Research Data Management MOOC

Prof. Dr. Basma Makhlouf-Shabou

Head of Master in Information Sciences

Geneva School of Business Administration

University of Applied Sciences and Arts Western Switzerland

President of olos.swiss

h e g

Haute école de gestion de Genève
Geneva School of Business Administratio

Agenda

Basic questions on publishing data

1. What we are talking about
2. Why it is so important
3. Who is concerned / involved in publishing data
4. What you may publish
5. When you may publish
6. Where to publish
7. How to publish …

What are we talking about?

- Data Publication/publishing

Data publication $\neq$ data sharing

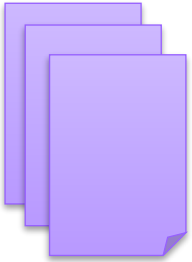
$\neq$ right to use and re-use them

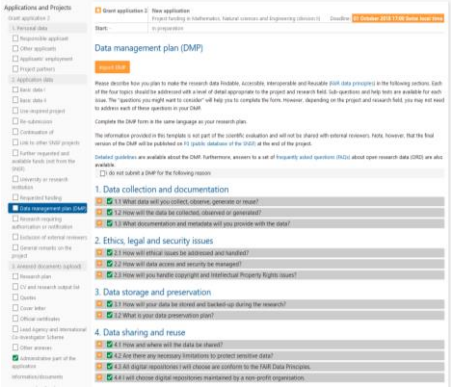
= to provide research data to enable its exploration. It would be possible with a specific action and/conditions (registration, login, mailing, etc.) or a predefined cost (licences).

Why it is so important?

FNSNF

FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION





Funding requirements

- Horizon 2020
- The biggest EU research programme: ~€80 billion over 7 years (2014-2020)
  - The preparation of a DMP is mandatory to receive research funding
  - The research data is **open by default**, while allowing opt-outs
- SNSF
  - Submission of DMPs is mandatory for (most) grant applications (since October 2017)
  - Researchers must **share** (at least) the data underlying their publications, to assure reproducibility

Publisher's requirements on Open Data

- Many journals require to publish the data underlying the published results
- Examples:
  - PLoS (obligation)
  - Nature journals (obligation)
  - American Chemical Society (encouragement)
  - Wiley journals (encouragement)
  - ...
- List of editorial policies on the Dryad website

OPEN DATA

Institutional requirements

UNIVERSITÉ DE GENÈVE

UNIVERSITY FACULTIES STUDENTS SERVICES

Portal | FR EN

RESEARCH DA

Plan Collect & Organi

INSTITUTIONAL POLICY ON

[to be translated]

Préambule

1. La présente politique instit
l'expérimentation ou dérivé
pour la recherche scientifiq
aux données de recherche
projets de recherche condu
internes.
2. La recherche scientifique ét
attention particulière à la **n**
recherches et des données
3. Afin de promouvoir et mair
reconnaît l'importance des
soutient le principe d'une **b**
conformité avec les normes
4. Les chercheurs et les cherch
possible la **maîtrise des dro**

Unil

UNIL | Université de Lausanne

UNIL L'Actu L'Agenda Campus life MyUNIL Faculties Centres EN FR Search

Open Science at UNIL

Strategy|Actions Plan | Open Access | Open research Data | COVID-19 & OS | Contact | More info

You are here: UNIL > Open Science at UNIL > Open research Data

Research data in general

Compliance & Requirements

How to manage your Data ?

Data Management Plan - DMP
Organization & description
Storage & security
Preservation & sharing

FAQ

Data Managemen

What is the Data Mar

Managing data for potential
Researchers must therefore
research project.

The Data Management Plar
collection, documentation, c
related to its use or reuse (l
obligations, sensitive data).
after the research project.

The DMP is a living docume
forms (e.g. electronic docun
on the discipline and resear

In practice, the DMP is the r
discovery, accessibility, inte

De facto, the DMP has becc
growing number of publishe
below).

Why should you writ

Performing a DMP is import

- It saves time and anti
etc.)
- It is sometimes mandatory and required by some public donors before funding is granted.

ETH Zurich > Services & resources English

ETH zürich Services & resources

News & events Organisation Employment & work Teaching Finance & controlling IT Services More services

Homepage > More services > A to Z > Research Data > Data Management Planning

Subnavigation

Data Management Planning

Introduction

A Data Management Plan (DMP) is a document, which describes the way data in a research project will be managed, through each step of the data lifecycle: starting from creation and/or collection of data, to analysis, publication, storage, sharing and reuse.

Data Management Plans

The Swiss National Science Foundation SNSF has published guidelines for Data Management Plans. As of October 2017 researchers are obliged to submit a Data Management Plan as an integral part of their research application.

ETH Library, jointly with the EPFL Library and DLCM project partners, prepared a step-by-step instruction document for filling out Data Management Plans for SNSF proposals.

Access the document here: DLCM template

For users of the openBIS data management system at ETH Zurich, Scientific IT

Data Management Plan Examples

Contact

Research Data @ ETH
ETH Library & IT Services

E-mail

News & Events

09.04.2020

Open Research Data Agreement SERI

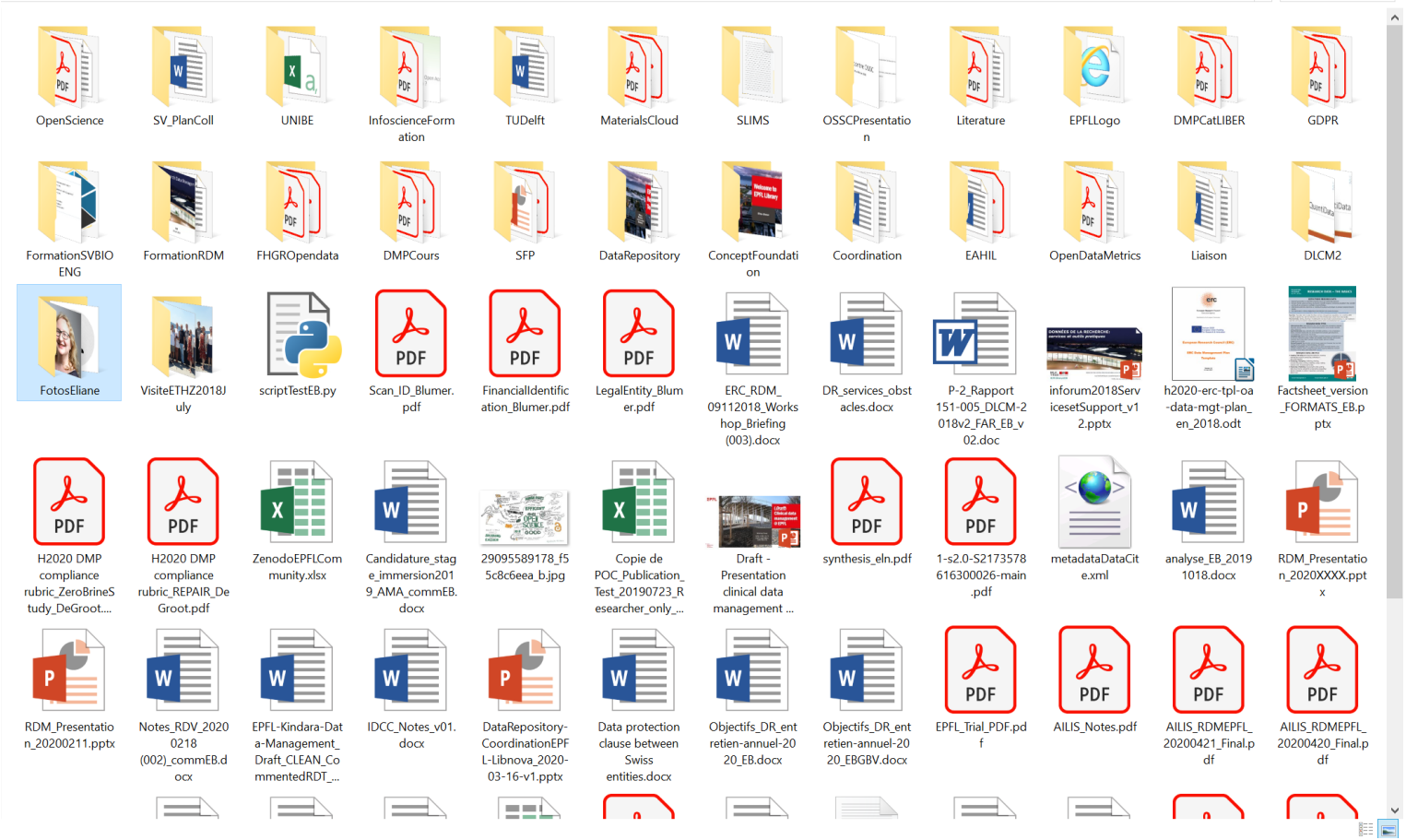
13.03.2020

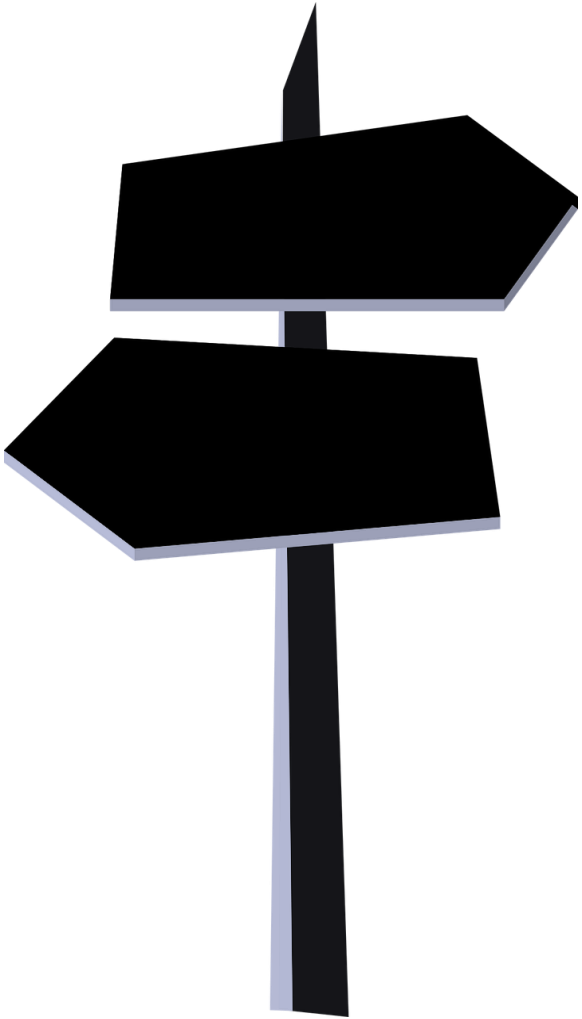
Register for the RDM Summer School in June 2020!

Your personal DMP horror story!

- “I do not expect to produce much data” ⍰ but now you have 250 Gb!
- “I will publish on line” ⍰ good intention, but where exactly?
- “I will publish everything open to the public ... I will consider not publishing everything” ⍰ the hell did I want to say here?!

  - No mention of metadata
  - No mention of access rights
  - No mention of data repositories
  - No mention of laptops, working stations, etc.
  - No mention of filename convention, or data classification plan/shema


OpenScience
SV_PlanColl
UNIBE
InfoscienceFormation
TUDelft
MaterialsCloud
SLIMS
OSSCPresentation
Literature
EPFLLogo
DMPCatLIBER
GDPR
FormationSVBIOENG
FormationRDM
FHGROpendata
DMPCours
SFP
DataRepository
ConceptFoundation
Coordination
EAHIL
OpenDataMetrics
Liaison
DLCM2
FotosEliane
VisiteETHZ2018July
scriptTestEB.py
Scan_ID_Blumer.pdf
FinancialIdentification_Blumer.pdf
LegalEntity_Blumer.pdf
ERC_RDM_09112018_Workshop_Briefing(003).docx
DR_services_obstacles.docx
P-2_Rapport151-005_DLCM-2018v2_FAR_EB_v02.doc
inforum2018ServicesetSupport_v12.pptx
h2020-erc-tpl-oa-data-mgt-plan_en_2018.odt
Factsheet_version_FORMATS_EB.pptx
H2020 DMP compliance rubric_ZeroBrineStudy_DeGroot....
H2020 DMP compliance rubric_REPAIR_DeGroot.pdf
ZenodoEPFLCommunity.xlsx
Candidature_stage_immersion2019_AMA_commEB.docx
29095589178_f55c8c6eea_b.jpg
Copie de POC_Publication_Test_20190723_Researcher_only_...
Draft - Presentation clinical data management ...
synthesis_eln.pdf
1-s2.0-S2173578616300026-main.pdf
metadataDataCite.xml
analyse_EB_20191018.docx
RDM_Presentation_2020XXXX.pptx
RDM_Presentation_20200211.pptx
Notes_RDV_20200218(002)_commEB.docx
EPFL-Kindara-Data-Management_Draft_CLEAN_CommentedRDT_...
IDCC_Notes_v01.docx
DataRepository-CoordinationEPFL-Libnova_2020-03-16-v1.pptx
Data protection clause between Swiss entities.docx
Objectifs_DR_entretien-annuel-2020_EB.docx
Objectifs_DR_entretien-annuel-2020_EBGBV.docx
EPFL_Trial_PDF.pdf
AILIS_Notes.pdf
AILIS_RDMEPFL_20200421_Final.pdf
AILIS_RDMEPFL_20200420_Final.pdf



Findable	Accessible	Interoperable	Reusable
discoverable with machine readable metadata, identifiable and locatable by means of a standard identification mechanism	available and obtainable to both human and machine	both syntactically parseable and semantically understandable, allowing data exchange and reuse among scientific disciplines, researchers, institutions, organisations and countries	sufficiently described and shared with the least restrictive licences, allowing the widest reuse possible across scientific disciplines and borders, and the least cumbersome integration with other data sources

Figure 1: Four foundational characteristics of FAIR

EUROPEAN COMMISSION 2018

FAIR Principles – in concrete…

- ***F* indable**
  Data and metadata are easy to find by both humans & computers.
  - Use metadata
  - Deposit (meta)data in repository/registry
  - Assign a persistent identifier (eg. DOI, HANDL, URN)
- ***A* ccessible**
  Machines & humans can readily access or download (meta)data.
  - As-open-as-possible access to your data (licensing, …)
  - Services with user-friendly interfaces
  - Make the metadata available after data deletion
- ***I* nteroperable**
  Data from different datasets are ready to be exchanged or combined.
  - Use open file format(s), whenever possible
  - Use standardized vocabularies/tags
  - Use cross-linking as much as possible
- ***R* eusable**
  (Meta)data are easily replicated / combined in future research.
  - Attach standardized license to your data (CC, GPL, …)
  - Capture provenance information as precisely as possible

More from the *GO FAIR Initiative*

Who may publish data?

- Research members (individuals and groups)
- Research lead
- Funders
- Academic entity: Institutional repository
- Publisher & editors

What data is worth publishing?

- **Data types**
  - Observational Data, such as sensor observations or interview notes
  - Experimental Data, such as gene sequences or microscopy
  - Simulation Data, such as climate models
  - Derived / Compiled Data, such as compiled database
  - Reference / Canonical Data, such as chemical structures or gene databases
  - Metadata, such as read-me files, file or folder names
  - Paradata, such data collected about interviews and the survey process

Metadata and where to find them

- Metadata **standards** & vocabularies
- **In-file** metadata (eg. *.docx* author, creation date, File tagging, etc.)
- Data **dictionaries** / Codebooks
- Folders & Files structure / **naming** convention
- **Versioning**
- **Readme** files
- **Discovery** metadata (eg. publication keywords)
- **DMP**

What data is worth publishing?

- **Data origin:** to be mentioned if you are reusing existing data (yours or third-party one). Add the reference of the source if relevant.
- **Format of raw data** as created by the device used, by simulation or downloaded: open standard formats should be preferred, as they maximize reproducibility and reuse by others and in the future
- **Format of curated data** (if applicable): open standard formats should be preferred [see *List of recommended file formats by* **EPFL** and/or **ETH Zurich**
- **Estimation of volume of raw and curated data**
- **Data appraisal:** choice of data regarding relevance, uniqueness, openness, usability, etc.

See also recommendations of EPFL https://www.epfl.ch/campus/library/services/services-researchers/#rdm

Focus: file formats

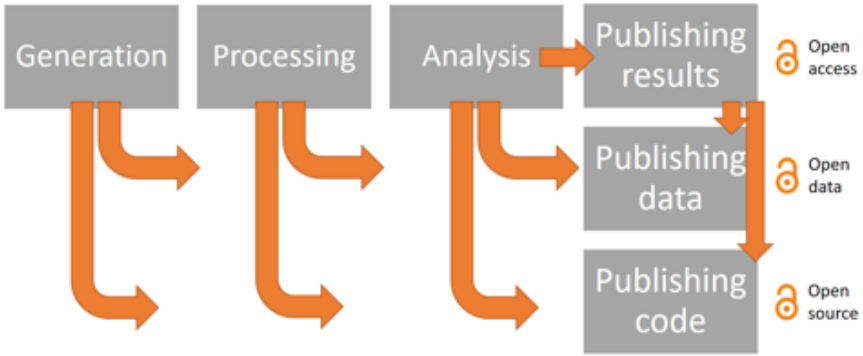
- File formats may be either proprietary or free and may be either unpublished or open.
- Publish, wherever possible in **OPEN formats**
- When sharing the data, make sure to include:
  - The necessary **software** to view the data (e.g. SPSS v.3; Microsoft Excel 97-2003)
  - Information about **version control**
  - If data are stored in one format during collection and analysis, and then transferred to another format for publication: **list out features that may be lost in data conversion such as system specific labels**

Source https://www.epfl.ch/campus/library/wp-content/uploads/2019/09/EPFL_Library_RDM_FastGuide_All.pdf

When to publish?

- Or only publish at the end of a project

- Researcher may start publish from the beginning of the cycle



Where to publish research data?

General Data Repositories

- Dataverse: A repository for research data that takes care of long-term preservation and good archival practices, while researchers can share, keep control of, and get recognition for their data.
- Zenodo: A repository service that enables researchers, scientists, projects, and institutions to share and showcase multidisciplinary research results (data and publications) that are not part of existing institutional or subject-based repositories. ⍰ free, maximum 50 GB / dataset, hosted by CERN
- Dryad: A repository that aims to make data archiving as simple and as rewarding as possible through a suite of services not necessarily provided by publishers or institutional websites. ⍰ 120$ for the first 20 GB and 50$ for additional GB, Non-profit organization
- Olos: Generic, non-profit, Swiss, fee per project and per TB, no limit in volume, licence CC BY, CC0 + any other desired license.

Adapeted form PLoS 2014

Data repositories examples

NAME	DISCIPLINE	NON-PROFIT / INSTITUT.	COUNTRY	FREE	MAX VOLUME	LICENSING
zenodo	Generic	✓ (CERN)		✓	50GB/dataset, ∞ datasets	CC, GNU, BSD
MATERIALSCLOUD	STI / Materials	✓ (EPFL)		✓	5GB General / 50GB AiiDa DB	CC-BY (MIT for Aiida)
figshare	Generic	✗ (Holtzbrinck Group)		Freemium	1 TB per dataset	CC0, CC-BY
DRYAD	Bio / Medical	✓ (?)		✗	?	CC0
The Dataverse Project	Generic	✓ (Harvard University)		✓	?	?
EUDAT	Generic	✓ (HORIZON 2020)		✓	?	CC (DARUP)
OIOS	Generic	✓		fee per project and per TB	No limit	licence CC BY, CC0 + any other desired licence

How to find repositories?

Directories of Research Data Repositories

- DataBib: Databib is a tool for helping people identify and locate online repositories of research data. Users and bibliographers create and curate records that describe data repositories that users can search.
- Re3data.org: Re3data is a global registry of research data repositories from different academic disciplines for researchers, funding bodies, publishers, and scholarly institutions.
- Force 11 Catalog: A dynamic inventory of web-based scholarly resources, a collection of alternative publication systems, databases, organizations and groups, software, services, standards, formats, and training tools.

Adapeted form PLoS 2014

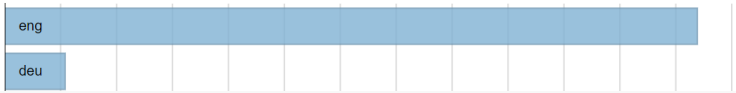
Spotlight: re3data.org

re3data.org

Search Browse Suggest Resources

Repository languages

- AID systems
- API
- Certificates
- Content types

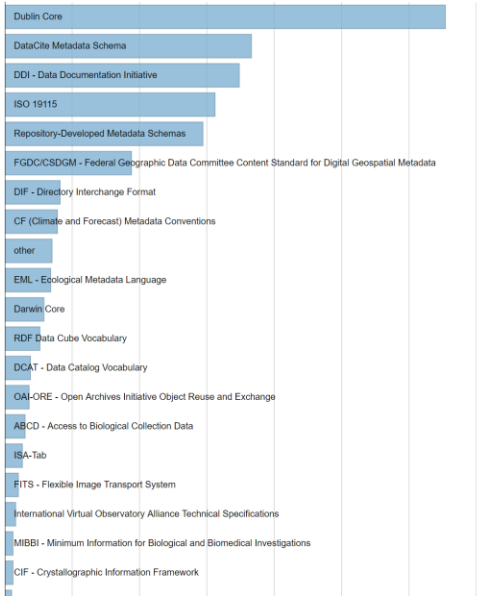


re3data.org

Search Browse Suggest Resources Contact DataCite

Metadata standards

- AID systems
- API
- Certificates
- Content types
- Data access
- Data access restrictions
- Database access
- Database access restrictions
- Database licenses
- Data licenses
- Data upload
- Data upload restrictions
- Enhanced publication
- Institution country
- Institution responsibility type
- Institution type
- Keywords
- Metadata standards
- PID systems
- Provider types
- Quality management
- Repository languages
- Software
- Subjects
- Syndications
- Repository types
- Versioning

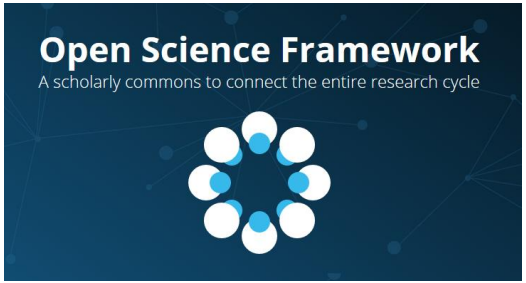


Other options for publication

- Data Papers: A data paper is a peer reviewed document describing a dataset, published in a peer reviewed journal. It takes effort to prepare, curate and describe data

- Open Science approach

F1000Research
Open for Science



The Open Science Training Handbook

Important note!

- Always chose repositories that are:
  - Non-commercial
  - Provide you with unique identifiers
  - Are compliant with standardization in metadata
  - Work in interoperability and are not in a silo
  - Are user-friendly
  - Are endorsed within a long-living institution

How to prepare the dataset for publishing?

- Choose your dataset, based on data type, origin, appraisal etc.
- Think about a logic and well-described structure of your folders and files
- In case of sensitive data: think of how to code
- Transform proprietary formats into open data formats wherever possible
- Choose an appropriate licence
- Recap of metadata: summarize all in a readme-file
- Find the right place to publish the data
- Enter metadata in compliance with the appropriate standards

Documentation summary: readme

- A README provides information about data file(s), granting better reusability
- It can be generated automatically and manually
- Its main content:
  - General information
  - Data and file overview
  - Sharing and access information
  - Methodological information

[Example structure: https://hmd.youmi-lausanne.ch/QpujObCLRSKxefhIJrohtA#](https://hmd.youmi-lausanne.ch/QpujObCLRSKxefhIJrohtA#)

CHANGED 2 MONTHS AGO
OWNED THIS NOTE
EDITABLE

README file: examples and best practice

Bullet point definition

- as the title suggests, 1st thing to look for and read on
- a form of first level documentation (easy to write, easy to read)
- supplies information about the other files in a directory for data or archive for computer software, so that data/software is correctly interpreted and used
  - by yourself at a later time
  - or by others at any time
- structural AND functional information, about both shape and content
- simple (plain) text file
- human readable
- called "README"
- at the root of your project
- hassle-free but some best practice apply (see below)

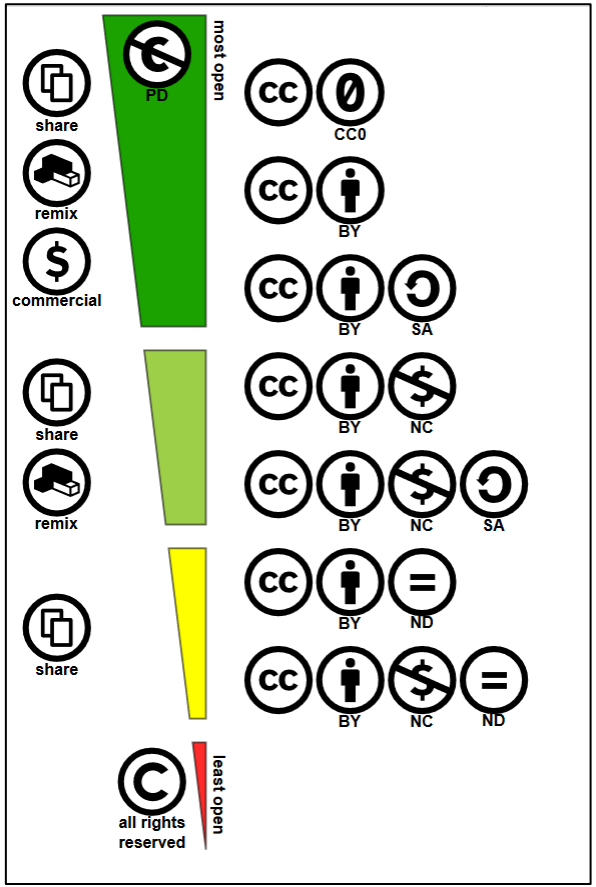
Bullet point best practice

- choose plain text or md
- name it `README.txt`, `README.md` ou simply `README`
- put it at the root of your project
- be consistent, always use the same template
- if needed create one README per dataset
- stick to a KISS layout and display, make full use of
  - blank lines

Adequate licences for data

- Creative Commons are:
  - Enforced by the author
  - Check platform's policy
  - On datasets (no data points)

- The Open Data Commons can be a viable option



Data coding: masking techniques

Pseudonymization
(working data, reversible)

- **PSEUDONYMIZATION**
  Replace data by identifiers. The key is kept separately & securely
- **ENCRYPTION**
  Encrypt the data & keep the key secure. Also for long-term preservation, not data publishing

Some tools:
- **R package:** sdcMicro
- **Java application:** ARX Data Anonymization Tool
- **Java application:** ARGUS
- **Platform:** Amnesia

Anonymization
(published data, irreversible)

- **GENERALIZATION**
  Diminish granularity by generalizing the variables. Appropriate for data too specific or unique records
- **SUPPRESSION**
  Destroy data or part of the outlier records. Appropriate for processing identifiers
- **ADD FAKE DATA**
  To prevent the identification of specific records, add fake data while preserving correlations
- **SHUFFLE**
  Shuffle data over one / several columns without compromising the utility of the data

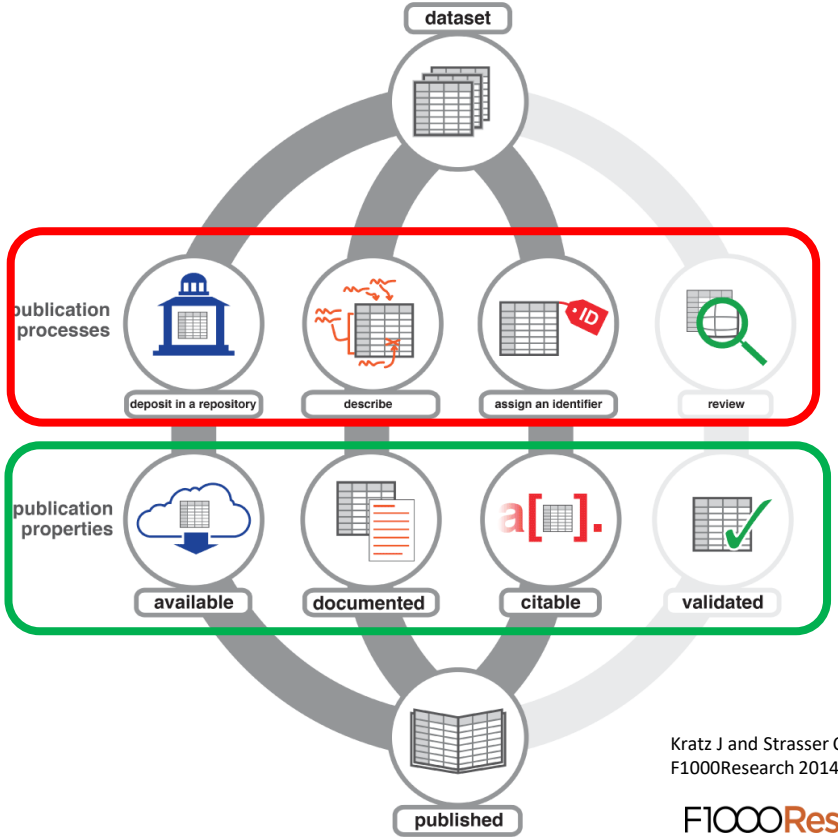
(Other: Differential Privacy, T-closeness, ...)

Adequate licences for code

Name	For	Main idea
MIT	Code	Short Permissive No warranty
Apache 2.0	Code	Permissive Patents allowed No warranty
GPL	Code	Copyleft license Patents allowed Viral
LGPL	Code	Sharing libraries under the same terms Mix of different licenses allowed
AGPL	Code	Strong copyleft Patents allowed Viral

Wikipedia : Opensource License

IN SHORT… To be published, datasets are typically deposited in a repository to make them available, documented...



Kratz J and Strasser C. Data publication consensus and controversies [version 3]. F1000Research 2014, 3:94 (doi: 10.12688/f1000research.3979.3)

F1000Research

Bibliography

- Kratz, J., & Strasser, C. (2014). Data publication consensus and controversies. F1000Research, 3. https://doi.org/10.12688/f1000research.3979.3
- Niemeyer, K. E., Smith, A. M., & Katz, D. S. (2016). The Challenge and Promise of Software Citation for Credit, Identification, Discovery, and Reuse. J. Data and Information Quality, 7(4), 16:1–16:5. https://doi.org/10.1145/2968452
- Wilkinson, M. D. , and al The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

- Acknowledgements to the RDM team from the EPFL Library and DLCM and OLOS Team

Thank you!

Research Data Management MOOC Powered by DLCM Mandated by swissuniversities

Contact us info@dlcm.ch

OLOS

https://olos.swiss