

UNIVERSITY OF OSLO

Sharing and archiving

research data

Live Håndlykken Kvale
University of Oslo Library
CC-BY-SA-4.0 2023



Agenda

- **Why archive research data?**

- Requirements
- Reasons for sharing
- Which data should be shared?

- **Data repositories**

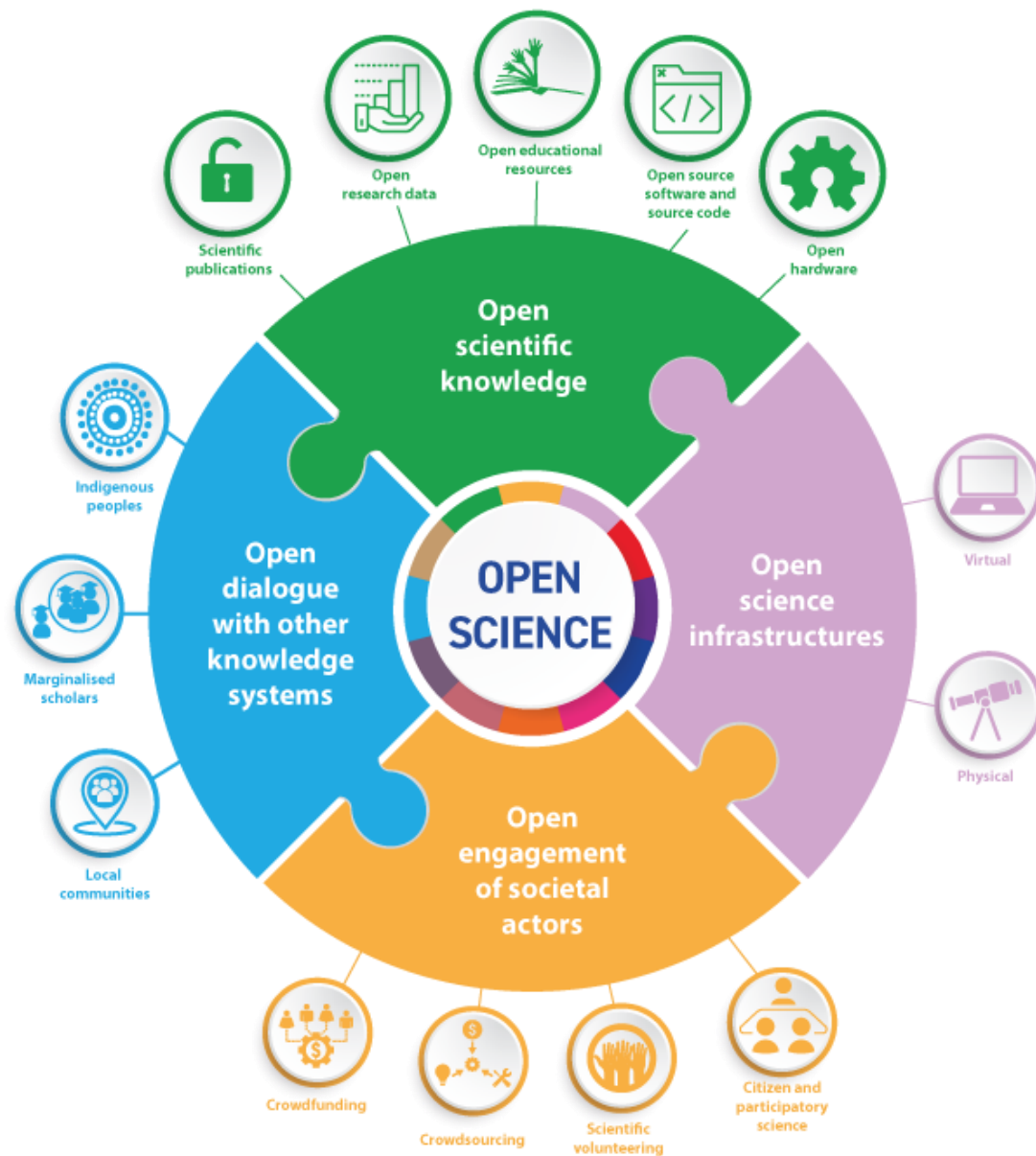
- Selecting repositories
- Levels of FAIR
- Types of repositories
- DataverseNO at UiO
- Data journals
- Archiving of code
- Finding a repository

- **What to consider**

- Preparing data
- What cannot be open
- PIDs
- Certification
- Licenses

- **Part 2:**

- **Examples**
- **Menti**
- **Key takeaways**
- **Q&A**



Research Data in Horizon Europe

- You must provide open access to research data under the principle 'as open as possible, as closed as necessary'.
- In general, you should deposit data generated or collected by the project as soon as possible after data production/generation or after adequate processing and quality control have taken place (for dynamic data, a snapshot of the data is enough).
- This should happen at the latest by the end of the project, and does not entail that data are immediately open, but rather that they have been deposited so that metadata information is available and hence information about the data is findable.
- Provide information via the repository about any research output or any other tools and instruments needed to re-use or validate the data.
- It is important that you check before depositing your research data that your chosen repository is technically capable of accepting the required metadata.

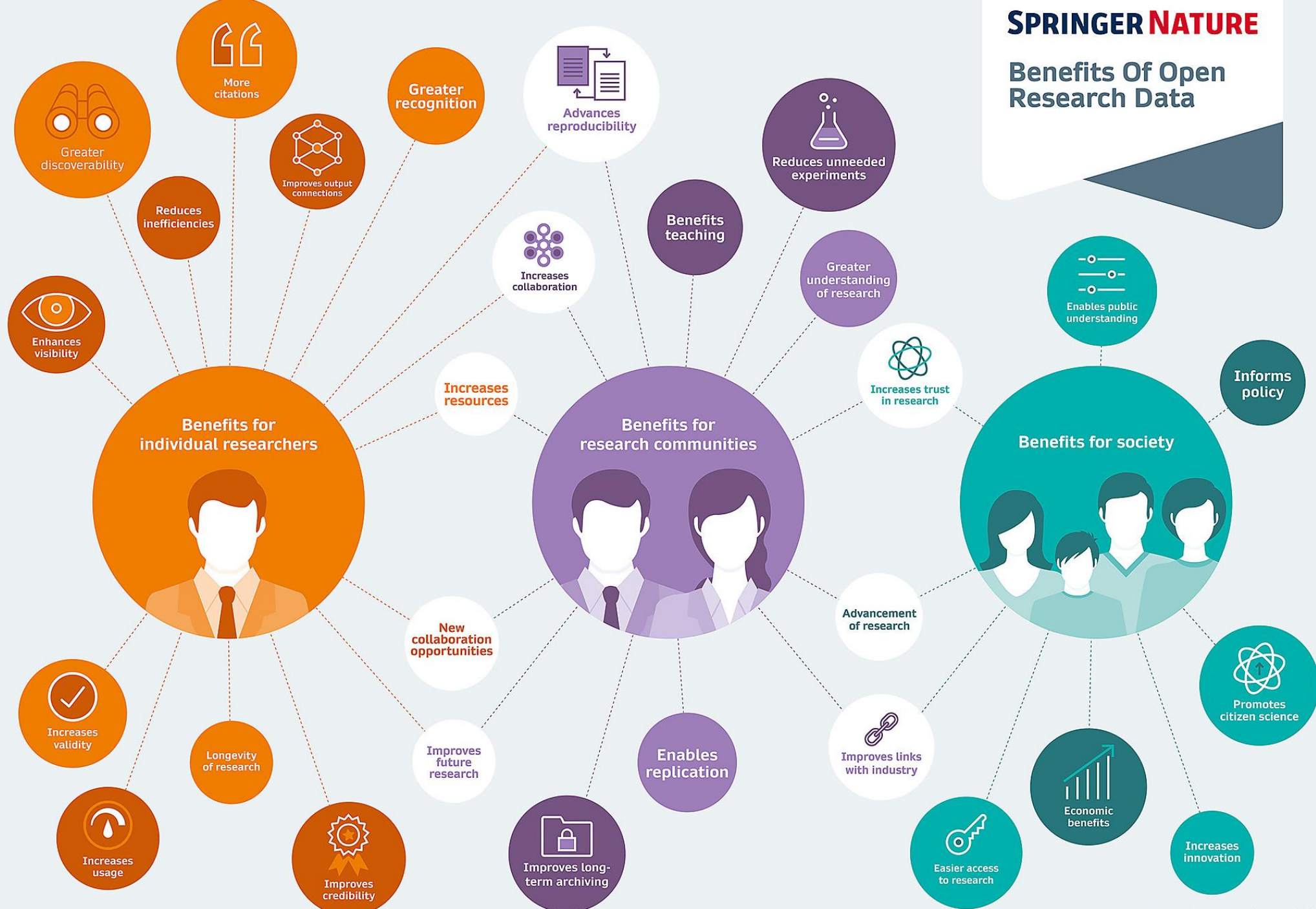
Local Requirements

Research data at the University of Oslo shall:

- be made openly available for further usage
- be made available at an early stage
- have metadata and be documented
- must be securely archived
- have licenses for access, reuse and redistribution
- made freely available (but the actual distribution cost should be covered)



Benefits Of Open Research Data



Reasons for sharing research data

External Factors

- Funder and publisher requirements
- Institutional requirements
- New assessment systems

Career Benefits

- Increased visibility
- More data reuse
- New collaborations
- Increased citations

Scientific Progress

- More robust research
- Enables verification of results
- Enables new collaborations across disciplines and borders
- Opens up for new uses of data
- Avoids duplication
- Easier to use data in teaching

Scientific Misconduct and the Myth of Self-Correction in Science

Wolfgang Stroebe^{1,2}, Tom Postmes², and Russell Spears²

¹Utrecht University, The Netherlands, and ²University of Groningen, The Netherlands

Abstract

The recent Stapel fraud case came as a shattering blow to the scientific community of psychologists and their image in the media and their collective self-esteem. The field responded with suggestions of how to prevent. However, the Stapel fraud is only one among many cases. Before basing recommendations would be informative to study other cases to assess how these frauds were discovered. The authors analyzed a sample of fraud cases to see whether (social) psychology is more susceptible to fraud than other disciplines. They evaluate whether the peer review process and replications work well in practice to detect fraud. There is evidence that psychology is more vulnerable to fraud than the biomedical sciences, and most frauds are detected through whistleblowers with inside information. On the basis of this analysis, the authors suggest a number of strategies to reduce the risk of scientific fraud.

Keywords

fraud, scientific misconduct, research integrity, replication, peer review

The news that the highly respected Dutch social psychologist Diederik Stapel had been accused of scientific misconduct and had admitted large-scale research fraud came as a terrible shock to the scientific community of social psychologists. Stapel was internationally renowned, and his work had received prestigious awards from the European Association of Social Psychology and the U.S. Society of Experimental Social Psychology. This scandal provided a field day for the international press, and psychology was portrayed as being highly vulnerable to scientific misconduct. The field responded with suggestions on how the risk of fraud could be reduced in the future (e.g., Crocker & Cooper, 2011; Mummendey, 2012; Roediger, 2012). However, the Stapel case, although very high profile, is only one of many fraud cases that were discovered in recent years. Instead of proposing changes on the basis of one case, it

the way that most of these frauds have been demonstrated demonstrate that the idea of the self-correction in science is a myth.

Notorious Cases of Research Fraud: A Review

The National Science Foundation (2001) defines research misconduct as fabrication, falsification, or plagiarism in performing, or reviewing research or in reporting results. Such misconduct must be committed knowingly, or in disregard of accepted practices. Fabrication of data involves totally inventing data that refers to manipulation of equipment or procedures that the research is not accurately represented



NMBU brukte åresvis på å granske mulig juks i en doktorgradsavhandling, men gjorde aldri jobben godt nok. Ikke før nå. (Foto: Lise Åserud / NTB)

Bortkommen banankasse funnet. Etter ni år er mulig jukesak avsluttet

KOMMENTAR: En bortkommen banankasse ble funnet i et kott. Tapte data var allikevel ikke slettet. NMBU har nå tatt en ny runde i sin ni år lange gransking av mulig juks i en doktorgradsavhandling.



Nina Kristiansen
JOURNALIST

Torsdag 22. september 2022 – 04:30



Det startet for ni år siden. En maraton av en sak som aldri fikk en avslutning. Verken granskerne eller NMBU gjorde en god nok jobb. Saken har versert i fagmiljøene som en verkebyll.

Nå kan saken ha fått en endelig slutt – i en ny gransking der data alle trodde var forsvunnet, har kommet til rette.

forskning.no rullet opp saken for to år siden. Da hadde rektor avsluttet saken, selv om mange i fagmiljøet mente den ikke var undersøkt grundig nok.

- Les hele saken fra 2019: [De så tegn til juks i en doktorgrad, men i løpet av tre granskninger har ingen undersøkt hele avhandlingen](#)

Oppsiktsvekkende resultater

Utoe_fyVB0U2LXce42n8UsAsV_9X5mPxcq1RiInHuOmrtkcy172mnojVVKXw2QxHaY69qYwheJ...aTBOIU5kxz5FKKDXe092KuSxMQwb6dYLcDT4WQaUJqowE57JdwmYGDQJUFH7sfXDRJ4w_Tk0

Utgave 1 – 2007

ARTIKLER

- › **Research misconduct: lessons to be learned?**
- › Investigation of scientific misconduct – some personal reflections
- › From Darsee to Sudbø: NLM's role in the retraction process
- › Can editors police scientific misconduct?
- › Playing by the rules – Scientific misconduct in a legal perspective
- › Can research institutions live up to their responsibilities?

Magne Nylenna

Research misconduct: lessons to be learned?

Michael 2007;4:7–9

«It can never happen here» has been the traditional saying in Norway when incidents of scientific dishonesty have been disclosed around the world. In a small country with a limited number of medical researchers, traditions for transparency and a strong belief in honesty, there has been a more or less naïve attitude to research misconduct.

In January 2006, on Friday 13th (1), the news was broken that a Norwegian scientist at Rikshospitalet-hospitalet, Jon Sudbø, had admitted to research misconduct in a recently published paper in *The Lancet* (2). It became a national sensation. The case made headline news in all major newspapers and television news, more than 330 media reports were registered over the first two weeks and the case received international attention.

At an early stage it became evident that the actual case, widely known as the Sudbø case, included fabrication, and a special Commission was appointed on 18 January to conduct an independent investigation. The Commission chaired by the Swedish epidemiologist, Professor Anders Ekblom, then presented an interim report on 30 June 2006 (2).

Many of Jon Sudbø's scientific publications are invalid due to the fabrication and manipulation of the raw data material, read the main conclusion of the Commission. Based on investigations into the history of Sudbø's scientific work, 38 published papers, the Commission found several breaches of scientific practice. Jon Sudbø, a dentist and physician, had been doing research on the early stages of oral cancer. One of his main questions was whether and to what extent different types of leukoplakia could be a risk factor for developing oral cancer. Sudbø's results had been published in high-profile international journals (1,3,4) and formed the basis for his PhD thesis. A series of flaws were, however, found in his data and the summing up by the Commission was harsh: «The Commission is of the opinion that the detected defects that have been exposed are too numerous, too great and too obvious to be attributed to errors, incompetence or the like; and that the raw data therefore appear to have been fabricated, falsified and adapted to the desired findings» (2).

The Sudbø case has been intensively discussed within the health care sector in Norway over the last year, and undoubtedly led to an increase in the awareness of research misconduct. Many institutions have reviewed their research programmes and routines. Supervisory and regulatory systems have been introduced.

The Sudbø case is also of interest from an international perspective. Learning from adverse events is a way to improve quality in all parts of medicine – research as well as patient treatment. What lessons can be learned from this and other revealed cases of scientific fraud for researchers, research institutions, scientific journals and other parties? Is a more detailed bureaucratic regulation of research the inevitable consequence of research misconduct being prevented through information campaigns? And who is really responsible for the prevention of research misconduct?

<https://journals.sagepub.com/doi/pdf/10.1177/1745691612460687>
<https://www.michaeljournal.no/article/2007/05/Research-misconduct-lessons-to-be-learned->
<https://forskning.no/forskningsetikk-veterinaermedisin/bortkommen-banankasse-funnet-etter-ni-ar-er-mulig-juksesak-avsluttet/2081272>

The New Information Retrieval Problem: Data Availability

- 1: Publication Pathway
- 2: Data Creator Pathway
- 3: Project Pathway
- 4: Phenomenon-Centric Search
- 5: Search by Institutional Repository at PIs organization

<https://doi.org/10.1002/pra2.796>

available. No awards from engineering and science and technology studies resulted in the retrieving data. In total, data was available from 11 awards out of the 56 awards that were analyzed.

TABLE 1. Research areas had varying levels data availability. In total, datasets from 11 out of 50 awards were found. Data was made available via cloud based services, data repositories, institutional repositories, or supplemental materials in publications.

Research Area	Number of Awards that Stored Retrievable Data	Location of Datasets	Pathway to find data
Oceanography	4	GitHub; Google Drive; Sequence Read Archive	1 and 2
Engineering	0	-	-
Science and Technology Studies	0	-	-
Cybersecurity	3	GitHub; UCF101, HMDB51, IEEE Dataport	1 and 2
Biology	4	Smithsonian Collections, Proteome, Dryad, GitHub, Zenodo; Institutional Repository at University	1, 2 and 3

We found a combination of storage sites for data. Data was found in domain repositories; generalist repositories; platforms like GitHub and Google Drive; and institutional repositories at universities. Coders use pathway one, two and three to find data.

An example

Our Initial Quality Check of your submission has now taken place. As a result, we need you to address the following points before your manuscript can progress any further:

*** As your data falls under the list of data types that must be deposited** (listed below), we ask that you please provide the relevant accession numbers and the name of the database in which your data is stored.

[.....]

Your availability of data and materials statement can take the following form: “The datasets generated and/or analysed during the current study are available in the [NAME] repository, [PERSISTENT WEB LINK OR ACCESSION NUMBER TO DATASETS]”.

Your paper has been placed back in the menu of the submitting author. To access it, please use the following link, making sure you log in with the same email address you registered with: [...]

[Explore content](#) [About the Journal](#) [Publish with us](#) [Sign up for alerts](#) [RSS feed](#)

[nature](#) > [scientific reports](#)

Publish with Scientific Reports

We're an open-access journal publishing rigorously peer-reviewed research from across the natural sciences, psychology, medicine and engineering.



Announcements

Journal Metrics

Scientific Reports has a 2-year impact factor: 4.996 (2021), and is the 5th most-cited journal in the world, with more than 696,000 citations in 2021*.

*2022 Journal Citation Reports® Science Edition (Clarivate Analytics, 2022).



Calls for Papers

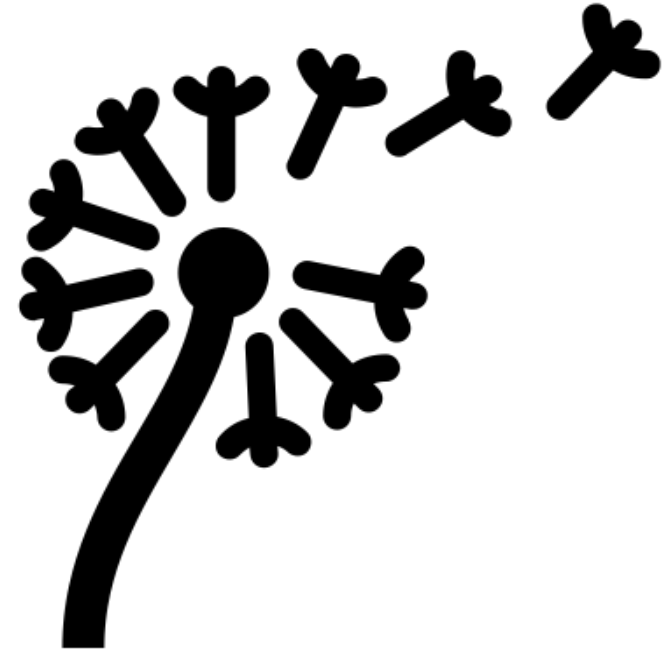
Highlight your research and enhance its visibility to your field by submitting to one of our Guest Edited Collections, led by our expert Editorial Board Members.

The median time from submission to acceptance for papers submitted to our Guest Edited Collections is 102 days.



Top 100 Collections

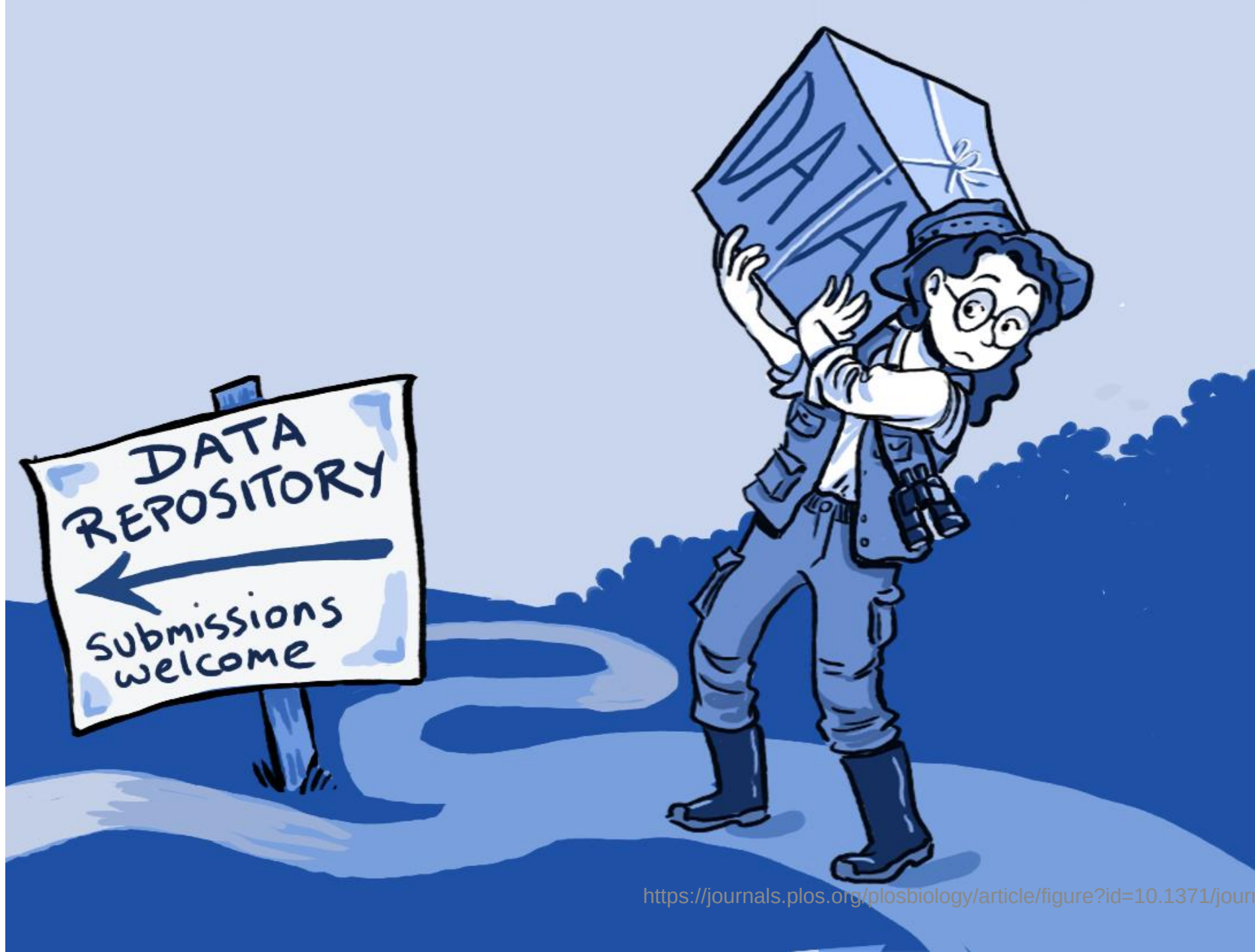
Strategy for archiving



Selecting data for archiving

- Does your dataset have a potential for reuse?
- (Inter-)national or historical importance
- Data quality
- Uniqueness or originality
- Size, scale, cost
- Innovativeness

How do I share data?



YES, a repository is the
best place to archive data

select a repository which...

- is domain specific if this exist in your field
- is a certified as trusted repository.
- supports persistent identifiers.
- offers curation.
- offers an informative landing page with metadata.
- attaches a licence.
- provide usage statistics.
- matches your particular data needs (formats, size, openness)
- provides guidance on how to cite deposited data.

Levels of FAIR

- Open data, but not FAIR
- FAIR metadata
- FAIR data Restricted access
- FAIR data Open access
- FAIR linked data Restricted access
- FAIR linked open data

METADATA

DATA



Levels of FAIR

- Open data, but not FAIR
- **FAIR metadata**
- FAIR data Restricted access
- FAIR data Open access
- FAIR linked data Restricted access
- FAIR linked open data

METADATA



DATA

Levels of FAIR

- Open data, but not FAIR
- FAIR metadata
- FAIR data Restricted access
- FAIR data Open access
- FAIR linked data Restricted access
- FAIR linked open data

METADATA



DATA



Levels of FAIR

- Open data, but not FAIR
- FAIR metadata
- FAIR data Restricted access
- **FAIR data Open access**
- FAIR linked data Restricted access
- FAIR linked open data

METADATA



DATA



Levels of FAIR

- Open data, but not FAIR
- FAIR metadata
- FAIR data Restricted access
- FAIR data Open access
- **FAIR linked data Restricted access**
- FAIR linked open data

METADATA



DATA



Levels of FAIR

- Open data, but not FAIR
- FAIR metadata
- FAIR data Restricted access
- FAIR data Open access
- FAIR linked data Restricted access
- **FAIR linked open data**

METADATA



DATA



Domain-specific data repositories



HEPData



General-purpose data repositories

The Zenodo logo consists of the word "zenodo" in a white, lowercase, sans-serif font, centered within a solid blue rectangular background.

Open Science Framework



National or institutional data archives



NIRD RESEARCH DATA ARCHIVE

Institutional data
archive



DataverseNO at UiO

- **Follow the guidance for the mandatory metadata fields**
- **Follow the curation guide**
 - Use the recommended file formats.
 - Use existing standard for describing data whenever possible.
- **Use ORCID as identifier for all authors**
- **The data gets a DOI**
 - A DOI is reserved at the draft stage
 - The DOI will not be activated and work until after the dataset is published.
 - If it turns out that the dataset cannot be published the DOI will never be activated.

ORCID



DataverseNO at UiO

- **DataverseNO is for Open data only**
 - Once a dataset is published, it is NOT possible to delete.
 - Make sure the data does not contain **personal or confidential information**.
- **Be detailed when writing the readme file.**
- **Choose an appropriate license.**
 - CC-0 is currently the default license in DataverseNO.
- The curator team will go through the metadata, and normally make suggestions for improvements.

DataverseNO at UiO

Contact the curator team if you:

- Are uncertain whether you hold the rights to the data.
- Need advice on which license to assign.
- Need to assign different licenses to different parts of the dataset.
- Need to share access with fellow researchers before publishing.
- Need to share the data for double-blind peer review.
- Have other needs regarding access.

The curation process

- After you submit your dataset, you will receive a curation report, a standardized document used by all DataverseNO institutions.
- Here we will inform you about required and/or recommended changes to be done before publication.
- We will return the dataset to you at the same time as you receive the curation report.
- You must resubmit the dataset after you have completed the revisions.
- We will publish the dataset when revisions have ensured the dataset is ready for publication.

DataverseNO Curation Report	
Author(s):	Kvale, Live
Dataset:	«Curation_Report_Replication_data_for-xxx_2020_xxxxx»»
Collection:	University of Oslo
Curator:	Elin Frøshaug
Date:	09.01.2023

DataverseNO aims to make published datasets as FAIR (Findable, Accessible, Interoperable, Reusable) as possible. In order for other researchers to be able to find, understand and reuse your data, it is important that you describe them in a good way before they are published. There are particularly two places in DataverseNO where such documentation is important:

1. In the metadata schema, you should enter as much relevant information as possible so that your dataset can be found via search engines such as Google Dataset Search.
2. The ReadMe file should provide an overview of your dataset and explain how you have collected and processed your data. This documentation serves as a guide to your dataset and enables others to reuse your data.

Below you will find suggestions for changes that will make your dataset more in line with the DataverseNO guidelines (see the [Deposit Guidelines](#)) and thus increase its value and the chance that it will be found and reused.

To carry out the changes, first navigate to the draft of your dataset (DRAFT) and then click **Edit**, and select **Metadata**, **Files** or **Terms**. After making the changes, click **Save Changes**. Before uploading an edited file, you need to delete the file which needs to be replaced. After uploading the new file, click **Save Changes**. When you have done the necessary changes in your dataset, click **Submit for Review** so I can review the dataset again.

METADATA (see the section "Metadata" in the Deposit Guidelines)	
Citation Metadata	
Title:	Replication_data_for-xxx_2020_xxxxx»
Author – Name:	You should write your name with your family name first: "Kvale, Live"
Author – Identifier:	We recommend adding your ORCID in the <i>Identifier</i> field (e.g. 0000-0001-1234-5678). Using an ORCID ensures that your research results are unambiguously linked to you as a researcher. Learn more about ORCID in this video , and get your own ORCID at http://orcid.org/ .



UNIVERSITY OF OSLO

University of Oslo

(University of Oslo)

DataverseNO >

[Contact](#) [Share](#)



[Advanced Search](#)

☒ [Dataverses \(0\)](#)

☒ [Datasets \(40\)](#)

☐ [Files \(787\)](#)

Publication Year

2023 (19)

2022 (7)

2021 (7)

2020 (5)

2018 (2)

Subject

[Earth and Environmental Sciences \(23\)](#)

[Social Sciences \(8\)](#)

[Medicine, Health and Life Sciences \(6\)](#)

[Arts and Humanities \(5\)](#)

[Physics \(4\)](#)

[More...](#)

Keyword Term

[Svalbard \(11\)](#)

1 to 10 of 40 Results

[Sort](#)

Danish-Norwegian Melusina Folkebøger



Nov 1, 2023

Tidemand, Sverre Christen Ospina, 2023, "Danish-Norwegian Melusina Folkebøger", <https://doi.org/10.18710/8MQOH7>, DataverseNO, V1

This dataset contains data for a study on Melusine-Folkebøger published in Denmark and Norway between 1613 and 1877. The purpose of this dataset is to provide a bibliographical catalogue of the various editions of this text from the two regions. Each entry consists of an "index-c..."

Replication Data for: The myth of the De Geer Zone



Oct 27, 2023

Koehl, Jean-Baptiste P., 2023, "Replication Data for: The myth of the De Geer Zone", <https://doi.org/10.18710/J98MLA>, DataverseNO, V1

High-resolution version of the figures in the following manuscript: "The myth of the De Geer Zone". High-resolution versions of the figures may be necessary to identify individual structures presented and discussed in the text.

Supplementary dataset for "Characterization of Human Stem Cell-Derived Hepatic Stellate Cells and Liver Sinusoidal Endothelial Cells During Extended in vitro Culture"



Sep 26, 2023

Wilhelmsen, Ingrid, 2023, "Supplementary dataset for "Characterization of Human Stem Cell-Derived Hepatic Stellate Cells and Liver Sinusoidal Endothelial Cells During Extended in vitro Culture"", <https://doi.org/10.18710/4MJLC3>, DataverseNO, V1

Data journals

scientific **data**



Data in Brief

Open access

**Research Data Journal
for the Humanities
and Social Sciences**

<https://www.nature.com/sdata/>

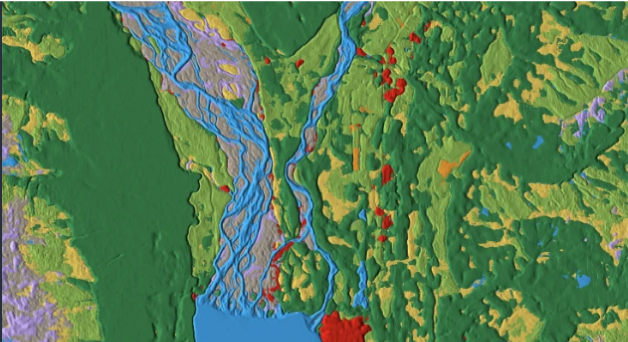
<https://www.sciencedirect.com/journal/data-in-brief>

<https://brill.com/view/journals/rdj/rdj-overview.xml>

[nature](#) > scientific data

Dynamic World, Near real-time global 10m land use land cover mapping

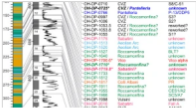
Christopher F. Brown, Steven P. Brumby ... Alexander M. Tait
Data Descriptor | 09 June 2022



Featured

Data Descriptor
[Open Access](#)
02 Sept 2021

[Lake Ohrid's tephrochronological dataset reveals 1.36 Ma of Mediterranean explosive volcanic activity](#)
Niklas Leicher, Blagio Giaccio ... Bernd Wagner



Data Descriptor
[Open Access](#)
12 Aug 2021

[OPERA tau neutrino charged current interactions](#)
N. Agafonova, A. Alexandrov ... C. S. Yoon



Announcements

[Open data in the COVID-19 pandemic](#)

A collection presenting a series of rapidly evolving resources that aggregate and bring cohesion to the massive volume of data being generated in the COVID-19 crisis



Search Scientific Data

All Subjects

Q

[Find out more about Scientific Data](#)



[Find the right repository for your data](#)



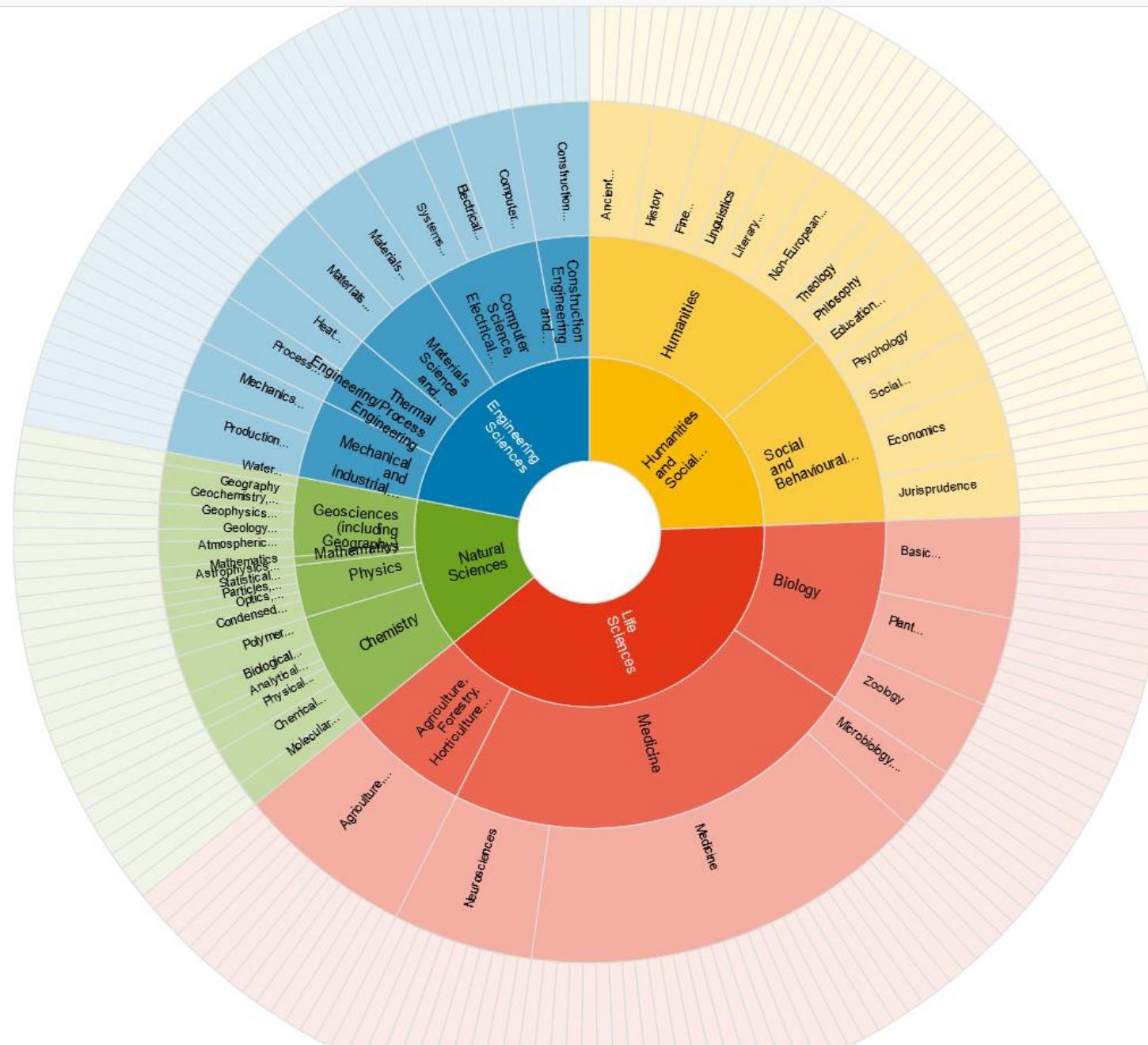
16.11.2023

Archiving of Code



Finding repositories and archives





Finding repositories and archives

Databases

A registry of knowledgebases and repositories of data and other digital assets.



Dr. Irene Salinas Remiro @DrSalinasLab · 27 Jan



Is there a public **repository** to submit 3D microscopy images for **data sharing** purposes? Zebrafish folks, what do you usually do? Thanks!



8



6



20



13.6K



Victoria E. Abraira @VAbraira · 23 Jan



Hey Neuro colleagues: what is your favorite **repository** for **data sharing**?



2



6



2,061



Selecting an archive

TRUSTworthy

- Should the data be openly available?
- Should the metadata be openly available?
- Presence of personal or confidential data can affect choice of archive / preservation solution
- What can the relevant archives offer for long term perspectives?
- Does the archive offer curation – control of metadata and updating of formats?

Trusted repository

The TRUST Principles for digital repositories

TRUSTworthy

TRUSTworthy repositories take responsibility for the stewardship of

- their data holdings and for serving their user community.
- Responsibility is demonstrated by:
 - Adhering to the designated community's metadata and curation standards, along with providing stewardship of the data holdings e.g. technical validation, documentation, quality control, authenticity protection, and long-term persistence.
 - Providing data services e.g. portal and machine interfaces, data download or server-side processing.
 - Managing the intellectual property rights of data producers, the protection of
 - sensitive information resources, and the security of the system and its content.
 -

Trusted repository

From Horizon Europe

TRUSTworthy

- Trusted repositories display specific characteristics of organisational, technical and procedural quality. These include services, mechanisms and/or provisions that are intended to secure the integrity and authenticity of their contents, thus facilitating their use and re-use in the short- and long-term. They have specific provisions in place and offer explicit information online about their policies, which define their services (e.g. acquisition, access, security of content, long-term sustainability of service including funding etc.).

In order to be considered ‘trusted’, a repository has to fall into one of the following categories:

- Certified repositories (e.g. CoreTrustSeal, nestor seal based on DIN 31644, ISO 16363);
- Internationally recognized discipline or domain-specific repositories commonly used and endorsed by the respective research communities;
- All other repositories (including but not restricted to general-purpose repositories and institutional repositories) that fulfil certain criteria as described in the [Horizon Europe Annotated Grant Agreement](#).

Certification



international, community based, non-governmental, and non-profit organization promoting sustainable and trustworthy data infrastructures

As open as possible,
as closed as necessary



EDITORIAL

Open science and sharing personal data widely – legally impossible for Europeans?

Giske Ursin^{a,b,c} and Heidi Beate Bentzen^{a,d} 

^aCancer Registry of Norway, Oslo, Norway; ^bInstitute of Basic Medical Sciences, University of Oslo, Oslo, Norway; ^cDepartment of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA; ^dNorwegian Research Center for Computers and Law, Faculty of Law, University of Oslo, Oslo, Norway

ARTICLE HISTORY Received 15 October 2021; accepted 15 October 2021

A requirement for having a research paper published in many medical journals is that the authors include a data sharing statement. Although the requirement from the International Committee of Medical Journal Editors is not very strict, simply requiring a statement [1], interpretation varies. Some journals essentially require that data must be *readily available* for other researchers for the paper to be accepted.

While most of us eagerly welcome open science and reuse of data to ensure reproducible science, the EU General Data Protection Regulation (GDPR) provides strong protection of privacy and rather restricts and counteracts open sharing of personal data [2]. Some editors will accept that data are not readily sharable with others than peer reviewers for legal reasons. However, editors of non-European journals will often object to a GDPR-compatible data sharing statement and, consequently and often at the last minute, reject the research paper.

Why is this an issue? How difficult is it for European researchers to share data with researchers in other parts of the world?

supplementary measures in place to protect the data. The European research institution will in collaboration with the data importer need to conduct a thorough assessment of the importer's country's laws to ensure that an EU level of data protection is obtained. Such assessments require sound knowledge of the EU Charter of Fundamental Rights, the GDPR, the Court of Justice of the European Union *Schrems II* judgment, and subsequent guidance from the European Data Protection Board, which comprises all Data Protection Authorities in the European Economic Area (EEA) [4]. Finally, the data exporter must be willing to take the risk that the national Data Protection Authority agrees that all requirements have indeed been met, as fines can be high if the institution makes a mistake. Whenever such transfer is possible to achieve, you are lucky! The only cost is that the legal and administrative work on your end has quadrupled compared to the pre-GDPR era.

The federal challenge

Info Tags Related		
Item Type	Journal Article	
Title	Open science and sharing personal data widely – legally impossible for Europeans?	
Author	Ursin, Giske	
Author	Bentzen, Heidi Beate	
Abstract		
Publication	Acta Oncologica	
Volume	60	
Issue	12	
Pages	1555–1556	
Date	2021–12–02	
Series		
Series Title		
Series Text		
Journal Abbr		
Language		
DOI	10.1080/0284186X.2021.1995894	
ISSN	0284–186X	
Short Title		
URL	https://doi.org/10.1080/0284186X.2021....	
Accessed	02/08/2022, 14:30:18	
Archive		
Loc. in Archive		
Library Catalogue	Taylor and Francis+NEJM	
Call Number		
Rights		
Extra	Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0284186X.2021.1995894 PMID: 34797207	
Date Added	02/08/2022, 14:30:18	
Modified	02/08/2022, 14:30:18	

Persistent identifiers (PIDs)



<http://urn.nb.no/URN:NBN:no-5678>

ORCID

ROR

DOI – (digital object identifier) Commonly used for datasets and publications
URN – Persistent identifiers used in DUO for theses and self-archived publications.
ORCID – identifies the researcher
ROR - Identifies the University

License makes data reusable

Preparing for archiving

- Do you have permission to share the data?
- Consistent, meaningful, and compatible file naming
- Choose accessible, patent-free, and open file formats
- Write a readme file where you describe the data.

Prepare your data

Before depositing your data in DataverseNO (including the different collections, e.g. UiT Open Research Data, TROLLing, etc.) you have to make sure your dataset(s) comply with our guidelines below. DataverseNO accepts only research data in digital formats. In brief, good practice for preparing research data for archiving may be summarized as follows:

- Use consistent and comprehensible file names (see section 1 below).
- Save your data in a preferred file format(s) (see section 2 below).
- Describe your data in a ReadMe file (see section 3 below).

For more detailed guidelines, see below:

- ▼ **1 File naming and organization**
- ▼ **2 Preferred file formats**
- ^ **3 How to describe your data**



In order for other researchers to be able to understand and reuse your data, it is essential that you describe them in a comprehensible and consistent manner before they are published. In DataverseNO, this kind of documentation must be provided in two ways, in the **metadata fields**, and in a separate **ReadMe file** which **must** be uploaded together with your data files:

▼ **Metadata**

^ **ReadMe file**



A **ReadMe file** is a more detailed user guide to your dataset so that other researchers are able to interpret, understand, and reuse your data, including information about how the dataset was created, how complete it is, and what kind of restrictions it has. The ReadMe file must minimally contain the following:

- Title of the dataset, DOI, contact information
- Methods
- Data and file overview
- Data-specific information
- Terms of Reuse



The Turing Way

Search this book...

Welcome

Guide for Reproducible Research

Overview

Open Research

Version Control

Licensing

Research Data Management

Research Data

Data Management Plan

The FAIR Principles and Practices

Personal data management

Data Storage and Organisation

Data Organisation in Spreadsheets

Documentation and Metadata

Data Curation

Sharing and Archiving Data

Data Article

Research Data Management Toolkits

Personal Impact Stories

Checklist

Resources

Reproducible Environments

BinderHub

Code quality

Code Testing

Code Reviewing Process

Reusable Code

Continuous Integration (CI)

Data Organisation in Spreadsheets

Spreadsheets, such as Microsoft Excel files, google sheets, and their Open Source alternative ([for instance](#)) [LibreOffice](#), are commonly used to collect, store, manipulate, analyse, and share research data. Spreadsheets are convenient and easy-to-use tools for organising information into an easy to write and easy to read forms for humans. However, one should use them with caution, as the use of an inappropriate spreadsheet is a major cause of mistakes in the data analysis workflow. There is a collection of [horror-stories](#) that tells how the use of spreadsheets can ruin analysis-based studies due to unexpected behaviour of the spreadsheet or error-prone editing processes. Some of these mishaps are not unique to spreadsheets, but many, such as [this](#) and [this](#), are.

Fortunately, most problems can be avoided with the following recommendations:

- Use spreadsheet in a text-only format (.csv or .tsv),
- Create tidy spreadsheets,
- Make spreadsheets consistent (with each other) and implement rules for data entries, and
- Avoid manipulating and analysing data in spreadsheet software (this includes copy-paste).

Spreadsheets are a powerful tool only if the dataset is collected and organised in specific formats that are usable for both the computers and researchers.

1. Avoid Non-Data Content

Spreadsheets are used for organising data in a tabular form. The subject, the object and the relationship between them are transformed into rows, cells and columns, respectively. For example, the subject: **experiment**, relationship: **was performed on the date**, and the object: **2020-06-06** gives one row for each experiment, one column for **date of experiment**, and the value **2020-06-06** in the cell. Unfortunately, spreadsheet programs allow you to add other kinds of contents to this, like color to specific cells. While it may help the researchers at some point, one needs to remember that this kind of **cell modification should not be considered as data**, primarily because they cannot be exported to other software.

As a simple rule, what can be exported in a text-only format, comma-separated values (CSV), or tab-separated values (TSV), can be considered as the data. Other functions should be avoided when using these programs for research data. This includes:

- changing font, color or borders,
- using functions,
- merging cells (this one is particularly problematic),
- using specific cell formats (especially dates, see below).

As a test for your spreadsheet compatibility with reproducible research, export your data from the spreadsheet to the CSV format and reopen it. If you can still get all the information that you stored in your sheet, then your data is fine.

Tip

If you want to use color to help with a rapid highlight in your document, create a new column to indicate which cells are highlighted (it becomes a part of your data). In addition to the visual feedback, you can now also use this information to filter or sort your data and get the highlighted cells quickly.

Contents

1. Avoid Non-Data Content

2. Tidy Format For Spreadsheets

3. Consistent Values

4. Data Manipulation and Analysis

Other Tips

Summary



Text documents

• Preferred format(s)

- PDF/A (.pdf)
- ODT (.odt)

• Non-preferred format(s)

- Microsoft Word (.doc)
- Office Open XML (.docx)
- Rich Text File (.rtf)
- PDF other than PDF/A (.pdf)

Plain text

- Unicode text (.txt)

- Non-Unicode text (.txt)

Markup language

- XML (.xml)
- HTML (.html)
- Related files: .css, .xslt, .js, .es

- SGML (.sgml)
- Markdown (.md)

Programming languages

- MATLAB
- NetCDF
- TextFabric

Spreadsheets

- ODS (.ods)
- CSV (.csv)

- Microsoft Excel (.xls)
- Office Open XML Workbook (.xlsx)
- PDF/A (.pdf)

Databases

- SQL (.sql)
- SIARD (.siard)
- CSV (.csv)

- Microsoft Access (.mdb, .accdb)
- dBase (.dbf)
- HDF5 (.hdf5, .he5, .h5)

Statistical data

- SPSS (.dat/.sps)
- STATA (.dat/.DO)
- R

- SPSS Portable (.por)
- SPSS (.sav)
- STATA (.dta)
- SAS (.7dat; .sd2; .tpt)

part 2

examples and menti

Key takeaways:

Contact us at:

research-data@uio.no

1. If you put your data in an archive, you get a **DOI**
2. Not all archives offer **curation** – if the data have a long-term value, choose an archive with curation
3. Add a **license** to your dataset and code so that others know what they can and cannot do with the data
4. If you share **code**, archive a frozen, citeable version (e.g. via GitHub - Zenodo connection)
5. Without metadata, documentation and structure, your data are not reusable

Links:

FAIR: <https://www.force11.org/fairprinciples>

Selecting an archive: <https://www.ub.uio.no/english/writing-publishing/data-archiving/selecting-archive.html>

DataverseNO: <https://dataverse.no/dataverse/uio>

NSD: <https://www.nsd.no/en/archiving-research-data/>

NIRD: <https://www.sigma2.no/research-data-archive>

Zenodo: <https://zenodo.org>

Figshare: <https://figshare.com>

OSF <https://osf.io>

Prepare data: <https://site.uit.no/dataverseno/deposit/prepare/>

Github citable code: <https://guides.github.com/activities/citable-code/>

Re3data:

Creative Commons Licenses: <https://creativecommons.org>

CoreTrustSeal:

UiO Research data management <https://www.uio.no/english/for-employees/support/research/research-data-management/>