Marjan Grootveld

Data Archiving and Networked Services - DANS
The Netherlands

@openaire_eu
@MarjanGrootveld
@DANSKNAW

“But can I trust your data?”

Swiss Research Data Day 2018,
ETH Zürich, 12 June 2018

OpenAIRE

“No progress without reuse”

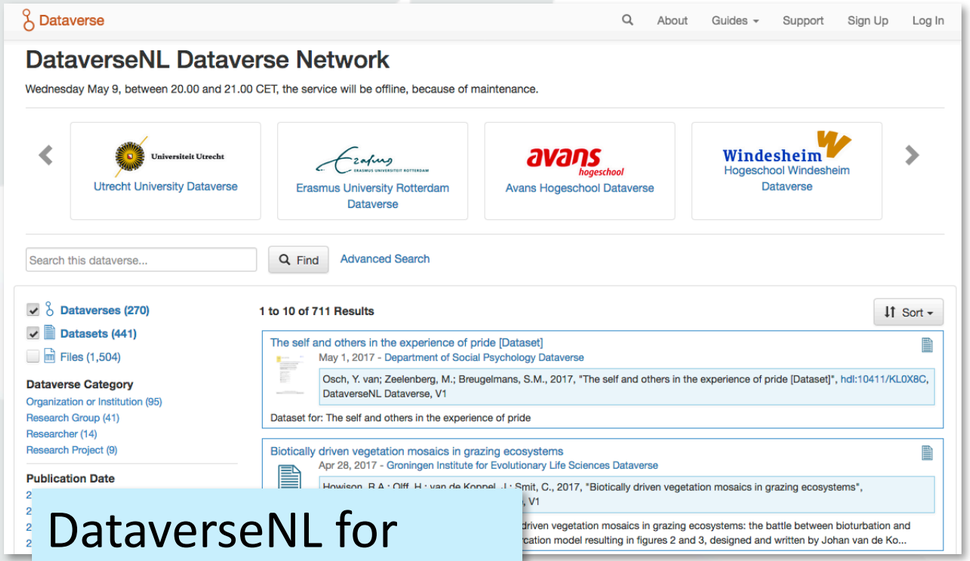
- **The potential of existing data**
- **The importance of documentation**
- **The support of trustworthy repositories**

What is DANS?

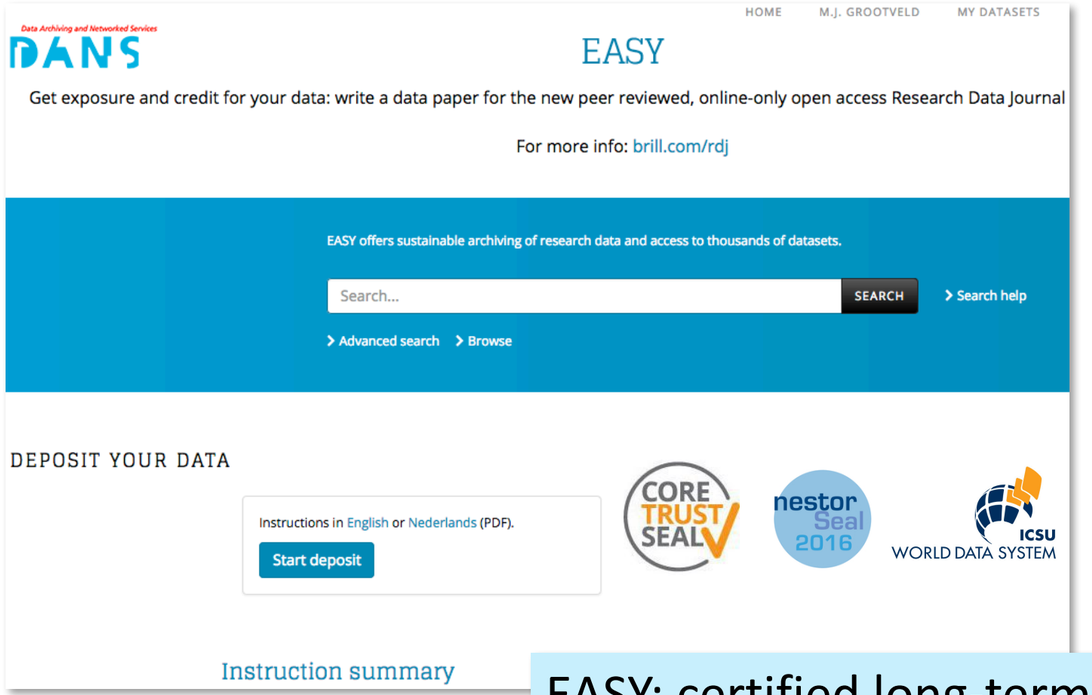
Mission: promote and provide permanent access to digital research resources

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

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



DataverseNL for short- and mid-term data storage



EASY: certified long-term Electronic Archiving System for self-deposit

NARCIS: Gateway to scholarly information in the Netherlands

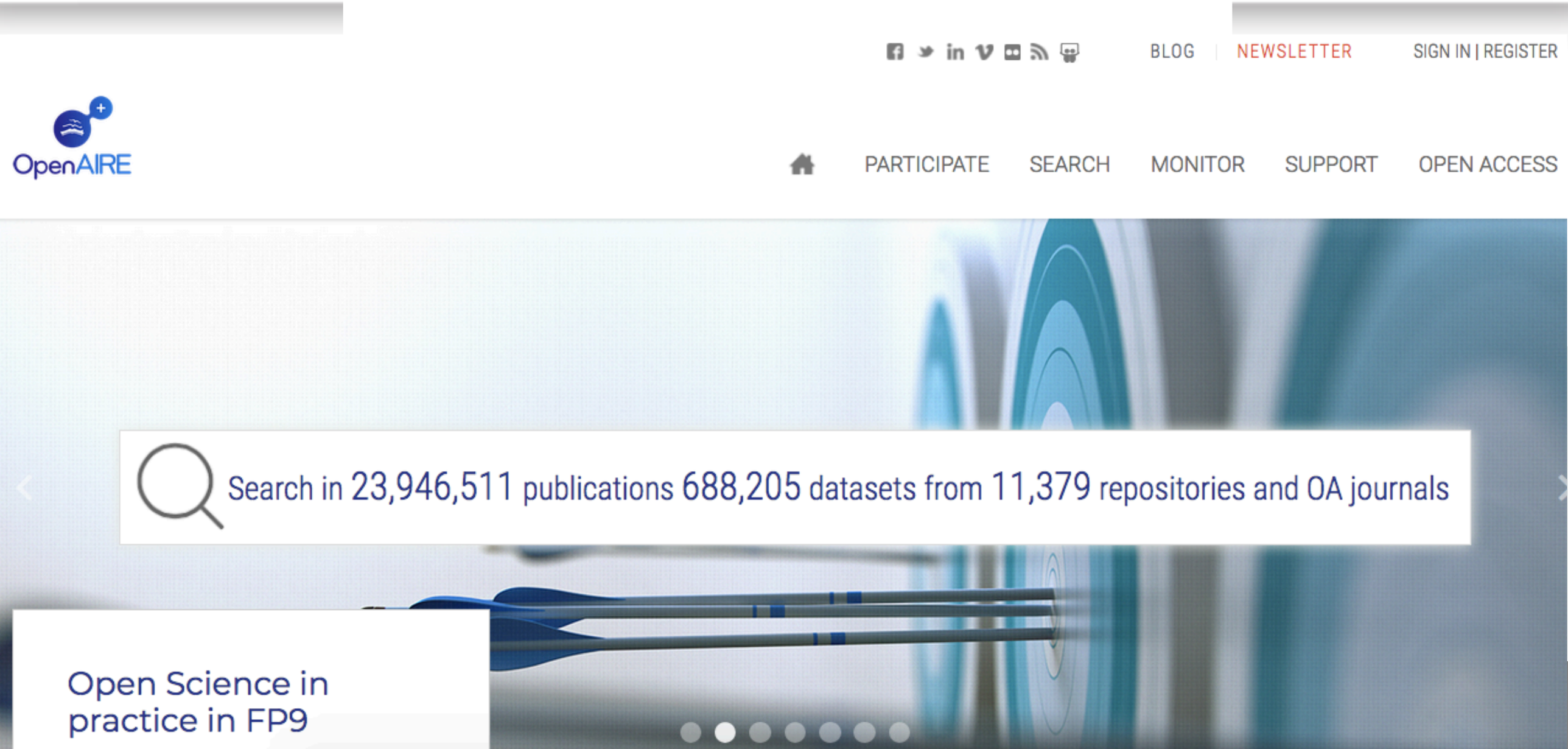
OpenAIRE

- Open Access Infrastructure for Research in Europe
- Funded by Horizon2020 to develop and maintain the infrastructure to support OA policy of the EU
- Supports H2020 OA mandates
  - 100% OA on scientific publications
  - Open Research Data Pilot
- A National Open Access Desk in each country
- 2018 – 2020: OpenAIRE Advance
- DANS leads taskgroup Research Data Management

SWITZERLAND

ANDRE HOFFMANN

OpenAIRE Network: www.openaire.eu



OpenAIRE RDM support (selection!)

- Briefing papers, factsheets, webinars, workshops, FAQs (with example DMPs)

- Information on:
  - Open Research Data Pilot
  - Creating a data management plan
  - Selecting a data repository

- Anonymisation tool *Amnesia*

https://www.openaire.eu/webinars/

https://www.openaire.eu/what-is-the-open-research-data-pilot

https://www.openaire.eu/support

https://www.openaire.eu/support/faq

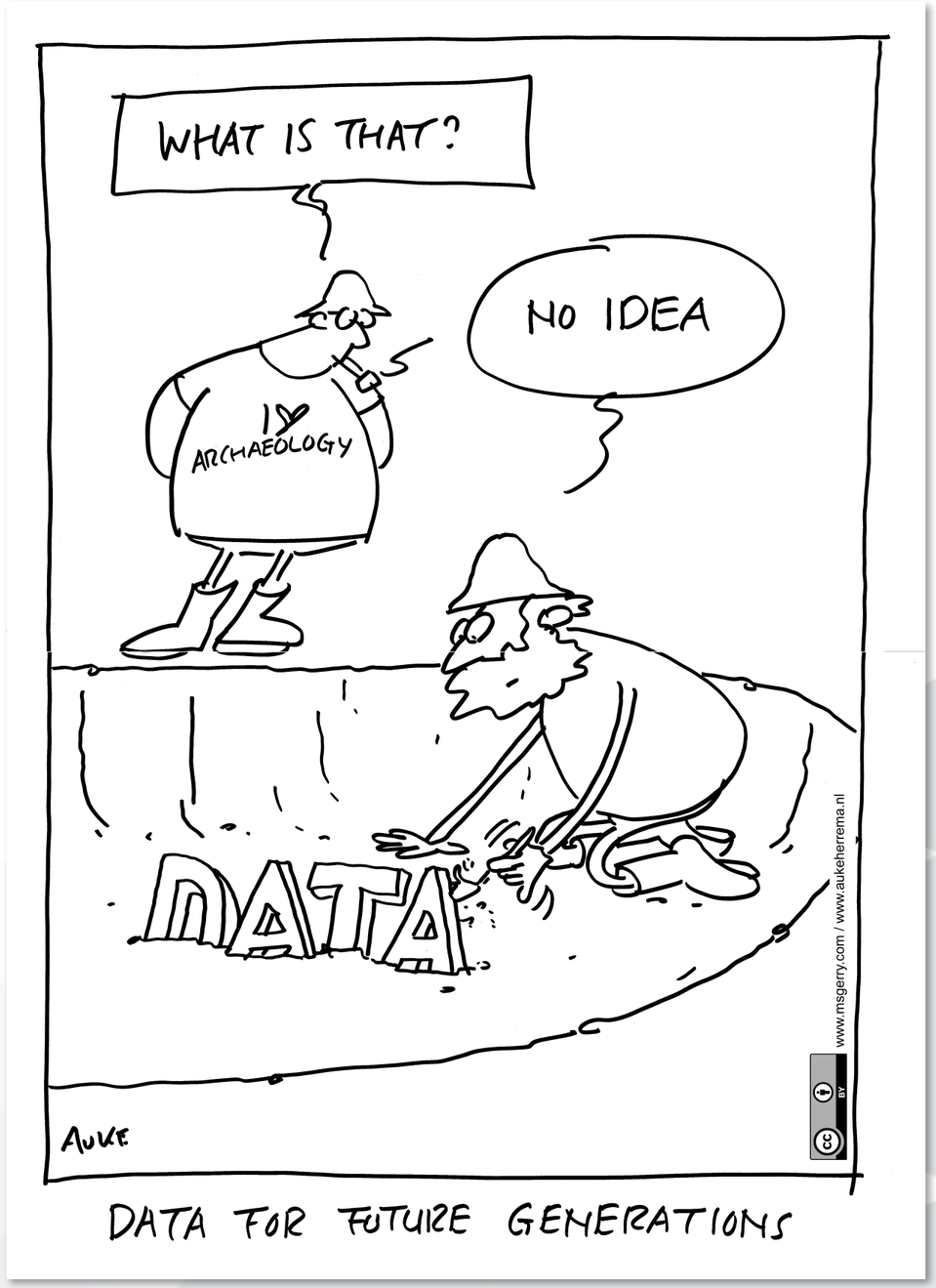
https://amnesia.openaire.eu/

European Commission

QUESTION:

Who has ever used data that was generated or collected by someone else?

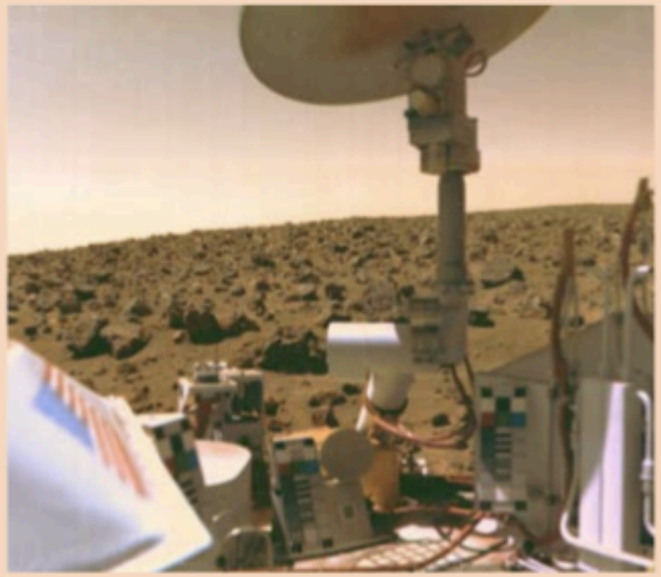
What is the *oldest* data that you have used?



Case study: Viking Lander data

When the US space agency NASA sent two Viking Landers to Mars in 1975 to find out whether life might exist on the red planet, it was assumed that the datasets painstakingly compiled by scientists at the time would be available for future generations of scientists on magnetic tape.

Yet, just a few decades later, despite the space agency's best efforts to keep the tapes in a climate-controlled environment, time has left them cracking and brittle. Furthermore, when scientists attempted to re-use some of the data in the late 1990s, they found that they could not decode the formats used. In the end they had to track down old printouts and retype everything.[1]



Courtesy NASA/JPL-Caltech

http://www.dpconline.org/docman/miscellaneous/advocacy/340-mind-the-gap-assessing-digital-preservation-needs-in-the-uk/file Data now available from https://pds-imaging.jpl.nasa.gov/volumes/viking.html

Viking Lander High Resolution Mosaics, Stereo Images and Range Data Sets - SDDPT		
The following are NOT PDS formatted volumes. They were produced by the Science Digital Data Preservation Task by copying data directly off of old, decaying tape media onto more stable CD-WO media. They have not been otherwise reformatted.		
vl_2011		tapes DNM_001-008, DNR_001-143, and DNS_001-016
vl_2012		tapes FNM_001-012 and FNS_001-024

Climatological database for the world's oceans

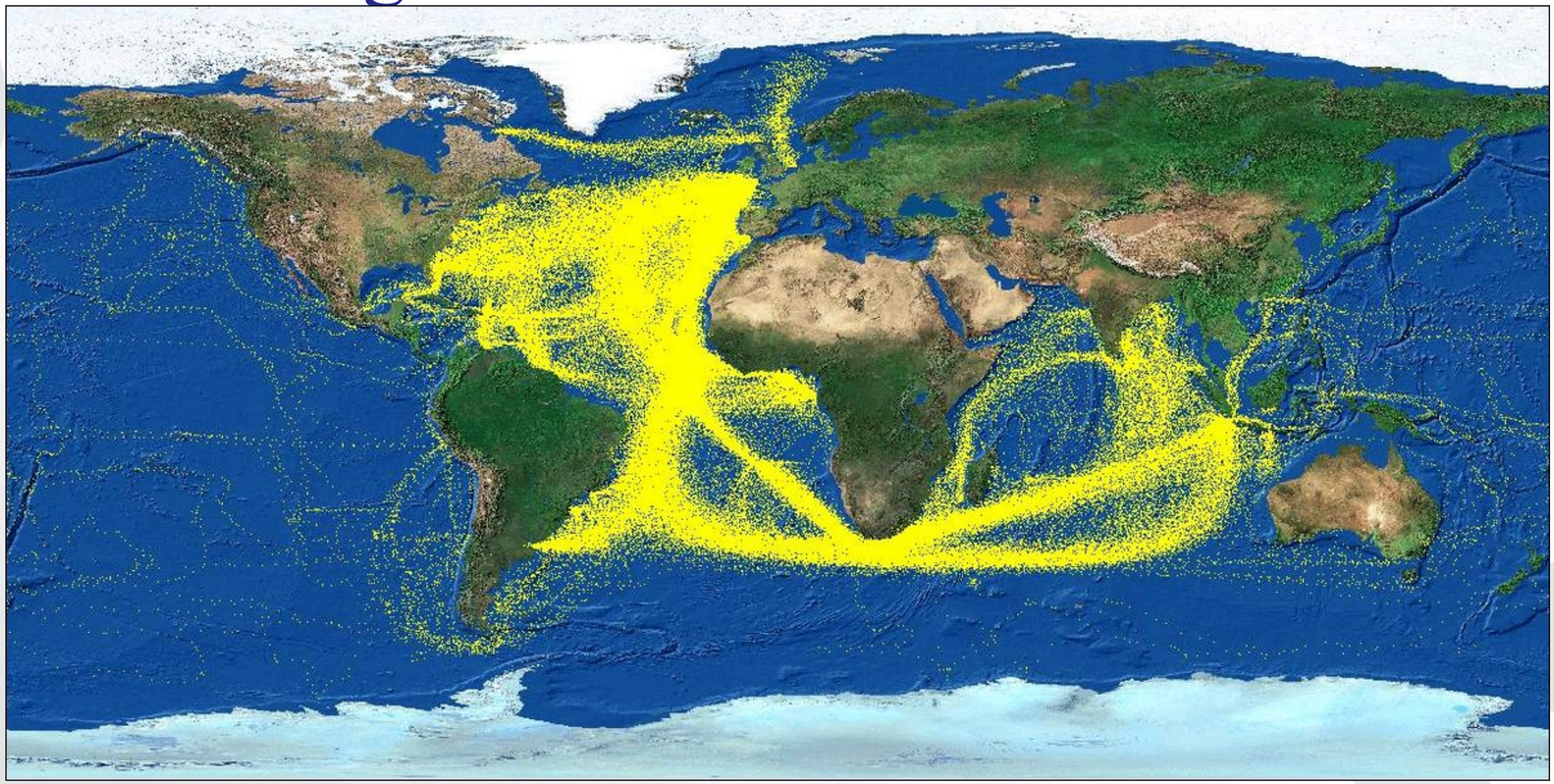
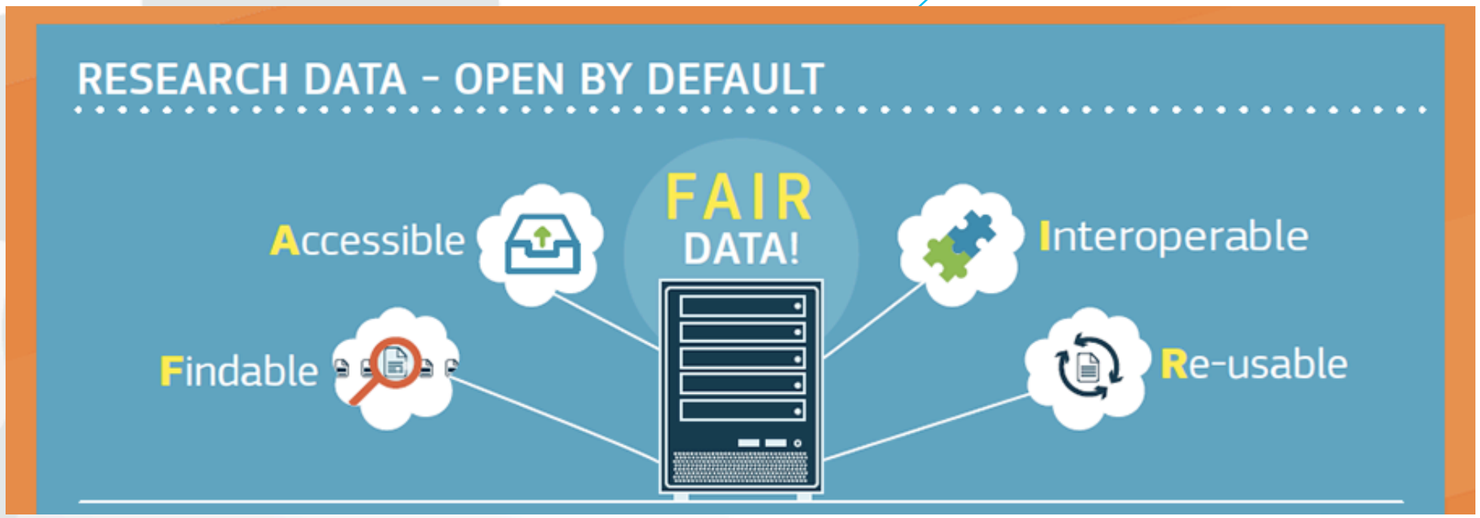


Image copied from https://www.knmi.nl/kennis-en-datacentrum/achtergrond/cliwoc
Every yellow dot represents a ship report.
Project website: http://pendientedemigracion.ucm.es/info/cliwoc/

Photo by rawpixel.com – CC0 - https://www.pexels.com/photo/group-of-people-holding-message-boards-1059116/



FNSNF
SWISS NATIONAL SCIENCE FOUNDATION

Research data should be freely accessible to everyone – for scientists as well as for the general public.

The SNSF agrees with this principle. Since October 2017, researchers have to include a data management plan (DMP) in their funding application for most of the funding schemes. At the same time, the SNSF expects that data generated by funded projects are publicly accessible in digital databases provided there are no legal, ethical, copyright or other issues.

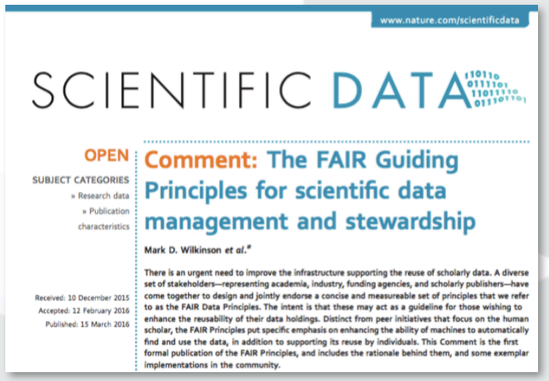
4. Data sharing and reuse		
	4.1 How and where will the data be shared? Questions you might want to consider - On which repository do you plan to share your data? - How will potential users find out about your data?	Consider how and on which repository the data will be made available. The methods applied to data sharing will depend on several factors such as the type, size, complexity and sensitivity [illegible] valued and acknowledged by other researchers. (This relates to the *FAIR Data Principles* F1, F3, F4, A1, A1.1, A1.2 & A2)
	4.2 Are there any necessary limitations to protect sensitive data? Questions you might want to consider: - Under which conditions will the data be made available (timing of data release, reason for delay if applicable)?	[illegible] the time of publication of the respective scientific output. Restrictions may be only due to legal, ethical, copyright, confidentiality or other clauses. Consider whether a non-disclosure agreement would give sufficient protection for confidential data. (This relates to the *FAIR Data Principles* A1 & R1.1)
	4.3 I will choose digital repositories that are conform to the FAIR Data Principles. [CHECK BOX]	The SNSF requires that repositories are conform to the FAIR Data Principles (Section 5 [illegible] searchers, SNSF's explanation of [illegible] FAIR Data Principles). If there are no repositories comp[illegible] ments in your research field, please deposit a copy of your data on a generic platform (see examples). If no data can be shared, this is a statement of principles.
	4.4 I will choose digital repositories maintained by a non-profit organisation. [RADIO BUTTON yes/no] → If the answer is no: "Explain why you cannot share your data on a non-commercial digital repository."	The SNSF supports the use of non-commercial repositories for data sharing. Costs related to data upload are only covered for non-commercial repositories.

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
http://www.snf.ch/en/theSNSF/research-policies/open_research_data/Pages/default.aspx

FAIR data principles

1. **Findable – Easy to find by both humans and computer systems and based on mandatory description of the metadata that allow the discovery of interesting datasets;**
2. **Accessible – Stored for long term such that they can be easily accessed and/or downloaded with well-defined license and access conditions (Open Access *when possible*), whether at the level of metadata, or at the level of the actual data content;**
3. **Interoperable – Ready to be combined with other datasets by humans as well as computer systems;**
4. **Re-usable – Ready to be used for future research and to be processed further using computational methods.**

http://www.dtls.nl/fair-data/
www.force11.org/group/fairgroup/fairprinciples
http://www.nature.com/articles/sdata201618


www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

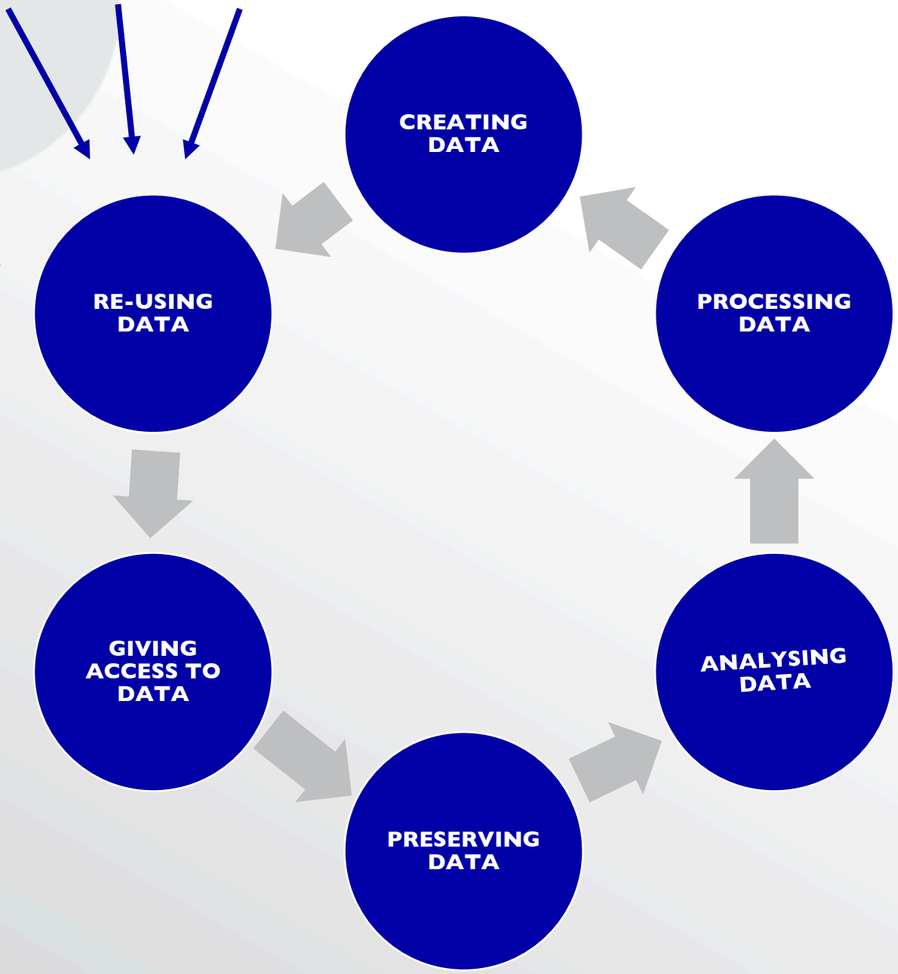
Mark D. Wilkinson et al.*

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

FAIR FOR REUSE

Focus on re-using data



How FAIR are your data?

Findable

It should be possible for others to discover your data. Rich metadata should be available online in a searchable resource, and the data should be assigned a persistent identifier.

- ☐ A persistent identifier is assigned to your data
- ☐ There are rich metadata, describing your data
- ☐ The metadata are online in a searchable resource e.g. a catalogue or data repository
- ☐ The metadata record specifies the persistent identifier

Accessible

It should be possible for humans and machines to gain access to your data, under specific conditions or restrictions where appropriate. FAIR does not mean that data need to be open! There should be metadata, even if the data aren't accessible.

- ☐ Following the persistent ID will take you to the data or associated metadata
- ☐ The protocol by which data can be retrieved follows recognised standards e.g. http
- ☐ The access procedure includes authentication and authorisation steps, if necessary
- ☐ Metadata are accessible, wherever possible, even if the data aren't

Interoperable

Data and metadata should conform to recognised formats and standards to allow them to be combined and exchanged.

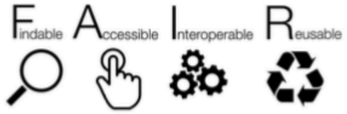
- ☐ …referably open formats
- ☐ …ds
- ☐ …ntologies are used where possible
- ☐ …her related data

"Lots of documentation is needed"

Reusable

Lots of documentation is needed to support data interpretation and reuse. The data should conform to community norms and be clearly licensed so others know what kinds of reuse are permitted.

- ☐ The data are accurate and well described with many relevant attributes
- ☐ The data have a clear and accessible data usage license
- ☐ It is clear how, why and by whom the data have been created and processed
- ☐ The data and metadata meet relevant domain standards



'How FAIR are your data?' checklist, CC-BY by Sarah Jones & Marjan Grootveld, EUDAT, https://doi.org/10.5281/zenodo.1065991 Image CC-BY-SA by SangyaPundir

EUDAT FAIR checklist - CC-BY Sarah Jones & Marjan Grootveld, EUDAT. https://doi.org/10.5281/zenodo.1065991

Metadata

- **Needed to locate the research data and get a first idea of the content.**
- **SNSF:**
  - **intrinsic metadata (e.g. author's name, content of dataset, associated publication, etc.)**
  - **submitter-defined metadata (e.g. definition of variable names, etc.)**
- **Use relevant standards to enable interoperability.**
- **Check which standards the long-term repository supports or expects.**

FAIRsharing.org
standards, databases, policies

https://fairsharing.org/

RDA
RESEARCH DATA ALLIANCE

DCC

http://rd-alliance.github.io/metadata-directory

https://rdamsc.dcc.ac.uk/

Extra: metadata tools: https://rdamsc.dcc.ac.uk/tool-index

Index of metadata tools

- AgriMetamaker
- ANZ-MEST (Metadata Entry and Search Tool)
- AVM Adobe Metadata Panels
- AVM Web Tool
- Bio-Formats
- CF Compliance Checker
- CIF2Cell
- CIM Comparator Tool
- CIM Questionnaire Generator
- CIM Viewer Tool
- CKAN
- CMOR (Climate Model Output Rewriter)
- Converis
- Darwin Core Archive Assistant
- Darwin Core Archive Validator
- Data Package libraries
- Data Package Validator
- Data Package Viewer
- Data Packagist
- DataCite Metadata Store API

http://www.snf.ch/en/theSNSF/research-policies/open_research_data/Pages/data-management-plan-dmp-guidelines-for-researchers.aspx

Documentation?

- Code book explaining the variables
- Study design
- Lab journal
- iPython or Jupyter notebook
- Statistical queries
- Software or instruments to understand or to reproduce the data
- Machine configurations
- Informed consent information
- Data usage licence
- …

In short: document and preserve everything that is needed to replicate the study – ideally following the standard in your discipline

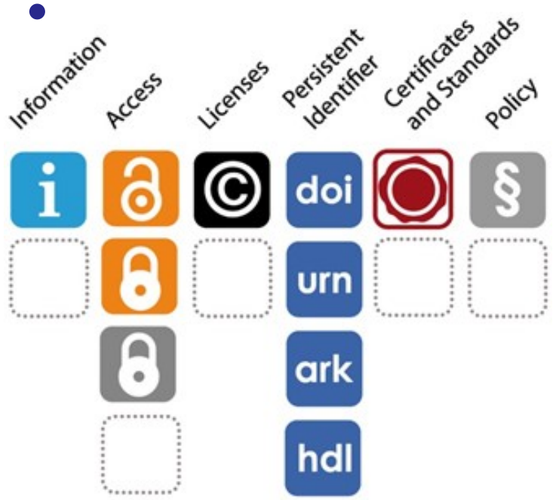
DATA REPOSITORIES

Where to find a repository?

1. Use an external data archive or repository already established for your research domain to preserve the data according to recognised standards in your discipline. More information for **selecting a data repository.**
2. If available, use an institutional research data repository, or your research group's established data management facilities.
3. Use a cost-free data repository such as Zenodo.

   zenodo
   Research. Shared.
4. Search for other research data repositories in http://re3data.org/

   re3data.org
   REGISTRY OF RESEARCH DATA REPOSITORIES



- More information: https://www.openaire.eu/opendatapilot-repository
- Zenodo: http://www.zenodo.org
- Re3data.org: http://www.re3data.org

How to select a repository?

For giving (i.e. archiving & sharing) and taking (i.e. reusing) data:

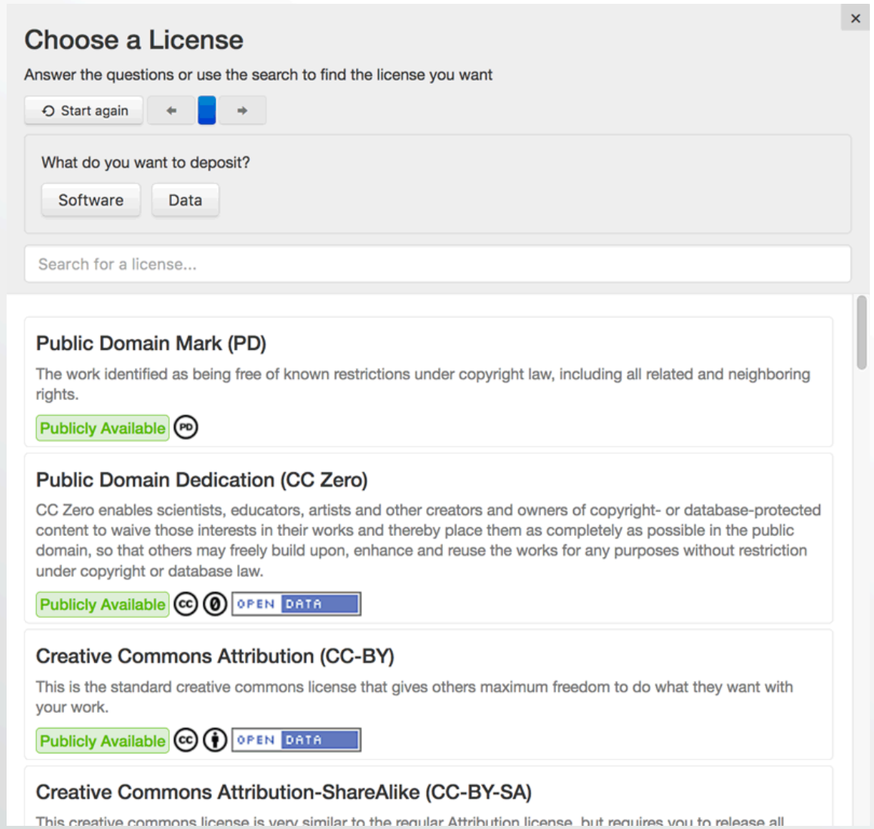
- **Certification as a 'Trustworthy Digital Repository' with an explicit ambition to keep the data available for the long term**
- **Matches your particular data needs: e.g. file formats accepted; mixture of open and restricted access; licences**
- **The costs for e.g. depositing the data, data documentation, and support**
- **Gives your submitted dataset a persistent and globally unique identifier for sustainable citations and to link back to particular researchers and grants**
- **Provides guidance on how to cite the deposited data**

Type	Preferred format(s)	Non-preferred format(s)
Text documents	• PDF/A (.pdf)	• ODT (.odt) • MS Word (.doc, .docx) • RTF (.rtf) • PDF (.pdf)
Plain text	• Unicode text (.txt)	• Non-Unicode text (.txt)
Markup language	• XML (.xml) • HTML (.html) • Related files: .css, .xslt, .js, .es	• SGML (.sgml)
Spreadsheets	• ODS (.ods) • CSV (.csv)	• MS Excel (.xls, .xlsx) • PDF/A (.pdf) • OOXML (.docx, .docm)
Databases	• SQL (.sql) • SIARD (.siard) • DB tables (.csv)	• MS Access (.mdb, .accdb) (v. 2000 or later) • dBase (.dbf) • HDF5 (.hdf5, .he5, .h5)
Statistical data	• SPSS Portable (.por) • SPSS (.sav) • STATA (.dta) • DDI (.xml) • data (.csv) + setup (.txt)	• SAS (.7dat; .sd2; .tpt) • R (* under examination)
Raster images	• JPEG (.jpg, .jpeg) • TIFF (.tif, .tiff) • PNG (.png) • JPEG 2000 (.jp2)	• DICOM (.dcm) (by mutual agreement)
Vector images	• SVG (.svg)	• Illustrator (.ai) • EPS (.eps)

https://dans.knaw.nl/en/deposit/information-about-depositing-data/before-depositing/file-formats?set_language=en

Licensing research data and software

EUDAT licensing wizard helps you pick licences for data & softwa



Choose a public license by answering some questions regarding access to your dataset.

Suggestions depend on several factors:

- **Type of data**
- **Original licenses**
- **Data consumer access and distribution rights**

(EUDAT services are provided through EOSC-hub.)

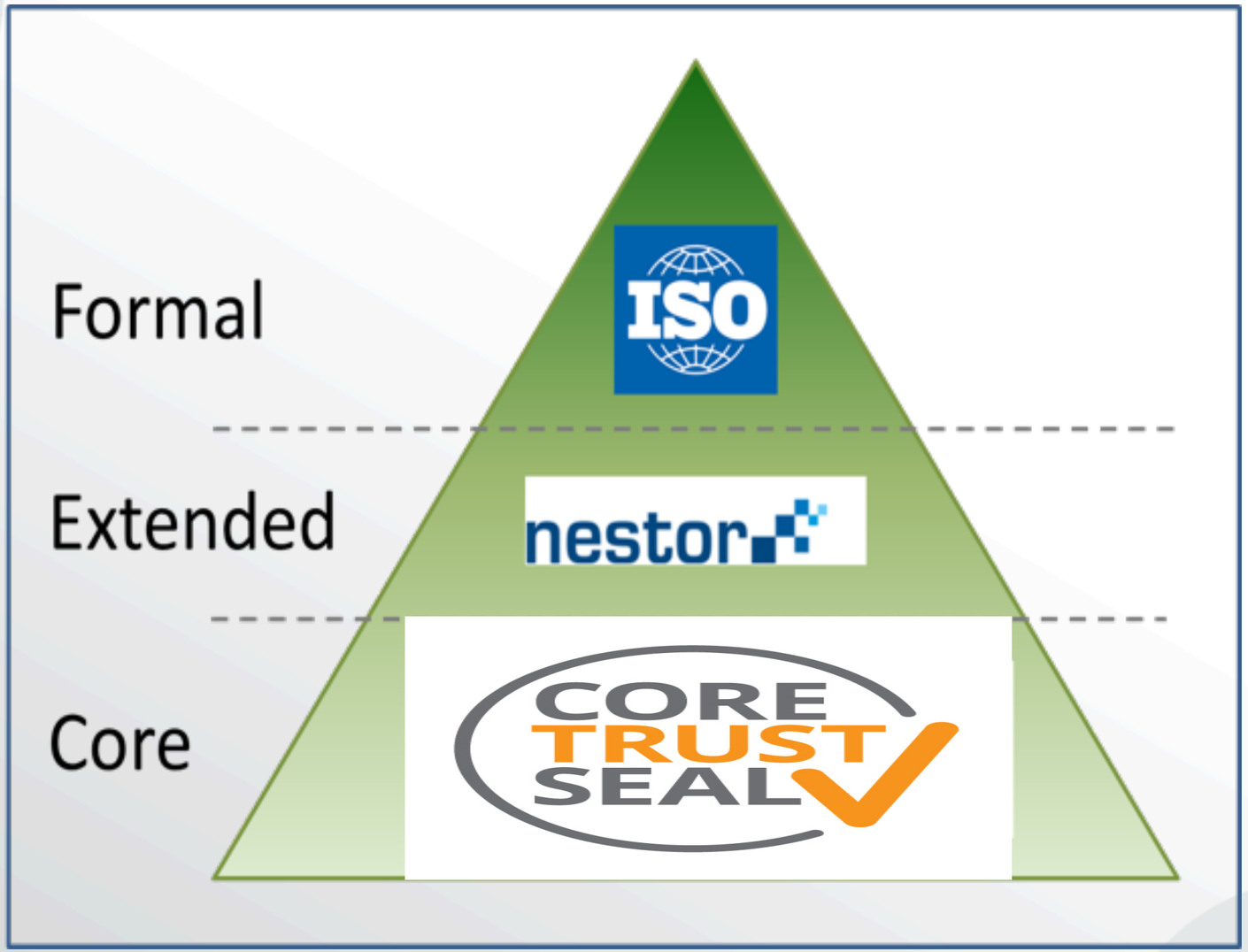
European Commission | http://ufal.github.io/public-license-selector/ and part of https://b2share.eudat.eu/

Short facts about Zenodo

- Catch-all repository for EU-funded research
- Up to 50 GB per upload
- Data stored in the CERN Data Center
- Persistent identifiers (DOIs) for every upload, with DOI versioning
- Includes article-level metrics
- Free for the long tail of science
- Open to all research outputs from all disciplines
- GitHub integration
- Easily add EC funding information and report via OpenAIRE

Zenodo: https://zenodo.org/

Standards of trust



Part of the requirements

R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.

R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

R7. The repository guarantees the integrity and authenticity of the data.

R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.

R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

https://www.coretrustseal.org/

DATA IN REPOSITORIES

GO-FAIR Initiative: FAIR Metrics Group

- ***Aim*: to define metrics enabling both qualitative and quantitative assessment of the degree to which online resources comply with the 15 Principles of FAIR Data**
- ***Founding Members:***
  - **Mark Wilkinson, Universidad Politécnica de Madrid**
  - **Susanna Sansone, University of Oxford**
  - **Michel Dumontier, Maastricht University**
  - **Peter Doorn, DANS**
  - **Luiz Olavo Bonino, VU/DTL**
  - **Erik Schultes, DTL**

https://www.dtls.nl/fair-data/fair-metrics-group/ and http://fairmetrics.org/

FAIR “light” assessment

Findable (defined by metadata (PID included) and documentation)

1. No PID nor metadata/documentation
2. PID without or with insufficient metadata
3. Sufficient/limited metadata without PID
4. PID with sufficient metadata
5. Extensive metadata and rich additional documentation available

Accessible (defined by presence of user license)

1. Metadata nor data are accessible
2. Metadata are accessible but data is not accessible (no clear terms of reuse in license)
3. User restrictions apply (i.e. privacy, commercial interests, embargo period)
4. Public access (after registration)
5. Open access unrestricted

Interoperable (defined by data format)

1. Proprietary (privately owned), non-open format data
2. Proprietary format, accepted by Certified Trustworthy Data Repository
3. Non-proprietary, open format = ‘preferred format’
4. As well as in the preferred format, data is standardised using a standard vocabulary format (for the research field to which the data pertain)
5. Data additionally linked to other data to provide context

DANS FAIRdat prototype: https://www.surveymonkey.com/r/fairdat

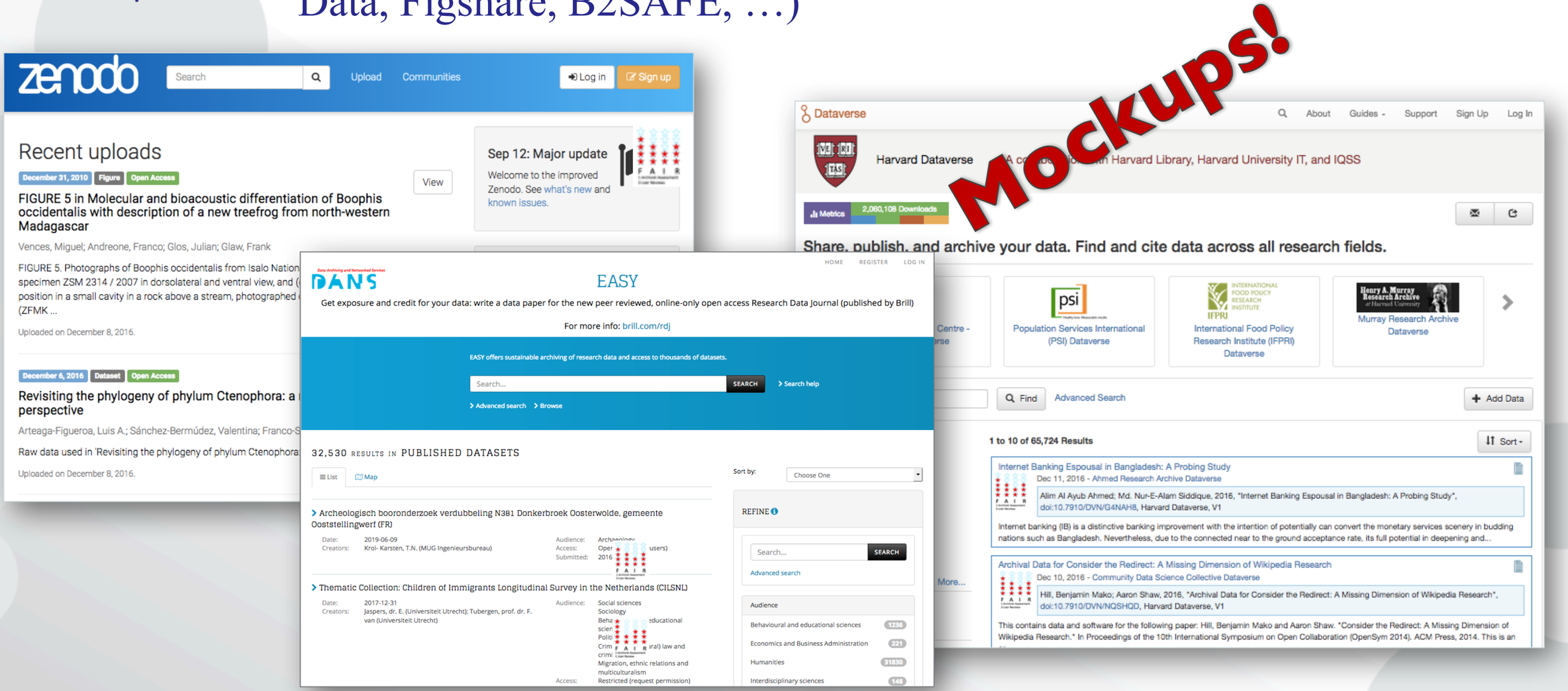
FAIR badge scheme

2 User Reviews
1 Archivist Assessment
24 Downloads

- FAIR as proxy for data “quality” or “fitness for (re-)use”
- We want to create a badge system using the FAIR principles to assess data sets in a Trustworthy Digital Repository
- Developing the data assessment tool: **FAIRdat**
- Score each FAIR dimension on a 5-point scale
- Operationalise the original principles to ensure no interactions among dimensions to ease scoring
- Consider Reusability as the resultant of the other three:
  - the average FAIRness as an indicator of data quality
  - (F+A+I)/3=R Assessment tool based on questionnaire to evaluate any dataset in any (trustworthy) repository by depositors, data specialists and users
- Manual and automatic scoring
- Prototype is being tested

DANS FAIRdat prototype: https://www.surveymonkey.com/r/fairdat

Display FAIR badges in any repository (Zenodo, Dataverse, Mendeley Data, Figshare, B2SAFE, …)



DANS FAIRdat prototype: https://www.surveymonkey.com/r/fairdat

FAIR data HLEG Recommendations and Action Plan – Published June 11, 2018

Rec. 10: Trusted Digital Repositories
Repositories need to be encouraged and supported to achieve CoreTrustSeal certification. The development of rival repository accreditation schemes, based solely on the FAIR principles, should be discouraged.

Rec. 29: Implement FAIR metrics
(…) Repositories should publish assessments of the FAIRness of datasets, where practical, based on community review and the judgement of data stewards.

Let’s learn to trust

- End of the year you have all downloaded and explored at least one dataset - from a repository outside your organisation
- Credit researchers and others who seek value in and add value to existing data
- Support FAIR and Open data in trustworthy repositories, for instance by
  - sticking to standards for data documentation and file formats (“replication packages”)
  - pushing your favourite repository to get certified: make processes and policies transparant
  - educating early-career researchers
    - learn the trade by replicating a study
    - manage data to make them FAIR and as open as possible

OpenAIRE

European Commission

Thank you!

www.openaire.eu

@openaire_eu

facebook.com/groups/openaire

linkedin.com/groups/OpenAIRE-3893548

marjan.grootveld@dans.knaw.nl