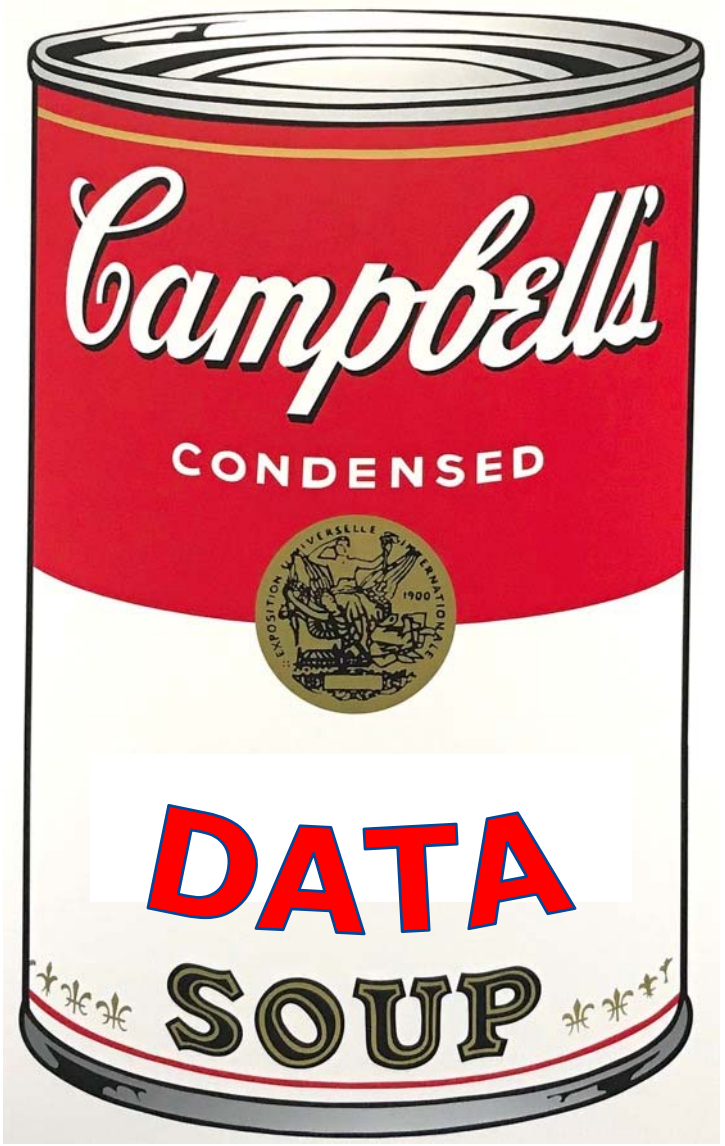
DANS

Data Archiving and Networked Services

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



Peter Doorn, Data Archiving and Networked Services (DANS)

twitter @pkdoorn @dansknaw

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

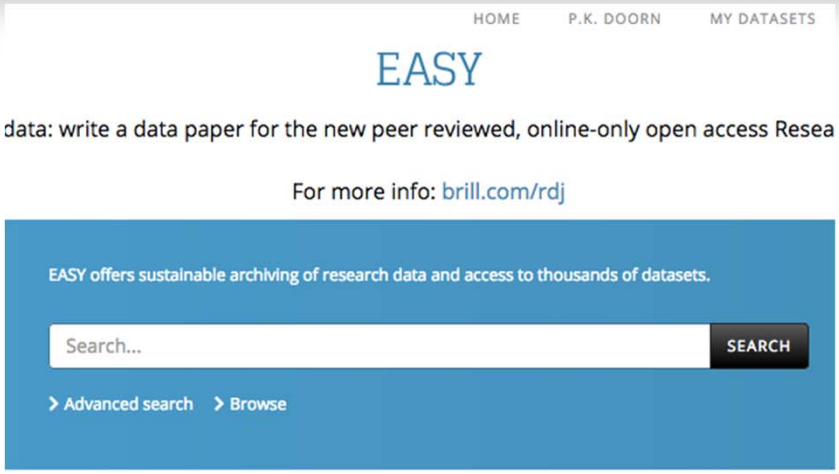
https://www.piql.com/arctic-world-archive/

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

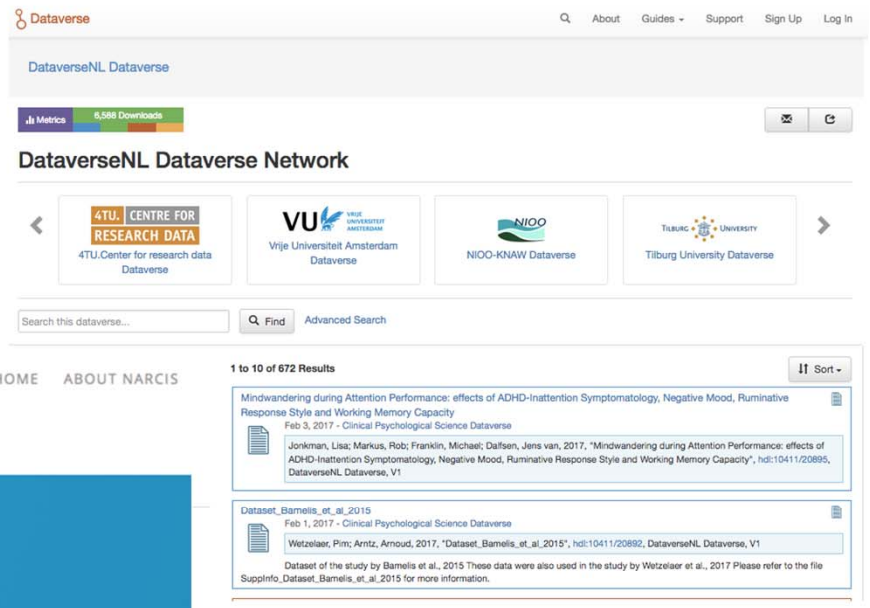
- How DANS archives data for long-term reuse
- Open access to data
- Restrictions to data access
- Big Data and the Long Tail of Data
- FAIR data and Research Data Management

DANS is about keeping data FAIR

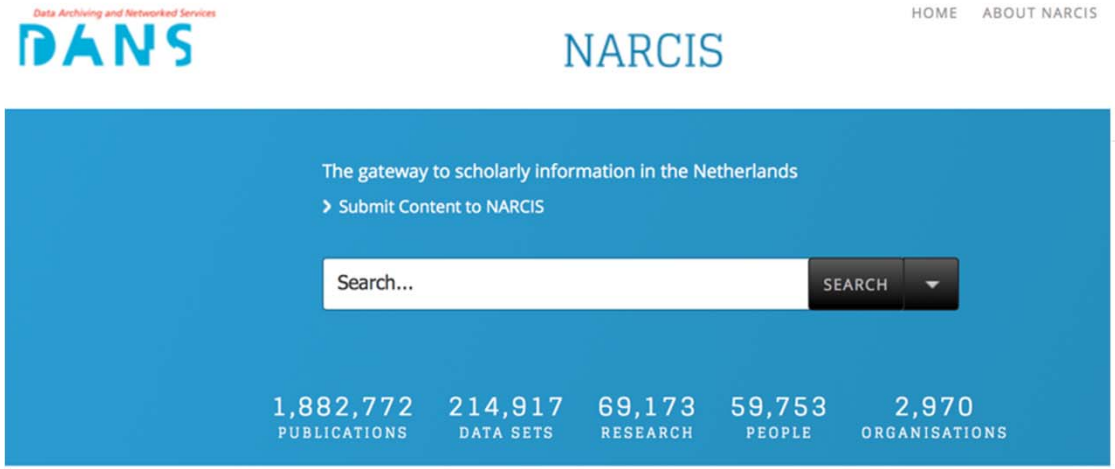
DANS core data services



EASY: certified Electronic Archiving System for self-deposit



DataverseNL: data repository at universities and other institutions



Data Archiving and Networked Services
DANS

NARCIS: Gateway to scholarly information In the Netherlands

https://easy.dans.knaw.nl
https://dataverse.nl
https://www.narcis.nl

Additional services

MENDELEY DATA Beta

Background Archive

DRYAD

ELSEVIER

Training & Consultancy

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

http://datasupport.researchdata.nl/

Work in Progress: Software Archive

netherlands eScience center

Inria

https://www.softwareheritage.org/

Research Data Journal for the Humanities and Social Sciences

TUTA SUB AEGIDE PALLAS · 1683 ·

BRILL

http://www.brill.com/rdj

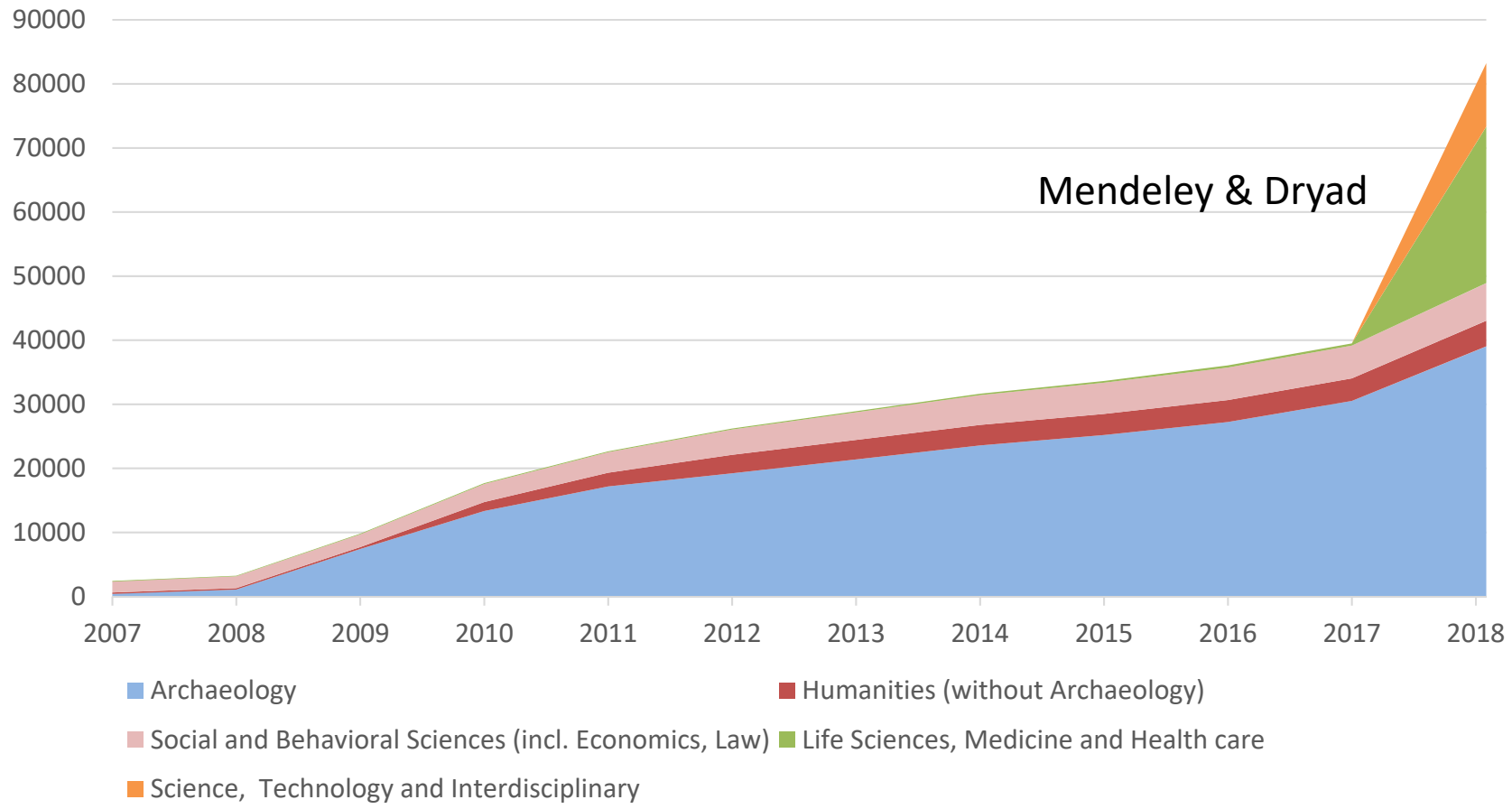
What do we have to offer?

- 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

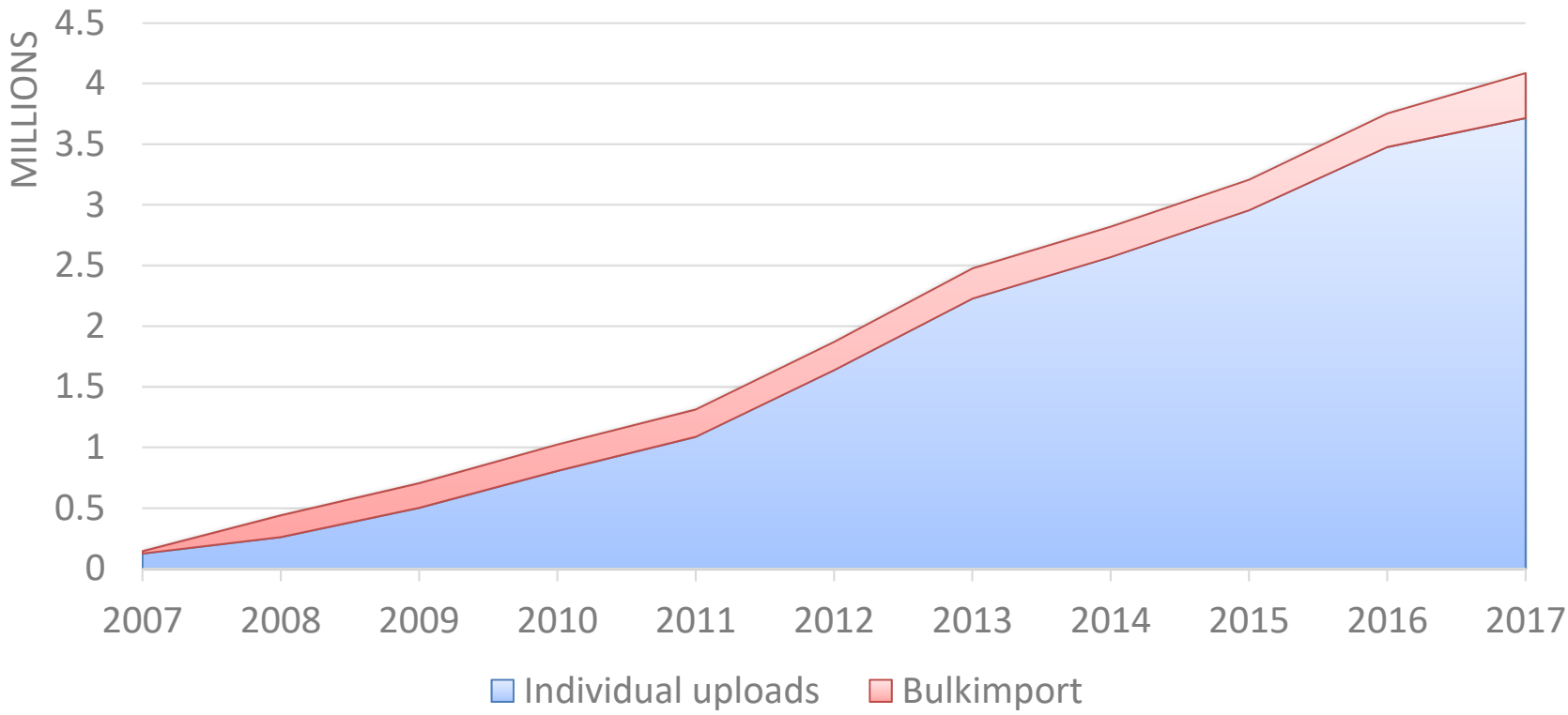
Most of our services are free for individual researchers.
Exceptions:

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

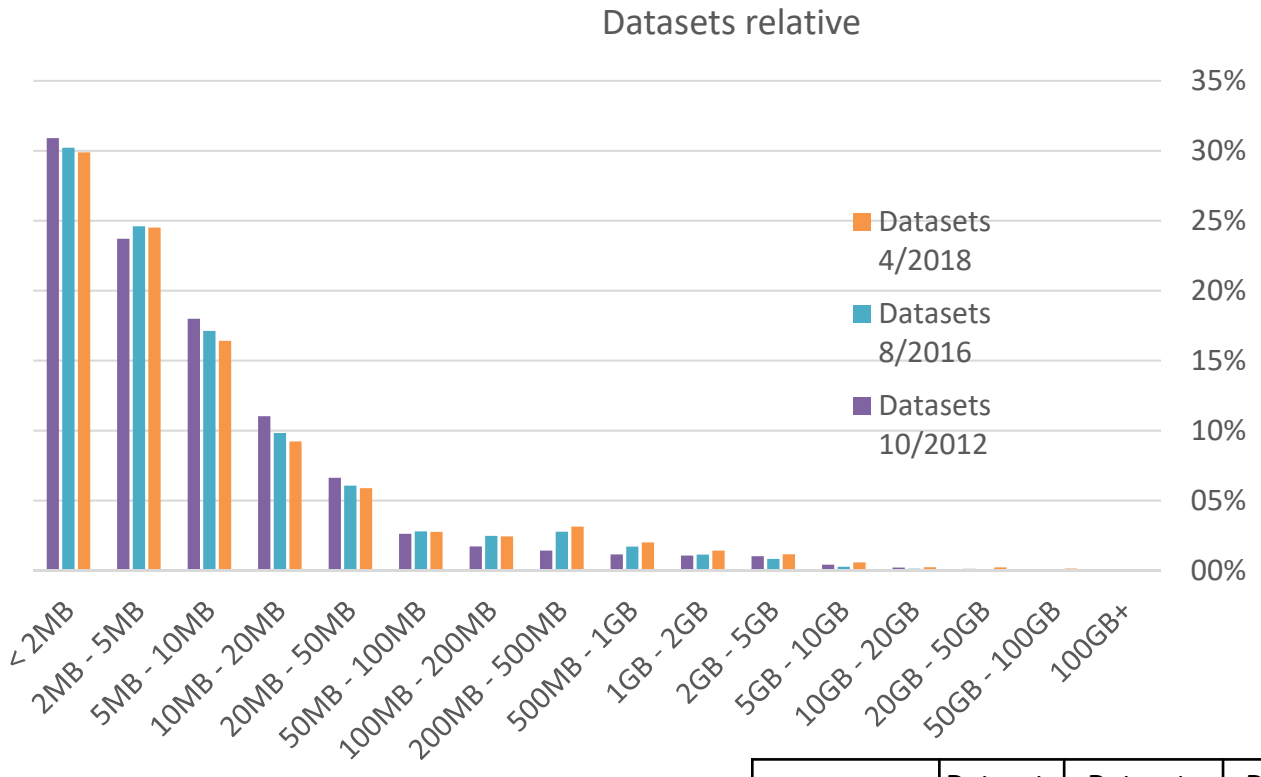
Datasets in EASY per discipline, 2007-2018

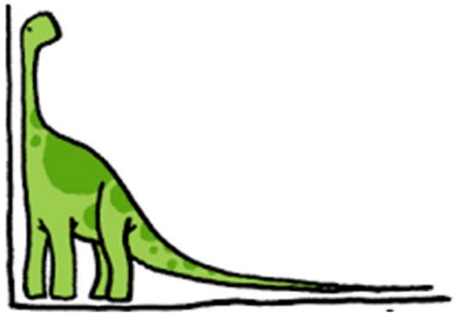


Growth of number of Files in EASY, 2007-2017



Datasets in DANS EASY archive according to size

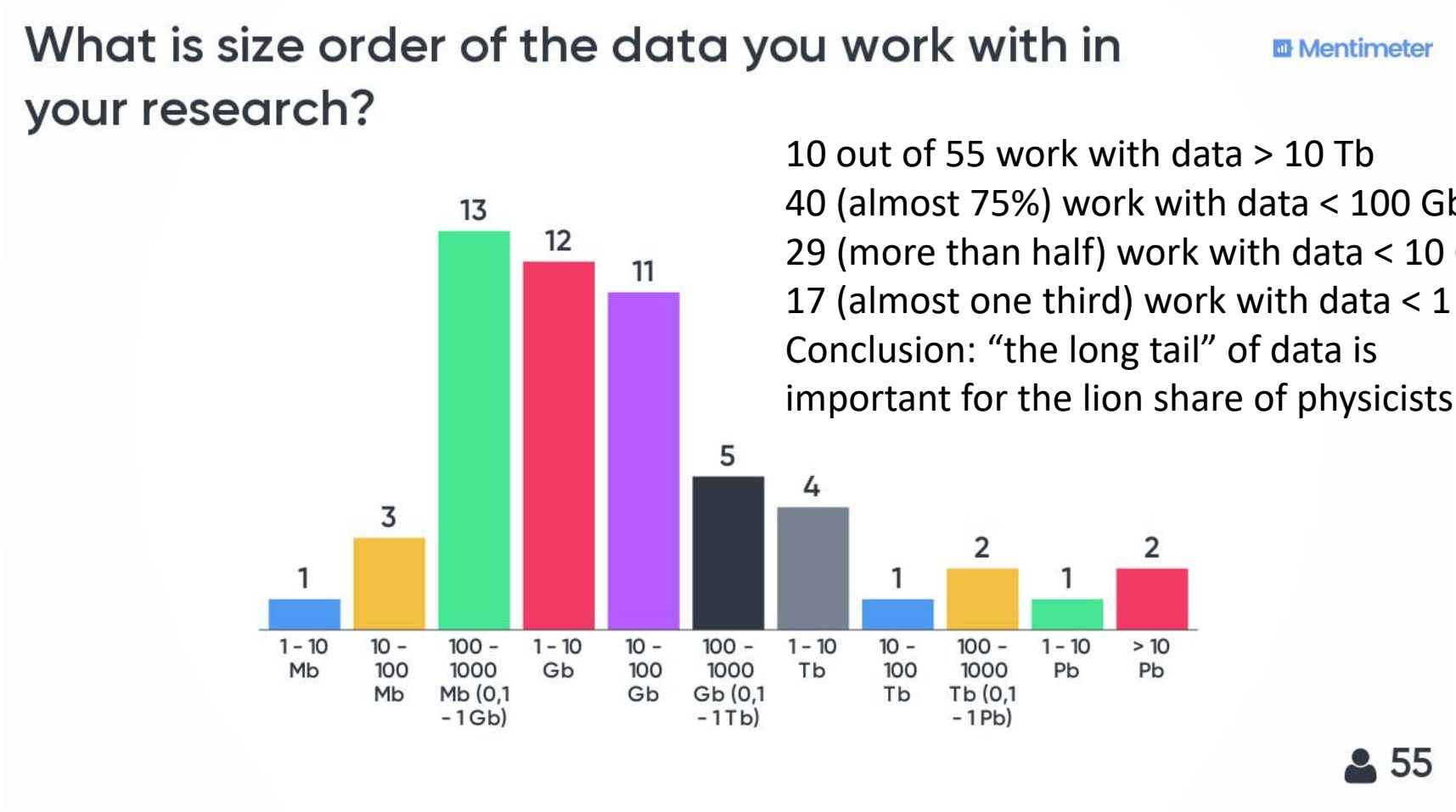




The long tail of research data

	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?

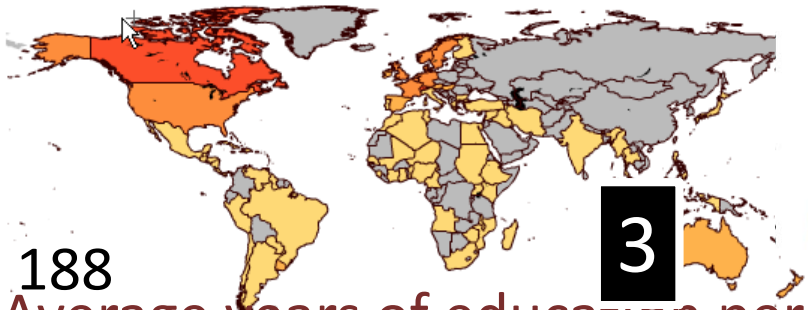


10 out of 55 work with data > 10 Tb
40 (almost 75%) work with data < 100 Gb
29 (more than half) work with data < 10 Gb
17 (almost one third) work with data < 1 Gb
Conclusion: “the long tail” of data is important for the lion share of physicists

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

Collaborative work: bringing together data from many scholars

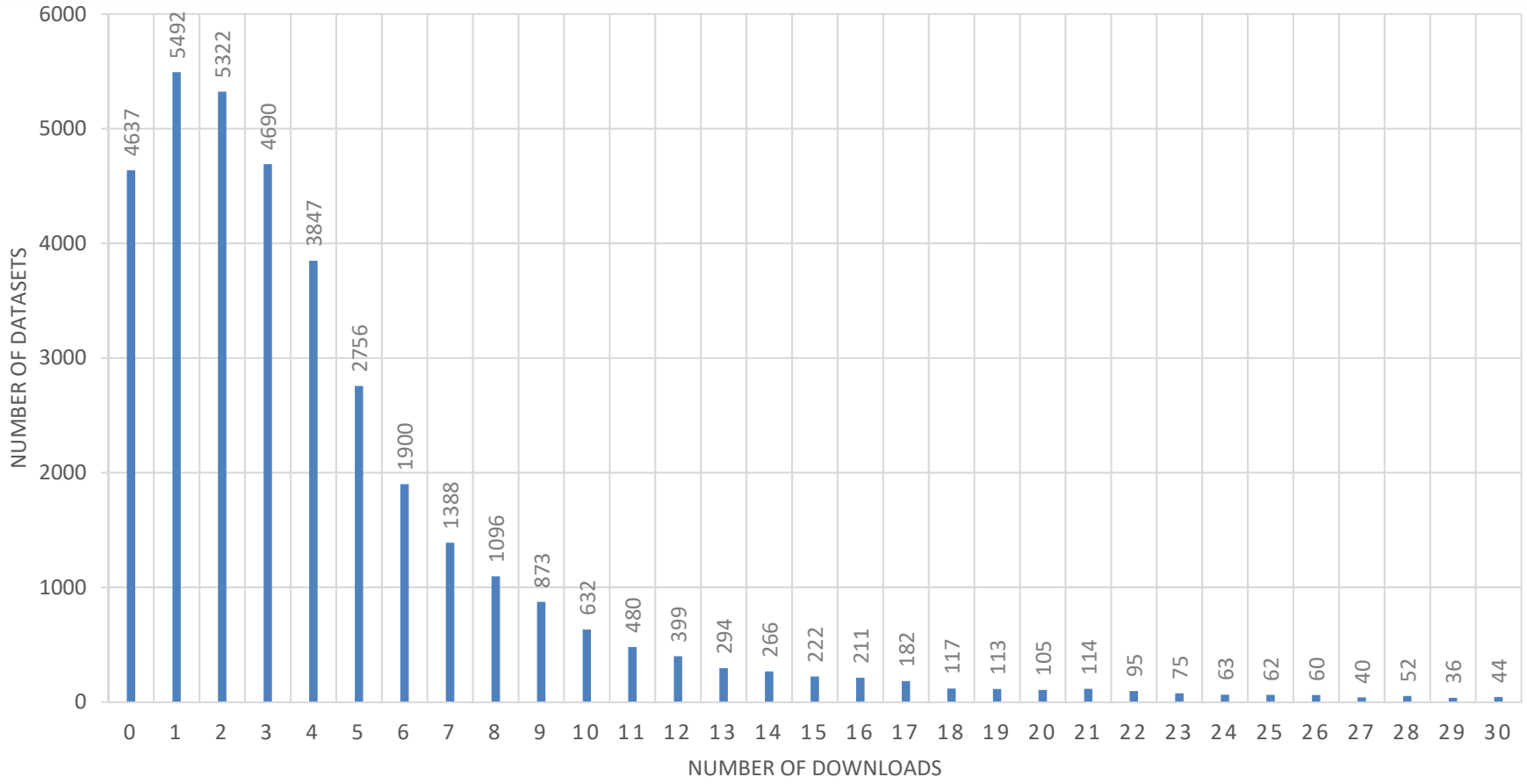
1. Historical shipping
2. Digitized censuses
3. Global inequality
4. Holocaust studies
5. Dendrochronology



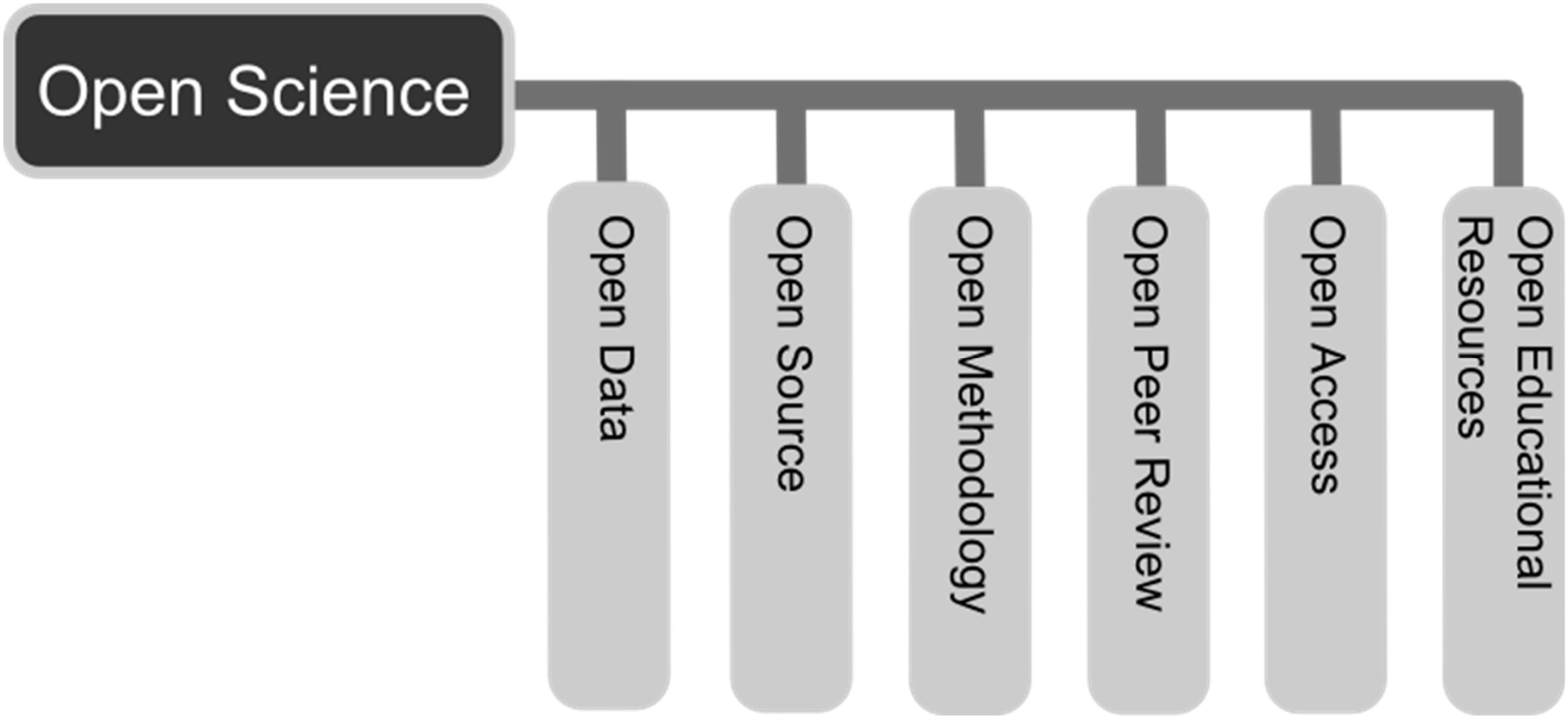
Top 10 of downloaded datasets from EASY

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

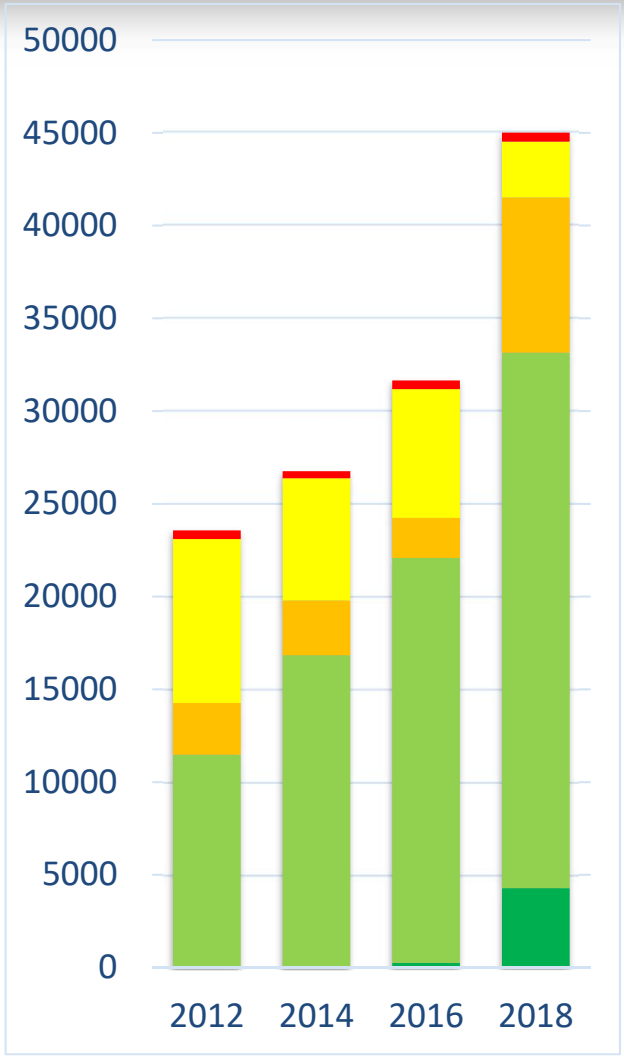
Frequency of 0-30 dataset downloads, 2007-2017

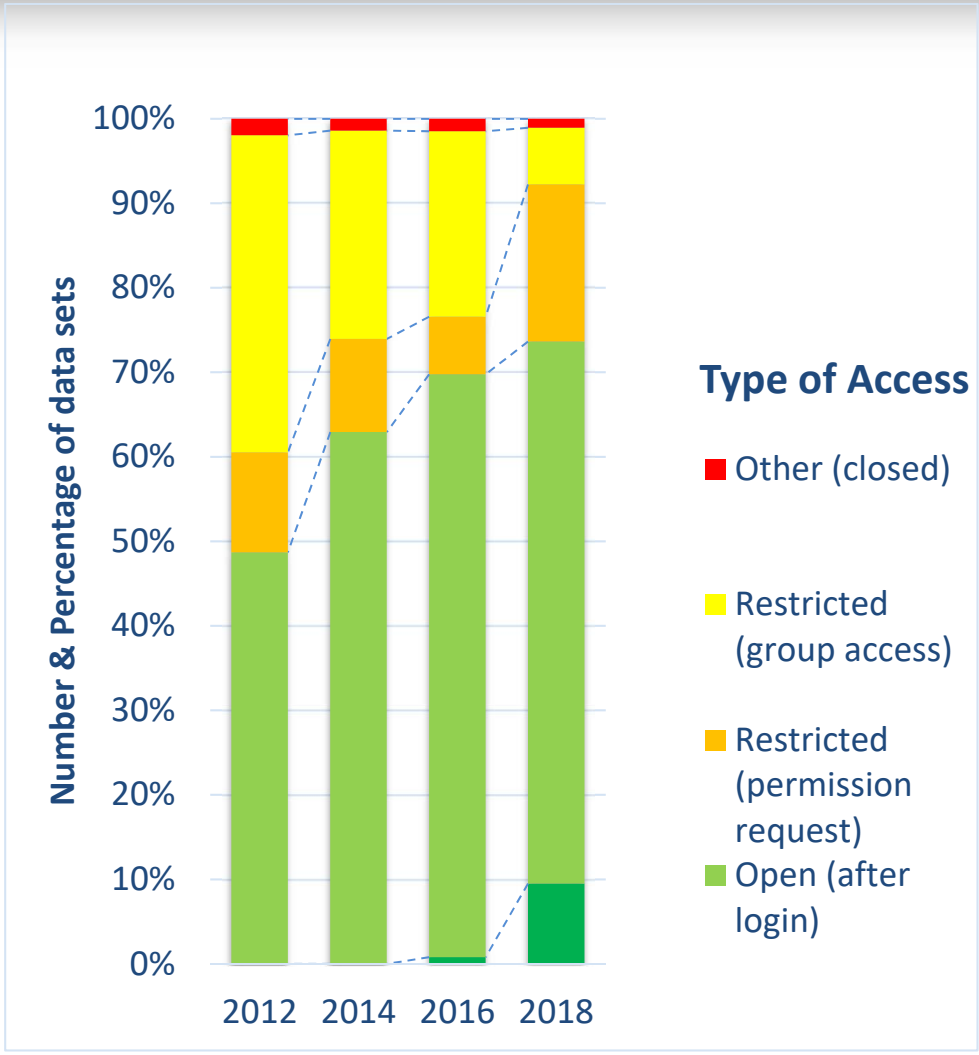


Six Principles of Open Science

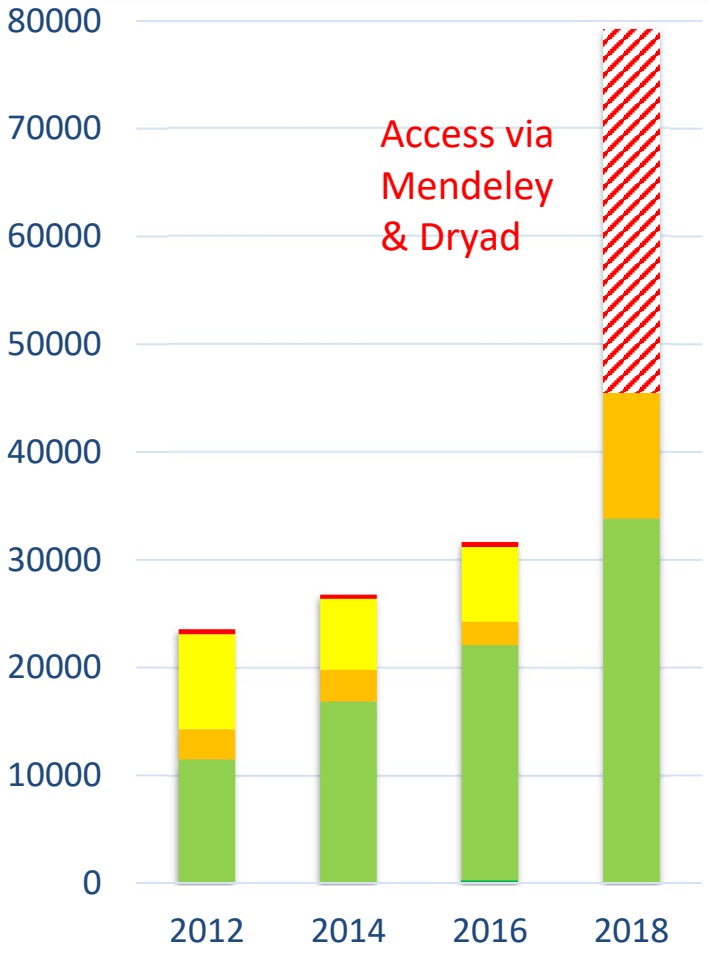


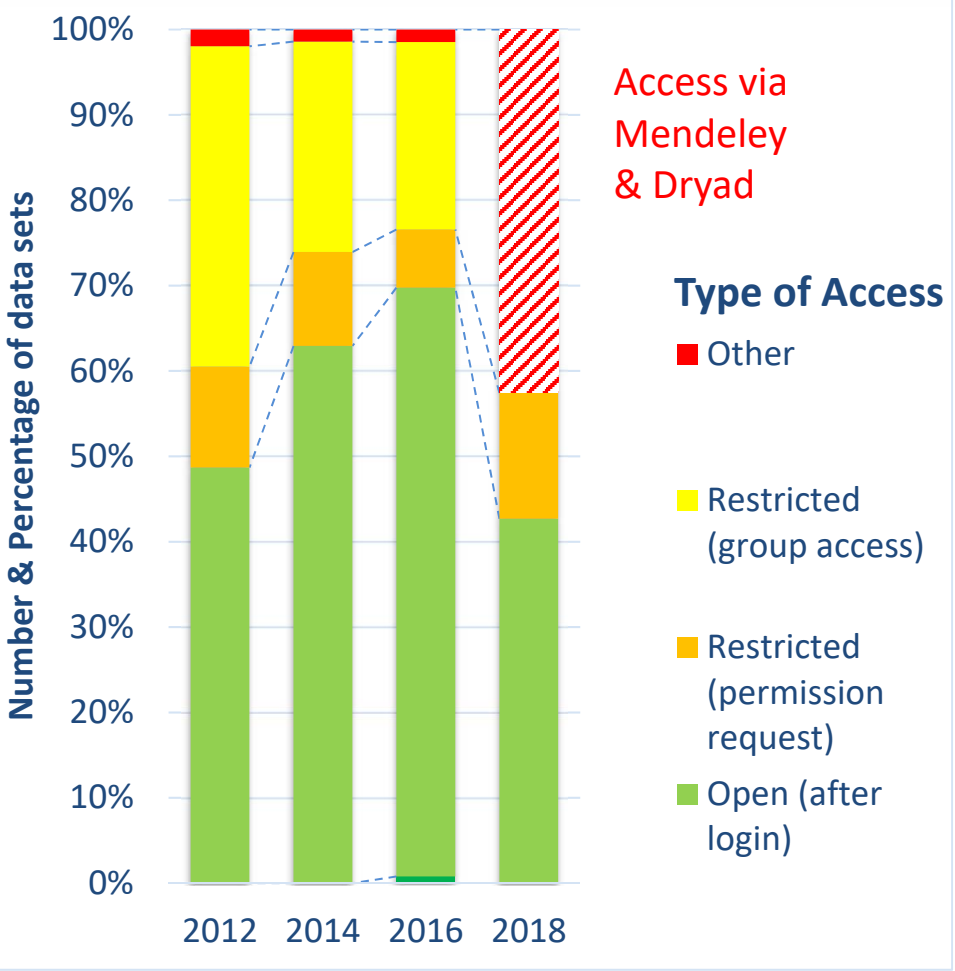
Access to Datasets in DANS archive 2012-2018 (without Mendeley & Dryad)



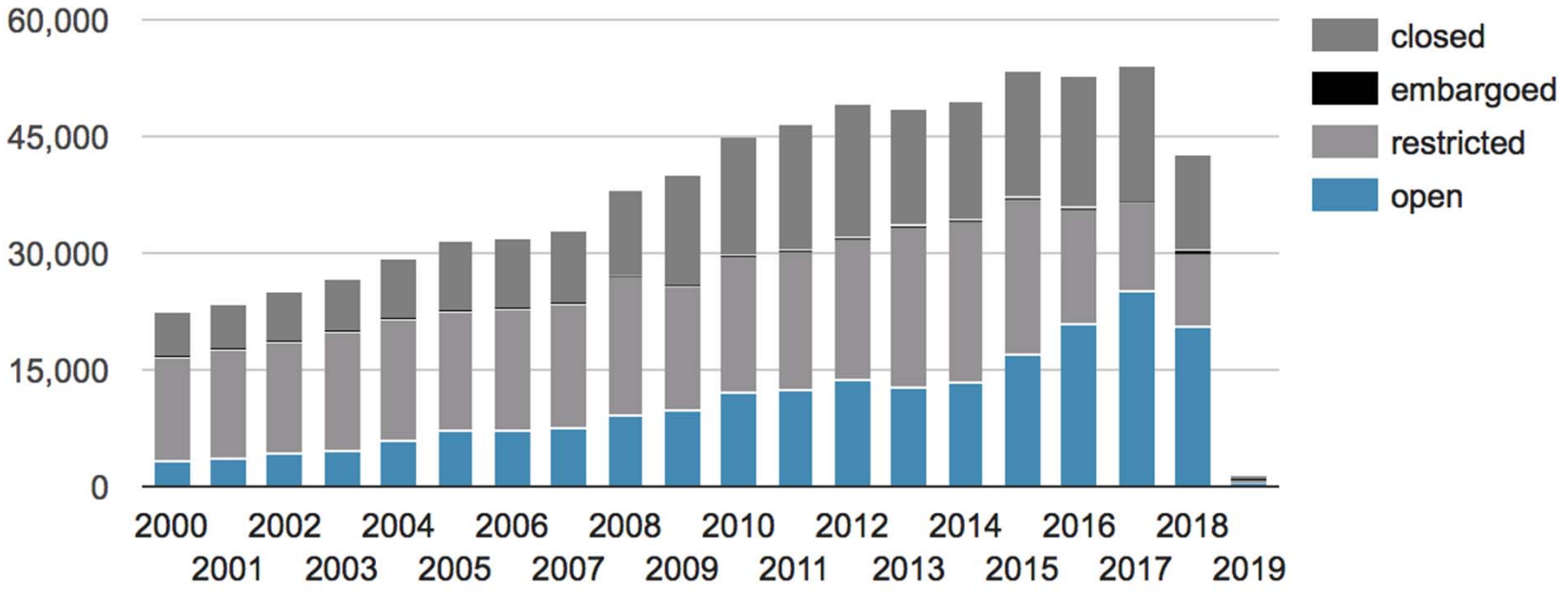


Access to Datasets in DANS archive 2012-2018 (including Mendeley & Dryad)

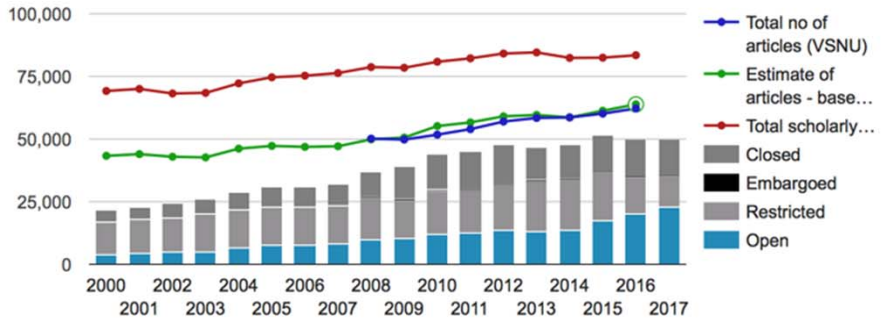




Open and Closed Access Articles in NARCIS per year of publication, 2000-now



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



Data Archiving and Networked Services
DANS

https://www.narcis.nl/metrics/

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

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

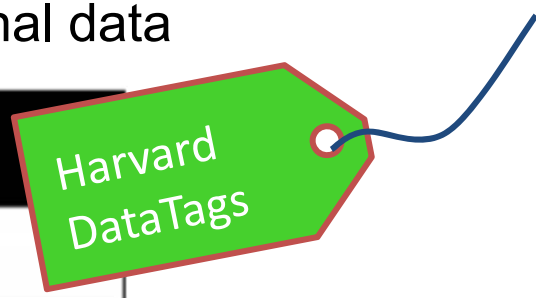
→ Data Tagging Approach, initially developed at Harvard

Background of DataTags approach @ Harvard

Sweeney & Crosas introduced the notion of a datatags repository

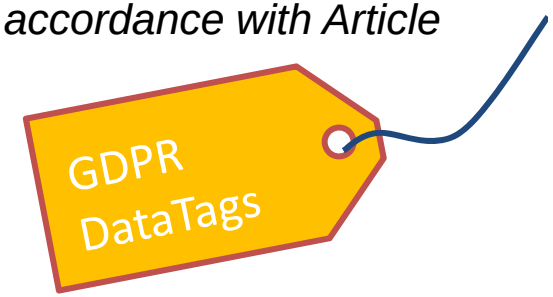
- Stores and shares data files in accordance with different security levels, access requirements and usage agreements
- Based on American laws and legislations of personal data

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



Our approach: Step by step

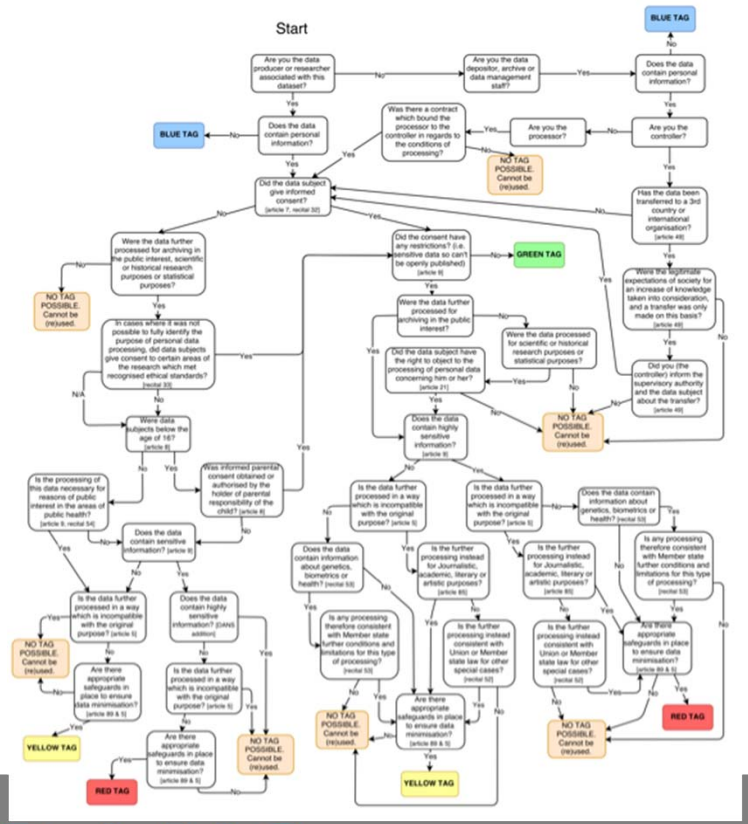
1. Identify the relevant articles of GDPR for research and archive purposes
   a) Example: Article 9(2) sets out the circumstances in which the processing of sensitive personal data (which is otherwise prohibited) may take place:
      - *Necessary for archiving purposes in the public interest, or scientific 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]

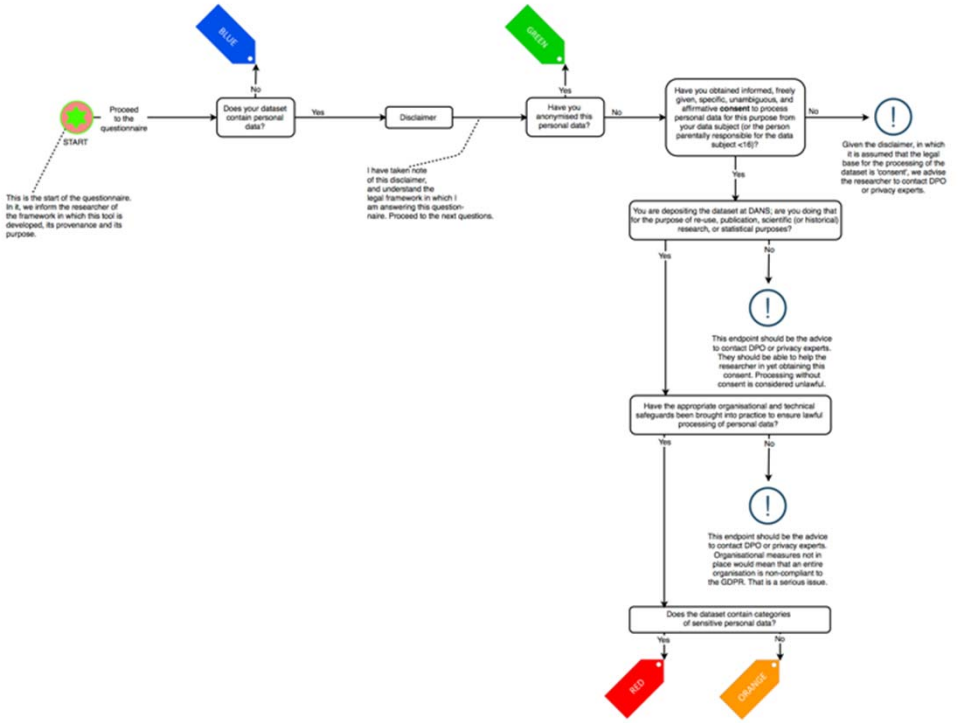


Decision tree evolution

GDPR DataTags

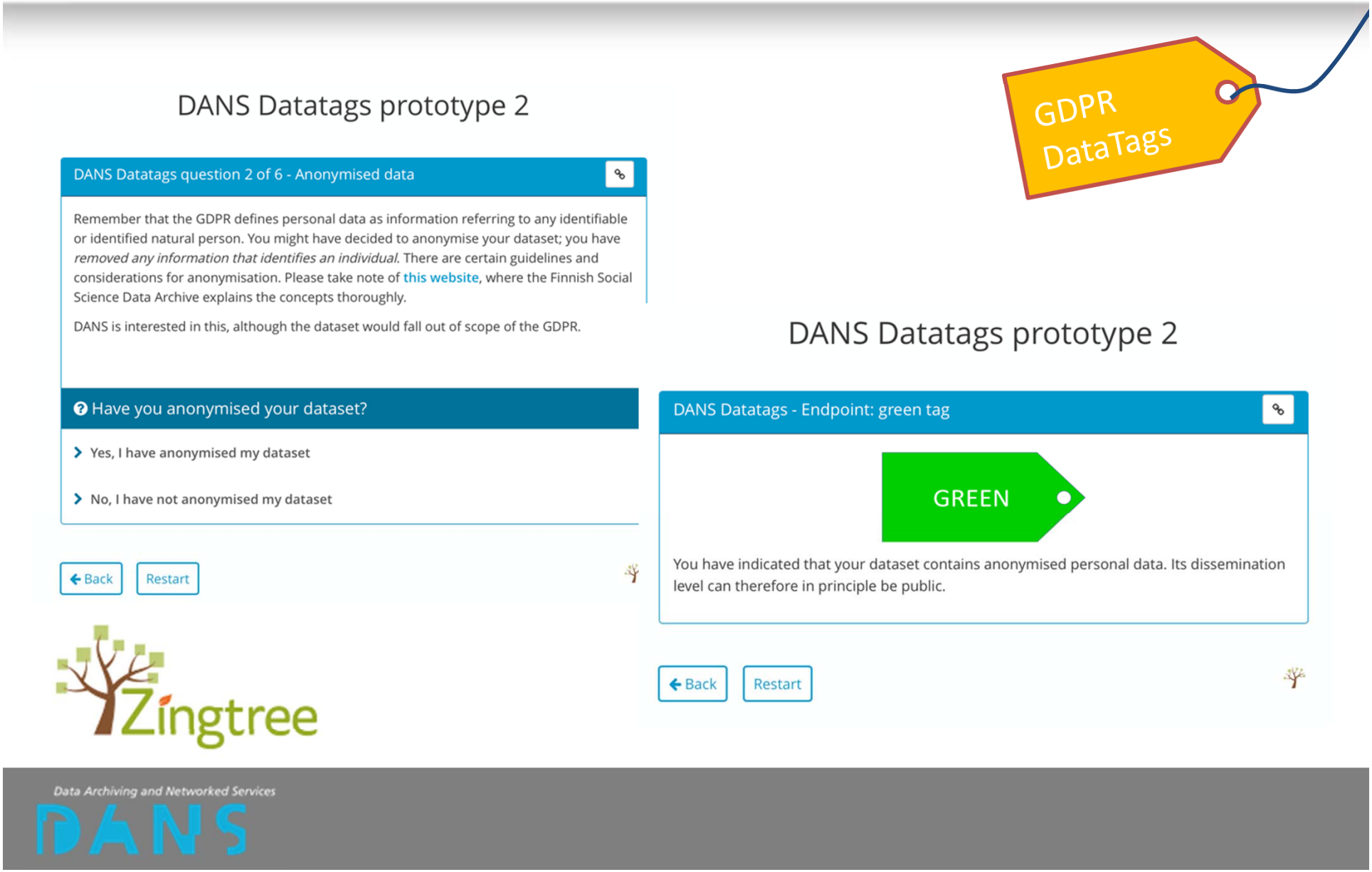
- Creating routes for questions, ending with tags
- Deciding on tag options and recommendations following each route
- Tree diagram and feedback





DANS

The questionnaire



FAIR Data Assessment

https://www.coretrustseal.org

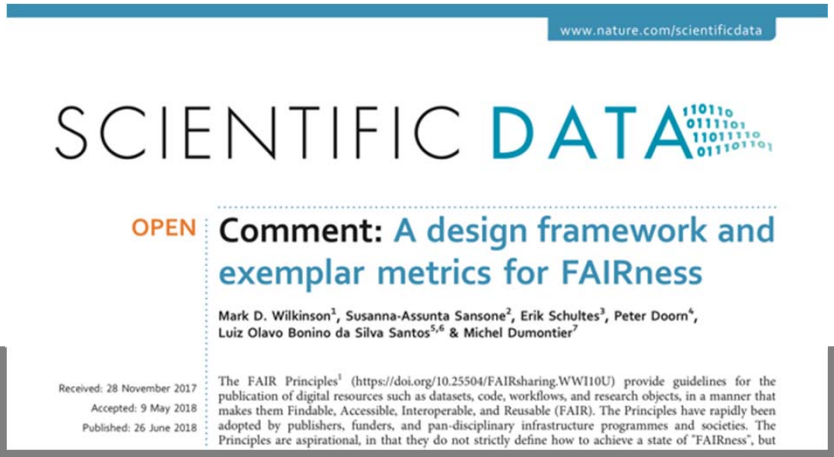
DSA Principles (for data repositories)	FAIR Principles (for data sets)
data can be **found** on the internet	**F**indable
data are **accessible**	**A**ccessible
data are in a **usable format**	**I**nteroperable
data are **reliable**	**R**eusable
data can be **referred** to	(citable)

Data Seal of Approval

ICSU WORLD DATA SYSTEM

FAIR checklist

Badging scheme

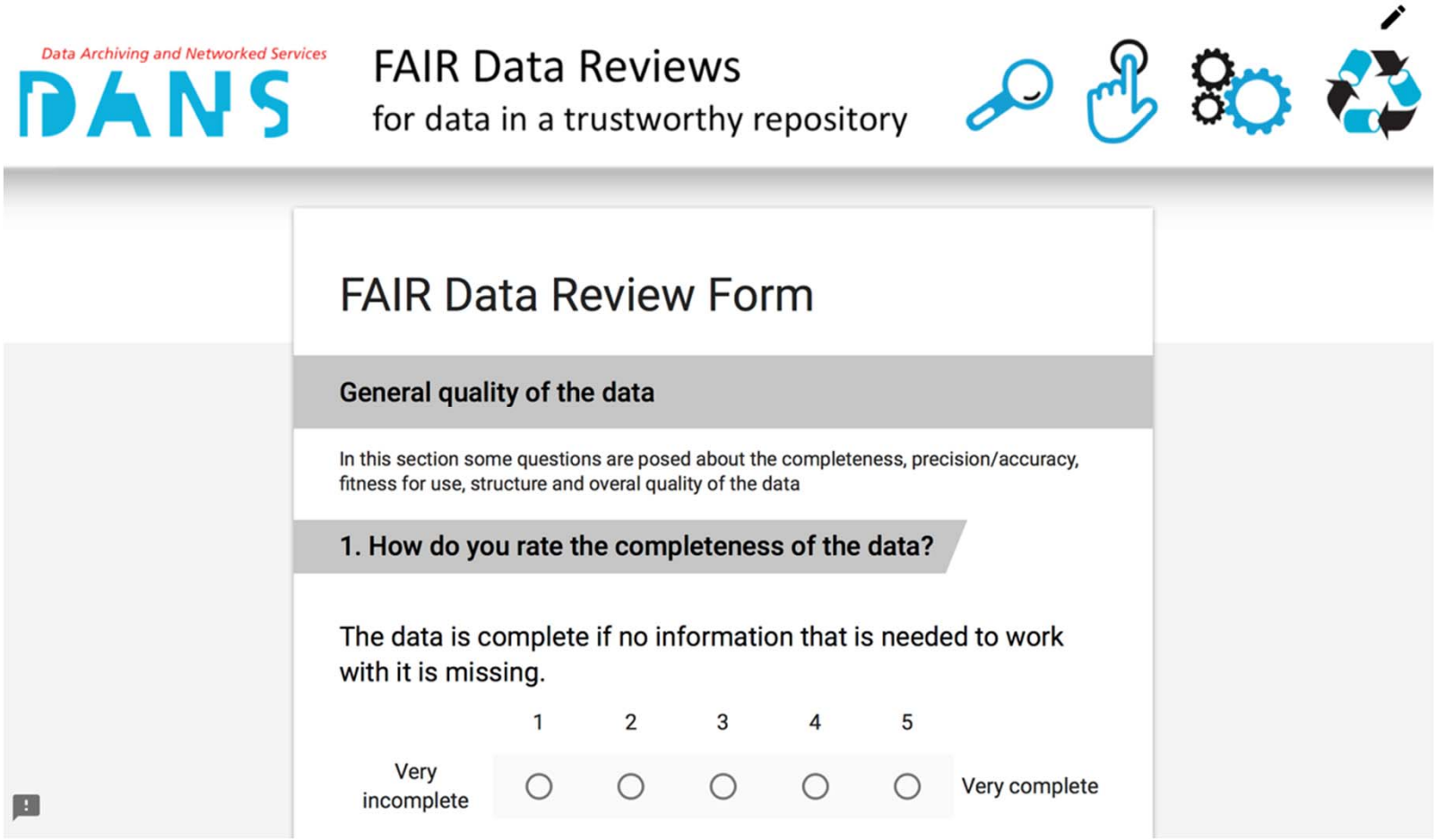


FAIR Metrics

Data Archiving and Networked Services DANS

http://fairmetrics.org

FAIR Data Reviews

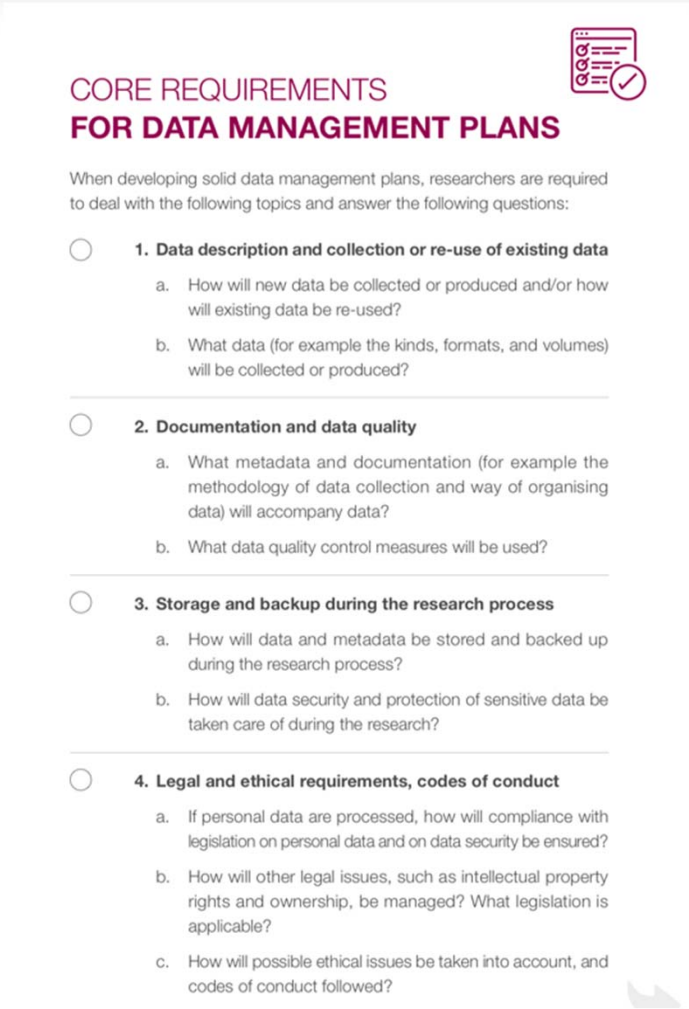


Research Data Management Requirements

Data Archiving and Networked Services
DANS

https://www.scienceeurope.org/policy/policy-areas/research-data/

Aligned requirements and simplified Data Management Plan process



CORE REQUIREMENTS
FOR DATA MANAGEMENT PLANS

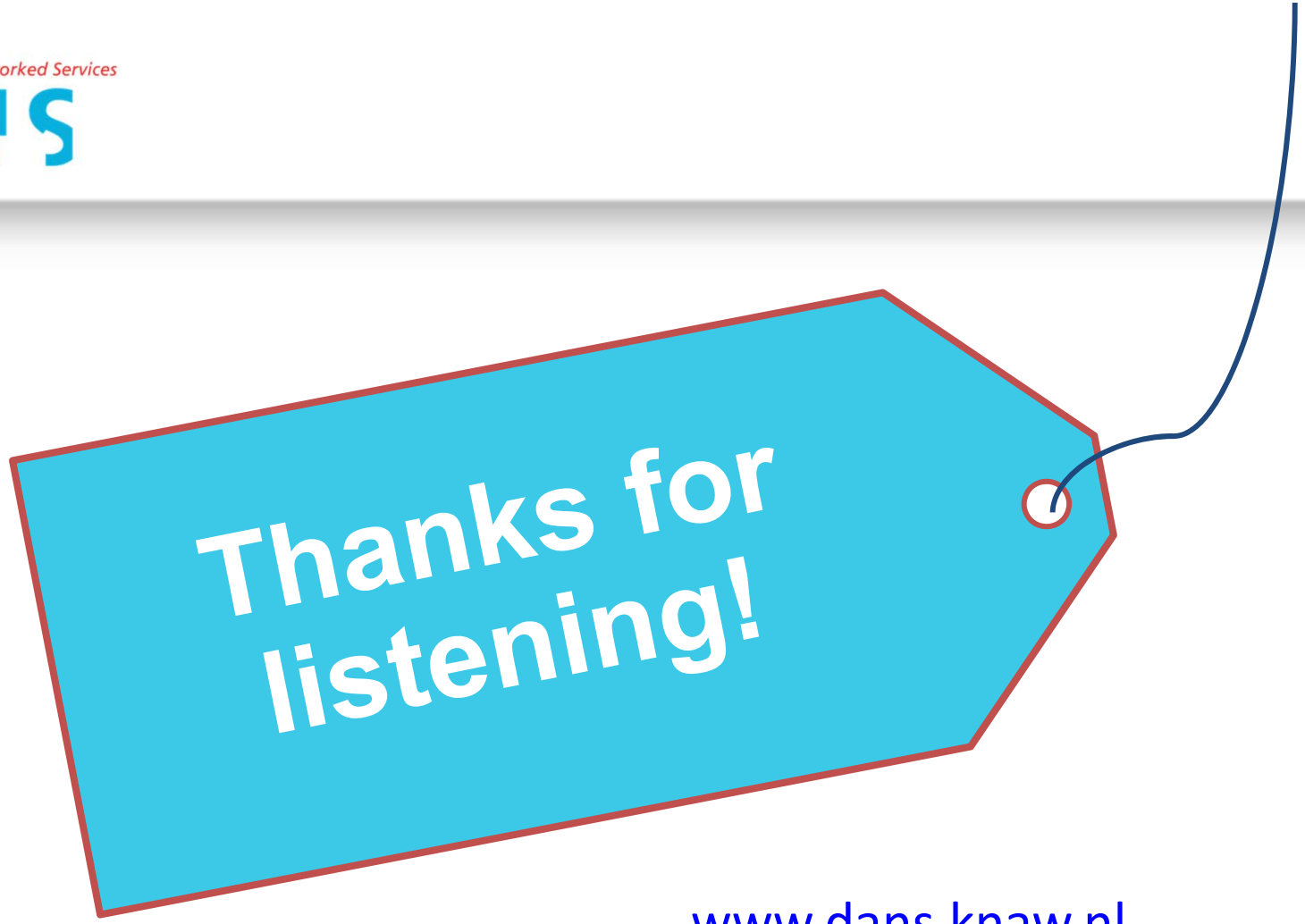
When developing solid data management plans, researchers are required to deal with the following topics and answer the following questions:

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

Data Archiving and Networked Services
DANS



www.dans.knaw.nl
peter.doorn@dans.knaw.nl