Data Management Planning & Research Data Management

Lars Jermiin, Systems Biology Ireland, UCD
Data Stewardship Manager (Elixir - Ireland)

ELIXIR-CONVERGE has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 871075.

1

Program (University of Galway)

13:00 – 15:00	DMP & RDM workshop – Part I
15:00 – 15:30	Coffee break
15:30 – 17:00	DMP & RDM workshop – Part II

2

Overview

This workshop covers the following topics:

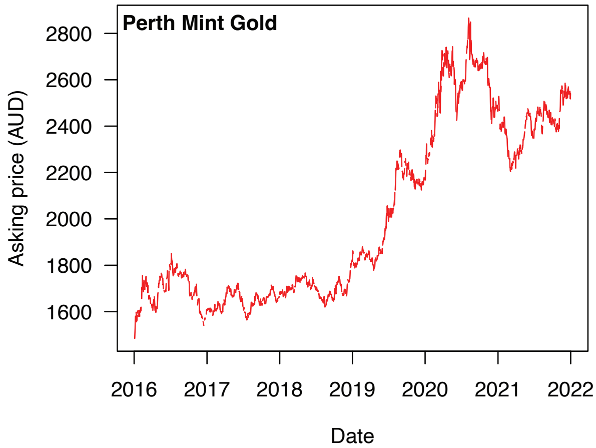
- The value of scientific data
- Loss of scientific data
- Open scientific data
- The FAIR data principles
- Data management planning (DMP)
- Research data management (RDM)

3

The value of scientific data

4

Gold

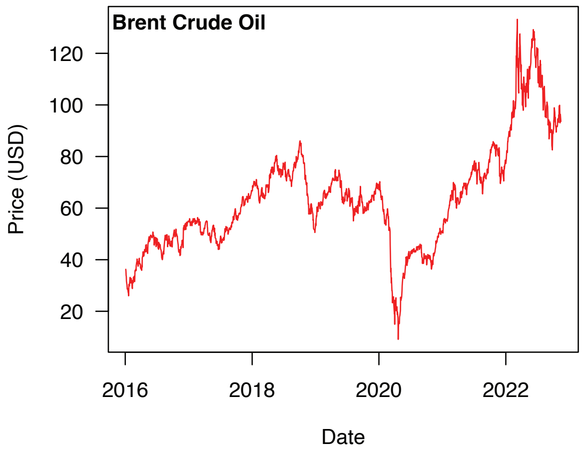


Note The price of gold fluctuates, partly in response to supply and demand

Source: Perth Mint (https://www.perthmint.com)

5

Crude oil



Note The price of Brent crude oil fluctuates, partly in response to supply and demand

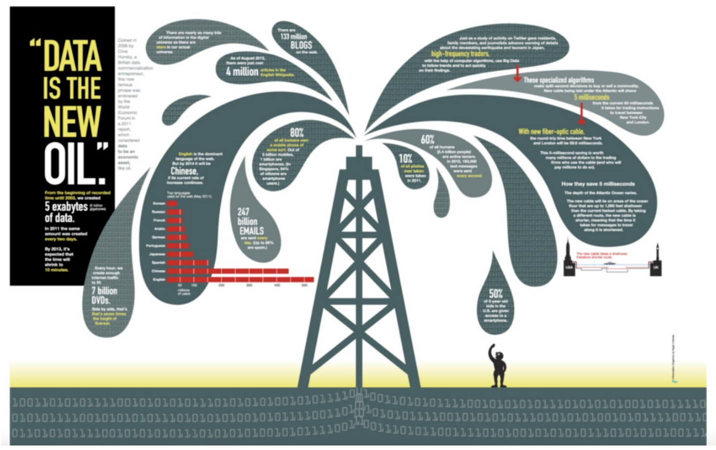
Source: British Petrol (https://www.bp.com/en_nz/new-zealand/home/products-and-services/bp-fuels/technical-information.html)

6

Question What is the value of scientific data?

7

Data is the new oil …



Note New businesses and professions have arisen aimed at mining data (e.g., Google, Facebook, etc)

8

Factors influencing the value of scientific data

- The cost of generating the data
- The provenance, nature, composition, and volume of the data
- The accuracy of the data
- The usefulness of the data (from the scientific and socio-economic perspectives)
- Supply and demand

Take-home message

Compared to gold and oil, it is much harder to determine **the value of scientific data** but, most sensibly, it **is set at the cost of replacing the original set of data**

9

The loss of scientific data

CONVERGE

10

Loss of scientific data

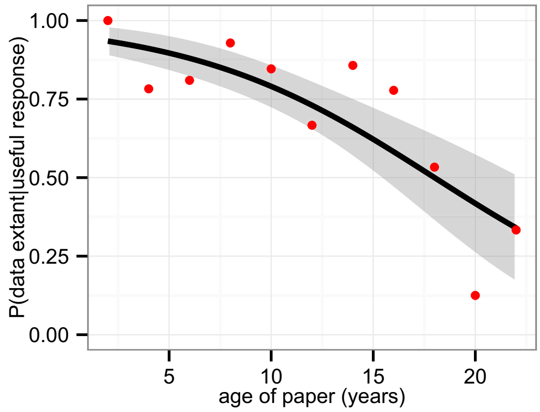
Question How fast does scientific data go missing?

Experiment Measure the availability of 2- to 22-year-old data from 516 studies of ecology

Result Availability of data is strongly affected by article age

Rate of loss 17%

Reason Obsolete e-mails and storage devices

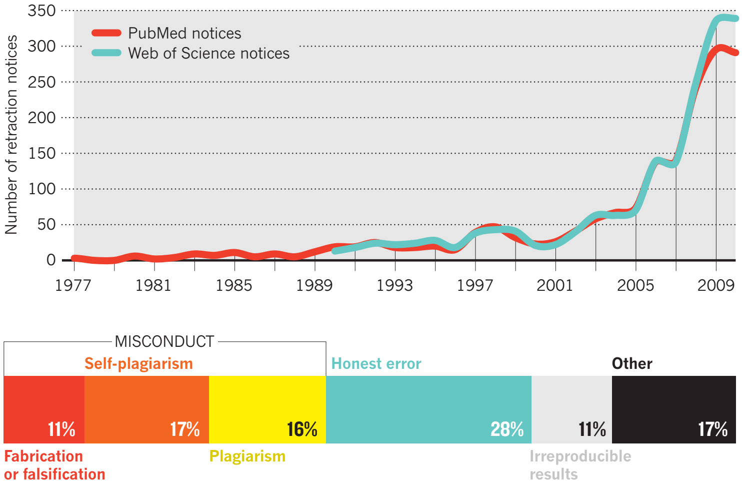


Note Policies mandating data archiving at publication are clearly needed

Source: Vines TH et al. (2014) **Current Biology** 24, 94-97

11

Other reasons for loss of data — retraction • 1



Statistics (period)

Rise in retractions	10-fold
Rise in publications	44%

Source: Van Noorden R (2011) **Nature** 478, 26-28

12

Other reasons for loss of data — retraction • 2

Misconduct accounts for the majority of retracted scientific publications

Ferric C. Fang[a,b,1], R. Grant Steen[c,1], and Arturo Casadevall[d,1,2]

"A detailed review of all **2,047 biomedical and life-science research articles indexed by PubMed as retracted** on May 3, 2012 revealed that

- 21.3% of retractions were **attributable to error**
- 67.4% of retractions were **attributable to misconduct**, including fraud or suspected fraud (43.4%), duplicate publication (14.2%), and plagiarism (9.8%)
- Incomplete, uninformative or misleading retraction announcements have led to a previous **underestimation of the role of fraud** in the ongoing retraction epidemic."

Source: Fang FC et al. (2012) **Proceedings of the National Academy of Science of the USA** 109, 17028-17033

13

Other reasons for loss of data — irreproducibility • 1

The Economics of Reproducibility in Preclinical Research

Leonard P. Freedman[1]*, Iain M. Cockburn[2], Timothy S. Simcoe[2,3]

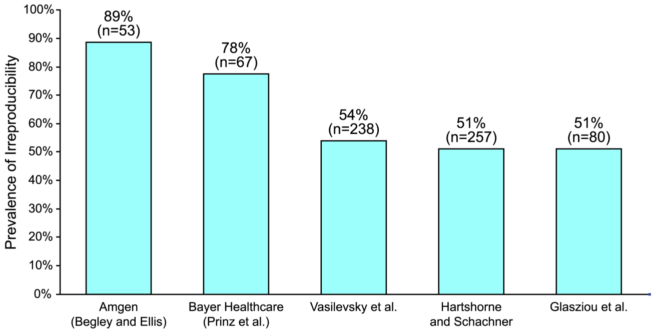


Fig 1. Studies reporting the prevalence of irreproducibility. Source: Begley and Ellis [6], Prinz et al. [7], Vasilevsky [8], Hartshorne and Schachner [5], and Glasziou et al. [9].

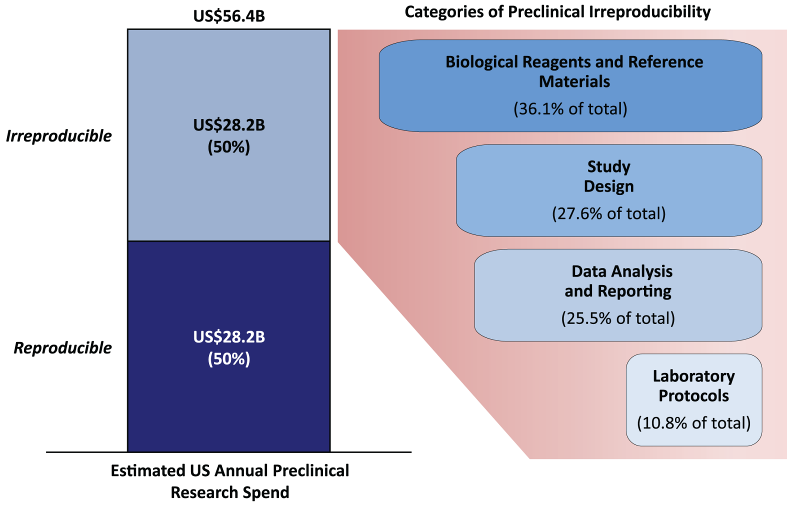
Primary courses of irreproducibility

1. Study design
2. Biological reagents and reference materials
3. Laboratory protocols
4. Data analysis and reporting

Source: Freedman LP et al. (2015) **PLoS Biology** 13, e1002165

14

Other reasons for loss of data — irreproducibility • 2

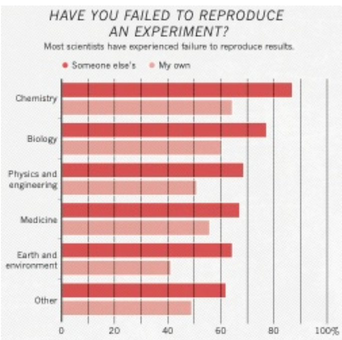


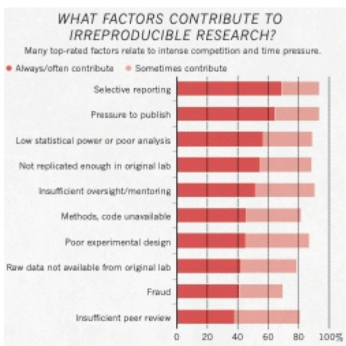
Source: Freedman LP et al. (2015) **PLoS Biology** 13, e1002165

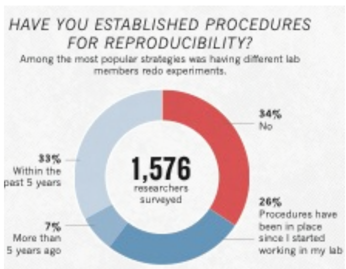
15

Survey — is there a reproducibility crisis?

Answers (1,576 respondents)	
Don’t know	7%
No	3%
Yes, a slight crisis	38%
Yes, a significant crisis	52%







Source: Baker M (2016) **Nature** 533, 452-454

16

Example 1 — possible scientific misconduct

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Letter to the Editor

Problems with data quality in the reconstruction of evolutionary relationships in the *Drosophila melanogaster* species group: Comments on Yang et al. (2012)

it is unclear how gaps were removed in the CDC6 gene. Moreover, we discovered additional problems when we analysed the verifiable data following the Yang et al. methodology. We recreated the Yang et al. study, using our best alignments, for all genes excepting ITS and H2S, using RaxML v1.3 (Stamatakis, 2006) and

Problems reported

1. "23 Genbank accession numbers presented in Table A1 of Yang et al. **are not for *Drosophila***. …"
2. "16 other Genbank accession numbers **are either listed twice within the above-mentioned table, or assigned to a different *Drosophila* species** in Genbank (searched 23 October 2013)."
3. "A further 21 sequences [cannot be aligned], **appearing to be either different genes or contamination**."

Note Yang et al. (2012) is still cited frequently, despite the problem reported in 2014

Source: Catullo RA & Oakeshott JG (2014) **Molecular Phylogenetics and Evolution** 78, 275-276

elixir CONVERGE

17

Example 2 — honest error announced in an Erratum

A Protein Domain-Based Interactome Network for *C. elegans* Early Embryogenesis

Mike Boxem,* Zoltan Maliga, Niels Klitgord, Na Li, Irma Lemmens, Miyeko Mana, Lorenzo de Lichtervelde, Joram D. Mul, Diederik van de Peut, Maxime Devos, Nicolas Simonis, Muhammed A. Yildirim, Murat Cokol, Huey-Ling Kao, Anne-Sophie de Smet, Haidong Wang, Anne-Lore Schlaitz, Tong Hao, Stuart Milstein, Changyu Fan, Mike Tipsword, Kevin Drew, Matilde Galli, Kahn Rhrissorrakrai, David Drechsel, Daphne Koller, Frederick P. Roth, Lilia M. Iakoucheva, A. Keith Dunker, Richard Bonneau, Kristin C. Gunsalus, David E. Hill, Fabio Piano, Jan Tavernier, Sander van den Heuvel, Anthony A. Hyman,* and Marc Vidal*
*Correspondence: m.boxem@uu.nl (M.B.), hyman@mpi-cbg.de (A.A.H.), marc_vidal@dfci.harvard.edu (M.V.)
http://dx.doi.org/10.1016/j.cell.2012.11.042

Problem reported

"Since publication of [Cell 134, 534-545; 2008], it has come to our attention that an error occurred in creating the initial data file cataloging the names and storage positions of the bait proteins tested in the yeast two-hybrid system. **The register for the bait names was shifted in an Excel file**, resulting in the assignment of an incorrect bait name to 37% of the published partners. …"

Source: Boxem M et al. (2012) **Cell** 151, 1633

elixir CONVERGE

18

Questions How many of you have asked for and obtained/downloaded scientific data?
How many of you have been able to analysed it?
How many of you failed to do so?
What were the reasons for this failure?
How many of you anticipate you will do so in the future?

19

Open scientific data

20

What is open data?

"Open Data (OD) is **an emerging term** in the process of **defining how scientific data may be published and re-used without price or permission barriers**."

Problem

"Scientists generally see published data as belonging to the scientific community, but many publishers claim copyright over data and will not allow its re-use without [prior] permission."

Implication "[A] major impediment to the progress of scholarship in the digital age."

Source: Murray-Rust P (2008) **Serials Review** 34, 52-64

21

OD — challenges & opportunities • 1

Challenges and Opportunities of Open Data in Ecology

O. J. Reichman,* Matthew B. Jones, Mark P. Schildhauer

"Ecology is a synthetic discipline benefiting from open access to data …. **Technological challenges exist**, however, due to the dispersed and heterogeneous nature of these data."

"**Standardization of methods** and **development of robust metadata** can increase data access but are not sufficient."

"**Reproducibility of analyses is also important**, and **executable workflows are addressing this issue by capturing data provenance**."

"**Sociological challenges, including inadequate rewards for sharing data, must also be resolved**."

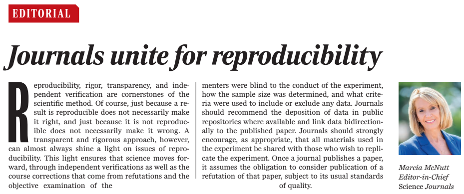
"**The establishment of well-curated, federated data repositories will provide a means to preserve data while promoting attribution and acknowledgement of its use**."

Source: Reichmann OJ et al. (2011) **Science** 331, 703-705

22

OD — challenges & opportunities • 2

- In June 2014, the **Nat. Inst. Health (US), *Science*, and *Nature* convened a meeting** of editors of +30 major journals, representatives from major funding agencies, and scientific leaders **to discuss principles and guidelines for preclinical biomedical research**
- The delegates **agreed on a common set of principles and guidelines in reporting preclinical research** that list **journal policies** and **author reporting requirements** to promote transparency and reproducibility (see URL)



EDITORIAL

Journals unite for reproducibility

Reproducibility, rigor, transparency, and independent verification are cornerstones of the scientific method. Of course, just because a result is reproducible does not necessarily make it right, and just because it is not reproducible does not necessarily make it wrong. A transparent and rigorous approach, however, can almost always shine a light on issues of reproducibility. This light ensures that science moves forward, through independent verifications as well as the course corrections that come from refutations and the objective examination of the menters were blind to the conduct of the experiment, how the sample size was determined, and what criteria were used to include or exclude any data. Journals should recommend the deposition of data in public repositories where available and link data bidirectionally to the published paper. Journals should strongly encourage, as appropriate, that all materials used in the experiment be shared with those who wish to replicate the experiment. Once a journal publishes a paper, it assumes the obligation to consider publication of a refutation of that paper, subject to its usual standards of quality.

Marcia McNutt
Editor-in-Chief
Science *Journals*

Science 346, 679 [2014]

THIS WEEK

EDITORIALS

CONSERVATION Saving species is far from a walk in the park p.8

WORLD VIEW Psychology gears up to check its workings p.9

BREAKFAST Chimps plan days to ensure they nab tastiest figs p.11

Journals unite for reproducibility

Consensus on reporting principles aims to improve quality control in biomedical research and encourage public trust in science.

Nature 515, 7 [2014]

Source: https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research

23

Center for Open Science (COS) — standards & guidelines

Standards	Not implemented	Level I	Level II	Level III
Data citation	No mention of data citation	Journal describes citation of data in guidelines to authors with clear rules and examples	…	…
Data transparency	Journal encourages data sharing, or says nothing	Article states whether data are available, and, if so, where to access them	…	…
Analysis code transparency	Journal encourages code sharing, or says nothing	Article states whether code is available, and, if so, where to access it	…	…
Materials transparency	Journal encourages materials sharing, or says nothing	Article states whether materials are available, and, if so, where to access them	…	…
Design & analysis reporting guidelines	Journal encourages design and analysis transparency, or says nothing	Journal articulates design transparency standards	…	…
Study preregistration	Journal says nothing	Article states whether preregistration of study exists, and, if so, where to access it	…	…
Preregistration of analysis plans	Journal says nothing	Article states whether preregistration of study exists, and, if so, where to access it	…	…
Replication	Journal discourages submission of replication studies, or says nothing	Journal encourages submission of replication studies	…	…

Note 2,007 journals, 25 publishers, and 92 societies have signed up to the Transparency and Openness Promotion (TOP) Guidelines at COS (http://cos.io/top)

Sources: https://topfactor.org/summary; Nosek BA et al. (2015) **Science** 348, 1422-1425

24

Example 3 — Public Library of Science (PLoS) and TOP • 1

- **PLoS has signed up to the TOP guidelines** (URL)
- PLoS journals require authors to **make all data necessary to replicate their study's findings publicly available** without restriction at the time of publication
- When specific legal or ethical restrictions prohibit public sharing of a data set, **authors must indicate how others may obtain access to the data**
- When submitting a manuscript, authors must include **a Data Availability Statement describing compliance with PLoS' data policy**
- If a manuscript is accepted for publication, **the Data Availability Statement will be published as part of the article**

Source: https://plos.org/open-science/open-data/

25

Example 3 — Public Library of Science (PLoS) and TOP • 2

- **Authors must share the "minimal data set" for their submission**. PLoS defines the minimal data set as the data required to replicate all findings reported in the article, as well as related metadata and methods
- PLoS also **requires authors to comply with field-specific standards for preparation, recording, and deposition of data**, when applicable
- For example, authors should submit the following data:
  - The **values behind the means, standard deviations and other measures reported**
  - The **values used to build graphs**
  - The **points extracted from images for analysis**
- PLoS **does not permit references to "data not shown"**

26

Example 3 — Public Library of Science (PLoS) and TOP • 3

- For studies involving human research participant data or other sensitive data, **PLoS encourages authors to share de-identified or anonymized data**
- When data cannot be publicly shared, **PLoS allows authors to make their data sets available upon request**
- PLoS will not consider manuscripts for which the following factors influence authors' ability to share data:
  - **Authors will not share data because of personal interests**, such as patents or potential future publications
  - **The conclusions depend solely on the analysis of proprietary data**. PLoS considers proprietary data to be data owned by individuals, organizations, funders, institutions, commercial interests, or other parties that the data owners will not share

27

EU announces that all scientific papers should be free by 2020

TECH 30 May 2016 By JOLENE CREIGHTON, FUTURISM

Sergei25/Shutterstock.com

"This week was a revolutionary week in the sciences … because some of the most prominent world leaders announced an initiative which asserts that **European scientific papers should be made freely available to all by 2020**."

Source: https://www.sciencealert.com/europe-announces-that-all-scientific-articles-should-be-freely-accessible-by-2020

28

Summary of the benefits of OD

Direct
- Preserve access to data
- Discover data
- Allow reuse or repurpose data
- Verify published research

Short-term
- Availability for review
- Availability for validation

To the author
- Protection against data entropy
- Improved methodologies
- Higher diffusion and visibility
- Higher citation rate of their publications
- Fulfillment of funding mandate

Indirect
- Redundant data collection
- Inefficient legacy data curation
- Burden of sharing-upon-request
- Studies cannot be completed

Long-term
- Persistent link with article data
- Increased impact per publication

To the scientific community and public
- More efficient use of research fundings
- Foster collaboration
- Accelerate innovation
- Educational opportunities
- Public trust in science

29

Questions How many of you have reviewed a manuscript submitted for publication?
Did you examine the associated data?
If you did not do so, then why not?
Did you survey the associated codes?
If you did not do so, then why not?

30

FAIR data principle

31

FAIR data principles • 1

- Designed by stakeholders from **academia**, **industry**, **funding agencies**, and **scholarly publishers**
- Put emphasis on enhancing **the ability of computational agents to automatically find and use** (digitalised) **data** and to **support its reuse** by individuals
- Rest on four foundational principles:
  - Findability
  - Accessibility
  - Interoperability
  - Reusability
- **Apply to digitalised research data objects** (e.g., data, codes, protocols, workflows) needed to ensure **transparency**, **reproducibility, and reusability** in research

Source: Wilkinson MD et al. (2016) **Scientific Data** 3, 160016

32

FAIR data principles • 2

- The main **barrier to expedient discovery and reuse of digitalise research objects is**
  - Not the lack of appropriate technology, but
  - The **lack of careful attention paid to digital data objects during their creation and storage**
- To overcome this barrier, we need to **render all digital research objects findable in special-purpose and general-purpose repositories using the metadata** assigned to each object
- The FAIR principles **apply to both human-driven and agent-driven activities**

33

FAIR data principles • 2

The challenge …

Depending on the amount and detail of information provided with a digital object, **the computational agent should be able to**:

1. **Identify the type of object** (with respect to both structure and intent)
2. **Determine if the object is useful** within the context of the agent's current task by interrogating metadata and/or data elements
3. **Determine if the object is usable**, with respect to its license, its consent, or other accessibility or use constraints
4. **Take the appropriate action**

(In much the same way that a human would)

34

FAIR data principles • 4

To be findable:

F1 Data and metadata are assigned a globally unique and persistent identifier

F2 Data are described with rich metadata (defined by R1)

F3 Metadata clearly and explicitly include the identifier of the data it describes

F4. Data and metadata are registered or indexed in a searchable resource

To be accessible:

A1 Data and metadata are retrievable by their identifier using a standardized communications 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 Data and metadata are accessible, even when the data are no longer available

To be interoperable:

I1 Data and metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation

I2 Data and metadata use vocabularies that follow FAIR principles

I3 Data and metadata include qualified references to other data and metadata

To be reusable:

R1 Data and metadata are richly described with a plurality of accurate and relevant attributes

R1.1 Data and metadata are released with a clear and accessible data usage license

R1.2 Data and metadata are associated with detailed provenance

R1.3 Data and metadata meet domain-relevant community standards

Source: Wilkinson MD et al. (2016) **Scientific Data** 3, 160016

35

Summary on FAIR data principles

Compliance with the FAIR data principles:

- Is **achieved by diligent annotation of digital data objects**, using metadata
- Is **achieved through consistent use of standard file formats**, with recognizable file name extensions (e.g., .csv, .tiff, .fst), which can be read and processed using open-source software
- Is **achieved through diligent record keeping** (keep a logbook)
- Is increasingly often a **requirement to operate in modern scientific environments** — so embrace it…

36

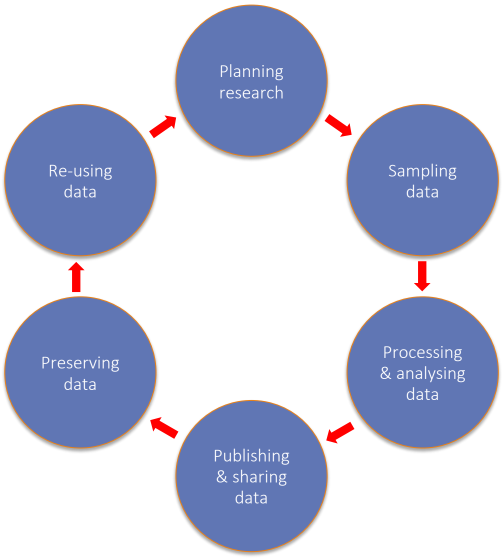
Questions How many of you use a lab/log book (digital or hard copy)?
How many years do you think you will need to keep the lab/log book?
Where will you store your lab/log book when you have finished a project?

37

Data management planning (DMP)

38

Scientific data have a live cycle

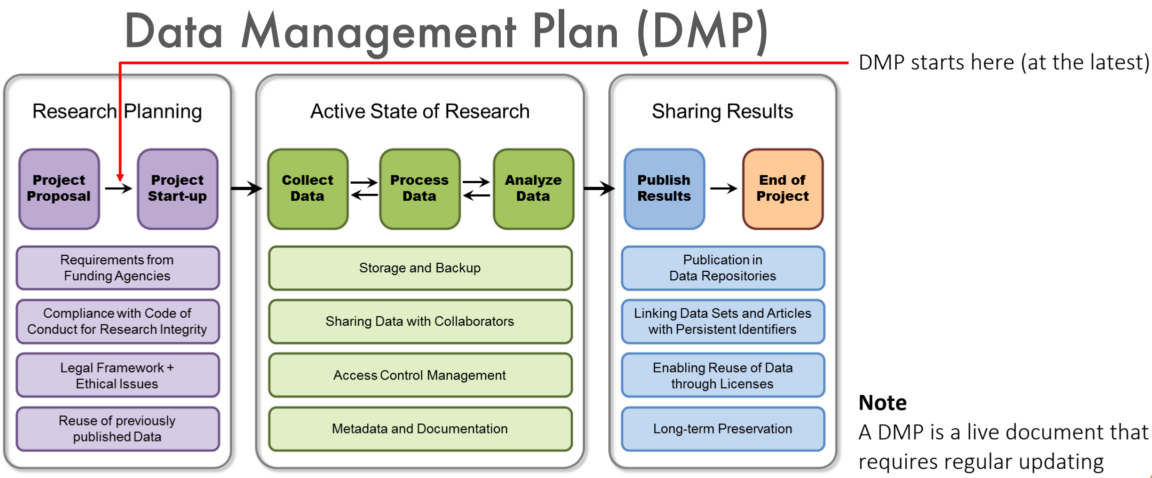


Notes

- At the sampling stage, we want to sample enough data, but no more than that
- At the following 4 stages, some data may become lost; we want to minimise that
- **Careful data management planning and research data management is essential**

39

Data management planning — in a nutshell…



Note
A DMP is a live document that requires regular updating

40

Data management plans • 1

Required by a growing number of funding agencies, including the

- European Research Research Council (ERC)
- Swiss National Research Foundation (SNRF)
- Health Research Board (HRB), Ireland
- Science Foundation Ireland (SFI)
- Austrian Science Fund (FWF)
- etc

European Research Council
Established by the European Commission

Swiss National
Science Foundation

Health Research Board

Science Foundation Ireland For what's next

Der Wissenschaftsfonds.

41

Data management plans • 2

- Many research funding agencies **use similar policies** (e.g., GDPR, code of conduct, Open Access to publications & data) **and elaborate guidelines** on how to write a DMP
- Some funding agencies cover the costs of enabling open access to publications and data (e.g., SNSF)
- **Preparation of DMPs is facilitated using web-based services**, such as **DMPOnline** and ELIXIR-Converge’s **RDMKit**
- Many research organisations have **local facilities and help in preparing DMPs** (e.g., University College Dublin)

42

Data management plans • 3 — resources & repositories

Electronic **lab notebooks**

- **SciNote** — free (https://www.scinote.net/)
- **Benchling Notebook** — free (https://www.benchling.com/benchling-eln)
- **Open Science Framework** — free (https://osf.io/)
- **BBEdit** — powerful & free text editor (https://www.barebones.com/products/bbedit/)

Repositories with version control for source code and documents

- **GitHub** — free public and private repositories (https://github.com/)
- **GitLab** — free public and private repositories (https://about.gitlab.com/)
- **GitKraken** — A GUI client for using GIT version control without the use of command line (https://www.gitkraken.com/)

43

Data management plans • 4 — data repositories

Data repositories

- **Figshare** — Free digital repository, max 5Gb, free to access (https://figshare.com/)
- **Zenodo** — Free digital repository, max 50Gb, free to access (https://zenodo.org/)
- **DRYAD** — Curated digital repository, max 20Gb, free to access (https://datadryad.org/stash)
- **Nature** — Guide on data repository (https://www.nature.com/sdata/policies/repositories)
- **PLoS** — Policy on data availability (https://journals.plos.org/plosone/s/data-availability)
- **Dataverse** — review (https://dataverse.org/blog/comparative-review-various-data-repositories)
- **re3data** — Registry of Research Data Repositories (https://www.re3data.org/)
- **FAIRsharing** — A curated resource on data and metadata standards, etc (https://fairsharing.org)
- **Open Access Directory** — Data repositories, partitioned by discipline (Archaeology, Astronomy, Biology, ..., Social Sciences (https://oad.simmons.edu/oadwiki/Data_repositories)

44

Data management plans • 5 — file formats, metadata, etc

File formats for long-term preservation

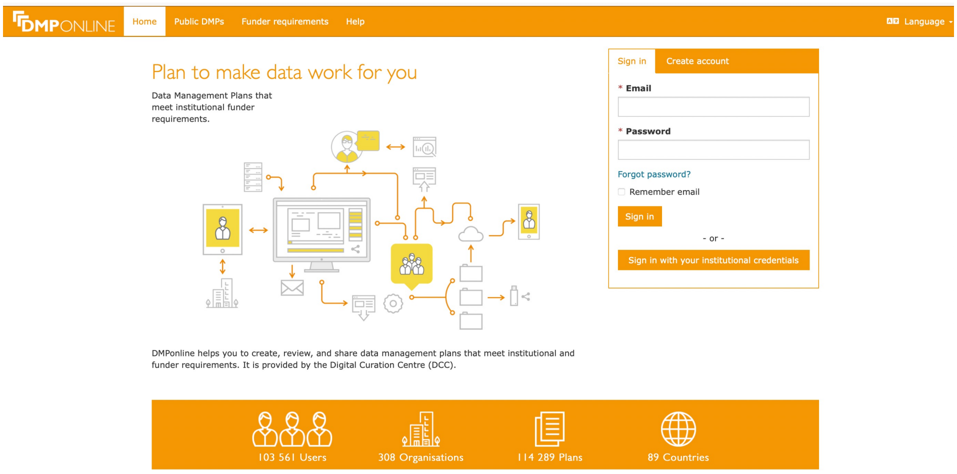
- **DRYAD** — recommended formats and guidelines (https://datadryad.org/stash/terms#formats)
- **BiUM** — useful info (https://www.bium.ch/en/publication-open-access/data-management/#5)

Metadata and README files, allowing data to be understood and reused

- **FAIRsharing** — A resource on field-specific metadata and format standards (https://fairsharing.org)
- **Digital Curation Center** — A resource on field-specific metadata and format standards (https://www.dcc.ac.uk/guidance/standards/metadata)
- **DataCite metadata schema** — Standard for describing general research data. Useful before data is stored (https://schema.datacite.org/meta/kernel-4.1/doc/DataCite-MetadataKernel_v4.1.pdf)
- **DataCite metadata generator** — tool used to generate a Readme XML file describing your datasets (https://github.com/mpaluch/datacite-metadata-generator)

45

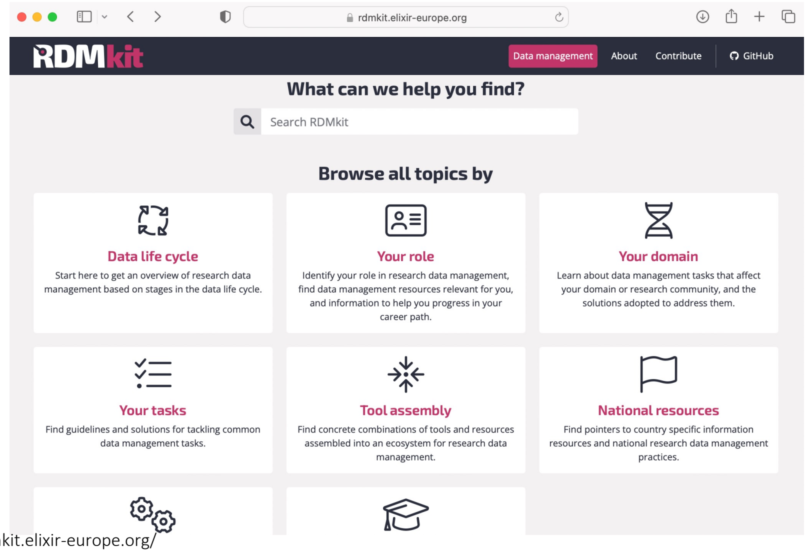
Data management plans • 6 — DMPOnline



Source: https://dmponline.dcc.ac.uk

46

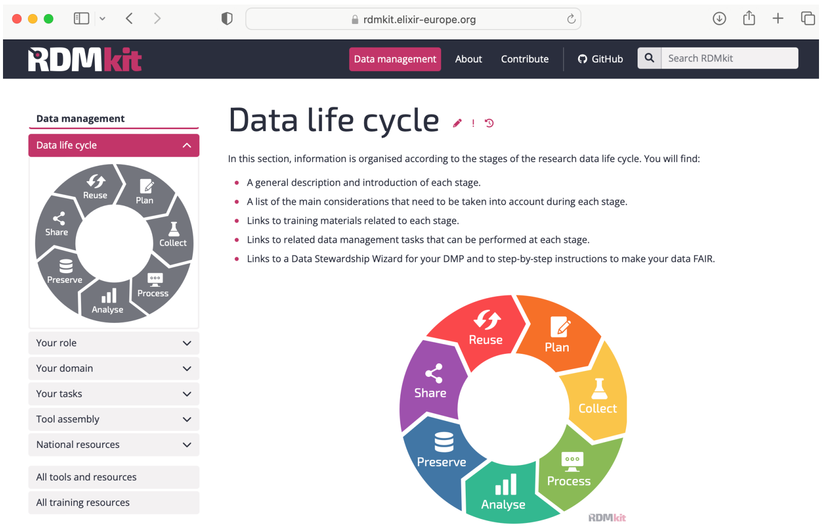
Data management plans • 7 — Elixir Converge RDMKit



Source: https://rdmkit.elixir-europe.org/

47

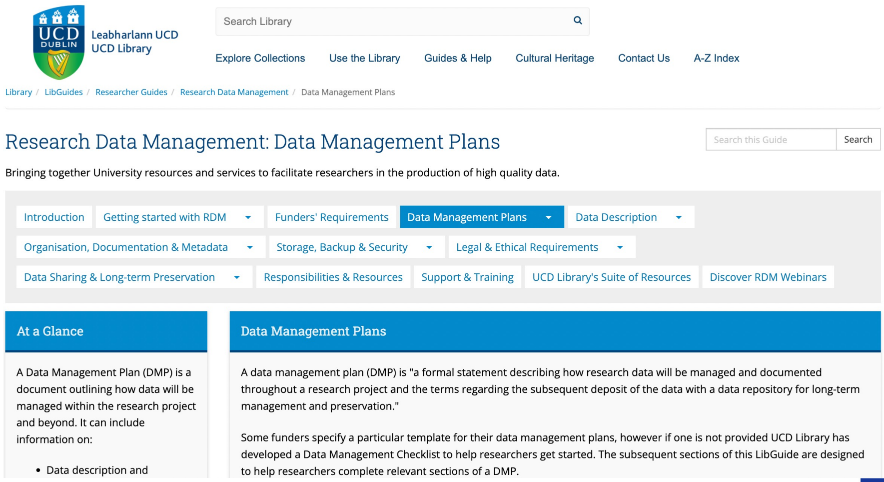
Data management plans • 8 — Elixir Converge RDMKit



Source: https://rdmkit.elixir-europe.org/data_life_cycle

48

Data management plans • 9 — RDM & DMP at UCD



Source: https://libguides.ucd.ie/data/dmp

49

Data management plans • 10 — typical questions

1. **Data collection and documentation**
   a. What data will you collect, observe, generate or reuse?
   b. How will the data be collected, observed, or generated?
   c. What documentation and metadata will you provide with the data?
2. **Ethics, legal and security issues**
   a. How will ethical issues be addressed and handled?
   b. How will data access and security be managed?
   c. How will you handle copyright and Intellectual Property Rights issues?
3. **Data storage and preservation**
   a. How will your data be stored and backed-up during the research?
   b. What is your data preservation plan?
4. **Data sharing and reuse**
   a. How and where will the data be shared?
   b. Are there any necessary limitations to protect sensitive data?
   c. Do the intended digital repositories conform to the FAIR Data Principles?
   d. What organisations will be maintaining the intended data repositories (profit/non-profit)?

50

Data management plans • 11 — things to consider

Types of data used in research project
- Materials / samples
- Protocols
- Codes/scripts/programs
- Raw data
- Processed data
- Results
- Notes/notebooks

Types and sizes of files used in project
- Cell microscopy images (.tiff, .jpg)
- Sequencing data (FASTQ, fasta, .fst)
- Figures and graphs (.pdf, .svg)
- Spreadsheets (.csv)
- Scripts (.sh, .r, .py)
- MS data (mzXML, PKL*)
- Interview videos (MP4)
- Protocols and instructions (.txt)
- Texts accompanying videos (.pdf)

51

Data management plans • 12 — assembling basic information

Types	Equipment	Software	Data storage format	Data archiving / sharing format	Volume

52

Data management plans — example 1 (cell. & microb. project)

Types	Equipment	Software	Data storage format	Data archiving / sharing format	Volume
Microscopy images					
Raw data: microscopy cell images	Zeiss LSM 710 Quasar	ZEN lite software	.liff	.tiff uncompressed, JPEG2000	500 GB
Secondary data: 3D Z-stack reconstructions and processed images		Imaris 7.2.1 software; Fiji/ImageJ; Adobe Photoshop CS5	.ims, .tif series, .PSD	.tiff uncompressed, JPEG2000	1 TB
Analysed data: cell quantifications		Imaris 7.2.1 software, Excel	.ims, .xlsx	.xlsx; .csv	3 GB
Raw data :time lapse video microscopy	Leica SP5	LAS AF Lite 4.0.11706	.czi files;.avi,.mov	MPEG-4; Motion JPEG 2000	500 GB
Analysed data: tracking function		Metamorph software 6.0	.xlsx	.xlsx; .csv	2 GB
Western Blots					
Raw data: cell images					1 GB
Analysed data: quantification					500 MB
					TOTAL =

Data management plans — example 2 (comp. biol. project)

Types	*Equipment*	*Software*	*Data storage format*	*Data archiving / sharing format*	*Volume*
Sequence data					
Nucleotide sequences (simulated)	Computer	INDELible V1.03 (Mol. Biol. Evol. 26, 1879-1888) Hetero2 v2.4 (Syst. Biol. 63:726-742) R v4.2.0 (www.R-project.org) SatuRation v1.0 (www.github.com/lsjermiin/SatuRation.v1.0/) SatuRationHeatMapper v1.0 (www.github.com/ZFMK/SatuRationHeatMapper/) RedundancyHeatMapper v1.0 (www.github.com/ZFMK/RedundancyHeatMapper/) FigTree v1.4.4 (tree.bio.ed.ac.uk/software/figtree/)	Text, PDF & script files: .fas (text) .fst (text) .txt (text) .csv (text) .sh (scripts) .R (scripts) .pdf (figures)	Text, PDF & scripts .fas (text) .fst (text) .txt (text) .csv (text) .sh (scripts) .R (scripts) .pdf (figures)	150 Mb
Nucleotide sequences (simulated)	Computer	IQ-TREE2 v2.1.2 (Mol. Biol. Evol. 37, 1530-1534)	Text files .iqtree (text) .bionj (text) .contree (text) .log (text) .mldist (text) .treefile (text) .nwk (text) .nex (text)	Text files .iqtree (text) .bionj (text) .contree (text) .log (text) .mldist (text) .treefile (text) .nwk (text) .nex (text)	600 Mb

Questions How many of you have written a DMP?
What was the worst part of it?
What was the best part of it?

55

Research Data Management (RDM)

56

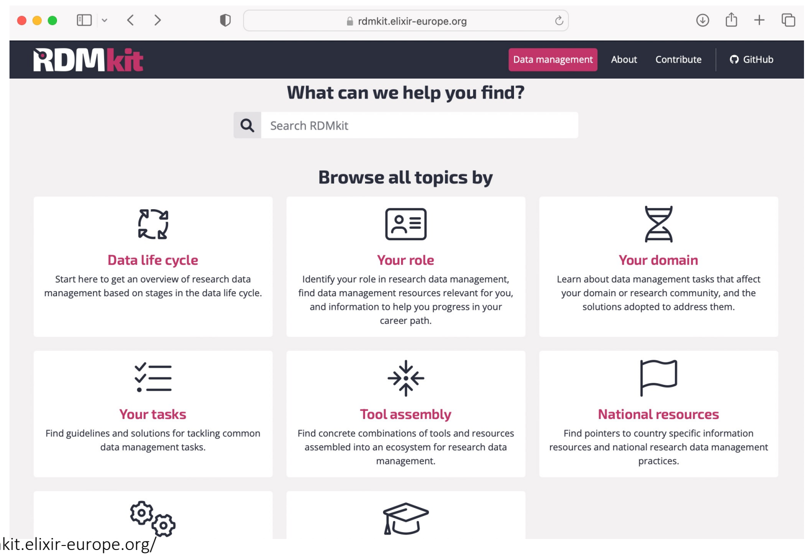
Research data management • 1

- Research in modern scientific communities often entails **collecting and analysing huge amounts of highly heterogeneous data** (qualitative and quantitative data in various forms)
- Growing demands for **transparency**, **reproducibility**, and **accountability** — coupled with the **FAIR Data Principles** — have led to **significant changes in how research is done**, on a small scale as well as on an industrial scale
- **RDM is now an integral part of R&D in modern societies**
- In Europe, **ELIXIR is a key innovator**, **enabler**, and **partner**

Source: https://elixir-europe.org

57

Research data management • 2



Source: https://rdmkit.elixir-europe.org/

58

Research data management • 3

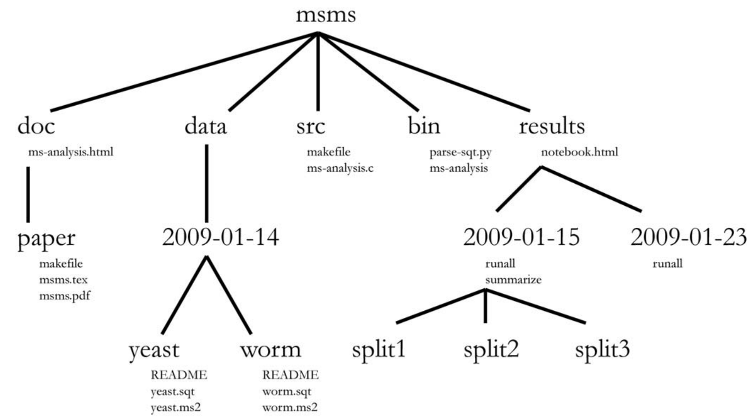
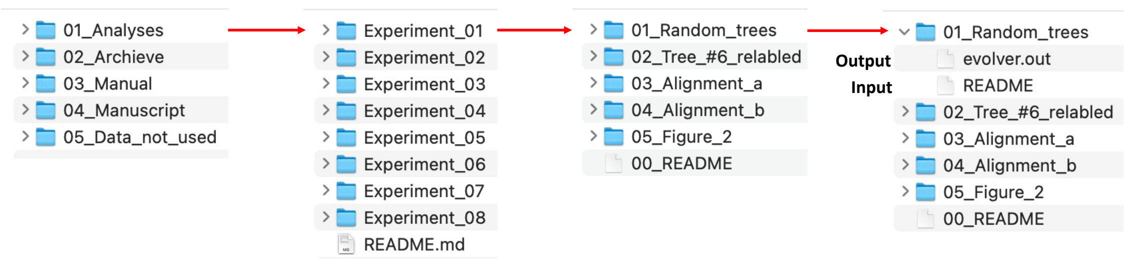


Figure 1. Directory structure for a sample project. Directory names are in large typeface, and filenames are in smaller typeface. Only a subset of the files are shown here. Note that the dates are formatted `<year>-<month>-<day>` so that they can be sorted in chronological order. The source code `src/ms-analysis.c` is compiled to create `bin/ms-analysis` and is documented in `doc/ms-analysis.html`. The `README` files in the data directories specify who downloaded the data files from what URL on what date. The driver script `results/2009-01-15/runall` automatically generates the three subdirectories `split1`, `split2`, and `split3`, corresponding to three cross-validation splits. The `bin/parse-sqt.py` script is called by both of the `runall` driver scripts.

Source: Noble WS (2009) **PLoS Computational Biology** 5, e1000424

59

Example 4 • RDM & directory structure



Notes

- The README file contains information about the software as well as the input instructions used
- The evolver.out file contains the output (i.e., results from running evolver with the instructions)

Source: Jermiin LS et al. (2023) **Systematic Biology** (in review)

60

Example 4 • RDM & record keeping

The README file

```
# This file includes the input and output from evolver, an interactive CMD-line program.
#
# Software    evolver (from the PAML program package)
# Version     4.8a
# Source      http://abacus.gene.ucl.ac.uk/software/paml.html
# Reference   Yang, Z. 2007. PAML 4: a program package for phylogenetic analysis by
#             maximum likelihood. Molecular Biology and Evolution 24: 1586-1591
#
# Text below the line ('---') was copied from the terminal. The first line is the command
# that causes the program to execute. The operator then gave the answer to 6 questions:
#   Q1 = 2
#   Q2 = 20
#   Q3 = 10 57
#   Q4 = 1
#   Q5 = 0.6 0.1 0.5 0.5
#   Q6 = 0
# The output is included in the associated file called evolver.out.
--
evolver
EVOLVER in paml version 4.8a, August 2014
Results for options 1-4 & 8 go into evolver.out

    (1) Get random UNROOTED trees?
    (2) Get random ROOTED trees?
    (3) List all UNROOTED trees?
    (4) List all ROOTED trees?
    (5) Simulate nucleotide data sets (use MCbase.dat)?
    (6) Simulate codon data sets      (use MCcodon.dat)?
    (7) Simulate amino acid data sets (use MCaa.dat)?
    (8) Calculate identical bi-partitions between trees?
    (9) Calculate clade support values (evolver 9 treefile mastertreefile <pick1tree>)?
    (11) Label clades?
    (0) Quit?
2
No. of species: 20

number of trees & random number seed? 10 57
Want branch lengths from the birth-death process (0/1)? 1

birth rate, death rate, sampling fraction, and mutation rate (tree height)?
0.6 0.1 0.5 0.5
```

Allows you to find and download correct version of the software used

Lists the answers given to the questions posted by the software

Start the program

Question 1

Answer 1

Etc …

61

Example 4 • RDM & directory structure

- 01_Analyses
- 02_Archieve
- 03_Manual
- 04_Manuscript
- 05_Data_not_used

01_Analyses →

- Experiment_01
- Experiment_02
- Experiment_03
- Experiment_04
- Experiment_05
- Experiment_06
- Experiment_07
- Experiment_08
- README.md

Experiment_01 →

- 01_Random_trees
- 02_Tree_#6_relabled
- 03_Alignment_a
- 04_Alignment_b
- 05_Figure_2
- 00_README

03_Alignment_a →

- 01_Random_trees
- 02_Tree_#6_relabled
- 03_Alignment_a
  - 01_INDELible
  - 02_IQ-TREE2
    - Pasta1_TRUE.fas
    - Pasta1_TRUE.fas.bionj
    - Pasta1_TRUE.fas.ckp.gz
    - Pasta1_TRUE.fas.contree
    - Pasta1_TRUE.fas.iqtree
    - Pasta1_TRUE.fas.log
    - Pasta1_TRUE.fas.mldist
    - Pasta1_TRUE.fas.splits.nex
    - Pasta1_TRUE.fas.treefile
    - Pasta1_TRUE.fas.treefile.nwk
    - Runner_IQTree2.sh
- 04_Alignment_b
- 05_Figure_2
- 00_README

Results (Pasta1_TRUE.fas … Pasta1_TRUE.fas.treefile.nwk)

Notes

- Use shell scripts (see below) to control as many analytical process as possible
- Ensure transparency about the software used (name, version, source, reference)

```
#!/bin/bash
# Software   IQTREE 2
# Version    2.1.2
# Source     http://www.iqtree.org
# Reference  Minh BQ et al. (2020) IQ-TREE 2: New models and efficient methods for
#            phylogenetic inference in the genomic era. Mol. Biol. Evol. 37, 1530-1534

iqtree2 -s Pasta1_TRUE.fas --seqtype DNA -m JC+I --ufboot 10000
```

← Runner_IQTree2.sh

Source: Jermiin LS et al. (2023) **Systematic Biology** (in review)

62

Example 4 • RDM & directory structure

```
> 01_Analyses          > Experiment_01       > Experiment_01
> 02_Archieve          > Experiment_02       v Experiment_02
> 03_Manual            > Experiment_03           > Table_1
> 04_Manuscript        > Experiment_04             README
> 05_Data_not_used     > Experiment_05       > Experiment_03
                       > Experiment_06       > Experiment_04
                       > Experiment_07       > Experiment_05
                       > Experiment_08       > Experiment_06
                         README.md           > Experiment_07
                                             > Experiment_08
                                               README.md
```

README is a log, describing how the results in Table 1 were obtained

Source: Jermiin LS et al. (2023) **Systematic Biology** (in review)

63

Example 4 • RDM & record keeping

The README file

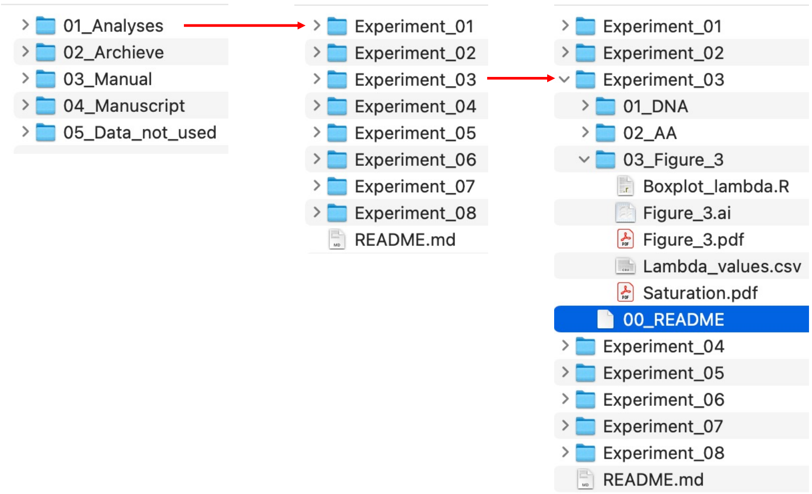
```
INFORMATION PERTAINING TO TABLE 1 OF THE MANUSCRIPT

Here "the manuscript" refers to

Jermiin LS, Meusemann K, Misof B, Shields DC. 2022. Quantifying the strength of the
historical signal in multiple sequence alignments of phylogenetic data. Systematic
Biology (in review)

-----------------------------------------------------------------------------------------

Objective: Determine how different metrics change as a function of the input data.

STEP 1
Generated nine directories:

./N1
./N2
./N3
./N4
./N5
./N6
./N7
./N8
./N9

one for each divergence matrix in the manuscript.

STEP 2
Within each of these directories we placed five similar files. For example, in ./N1 we
placed:

Runner_1-Make_data.sh
Parameter_File_list.txt
Parameters_info.txt
Sites_info.txt
Tree_0.1.nwk

Runner_1-Make_data.sh is a shell script, which calls Hetero2 v2.4 (Jayaswal et al. 2014;
Syst. Biol. 63, 726-742), a program that reads the parameters from the other four files,
and then simulate evolution of a nucleotide sequence on a 2-tipped tree, which is stored
in Tree_0.1.nwk.

The numbers in the name of the .nwk file represent the edge lengths in the 2-tipped tree.

STEP 3
Within each directory we ran the shell script and obtained an alignment with nucleotides.
For example, in ./N1 we obtained Tree_0.1.fst.

STEP 4.
Using SeaView v5.0.4 (Gouy et al. 2010; Mol. Biol. Evol. 27, 221-224), we obtained the
divergence matrix from each .fst file in the nine directories. These divergence matrices
were transferred to spreadsheet

./Table_1/Matrices.xlsx

Estimates of d_obs, d_ran, b_1, lambda, and  d_obs/b_1 were obtained using equations
embedded in Matrices.xlsx. Relevant numbers from Matrices.xlsx were transferred to the
manuscript, including Table 1.

END
```

Note Each STEP describes one of the actions taken, leading to the results in ./Table_1/Matrices.xlsx

Source: Jermiin LS et al. (2023) **Systematic Biology** (in review)

64

Example 4 • RDM & directory structure



Directories with scripts and input files needed to do simulations

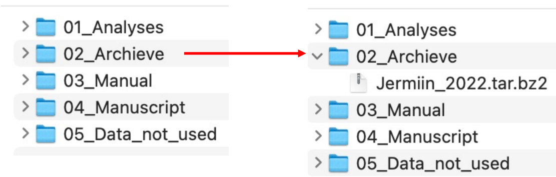
The R script reads the results in the .csv file and produces four graphs in the file called Saturation.pdf.

The graphs are transferred to an Adobe Illustrator file (i.e., Figure_3.ai) for final editing.

Source: Jermiin LS et al. (2023) **Systematic Biology** (in review)

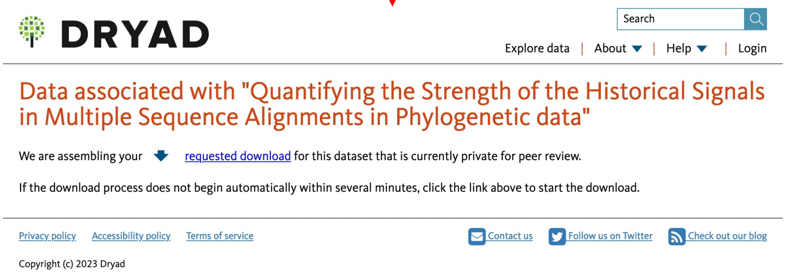
65

Example 4 • RDM & repository



Archive containing all material stored in 01_Analyses

Upload archive to repository



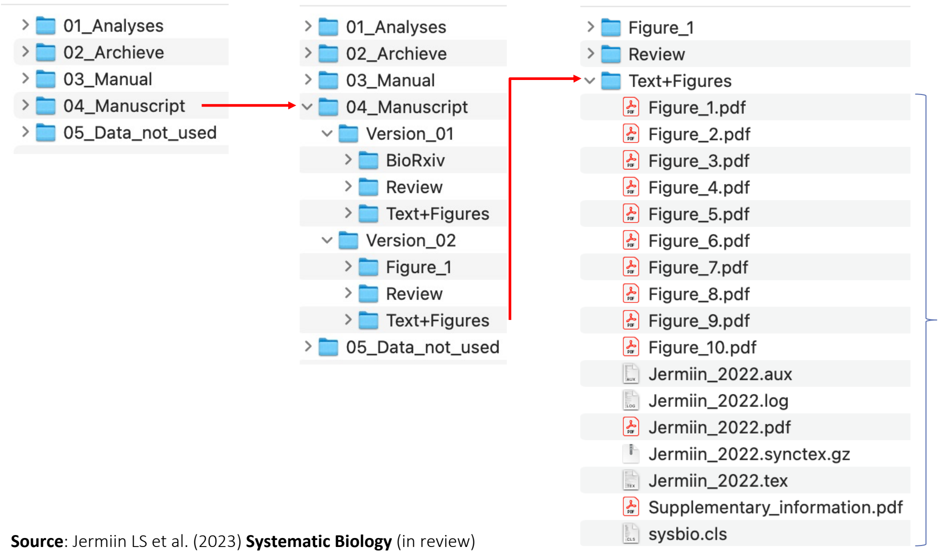
Notes

- In this case, we used DRYAD (we could have used Zenodo)
- **Provide relevant metadata**, outlining what is included in the archive
- **The repository returns a unique DOI** (digital object identifier)
- **Include the DOI in the paper** (creates a permanent link between the archive and the paper)

Source: Jermiin LS et al. (2023) **Systematic Biology** (in review)

66

Example 4 • RDM & directory structure

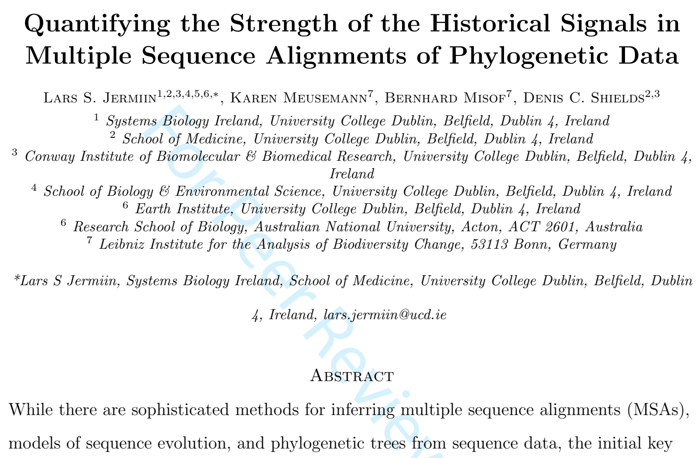


Manuscript, version 2

Source: Jermiin LS et al. (2023) **Systematic Biology** (in review)

67

Example 4 • RDM & manuscript



Quantifying the Strength of the Historical Signals in Multiple Sequence Alignments of Phylogenetic Data

Lars S. Jermiin[1,2,3,4,5,6,*], Karen Meusemann[7], Bernhard Misof[7], Denis C. Shields[2,3]
[1] *Systems Biology Ireland, University College Dublin, Belfield, Dublin 4, Ireland*
[2] *School of Medicine, University College Dublin, Belfield, Dublin 4, Ireland*
[3] *Conway Institute of Biomolecular & Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland*
[4] *School of Biology & Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland*
[5] *Earth Institute, University College Dublin, Belfield, Dublin 4, Ireland*
[6] *Research School of Biology, Australian National University, Acton, ACT 2601, Australia*
[7] *Leibniz Institute for the Analysis of Biodiversity Change, 53113 Bonn, Germany*

**Lars S Jermiin, Systems Biology Ireland, School of Medicine, University College Dublin, Belfield, Dublin 4, Ireland, lars.jermiin@ucd.ie*

Abstract

While there are sophisticated methods for inferring multiple sequence alignments (MSAs), models of sequence evolution, and phylogenetic trees from sequence data, the initial key

Source: Jermiin LS et al. (2023) **Systematic Biology** (in review)

68

Research data management • benefits

Although **good RDM requires time and attention to detail**, it is also likely to

- Make it easier and faster to **recall what you did** months or years ago
- Make it easier and faster to **respond to enquiries and peer reviews**
- Improve your **standing in collaborative research projects**
- Improve the **quality of your research and research output**
- Increase your **scientific impact** (e.g., though citations)
- Improve the **transparency and reproducibility of your research**
- **Safeguard you** against accusations of engaging in fraudulent research practices

69

Thank you

Acknowledgement

I am grateful to Drs Vassilios Ioannidis and Cécile Lebrand (Université de Lausanne) and Grégoire Rossier (Swiss Institute of Bioinformatics) for sharing their knowledge on DMP and RDM, and for allowing me reuse some of their slides from a workshop on RDM and DMP (Bern, 26-27 October 2022).

For further details, contact

lars.jermiin@ucd.ie

70