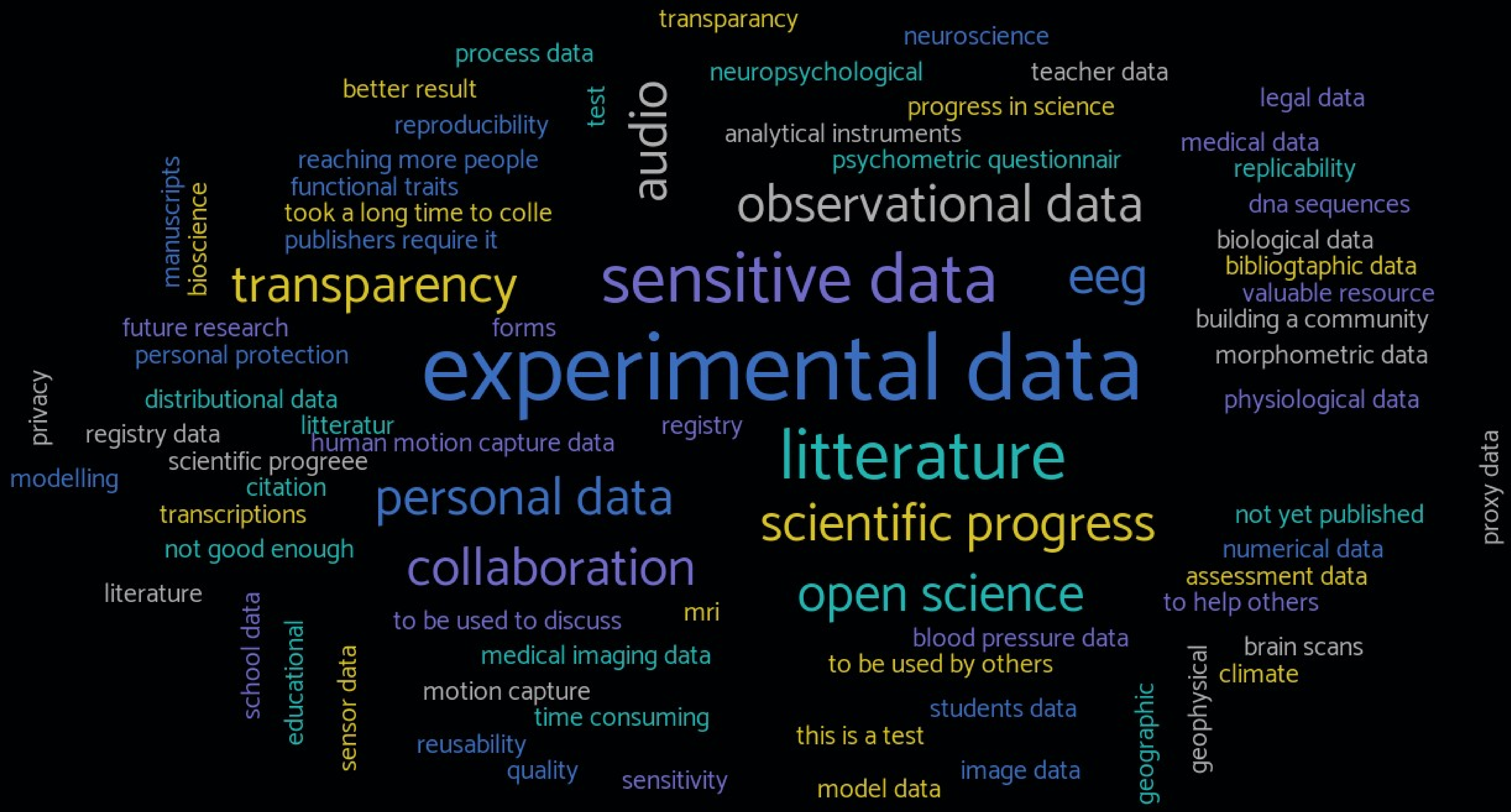
Sharing and archiving research data

Live Håndlykken Kvale and Agata Bochynska

May 20, 2021

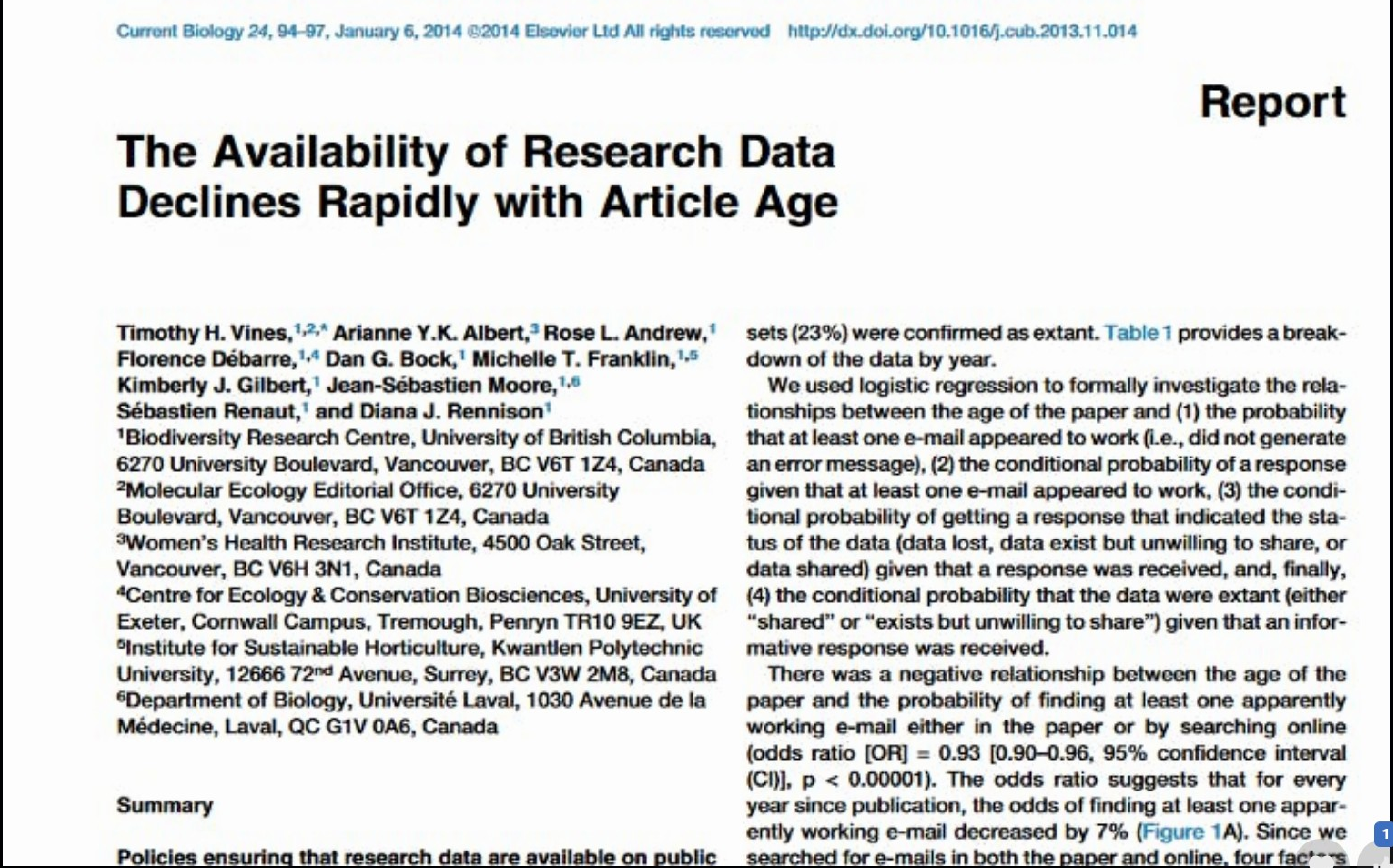
9:00 AM–10:30 AM, Zoom

What types of data do you work with?



transparancy
neuroscience
process data
neuropsychological
teacher data
better result
legal data
progress in science
reproducibility
test
analytical instruments
audio
medical data
reaching more people
psychometric questionnair
replicability
functional traits
manuscripts
bioscience
took a long time to colle
observational data
dna sequences
publishers require it
biological data
bibliogtaphic data
transparency
sensitive data
eeg
valuable resource
building a community
future research
forms
personal protection
experimental data
morphometric data
distributional data
physiological data
privacy
registry
litteratur
registry data
human motion capture data
scientific progreee
litterature
modelling
citation
personal data
proxy data
transcriptions
not yet published
scientific progress
not good enough
numerical data
assessment data
collaboration
literature
open science
to help others
to be used to discuss
mri
school data
blood pressure data
brain scans
medical imaging data
to be used by others
climate
educational
sensor data
motion capture
geophysical
students data
time consuming
geographic
this is a test
reusability
quality
image data
sensitivity
model data

58



Current Biology 24, 94–97, January 6, 2014 ©2014 Elsevier Ltd All rights reserved http://dx.doi.org/10.1016/j.cub.2013.11.014

Report

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. Rennison[1]
[1]Biodiversity Research Centre, University of British Columbia, 6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada
[2]Molecular Ecology Editorial Office, 6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada
[3]Women's Health Research Institute, 4500 Oak Street, Vancouver, BC V6H 3N1, Canada
[4]Centre for Ecology & Conservation Biosciences, University of Exeter, Cornwall Campus, Tremough, Penryn TR10 9EZ, UK
[5]Institute for Sustainable Horticulture, Kwantlen Polytechnic University, 12666 72nd Avenue, Surrey, BC V3W 2M8, Canada
[6]Department of Biology, Université Laval, 1030 Avenue de la Médecine, Laval, QC G1V 0A6, Canada

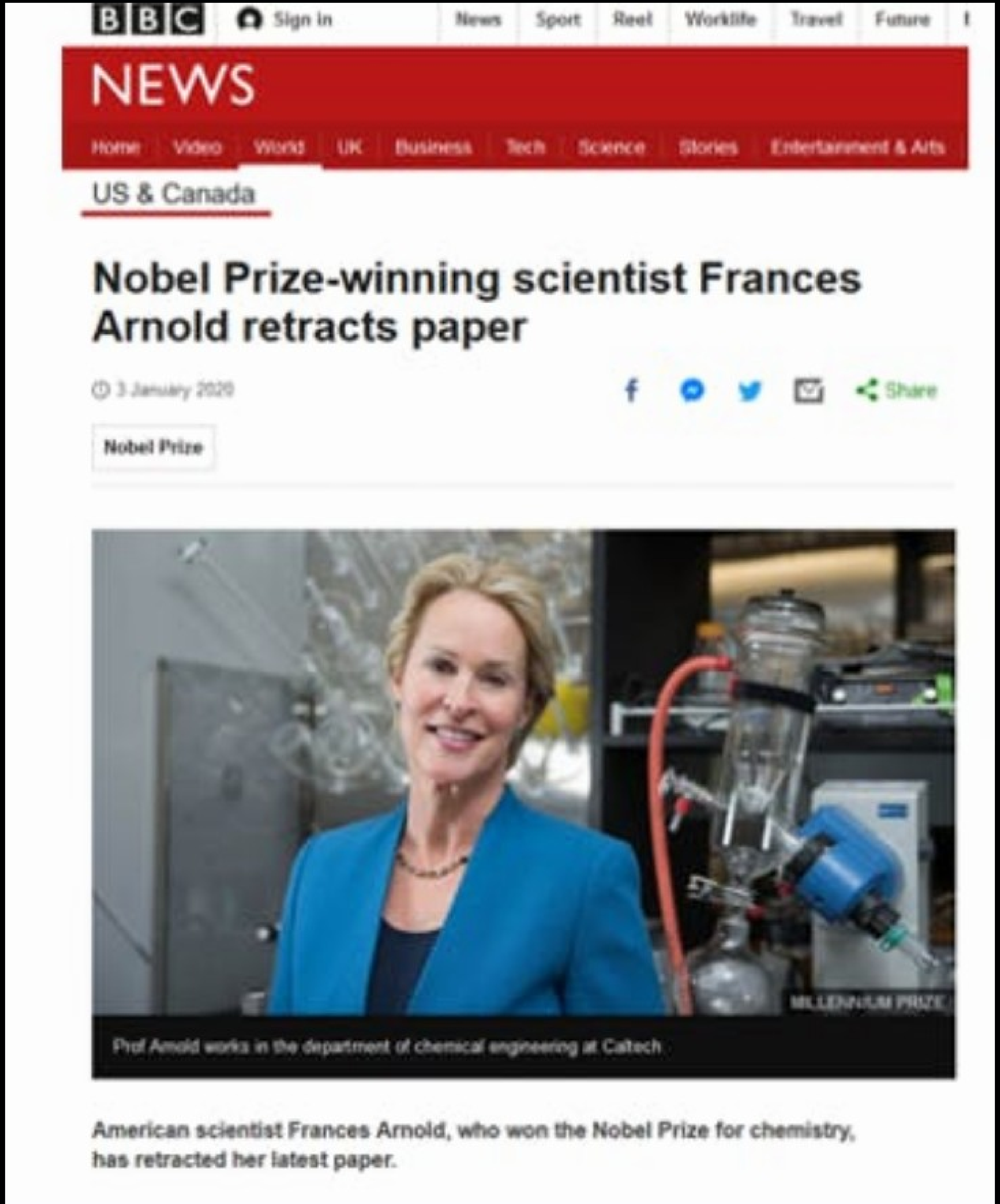
Summary

Policies ensuring that research data are available on public

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 fac



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.

Source: https://www.bbc.com/news/world-us-canada-50989423

Why would you like to share your data?

support the results

follow up study

credibility

mandatory

future research

optimise knowledge growth

open research

training material

reproducibility

it’s messy

sensitive data

open science

transparency

transparancy

funding org requirement

fair

privacy

requirement

inform other research

alternative analyses

expectation of funders

collaboration

progress in research

26

Why would you not share your data?

not good enough

takes time to prepare

technical problems

personal data protection

supervisor does not allow

loosing context

publications

exclusivity

sensitive data

problems with anonymity

unpublished work

gdpr

working on it actively

identifiability

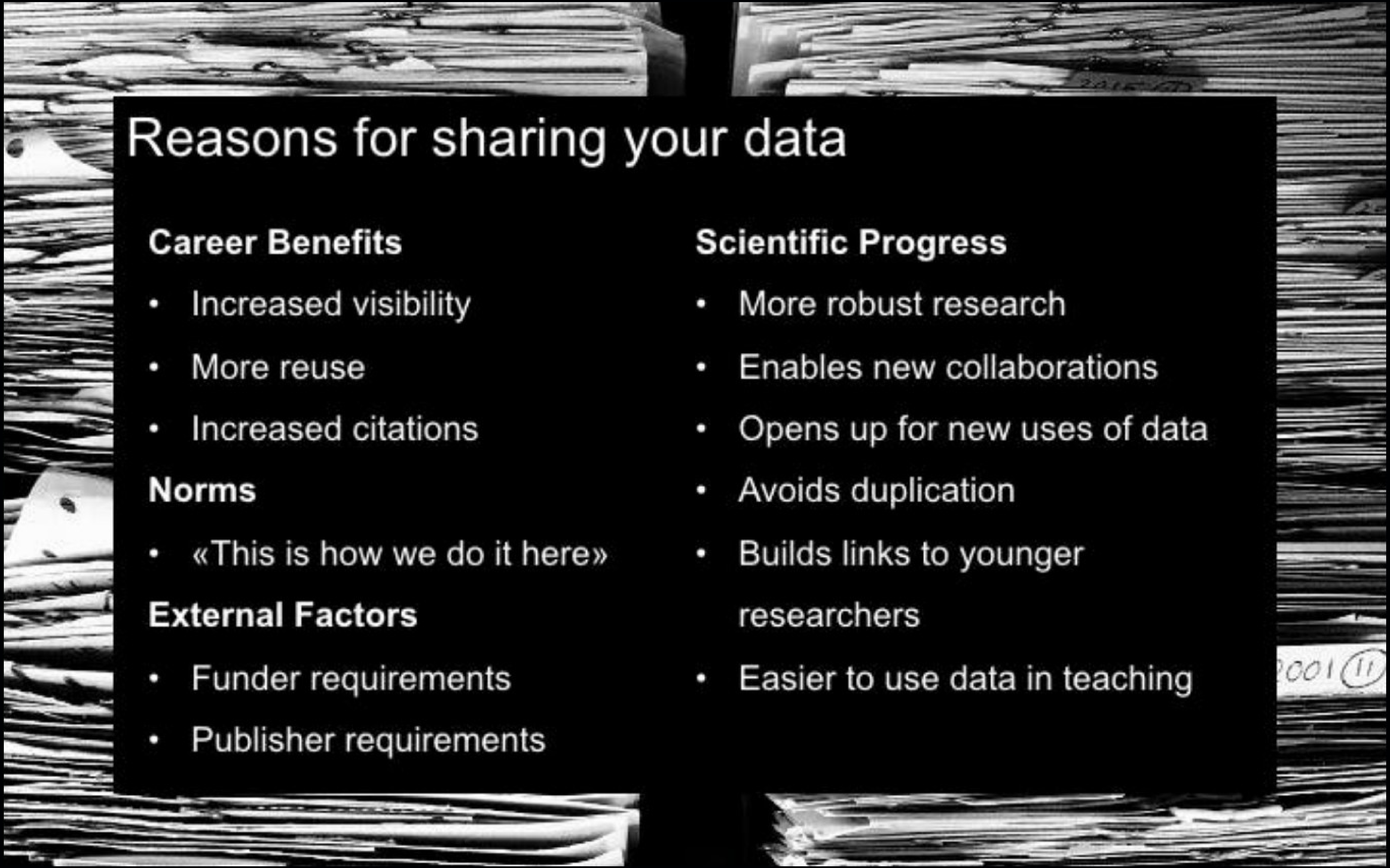
data of economical value

patents

it's messy

workload

privacy



Reasons for sharing your data

Career Benefits

- Increased visibility
- More reuse
- Increased citations

Norms

- «This is how we do it here»

External Factors

- Funder requirements
- Publisher requirements

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

Source: https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/6.-Archive-Publish/Towards-archiving-publication

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 […].

Source: The Research Council of Norway. Policy for open science 2020

Image by: opensourceway

What are the advantages of open science?

- reproducibility
- Reproducibility
- Transparency
- Visibility
- Interdisciplinary research
- Saves cost
- Replicability and reproducibility
- It’s fair, transparent, fuels new research analyses.
- Accessibility to more knowledge

19

What are the advantages of open science?

better for the world that we share and help each other

transparency

Not re-doing the same stuff over and over again

Collaboration

scientific progress

Science is based on sharing knowledge

Transparency reusability

Share, avoid duplication, more citations, more collaborations...

Truth-seeking

19

What are the advantages of open science?

closed science is no science

19

RESEARCH DATA – OPEN BY DEFAULT

FAIR DATA!

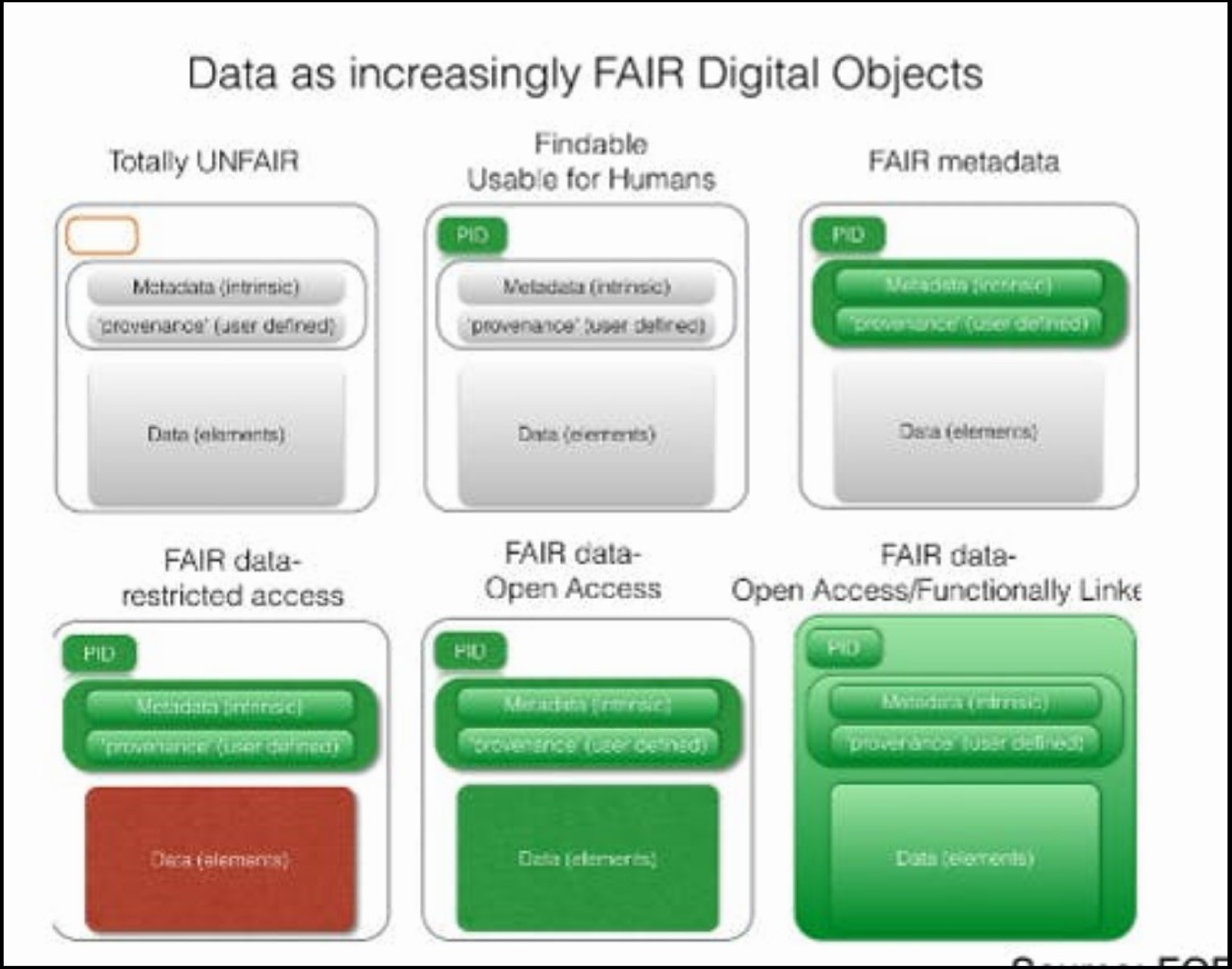
Findable

Accessible

Interoperable

Re-usable

Data as increasingly FAIR Digital Objects



Totally UNFAIR

Metadata (intrinsic)

'provenance' (user defined)

Data (elements)

Findable
Usable for Humans

PID

Metadata (intrinsic)

'provenance' (user defined)

Data (elements)

FAIR metadata

PID

Metadata (intrinsic)

'provenance' (user defined)

Data (elements)

FAIR data-
restricted access

PID

Metadata (intrinsic)

'provenance' (user defined)

Data (elements)

FAIR data-
Open Access

PID

Metadata (intrinsic)

'provenance' (user defined)

Data (elements)

FAIR data-
Open Access/Functionally Linke

PID

Metadata (intrinsic)

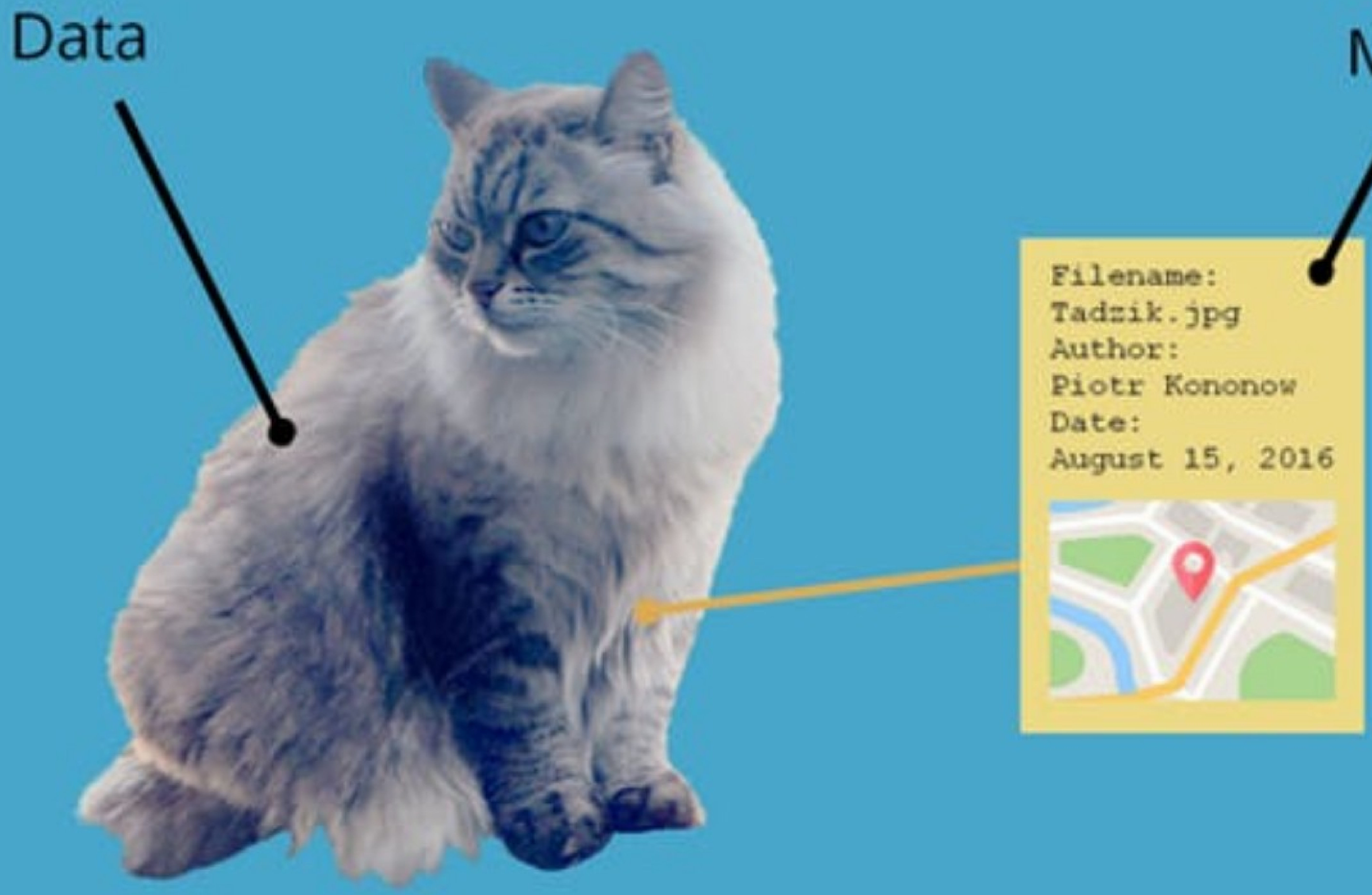
'provenance' (user defined)

Data (elements)

Source: FOR

Source: FORCE11.org

3



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

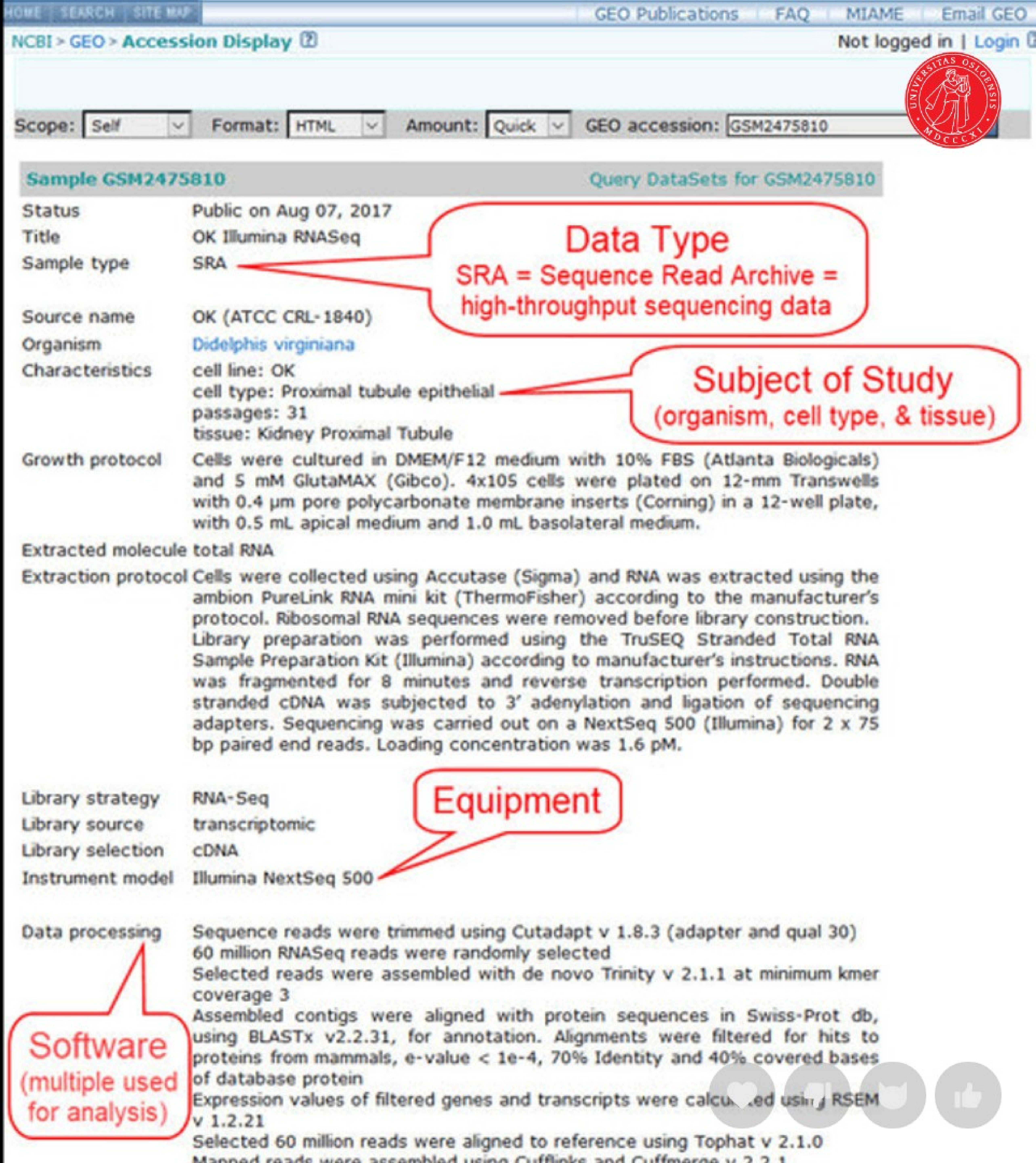
https://dataedo.com/kb/data-glossary/what-is-metadata

Descriptive metadata in a research data context

- Tittle (on 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



HOME | SEARCH | SITE MAP
GEO Publications | FAQ | MIAME | Email GEO
NCBI > GEO > Accession Display
Not logged in | Login

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	Didelphis virginiana
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). 4x105 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

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

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

Equipment

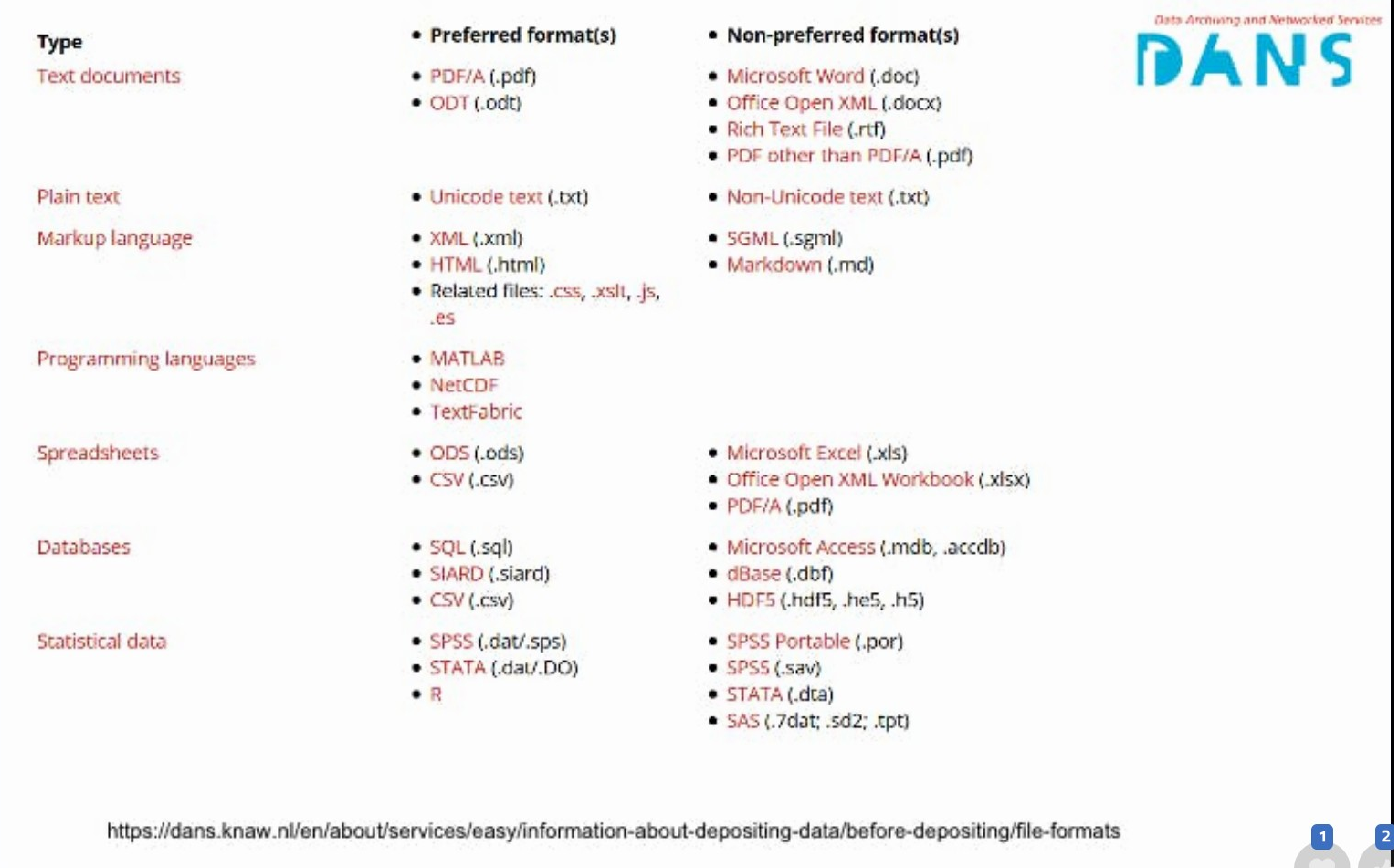
Software
(multiple used for analysis)

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)



Data Archiving and Networked Services

DANS

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) • HDF5 (.hdf5, .he5, .h5)
Statistical data	• SPSS (.dat/.sps) • STATA (.dat/.DO) • R	• SPSS Portable (.por) • SPSS (.sav) • STATA (.dta) • SAS (.7dat; .sd2; .tpt)

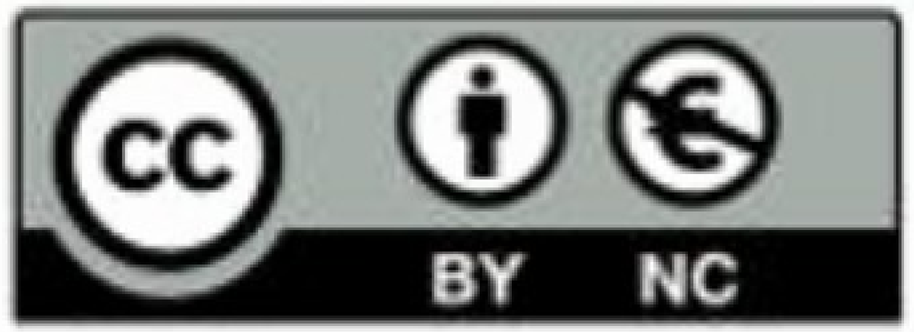
https://dans.knaw.nl/en/about/services/easy/information-about-depositing-data/before-depositing/file-formats

Licensing your data

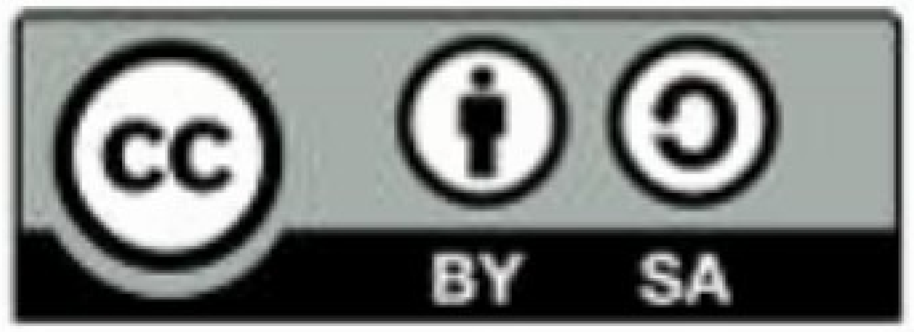
A license agreement is a legal arrangement between the creator/depositor of the data set and the data repository, signifying what a user is allowed to do with the data.

Creative Commons licenses are often used

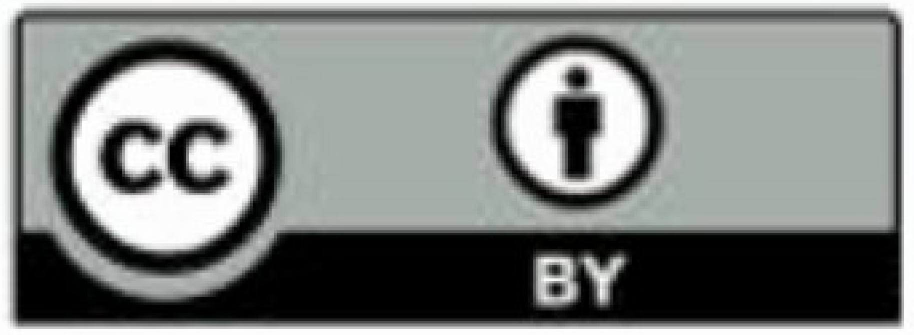
Legal Code
Human Readable
Machine Readable



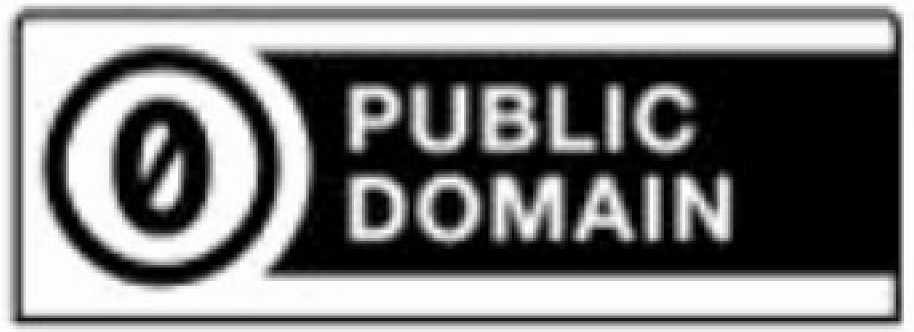
- Credit must be given to the creator
 – Only noncommercial uses of the work are permitted



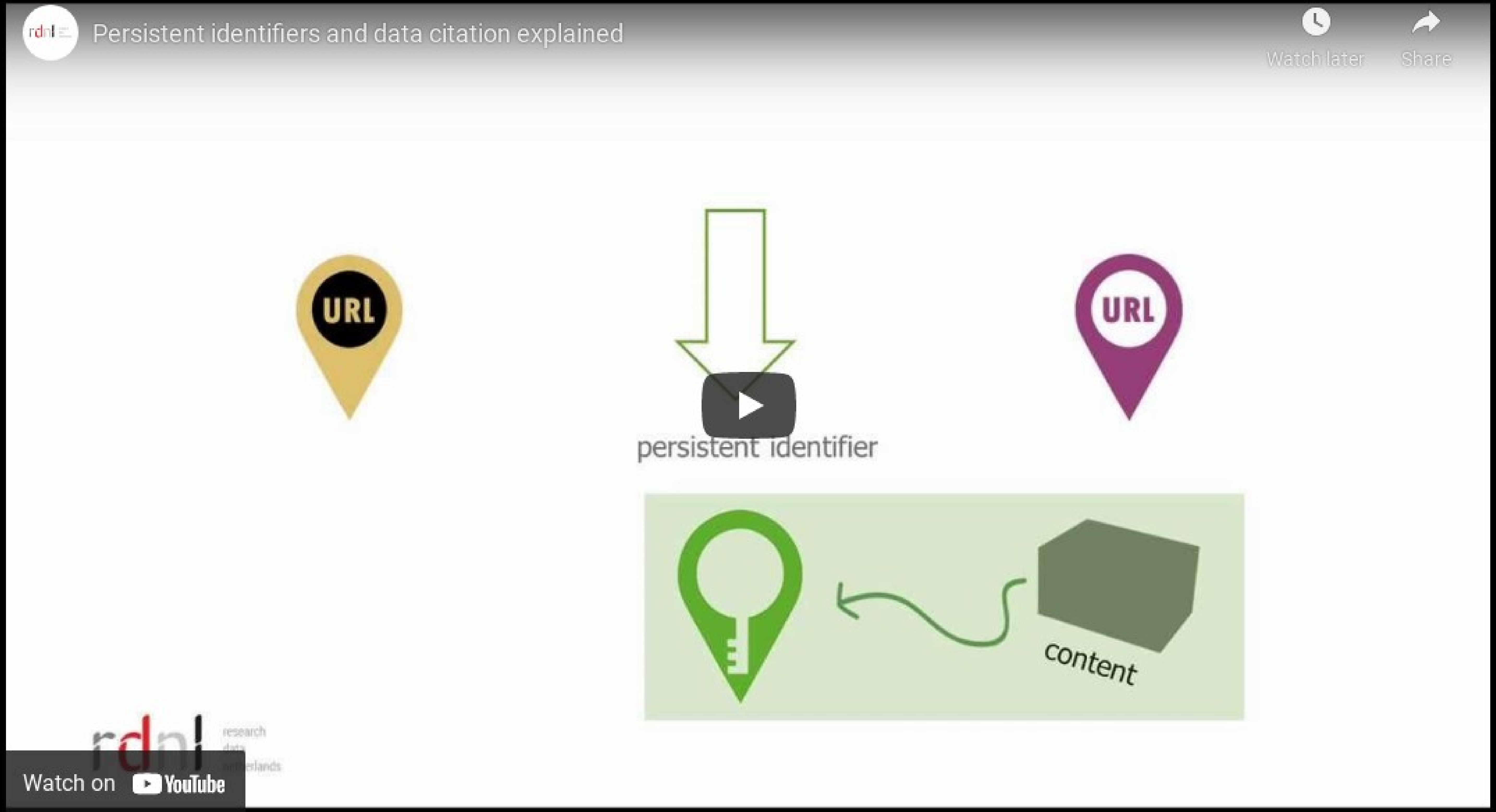
– Credit must be given to the creator
– Adaptations must be shared under the same terms



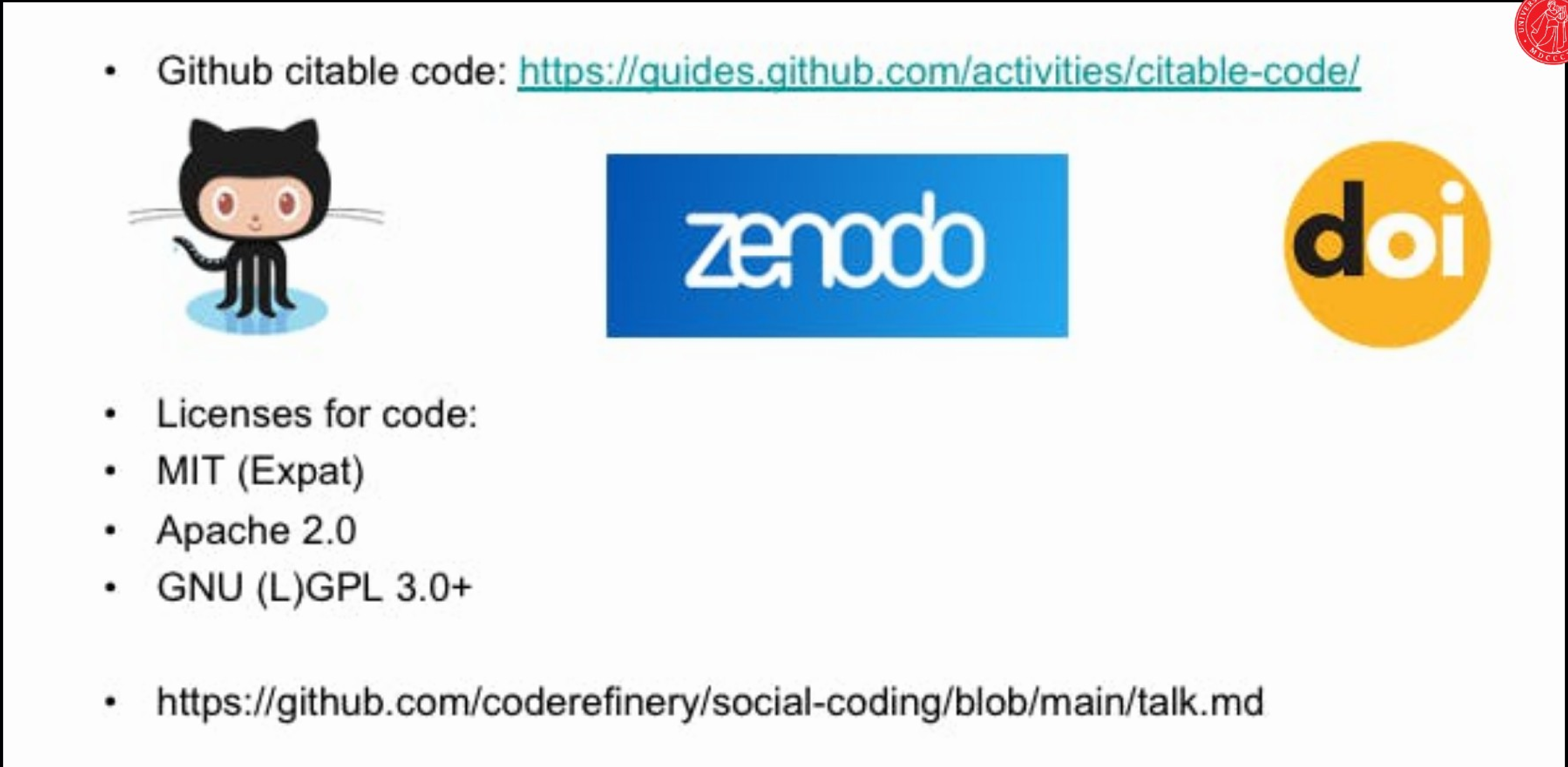
- Credit must be given to the creator



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


Persistent identifiers and data citation explained
Watch later
Share
URL
URL
persistent identifier
content
rdnl research data netherlands
Watch on YouTube

Persistent identifiers and data citation explained



- Github citable code: https://guides.github.com/activities/citable-code/
- Licenses for code:
- MIT (Expat)
- Apache 2.0
- GNU (L)GPL 3.0+
- https://github.com/coderefinery/social-coding/blob/main/talk.md

Reusable code

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

7

minutes break

Case 1

A shares movement

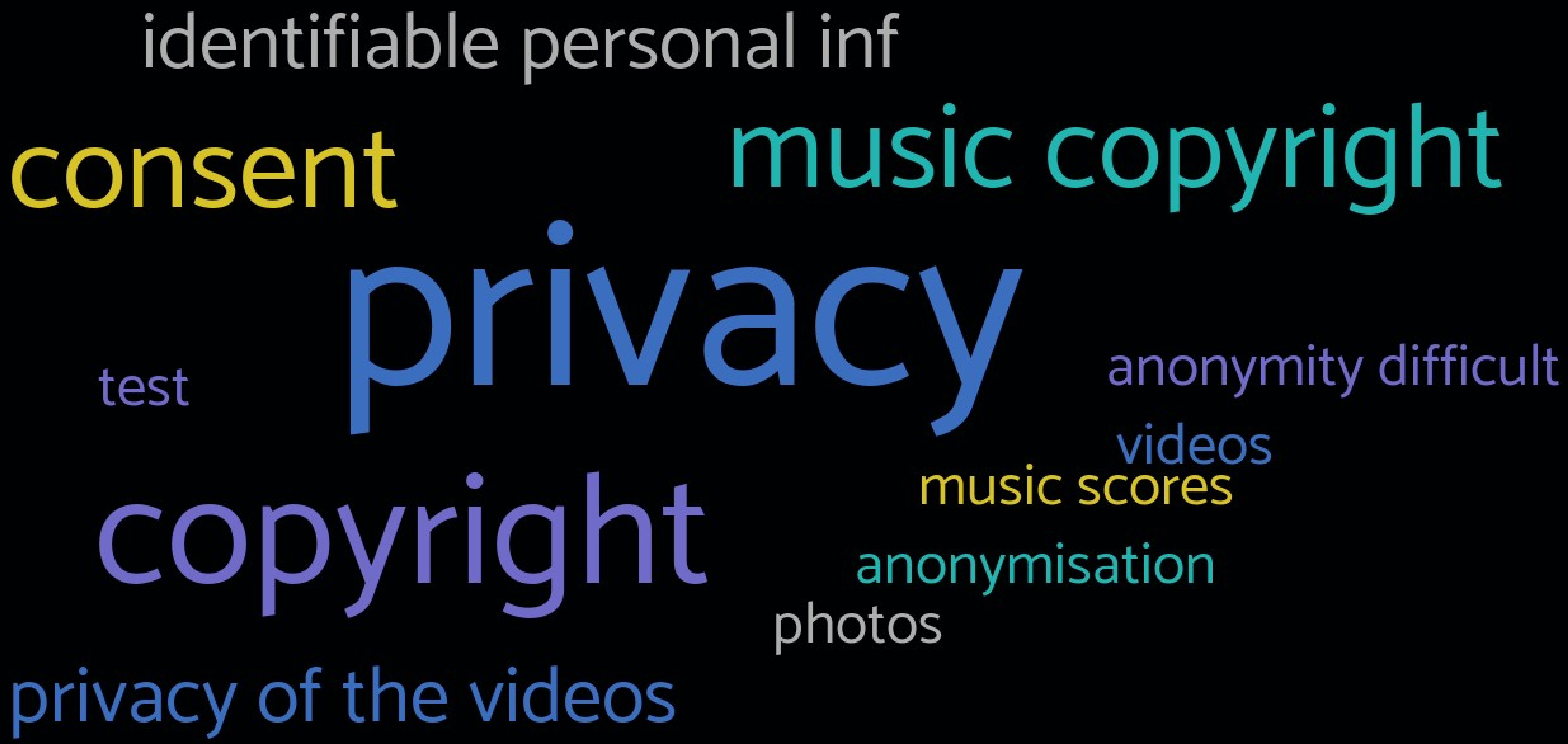
A investigates humans interacting with music

- → 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 form 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?



identifiable personal inf

consent

music copyright

privacy

test

anonymity difficult

videos

music scores

copyright

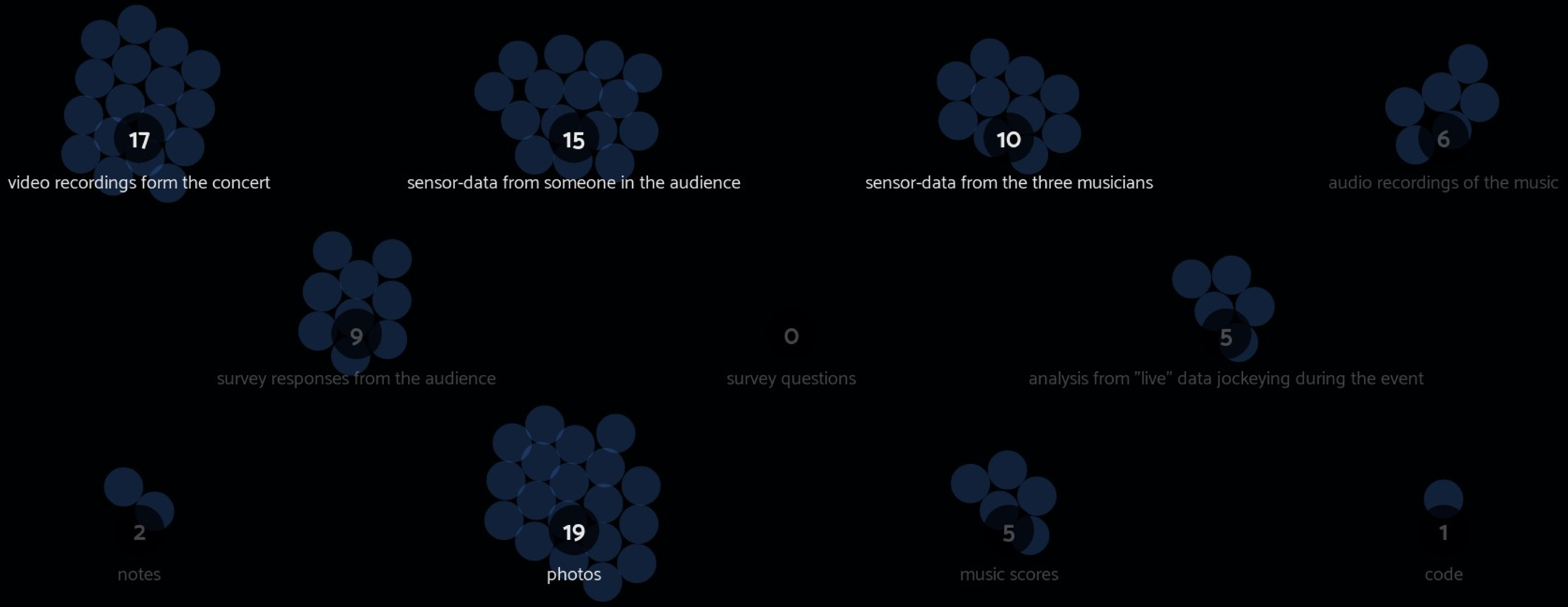
anonymisation

photos

privacy of the videos

17

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



24

The approach chosen by A

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

Case 2

∅ 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?

It is not Ø's data, it's the participants'

Sensitive data!

Ø might have right to use the data, but he does not own the data

Ø does not own the data, his institution does

do you mean sharing data while he is still working on the data?

Confidentiality agreement

He keeps data private thought its collection has been funded by a public institution.

risk of further trauma in participants

have the people beining inteviewed agreed to the data being used by other researches?

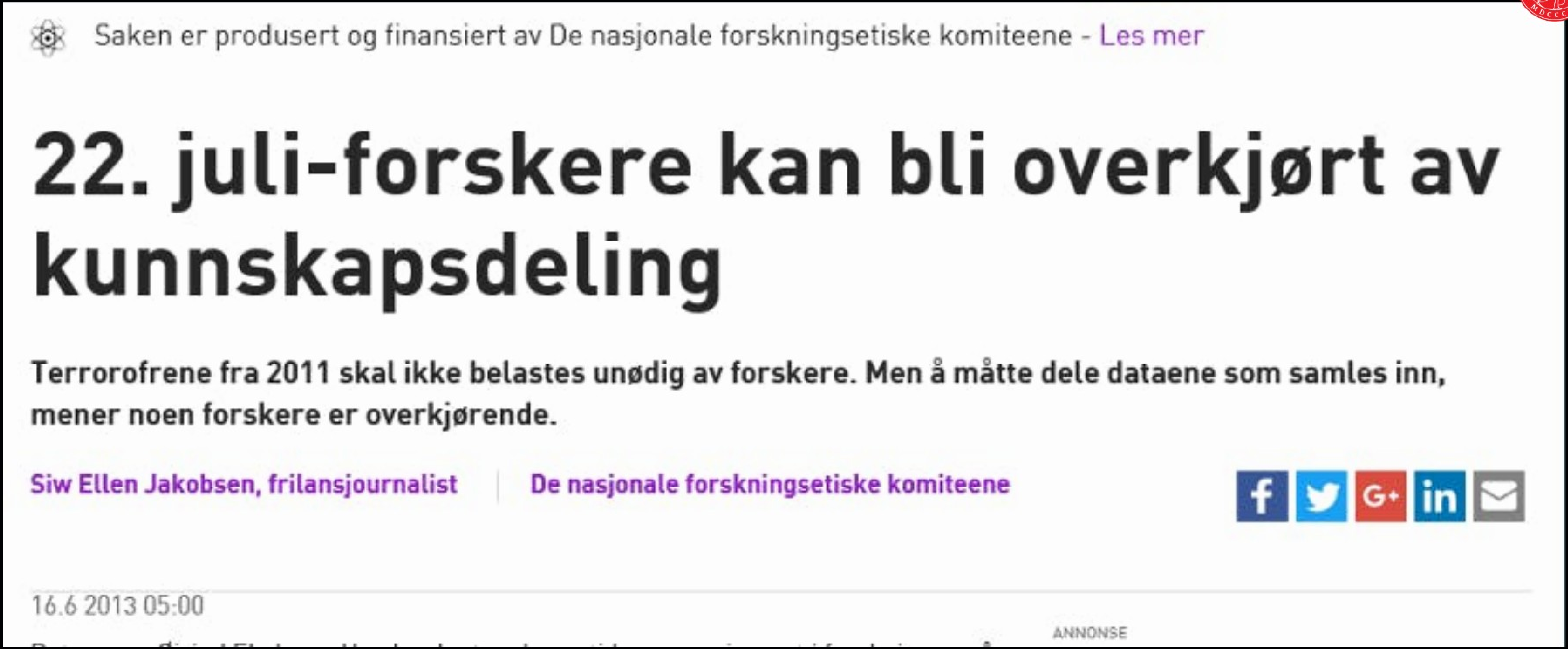
11

What ethical challenges occur in this situation?

How to provide sufficient context for transparency to be true?

Making the data non-identifiable

11



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

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

Regarding reuse and sharing of data

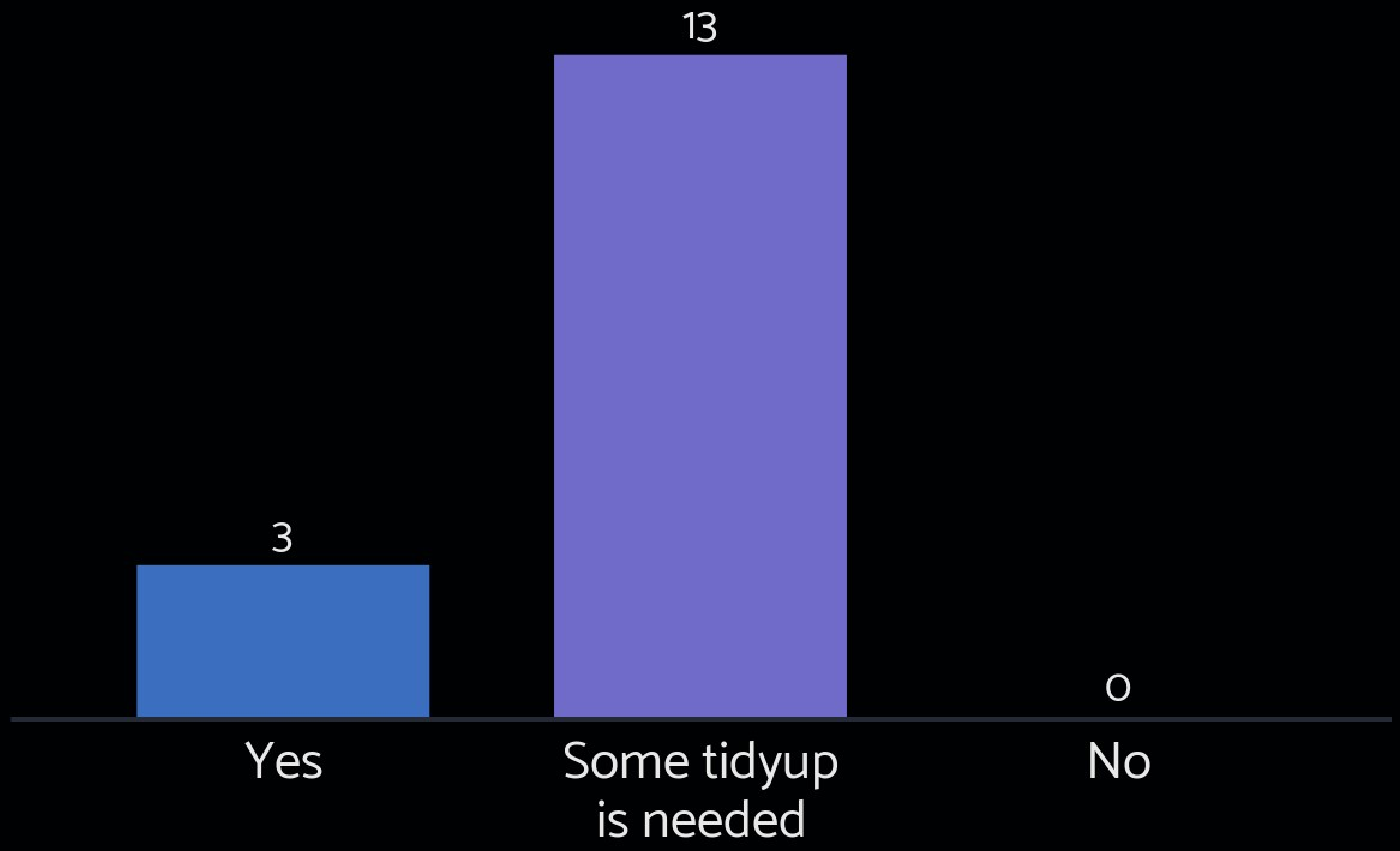
1) The coordination group recommends that in future research common data platforms should be created and used across subjects and institutions.

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 trough 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

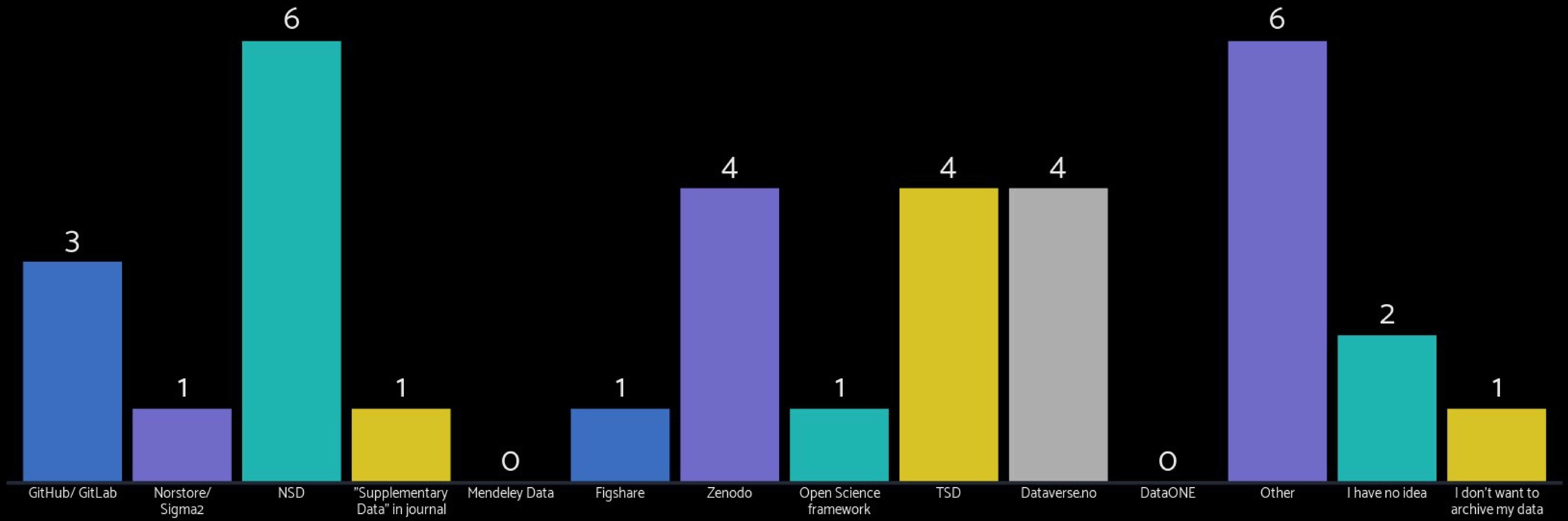
https://nsd.no/nsddata/22juli/datasett.html?a=/nsddata/22juli/datasett/datasett0009.html

Are you ready to share your data?



14

Where do you plan to archive your research data?



16

Ask us

0 questions

0 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