Sharing research data: a FAIRytale?

Ingrid Dillo

Deputy Director DANS, Netherlands

Open Science Conference, University of Ljubljana, 22 May 2019

DANS is about keeping data FAIR

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

https://dans.knaw.nl

Policy makers

'Just as the Netherlands is a knowledge society, science is a knowledge community. Both benefit from the free exchange of information. For that reason alone you have my fullest support in your quest for Open Science.'

Ingrid van Engelshoven, minister of Education, Culture and Science

MAKING OPEN SCIENCE A REALITY

OECD

Realising the European Open Science Cloud

First report and recommendations of the Commission High Level Expert Group on the European Open Science Cloud

European Commission

DATA SHARING
OPEN SERVICES
LINKING DATA
CONNECTING DISCIPLINES
CONNECTING SCIENTISTS
BETTER SCIENCE
SUSTAINABLE

Research and Innovation

open science and **H2020 Responsible Research and**
...n

2020 open access mandate
...en access to peer-reviewed articles from co-funded projects; open access to proceedings, grey literature „strongly encouraged"; Open Research Data

research funding
...U Member States mandates aligned with open access in H2020 by the end of 2014, Commission Recommendation of 17. 7. 2012 on access to and preservation of scientific information

Government of the Republic of Slovenia adopted:
· **National strategy of open access to scientific publications and research data in Slovenia 2015–2020** (2015): The same provisions as in H2020 – a Slovenian researcher fulfils the requirements of H2020 and national funding in the same way
· **Action plan for the implementation of the national strategy** (2017)
Till 29 November 2017: **public consultation** on draft Research and Development Act (Articles 21-25 on open science and open access)

Open access in ERA & in Slovenia

Why data sharing is important

Replication and validation of research outcomes
(scientific integrity and transparency)

nature

International weekly journal of science

News

Report finds massive fraud at Dutch universities

Investigation claims dozens of social-psychology papers contain faked data.

The New York Times

Fraud Case Seen as a Red Flag for Psychology Research

By BENEDICT CAREY
Published: November 2, 2011

A well-known psychologist in the Netherlands whose work has been published widely in professional journals falsified data and made up entire experiments, an investigating committee has found. Experts say the case exposes deep flaws in the way science is done in a field, psychology, that has only recently earned a fragile respectability.

ScienceInsider

Breaking news and analysis from the world of science policy

Report: Dutch 'Lord of the Data' Forged Dozens of Studies (UPDATE)

by Gretchen Vogel on 31 October 2011, 7:05 PM | 34 Comments

SPIEGEL ONLINE

Niederlande

Renommierter Psychologe gesteht Fälschungen

---

theguardian

Public sector can combat fraud with data sharing

Outsourcing is not the only thing to blame for procurement fr... says **Graham Kemp**, and the public sector needs to view da... less as a security risk but knowledge to be shared

---

PLoS one

accelerating the publication of peer-reviewed science

Home   Browse Articles   About   For Readers   For Authors and Reviewers

Search a

RESEARCH ARTICLE   OPEN ACCESS

How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data

Why data sharing is important

Re-use of data (efficiency, return on investment, standing on the shoulders of others)

The 19th-century whaling logbooks that could help scientists understand climate change

The public are being enlisted to read through detailed logs of whaling ships which include records of ice flows and weather conditions

Hunters with their catch, circa 1900. Whalers' logbooks are providing detailed information about the climate and ice patterns many years ago. Photograph: Hulton Archive/Getty Images

Maritime historians, climate scientists and ordinary citizens are coming together on a project to study the logbooks of 19th-century whaling ships to better understand modern-day climate change and Arctic weather patterns.

https://www.theguardian.com/environment/2015/dec/17/the-19th-century-whaling-logbooks-that-could-help-scientists-understand-climate-change

16 year study suggests air temperature is external trigger for heart attack

28 Aug 2017

Topic(s): *Environmental and Cardiovascular Disease;*

Barcelona, Spain – 28 Aug 2017: A 16 year study in more than 280 000 patients has suggested that air temperature is an external trigger for heart attack. The findings are presented today at ESC Congress. (1)

"There is seasonal variation in the occurrence of heart attack, with incidence declining in summer and peaking in winter," said first author Dr Moman A. Mohammad, from the Department of Cardiology at Lund University, Skane University Hospital, Lund, Sweden. "It is unclear whether this is due to colder temperatures or behavioural changes."

This nationwide, 16 year, observational study led by Prof David Erlinge from Lund University, is the largest to investigate the association between heart attack incidence and weather conditions such as air temperature, sunshine duration, precipitation, and air pressure.

Using the Swedish myocardial infarction registry (SWEDEHEART), all consecutive heart attacks treated at a coronary care unit between 1 January 1998 and 31 December 2013 were included in the study. The investigators studied the specific weather conditions during which heart attacks occurred using local meteorological data from hundreds of weather stations in the Swedish Meteorological and Hydrological Institute (SMHI).

..and what about the researchers?

How supportive would you be of a national mandate for open data?

2.4% 5.6% 14.3% 26.1% 51.6%

Strongly oppose　Somewhat oppose　Neutral　Somewhat support　Strongly support

Source: The State of Open Data, Digital Science Report (2016).Retrieved: December 23, 2016 . Figures have been redrawn from the originals.

Hesitance in reality

THE DATA ARE MINE!

IVORY TOWER 07 SCIENCE

SCENE FROM THE PAST ?

**Top 4 reasons why researchers
are hesitant to share their data**

1 **50%** - Intellectual property or confidentiality issues

2 **31%** - Ethical concerns

3 **23%** - I am concerned about misinterpretation or misuse of my research

4 **22%** - I am concerned that my research will be scooped

https://authorservices.wiley.com/asset/photos/licensing-and-open-access photos/Wiley%20Global%20Data%20Sharing%20Infographic%20June%202017.pdf

Enabling and promoting data sharing

1. Infrastructure: VREs, trustworthy data repositories

2. Awareness raising and training of researchers

3. Culture change
   - Influence of sharing norms within direct research circle
   - Professional rewards for data sharing

   - External drivers:
     - Funder policies
     - Publishers requirements (DAPs)

     http://repository.jisc.ac.uk/5662/1/KE_report-incentives-for-sharing-researchdata.pdf

FAIR Data Principles

During the 2014 workshop "Designing a data FAIRport" for the life sciences in Leiden a minimal set of community-agreed guiding principles were formulated.
The FAIR Data Principles:

- Easy to find by both humans and machines based on metadata

- With well-defined use license and access conditions (Open Access if possible)

- Ready to be linked with other datasets

- Ready to be re-used for future research and to be processed further using computational methods and tools

FAIR guiding principles

Menu

- About FAIR Data
- Find FAIR Data tools
- Find FAIR Data knowledge & expertise
- The FAIR Data Principles explained
- Bring Your Own Data workshops
- FAIR Data hackathons
- FAIR Data training

FAIR Metrics Group

How FAIR is your data ?

The FAIR Metrics Group is collaborating with a broad set of stakeholders to define metrics enabling both qualitative and quantitative assessment of the degree to which online resources comply with the 15 Principles of FAIR Data as they were originally stated in the "The FAIR Guiding Principles for scientific data management and stewardship".

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

Website

Information about the FAIR Metrics Group can be found at http://fairmetrics.org.

Status

In the first 2 day face to face meeting (2-4 June), this group has set a framework for metrics

https://www.dtls.nl/fair-data/fair-metrics-group/

DANS

The internet of FAIR data and services

Different implementations of FAIR data

Creation

Requirements for new data creation

RISK ALERT

Assessment

Establishing the profile for
existing data

Transformation

Transformation tools to make
data FAIR (Go-FAIR
initiative)

Data Archiving and Networked Services

DANS

Europe and FAIR

EUROPEAN COMMISSION
Directorate-General for Research & Innovation

H2020 Programme

Guidelines on

FAIR Data Management in Horizon 2020

Final Report and Action Plan
from the European
Commission Expert Group
on FAIR Data

**TURNING
FAIR INTO
REALITY**

2018

https://publications.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en/format-PDF/source-80611283

Key Points: To make FAIR a reality …

- Report takes a holistic approach, not a data centric approach
- Need to address the enabling practices and technologies – not just focus on the data and its attributes
- Need to consider all digital outputs (data, code, metadata etc)
- Objective is to make data and other digital research outputs FAIR for humans and machines.
- Needs: concept of FAIR digital objects, FAIR ecosystem, interoperability frameworks for disciplines and across disciplines, FAIR services including trusted digital repositories, skills, metrics and sustainable funding.

FAIRsFAIR

Data Archiving and Networked Services

DANS

Everybody loves FAIR!

Everybody wants to be FAIR, but:
- what does that mean?
- how to put the principles into practice?

The concept of FAIR: what does it really mean?

Good RDM ??

The concept of FAIR: how to put into practice?

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

15 Criteria

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

See: http://datafairport.org/fair-principles-living-document-menu and
https://www.force11.org/group/fairgroup/fairprinciples

OPEN SCIENCE

SCIENCE 1.0

Towards Open Access

Towards Fair Data

Citizen Science

SCIENCE 2.0

skills   rewards   metrics

Stimulation and support

Guidelines to FAIRify Data

http://www.parthenos-project.eu/portal/policies_guidelines

Aim and users

- 20 guidelines structured around the letters of **FAIR: Findable, Accessible, Interoperable, Reusable**

- Intended users are:
  - data producers / researchers who need **clear and simple guidelines** on how to start with RDM
  - RIs and Data Archives

- Intended to be a **first entry point** for good RDM practices

Invest in people and infrastructure

FINDABLE

Research data should be easy to find by both humans and computer systems and based on mandatory descriptions of the metadata that allows the discovery of interesting datasets.

2 Use persistent identifiers

Locating data is a necessary condition for any other step from access to reuse. To be findable, any data object and dataset should be uniquely and persistently identifiable over time with a persistent identifier (PID). A PID continues to work even if the web address of a resource changes. PIDs can take different forms, such as a Handle, DOI, PURL, or URN.

💡 *Reference the PID which was assigned to your dataset in your research output.*

⚙️ *Select the appropriate form of persistent identification schema and assign a PID to every resource. Use the PID Guide from NCDD to decide on the right PID for your research infrastructure.*

3 Cite research data

If research data have a persistent identifier and are cited in accordance with community standards, the corresponding data objects or datasets are more easily found.

💡 *Get acquainted with data citation guidelines that are specific to your field or discipline and cite research data accordingly.*

⚙️ *Provide information about best practices in data citation to research communities and make it easy for data users to cite data, e.g. by using a standardised button which says 'How to cite this dataset'.*

4 Use persistent author identifiers

A persistent author identifier (e.g. VIAF, ISNI or ORCID) helps to create linkages between datasets, research activities, publications and researchers and allows recognition and discoverability.

💡 *Distinguish yourself from any other researcher or research group. Apply for an author identifier if you do not already have one and reference it in your dataset.*

⚙️ *Reference author identifiers in the metadata.*

5 Choose an appropriate metadata schema

Metadata is essential in making data findable, especially the metadata which is used for citing and describing data. A metadata schema is a list of standardised elements to capture information about a resource, e.g. a title, an identifier, a creator name, or a date. Using existing metadata schemas will ensure that international standards for data exchange are met.
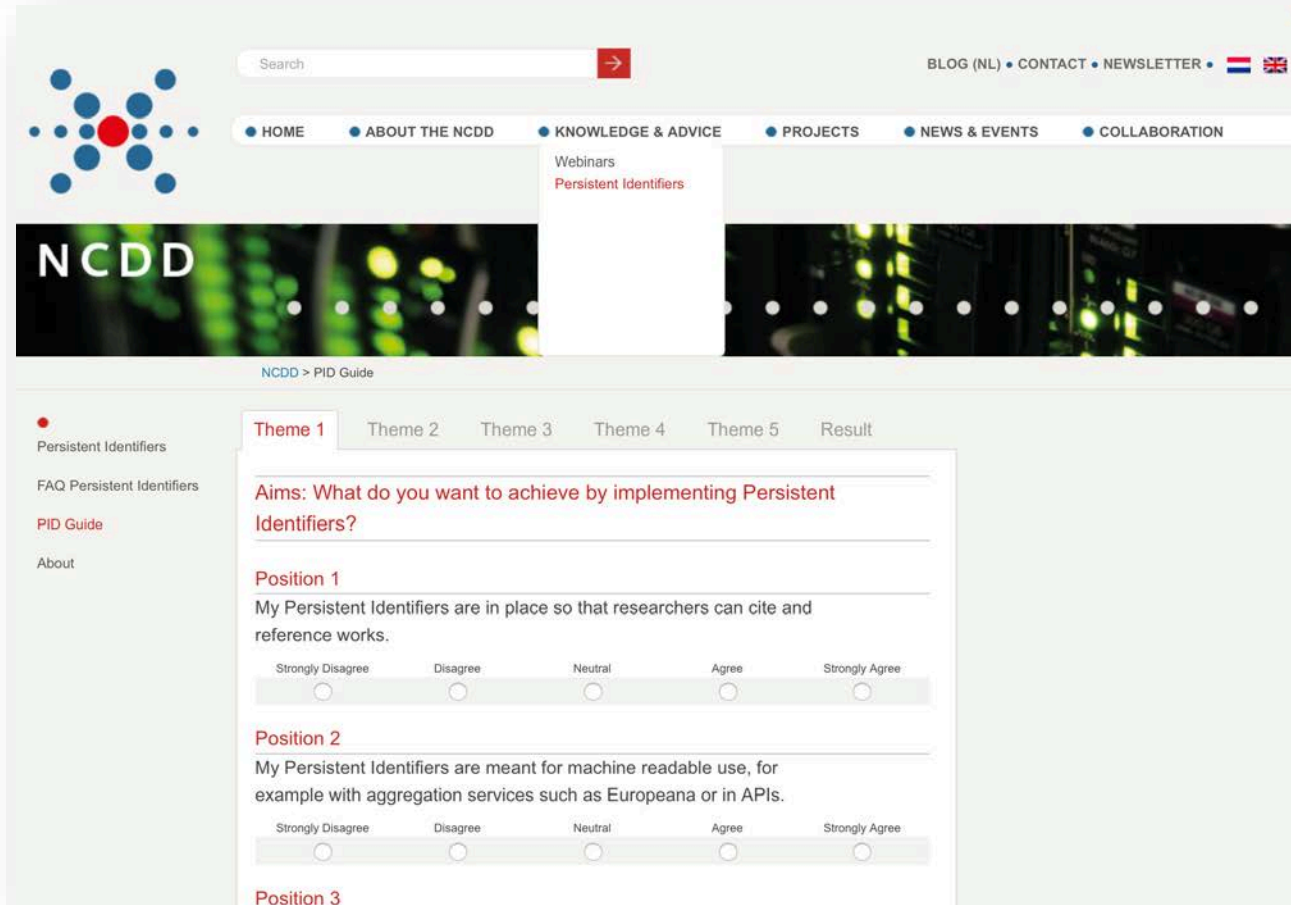
💡 *To enable the discovery of content, describe research data as consistently and completely as possible. Include enough information for the data to be accessed and understood later on. If possible, use an existing metadata schema which fits the type of data object or dataset you are describing.*

⚙️ *Clearly state which metadata schema you apply and recommend to the research community. To enrich datasets at data deposit, consider having a data submission form which collects additional metadata, e.g. about the provenance of the data.*

PID Guide

Data Archiving and Networked Services
DANS

ACCESSIBLE

Research data should be easily accessible and retrievable with well-defined access conditions using standardised communication protocols.

6 Choose a trustworthy repository

A certified repository offers a trustworthy home for datasets. Certification is a guarantee that data are stored safely, and will be available, findable and accessible over the long-term. Examples of certification standards are CoreTrustSeal, nestor seal and ISO 16363 certification.

Make your data accessible through a trustworthy repository. In addition, if you follow the repositories' standards (on preferred file formats, metadata schemas etc.) you can make sure that all requirements for making data FAIR are met.

Clearly state the level of certification on your website. If you are not (yet) certified, state how you plan to ensure availability, findability, accessibility and reusability in the long-term.

7 Clearly state accessibility

Access information specifies how a data user may access a dataset. When depositing data in a data repository, it should be clear which access options a data depositor can choose.

When choosing an access option, consider legal requirements, discipline-specific policies and ethics protocols when applicable. Choose Open Access when possible. When you collect personal data, ask yourself whether it contains any information which might lead to participants' identities being disclosed, what participants consented to and which measures you have taken to protect your data. If your data cannot be published in Open Access, the metadata should be, allowing data discovery.

Encourage (meta)data to be published in Open Access. Cleary state restricted access options for sensitive (meta)data that should not be part of the publicly accessible (meta)data. In this case, strive to make the (meta)data available through a controlled and documented access procedure.

8 Use a data embargo when needed

During a data embargo period, only the description of the dataset is published. The data themselves are not accessible. The full (meta)data will become available after a certain period of time.

Clearly state why and for what period a data embargo is needed. Make the (meta)data openly available as soon as possible.

Specify whether a data embargo is allowed and what conditions apply.

9 Use standardised exchange protocols

By using standardised exchange protocols, research infrastructures can make (meta)data publicly accessible and harvestable by e.g. search engines, vastly improving accessibility.

Use standardised protocols such as SWORD, OAI-PMH, ResourceSync and SPARQL. Convert metadata schemas into XML/RDF. Maintain a registry for protocol endpoints, the path at which research data can be accessed, and publish them.

Registry of Research Data Repositories

http://re3data.org

https://www.coretrustseal.org

To speed up discovery and uncover new insights, research data should be easily combined with other datasets by humans as well as computer systems.

INTEROPERABLE

10 Establish well documented machine-actionable APIs

Well documented and machine-actionable APIs - a set of subroutine definitions, protocols, and tools for building application software - allow for automatic indexing, retrieval and combining of (meta)data from different data repositories.

Document APIs well and make it possible to deliver the schema of the (meta)data model. Consider showing examples of how to successfully mine data from different endpoints and combine them into new data sets usable for new research.

11 Use open well-defined vocabularies

The description of metadata elements should follow community guidelines that use open, well defined and well known vocabularies. Such vocabularies describe the exact meaning of the concepts and qualities that the data represent.

Use vocabularies relevant to your field, and enrich and structure your research output accordingly from the start of your research project.

Give examples of vocabularies the research community may use, based on research domain specifics.

12 Document metadata models

Clearly documenting metadata models helps developers to compare and make mappings between metadata.

Publish the metadata models in use in your research infrastructure. Document technical specifications and define classes (groups of things that have common properties) and properties (elements that express the attributes of a metadata section as well as the relationships between different parts of the metadata). For metadata mapping purposes, list the mandatory and recommended properties.

13 Prescribe and use interoperable data standards

Using a data standard backed up by a strong community, increases the possibility to share, reuse and combine data collections.

Check with the repository where you want to deposit your data what data standards they use. Structure your data collection in this format from the start of your research project.

Clearly specify which data standard your institution uses, pool a community around them and maintain them especially with a perspective on interoperability. Good examples are CMDI (language studies) and the SIKB0102 Standard (archaeology).

14 Establish processes to enhance data quality

To boost (meta)data quality and, therefore, interoperability, establish (automatic) processes that clean up, derive and enrich (meta)data.

Establish procedures to minimise the risk of mistakes in collecting data. E.g. choose a date from a calendar instead of filling it in by hand.

Invest in tools to help clean up (meta)data and to convert data into standardised and interoperable data formats. Combine efforts to develop workflows and software solutions for such automatic processes, e.g. by using machine learning tools.

15 Prescribe and use future-proof file formats

All data files held in a data repository should be in an open, international, standardised file format to ensure long-term interoperability in terms of usability, accessibility and sustainability.

From the start of your research project think about future-proof file formats. Use preferred formats which are recommended by the data repository and are independent of specific software, developers or vendors.

Encourage the use of formats that are considered suitable for long-term preservation such as PDF-A, CSV and MID/MIF files. Provide an easy-to-find and detailed overview of accepted file formats.

Information about depositing data

Before depositing

Archaeological exchange protocol

File formats

Text documents

Plain text

Markup language

Spreadsheets

Databases

Statistical data

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

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. Depositing 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. DANS favours the use of preferred formats and recommends depositors to try to deposit data as much as possible in preferred formats.

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.

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) • 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)	• MS Access (.mdb, .accdb) (v. 2000 or

Research data should be ready for future research and future processing, making it self-evident that findings can be replicated and new research effectively builds on already acquired, previous results.

REUSABLE

16 Document data systematically

To make clear what can and what cannot be expected in a dataset or repository, data should be systematically documented. Being transparent about what's in the data and what isn't facilitates trust and, consequently, data reuse.

Provide codebooks, including a description of methodology, a list of abbreviations, a description of gaps in the data, the setup of the database, etc.

17 Follow naming conventions

Following a precise and consistent naming convention - a generally agreed scheme to name data files - makes it significantly easier for future generations of researchers to retrieve, access and understand data objects and datasets.

Consult the policies and best practices for your research discipline or domain to find the most suitable naming convention.

Clearly state best practices to create and apply specific file naming conventions.

18 Use common file formats

By using standardised file formats that are widely used in your community, reusability is increased.

Use current popular file formats next to archival formats to share your data, e.g. Excel (xlsx) and CSV or ESRI Shapefiles next to MID/MIF files.

Publish the data in popular formats next to the archival format if they are not the same.

19 Maintain data integrity

Research data which were collected should be identical to the research data which are accessed later on. To ensure data authenticity, checks for data integrity should be performed.

Implement a method for version control. The guarantee that every change in a revised version of a dataset is correctly documented, is of integral importance for the authenticity of each dataset.

To identify if a file has been modified, it is essential to record provenance - the origin of the data plus any changes made over time - and to compare any copy with the original. A data integrity check can be performed by means of a fingerprint such as a checksum, or by a direct comparison of two files. Provide a mechanism to address different versions, for example by adding the version to the identifier as a search parameter.

20 License for reuse

To permit the widest reuse possible of (meta)data, it should be clear who the (meta)data rights holder is and what licence applies.

Make sure you know who the (meta)data rights holder is before publishing your research data.

Communicate the (meta)data licence and reuse options transparently and in a machine-readable format. To improve interoperability, try to map your licences to frameworks which are already widely adopted such as Creative Commons.

Love to share data

February 14, 2019
Open day on open science
DANS, The Hague

dans.knaw.nl/love-to-share-data

DANS   OpenAIRE | ADVANCE   PNN

"Research Data Management...
Everybody talks about it
but nobody does it."

It's Love Data Week 2019!
Don't stay alone with your Research Data.

I L♥VE
RESEARCH
DATA

DATA

Love data management

Webinar on the Expert tour guide on Data Management

Benefits of data management

The concept of Data Management implies

PLAN
DISCOVER   ORGANISE & DOCUMENT
PUBLISH   PROCESS
PROTECT   STORE

cessda eric

» How to handle, organize, structure and store research data
» Takes into account technical, organizational, structural, legislative and
  sustainability aspects
» Clear structure of how data is going to be managed
» Might involve some additional work at an early stage

METADATA IS A
LOVE NOTE
TO THE FUTURE

Become a member of the Research Data Alliance!

Thank you for listening!

ingrid.dillo@dans.knaw.nl
www.dans.knaw.nl

https://www.fairsfair.eu
https://www.rd-alliance.org