Research Data Management

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.”
«…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?

- **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.

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](https://www.ethz.ch/content/dam/ethz/main/research/pdf/forschungsethik/Broschure.pdf)
• [...] 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/133en.pdf](https://rechtssammlung.sp.ethz.ch/Dokumente/133en.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 the data** will be documented and described?
- **Where** the data will be stored?
- **Who will be responsible** for data security and backup?
- **Which data** will be shared and/or preserved?
- **How the data** will 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](http://www.snf.ch/de/derSnf/forschungspolitische_positionen/open_research_data/Seiten/default.aspx)
- **Horizon2020 EU funding programme**
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

- **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](https://dmponline.dcc.ac.uk)

**Collection of DMP examples:**
Data should be FAIR

Findable

Accessible

Interoperable

Reusable

CC-BY-SA 4.0 Sangya Pundir
https://upload.wikimedia.org/wikipedia/commons/a/aa/FAIR_data_principles.jpg
Best practices for personal data management
"A story told in file names":


Copyright: Jorge Cham
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

PUBLIC SERVICE ANNOUNCEMENT:
Our different ways of writing dates as numbers can lead to online confusion. That’s why in 1988 ISO set a global standard numeric date format:

This is the correct way to write numeric dates:

2013-02-27

The following formats are therefore discouraged:

02/27/2013 02/27/13 27/02/2013 27/02/13 20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013.ii.27 2/7/13 2013.158904109 MMXIII-II-XVII MMXIII 2078 1330300800
((3+3)*((1+1)+1)-1)*3/3-1/3 2078
10/1101/1101 02/27/20/13 0\ 1\ 237
6\ 7\ 8

© XKCD
My PhD

Admin
- Contracts
- Budget

Lab Gear
- Conference Travel

Reviews

Proposals

Publications

Writing

My PhD

Academic

Modelling
- Source Code
- Input Data
- Output Data

Lab Data
- Exp. 1
- Exp. 2

Paper 1
- Paper 2
- TeX Src
- Images

Original
- Modified

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://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.

Source: [https://doi.org/10.22010/ethz-exp-0002-en](https://doi.org/10.22010/ethz-exp-0002-en)
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 choosing 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/

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!

https://www.dropbox.com

https://www.wetransfer.com
openBIS – ELN-LIMS offered by ETH Scientific IT Services

openBIS ELN-LIMS is an integrated:

- Inventory management system
- Data management system

https://openbis-eln-lims.ethz.ch
Services at ETH Library

- ETH Data Archive ([http://www.library.ethz.ch/Digital-Curation](http://www.library.ethz.ch/Digital-Curation))
  - Long term preservation of data
  - Not for mass storage and active data
- DOI registration ([http://www.library.ethz.ch/DOI-Desk-EN](http://www.library.ethz.ch/DOI-Desk-EN))
IT services and ETH transfer

IT Services

- Storage provisioning, usually via your IT Support Group
  - NAS (Networked Attached Storage) and HSM (Hierarchical Storage Management)
  - 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