MDC MAX DELBRÜCK CENTER FOR MOLECULAR MEDICINE IN THE HELMHOLTZ ASSOCIATION

Webinar "Uloga časopisa u poticanju pristupa istraživačkim podacima", 5. srpanj 2022.
Organizatori: Udruga Znak i CROSSDA

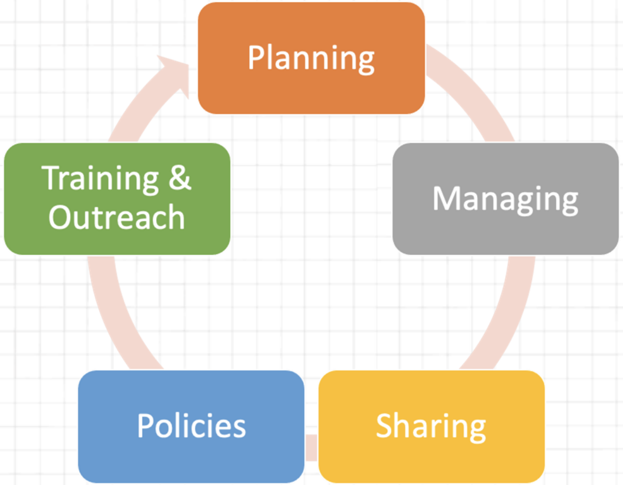
Requirements from publishers for data management and storage

Inga Patarcic, PhD
inga.patarcic@mdc-berlin.de

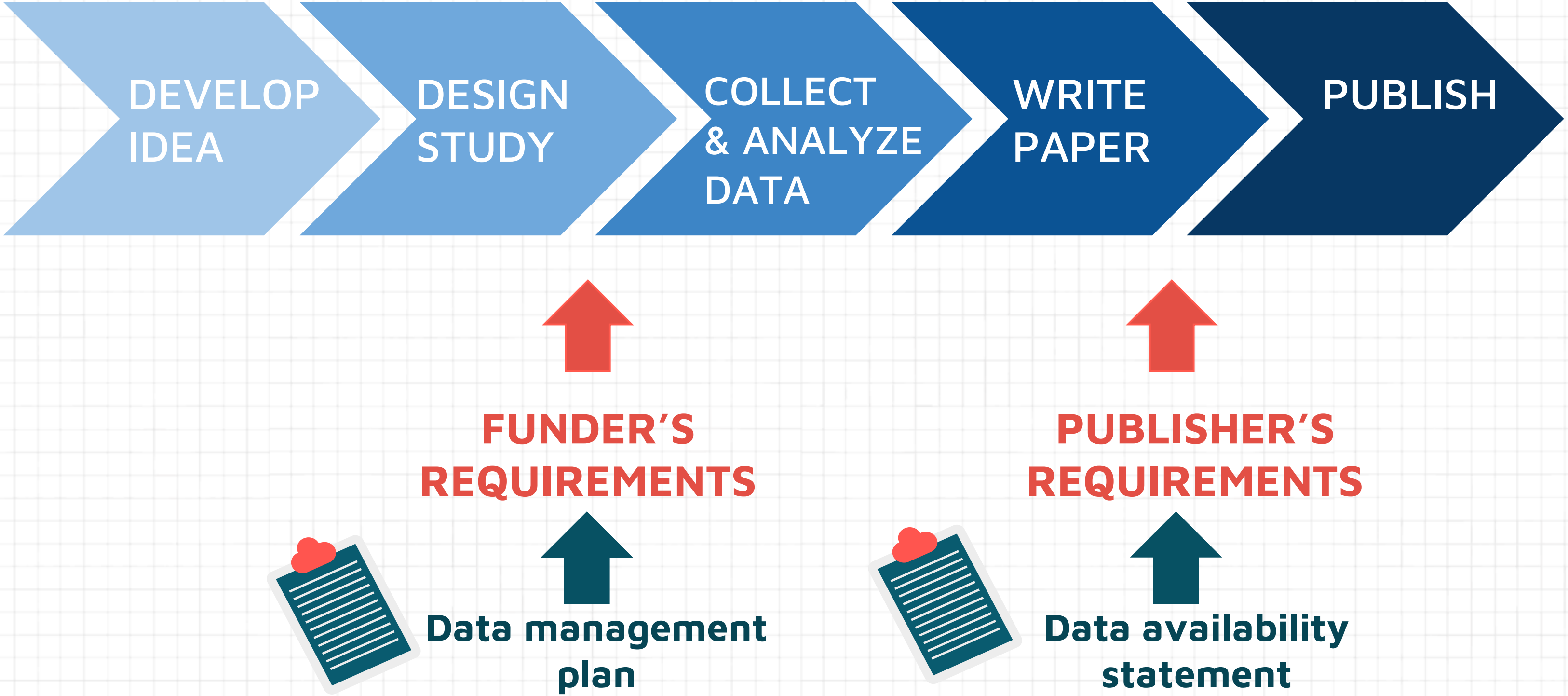
05.07.2022, Berlin

Working as a RDM

- → **67 RDM** units in Germany (forschungsdaten.org).
- → Plenty of positions in **Nederlands and the UK**
- → **No** specific university education
  - ◆ Research Data Management and Sharing on coursera
  - ◆ ORION MOOC for Open Science in the Life Sciences
  - ◆ Mantra- Free online Research Data Management Training
  - ◆ FOSTER- Open Science Training Courses
  - ◆ Datatree Free online course on research data management
  - ◆ Open Science MOOC



MOTIVATION: Requirements from funders and publishers for data management and storage



DEVELOP IDEA

DESIGN STUDY

COLLECT & ANALYZE DATA

WRITE PAPER

PUBLISH

FUNDER'S REQUIREMENTS

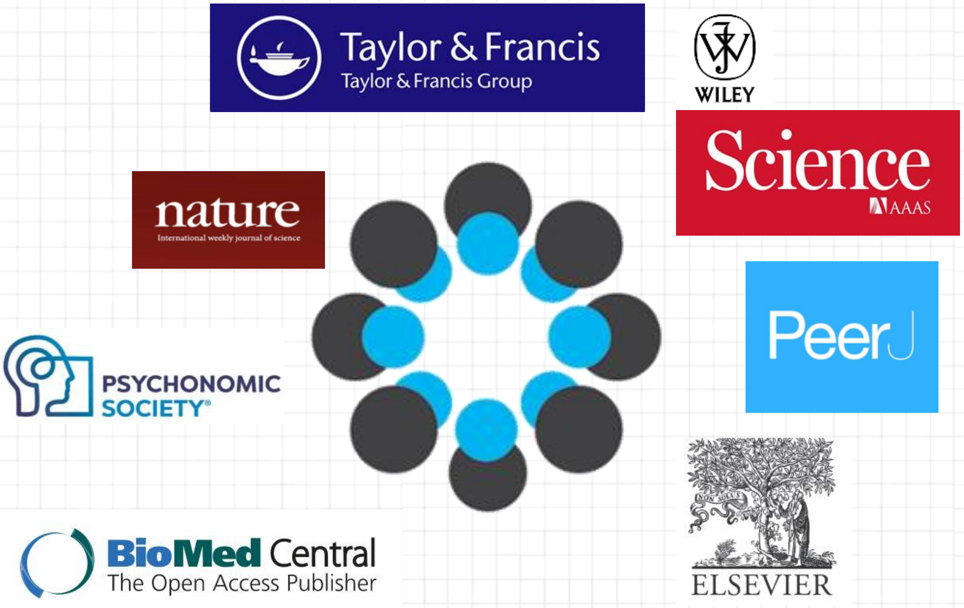
Data management plan

PUBLISHER'S REQUIREMENTS

Data availability statement

Requirements of the journals

Unifying journal’s TOP (transparency and openness) guidelines



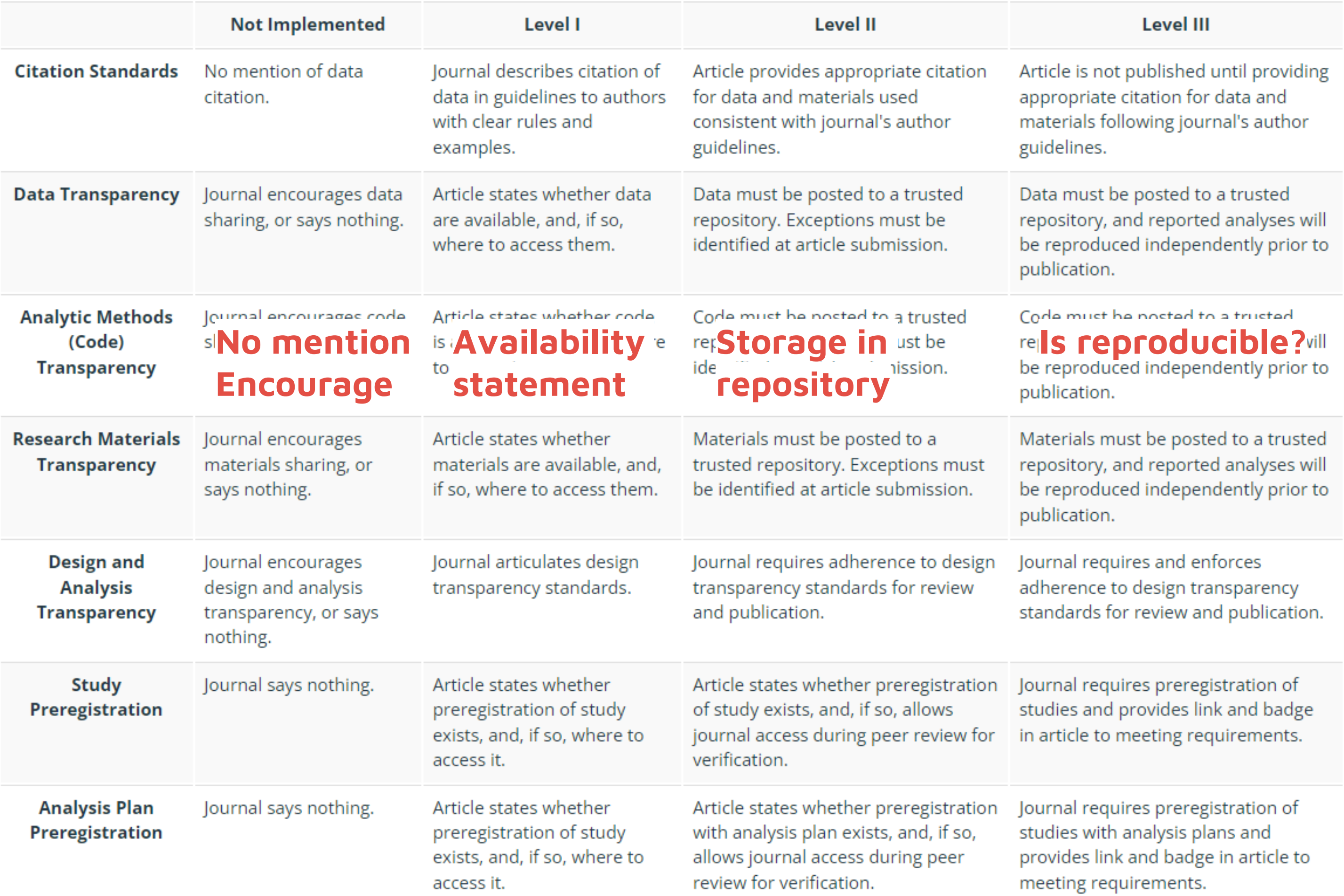
Journals implementing TOP guidelines (Center for Open Science)

Eight policies: Data citation; Data Transparency; Materials Transparency, Code Transparency; Design and Analysis; Study Preregistration; Preregistration of Analysis Plans; Replication Policies

Modular: Use the policies that you are ready to implement

Three Tiers: Low barrier to entry. Room for improvement over time.

View the full TOP Guidelines: osf.io/xd6qr



	Not Implemented	Level I	Level II	Level III
Citation Standards	No mention of data citation.	Journal describes citation of data in guidelines to authors with clear rules and examples.	Article provides appropriate citation for data and materials used consistent with journal's author guidelines.	Article is not published until providing appropriate citation for data and materials following journal's author guidelines.
Data Transparency	Journal encourages data sharing, or says nothing.	Article states whether data are available, and, if so, where to access them.	Data must be posted to a trusted repository. Exceptions must be identified at article submission.	Data must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.
Analytic Methods (Code) Transparency	Journal encourages code sl No mention Encourage	Article states whether code is e to Availability statement	Code must be posted to a trusted rep ust be ide ission. Storage in repository	Code must be posted to a trusted re vill be reproduced independently prior to publication. Is reproducible?
Research Materials Transparency	Journal encourages materials sharing, or says nothing.	Article states whether materials are available, and, if so, where to access them.	Materials must be posted to a trusted repository. Exceptions must be identified at article submission.	Materials must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.
Design and Analysis Transparency	Journal encourages design and analysis transparency, or says nothing.	Journal articulates design transparency standards.	Journal requires adherence to design transparency standards for review and publication.	Journal requires and enforces adherence to design transparency standards for review and publication.
Study Preregistration	Journal says nothing.	Article states whether preregistration of study exists, and, if so, where to access it.	Article states whether preregistration of study exists, and, if so, allows journal access during peer review for verification.	Journal requires preregistration of studies and provides link and badge in article to meeting requirements.
Analysis Plan Preregistration	Journal says nothing.	Article states whether preregistration of study exists, and, if so, where to access it.	Article states whether preregistration with analysis plan exists, and, if so, allows journal access during peer review for verification.	Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.

SOURCE: cos.io/top

An example of TOP guidelines for individual journals

	Level 1	Level 2	Level 3
Data Citation	✓		
Data Transparency	✓		
Materials Transparency	✓		
Code Transparency		✓	
Design & Analysis		✓	
Study Preregistration	✓		
Analysis Preregistration	✓		
Replication	✓		

nature
International weekly journal of science

An example of TOP guidelines for individual journals

	Level 1	Level 2	Level 3
Data Citation		✓	
Data Transparency		✓	
Materials Transparency	✓		
Code Transparency		✓	
Design & Analysis	✓		
Study Preregistration	✓		
Analysis Preregistration	✓		
Replication	✓		

http://www.sciencemag.org/authors/science-journals-editorial-policies

Data Transparency

OPEN DATA

Level 1: Disclosure

Article states whether or not data are available. If so, give URL.

THE EMBO JOURNAL

EMBOpress

Cell Reports

Cell

PNAS

nature International weekly journal of science

Level 2: Mandate

Data MUST be posted in a trusted repo (exceptions permitted for legal or ethical constraints).

PLOS

SCIENTIFIC DATA

eLIFE

Science AAAS

GENOME RESEARCH

PeerJ

F1000Researc

Open for Scien

Level 3: Verified Mandate

Level 2 + Can results be replicated?

Reported analysis will be reproduced independently prior to publication

Journal of Experimental Political Science JEPS

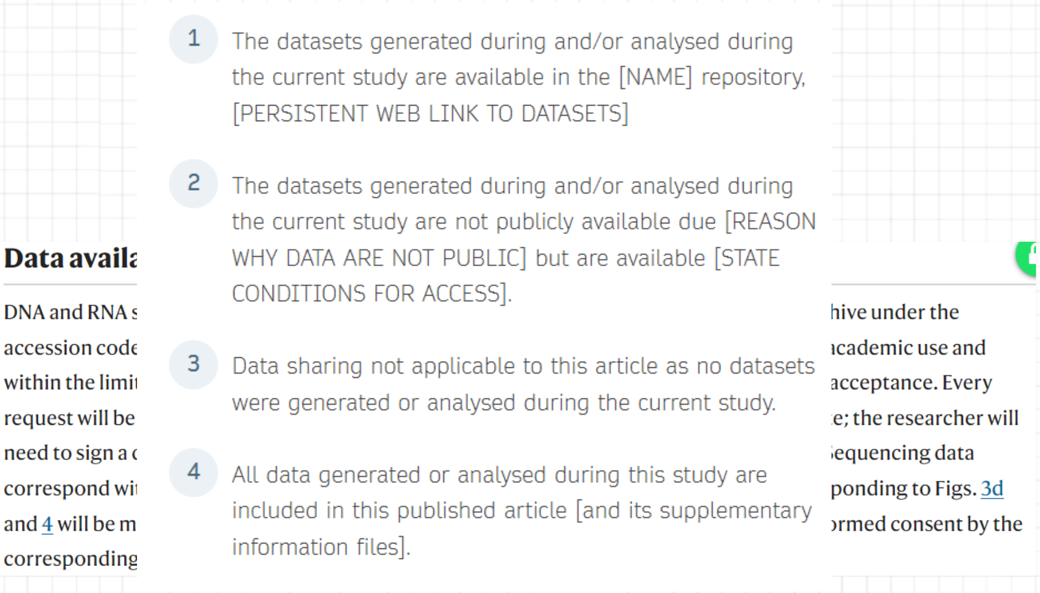
Meta-Psychology

Journal of Peace Research 2019 Special Issue

IREE International Journal for Re-Views in Empirical Economics

Data Transparency: Level 1 (Springer Nature)

- Data available at submission
- Must include a **data availability statement**
- min dataset may be provided through preferred deposition in repositories
- Providing large datasets in supplementary information is strongly discouraged



1 The datasets generated during and/or analysed during the current study are available in the [NAME] repository, [PERSISTENT WEB LINK TO DATASETS]

2 The datasets generated during and/or analysed during the current study are not publicly available due [REASON WHY DATA ARE NOT PUBLIC] but are available [STATE CONDITIONS FOR ACCESS].

3 Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

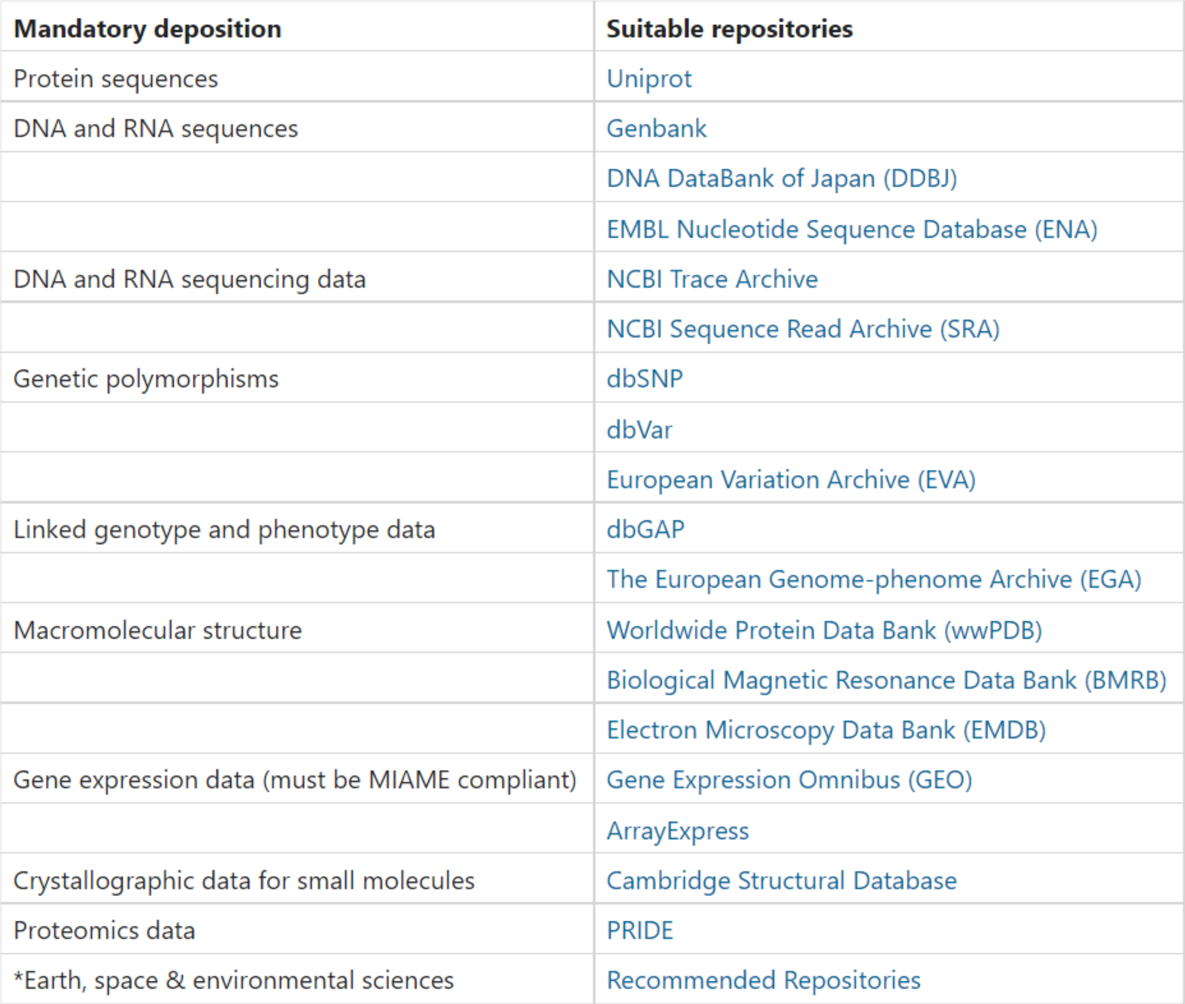
4 All data generated or analysed during this study are included in this published article [and its supplementary information files].

Data availa

DNA and RNA s
accession code
within the limit
request will be
need to sign a c
correspond wi
and 4 will be m
corresponding

hive under the
academic use and
acceptance. Every
e; the researcher will
equencing data
ponding to Figs. 3d
ormed consent by the

Mandates for the specific datasets (Springer Nature)



Mandatory deposition	Suitable repositories
Protein sequences	Uniprot
DNA and RNA sequences	Genbank
	DNA DataBank of Japan (DDBJ)
	EMBL Nucleotide Sequence Database (ENA)
DNA and RNA sequencing data	NCBI Trace Archive
	NCBI Sequence Read Archive (SRA)
Genetic polymorphisms	dbSNP
	dbVar
	European Variation Archive (EVA)
Linked genotype and phenotype data	dbGAP
	The European Genome-phenome Archive (EGA)
Macromolecular structure	Worldwide Protein Data Bank (wwPDB)
	Biological Magnetic Resonance Data Bank (BMRB)
	Electron Microscopy Data Bank (EMDB)
Gene expression data (must be MIAME compliant)	Gene Expression Omnibus (GEO)
	ArrayExpress
Crystallographic data for small molecules	Cambridge Structural Database
Proteomics data	PRIDE
*Earth, space & environmental sciences	Recommended Repositories

SOURCE: https://www.springernature.com/gp/authors/research-data-policy/recommended-repositories

Data transparency: Level 2

Acceptable Data Access Restrictions (PLOS)

Acceptable restrictions on public data sharing are detailed below. For an author to be the **sole named individual responsible for ensuring data access** is **not acceptable**.

- When **third-party data** cannot be publicly shared, authors must provide all information necessary for interested researchers to apply to gain access to the data.
- For studies involving **human research participant data** or other **sensitive data**, we encourage authors to share de-identified or anonymized data. However, when data cannot be publicly shared, we **allow authors to make their data sets available upon request**.

The following are examples of data that should not be shared:

- ➢ Name, initials, physical address
- ➢ Internet protocol (IP) address
- ➢ Specific dates (birth dates, death dates, examination dates, etc.)
- ➢ Contact information such as phone number or email address
- ➢ Location data

Data Transparency: Level 3

Verified Mandate

Can results be replicated using your data prior to publication?

Journal of Experimental Political Science

JEPS

Journal of Peace Research

2019 Special Issue

Journals with data transparency Level 3:

[1] Meta-Psychology
[2] Journal of Experimental Political Science
[3] International Journal for ReViews in Empirical Economics
[4] American Journal of Political Science
[6] Journal of Legal Studies
[7] Journal of Peace Research
[8] AEJ: Applied Economics
[9] AEJ: Economic Policy
[10] AEJ: Macroeconomics
[11] American Economic Review
[12] Economic Policy
[13] Journal of Economic Perspectives
[14] Political Science Research and Methods
[15] The Journal of Politics
[16] AEA Papers & Proceedings
[17] AEJ: Microeconomics
[18] AER: Insights
[19] Biometrical Journal
[20] Economic Journal
[21] Journal of Economic Literature
[22] International Organization
[23] International Studies Quarterly

Code Transparency

Level 1: Disclosure

Article state whether or not code is available. If so, give URL.

THE EMBO JOURNAL
EMBOpress
Science AAAS
Cell Reports
Cell
PNAS
nature International weekly journal of science

Level 2: Mandate

Code MUST be posted in a trusted repo (exceptions permitted for legal or ethical constraints).

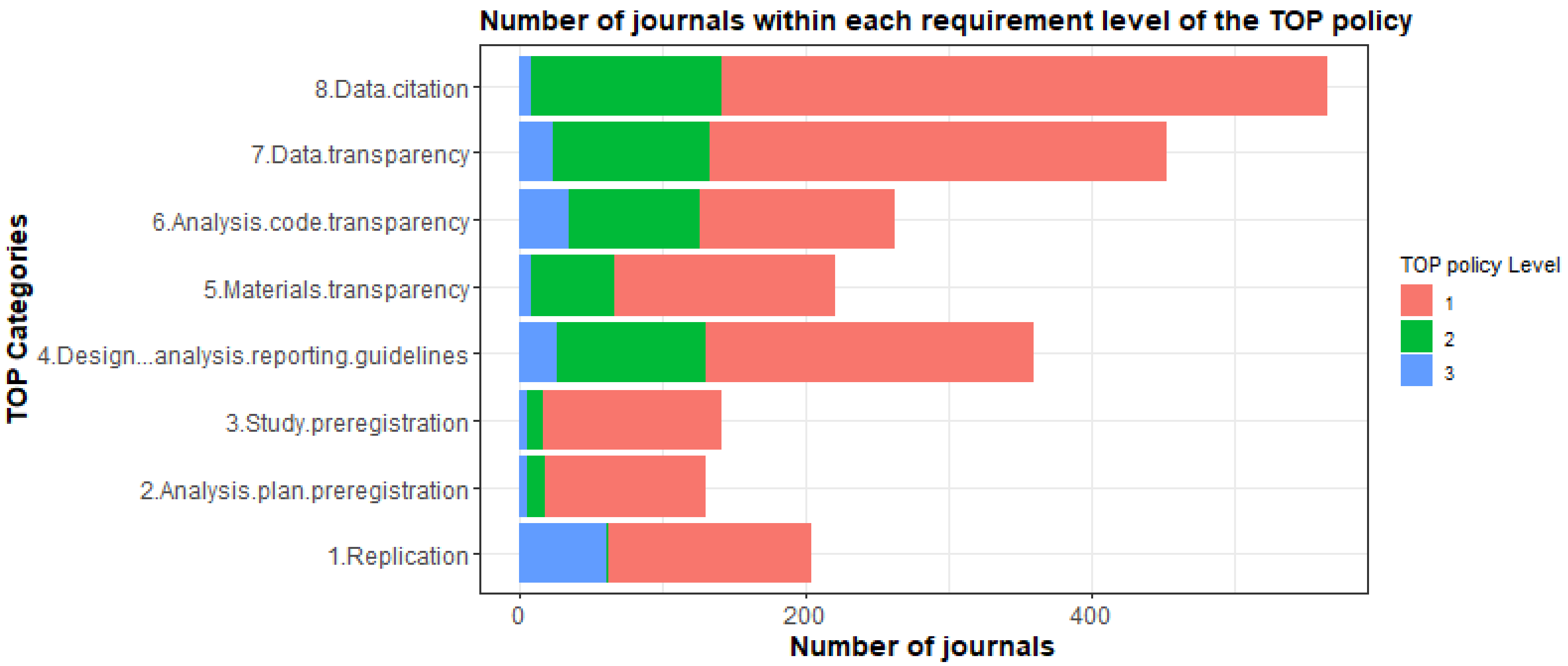
PLOS
F1000Research Open for Science
eLIFE
GENOME RESEARCH
PeerJ

Level 3: Verified Mandate

Level 2 + Can code be replicated? Reported analysis will be reproduced independently prior to publication

(GIGA)n SCIENCE
PLOS COMPUTATIONAL BIOLOGY

The current state of journal requirement trends



DATA SOURCE: https://osf.io/qatkz/ CODE: https://github.com/IngaPa

Comparison of different data policies across journals

Journal	Data Citation	Data Transp.	Material Transp.	Code Transp.	Design & Analysis	Study Prereg	Analysis Prereg.	Replication	Sum
Ncomms, Nature, Nature Machine Intelligence, Genome Biology	1	1	3	1	2	1	1	1	11
Bioinformatics	2	2	0	2	2*	0	0	0	8
NAR	0	2	2	2	0	0	0	0	6
F1000	1 or 2	1 or 2**	1 or 2	1 or 2***	2	3	3	NA	16
Gigascience	2	2	2	3***	3*	0	0	0	12
Science	2	2	2	1	1	1	1	1	11
PNAS	1	1	1	1	1	1	1	1	8
Cell, Cell Reports	2	1	2	1	2	1	1	1	11
Biophysical Journal	0	2	2	2	2	0	0	0	8
PLOS Computational Biology	1	2	2	2-3****	2	0	0	0	13
The EMBO Journal	1	1	1	1	1	0	0	0	5
PLOS Biology	1	2	2	2	2	0	0	NA	9
Nature Methods	1	2	1	1	2	1	1	1	10
eLife	2	2	2	2	2	1	1	1	13
Genome Research	2	2	2	2	0	0	0	0	8

Open for discussion

Thank you for your attention!

MDC MAX DELBRÜCK CENTER FOR MOLECULAR MEDICINE IN THE HELMHOLTZ ASSOCIATION
Planning
Managing
Sharing
Policies
Training & Outreach

inga.patarcic@mdc-berlin.de

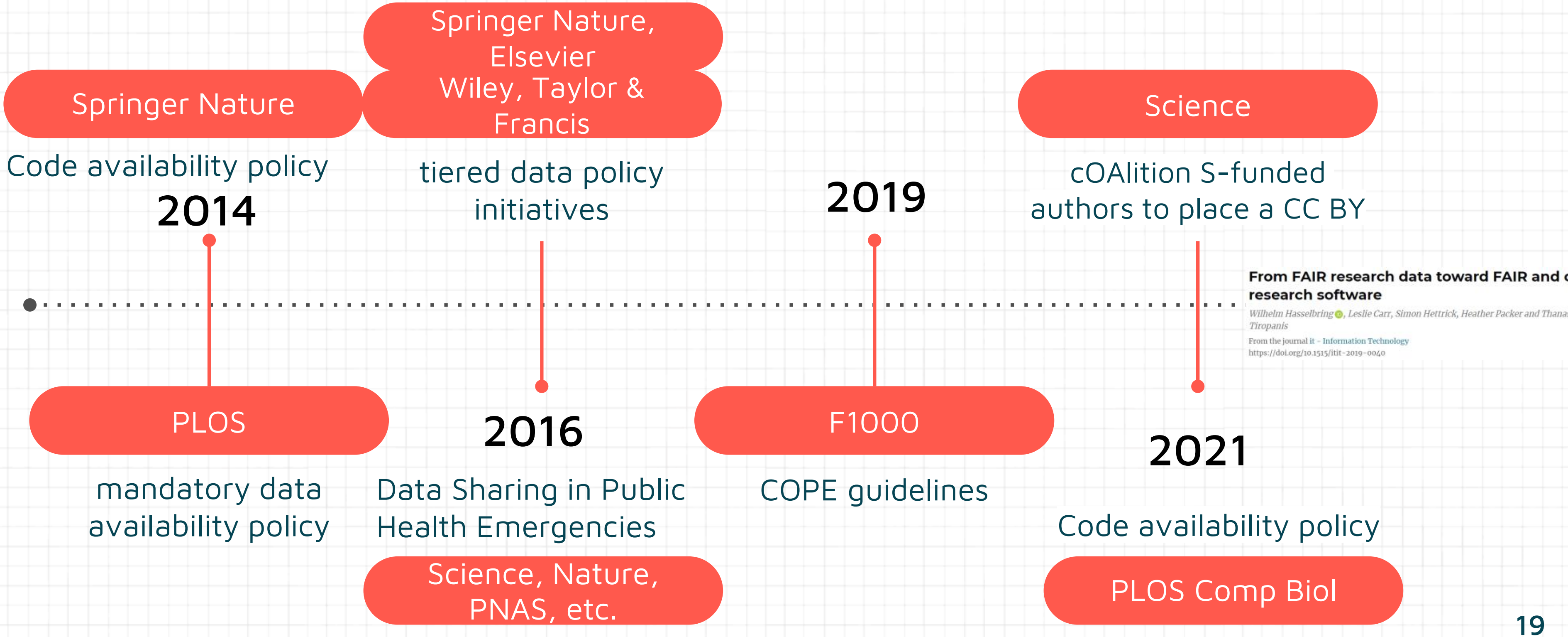
Presentation template by Slidesgo, Flaticon and Freepik

Code Transparency: Level 2

Science
AAAS

- All computer **code** central to the findings being reported **should be available** to readers to ensure reproducibility
- Commercially available software or publicly archived source code should be **appropriately referenced** (with the version included).
- Author-written **source code** should be **archived** in a permanent public repository prior to publication and likewise cited (see Data and Code Deposition).
- In exceptional cases where for example, security concerns, legal restrictions, or proprietary hardware preclude sharing of custom code, an alternate means of ensuring reproducibility must be arranged with the editor. Preferred option: **include pseudocode that fully clarifies the underlying algorithms**; this pseudocode will be subject to peer review and may require further elaboration in accord with reviewers' feedback.
- The reason(s) for the **code restrictions** in these special cases **should be explained** clearly in the acknowledgments.

Increasing levels of demands



2014

Springer Nature

Code availability policy

PLOS

mandatory data availability policy

2016

Springer Nature, Elsevier Wiley, Taylor & Francis

tiered data policy initiatives

Data Sharing in Public Health Emergencies

Science, Nature, PNAS, etc.

2019

F1000

COPE guidelines

Science

cOAlition S-funded authors to place a CC BY

2021

Code availability policy

PLOS Comp Biol

From FAIR research data toward FAIR and c research software

Wilhelm Hasselbring, Leslie Carr, Simon Hettrick, Heather Packer and Thanas Tiropanis

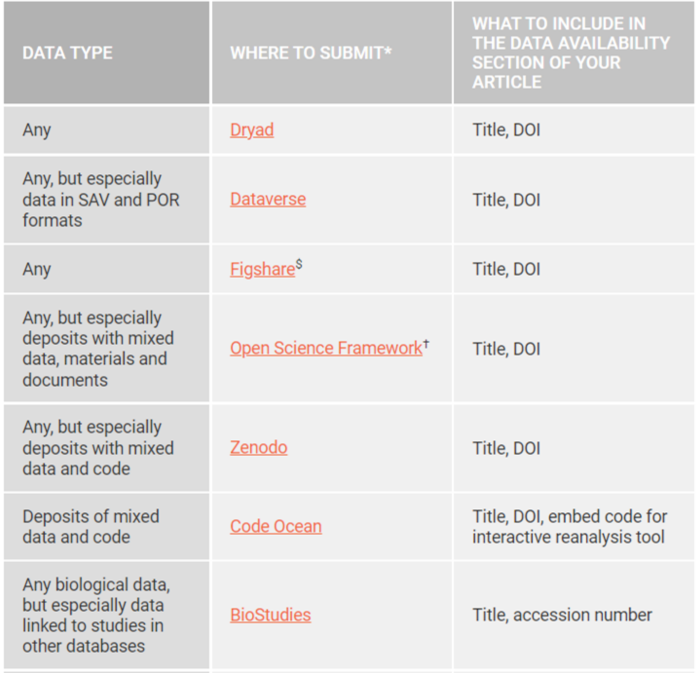
From the journal it - Information Technology

https://doi.org/10.1515/itit-2019-0040

Generalist repositories

Recommended by Springer Nature (below) and F1000 (right):

- Dryad Digital Repository
- figshare
- Harvard Dataverse
- Open Science Framework
- Science Data Bank
- Zenodo



DATA TYPE	WHERE TO SUBMIT*	WHAT TO INCLUDE IN THE DATA AVAILABILITY SECTION OF YOUR ARTICLE
Any	Dryad	Title, DOI
Any, but especially data in SAV and POR formats	Dataverse	Title, DOI
Any	Figshare[$]	Title, DOI
Any, but especially deposits with mixed data, materials and documents	Open Science Framework[†]	Title, DOI
Any, but especially deposits with mixed data and code	Zenodo	Title, DOI
Deposits of mixed data and code	Code Ocean	Title, DOI, embed code for interactive reanalysis tool
Any biological data, but especially data linked to studies in other databases	BioStudies	Title, accession number

Source: https://f1000research.com/

Charité monographie requirements

1. An accurate and complete copy of the primary dataset of the underlying works must be submitted in digital form, which enables the allocation of results to the relevant primary data.
2. In the case of data stored on a server, it must be disclosed how the data can be accessed and storage of the data for 10 years after initiation of the doctoral examination procedure must be ensured.
3. Primary datasets containing personal data must be suitably pseudonymised. The relevant allocation table must be kept by the doctoral candidate for 10 years and presented upon request.

https://promotion.charite.de/promotionsverfahren/po_2017/eroeffnung/monographie/

https://promotion.charite.de/en/procedure/regulations_2017/initiation_of_the_doctoral_examination_procedure/monograph/

Data pre-processing steps such as transformations, re-coding, re-scaling, normalization, truncation, and handling of below detectable level readings and outliers should be fully described; any removal or modification of data values must be fully acknowledged and justified.

The number of sampled units, n, upon which each reported statistic is based must be stated.

For continuous variables, distributions should be described using graphical displays such as scatterplots, boxplots, or histograms or by reporting measures of central tendency (e.g., mean or median) and dispersion (e.g., SD, interquartile range).

For continuous variables that are approximately normally distributed, mean and SD are suitable measures for center and dispersion, respectively.For continuous variables with asymmetrical distributions, median and range (or interquartile range) are preferred to mean and SD.All measures of central tendency or dispersion that are used should be identified.

For very small samples sizes (e.g., n < 20), presentation of all data values in tabular format is desirable unless presentation would violate restrictions for privacy or confidentiality for human subjects.

Methods used for conducting statistical tests (e.g., t-test, Wilcoxon signed rank test, Wald test of regression coefficient) and for constructing confidence intervals (e.g., normal-based 95% CI: mean ± 2SD, likelihood ratio-based interval) should be clearly stated.

The testing level (alpha) and whether one-sided or two-sided testing was used should be reported for each statistical test; typically, two-sided testing is appropriate, but if one-sided testing is used its use should be justified.

Adjustments made to alpha levels (e.g., Bonferroni correction) or other procedure used to account for multiple testing (e.g., false discovery rate control) should be reported.

When Bayesian analyses are conducted, any assumptions made for prior distributions must be fully described.Sufficient information should be supplied to allow readers to judge whether any assumptions necessary for the validity of statistical approaches (e.g., data are normally distributed, survival data are consistent with proportional hazards in a Cox regression model) have been verified.

An accounting of missing data values should be provided; if imputed data values are used in statistical analyses, the methods used for imputation should be fully described.

Authors should present results in a complete and transparent fashion so that stated conclusions are backed by appropriate statistical evaluation and limitations of the study are frankly discussed

Point estimates of population parameters (e.g., mean, correlation coefficient, slope) or comparative measures (e.g., mean difference, odds ratio, hazard ratio) should be accompanied by a measure of uncertainty such as a standard error or a confidence interval.

Guidelines for ML and statistics

Journals with written guidelines:

- Bioinformatics
- Gigascience
- Science
- PLOS Computational Biology
- EMBO Journal

- Papers using **leave-one-out will be editorially rejected** unless there is a special circumstance (PLOS Comp Biol)

- Use of **literate programming outputs**, such as **Jupyter notebooks**, and/or publishing environments, such as **Code Ocean** or **Gigantum** (GigaScience)

- Provide support / **maintenance** for **min 3yrs** post publication (Bioinformatics)

- Data pre-processing steps such as transformations, re-coding, re-scaling, normalization, truncation, and handling of below detectable level readings and outliers should be fully described; any removal or modification of data values must be fully acknowledged and justified (Science).

- the manuscript **should be available for testing by reviewers**, and also include test data (GigaScience)

Repositories for unstructured and/or Large Data*

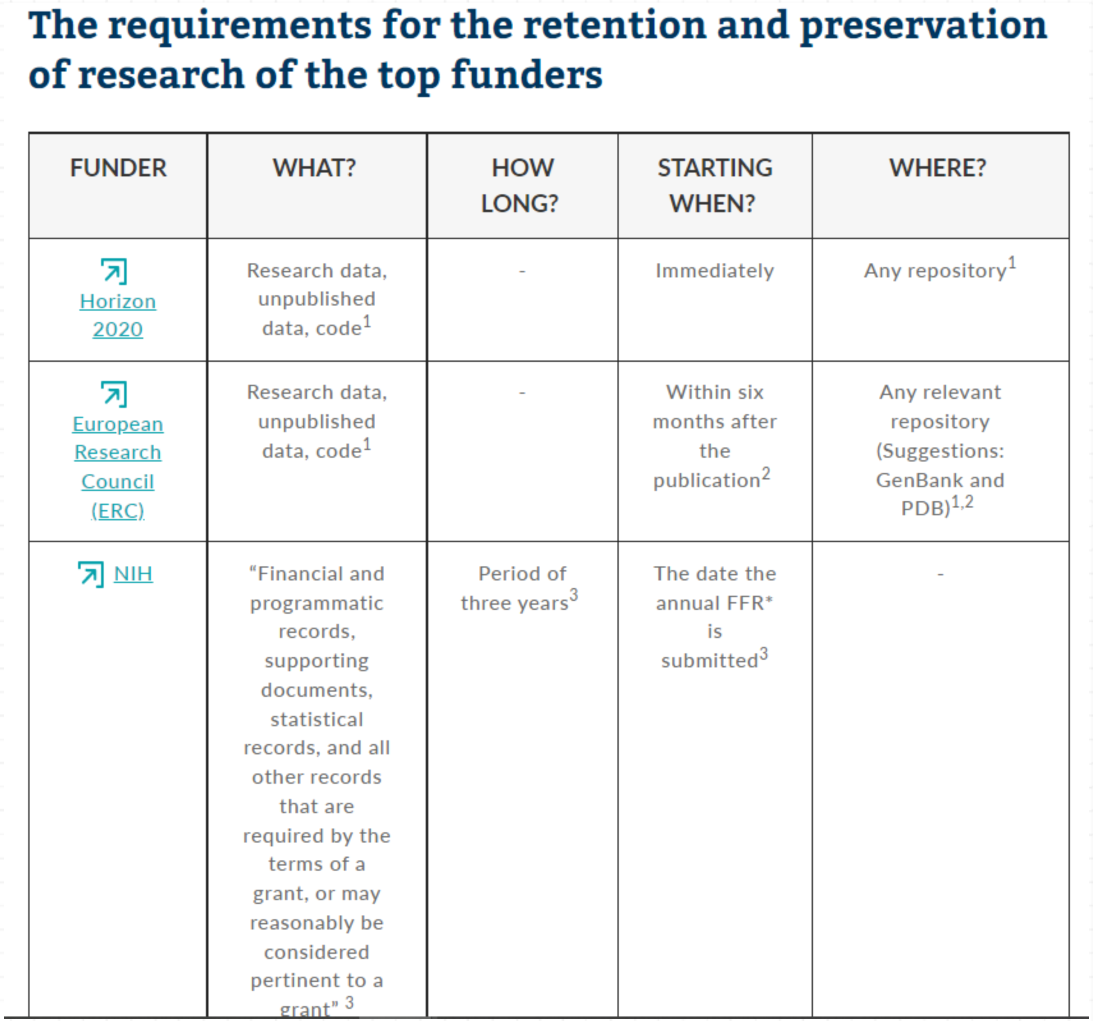
Recommended by PLOS Comp Biol:

- BioStudies
- CSIRO Data Access Portal
- Data Archiving and Networking Services (DANS)
- GigaDB
- SimTK
- Swedish National Data Service

*PLOS would like to thank the Open Access Nature Publishing Group journal, *Scientific Data*, for their own list of recommended repositories

Information about general purpose repositories

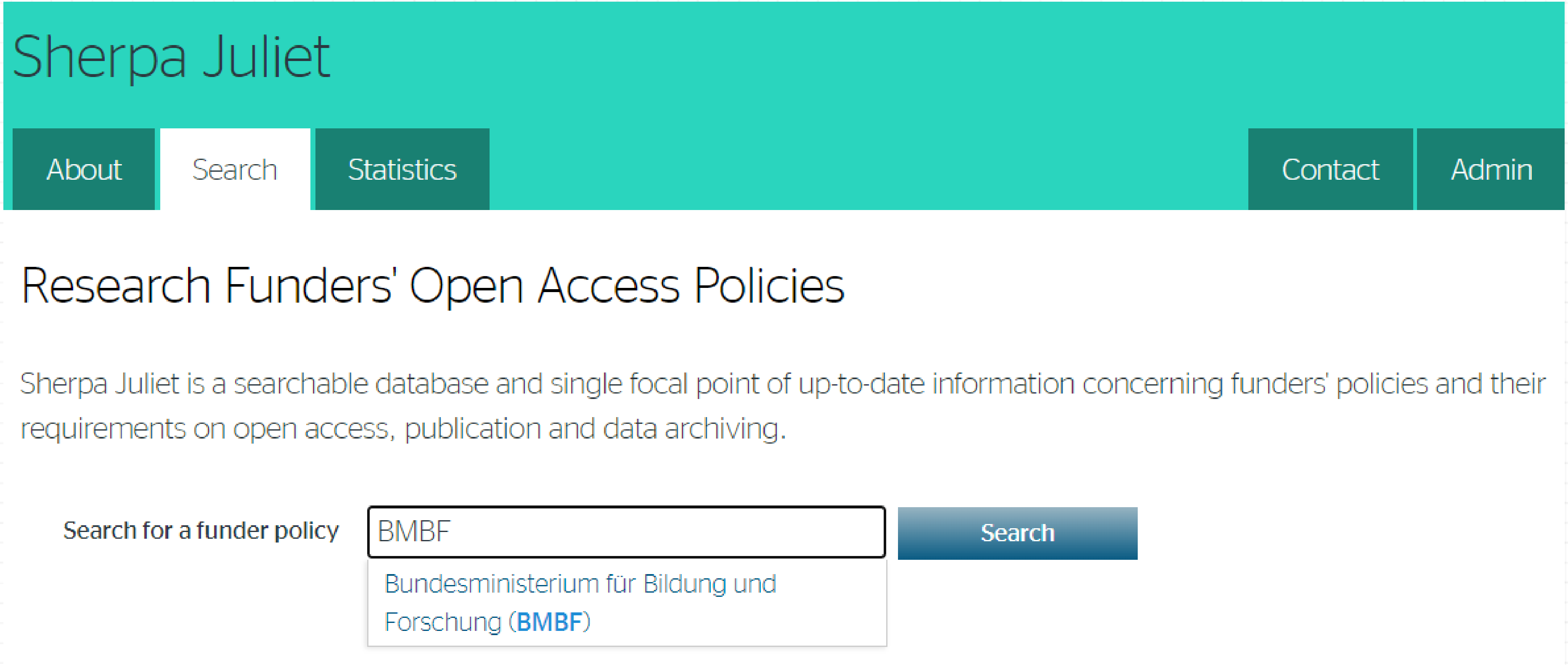
Repository	Costs	Size	Retention period
Dryad Digital Repository	DPC per data submission is $120 + charges excess storage fees for data totaling over 50GB ($50 for each additional 10GB)	300GB per data publication individual files should not exceed 10GB.	permanently archived and available through the California Digital Library's Merritt Repository.
figshare	250GB/$745 to 5TB/$11,860	250GB - 5TB	AWS, lifetime of the repository
Harvard Dataverse	free	up to 1TB and files of up to 2.5GB	NA
Open Science Framework	free	individual files must be 5GB, public projects and components to 50 GB	Always welcome to deactivate or delete your account. Uses Google Cloud (optional storage in Frankfurt).
Science Data Bank	Do not charge individual users at present. All fees incurred throughout the publishing process will be waived. However, we reserve the rights to charge users for data storage, review and publishing,	Any	Do not guarantee life-long free services. China
Zenodo	free	<50GB per dataset. If bigger, contact	Lifetime of the repository = lifetime of the host laboratory CERN (exp. programme defined for the next 20 years at least.)
Code Ocean	Free version: 20Gb, 10h/month	NA	NA



The requirements for the retention and preservation of research of the top funders

FUNDER	WHAT?	HOW LONG?	STARTING WHEN?	WHERE?
Horizon 2020	Research data, unpublished data, code[1]	-	Immediately	Any repository[1]
European Research Council (ERC)	Research data, unpublished data, code[1]	-	Within six months after the publication[2]	Any relevant repository (Suggestions: GenBank and PDB)[1,2]
NIH	"Financial and programmatic records, supporting documents, statistical records, and all other records that are required by the terms of a grant, or may reasonably be considered pertinent to a grant" [3]	Period of three years[3]	The date the annual FFR* is submitted[3]	-

SOURCE: https://www.mdc-berlin.de/research-data-management



Research Funders' Open Access Policies

Sherpa Juliet is a searchable database and single focal point of up-to-date information concerning funders' policies and their requirements on open access, publication and data archiving.

Search for a funder policy: BMBF

Bundesministerium für Bildung und Forschung (**BMBF**)

https://v2.sherpa.ac.uk/juliet/

- Projects **must** take **measures** to enable third parties to access, mine, exploit, reproduce and disseminate (free of charge to users) this research data, for example CC-BY or CC0 Licenses

- Authors are **strongly encouraged** to **provide open access** to monographs, books, conference proceedings and grey literature

- Participants **must provide** an initial **data management plan** within the first 6 months of the project starting

- **ensure open access** to the **publication** within the embargo period of **six months** of publication - 12 months in case of the social sciences and humanities

—Horizon 2020

Mandatory open science practices required by HE:

- **open access** to scientific publications
- responsible **management** of research data in line with the **FAIR principles** through the generalised use of **data management plans**, and open access to research data under the principle **'as open as possible, as closed as necessary'**
- information about the research **outputs**/tools/instruments needed to **validate** the conclusions of scientific publications or to validate/re-use research data
- **digital or physical access** to the results needed to validate the conclusions of scientific publications, unless exceptions apply
- in cases of **public emergency**, if requested by the granting authority, **immediate open access to all research** outputs under open licenses or, if exceptions apply, access under fair and reasonable conditions to legal entities that need the research outputs to address the public emergency

—Horizon Europe

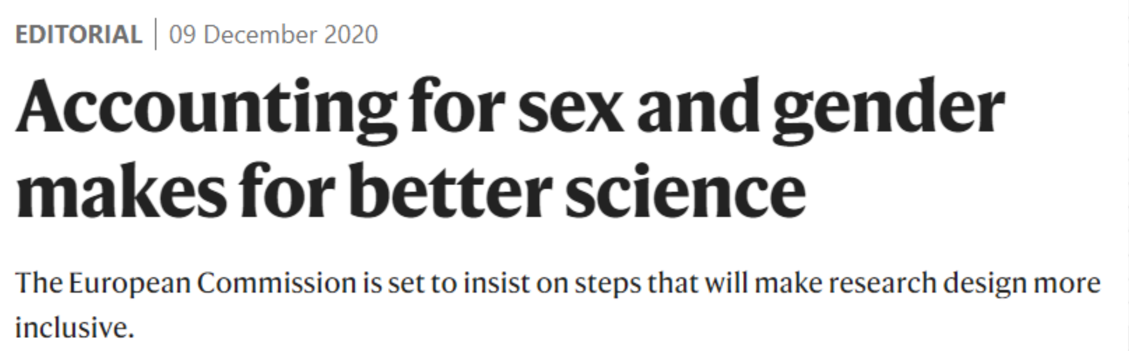
Requirements for projects generating or reusing data.

- Provide outline the measures planned in the project that tend to **increase reproducibility**
- **DMP** template (usually not required at submission stage) HERE
- **OA without embargo** under open licenses (such as CC),
- Information in the **repository + Open peer review** encouraged
- OA to **software, models, algorithms, workflows, protocols, simulations, electronic notebooks** and others is not required but strongly recommended.
- Access to `physical' results like cell lines, biospecimens, compounds, materials, etc. is also strongly encouraged.
- **citizen**, civil society and end-user engagement should be implemented in project, if approp.
- early and open sharing of research (preregistration, registered reports, pre-prints, or crowd-sourcing) encouraged

Did you know!?

Horizon Europe mandates a gender equality plan (GEP) for grant proposals.

*The Canadian Institutes of Health Research began to mandate that analyses of sex and gender be included in grant applications in 2010, and the US National Institutes of Health followed suit in 2016.



EDITORIAL | 09 December 2020

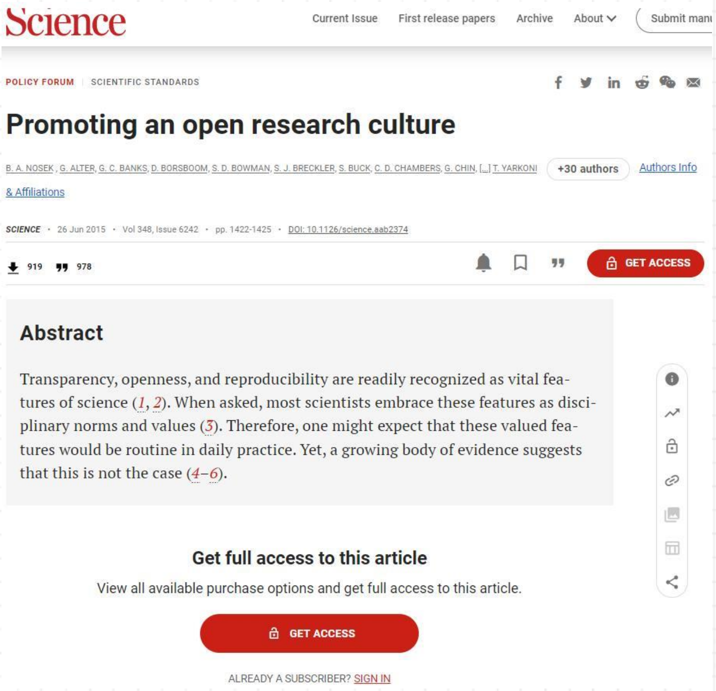
Accounting for sex and gender makes for better science

The European Commission is set to insist on steps that will make research design more inclusive.

Shared goal:
Improvement of data transparency, openness, availability

Data availability allows and facilitates:

- ➢ **Validation**, **replication**, **reanalysis**, new analysis, **reinterpretation** or inclusion into meta-analyses;
- ➢ **Reproducibility** of research;
- ➢ **Increase** the **value** of the investment made in funding scientific research;
- ➢ **Reduction** of the **burden on authors** in finding old data;
- ➢ **Enhance visibility** and ensuring **recognition** for authors, data producers and curators.



POLICY FORUM SCIENTIFIC STANDARDS

Promoting an open research culture

B. A. NOSEK, G. ALTER, G. C. BANKS, D. BORSBOOM, S. D. BOWMAN, S. J. BRECKLER, S. BUCK, C. D. CHAMBERS, G. CHIN, [...] T. YARKONI +30 authors Authors Info & Affiliations

SCIENCE • 26 Jun 2015 • Vol 348, Issue 6242 • pp. 1422-1425 • DOI: 10.1126/science.aab2374

919 978

Abstract

Transparency, openness, and reproducibility are readily recognized as vital features of science (1, 2). When asked, most scientists embrace these features as disciplinary norms and values (3). Therefore, one might expect that these valued features would be routine in daily practice. Yet, a growing body of evidence suggests that this is not the case (4–6).

Get full access to this article

View all available purchase options and get full access to this article.

SOURCE: https://journals.plos.org/plosone/s/data-availability