EPFL

DOI: 10.5281/zenodo.3327829

Research Data Management

RESEARCH DATA

Fast Guide #01

DEFINITIONS

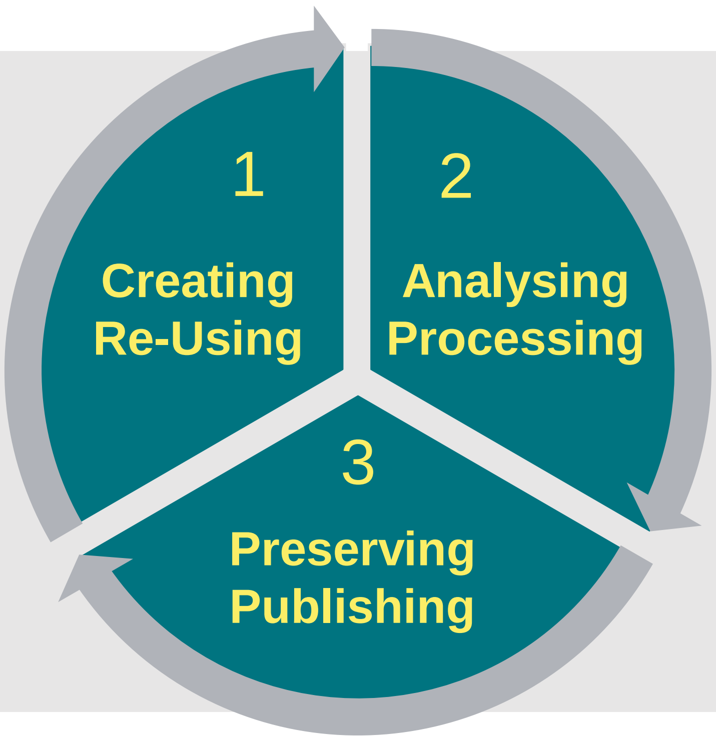
- ✓ Material generated or collected during the course of conducting research [1]
- ✓ Factual records used as primary sources for scientific research, commonly accepted in the scientific community as necessary to validate research findings [2]
- ✓ Information collected, observed, or created, for analysis purposes or produce original results [3]
- ✓ Any information in binary digital form derived from the research process [4]

RESEARCH DATA LIFECYCLE

1 **Creating / Re-using**. Plan data collection; Locate existing data sources; Produce, Collect or Document data

2 **Analyzing / Processing**. Validate data; Clean data; Transform data; Create metadata; Use or Create analysis tools; Visualize and Interpret data

3 **Preserving / Publishing**. Review data; Convert data into formats suited for preservation; Deposit data and metadata in archive / repository; Promote data re-use



RESEARCH DATA TYPES

- **Observational Data**. Data captured in-situ, can’t be recaptured, recreated or replaced.
  Examples: Sensor readings; Sensory (human) observations; Survey results; Interview notes / transcripts
- **Experimental Data**. Data collected under controlled conditions, in situ or laboratory-based, should be reproducible, but can be expensive. *Examples*: Gene sequences; Chromatograms; Spectroscopy; Microscopy
- **Simulation Data**. Use of a model to study an actual or theoretical system, where the input can be more important than output data. *Examples*: Climate models; Economic models; Biogeochemical models
- **Derived / Compiled Data**. Reproducible, but can be very expensive.
  Examples: Derived variables; Compiled database; 3D models
- **Reference / Canonical Data**. Static or organic collection [peer-reviewed] datasets, probably published and/or curated. *Examples*: Gene sequence databanks; Chemical structures; Census data; Spatial data portals [5]
- **Metadata**. Structured information associated with data for purposes of discovery, description, use, management, and preservation. *Examples*: Read-me files; Publication keywords; File and folder names [7]

RAW DATA

Raw data refer to data that have not been changed since acquisition, e.g. a real-time GPS-encoded navigation file, and the initial time-series file of temperature values from a heat probe.

PROCESSED DATA

Editing, cleaning or modifying the raw data results in processed data, e.g. raw multibeam data files can be processed to remove outliers and to correct sound velocity errors [6].

[1] www.ed.ac.uk/information-services/research-support/research-data-service

[2] www.oecd.org/sti/sci-tech/38500813.pdf

[3] www.ed.ac.uk/information-services/research-support/data-management

[4] www.degruyter.com/view/product/430793

[5] http://guides.library.stonybrook.edu/research-data-services/types

[6] www.marine-geo.org/help/data_FAQ.php

[7] www.epfl.ch/campus/library/wp-content/uploads/2019/09/EPFL_Library_RDM_FastGuide_All.pdf#page=5

cc BY NC SA

EPFL

DOI: 10.5281/zenodo.3327829

Research Data Management

FAIR Data Principles [1]

Fast Guide #02

Data and metadata are easy to find by both humans and computers.

F
FINDABLE

F1 (Meta)data are assigned a globally unique and persistent identifier.

F2 Data are described with rich metadata.

F3 Metadata clearly and explicitly include the identifier of the data they describe.

F4 (Meta)data are registered or indexed in a searchable resource.

DESCRIBE

Describe provenance, usage and organization of data with standardized metadata (RDA standards, DataCite, DublinCore). Make metadata available even if data is not.

Both humans and computers can readily access or download datasets.

A
ACCESSIBLE

A1 (Meta)data are retrievable by their identifier using a standardized communication protocol:

- A1.1 the protocol is open, free and universally implementable;
- A1.2 the protocol allows for an authentication and authorization procedure where necessary.

A2 Metadata are accessible, even when the data are no longer available.

OPEN

Open your data using standardized licenses. Limitations may apply to the openness (ex. embargo). Disclose files in open formats, even alongside proprietary formats.

Data from different datasets are prepared to be combined or exchanged.

I
INTEROPERABLE

I1 (Meta)data use a formal, accessible, shared and broadly applicable language for knowledge representation.

I2 (Meta)data use vocabularies that follow FAIR principles.

I3 (Meta)data include qualified references to other (meta)data.

LINK

Use persistent identifiers for datasets (ex. DOI, HANDL, URN) and tag all the metadata with the same identifiers. Cross-link datasets with linked-data standards (RDF).

Published data can be easily combined / replicated in future research.

R
REUSABLE

R1 (Meta)data are richly described with a plurality of accurate and relevant attributes:

- R1.1 (meta)data are released with a clear and accessible data usage license;
- R1.2 (meta)data are associated with detailed provenance;
- R1.3 (meta)data meet domain-relevant community standards.

PUBLISH

Deposit datasets in data repositories, favoring services with user-friendly interfaces. Make sure to chose a FAIR-compliant data repository, also for the relative code.

> “Data should be as open as possible, as closed as necessary.”
>
> Carlos Moedas, EU Commissioner

How FAIR are your data?
Take the self-assessment test [2]

Did you know?

46%	of researchers are aware of the existence of FAIR principles [3]
20-50%	increased citation for articles linked to associated data [4, 5]
32%	of time spent to find data could be saved if FAIR principles were applied [6]

[1] FAIR principles: go-fair.org/fair-principles

[2] FAIR self-assessment tool: ands-nectar-rds.org.au/fair-tool

[3] State of Open Data 2019: doi.org/10.6084/m9.figshare.9980783.v2

[4] Open Data Citation Advantage: sparceurope.org/open-data-citation-advantage

[5] The citation advantage of linking publications to research data: arxiv.org/abs/1907.02565

[6] Cost-benefit analysis for FAIR research data: https://op.europa.eu/s/n75s

cc BY NC SA

EPFL

DOI: 10.5281/zenodo.3327829

Research Data Management

COST OF RDM

Fast Guide #03

RDM activities to budget

DATA MANAGEMENT PLAN
DMP writing; DMP revision; DMP publishing

COLLECTION
Databases and software; Data formatting; Data organization; Data transfer

ACTIVE MANAGEMENT
Electronic Lab Notebook (ELN); Laboratory Information Management System (SLIMS); Data sharing platform

DOCUMENTATION
Data description and metadata; Documentation and transcription

STORAGE / BACK-UP
Data back-up; Data storage; NAS; Cloud

ACCESS / CONTROL
Access control; Data security; Sensitive data protection; 3rd party data

SHARING
Anonymization; Copyright assessment; Data cleaning; Data publishing

PRESERVATION / ARCHIVING
Data preparation; Long-term preservation; Data repository

HR + HARDWARE + SOFTWARE + SECURITY + PUBLISHING

Time

Costs cumulate and can increase over time

- Do not overcomplicate your **processes**
- Do not adopt too many **tools**

Did you know?

RDM costs can be eligible for funding applications

- **SNSF.** Costs for Open Research Data (ORD) activities, to grant access to data in non-commercial repositories (if no legal, ethical, copyright or other issues exist). The SNSF may generally allocate up to CHF 10,000 [7].
- **ERC / H2020 / MSCA.** Costs related to Open Access to research data (APC, RDM, curation, storage costs, …) are eligible for reimbursement during the project [7].

Percentage relative to the total funds requested to SNSF in 2018 for ORD activities [8] **0.2%**

+5% RDM cost on the total project expenditure expected to properly manage and steward data [1]

Cost reduction expected for projects tackling poor data quality, redundant data, and lost data [2] **-15%**

TOOLS / RESOURCES

- RESEARCH OFFICE BUDGET TEMPLATES
  The RDM costs are already listed in the budget templates available on the EPFL-ReO website [3]
- COST CALCULATOR
  Try out the EPFL Library's online tool to calculate the storage costs for your research project [4]
- QUESTIONS TO CONSIDER
  An overview of possible costs per research activity is presented by the Utrecht University [5]
- CHRONOS
  Timekeeping for research projects, to justify eligible personnel costs to the funding bodies [6]

[1] ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf (p.17) / [2] www.usgs.gov/products/data-and-tools/data-management/value-data-management / [3] research-office.epfl.ch/research-funding / [4] rdmepfl.github.io/costcalc / [5] www.uu.nl/en/research/research-data-management/guides/costs-of-data-management / [6] www.epfl.ch/research/services/manage-projects/chronos / [7] www.epfl.ch/research/services/fund-research/funding-opportunities/research-funding / [8] doi.org/10.5281/zenodo.3618209 / Icons: www.flaticon.com/packs/business-seo

Bibliothèque Contact and info go.epfl.ch/rdm researchdata@epfl.ch

CC BY NC SA

EPFL

DOI: 10.5281/zenodo.3327829

Research Data Management

FILE FORMATS

Fast Guide #04

DEFINITION

A file format is a standard way to encode data for storage in a computer file. It follows a protocol that specifies how bits are used to encode information in a digital storage medium. File formats may be either proprietary or free and may be either unpublished or open [1].

When listing out the data formats you will be using, 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 preservation: list out features that may be lost in data conversion such as system specific labels

When selecting file formats for archiving, the formats should ideally be:

- Non-proprietary, unencrypted, uncompressed, commonly used by the research community
- Compliant to an open, documented standard: interoperable among diverse platforms and applications, fully published and available royalty-free, fully and independently implementable by multiple software providers on multiple platforms without any intellectual property [2]

TYPE OF DATA	APPROPRIATE	ACCEPTABLE	DEPRECATED
Tabular (extensive metadata)	CSV – HDF5	TXT – HTML – TEX – FASTQ [3] – POR	
Tabular (minimal metadata)	CSV – TAB – ODS – SQL – TSV	XML (if appropriate DTD) – XLSX	XLS – XLSB
Textual / Presentation	TXT – PDF – ODT – ODM – TEX – MD – HTM – XML – EXTXYZ [4] – ODF	PPTX – RTF – DOCX – PDF (with embedded forms) – EPS – IPF	DOC – PPT – DVI – PS
Code / Computation	M – R – PY – IYPNB – RSTUDIO – RMD – NETCDF – AIML	SDD	MAT – RDATA
Image & Spectroscopy	TIF – PNG – SVG – JPEG – FITS	JCAMP – JPG – JP2 – TIF – TIFF – PDF – GIF – BMP – DM3 – OIR – LSM [5]	INDD – AIT – PSD – SPC
Audio	FLAC – WAV – OGG – MXL – MIDI – MEI – HUMDRUM	MP3 – AIF	
Video	MP4 – MJ2 – AVI – MKV	OGM – MP4 – WEBM	WMV – MOV – QT
Geospatial	NETCDF – tabular GIS attribute data – SHP – SHX – DBF – PRJ – SBX – SBN – POSTGIS – TIF – TFW – GEOJSON	MDB – MIF	
3D structures & images	X3D – X3DV – X3DB – PDF3D – POV – PDBML	DWG – DXF – PDB	PXP
Generic	XML – JSON – RDF		

[1] en.wikipedia.org/wiki/File_format / [2] library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-formats / [3] justsolve.archiveteam.org/index.php/Scientific_Data_formats / [4] wiki.fysik.dtu.dk/ase/ase/io/io.html / [5] docs.openmicroscopy.org/bio-formats/5.8.2/supported-formats.html

BY NC SA

EPFL

DOI: 10.5281/zenodo.3327829

Research Data Management

METADATA

Fast Guide #05

> "Metadata is structured information associated with an object for purposes of discovery, description, use, management, and preservation"
>
> (National Information Standards Organization, 2008)

METADATA is

- UBIQUITOUS, PROLIFERATIVE
- EMBEDDED OR SUPPLEMENTAL
- AUTOMATIC OR MANUAL
- FOR INTER-OPERABILITY

5 METADATA FAMILIES

- ● USE (ex. number of downloads)
- ● DESCRIPTIVE (ex. title, author, keywords)
- ● ADMINISTRATIVE (ex. publication date, license)
- ● TECHNICAL (ex. version of producing device)
- ● PRESERVATION (ex. last checksum date)

R E S E A R C H

- Find datasets / code using metadata
- Plan & Design metadata
- Acquire / Create metadata
- Clean metadata, control quality
- FAIR-ify (normalize, enrich, reconcile)
- Choose FAIR compliant repositories
- Carefully fill-in forms
- Publish metadata (along data)
- Allocate / Add PIDs
- Interlink your and others' PIDs
- Hand-over for long term preservation

Be systematic, consistent, thorough

From spreadsheets to databases and semantic web knowledge bases, the MORE metadata you have, the BETTER data management system you need

FAIR data, good QUALITY linked (open) data, mainly relies on rich, detailed, qualified, shared, standardized metadata

Metadata and metadata STANDARDS creation, adoption and maintenance is a JOINT EFFORT within and between interest-based communities

Elements to Build (Your own) Strong METADATA

FORMAT, TECHNICAL, INTERCHANGE STANDARDS: exif, IPTC, instrumentation specific standards…

NORMS, STANDARDS, REFERENCES: ISO 8601, ISO 639-1, ISO 3166-1, thesaurii, vocabularies, authorities…

CONTENT MODELS: ISA (Investigation-Study-Assay) framework, Force11 Software citation principles

STRUCTURE STANDARDS & SCHEMAS : INSPIRE, SDMX, Darwin Core, Dublin Core, PROV model, Datacite…

Give it a try

Source code citation: CodeMeta metadata generator

Dataset citation: Datacite metadata generator

Useful resources

www.dcc.ac.uk/resources/metadata-standards / fairsharing.org / bartoc.org / lov.linkeddata.es

Bibliothèque

Contact and info go.epfl.ch/rdm researchdata@epfl.ch

cc BY NC SA

EPFL

DOI: 10.5281/zenodo.3327829

Research Data Management

CODE AS DATA

Fast Guide #06

As for data, when working with code, good management practices are needed.
The publication of code is crucial to understand, validate, reuse and repeat the research.

CODE MANAGEMENT – TIPS & TRICKS

VERSIONING

Versioning systems are powerful code management tools. The most used is Git, it's free and open:

- You can manage different versions of your code, track and undo changes as needed.
- Connected to a repository, your code and its changes are automatically backed up
- You can also work in team on the same code

SHARING

In order to share your code and make it visible, repositories provide various services like version management system, wikis, task management and issues tracking, one of the most known is Github.

- EPFL provides c4science.ch for code versionning. Data are stored in Switzerland.
- EPFL provides also gitlab.epfl.ch (open-source github). Data are stored at EPFL.

DESCRIBING

README documentation is a really important part of coding. It allows you to explain your code, for you and others. You should add rich metadata and documentation (README, LICENSE, comments on code...) on any publication of the code.

Some tools like sphinx-doc.org and doxygen.nl can help you by going through your code and generating a preformated documentation.

LICENSING

As for data, it is important to explain how your code can be used or cited by others (with related restrictions). You have at least three options:

- Open source licenses (permissive as MIT or GPL)
- Academic licenses (restrict commercial usage)
- Commercial licenses (reserve commercial usage)

PUBLISHING

Don't forget to generate a DOI to uniquely identify a version of your software and to easily cite it.

Most data repositories (like Zenodo) generate a DOI for your deposit of code.

TIP: Zenodo provides an integration with Github.

PRESERVING

Preservation is important for keeping your work secure and also for scientific validation.

- c4science is an EFPL solution to preserve your code as an backup solution.
- For the long term, you can still use a generic data repository (like Zenodo).
- If you use another code repository, you can always copy it on c4science.

BY NC SA

EPFL

DOI: 10.5281/zenodo.3327829

Research Data Management

ELN – Electronic Lab Notebook

Fast Guide #07

ELN vs. Paper notebook

- Better knowledge transmission internally and externally
- Automated backup and centralized storing location for improved preservation
- Work uniformization by proposing templates and sharing between members

When considering an ELN for your lab, make sur to answer the following questions …

PRACTICAL

- Are the **storage** method and **location** adequate for my research?
- If your ELN is **cloud based**, where are your data hosted and who can access them?
- Can I have a **connected computer** where I need to use the ELN?
- Do I work with **pattern** and does my ELN support it?
- Do I need **support** (hotline, ...)?
- Do I need some **specific tools**?

INTERFACE

- Do I find the interface **suitable** for me?
- Is it compatible with **mobile** devices?
- Do I need a **sample**/laboratory management?

IMPORT EXPORT

- Can I **import** my previous notes?
- What are the import options, is there an **API**?
- Can I **export** my data in an open way?
- What are the export options, is there an **API**?
- What are the export **formats**?
- Do I have data **volume** limitation?
- What is the best ELN **business plan** for me? §

INTEROPERABILITY

- Is the ELN **compatible** with software I use to generate data?
- Is the compatibility limited to the **import** of data I generate?
- Can I use my **cloud** software (Google Drive, Mendeley, ...) with the ELN?
- Can I **integrate** new services with the ELN?
- Can I use **repositories** from within the ELN?
  - Data: Zenodo, MaterialsCloud, Figshare, ...
  - Code: C4science, EPFL GitLab, …

ELN @ EPFL

- EPFL ELN: Chemistry Notebook, developped by Luc Patiny (eln.epfl.ch)
- SLIMS: Commercial solution proposed by the School, integrating a sample management and different services for biologists (https://sv-it.epfl.ch/page-120709-fr-html/lims)
- OTHERS: The EPFL Library Research Data team can help you choosing or implement a viable ELN

More resources

ELN comparison table: datamanagement.hms.harvard.edu/electronic-lab-notebooks

RDM software: www.epfl.ch/campus/library/services/services-researchers/rdm-software

BY NC SA

EPFL

DOI: 10.5281/zenodo.3327829

Research Data Management

PERSONAL DATA MANAGEMENT

Fast Guide #08

PERSONAL DATA is …

… all information relating to an identified or identifiable person [1]

EXAMPLES

Name, Date of birth, Address, Photos, Videos, IP address, GPS coordinates, Biometric data, Genomic data, etc.

SENSITIVE PERSONAL DATA

“data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation” [4]

FEDERAL ACT ON DATA PROTECTION (FADP) [1]

applies to projects conducted in Switzerland, with additional laws for research involving human beings (Human Research Act)[2]

Principles: Good faith, Lawfulness, Proportionality, Exactitude, Security

- **Data collected on internet** [3] is still submitted to restrictions, even if published by the subjects
- **Hash** the identifiers if the project goals can be reached without them, and restricted access right to the **pseudomisation** key
- You need to assess the **risk of reidentification**
- **Inform the subjects** about the contact details of your unit, the purposes of your data collection, the recipients of the personal data, their right to access to their personal data, and the likely consequences if they refuse to provide their personal data
- Anonymized data received from a **third party** still requires the subject to be informed of this new use
- Legal consent for subjects **under 18** years is required to collect their data
- Personal data can be **published** only if the subjects consent to the publication, but in no case if they are sensitive
- Guarantee these subjects’ **minimal rights**: rights of access, modification, erasure

GENERAL DATA PROTECTION REGULATION (GDPR) [4]

for projects using personal data of subjects who are in the EU, with some derogations for scientific or statistical purposes (art.89)

Principles: Lawfulness, Data minimization, Accuracy, Storage limitation, Integrity, Transparency, Privacy-by-design, Confidentiality, Accountability

- You need a **description** of how you will implement the principles
- If data processing or storage are **outsourced**, document external services’ GDPR compliance
- In the event of a **data breach**, notify immediately the VPSI or the DSPS
- **Inform the subjects** about their rights to modify their data, restrict the use of their data and withdraw their participation, plus extensive information about the data collection/processing
- Provide a **Data Protection Impact Assessment** [5] if the project may result in a high risk, i.e. if involving data processed on a large scale, innovative data use, sensitive data, vulnerable subjects, data transfers outside of the EU, etc.
- Any transfer of the personal **data abroad** is secured only for transfers to countries [6,7] whose legislation ensures an adequate level of protection
- Guarantee these subjects’ **minimal rights**: right of access, rectification, portability, to object, erasure

Any doubt? Contact the EPFL Human Research Ethics Committee [8]

[1] www.admin.ch/opc/en/classified-compilation/19920153/index.html#a3 / [2] www.admin.ch/opc/en/classified-compilation/20061313/index.html /
[3] www.edoeb.admin.ch/edoeb/fr/home/protection-des-donnees/Internet_und_Computer/services-en-ligne/medias-sociaux.html /
[4] http://data.europa.eu/eli/reg/2016/679/2016-05-04 / [5] https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611236 /
[6] www.edoeb.admin.ch/dam/edoeb/fr/dokumente/2018/staatenliste.pdf.download.pdf/20181213_Staatenliste_f.pdf /
[7] ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en /
[8] www.epfl.ch/research/ethic-statement/human-research-ethics-committee

Bibliothèque Contact and info go.epfl.ch/rdm researchdata@epfl.ch

cc BY NC SA

EPFL

DOI: 10.5281/zenodo.3327829

Research Data Management

DATA MASKING

Fast Guide #09

Data masking, or data obfuscation, is the process of **hiding original data** with modified content [1]

ADVANTAGES

WHY IT'S WORTH

- Complies with law
- Makes data sharable
- Prevents data misuse
- Makes data publishable

APPLICABILITY

TESTS ON HUMANS / SENSITIVE DATA

- Name, identification number, location data, online identifier, etc.
- Factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity

PSEUDONYMIZATION ≠ ANONYMIZATION

PSEUDONYMIZATION

(FOR ACTIVE DATA)

REVERSIBLE

● REPLACING

Replace data by identifiers. The key is stored separately and securely.

● ENCRYPTING

Encrypt the data and store the key securely. Appropriate for long-term preservation, not for data publishing.

ANONYMIZATION

(FOR PUBLISHED DATA)

IRREVERSIBLE

● GENERALIZING

Diminish granularity by generalizing the variables. Appropriate for data too specific or unique records.

● SHUFFLING

Shuffle data over one / several columns without compromising their utility.

● FAKING

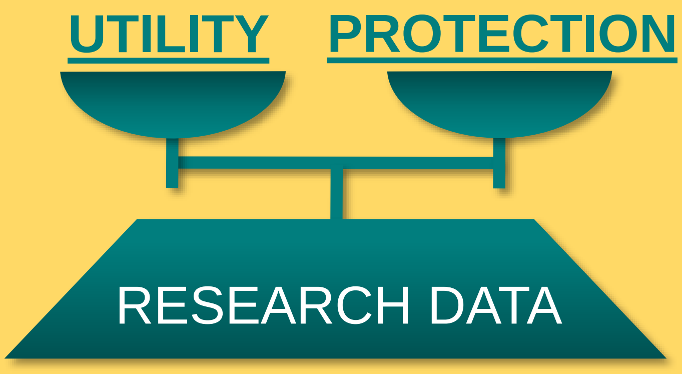
Prevent the identification of specific records, adding fake data while preserving correlations.

● REMOVING

Suppress data or part of the outlier records. Appropriate for processing identifiers.

3RD PARTY DATA

- Do you use **commercial dataset**? Or **collaborate** in a joint research?
- **Define a contract** for data use, sharing and publication.
- **Distinguish** between research authors and data owners.



HINT

Mitigate the identification risk, but preserve the data utility for research.

SOME TOOLS

MASK IDENTITY OR ASSESS IDENTIFICATION RISKS

- ARX Data Anonymization Tool (Java) [2]
- Amnesia (online) [3]
- ARGUS (Java) [4]
- sdcMicro (R) [5]
- Differential privacy queries (SQL) [6]
- Faker (Python) [7]

SUPPORT AND LAWS

- EPFL Research Ethics [8]
- Federal Act on Data Protection [9]
- Human Research Act [10]
- GDPR [11]

[1] en.wikipedia.org/wiki/Data_masking / [2] arx.deidentifier.org / [3] amnesia.openaire.eu / [4] qosient.com/argus/anonymization.shtml / [5] cran.r-project.org/web/packages/sdcMicro/index.html / [6] github.com/uber/sql-differential-privacy / [7] faker.readthedocs.io/en/master / [8] www.epfl.ch/research/ethic-statement / [9] admin.ch/opc/en/classified-compilation/19920153/index.html / [10] admin.ch/opc/en/classified-compilation/19920153/index.html / [11] eur-lex.europa.eu/eli/reg/2016/679/oj / Icons: cran.r-project.org/web/packages/sdcMicro/index.html

BY NC SA

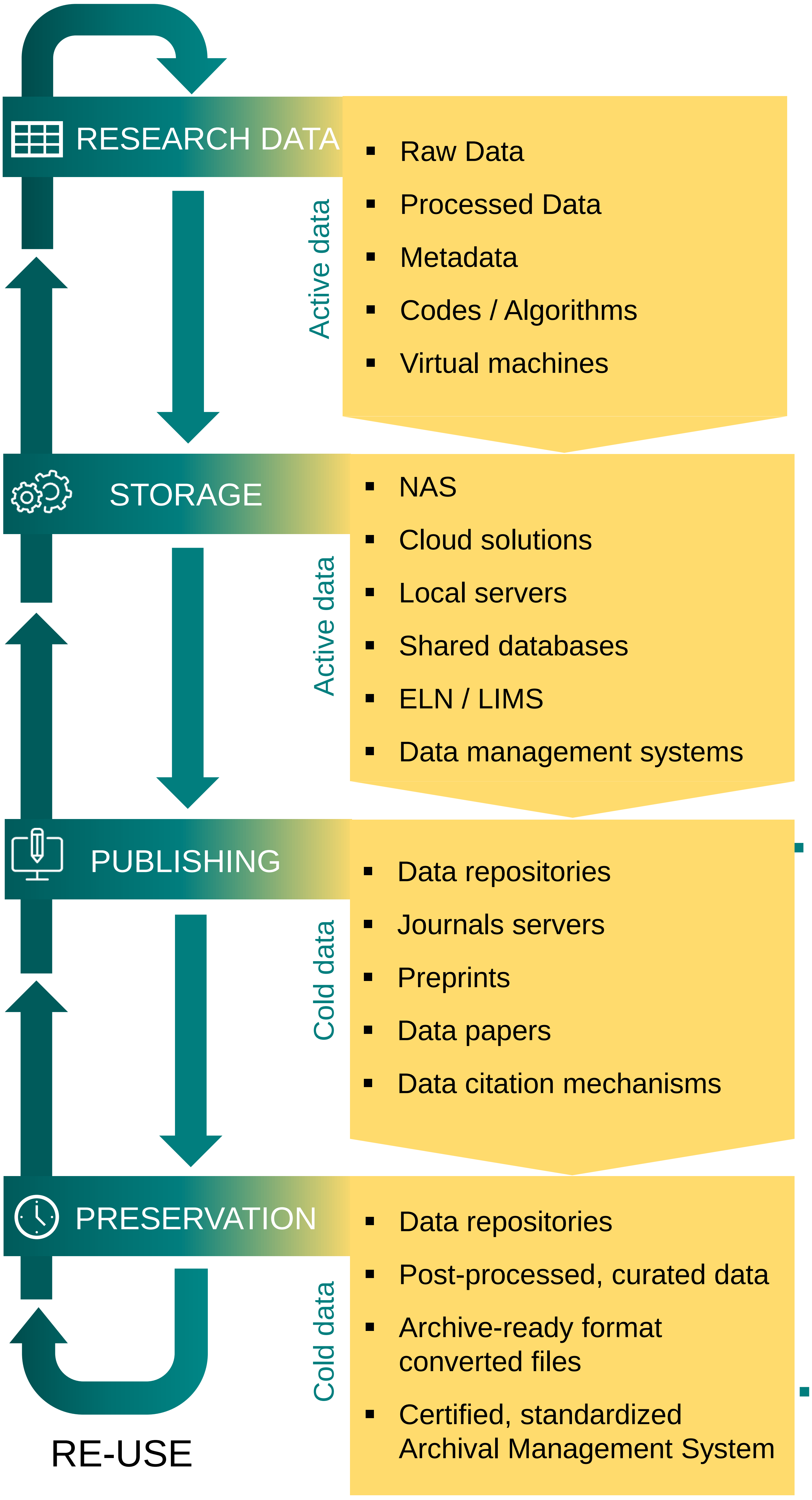
EPFL

DOI: 10.5281/zenodo.3327829

Research Data Management

STORE, PUBLISH, PRESERVE DATA

Fast Guide #10



RESEARCH DATA

Active data

- Raw Data
- Processed Data
- Metadata
- Codes / Algorithms
- Virtual machines

STORAGE

Active data

- NAS
- Cloud solutions
- Local servers
- Shared databases
- ELN / LIMS
- Data management systems

PUBLISHING

Cold data

- Data repositories
- Journals servers
- Preprints
- Data papers
- Data citation mechanisms

PRESERVATION

Cold data

- Data repositories
- Post-processed, curated data
- Archive-ready format converted files
- Certified, standardized Archival Management System

RE-USE

STAKEHOLDERS

- Research teams
- Institutions
- Funders
- Research partners
- Private partners
- IT services providers

Publishing conditions

- Data ownership
- Stakeholders consent
- Compliance with protection laws
- Ensuring data integrity
- Providing appropriate metadata
- Clarifying reuse licensing
- Setting up embargoes and sampling rules, if needed

Preserving criteria

- Historical and scientific data value
- Data quality and uniqueness
- Reliability of sources
- Data preparation cost
- Repository and maintenance cost
- Deposit responsibility

How long to preserve?

- At least 10 years for the SNSF
- Evaluate preserving criteria
- Mind any retention and disposal schedules
- Stick to administrative and legal requirements

KEEP IN MIND: Store ≠ Backup ≠ Preserve ≠ Publish ≠ Archive

BY NC SA

EPFL

DOI: 10.5281/zenodo.3327829

Research Data Management

DMP – DATA MANAGEMENT PLAN

Fast Guide #11

WHY A DMP?

COMPLIANCY Requested by research funders (public or private), a DMP enhances research reproducibility and the use of public funds.

TRANSPARENCY Usually published when the funding period ends, a DMP completes the research results with the information on data, software, protocols, sources, etc.

FORECAST To anticipate costs (materials and software) and identify risks (eg. data loss, incompatible formats, security). DMPs allow institutions to better allocate services.

STREAMLINE To reduce risks of data loss and the efforts of reverse engineering for new collaborators. A DMP boosts data reuse in the lab and outside.

Target the reproducibility of research results!
Anticipate questions about data in your projects.

WHAT'S IN A DMP?

DESCRIPTION Data types, formats, size.

COLLECTION Sources, experiments, analysis, simulations.

CURATION Metadata, naming, datasets structures.

STORAGE Active data, sharing tools, preservation.

RISKS Access rights, anonymization, ethics assessment.

PUBLICATION Data licenses, data repositories, IP.

COSTS For RDM: refer to Fast Guide #03.

Not just administrative hurdle! Use your DMP as reference tool for in-lab discussions & decisions.

WHEN A DMP?

IDEALLY At the conception of your research project.

USUALLY When requesting funds.

REALLY ASAP, but it is never too late.

The DMP is a living document!
Keep it up-to-date throughout the project.

A DMP IS…

… a written document describing how data of a research project is managed during its life-cycle

FUNDERS REQUIRING A DMP

- SNSF
- H2020 (ERC, FET, MSCA, ...)
- EPFL (some internal projects)
- AXA Research Fund
- U.S. Federal Grants
- Wellcome Trust
- Ligue Vaudoise Contre le Cancer
- CCR-pro

DMP TEMPLATES [1]

- SNSF DMP
  A template based on SNSF Open Research Data Policy, with added guiding examples.
- ERC DMP
  A template based on the FAIR principles, with added guiding examples.
- MSCA DMP
  Suggested for Marie Skłodowska-Curie Actions' applicants.
- NCCR RDM STRATEGY
  More than a DMP, it describes the data management for all the projects of a NCCR.

[1] www.epfl.ch/campus/library/services/services-researchers/rdm-guides-templates / Icons: www.flaticon.com/packs/essential-set-2

BY NC SA

EPFL

DOI: 10.5281/zenodo.3327829

Research Data Management

DATA & CODE LICENSING

Fast Guide #12

LICENSING applies to…

- both data and code
- collected, processed, aggregated, augmented data/code
- original work, or from another researcher

At EPFL?

EPFL owns the original data & code, but the authors can use it for research and IP [8]

WHY licensing data & code?

- Collaboration with other researchers, research groups, institutes
- Reuse of others' work, to generate new information or data/code
- Clarify ownership and authorship
- Differentiate allowed use of original or derived work
- Share / Publish your work with clear usage rights

LICENSES FOR DATA

CC-ZERO	No restrictions [1]
CC-BY	Mandatory citation
CC-BY-SA	Share Alike (Viral), Mandatory citation
CC-BY-ND	No Modifications, Mandatory citation
CC-BY-NC	Non Commercial, Mandatory citation
CC-BY-NC-SA	Non Com., No Mod, Mandatory cit., Viral
ODBL	Open Access specific for databases [2]
MICRODATA RESEARCH LICENSE	Unit-level data [3,4]

LICENSES FOR CODE [5]

MIT	Short term, Permissive, No warranty
APACHE	Permissive, Patents allowed, No warranty
GPL	Copyleft license, Patents allowed, Viral
LGPL	Libraries Sharing, Licences mix allowed
AGPL	Strong copyleft, Patents allowed, Viral
BSD	All code by one organisation, GPL mix not allowed

The protection of data by law is **not harmonized internationally**, but varies depending on the specific country. Licenses have not all the same international recognition.

NEEDS and ISSUES [6]

Data licenses present several issues:

- Data ownership and use
- The treatment of original and derived data

This is because a researcher may:

- Receive or collect and compile data from another researcher
- Generate information or data from the other researcher's data on that other researcher's, or its own, behalf

Moreover, a researcher may want to:

- Analyze/reuse another researcher's data
- Process/aggregate data for own research, using the processed data
- Licensing the processed, aggregated, augmented data from another researcher

Any doubt? Contact the EPFL Technology Transfer Office [7]

[1] creativecommons.org / [2] opendatacommons.org / [3] microdata.worldbank.org/index.php/terms-of-use / [4] www.w3.org/TR/microdata / [5] choosealicense.com/appendix / [6] Sharing Research Data and Intellectual Property Law: A Primer https://doi.org/10.1371/journal.pbio.1002235 / [7] www.epfl.ch/research/services/units/technology-transfer-office / [8] www.admin.ch/opc/en/classified-compilation/19910256/index.html#a36 /

Icons: www.flaticon.com/packs/development-62

Bibliothèque   Contact and info   go.epfl.ch/rdm   researchdata@epfl.ch

cc BY NC SA