Research Data Management
PhDs SEMINAR
From Research to Publication

12.06.18 : MODULE 4
Research data management

Meet & greet

8:45
Learn
What is open science?
Practice
□ Build the data lifecycle

9:30
Learn
How to produce data
Practice
□ Write a README.txt

10:00
Break

10:30
Learn
How to share your data?
Practice
□ Play the RDM Poker
□ Publish a dataset

10:45
Learn
How to reuse data?

12:00
Get things done: Working with a DMP
Practice
□ Experiment RDM Lego

12:30
Up to speed with Zotero

13:00
4.1 Context

Section objectives

✓ The participant understands the link between open science and research data
✓ The participant knows the main steps of the data lifecycle
Open science

**What is it?**

*Open science is the movement to make scientific research, data and dissemination accessible to all levels of an inquiring society.*

FOSTER – Facilitate Open Science Training for European Research

**What to open?**

- Open methodology
- Open source
- Open data
- Open access
- Open peer review
- Open educational resources

→ **Quality** of science
Open 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.

OECD Principles and Guidelines for Access to Research Data from Public Funding, p.13

Why would you make them open?

Think by group. 1 group arguments PRO. 1 group arguments AGAINST.

- Accessible
- Free to all users
- Transparent
- Faster
- Reproducible (less money, less effort)
Stakes

- How to?
- Open science
- Copyright rules
- DPC
- Condition for grants
- Public funding
- Access for all
1) Place the stages in the correct order (white)

2) Find which actions are taken during each stage (orange)

3) Who are the stakeholders for each stage? (blue)

4) What difficulty do you foresee for each stage?
The data lifecycle

CREATE
Researcher/funder
- Collecting
- Writing a DMP

REUSE
Researcher
- Searching
- Reusing
- Citing

PROCESS
Researcher
- Transcribing
- Cleaning
- Describing

PROVIDE ACCESS
Researcher/funder/publisher
- Sharing
- Publishing
- Licensing

ANALYZE
Researcher
- Analyzing
- Interpreting
- Producing research outputs

PRESERVE
Researcher
- Formatting
- Migrating
- Depositing
4.2 How To Produce Data?

Section objectives

✓ The participant is aware of what to take into account when creating data
Research data storing

Planning → Active data → Long term preservation

Selection

File Format?
Storage place?

Store

Selection

Description

Preserve

Renderability?
Integrity? Metadata?
Persistent identifiers?
Appraisal?

Back up
Metadata Standards

- 15 fields
- simple and easy
- not perfect.

- Refined terms (>50)
  - relation -> isPartOf, hasVersion
  - coverage -> temporal, spatial
- And Schemes
  - subjects: LCSH (Thesaurus)
  - languages: RFC1766 (en, de, fr)
  - Date: ISO8601 (YYYY-MM-DD)

And more - disciplinary specific metadata standards:
See http://rd-alliance.github.io/metadata-directory/
Exercise

1. Choose one of the articles provided in class and find the related datasets!

2. Open the dataset and see what it contains.

3. Write a README.txt based on the dataset.
4.3 How To Share Data?

Section objective

✓ The participant knows **how** and **where** research data can be published
RDM Poker Game

1. Choose 3 blue cards
2. Choose 1 green card
3. Choose 1 red card

Flip your cards and see what poker hand you get! Do you have a flush?
RDM Poker Game

«Personal Daily Pragmatic Flush»

EASE THE REUSE OF YOUR DATA

REDUCE ERRORS

SAVE TIME

PERSONAL DAILY MANAGEMENT

PERSONAL HD
«Lab Daily Pragmatic Hand»

SUPPORT INTERNAL LAB COLLABORATION

LAB DAILY MANAGEMENT

SHARED LAB/SECTION STORAGE SPACE

*can be completed with others advantages... 😊*
RDM Poker Game

«Collaborative Daily Pragmatic Flush»

HELP COLLABORATION WITH PARTNERS (OUTSIDE YOUR INSTITUTION)

EASE COMPLIANCE WITH REQUIREMENTS

ALLOW OTHERS TO REUSE YOUR DATA

INTERINSTITUTIONAL COLLABORATIVE DAILY MANAGEMENT

REPOSITORY

there are other possible combinations...

BIBLIOTHEQUE EPFL ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
RDM Poker Game

«Towards Open Science Flush»

- Enhance reproducibility
- Promote free access to science
- Enforcing high scientific ethical standards
- Repository
RDM Poker Game

«Higher possible impact Flush»

GET MORE CITATIONS

BROADER DISSEMINATION OF YOUR DATA

HIGHER VISIBILITY & IMPACT

CONCURRING WITH SCIENTIFIC GOOD PRACTICES OF MY INSTITUTION

DATA PAPER/JOURNAL
Exercise

1. Download data from go.epfl.ch/RDMdataset

2. Publish the dataset on sandbox.zenodo.org

3. Once you are done, write the reference to cite your dataset!
Research data storing

Planning ➔ Active data ➔ Long term preservation

Selection

File Format?
Storage place?

Store

Selection
Description

Preserve

Renderability?
Integrity?
Metadata?
Persistent identifiers?
Appraisal?

Back up
## Comparison: DRYAD – FIGSHARE – ZENODO

<table>
<thead>
<tr>
<th></th>
<th>DRYAD</th>
<th>FIGSHARE</th>
<th>ZENODO</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nonprofit organization</strong></td>
<td>Digital Science (McMillan – SpringerNature)</td>
<td>CERN &amp; OpenAire</td>
<td></td>
</tr>
<tr>
<td><strong>multidisciplinary</strong></td>
<td>multidisciplinary</td>
<td>multidisciplinary</td>
<td></td>
</tr>
<tr>
<td><strong>multiformat</strong></td>
<td>multiformat</td>
<td>multiformat</td>
<td></td>
</tr>
<tr>
<td><strong>CC0</strong></td>
<td>CC0 (data)</td>
<td>license to choose (CC, GNU, BSD...)</td>
<td></td>
</tr>
<tr>
<td><strong>Fee (not for «low-income countries»)</strong></td>
<td>Free of charge</td>
<td>Free of charge</td>
<td></td>
</tr>
<tr>
<td>• Simple search (keywords, author, title, DOI)</td>
<td>• Simple search (title, tags, authors...)</td>
<td>• Simple search</td>
<td></td>
</tr>
<tr>
<td>• Advanced search: add refinement (on all fields)</td>
<td>• Advanced search: document, licence.</td>
<td>• Filter by document type (poster, dataset, presentation...)</td>
<td></td>
</tr>
<tr>
<td>• List of results: filter by author, topic, date, publication</td>
<td>• “Browse” by topic</td>
<td>• List of results: filter by document type, author, access, European Commission projects</td>
<td></td>
</tr>
<tr>
<td>• Ranking of results</td>
<td>• Ranking of results (most recent or most popular first)</td>
<td>• Ranking of results</td>
<td></td>
</tr>
</tbody>
</table>
## Comparison DRYAD – FIGSHARE – ZENODO

<table>
<thead>
<tr>
<th>Pros</th>
<th>Pros</th>
<th>Pros</th>
</tr>
</thead>
<tbody>
<tr>
<td>• link data &lt;-&gt; article</td>
<td>• link data &lt;-&gt; article</td>
<td>• altmetrics</td>
</tr>
<tr>
<td>• data can be updated</td>
<td>• altmetrics</td>
<td>• Personal/EPFL collections</td>
</tr>
<tr>
<td>• view of complete metadata</td>
<td>• data can be updated</td>
<td>• Login with ORCID</td>
</tr>
<tr>
<td>• «submission integration»</td>
<td>• # of views</td>
<td>• Github integration</td>
</tr>
<tr>
<td>• Long-term preservation</td>
<td>• # of shares</td>
<td>• DOI versioning</td>
</tr>
<tr>
<td></td>
<td>• negatives data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Collaboratives spaces</td>
<td></td>
</tr>
<tr>
<td>Cons</td>
<td>Cons</td>
<td>Cons</td>
</tr>
<tr>
<td>DPC: 120$/submission</td>
<td>owned by a big commercial company</td>
<td>search tool to be completed</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>On the roadmap</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Advanced search syntax</td>
</tr>
</tbody>
</table>
Global registry of research data repositories for all academic disciplines

Detailed information on more than 2’000 repositories

Many filters to identify the most suitable repositories
    Subjects, Data access, Data license,
    Metadata standards, PID systems, Repository type...

http://www.re3data.org/
Research data publishing

“It is the release of research data, associated metadata, accompanying documentation, and software code [...] for re-use and analysis in such a manner that they can be discovered on the Web and referred to in a unique and persistent way.

Data publishing occurs via dedicated data repositories and/or (data) journals which ensure that the published research objects are well documented, curated, archived for the long term, interoperable, citable, quality assured and discoverable – all aspects of data publishing that are important for future reuse of data by third party end-users.”

Data paper, data journal

Papers in journals

Data in a repository

Link to dataset*

Compliance with publishers’ requirements

Describe your data
Show their value
Provide access
No new hypotheses or new interpretations
Specific structure of the data paper
Peer-review and article processing charge

Data papers in (data) journals

Link to dataset*

*or data as supplementary material
Data paper

**Data in Brief (Elsevier)**

- Title / Authors / Affiliations
- Abstract
- Specification tables (Subject area, type of data, data acquisition, data format, experimental factors, experimental features, data accessibility)
- Value of data
- Data (Description of the data shared with the data article, to give the reader context before the description of materials and methods)
- Experimental design, materials and methods
- Acknowledgements
- References

**Geoscience Data Journal (Wiley)**

- Title / Authors / Affiliation
- Abstract / Key words
- Dataset (name, dataset, unique identifier)
- Data production methods (description of the experimental set-up for the acquisition of the data)
- Dataset location and format (Description of location, format and accessibility of the dataset)
- Dataset use and reuse (description of actual and potential use for the dataset)
- Acknowledgements
- References

**Drivers:** enhanced visibility, increased citation rate, validation and reproducibility of research, errors and fraud reduction, funders’ and publishers’ compliance, ...

**Barriers:** time consuming, associated costs, confidentiality issues, concerns for data misuse, not (yet?) enough credits and reward,...
4.4 How To Reuse Data?

Section objectives

☑ The participant understands the challenges he/she will face in reusing data
☑ The participant is aware of how to deal with them
MATERIAL & METHODS

1. Take 2 provided sets of LEGO®
2. Combine them into the coolest aircraft ever!

You have 5 minutes.
DISCUSSION

Let’s consider LEGO® as a dataset and discuss at least the following points.

> How would you like to be cited for this dataset?
> How would you like others to re-use this dataset?
> How would you label third party content included in your work?

...
How to cite a dataset

Minimum requirements (according to DataCite)
Creator (year of publication): title. publisher. (resource type). Identifier (DOI as permanent url)
Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. Geological Institute, University of Tokyo. (Dataset)
http://dx.doi.org/10.1594/PANGAEA.726855

Optional (but useful) element
Version
v. 2.1

Give as much information as possible!
4.5 Getting things done: **THE** tool

Section objectives

- The participant discovers how a Data Management Plan (DMP) works
Data Management Plan (DMP)

It is a written document that describes the data you expect to acquire or generate during the course of a research project, how you will manage, describe, analyze, and store those data, and what mechanisms you will use at the end of your project to share and preserve your data.

Stanford University Libraries (2016)

Why

• Streamline & formalize the data management
• Save time - Encourage re-use - Identify weaknesses
• Anticipate on costs
• Comply with funders or institutions’ requirements

When

• In the early stages of a research project (but it is never too late to develop a DMP)
• It is a living document to be maintained and kept up-to-date throughout the project
Data Management Plan (DMP)

Description of the data
1.1 Type of study
1.2 Type of data
1.3 Format and scale of the data

Data collection / generation
2.1 Methodologies for data collection / generation
2.2 Data quality and standards

Data management, documentation and curation
3.1 Managing, storing and curating data
3.2 Metadata standards and data documentation
3.3 Data preservation strategy and standards

Data security and confidentiality
4.1 Formal information/data security standards
4.2 Main risks to data security

Data sharing and access
5.1 Suitability for sharing
5.2 Discovery by potential users of the research data
5.3 Governance of access
5.4 The study team’s exclusive use of the data
5.5 Restrictions or delays to sharing
5.6 Regulation of responsibilities of users
Funders’ requirements

Horizon 2020 (European Commission) – Open Research Data Pilot
Once a funded project has started...
  • Obligation to submit a DMP (as a deliverable) within the first 6 months
    of the project (to be updated whenever significant changes arise)
  • Obligation to deposit underlying data in a research data repository
    (subject-based/thematic, institutional or centralised).

Swiss National Science Foundation
In the funding application...
  • Obligation to include a DMP
Once a funded project has started...
  • Obligation to make underlying data publicly accessible
    in non-commercial, digital databases
    (provided there are no legal, ethical, copyright or other issues)
Research Data Management

Further readings


This bibliography is regularly updated: go.epfl.ch/RDM

For more information on EPFL storage options: http://go.epfl.ch/epflstorage
STAY CONNECTED

Library Training Team
Noémi Cobolet
Raphaël Grolimund
Caroline Salamin
Mathilde Panes
formations.bib@epfl.ch

Research Data at EPFL Library
Lorenza Salvatori
Eliane Blumer
https://researchdata.epfl.ch/

Research Data Management by EPFL Library (2017)

library.epfl.ch
facebook.com/EPFL.library
youtube.com/epfllibrary
@EPFLlibrary
Research Data Management

Credits

These course notes reuse icons published under CC BY 3.0 license on thenounproject.com:

Icons from the Noun Project:
Atom, By Michael Thompson, US
Options, By Mert Güler, TR
Money, By João Gabriel Mendes Corrêa, BR
Justice, By Márcio Duarte, BR
People, By Gregor Cresnar, Business: Marketing Vol. 5 Collection
Government, By Vinny Bucchino
Diamond, By Ferran Brown
Club, By Mister Pixel
Spade, By Yeong Rong Kim
Research by Becris
Magazine by Glenis Gómez Rios
Repository by lastspark
Quote by A.Varghese
Toolkit by Brian Ejar
Law by Marie Van den Broeck
CC0 Heart
Data, by Aybigé