

# DANS

## *Data Archiving and Networked Services*

**Canned data, best before... keeping research data OPEN and FAIR**

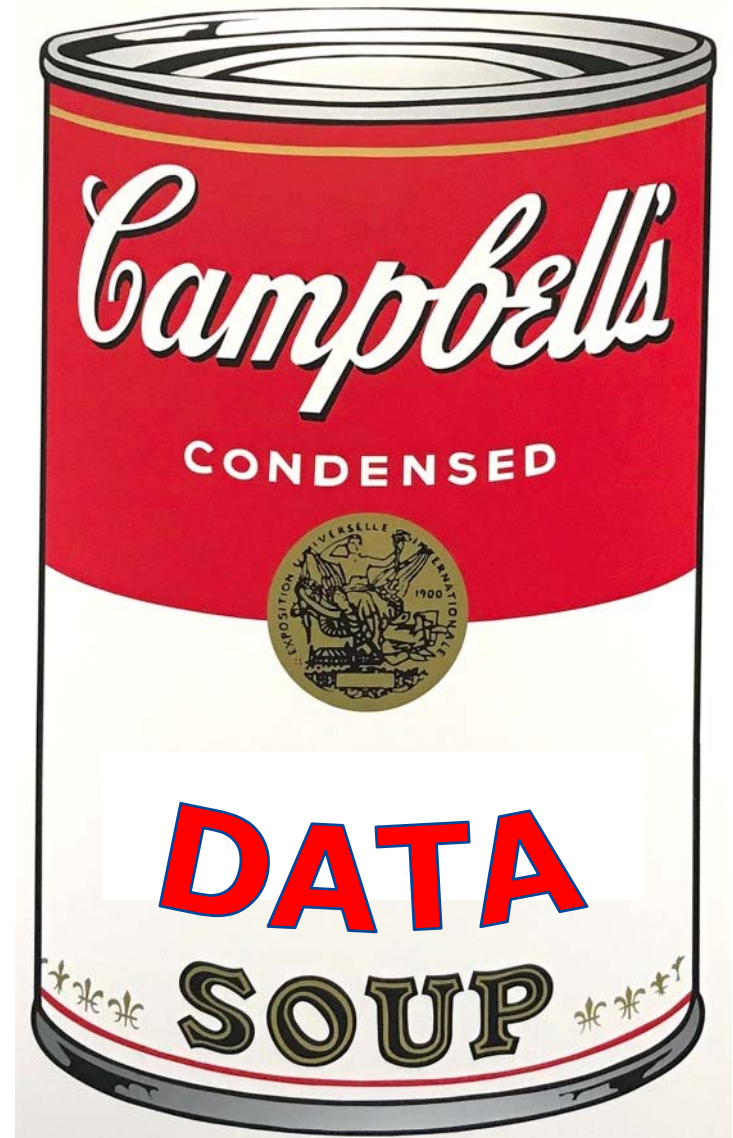
Peter Doorn, Data Archiving and Networked Services (DANS)



@pkdoorn @danskaw

Session "Open science and open data"  
Physics@Veldhoven  
22 January 2019

DANS is an institute of KNAW and NWO



*Driven by data*

# Canning data is not that crazy...



... if you know there is also deep frozen data... in the Arctic World Archive on Svalbard, where also the Global Seed Vault is located



Data Archiving and Networked Services

**DANS**

<https://www.piql.com/arctic-world-archive/>

However, I am not going to talk on deep freezing or canning data, but about:



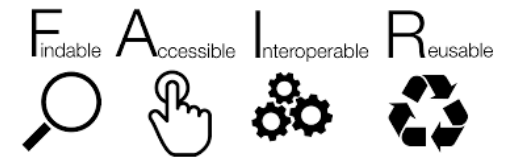
- How DANS archives data for long-term reuse
- Open access to data
- Restrictions to data access
- Big Data and the Long Tail of Data
- FAIR data and Research Data Management



# DANS is about keeping data FAIR



<https://dans.knaw.nl>



Mission: promote  
and provide  
permanent  
access to digital  
research  
resources



Institute of  
Dutch Academy  
and Research  
Funding  
Organisation  
(KNAW & NWO)  
since 2005

First predecessor  
dates back to  
1964 (Steinmetz  
Foundation),  
Historical Data  
Archive 1989

Data Archiving and Networked Services

**DANS**

# DANS core data services

HOME P.K. DOORN MY DATASETS

## EASY

data: write a data paper for the new peer reviewed, online-only open access Resea

For more info: [brill.com/rdj](http://brill.com/rdj)

EASY offers sustainable archiving of research data and access to thousands of datasets.

Search...

> Advanced search > Browse

Data Archiving and Networked Services

## DANS

## NARCIS

The gateway to scholarly information in the Netherlands

> Submit Content to NARCIS

Search...

1,882,772	214,917	69,173	59,753	2,970
PUBLICATIONS	DATA SETS	RESEARCH	PEOPLE	ORGANISATIONS

EASY: certified Electronic Archiving System for self-deposit

Dataverse

DataverseNL Dataverse

6,588 Downloads

### DataverseNL Dataverse Network

4TU. CENTRE FOR RESEARCH DATA  
4TU.Center for research data Dataverse

VU Vrije Universiteit Amsterdam  
Vrije Universiteit Amsterdam Dataverse

NIOO-KNAW Dataverse

Tilburg University  
Tilburg University Dataverse

Search this dataverse...  Advanced Search

HOME ABOUT NARCIS

1 to 10 of 672 Results

Mindwandering during Attention Performance: effects of ADHD-inattention Symptomatology, Negative Mood, Ruminative Response Style and Working Memory Capacity  
Feb 3, 2017 - Clinical Psychological Science Dataverse

Jonkman, Lisa; Markus, Rob; Franklin, Michael; Dalfsen, Jens van, 2017, "Mindwandering during Attention Performance: effects of ADHD-Inattention Symptomatology, Negative Mood, Ruminative Response Style and Working Memory Capacity", hdi:10411/20895, DataverseNL Dataverse, V1

Dataset\_Bamelis\_et\_al\_2015  
Feb 1, 2017 - Clinical Psychological Science Dataverse

Weitzelaar, Pim; Arntz, Amoud, 2017, "Dataset\_Bamelis\_et\_al\_2015", hdi:10411/20892, DataverseNL Dataverse, V1

Dataset of the study by Bamelis et al., 2015 These data were also used in the study by Weitzelaar et al., 2017 Please refer to the file SuppInfo\_Dataset\_Bamelis\_et\_al\_2015 for more information.

DataverseNL: data repository at universities and other institutions

# Additional services



Background Archive



Training & Consultancy

<http://datasupport.researchdata.nl/>



<https://data.mendeley.com/>  
<https://datadryad.org>

Work in Progress: Software Archive



<https://www.softwareheritage.org/>

Research Data Journal for the Humanities and Social Sciences



BRILL

<http://www.brill.com/rdj>



# What do we have to offer?

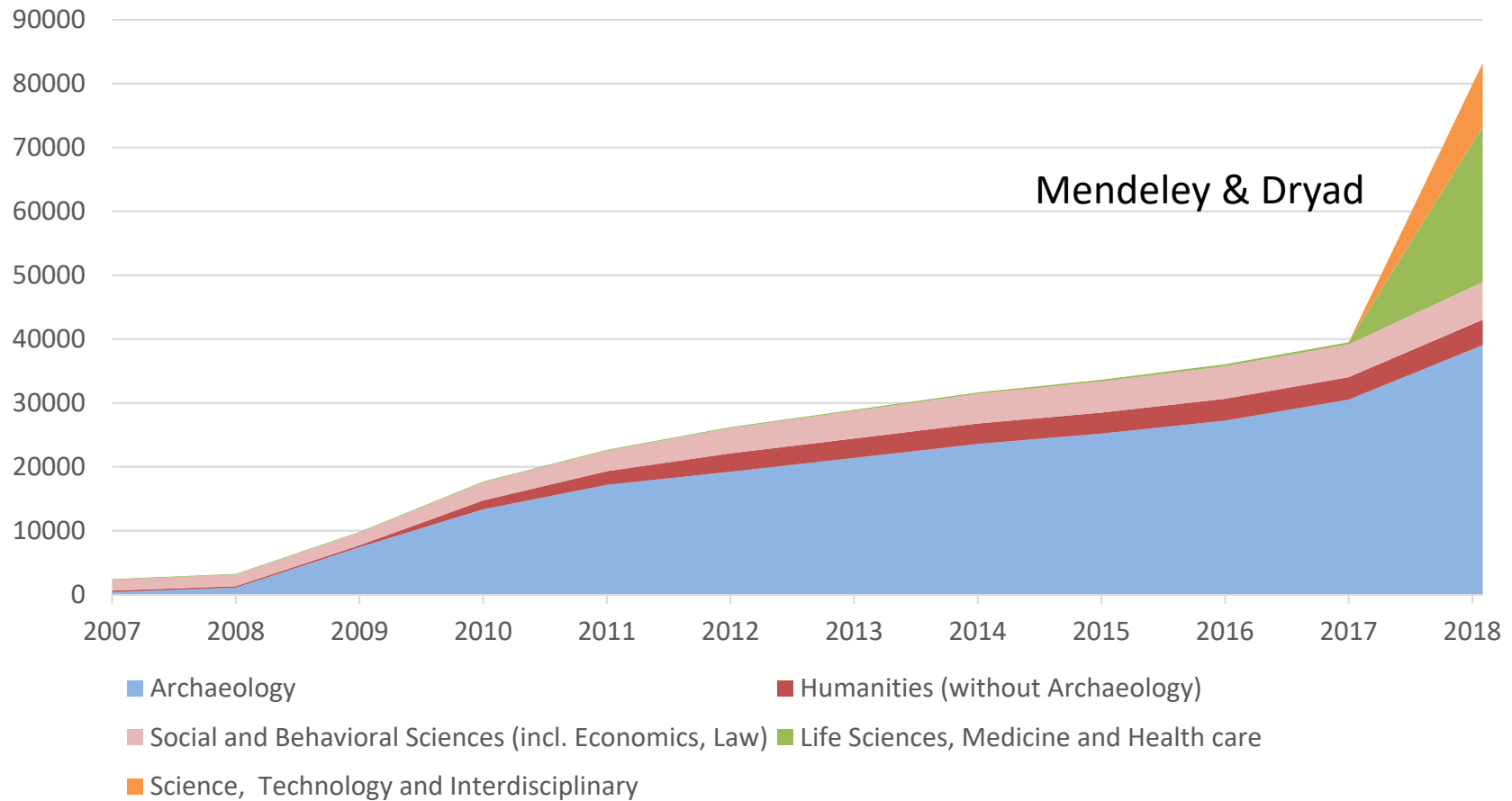
- Find and reuse existing research data and other resources
- Store your data securely and FAIR for:
  - Data management during a research project
  - Certified permanent archiving afterward
  - Sharing data according to your wishes and institutional policies:  
"Open if possible, protected if needed"
- Expert advice on data management

Most of our services are free for individual researchers.

Exceptions:

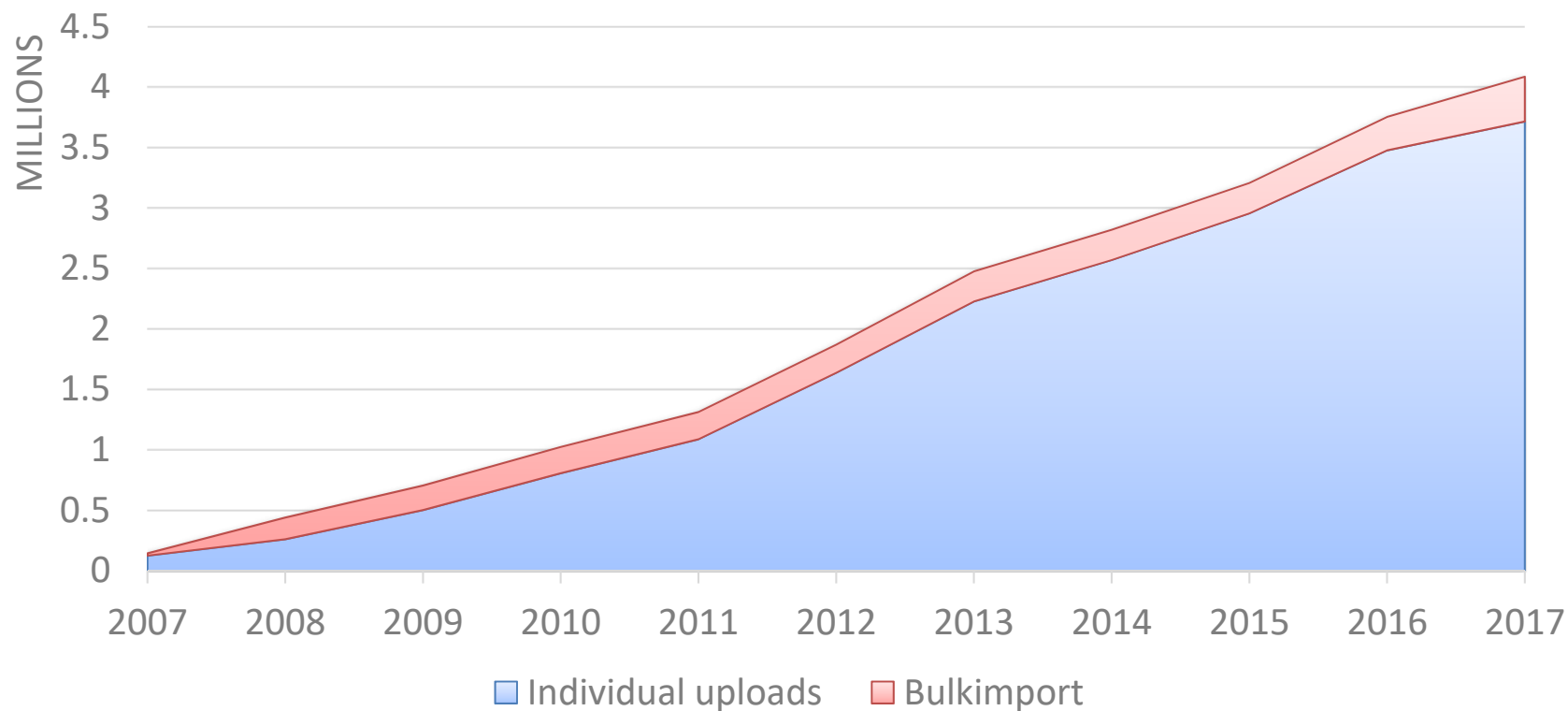
- Big volumes of data require separate treatment and budgets
- Involvement of DANS data experts in projects / consultancy
- Institutional arrangements

# Datasets in EASY per discipline, 2007-2018



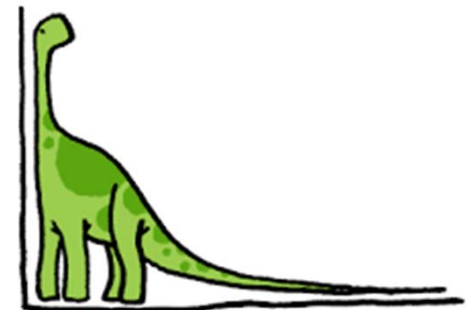
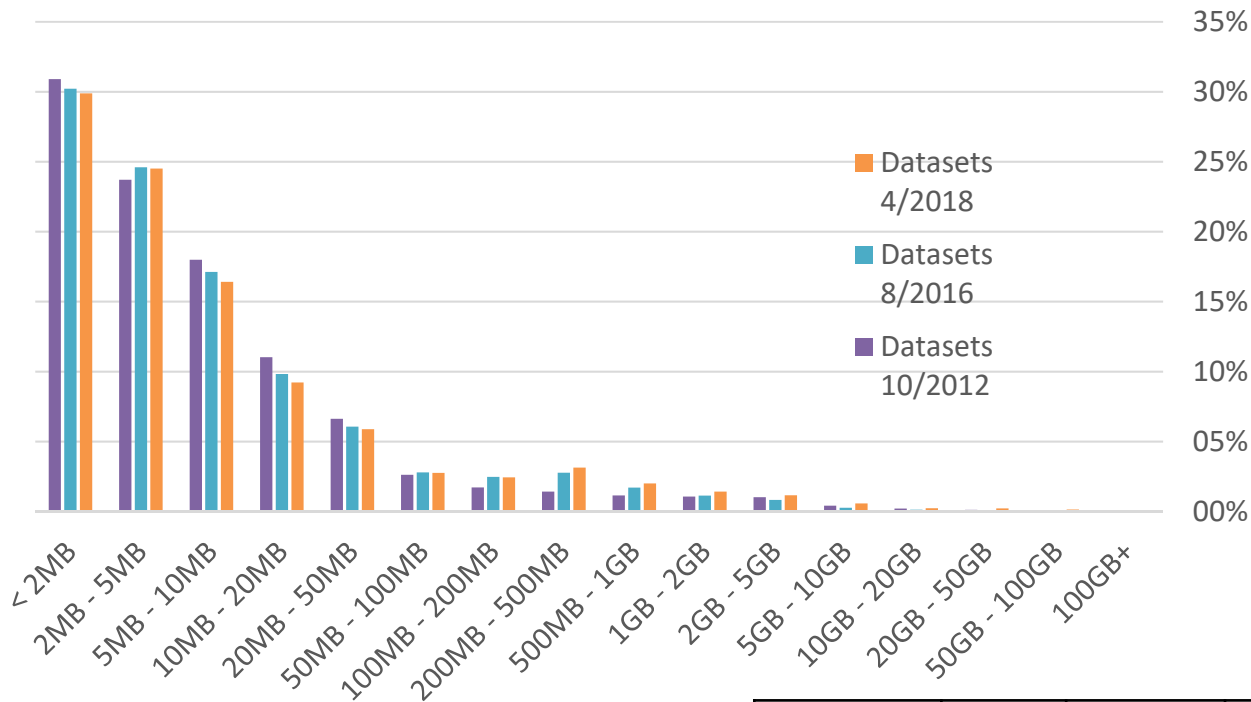


# Growth of number of Files in EASY, 2007-2017



# Datasets in DANS EASY archive according to size

Datasets relative



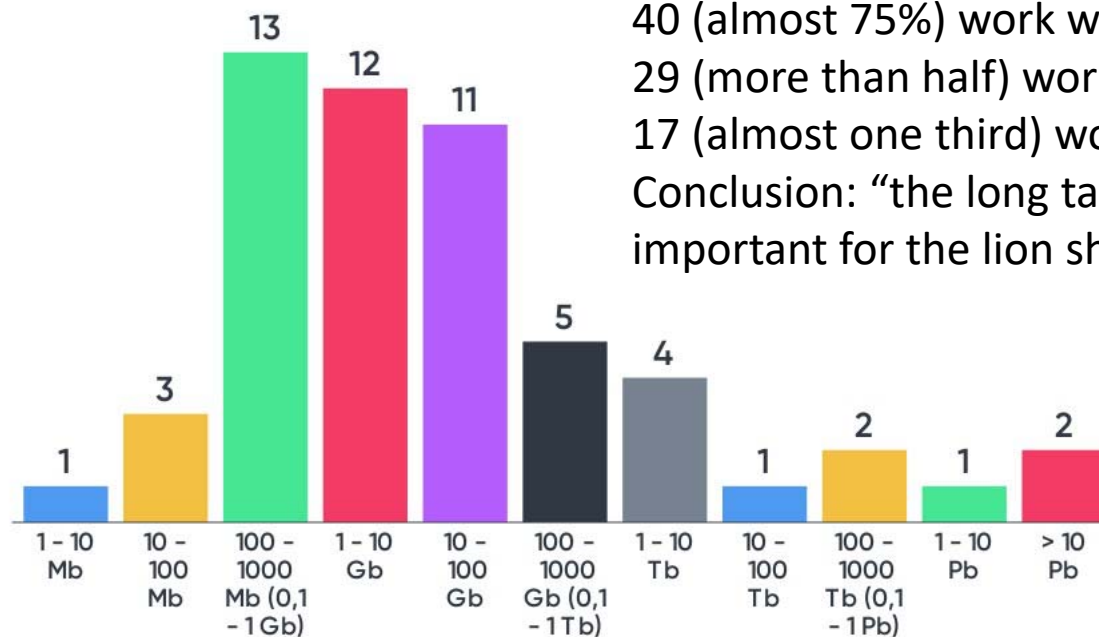
The long tail of research data

	Datasets 10/2012	Datasets 8/2016	Datasets 4/2018
> 1 Gb	2,8%	2,5%	3,8%
> 2Gb	1,8%	1,3%	2,3%

# How does that compare to this Physics Community?

What is size order of the data you work with in your research?

Mentimeter



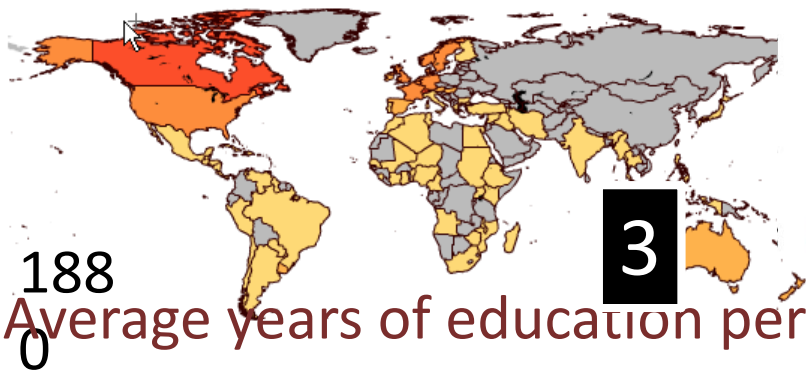
10 out of 55 work with data > 10 Tb  
40 (almost 75%) work with data < 100 Gb  
29 (more than half) work with data < 10 Gb  
17 (almost one third) work with data < 1 Gb  
Conclusion: “the long tail” of data is important for the lion share of physicists

55

# Long-tail data remains typical for the humanities (and for many other disciplines)

## Collaborative work: bringing together data from many scholars

1. Historical shipping
2. Digitized censuses
3. Global inequality
4. Holocaust studies
5. Dendrochronology



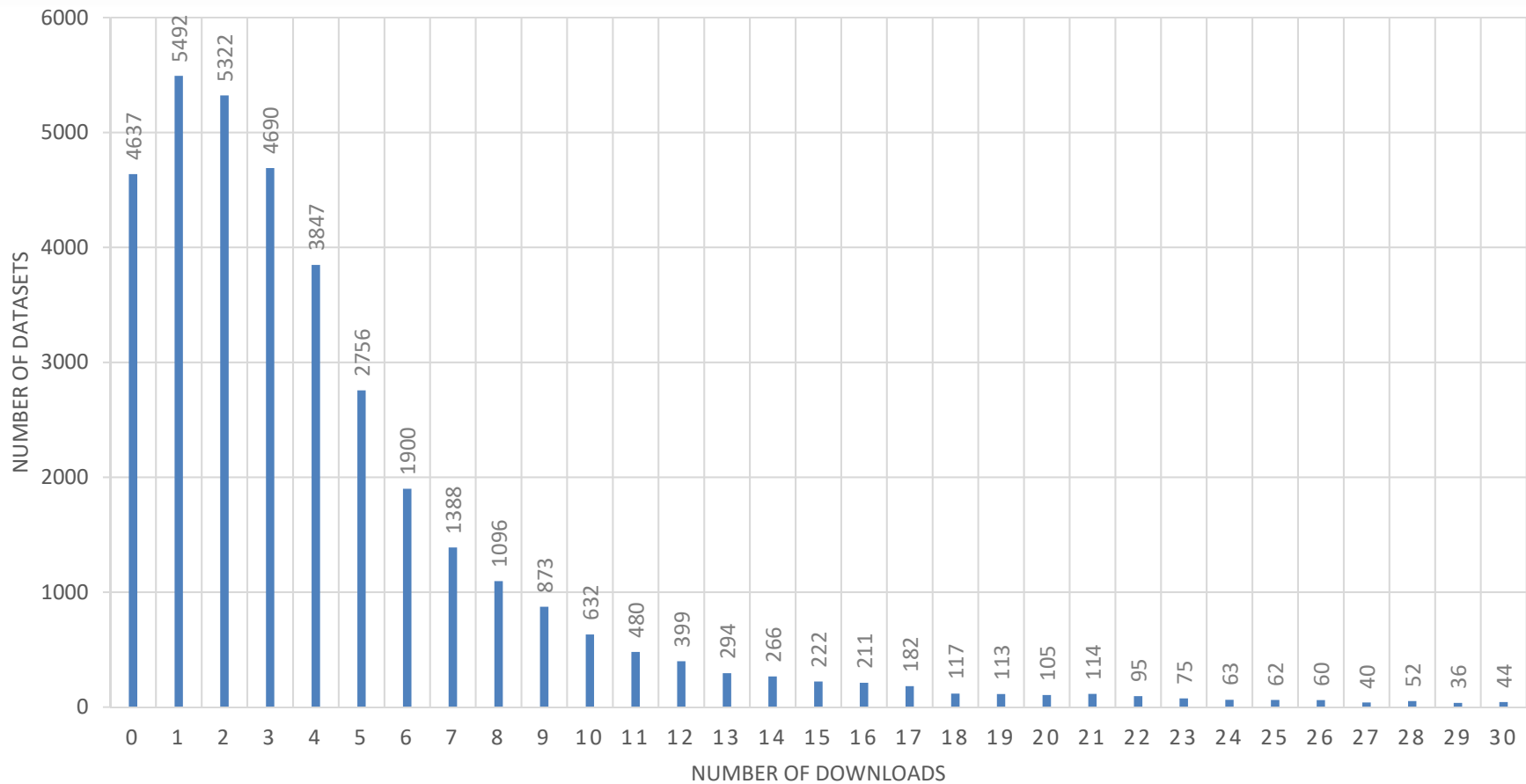
188  
0  
Average years of education per capita



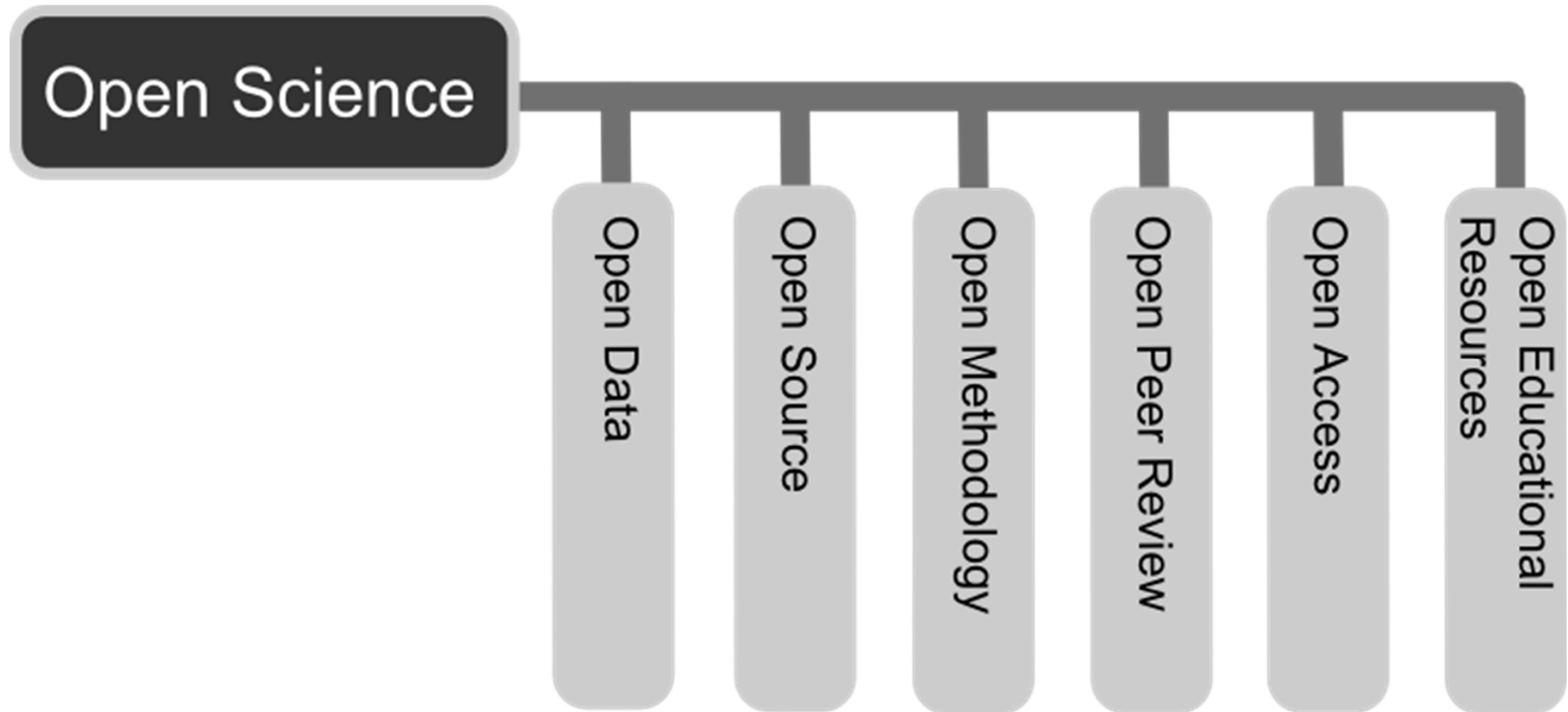
# Top 10 of downloaded datasets from EASY

Rank	Title of Dataset	Dataset Downloads	Downloaded Files	Persistent Identifier
1	Nationaal Kiezersonderzoek, NKO 2006	1125	5398	<a href="https://nbn-resolving.org/urn:nbn:nl:ui:13-4zd-x4e">urn:nbn:nl:ui:13-4zd-x4e</a>
2	De steentijd van Nederland	1108	1452	<a href="https://nbn-resolving.org/urn:nbn:nl:ui:13-tg4-mof">urn:nbn:nl:ui:13-tg4-mof</a>
3	Brabant cohort - derived student file	884	2161	<a href="https://nbn-resolving.org/urn:nbn:nl:ui:13-zgkg-jv">urn:nbn:nl:ui:13-zgkg-jv</a>
4	Nationaal Kiezersonderzoek, 2010 - NKO 2010	829	4279	<a href="https://nbn-resolving.org/urn:nbn:nl:ui:13-9x4l-vy">urn:nbn:nl:ui:13-9x4l-vy</a>
5	Netherlands Longitudinal Lifecourse Study - NELLS First Wave - 2009 - versie 1.3	742	1920	<a href="https://nbn-resolving.org/urn:nbn:nl:ui:13-54c-ue">urn:nbn:nl:ui:13-54c-ue</a>
6	Geological-Geomorphological map of the Rhine-Meuse delta, The Netherlands	736	34645	<a href="https://nbn-resolving.org/urn:nbn:nl:ui:13-nqjn-zl">urn:nbn:nl:ui:13-nqjn-zl</a>
7	Nationaal Kiezersonderzoek 2012 - NKO 2012	684	1339	<a href="https://nbn-resolving.org/urn:nbn:nl:ui:13-93iu-8p">urn:nbn:nl:ui:13-93iu-8p</a>
8	Nationaal Kiezersonderzoek, NKO 2002 2003	616	1197	<a href="https://nbn-resolving.org/urn:nbn:nl:ui:13-hvz-17u">urn:nbn:nl:ui:13-hvz-17u</a>
9	WoON2012: release 1.0 - WoonOnderzoek Nederland 2012 (voor overheid, universiteiten en overige partijen)	603	5043	<a href="https://nbn-resolving.org/urn:nbn:nl:ui:13-60fd-6i">urn:nbn:nl:ui:13-60fd-6i</a>
10	International Crime Victims Surveys - ICVS - 1989, 1992, 1996, 2000, 2005	568	3026	<a href="https://nbn-resolving.org/urn:nbn:nl:ui:13-wx0-h0o">urn:nbn:nl:ui:13-wx0-h0o</a>

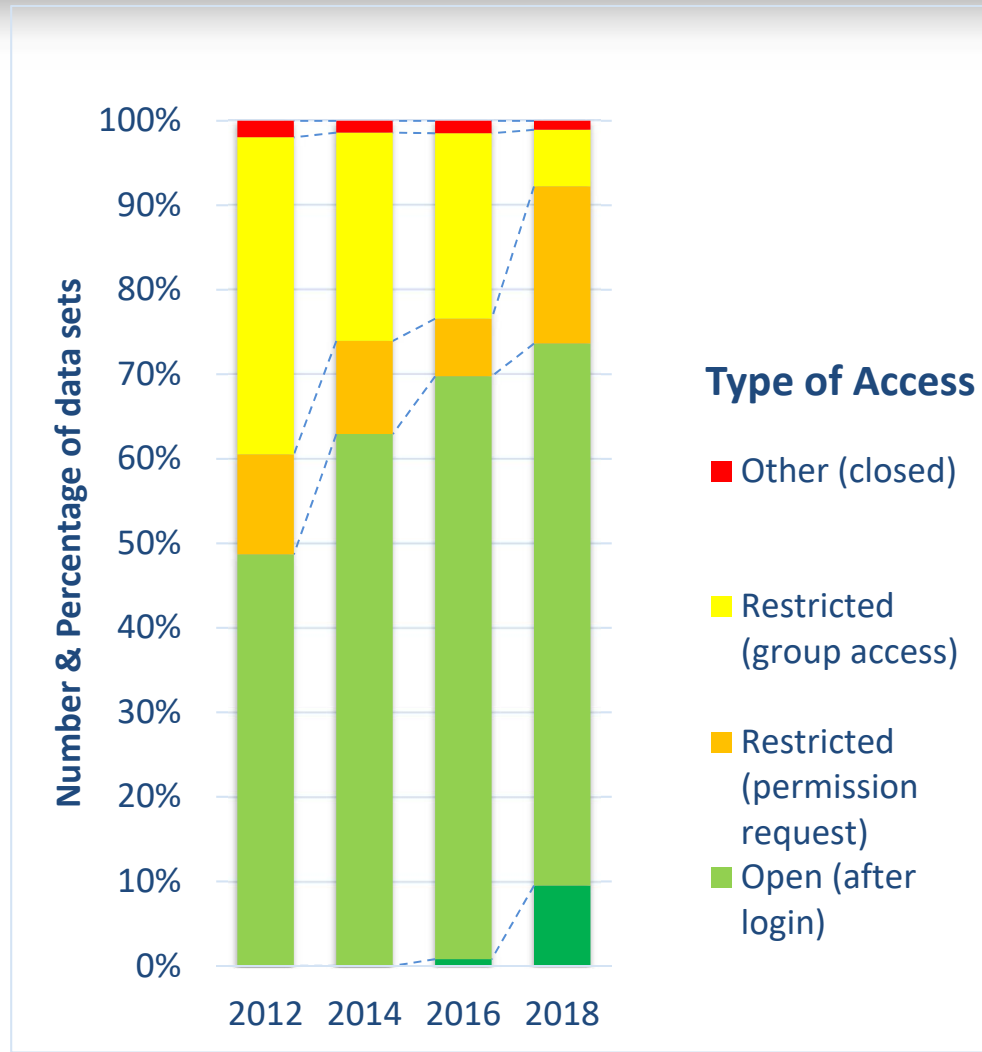
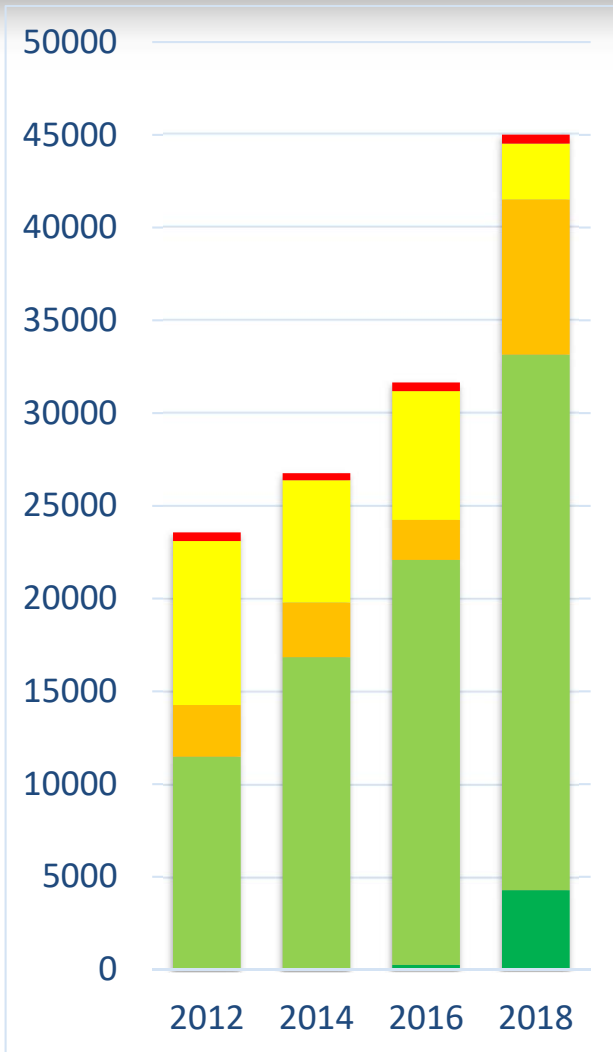
# Frequency of 0-30 dataset downloads, 2007-2017



# Six Principles of Open Science

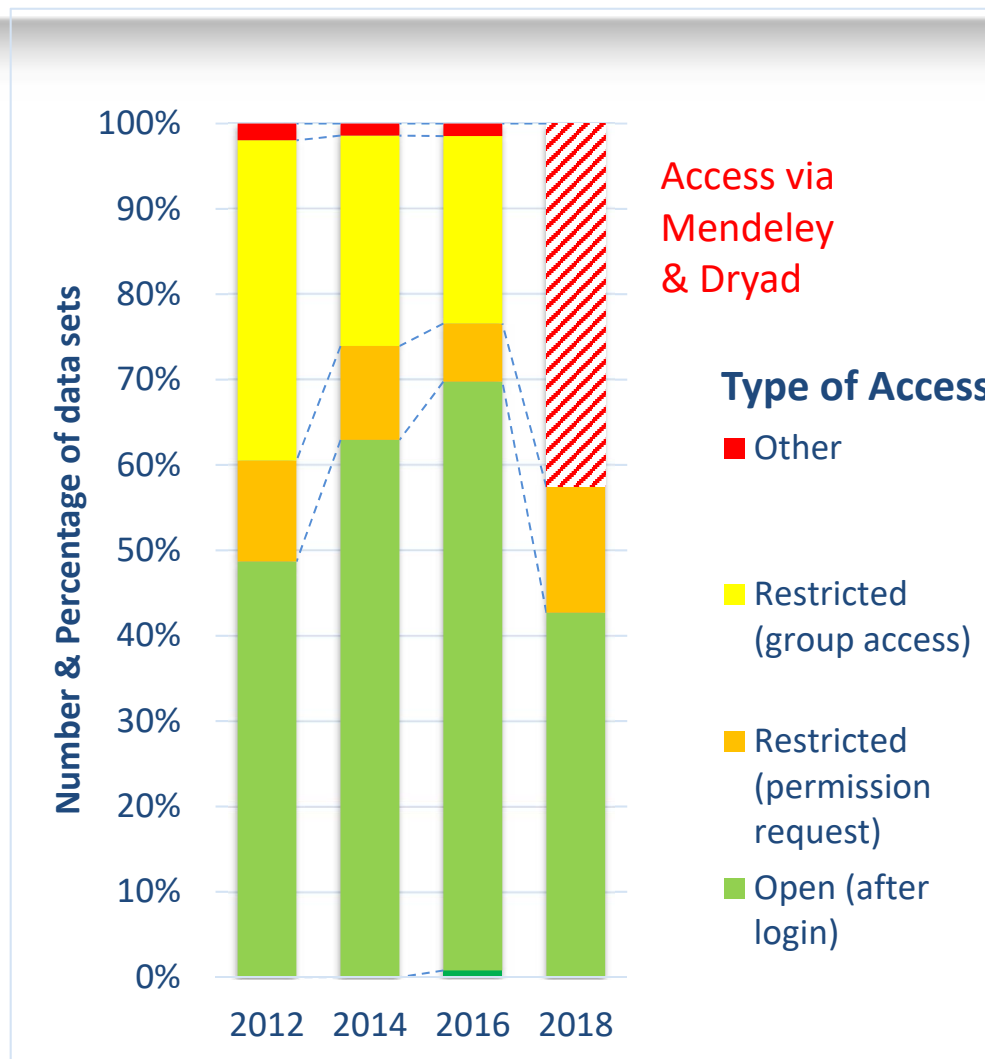
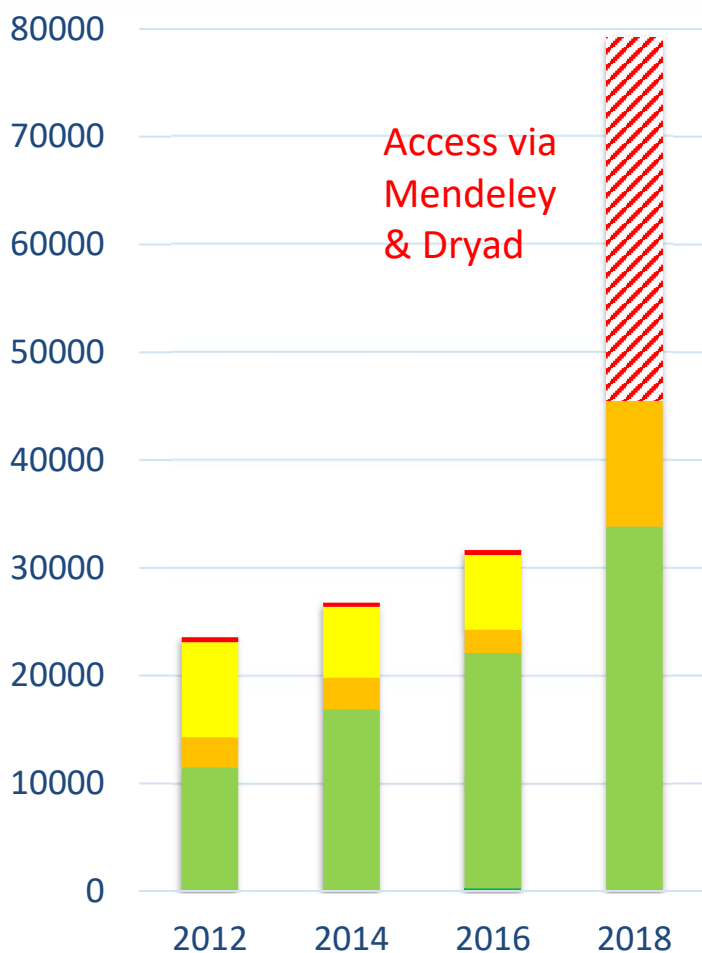


# Access to Datasets in DANS archive 2012-2018 (without Mendeley & Dryad)

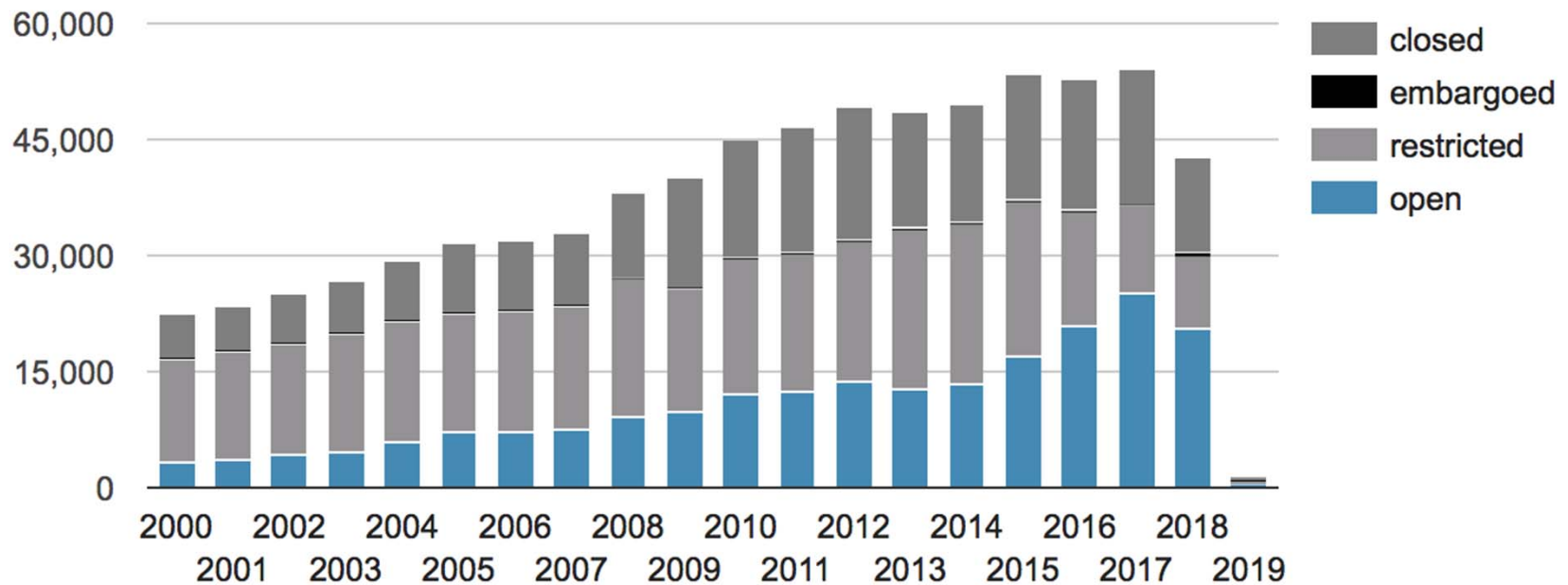




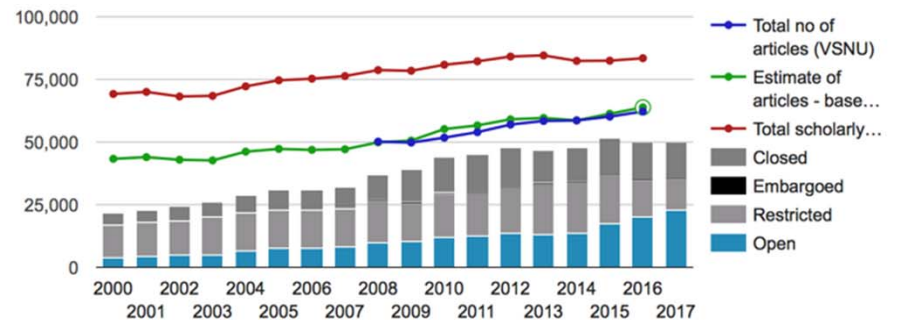
# Access to Datasets in DANS archive 2012-2018 (including Mendeley & Dryad)



# Open and Closed Access Articles in NARCIS per year of publication, 2000-now



Estimates of what we miss, currently about 10,000 articles (20%)



# Personal Data and the DANS archive

- Researcher uploading data is primarily responsible
- DANS can only check marginally
- Tool needed to support decisions on required data protection – compliant with GDPR and national legislation
- Use Harvard's DataTags as starting point



DataTags



<http://datatags.org/>

# Privacy: GDPR and Datatags



- General Data Protection Regulation EU – Passed 14 April 2016
  - New European “Law” from 25 May 2018 onward:
    - Data minimisation required
    - Informed consent important
    - Data Protection Officer mandatory, data protection impact assessment (DPIA)
    - Right to know (e.g. data leakages), right to be forgotten
    - High fines for trespassing (data leakage!)
  - Implications for sharing data on human subjects?
    - Researchers (and their employers) don’t know
    - Data repositories don’t know
- Data Tagging Approach, initially developed at Harvard



# Background of DataTags approach @ Harvard

Sweeney & Crosas introduced the notion of a datatags repository

- Stores and shares data files in accordance with different security levels, access requirements and usage agreements
- Based on American laws and legislations of personal data

Tag Type	Description	Security Features	Access Credentials
Blue	Public	Clear storage, Clear transmit	Open
Green	Controlled public	Clear storage, Clear transmit	Email- or OAuth Verified Registration
Yellow	Accountable	Clear storage, Encrypted transmit	Password, Registered, Approval, Click-through DUA
Orange	More accountable	Encrypted storage, Encrypted transmit	Password, Registered, Approval, Signed DUA
Red	Fully accountable	Encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA
Crimson	Maximally restricted	Multi-encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA



Harvard DataTags

# Our approach: Step by step

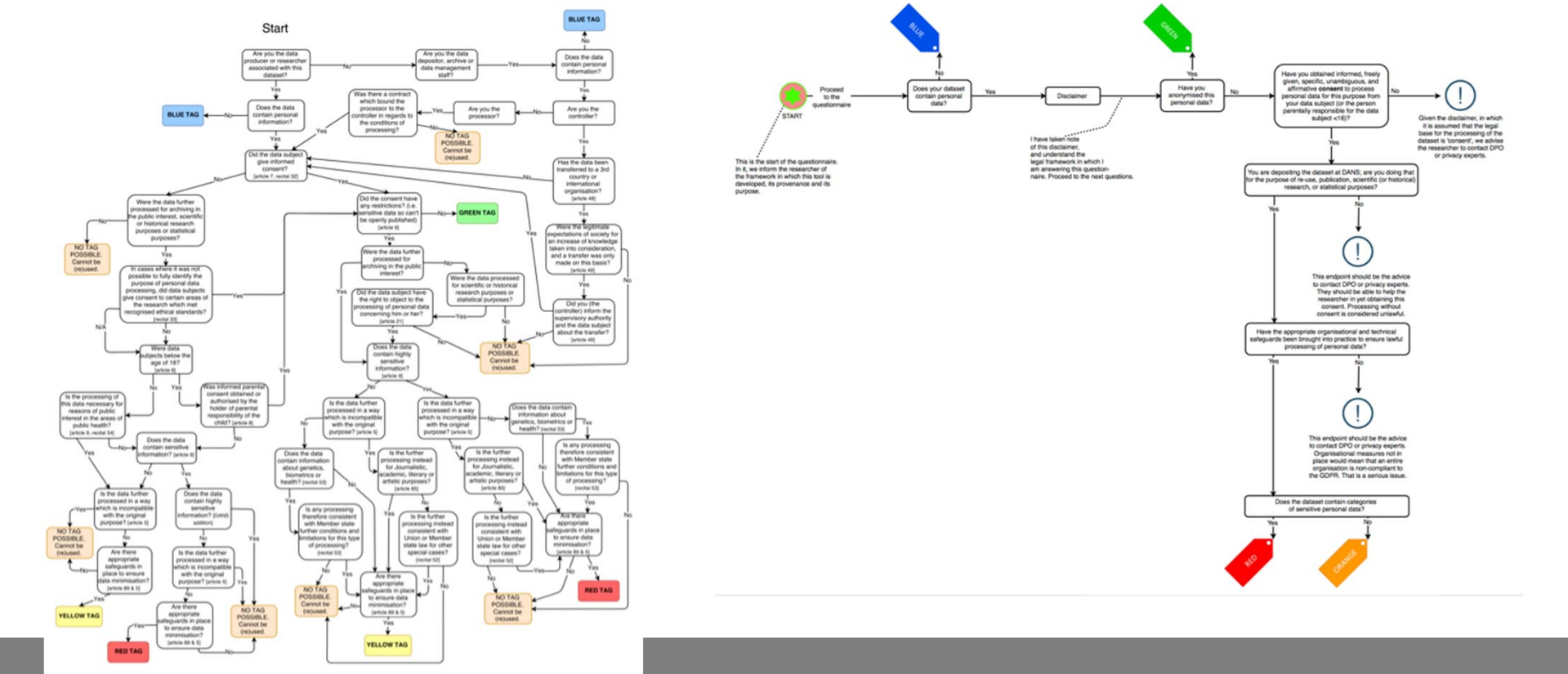
1. Identify the relevant articles of GDPR for research and archive purposes
  - a) Example: Article 9(2) sets out the circumstances in which the processing of sensitive personal data (which is otherwise prohibited) may take place:
    - *Necessary for archiving purposes in the public interest, or scientific and historical research purposes or statistical purposes in accordance with Article 89(1).*
  - b) Article 17 - right to be forgotten
2. Define access levels/tags for data protection
3. Create a decision tree
4. Transformation of relevant articles into questions
  - a) Were the data processed for archiving in the public interest, scientific or historical research purposes or statistical purposes?
  - b) Would you consider the dataset to contain sensitive personal information? [article 9]



# Decision tree evolution



- Creating routes for questions, ending with tags
- Deciding on tag options and recommendations following each route
- Tree diagram and feedback



# The questionnaire

## DANS Datatags prototype 2



### DANS Datatags question 2 of 6 - Anonymised data

Remember that the GDPR defines personal data as information referring to any identifiable or identified natural person. You might have decided to anonymise your dataset; you have *removed any information that identifies an individual*. There are certain guidelines and considerations for anonymisation. Please take note of [this website](#), where the Finnish Social Science Data Archive explains the concepts thoroughly.

DANS is interested in this, although the dataset would fall out of scope of the GDPR.

### Have you anonymised your dataset?

- > Yes, I have anonymised my dataset
- > No, I have not anonymised my dataset

← Back

Restart



## DANS Datatags prototype 2

### DANS Datatags - Endpoint: green tag

GREEN

You have indicated that your dataset contains anonymised personal data. Its dissemination level can therefore in principle be public.

← Back

Restart



# FAIR Data Assessment



## FAIR checklist

DSA Principles (for data repositories)	FAIR Principles (for data sets)
data can be <b>found</b> on the internet	Findable
data are <b>accessible</b>	Accessible
data are in a <b>usable format</b>	Interoperable
data are <b>reliable</b>	Reusable
data can be <b>referred to</b>	(citable)



WORLD DATA SYSTEM

Is your data



FAIR enough?



### Checklist to evaluate FAIRness of data(sets)

You would like to deposit one or several dataset(s) at a digital repository but you are not sure whether the information you provide is sufficient and in line with the principles of FAIR (Findable, Accessible, Interoperable, Reusable)? This checklist helps you assess the quality (FAIRness) of your dataset(s) and the trustworthiness of the repository that you have chosen.

- The assessment will cover four levels:
1. The data repository you are planning to use
  2. The metadata with which you describe your dataset
  3. The dataset itself
  4. The data files of which your dataset consists



Badging scheme

[www.nature.com/scientificdata](http://www.nature.com/scientificdata)

## SCIENTIFIC DATA

OPEN

### Comment: A design framework and exemplar metrics for FAIRness

Mark D. Wilkinson<sup>1</sup>, Susanna-Assunta Sansone<sup>2</sup>, Erik Schultes<sup>3</sup>, Peter Doorn<sup>4</sup>, Luiz Olavo Bonino da Silva Santos<sup>5,6</sup> & Michel Dumontier<sup>7</sup>

Received: 28 November 2017  
Accepted: 9 May 2018  
Published: 26 June 2018

The FAIR Principles<sup>1</sup> (<https://doi.org/10.25504/FAIRsharing.WW110U>) provide guidelines for the publication of digital resources such as datasets, code, workflows, and research objects, in a manner that makes them Findable, Accessible, Interoperable, and Reusable (FAIR). The Principles have rapidly been adopted by publishers, funders, and pan-disciplinary infrastructure programmes and societies. The Principles are aspirational, in that they do not strictly define how to achieve a state of "FAIRness", but

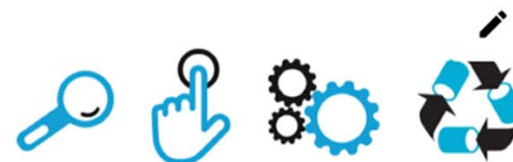
## FAIR Metrics

<http://fairmetrics.org>

# FAIR Data Reviews



FAIR Data Reviews  
for data in a trustworthy repository



## FAIR Data Review Form

### General quality of the data

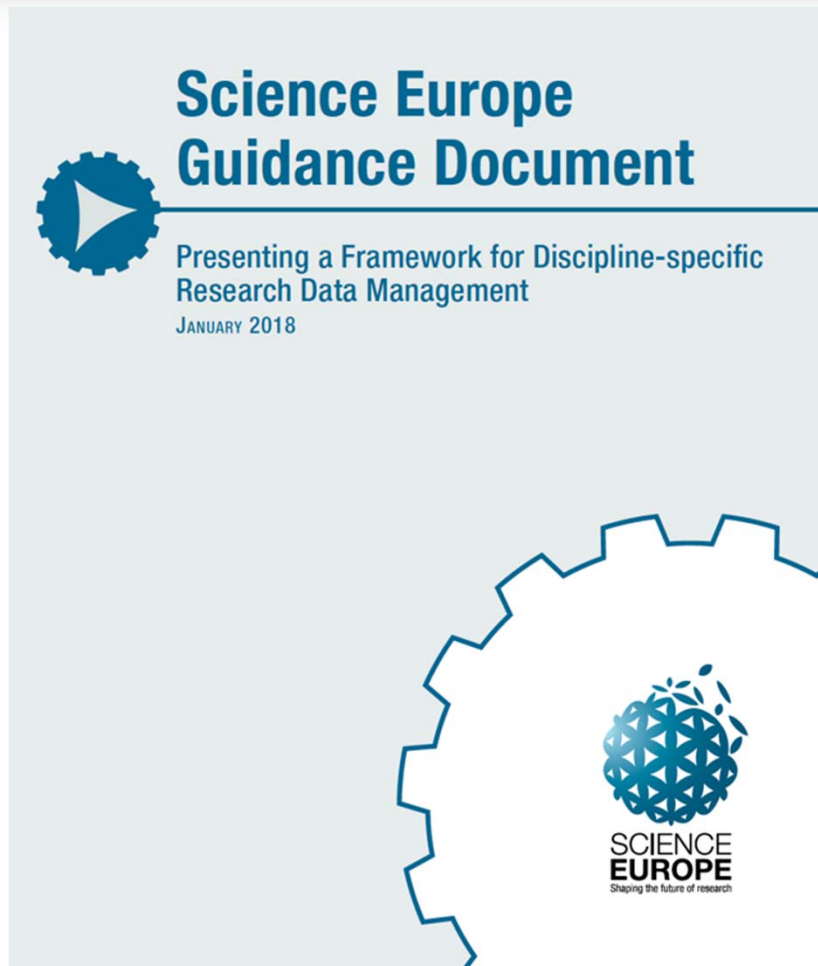
In this section some questions are posed about the completeness, precision/accuracy, fitness for use, structure and overall quality of the data

#### 1. How do you rate the completeness of the data?

The data is complete if no information that is needed to work with it is missing.

	1	2	3	4	5	
Very incomplete	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very complete

# Research Data Management Requirements



# Aligned requirements and simplified Data Management Plan process

## CORE REQUIREMENTS FOR DATA MANAGEMENT PLANS



When developing solid data management plans, researchers are required to deal with the following topics and answer the following questions:

- 1. Data description and collection or re-use of existing data**
  - a. How will new data be collected or produced and/or how will existing data be re-used?
  - b. What data (for example the kinds, formats, and volumes) will be collected or produced?

---

- 2. Documentation and data quality**
  - a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany data?
  - b. What data quality control measures will be used?

---

- 3. Storage and backup during the research process**
  - a. How will data and metadata be stored and backed up during the research process?
  - b. How will data security and protection of sensitive data be taken care of during the research?

---

- 4. Legal and ethical requirements, codes of conduct**
  - a. If personal data are processed, how will compliance with legislation on personal data and on data security be ensured?
  - b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?
  - c. How will possible ethical issues be taken into account, and codes of conduct followed?



## Domain Data Protocols

- to be formulated by research communities
- to be endorsed by research funders
- principle: comply or explain
- reduces need for individual data management plans
- simplifies evaluation of DMPs by funders



**Thanks for  
listening!**

[www.dans.knaw.nl](http://www.dans.knaw.nl)

[peter.doorn@dans.knaw.nl](mailto:peter.doorn@dans.knaw.nl)