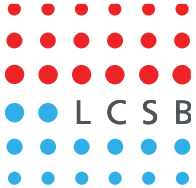
Data Management Planning

Pinar ALPER

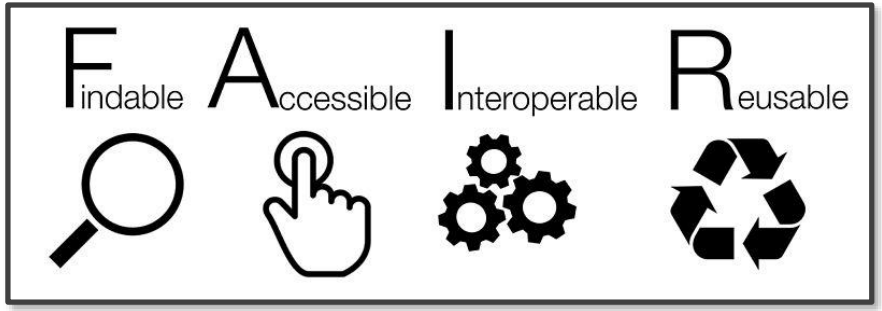


Training on "Best practices in research data management and stewardship"

14 June 2021

What is a DMP?

- formal document asking you to document your “(good) data management”
- projects with (good) data management produce “FAIR”, such data would have longevity.



Wilkinson M, Dumontier M et al. Nature Scientific Data 2016. "The FAIR Guiding Principles for scientific data management and stewardship"

Why DMPs?

— Funder requirement

- Ensure data management via DMPs.
- Researchers are accountable for how data is treated during and after the project.

- timely release of data - once patents are filed or on (acceptance for) publication
- (open) data sharing - minimal or no restrictions if possible
- preservation of data - typically 5-10+ years

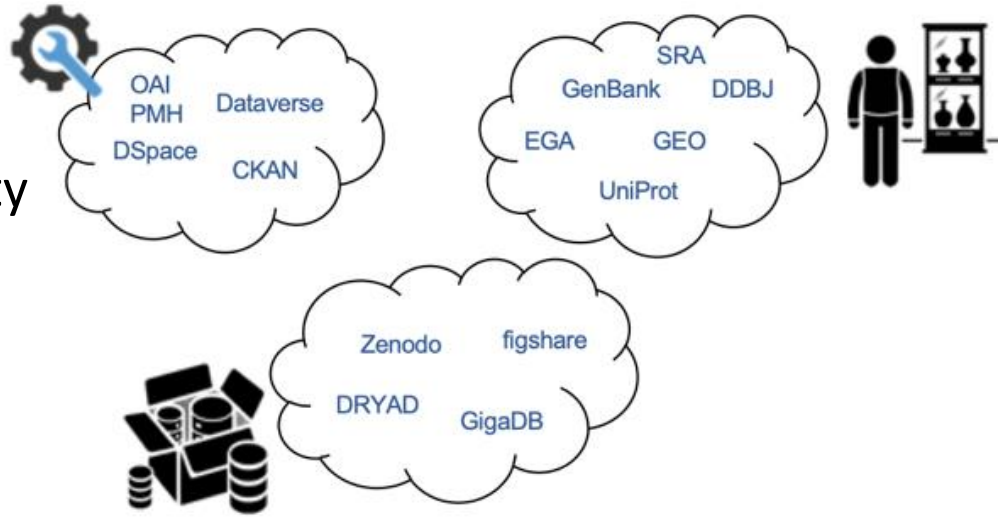
Introduction to Data Management. Joy Davidson. UK Digital Curation Centre. 2015

Practical Guide to the International Alignment of Research Data Management. Science Europe November 2018

Data as a 2nd class citizen

—decades of efforts, varying levels of "FAIR"ness

- Maybe on publication in peer-review (if considered)
- Buried in "Materials and Methods", PDFs, ZIPs
- Not consistently preserved.
- Low interoperability and re-usability

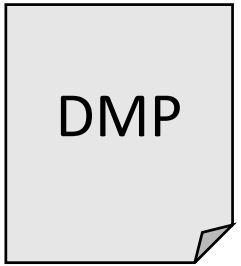


"Turning FAIR into reality" Final Report from the EC Expert Group on FAIR Data

Making data a 1st class-citizen in research

- A change in research culture and funding
- DMP is one such intervention

+



+

Data Steward

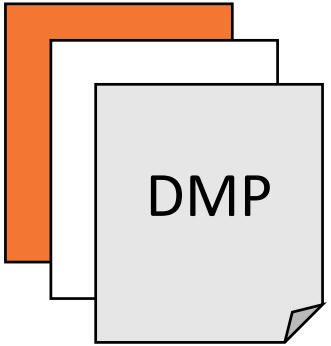
A DMP is shaped by

(Funder)
Requirements

Host
Infrastructure

Consortium/
Project

Your
Research



tip

An example goes a long way. Make use of DMP archives.

Avoid copy-paste.

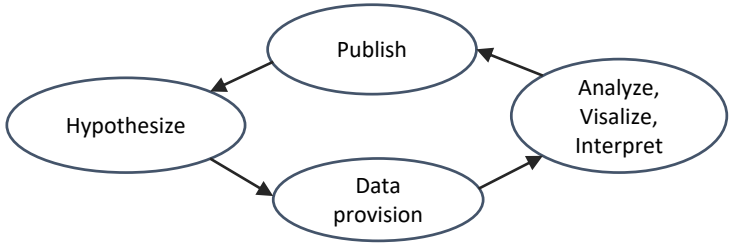
Be short and specific.

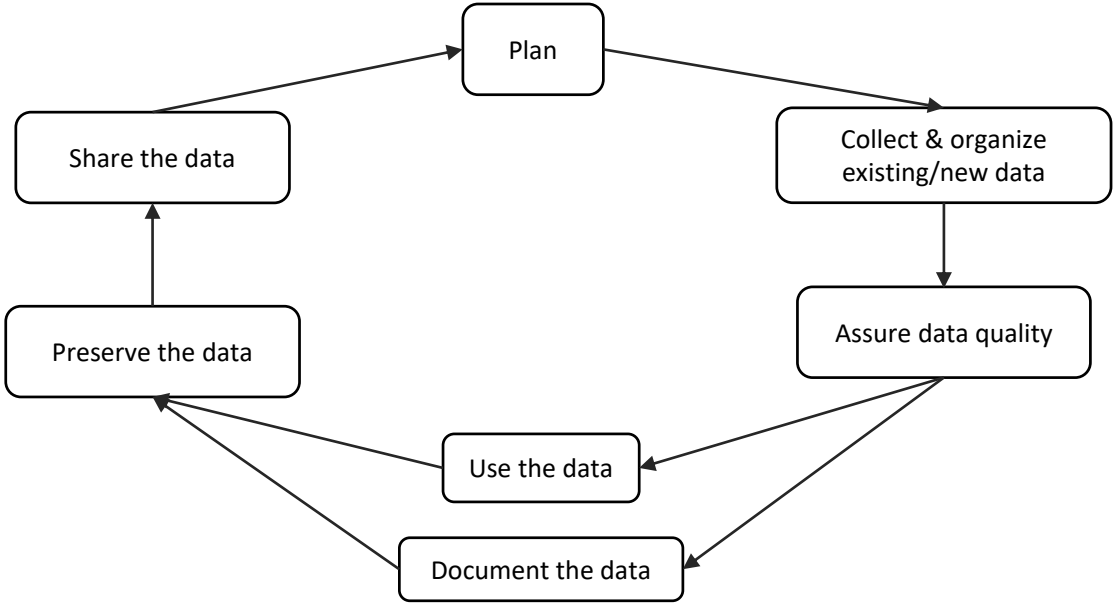
https://phaidra.univie.ac.at/o:1140797

LCSB

elixir LUXEMBOURG

The DMP world view





Adapted from "Ten Simple Rules for Creating a Good Data Management Plan". W Michener

Identify data

- What constitutes "data"
  - Primary/Derived data
  - Research Record
  - Accompanying documents

- Type, structure, format, estimated size
  - Type: Text, numeric, synthetic, image
  - Format: generic/discipline-specific

tip

To avoid data creep, identify data as early as and as thoroughly as possible, ideally during consortium setup.

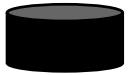
Use of open and standard formats for preservation. **MS Excel vs CSV.**

Use of proprietary formats must be justified. Normally, not suitable for preservation.

Collection and/or re-use of data

- Re-use of existing data

from-repo



from-collaborator

- Newly generated data

from-cohort

  - Sources,
  - Process of collection; instrument, kit, software, method
  - Periods of capture and updates

tip

High value data, e.g. one-time events, costly collection, validation studies

Identify data utility during and after project, potential re-use.

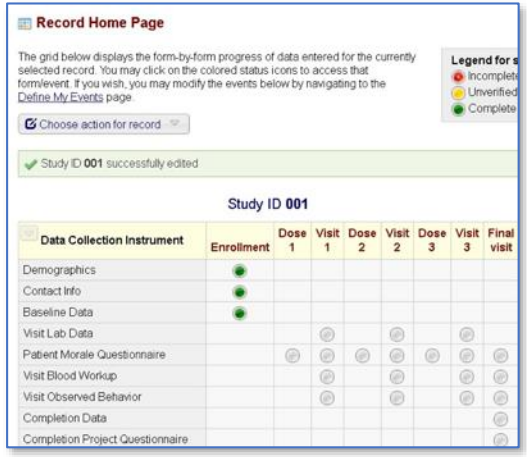
Highlight re-use, justify generation.

Data Processing, Quality Assurance and Control

- Data Quality is observed as a factor increasing data-reuse
- Automated or manual QA/QC measures
  - tool/pipeline/dashboard
  - training, standards
  - calibration, repeated samples, peer-review

data analysis

data wrangling



tip

Strong statements on potential re-usability of data will bring about increased expectation on QA/QC processes

LCSB

elixir LUXEMBOURG

Documentation of data

- Metadata: Information enabling the read and interpretation of data.
  - It is a requirement for publicly shared data.
  - It is commonly asked for during (data) peer-review.
  - What metadata will you record for your project's data
    - Bibliographic
    - Domain specific
- **Provenance** is THE key piece of metadata enabling the ultimate re-use of data.

"the origin, source; the history of ownership of a valued object or work of art or literature" Mirriam Webster

Bibliographic Metadata

DRYAD

Ex

Data from: Temporal enhancer profiling of parallel lineages identifies AHR and GLIS1 as regulators of mesenchymal multipotency

Sinkkonen, Lasse

Publication date: April 24, 2019
Publisher: Dryad
https://doi.org/10.5061/dryad.r32t3

Citation

Sinkkonen, Lasse (2019), Data from: Temporal enhancer profiling of parallel lineages identifies AHR and GLIS1 as regulators of mesenchymal multipotency, Dryad, Dataset, https://doi.org/10.5061/dryad.r32t3

Usage Notes

Time point-specific gene regulatory networks of mesenchymal differentiation
The full matrix of the time point-specific TF-target gene interactions for 6 time points of both adipocyte and osteoblast differentiation of bone marrow msenchymal stem cells as derived and used as input for EPIC-DREM analysis in publication by Gerad et al.
EPIC-DREM_Input_GRNs-Adipo_Osteo.zip

References

This dataset is supplement to https://doi.org/10.1093/nar/gky1240

License

LCSB

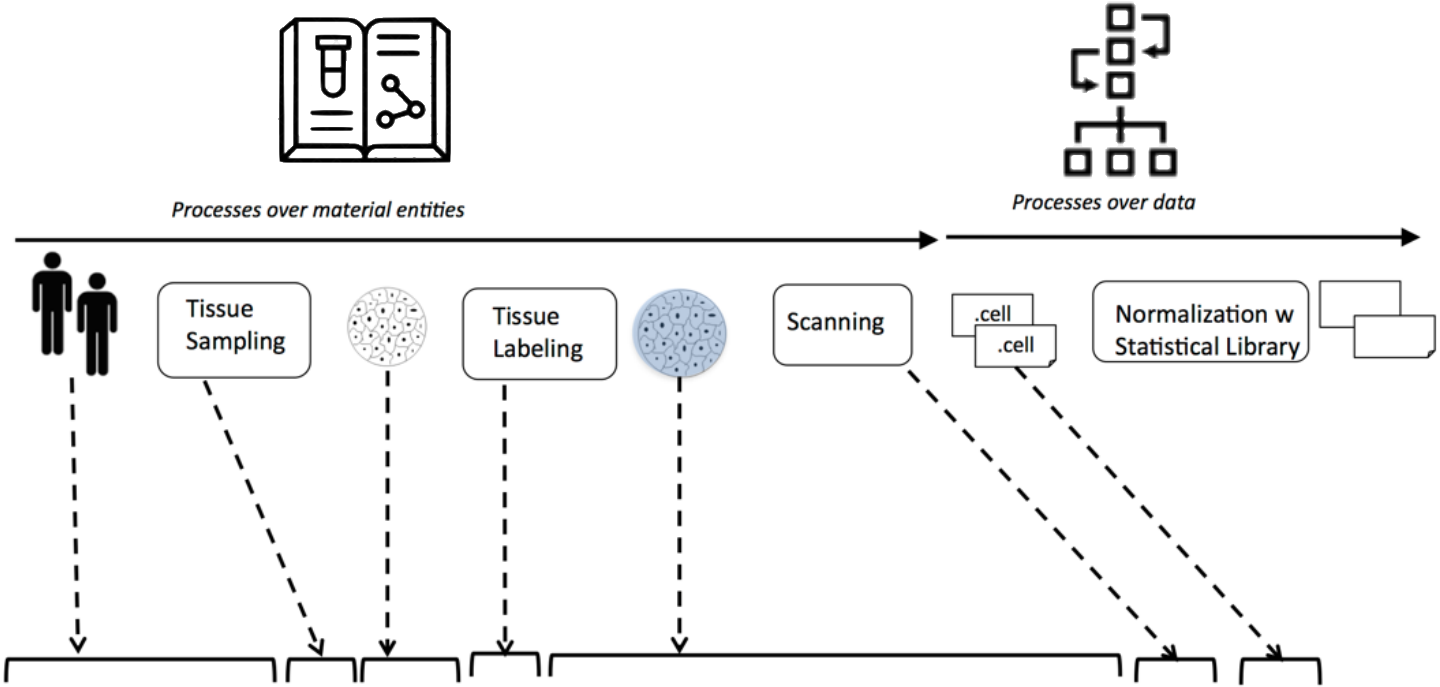
elixir LUXEMBOURG

Documentation of data

- Metadata: Information enabling the read and interpretation of data.
  - It is a requirement for publicly shared data.
  - It is commonly asked for during (data) peer-review.
  - What metadata will you record for your project's data
    - Bibliographic
    - Domain specific
- **Provenance** is THE key piece of metadata enabling the ultimate re-use of data.

"the origin, source; the history of ownership of a valued object or work of art or literature" Mirriam Webster

Data provenance



Source Name	Organism	Age	Unit	Protocol REF	Sample Name	Protocol REF	Labeled Extract Name	Label	Term Source REF	Term Accession Number	Protocol REF	Data File
H1	H.Sapiens	35	Years	Sampling	H1.Samp1	Labeling	H1.Samp1	biotin	CHEBI	http://purl.obolibrary.org/obo/CHEBI_15956	Scanning	H1.Samp1.cel
H1	H.Sapiens	35	Years	Sampling	H1.Samp2						Scanning	H1.Samp2.cel
H2	H.Sapiens	35	Years	Sampling	H2.Samp1	Labeling	H2.Samp1	biotin	CHEBI	http://purl.obolibrary.org/obo/CHEBI_15956	Scanning	H2.Samp1.cel

. . .

Basic domain-specific metadata

ENA Beta
European Nucleotide Archive

Home | Submit | Search | Rulespace | About | Support

You are using the new ENA Browser. To see the corresponding view in the old ENA Browser, please click https://www.ebi.ac.uk/ena/data/view/PRJEB

Project: PRJEB20933

We have generated time-series transcriptomic and epigenomic data during the differentiation of bone mar
shared mesenchymal precursor cells using RNA-seq and ChIP-seq for several histone modifications,
networks underlying these differentiation processes, and to better understand their dynamics over time
different time points during both of the 15-day differentiation processes and active enhancers (H3K27a
regions (H3K36me3) were mapped in both lineages. The identified time point-specific open chromatin
factor binding affinities and a novel machine learning approach was used to build dynamic regulatory n
series data. In parallel, to further prioritize the identified regulatory genes we mapped super-enhancers v

Show More

Secondary Study Accession:	ERP023143
Study Title:	Temporal epigenomic and transcriptomic profiling of mesenchymal differenti
Center Name:	Life Sciences Research Unit
Study Name:	Temporal epigenomic and transcriptomic profiling of mesenchymal differenti

Related ENA Records

Result	Count
Experiment	269
Run	269

Experiment: ERX2068030

Illumina HiSeq 2000 sequencing

Organism:	Mus musculus (house mouse)
Experiment Accession:	ERX2068030
Sample Accession:	SAMEA104124576
Instrument Platform:	ILLUMINA
Instrument Model:	Illumina HiSeq 2000

Show More

Read Files

Show selected columns

Download report: JSON TSV

Download Files as ZIP

Download selected

Download All

Study Accession	Sample Accession	Experiment Accession	Run Accession	Tax Id	Scientific Name	FASTQ FTP
PRJEB20933	SAMEA104124576	ERX2068030	ERR2008267	10090	Mus musculus	ERR2008267.fastq.gz

not the complete provenance

LCSB

elixir LUXEMBOURG

Minimum Information about a high-throughput SEQuencing Experiment

ENA
European Nucleotide Archive

Home | Search & Browse | Submit & Update | Software | About ENA | Support

ENA > Submit & Update > Epigenomics submissions

Submitting epigenomic data

This page details a checklist of minimal information that we expect from data submitters to the European Nucleotide Archive (ENA) when describing raw data sets from next generation sequencing platforms used in high-throughput studies of epigenetic features. We present this checklist in order to practically assist those preparing their data for submission to the ENA. We do not propose that the information described as mandatory in the list below is necessarily sufficient for successful reproduction of experimental findings and wish to note that the broader reporting standard, MINSEQE, exists that serve this purpose. Since information additional to the minimal checklist presented here may be required for MINSEQE compliance and to raise the level of utili sources will also be in use by our epigenomics data submitters and that the several components of reporting.

Checklist fields for study

Field
Mandatory fields
M-1. Study title
M-2. Investigator name
M-3. Investigator e-mail
M-4. Center name
M-5. Study description
Recommended fields
R-1. Study type
Optional fields
O-1. Release date

Checklist fields for sample and sample processing

Field	Description
Mandatory fields	
M-1. Taxonomic identifier	Species or infraspecies taxonomic name of the sar the NCBI Taxonomy. More information about taxor
M-2. Strain name	Strain name of the sampled organism, for prokary
M-3. Cell line	Name of the cell line, if used.
Recommended fields	
R-1. Organ or tissue source	Organ or tissue source of the sampled material.
R-2. Epitope tag	Details of epitope tagging approach, if used, inclu level.
R-3. Cell line growth conditions	Cell line growth conditions and characteristics, suc density.
R-4. Physical sample source	Physical source of sample, such as stock centre ar
Optional fields	
O-1. Phenotype attributes	Phenotypic attributes of the sampled organism of

Checklist fields for sequencing data

Field	Description
Mandatory fields	
M-1. Data files	Fastq-formatted data files or aligned BAM files, in which case the sequence should be indicated within the BAM file.
M-2. MD5 checksum	MD5 checksum for each data file.

Checklist fields for sequencing library

Field	Description
Mandatory fields	
M-1. Experimental design description	A brief experimental design description.
M-2. Epigenomics method	The epigenomics method that has been used, such as *ChIP*
M-3. Library source	The library source; expected to be *genomic.*
M-4. Library selection	The method of library selection, such as *5-methylcytidine a*
M-5. Antibody name	Antibody name, if used.
M-6. Library layout	The library layout; expected to be unpaired reads.
M-7. Platform/Model	Sequencing vendor platform and instrument model, such a
Recommended fields	
R-1. Post amplification validation	Description of post-amplification validation steps to ensure
R-2. Antibody lot number	The antibody lot number.
R-3. Antibody provider	The source of the antibody.

tip

Proper documentation will take time, poor documentation reduces the re-usability of data.

Documentation of data

- Where will you store metadata during and after the project?
  - Lab book, virtual laboratory app/database, readme, good old filename,
- Any applicable metadata standards,
- e.g. data dictionaries, standard identifiers, minimum information guidelines
  - MINSEQE: Minimum Information about a high-throughput SEQuencing Experiment
  - MIAMI: Minimum Information about a Microarray Experiment
  - MIARE: Minimum Information About an RNA interference Experiment

LCSB

tip

Aim to collect provenance during data generation.
Aim to automate provenance collection. E.g. Analysis with several samples and/or runs.

FAIRsharing.org
standards, databases, policies

Ethics Compliance

- Is ethics/IRB approval required and applied for?
  - H2020 Ethics self-assessment checklist
- All research with human data and biosamples requires an ethics approval.
- For projects with human data ethics committee may ask for:
  - DMP or DMP paragraph
  - Consent form template and subject information sheet
  - The data protection concept and/or the Data Protection Impact Assessment

Legal Compliance

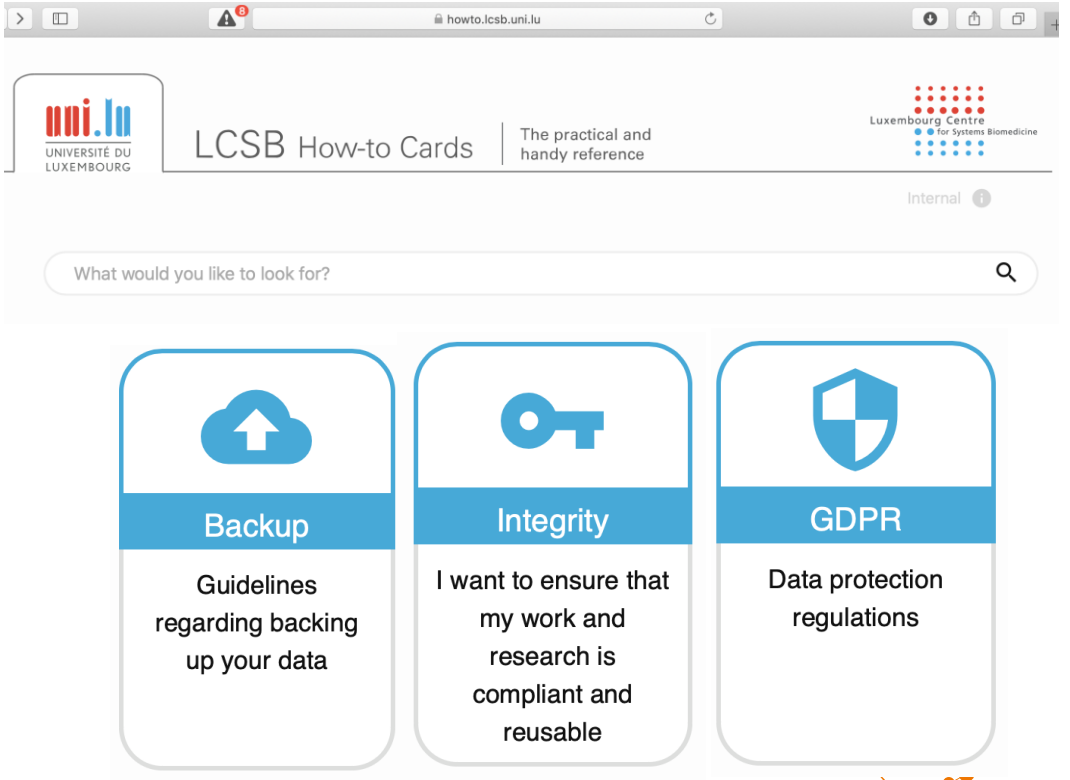
- What are the applicable regulations to your data? **GDPR**, **IP Laws**.
- GDPR: What Data Protection Concept will be applied?
  - For some projects a Data Protection Impact Assessment
  - Arrangements for the recording of Data Processing
  - Arrangements to handle data subjects' requests?
- More on tomorrow's session on "Data protection in research"

tip

Seek support from Data Protection Officer, Legal Office and Ethics Board during the planning phase

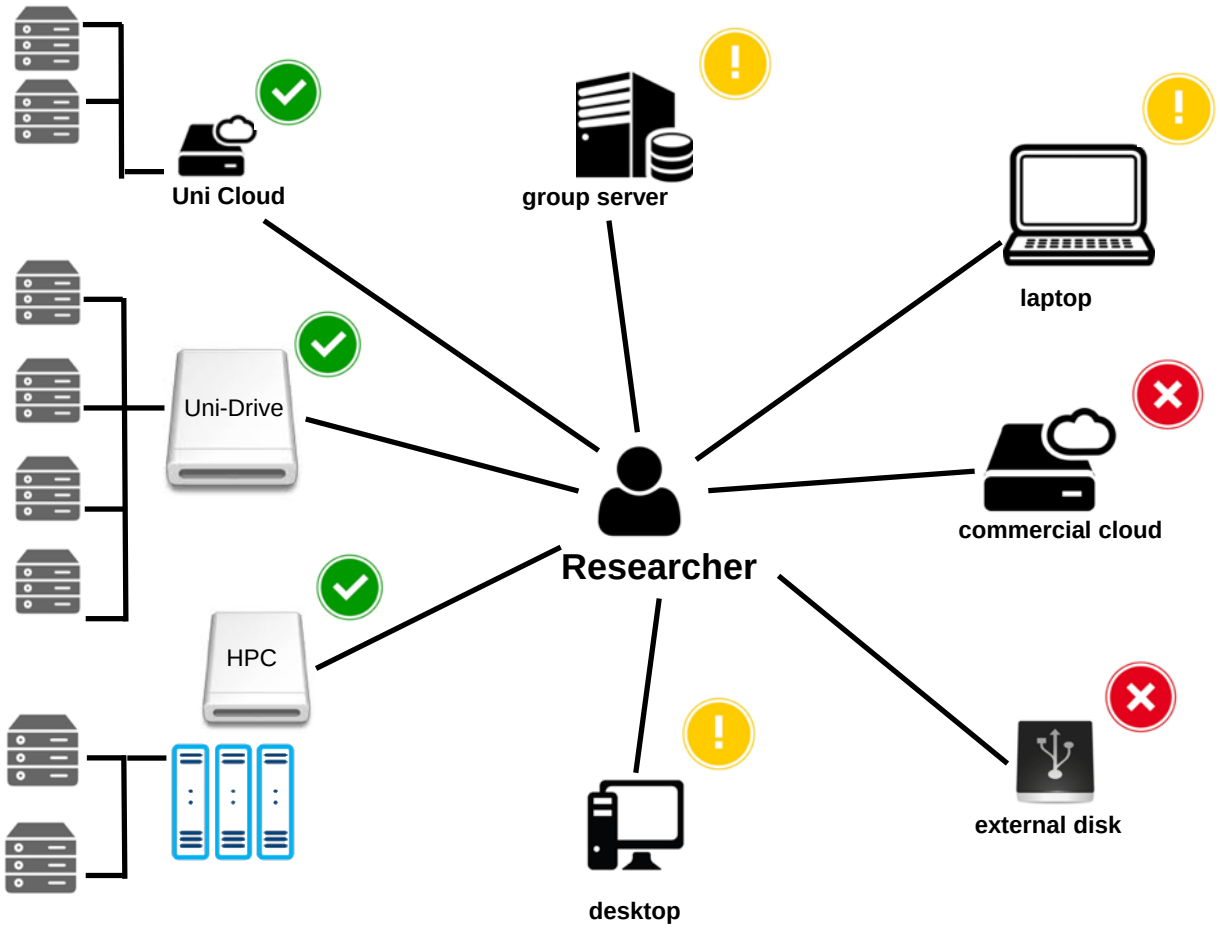
Storage and backup during research

- Where will data be stored during the project.
  - **Institutional** and/or **project-specific** resources?
  - Are there policies, guidelines? (Storage, Backup, Retention, Deletion)
- What is the backup and recovery process?



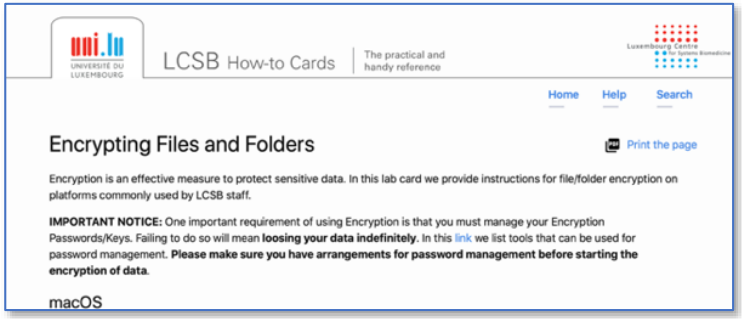
tip Data & IT Support Offices can give you directions.

Example guidance…



Security & Privacy

- Does your institute's data centre have IT security certification?
  - Encryption, Access Control, Password Management, Single Sign On, Multi-Factor Authentication, endpoint protection
  - If privacy is a concern then anonymisation, pseudonymisation
- In case of no certifications policy and guidelines play a role.



tip

Refer to standards where possible.
e.g. ISO 27001, NIST SP 800-37, Privacy-By-Design Certifications

Preservation

Data Archive

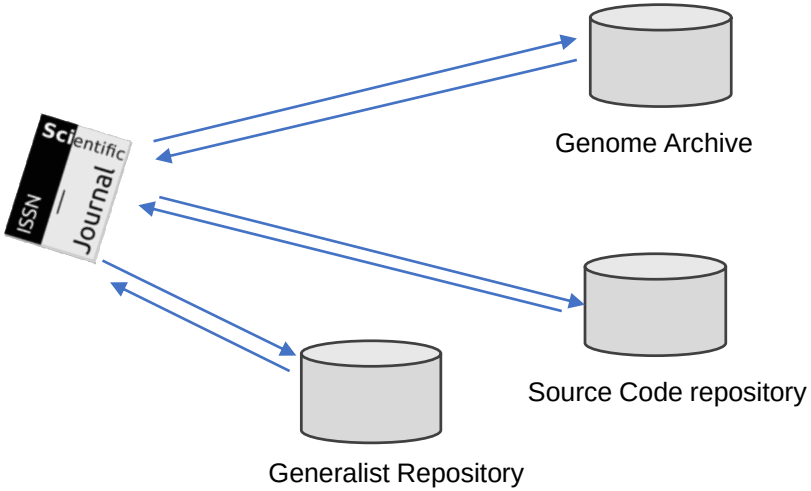
- Storage != Preservation
  - “preservation is the act of conserving and maintaining both the safety and integrity of data.” wikipedia
- In the DMP you should identify
  - Which data will be preserved after the project? What is the retention policy?
  - Which data will not be preserved (needs to be destructed, e.g. due to storage restrictions etc.)
  - Data preserved for how long? In what form?
  - Where;
    - Generalist repository
    - Community database
    - Institutional repository

re3data.org

FAIRsharing.org
standards, databases, policies

Preservation

- Persistent identifiers: DOI, accession numbers?
- Will there be multiple releases/versions for data and code?
- Preserving in multiple repositories; will there be data landing pages?



Copyright, Licensing, Access

- Copyright: Legal term for ownership
- Particularly important in public-private partnerships
- Example ownership policy:

Work	**Owner**
Literature	Researcher
Software	Institute
Data(base)	Institute

Copyright, Licensing, Access

- License: Terms under which others may use copyrighted material

Open Data Commons *Legal tools for Open Data*

- PDDL - Public Domain Dedication and License (PDDL)
- ODC-By - Attribution License
- ODC-ODbL - Open Database License

creative commons

- CC0 - Universal (v 1.0) Public Domain Dedication
- CC BY - Attribution 4.0 International
- CC BY SA - Attribution-Share Alike 4.0 International

tip

https://eudat.eu/services/userdoc/b2share-usage
http://www.dcc.ac.uk/resources/how-guides/license-research-data

Sharing and Access Levels

- If, when and how will you share data?

  Community/Generalist Repository, a Data Paper
- Any embargo periods?
- Can data be accessed by everyone in the public domain?
  - Will Registered/Controlled access be adopted?
  - Data Access Orchestration process,
  - Data Access Comittee

tip

Data sharing plan may be detailed at later stages of a project.

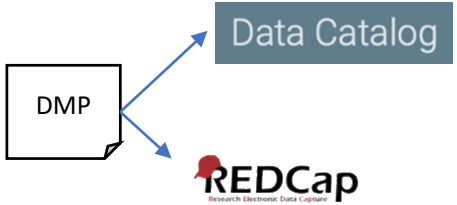
Some repositories handle preservation & DAC workflow, whereas others only provide preservation.

DMP as a living document

- You will be expected to update your DMP
- DMP paragraph, at proposal stage
- First full DMP, often at month 6
- Thereafter;
  - Periodic review
  - New data
  - Change in policy
  - Change in consortium co

tip

Separate frequently changing parts of DMP into "dynamic references"



Budgeting for Data Management

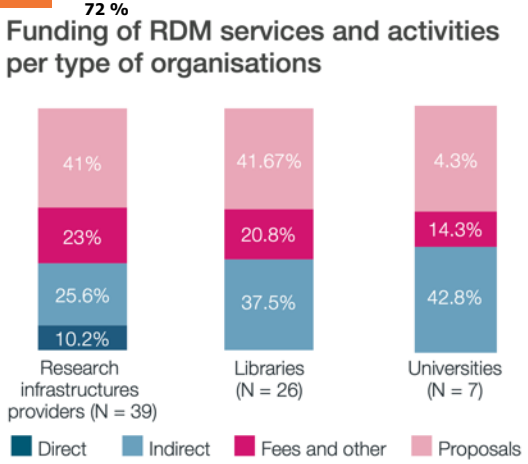
— 5% of total project costs

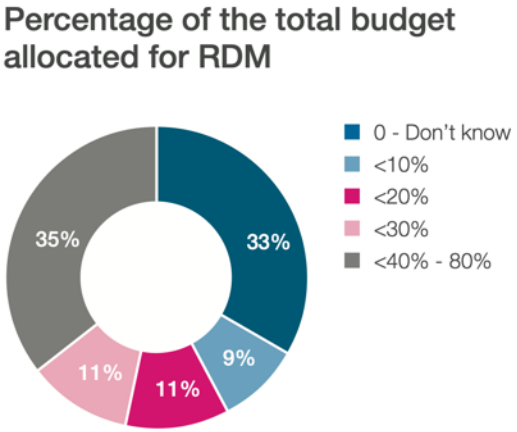
An 2016 survey:

- Infra providers, libraries, universities
- What percentage of total budget of your organization is allocated for RDM?

A commonly cited recommendation:

- An overall average of **5% of the total project costs** ……to sustain and share data”





*** Funding research data management and related infrastructures.* Knowledge Exchange and Science Europe briefing paper. May 2016

tip

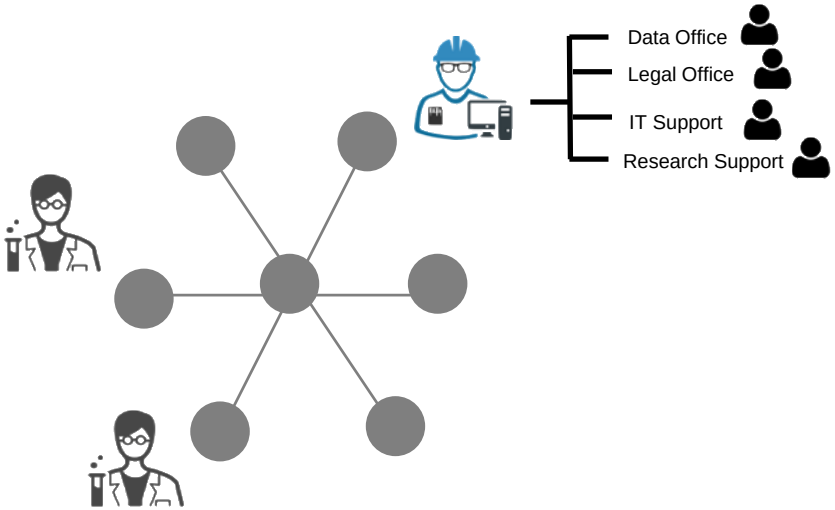
Search keywords “Research Data Management” + “Costing”: https://www.openaire.eu/how-to-comply-to-h2020-mandates-rdm-costs

Paid data management software also goes in to the DMP budget.

** *Funding research data management and related infrastructures.* Knowledge Exchange and Science Europe briefing paper. May 2016

Roles and responsibilities

— DMP is primarily a responsibility of researchers



- In big consortia one partner may take on the DMP responsibility
- Make sure partners responsibilities are documented in the DMP
- Gets support from institutional data support offices

DMP templates

- Skeletal documents containing the necessary headings for DMPs required by funders or organisations
- Template could be presented as a list of questions.
- Templates can be "machine-actionable" allowing export-import among tools.
- Commonly used templates:
  - Science Europe DMP Template
  - EU H2020 and ERC Templates
  - Machine-actionable DMPs Common Standard

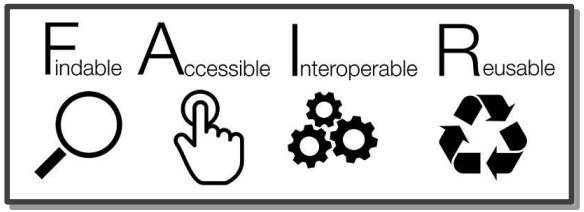
tip

For a one-stop-shop of DMP guidance links check out
https://rdmkit.elixir-europe.org/data_management_plan.html

FNR Policy on Research Data Management

- The Luxembourg National Research Fund (FNR)
- Applies to all FNR-funded projects, 1 January 2021 onwards
- Provides a DMP Template aligned with Science Europe Guidelines
- "We expect researchers to maximise the availability of research data with as few restrictions as possible.... The key principle that applies is "as open as possible, as closed as necessary.""
- "Research data should be deposited in a trusted repository in such a way that the data are as findable, accessible, interoperable and reusable (FAIR) as possible."

OPEN ACCESS



Wilkinson M, Dumontier M et al. Nature Scientific Data 2016. "The FAIR Guiding Principles for scientific data management and stewardship"

https://www.fnr.lu/open-science-new-fnr-policy-on-research-data-management/

DMP tools

- Software tools used to support the Data Management Planning process

Researcher

Funder compliant DMP writing
DMP sharing, authorship credit
Learning RDM

Data Steward

RDM Teaching/Awareness raising
Funder template dissemination
DMP repository establishment

DMP tools

- roadmap (UK DCC, UC3)

  https://dmponline.dcc.ac.uk

  https://github.com/DMPRoadmap/roadmap/wiki

Morning practical.

- DSW (ELIXIR)

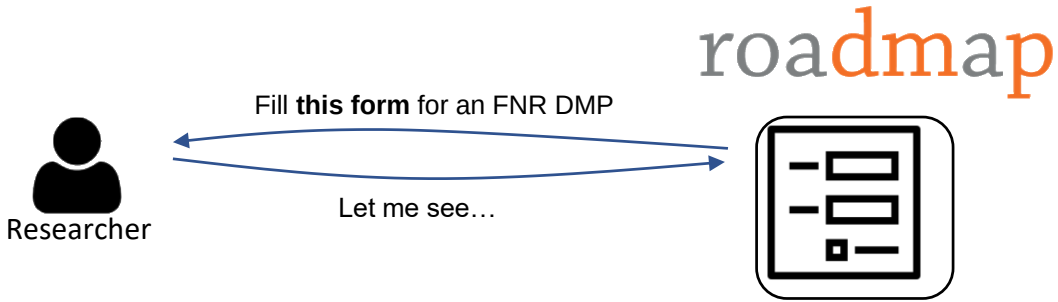
  https://ds-wizard.org

  https://github.com/ds-wizard

Afternoon practical.

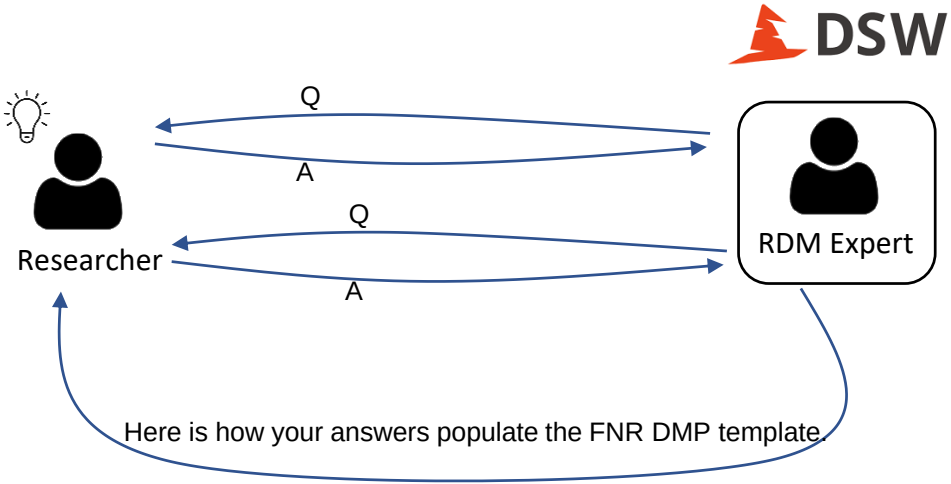
DMP Roadmap

- Template-based approach



Data Stewardship Wizard

- “Expert system” approach
- Contains decision trees
- Great for DMP beginners
- DMP is a side product of the learning experience



Should you use a DMP tool?

- Not a substitute to collaborative document authoring tools.
- Not a substitute to survey/questionnaire tools.
- Not integrated to grant submission systems, yet.

roadmap

Pool DMP templates and instances

DSW

Design questionnaires to train and collect information

DMP, future trends

- Reproducibility
  - Source code management and sharing
  - Reproducibility of computational analyses

RDAP Review

EDITOR'S SUMMARY

With reproducibility of research becoming a leading issue in academia, libraries are examining their role in promoting data and information transparency. The National Science Foundation's requirement for data management plans in research projects, grant applications stressing evidence of unbiased results and scholars' demands for standards for reproducibility together highlight the need for attention to the issue.

Is Research Reproducibility the New Data Management for Libraries?

by Cynthia R.H. Vitale

Research reproducibility has become a hot topic among academics in the last few years. With organizations such as Retraction Watch cataloging retractions of peer-reviewed literature, replication studies finding many research outcomes to not be reproducible [1, 2] and journals signing on to transparency polices [3, 4], requirement, libraries and library organizations were building socio-technical infrastructure for data management services, and more broadly, E-Science support, in the information science profession. Major professional organizations, such as the Association for Information Science and Technology (ASIS&T), the Association of

DMP, future trends

98 %

- Stronger sharing requirement
  - “shared scientific data should be made accessible as soon as possible, and no later than the time of an associated publication, or the end of performance period, whichever comes first.”

- Controlled-access for all human data
  - “access to scientific data derived from humans should be controlled, even if de-identified and lacking explicit limitations on subsequent use.”

NEWS: New NIH Policy on Data Management and Sharing

On October 29, 2020, NIH issued theÂ *NIH Policy for Data Management and Sharing* which will require NIH funded researchers to prospectively submit a plan outlining how scientific data from their research will be managed and shared. This policy will be effective January 25, 2023 and at that time will replace the 2003 NIH Data Sharing Plan. Learn more about the new policy.

Thank you!