

UNIVERSITY OF OSLO

Sharing and archiving research data

Live H. Kvale and Agata Bochynska

University of Oslo Library

26.10.2021



Go to www.menti.com and use the
code 4082 4837

The Availability of Research Data Declines Rapidly with Article Age

Timothy H. Vines,^{1,2,*} Arianne Y.K. Albert,³ Rose L. Andrew,¹ Florence Débarre,^{1,4} Dan G. Bock,¹ Michelle T. Franklin,^{1,5} Kimberly J. Gilbert,¹ Jean-Sébastien Moore,^{1,6} Sébastien Renaut,¹ and Diana J. Rennison¹

¹Biodiversity Research Centre, University of British Columbia, 6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada

²Molecular Ecology Editorial Office, 6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada

³Women's Health Research Institute, 4500 Oak Street, Vancouver, BC V6H 3N1, Canada

⁴Centre for Ecology & Conservation Biosciences, University of Exeter, Cornwall Campus, Tremough, Penryn TR10 9EZ, UK

⁵Institute for Sustainable Horticulture, Kwantlen Polytechnic University, 12666 72nd Avenue, Surrey, BC V3W 2M8, Canada

⁶Department of Biology, Université Laval, 1030 Avenue de la Médecine, Laval, QC G1V 0A6, Canada

Summary

Policies ensuring that research data are available on public archives are increasingly being implemented at the government [1], funding agency [2–4], and journal [5, 6] level. These policies are predicated on the idea that authors are poor stewards of their data, particularly over the long term [7], and indeed many studies have found that authors are often unable or unwilling to share their data [8–11]. However, there are no systematic estimates of how the availability of research data changes with time since publication. We therefore requested data sets from a relatively homogenous set of 516 articles published between 2 and 22 years ago, and found that availability of the data was strongly affected by

sets (23%) were confirmed as extant. [Table 1](#) provides a breakdown of the data by year.

We used logistic regression to formally investigate the relationships between the age of the paper and (1) the probability that at least one e-mail appeared to work (i.e., did not generate an error message), (2) the conditional probability of a response given that at least one e-mail appeared to work, (3) the conditional probability of getting a response that indicated the status of the data (data lost, data exist but unwilling to share, or data shared) given that a response was received, and, finally, (4) the conditional probability that the data were extant (either “shared” or “exists but unwilling to share”) given that an informative response was received.

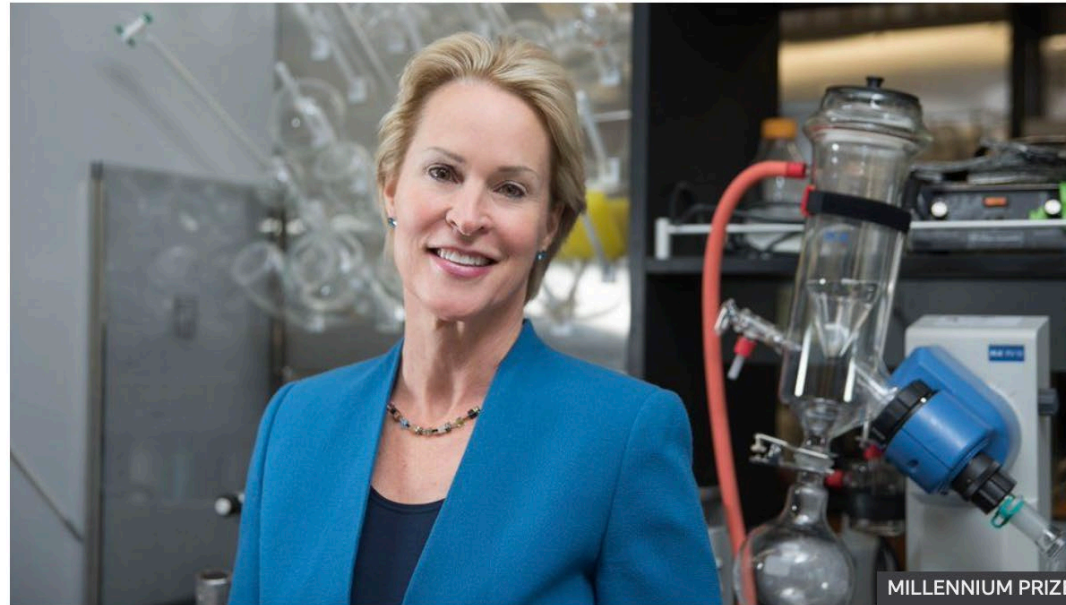
There was a negative relationship between the age of the paper and the probability of finding at least one apparently working e-mail either in the paper or by searching online (odds ratio [OR] = 0.93 [0.90–0.96, 95% confidence interval (CI)], $p < 0.00001$). The odds ratio suggests that for every year since publication, the odds of finding at least one apparently working e-mail decreased by 7% ([Figure 1A](#)). Since we searched for e-mails in both the paper and online, four factors contribute to the probability of finding a working e-mail: (1) the number of e-mails in the paper and (2) the chance that any of those worked and (3) the number of e-mails we could find by searching online and (4) the chance that any of those worked. The total number of e-mail addresses we found in the paper decreased with age (Poisson regression coefficient = -0.07 , $SE = 0.01$, $p < 0.0001$) from an average of 1.17 in 2011 to 0.42 in 1991 ([Figure 2A](#)), and there was a slight positive effect of article age on the number of e-mails we found online (Poisson regression coefficient = 0.015 , $SE = 0.007$, $p < 0.05$; [Figure 2C](#)). Moreover, the chance that an e-mail found in the

Nobel Prize-winning scientist Frances Arnold retracts paper

🕒 3 January 2020



Nobel Prize



Prof Arnold works in the department of chemical engineering at Caltech

American scientist Frances Arnold, who won the Nobel Prize for chemistry, has retracted her latest paper.

Prof Arnold shared the award with George P Smith and Gregory Winter for their research on enzymes in 2018.

A subsequent paper on enzymatic synthesis of beta-lactams was published in the journal Science in May 2019.

It has been retracted because the results were not reproducible, and the authors found data missing from a lab notebook.

Reproduction is an essential part of validating scientific experiments. If an experiment is a success, one would expect to get the same results every time it was conducted.

Go to www.menti.com and use the
code 4082 4837

What types of data do you work with?



Reasons for sharing your data

Career Benefits:

- Increased visibility
- More reuse
- Increased citations

External Factors:

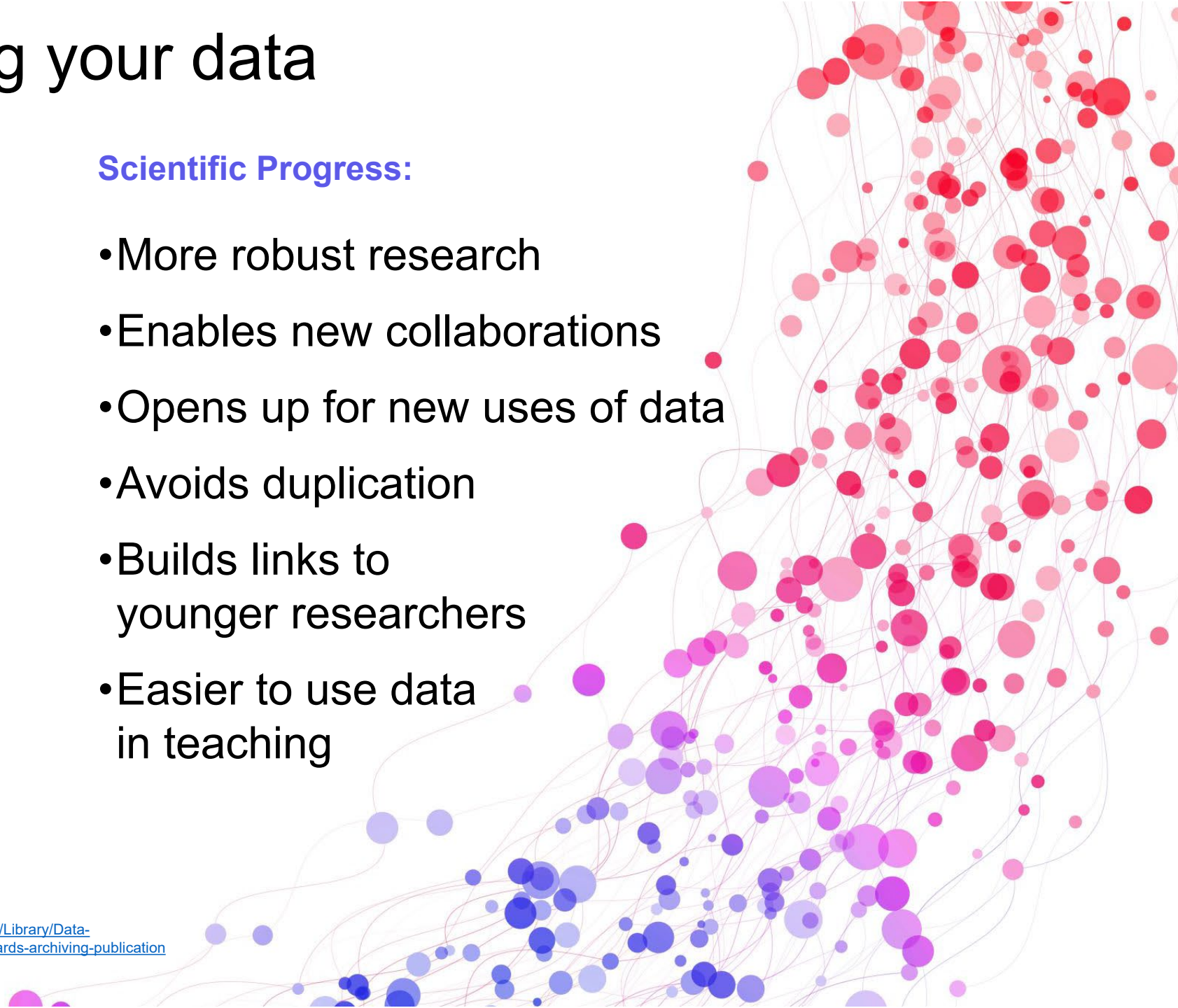
- Funder requirements
- Publisher requirements

Norms:

«This is how we do it here»

Scientific Progress:

- More robust research
- Enables new collaborations
- Opens up for new uses of data
- Avoids duplication
- Builds links to younger researchers
- Easier to use data in teaching



Go to www.menti.com and use the
code 4082 4837

Why would you like to share your data?



Why would you not share your data?

conflicts of interest
lots of work reuse of not open data
commercial value
afraid of unjust use privacy rights
problem to anonymise data
sensitive material legal issues
safeguarding participants
lack of infrastructure
copyrighted material

Open science

Open science means transparency and knowledge-sharing in research processes to make knowledge accessible across academic groups, sectors and national boundaries. The concept of open science encompasses the entire research process [...].



Why open science?

improved **quality** of research by **building** on previous works and compiling research data in new ways

transparency in the research process and better opportunity for **verifiability** of scientific results

increased **cooperation** and less duplication of research work

increased **innovation** in the private and public sectors

efficiency improvement and better utilisation of public funds

“My main argument for opening all parts of the process is that is “sharpening” the research process. You cannot be sloppy if you know that it will be exposed”.

Alexander R. Jensenius, 2020

Go to www.menti.com and use the
code 4082 4837

What are the advantages of open science?

cost reduction of science

More people have access to research

easier to spot errors

Transparency

sharing is caring

increased collaboration

increased knowledge on a specific topic

Science moves forward

Build on previous data ("standing on the shoulders of giants")

What are the advantages of open science?

discoveries can be made faster

More fair to 3rd world researchers

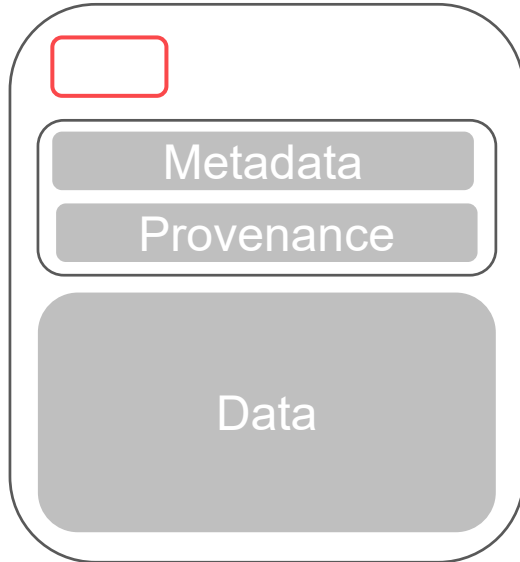
researchers with less access to funds are included

RESEARCH DATA - OPEN BY DEFAULT

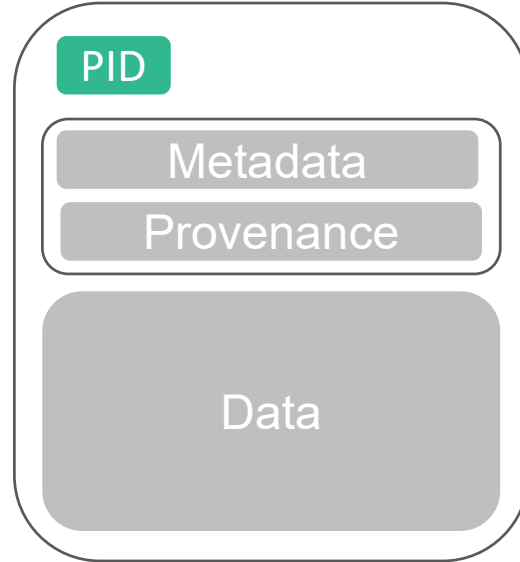


Levels of FAIR

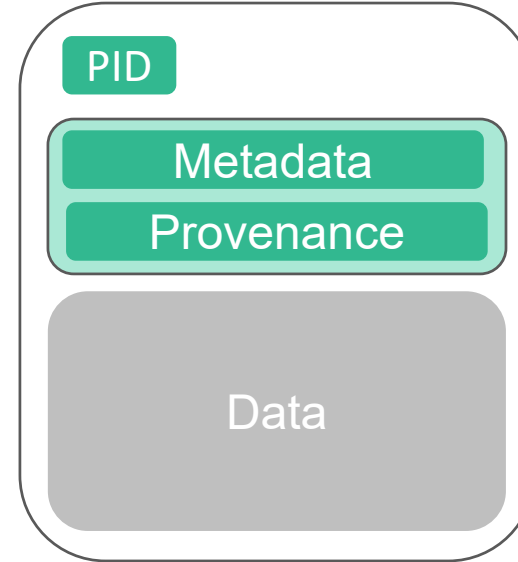
Totally UNFAIR



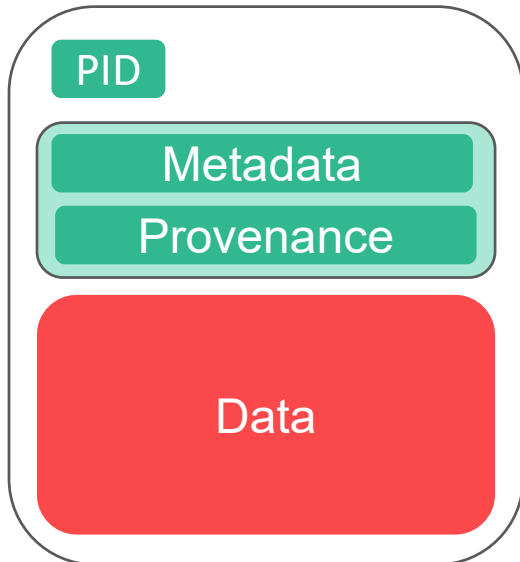
Findable
Usable for humans



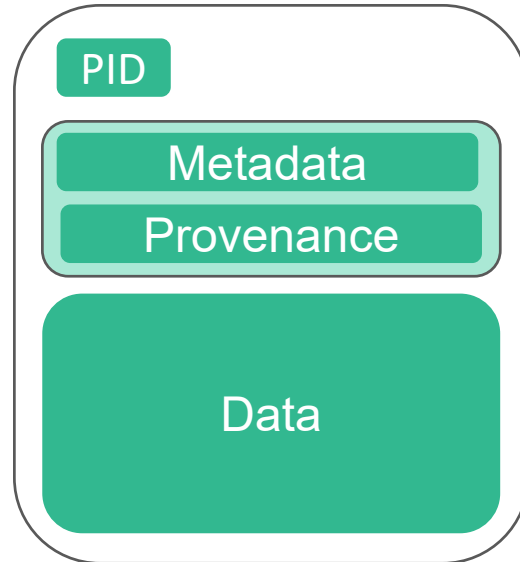
FAIR metadata



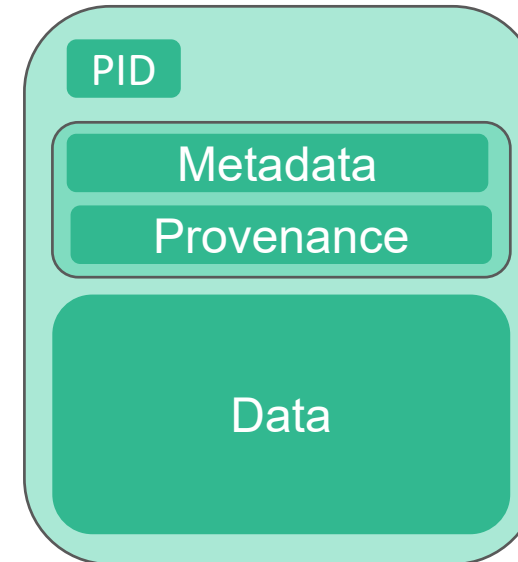
FAIR data
Restricted access



FAIR data
Open access



FAIR data
Open access and functionally linked

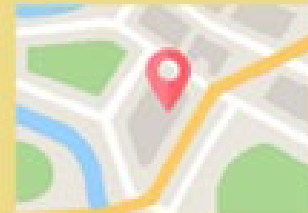


Data



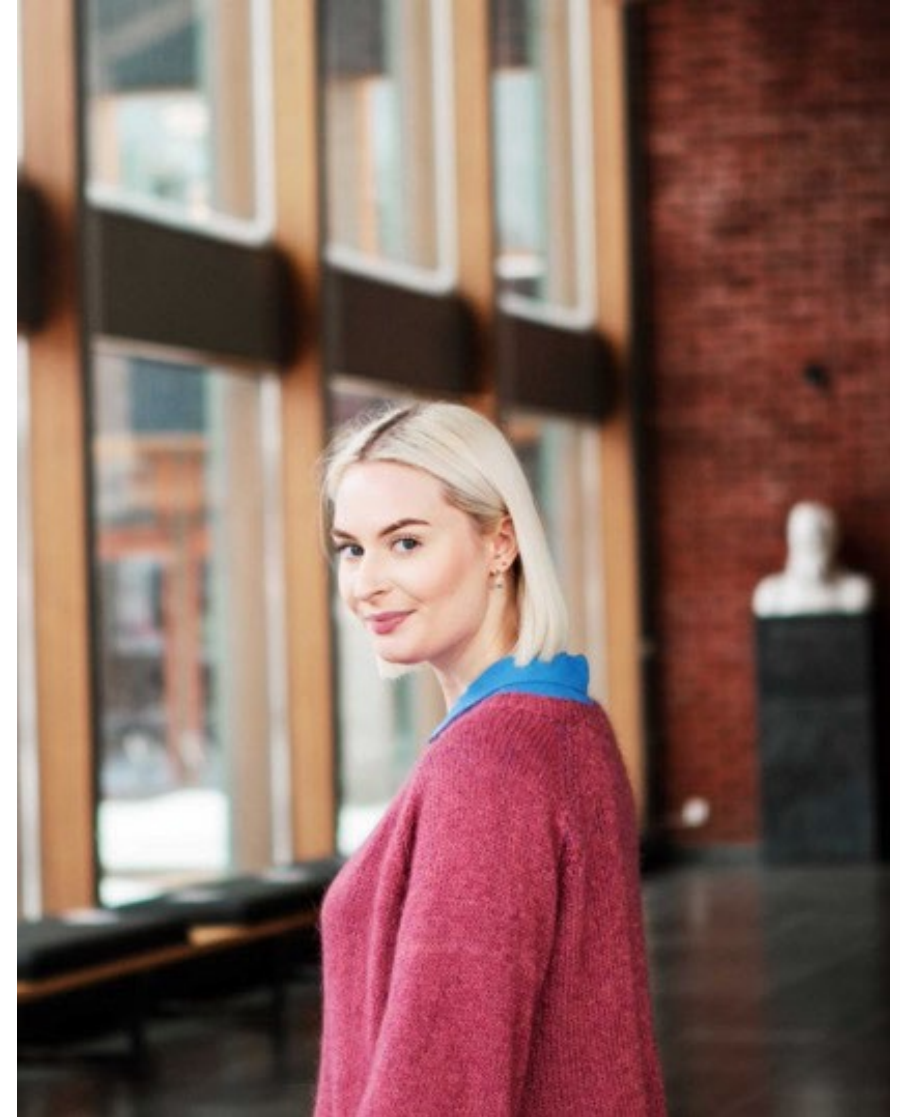
Metadata

Filename:
Tadzik.jpg
Author:
Piotr Kononow
Date:
August 15, 2016



Descriptive metadata in a research data context

- Title (on the project and files)
- Author (creator, copyright holder)
- Publication year (and year(s) of data collection)
- Persistent identifier (DOI)
- Location (preferably coordinates)
- Which publication(s) the datasets are used in
- And so on...



Other metadata

- Column units in a table
- Configurations and specifications of the instruments used in the data collection
- Software-specific information

NCBI GEO Accession Display

Scope: Self Format: HTML Amount: Quick GEO accession: GSM2475810

Sample GSM2475810 Query DataSets for GSM2475810

Status	Public on Aug 07, 2017
Title	OK Illumina RNASeq
Sample type	SRA
Source name	OK (ATCC CRL-1840)
Organism	<i>Didelphis virginiana</i>
Characteristics	cell line: OK cell type: Proximal tubule epithelial passages: 31 tissue: Kidney Proximal Tubule
Growth protocol	Cells were cultured in DMEM/F12 medium with 10% FBS (Atlanta Biologicals) and 5 mM GlutaMAX (Gibco). 4x10 ⁵ cells were plated on 12-mm Transwells with 0.4 µm pore polycarbonate membrane inserts (Corning) in a 12-well plate, with 0.5 mL apical medium and 1.0 mL basolateral medium.
Extracted molecule	total RNA
Extraction protocol	Cells were collected using Accutase (Sigma) and RNA was extracted using the ambion PureLink RNA mini kit (ThermoFisher) according to the manufacturer's protocol. Ribosomal RNA sequences were removed before library construction. Library preparation was performed using the TruSeq Stranded Total RNA Sample Preparation Kit (Illumina) according to manufacturer's instructions. RNA was fragmented for 8 minutes and reverse transcription performed. Double stranded cDNA was subjected to 3' adenylation and ligation of sequencing adapters. Sequencing was carried out on a NextSeq 500 (Illumina) for 2 x 75 bp paired end reads. Loading concentration was 1.6 pM.
Library strategy	RNA-Seq
Library source	transcriptomic
Library selection	cDNA
Instrument model	Illumina NextSeq 500
Data processing	Sequence reads were trimmed using Cutadapt v 1.8.3 (adapter and qual 30) 60 million RNASeq reads were randomly selected Selected reads were assembled with de novo Trinity v 2.1.1 at minimum kmer coverage 3 Assembled contigs were aligned with protein sequences in Swiss-Prot db, using BLASTx v2.2.31, for annotation. Alignments were filtered for hits to proteins from mammals, e-value < 1e-4, 70% Identity and 40% covered bases of database protein Expression values of filtered genes and transcripts were calculated using RSEM v 1.2.21 Selected 60 million reads were aligned to reference using Tophat v 2.1.0 Mapped reads were assembled using Cufflinks and Cuffmerge v 2.2.1 Sequences of Cuffmerge were assembled transcripts extracted using gffread tool of Cufflinks software suite Expression values of assembled genes and transcripts were calculated using

Data Type
SRA = Sequence Read Archive = high-throughput sequencing data

Subject of Study
(organism, cell type, & tissue)

Equipment

Software
(multiple used for analysis)

Selecting data for archiving

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



Preparing for archiving

- Determine scientific relevance and need for archiving long-term
- Consistent, meaningful, and compatible file naming
- Choose accessible, patent-free, and open file formats
- Make sure you have the necessary documentation (and metadata)
- Reduce complexity by grouping large groups of similar files in zip bundles to make upload and download easier
- Presence of personal or confidential data can affect choice of archive
- Consider size limitations when choosing an archive (e.g. some archives have a limit of 10 to 50 GB per dataset)

You are here: [Home](#) > [About DANS](#) > [Services](#) > [EASY](#) > [Information about depositing data](#) > [Before depositing](#) > File formats

Services

DataverseNL

EASY

Information about depositing data

Before depositing

Archaeological exchange
protocol

File formats

Text documents

Plain text

Markup language

Spreadsheets

Databases

Statistical data

Programming languages

Images (raster)

Images (vector)

Audio

Video

Computer Aided Design
(CAD)

Geographical information
(GIS)

Images (georeference)

Raster grid

3D

RDF

Computer Assisted
Qualitative Data Analysis
(CAQDAS)

Abbreviations and
acronyms

During depositing

After depositing

About reusing data

NARCIS

Training and consultancy

Organisation and policy

Research and Innovation

File formats

Preferred formats are file formats of which DANS is confident that they will offer the best long-term guarantees in terms of usability, accessibility and sustainability. Deposits of research data in preferred formats will always be accepted by DANS.

Non-preferred formats are file formats that are widely used in addition to the preferred formats, and which will be moderately to reasonably usable, accessible and robust in the long term.

Preferred formats will be automatically accepted by DANS.

As a general guideline, DANS believes that the file formats best suited for long-term sustainability and accessibility:

- Are frequently used
- Have open specifications
- Are independent of specific software, developers or vendors

In practice, it is not always possible to use formats which satisfy all of these criteria.

It may be desirable to make certain original data available in 'Non-preferred format(s)' because these can be characterized as current usage formats. Examples include Esri Shapefiles, Microsoft Access databases, SPSS .sav files. DANS then asks you to deposit your data in these original formats as well as in Preferred formats aimed at long-term sustainability.

If your data are stored in other formats than those mentioned below, please contact DANS at info@dans.knaw.nl.

Type	Preferred format(s)	Non-preferred format(s)
Text documents	• PDF/A (.pdf) • ODT (.odt)	• 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) • HDFS (.hdf5, .he5, .h5)
Statistical data	• SPSS (.dat/.sps) • STATA (.dat/.DO) • R	• SPSS Portable (.por) • SPSS (.sav) • STATA (.dta) • SAS (.7dat; .sd2; .tpt)
Raster images	• JPEG (.jpg, .jpeg) • TIFF (.tif, .tiff) • PNG (.png) • JPEG 2000 (.jp2) • DICOM (.dcm)	
Vector images	• SVG (.svg)	• Adobe Illustrator (.ai)

Licensing

- A license agreement is a legal arrangement between the creator/depositor of the data set and the data repository, signifying what a user is allowed to do with the data.
- Often Creative Commons licenses are often used, more licenses can be found at: <https://spdx.org/licenses/>



Recommendations regarding licensing and Dedication to Public Domain

1. A licence or Dedication to Public Domain, should reflect conditions further use of the data.
2. Research data should be assigned with an standard international license which:
 - Have as few access and use restrictions possible
 - to the largest extend promote principles of juridical interoperability for research data.
 - Is human and machine readable
3. Metadata should always have a Dedication to Public Domain assigned
4. Standard licenses for copyright should only be used for material filling the condition for copyright or database protection
5. In situation were different juridical conditions are applicable to different parts of a dataset, one should differentiate and license the different parts accordingly. This to avoid that data are assigned to a more restrictive licenses because some of the data require this.
6. The institution have a responsibility to ensure that data produces by employees trough their work are equipped with a licence or Dedication to Public Domain described in 1.
7. Reuse of data from others should always happen according to research ethical norms and subject specific standards.



- Credit must be given to the creator
- Only noncommercial uses of the work are permitted



- Credit must be given to the creator
- Adaptations must be shared under the same terms



- Credit must be given to the creator



- is a public dedication tool, which allows creators to give up their copyright and put their works into the worldwide public domain.

PDM – public domain mark – no known copyright to the material

<https://youtu.be/PgqtiY7oZ6k>

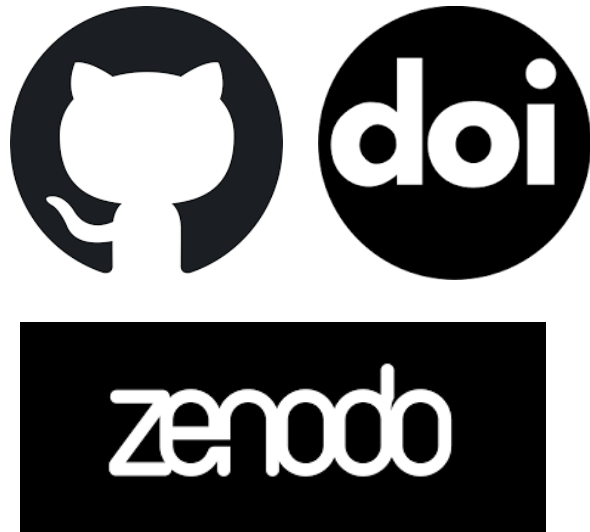
Reusable code

Licenses for code:

MIT (Expat)

Apache 2.0

GNU (L)GPL 3.0+



Making Your Code Citable

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

Social coding and open software

<https://github.com/coderefinery/social-coding/blob/main/talk.md>

Selecting an archive

- Should the data be openly available?
- Should the metadata be openly available?
- What can the relevant archives offer for long term perspectives?
- Does the archive offer curation – control of metadata and updating of formats?

Types of data archives

- Domain-specific
- General purpose
- Institutional
- Supplementary material to an article
- Data paper

Different types of archives

- NSD
- Dataverse.no
- Zenodo
- Spesialsamlinger uio
- Figshare usn
- <https://brill.com/view/journals/rdj/rdj-overview.xml>
- https://springernature.figshare.com/articles/dataset/Metadata_record_for_A_global_occurrence_database_of_the_Atlantic_blue_crab_Callinectes_sapidus/14077112
- <https://www.re3data.org>

The background of the slide is decorated with a pattern of small, semi-transparent dots in shades of yellow, orange, and light pink, arranged in a somewhat random but dense manner.

Break

After the break we will show two cases using menti

Go to www.menti.com and use the
code 4082 4837

Case 1

A shares movement



A investigates humans interacting with music

- Professional musicians participating with self-compound music
- Data collection from multiple sources
- Audience informed that the concerts are also research projects
- A wishes to do science as open as possible



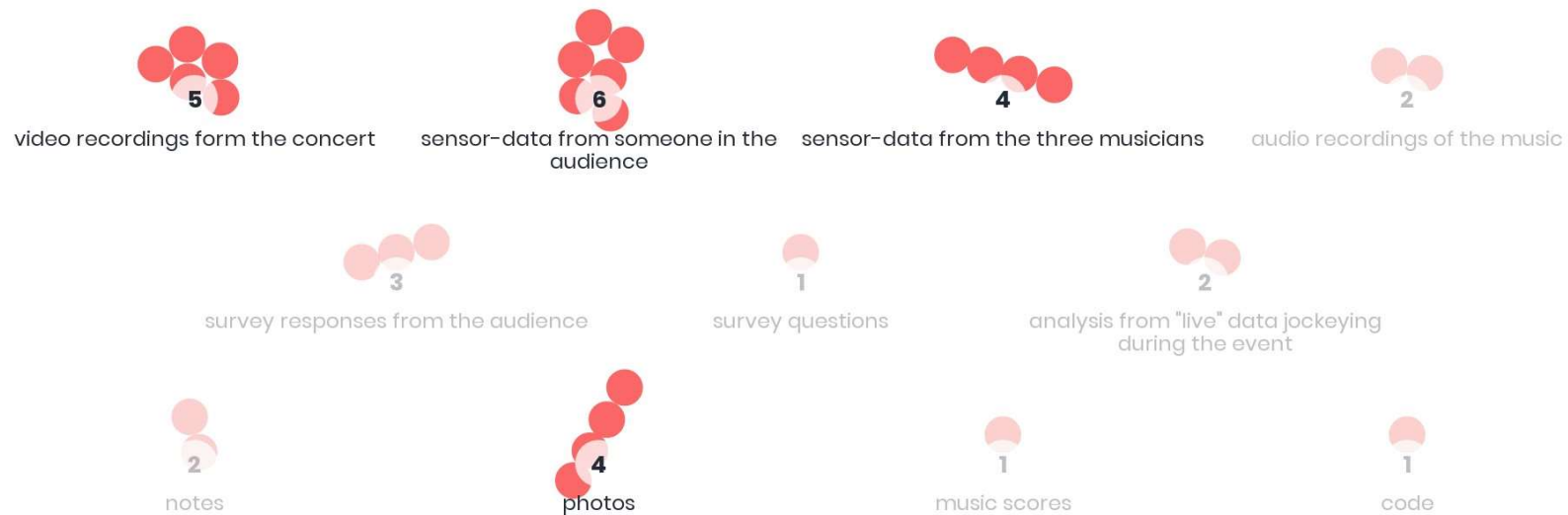
Data from A's project consists of:

- video recordings from the concert
- sensor-data from someone in the audience
- sensor-data from three musicians
- audio recordings of the music
- survey responses from the audience
- survey questions
- analysis from "live" data jockeying during the event
- notes
- photos
- music scores
- code

What legal challenges might A encounter?

copyright
personal data
gdpr

Which data types are likely to be sensitive/special categories?



The approach chosen by A

<https://www.uio.no/ritmo/english/news-and-events/events/musiclab/2019/utopia/index.html>



Case 2

Ø does not want to share



Ø conducted interviews with adolescents

- Ø is a specialist in psychiatry and a researcher at U
- The adolescents are in a vulnerable situation
- The interviews are unique and of huge value for both research and as historical documents
- Because it is difficult to talk about the trauma, other researchers want to reuse the data and not interview the adolescents again
- Ø claims he has a unique right to the material, and does not want to share it with anyone



What ethical challenges occur in this situation?

anonymisation

data are not reusable

Consent



 Saken er produsert og finansiert av De nasjonale forskningsetiske komiteene - [Les mer](#)

22. juli-forskere kan bli overkjørt av kunnskapsdeling

Terrorofrene fra 2011 skal ikke belastes unødig av forskere. Men å måtte dele dataene som samles inn, mener noen forskere er overkjørende.

Siw Ellen Jakobsen, frilansjournalist

De nasjonale forskningsetiske komiteene



16.6 2013 05:00

ANNONSE

A thorough ethical evaluation conclude that the data should be shared with other researchers



Regarding reuse and sharing of data



1) The coordination group recommends that in future research common data platforms should be created and used across subjects and institutions.

2) It also recommends that the metadata for July 22nd research from different studies is gathered at NSD, and that a portal for 22nd of July research is created through CRISTin. This can be done within the research environment at NSD, with some support to cover the cost of metadata creation.

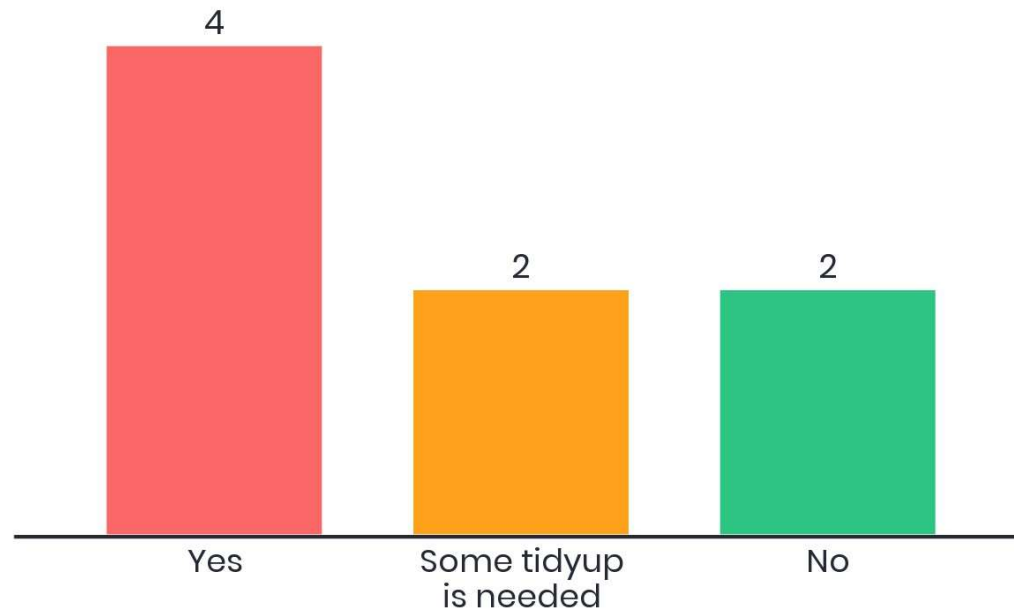


Research data on July 22nd events are today archived at NSD

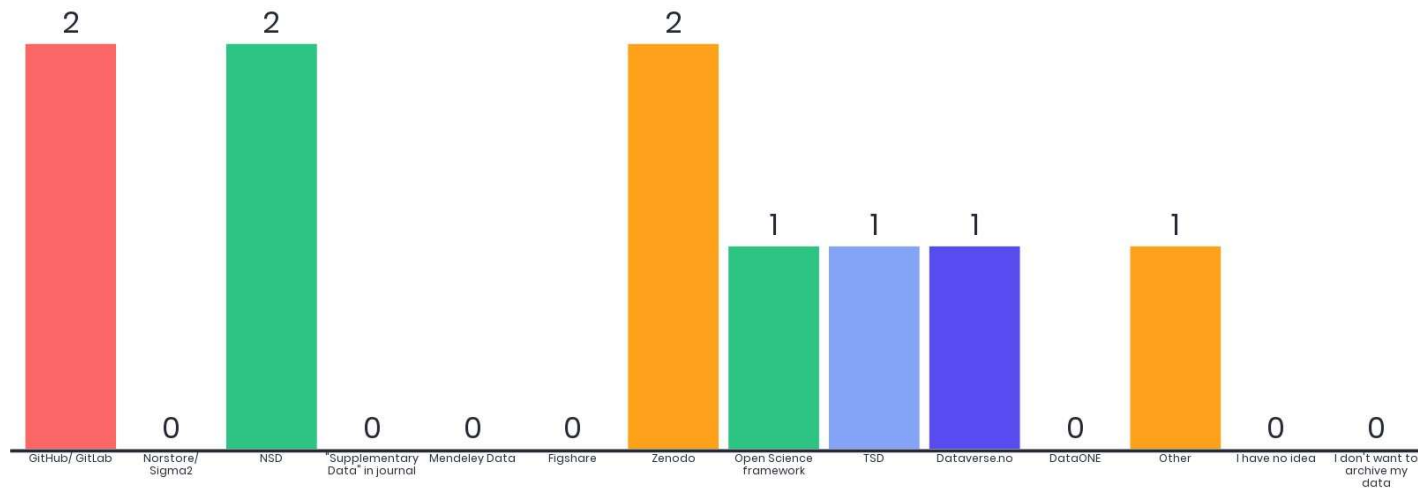
<https://nsd.no/nsddata/22juli/datasett.html?a=/nsddata/22juli/datasett/datasett0009.html>



Are you ready to share your data?



Where do you plan to archive your research data?



Ask us



2 questions
2 upvotes



Thank you for your attention

Contact research-data@uio.no for
questions

Agata Bochynska, Ivana Malovic,
Solveig Sørbø, Live H. Kvale



Sources

- Bjerknes Centre. Bjerknes Climate Data Centre. <https://www.bcdc.no/>
- Center for Open Science. Open Science Framework. <https://osf.io/>
- Creative Commons. <https://creativecommons.org/>
- DataverseNO. <https://dataverse.no/>
- DataverseNO. Prepare your data. <https://site.uit.no/dataverseno/deposit/prepare/>
- Mons, Barend et al. "Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud." Inf. Services and Use 37 (2017): 49–56. <https://doi.org/10.3233/ISU-170824>
- Re3data. <https://www.re3data.org/>
- Sigma2. NIRD research data archive. <https://archive.sigma2.no/>
- Vines, Timothy H. et al. «The Availability of Research Data Declines Rapidly with Article Age." Current biology 24 (2014): 94–97. <http://dx.doi.org/10.1016/j.cub.2013.11.014>
- Zenodo. <https://zenodo.org/>
- UB digital <https://bibsys-almaprimo.hosted.exlibrisgroup.com/primo-explore/collectionDiscovery?vid=UIO>
- Github citable code: <https://guides.github.com/activities/citable-code/>
- Licenses for code: <https://github.com/coderefinery/social-coding/blob/main/talk.md>
- Photos from Pixabay if not otherwise indicated. Scientist: RAEng_Publications.
- 22nd of July research: www.forskningsetikk.no/globalassets/dokumenter/x22.-juli-forskning/sluttrapport-koordineringsgruppa-for-22.-juli-forsking.pdf
- The Practice of Reproducible Research: <http://www.practicereproducibleresearch.org>

