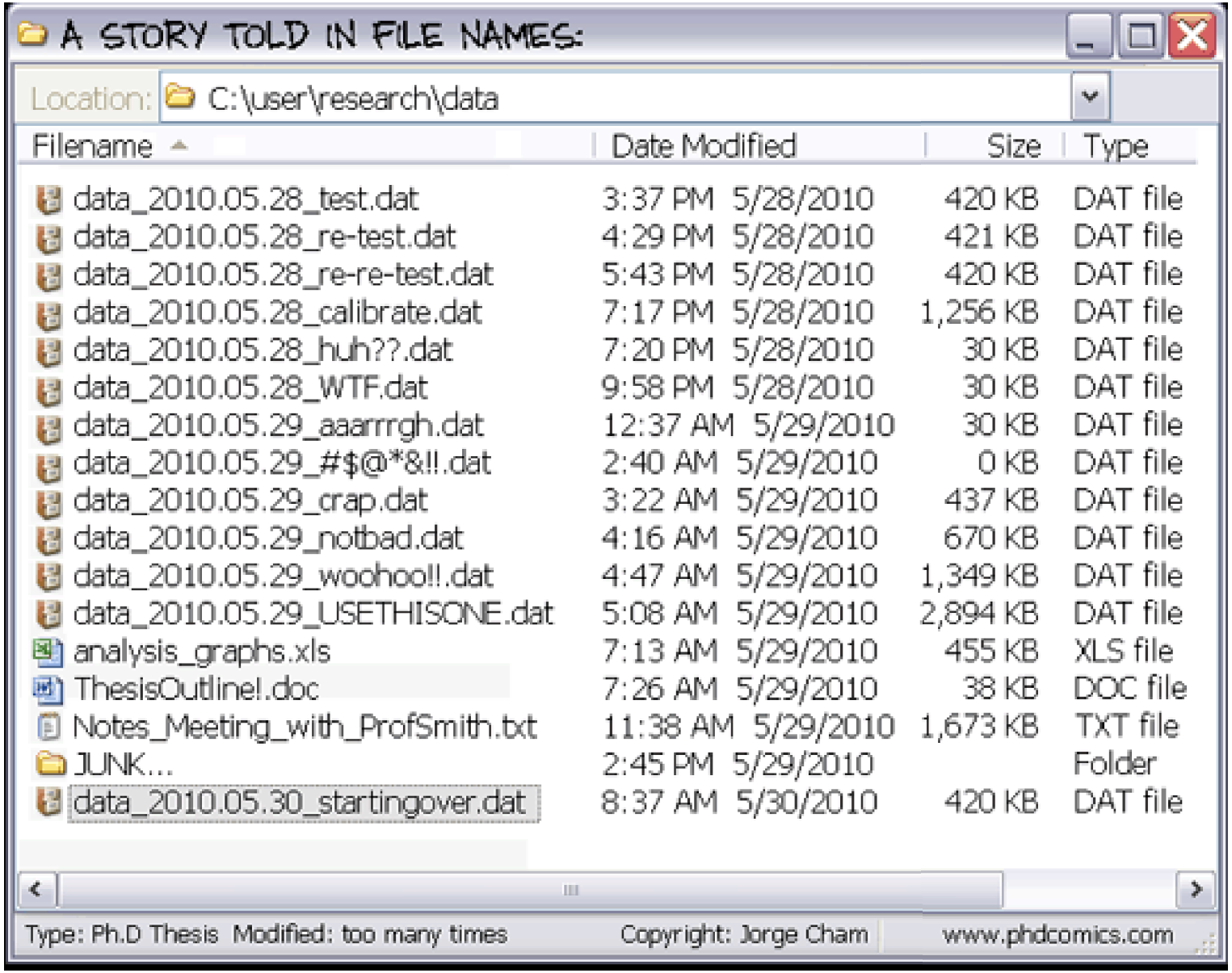
Data should be FAIR



**GARBAGE IN,
GARBAGE OUT!**

Best practices for personal data management

How NOT to do it…



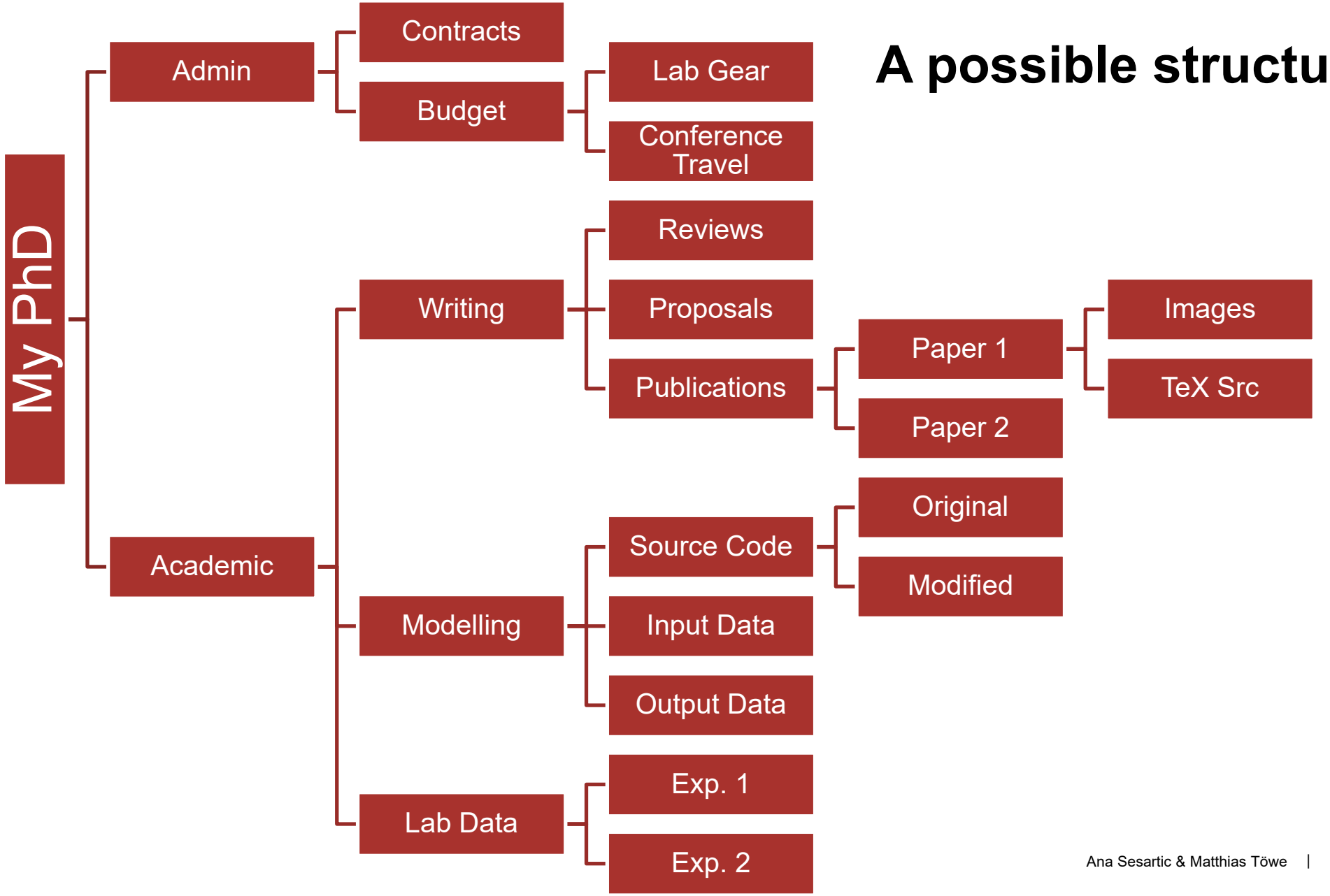
"A story told in file names" from
"Piled Higher and Deeper" by Jorge Cham
www.phdcomics.com
Source:
http://www.phdcomics.com/comics/archive.php?comicid=1323

Try this instead…

- **Keep stuff together that belongs together**
- Keep **path names** short
  - < 255 characters
- **File names** should
  - Reflect content and be unique
  - Use only ASCII characters (no diacritic characters)
  - No spaces
  - Lowercase or camel case (LikeThis)
- Careful! Not all systems are **case sensitive**!
  - UNIX: case sensitive
  - Win/Mac: mostly case insensitive
  - Assume that this, THIS and tHiS are the same.

- Write **dates** like this: **YYYY-MM-DD**

A possible structure…



File organisation tips

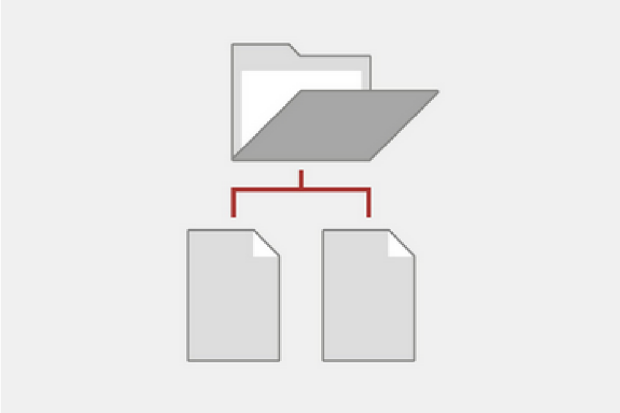
- Aim for a logical organisation, keeping things together that belong together
- Have a clear and consistent naming convention that suits your purposes
- Document your structure in a README text file

For further file and folder organisation tips, see:

- http://www.data.cam.ac.uk/data-management-guide/organising-your-data
- http://www.wur.nl/en/Expertise-Services/Data-Management-Support-Hub/Browse-by-Subject/Organising-files-and-folders.htm
- http://datalib.edina.ac.uk/mantra/organisingdata/

Preferences for file formats

- **Open standards** (non proprietary)
  - If proprietary, convert or if not possible include data viewer
- **Well documented**
- **Widely used and supported** by many tools
- **Uncompressed** (or at least losslessly compressed)
- **Unencrypted**
- When in doubt, **keep original and create a copy** in an open or exchange format
- Don’t rely on **file extensions**
- Consider that **data might be used in different operating systems**



1. Organise and standardise

Establish a file and folder structure that works for you and use it consistently.

2. Identify

Determine which files need to be preserved.

3. Automate backups

Create automated backups and keep them both locally and off-site.

4. Know the lifespan

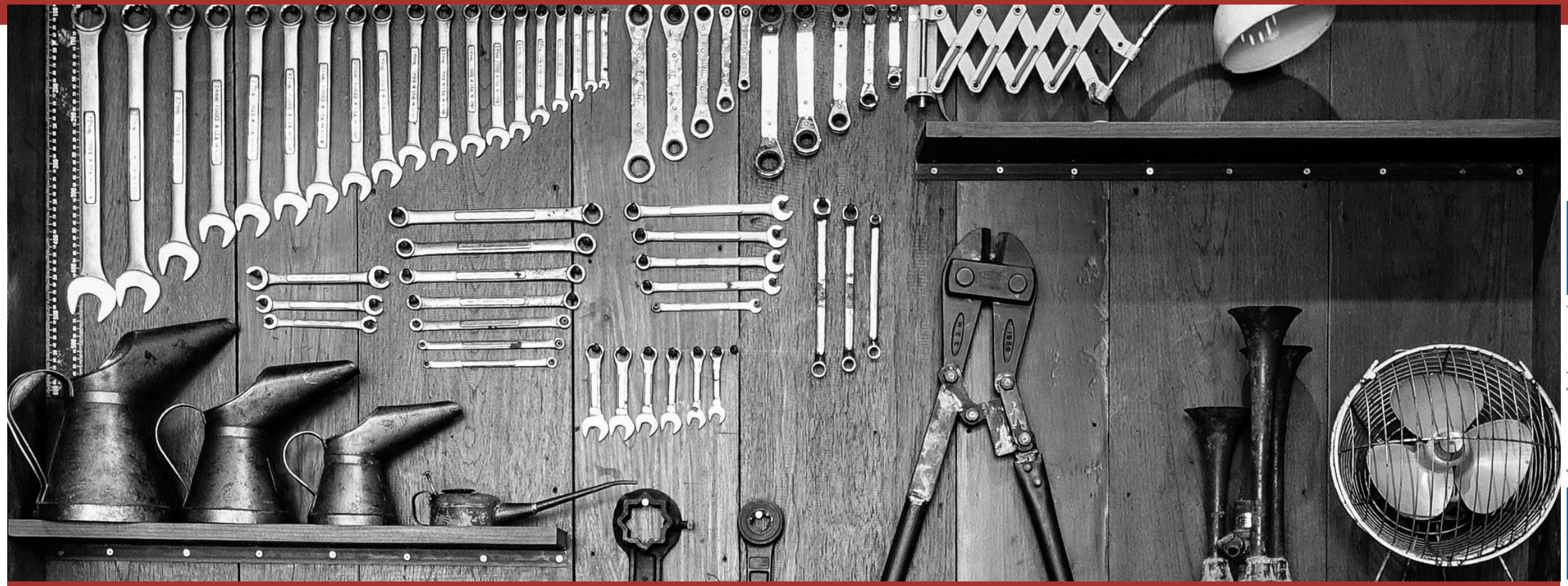
Know the lifespan of your data carriers and re-copy your data to new ones in time.

5. Use simple tools

When collaborating, agree on simple workflows and backup tools. Don't forget to document the context of your data.

6. Use open file formats

Use open file formats and don't compress data to ensure its compatibility with different operating systems.



Tools

Group discussion: current practice

- **Versioning**:
  How do you currently handle it? What works well? What went wrong?
- **Naming conventions**:
  Do you have any? Which rules apply?
- **Sharing**:
  Which tools or services do you use? What are your experiences?
- **Literature Management**:
  Which tools do you use? What are their pros and cons?

Criteria for chosing services and tools

- Where will your data reside?
- Which legislation applies, e.g. in terms of data protection?
- Is the service sustainable?
- Do you trust the provider?
- Who else can access and use which of your data?
- How can you get your data back?
- Is a certain license required?
- Are there immediate or longer term costs?

Example: Collaboration - Sharing

Recommended

- Data stored in Switzerland
- Security regulations fulfilled

https://polybox.ethz.ch

https://cifex.ethz.ch/

SWITCH

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

https://www.switch.ch/filesender

Only conditionally recommended

- Data stored in EU/USA
- Security regulations only partially fulfilled
- Never store sensitive / private data there!

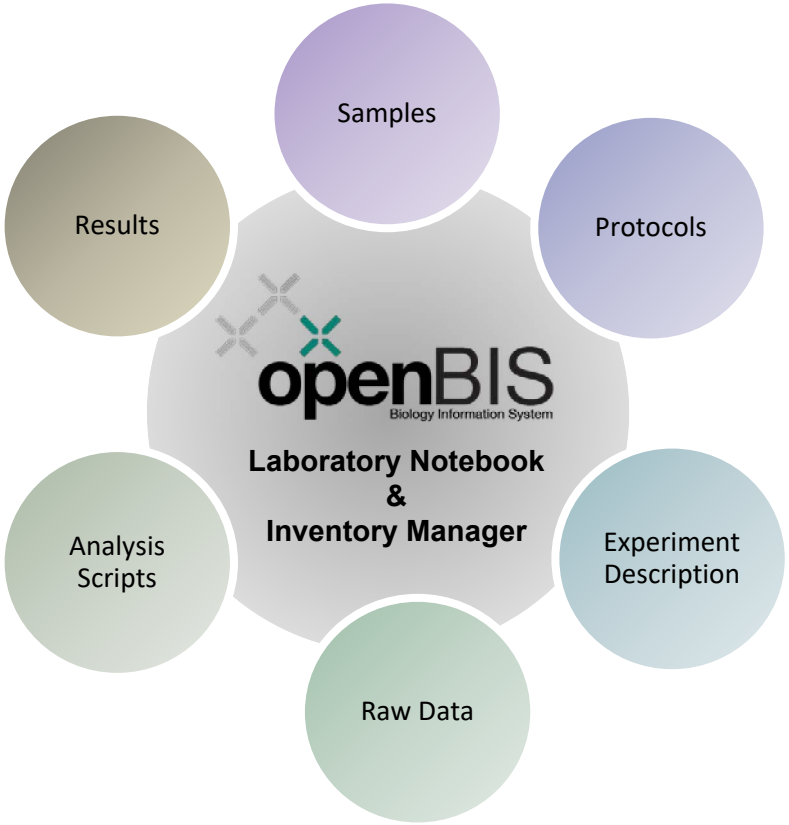
Dropbox

https://www.dropbox.com

wetransfer

https://www.wetransfer.com

openBIS – ELN-LIMS offered by ETH Scientific IT Services



https://labnotebook.ch/

openBIS ELN-LIMS is an integrated:

Inventory management system

Notebook

Data management system

Slide by Caterina Barillari – Scientific IT Services, ETH Zurich

ETH Services

Services at ETH Library

- ETH E-Collection (http://e-collection.library.ethz.ch/index.php?lang=en)
- ETH E-Citations (http://e-citations.ethbib.ethz.ch/index.php?lang=en)
- ETH Data Archive (http://www.library.ethz.ch/Digital-Curation)
  - Long term preservation of data
  - Not for mass storage and active data
- Open Access (http://www.library.ethz.ch/en/Open-Access) including payment of Article Processing Charges (APCs) with a range of publishers
- DOI registration (http://www.library.ethz.ch/DOI-Desk-EN)
- ORCID (http://www.library.ethz.ch/en/ORCID)

Will be merged into **«Research Collection»** and allow publication of documents and data as of mid-June 2017

IT services and ETH transfer

IT Services

- Storage provisioning, usually via your IT Support Group
  - NAS (Networked Attached Storage) and HSM (Hierarchical Storage Management) https://www.ethz.ch/services/en/it-services/catalogue/storage/nas.html
  - LTS (Long-Term Storage) https://www.ethz.ch/services/en/it-services/catalogue/storage/lts.html
- openBIS ELN-LIMS https://openbis-eln-lims.ethz.ch/

ETH-Transfer https://www.ethz.ch/en/the-eth-zurich/organisation/staff-units/eth-transfer.html

- Software disclosure workflow with ETH Data Archive
- Advice on Intellectual Property, Patents, Licensing of Software etc.

Trainings

- Training courses and workshops on information research, reference management, data management, scientific writing and open access by the ETH-Library:
  http://www.library.ethz.ch/en/Services/Training-courses-guided-tours
- Courses offered by the ETH Information Center for Chemistry/Biology/Pharmacy:
  http://www.infozentrum.ethz.ch/en/whats-up/events/
- Further topics on demand

Questions?

Dr. Ana Sesartic
ana.sesartic@library.ethz.ch

Dr. Matthias Töwe
matthias.toewe@library.ethz.ch

Digital Curation Office
ETH Library
ETH Zurich
http://www.library.ethz.ch/Digital-Curation