FAIR data in trustworthy repositories:
How do I organise and preserve my research data?

Workshop Digital Humanities – the perspective of Africa

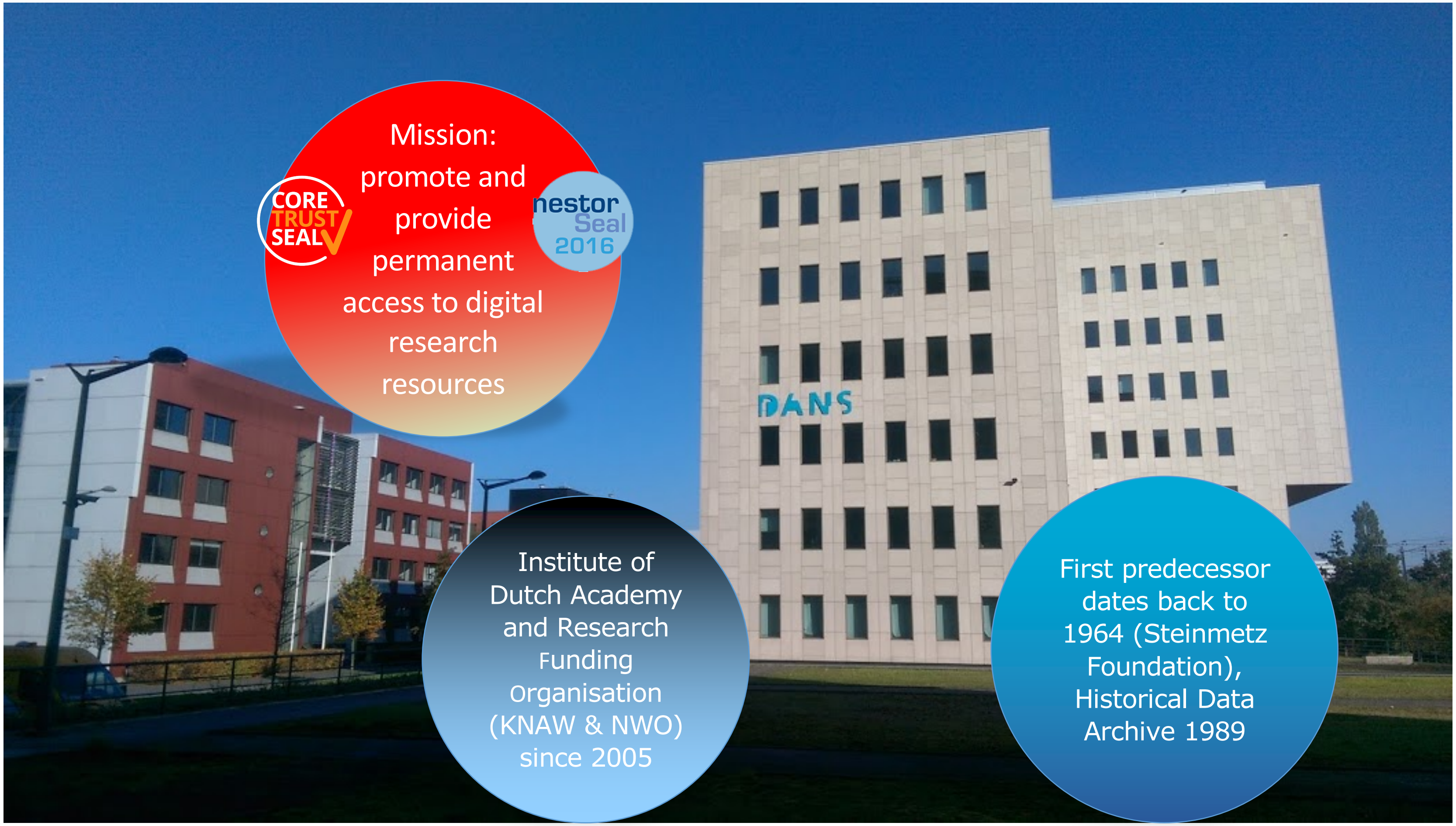
Marjan Grootveld

July 1st, 2019

1. What is DANS?
2. What is data?
3. Exercise 1: organise your data
4. Trustworthy data repositories
5. Exercise 2: find a relevant repository for your data
6. Take-aways from this session

https://www.pexels.com/photo/purple-blue-green-pink-orange-and-yellow-highlighter-159659/

DANS https://dans.knaw.nl/nl

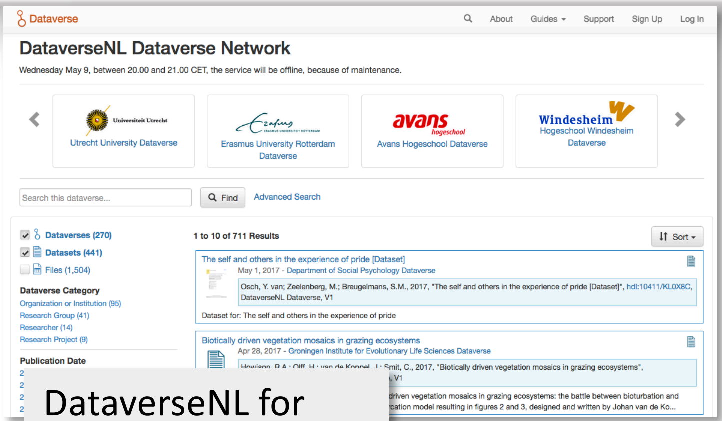


Mission: promote and provide permanent access to digital research resources

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

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

Core DANS services



DataverseNL Dataverse Network

Wednesday May 9, between 20.00 and 21.00 CET, the service will be offline, because of maintenance.

Universiteit Utrecht
Utrecht University Dataverse
Erasmus University Rotterdam Dataverse
avans hogeschool
Avans Hogeschool Dataverse
Windesheim
Hogeschool Windesheim Dataverse
Search this dataverse...
Find
Advanced Search

☑ Dataverses (270)
☑ Datasets (441)
☐ Files (1,504)

Dataverse Category
Organization or Institution (95)
Research Group (41)
Researcher (14)
Research Project (9)

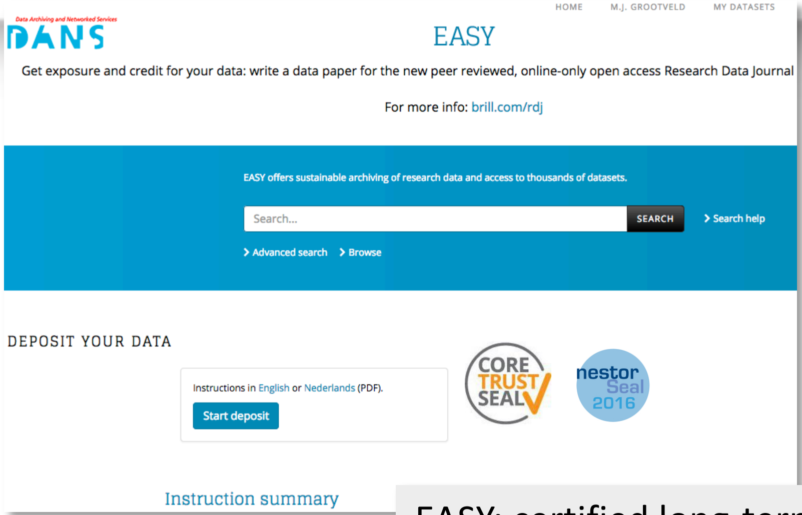
Publication Date

1 to 10 of 711 Results

The self and others in the experience of pride [Dataset]
May 1, 2017 - Department of Social Psychology Dataverse
Osch, Y. van; Zeelenberg, M.; Breugelmans, S.M., 2017, "The self and others in the experience of pride [Dataset]", hdl:10411/KL0X8C, DataverseNL Dataverse, V1
Dataset for: The self and others in the experience of pride

Biotically driven vegetation mosaics in grazing ecosystems
Apr 28, 2017 - Groningen Institute for Evolutionary Life Sciences Dataverse
Howison, R.A.; Olff, H.; van de Koppel, J.; Smit, C., 2017, "Biotically driven vegetation mosaics in grazing ecosystems", , V1
driven vegetation mosaics in grazing ecosystems: the battle between bioturbation and cation model resulting in figures 2 and 3, designed and written by Johan van de Ko...

DataverseNL for short- and mid-term data storage



Get exposure and credit for your data: write a data paper for the new peer reviewed, online-only open access Research Data Journal

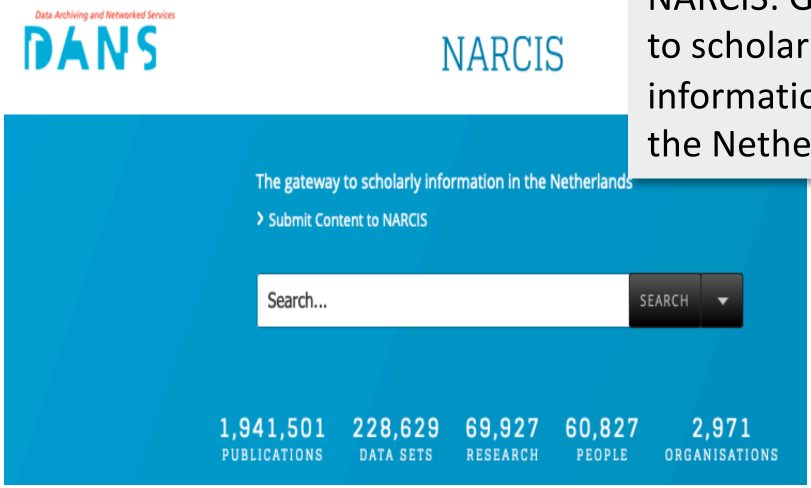
For more info: brill.com/rdj

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

DEPOSIT YOUR DATA

Instructions in English or Nederlands (PDF).

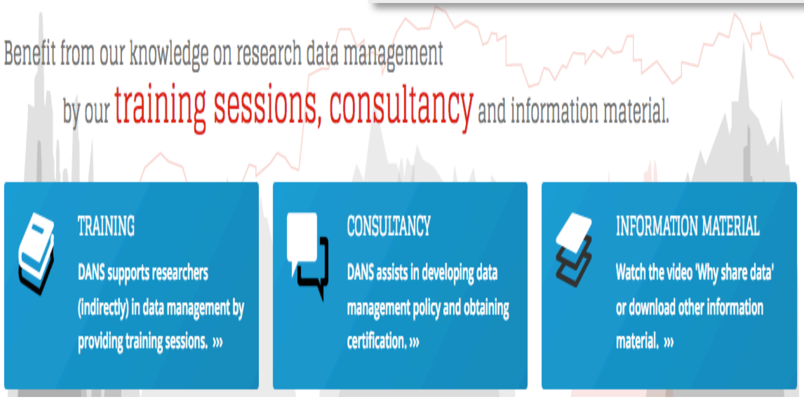
EASY: certified long-term Electronic Archiving System for self-deposit



The gateway to scholarly information in the Netherlands

1,941,501	228,629	69,927	60,827	2,971
PUBLICATIONS	DATA SETS	RESEARCH	PEOPLE	ORGANISATIONS

NARCIS: Gateway to scholarly information in the Netherlands


Benefit from our knowledge on research data management by our training sessions, consultancy and information material.

TRAINING
DANS supports researchers (indirectly) in data management by providing training sessions. »

CONSULTANCY
DANS assists in developing data management policy and obtaining certification. »

INFORMATION MATERIAL
Watch the video 'Why share data' or download other information material. »

Data management training and consultancy

Partner in (inter)national projects:

EUDAT

EOSC pilot
The European Open Science Cloud for Research Pilot Project

EOSC-hub

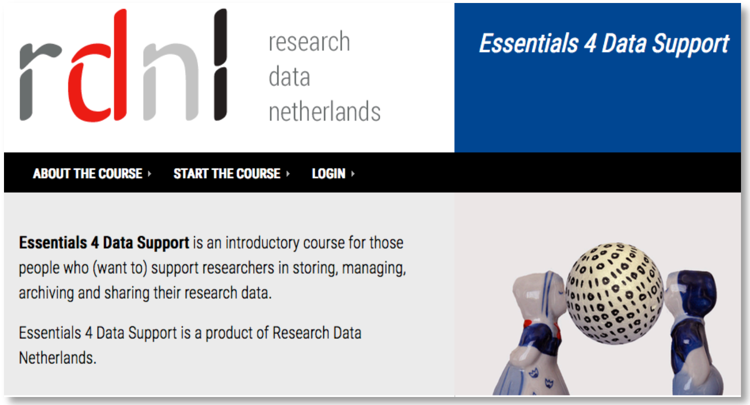
OpenAIRE | ADVANCE

FAIRSFAIR
Fostering Fair Data Practices in Europe

FAIR and Open data

Data management planning

Trustworthy digital repositories


rdnl research data netherlands

Essentials 4 Data Support

Essentials 4 Data Support is an introductory course for those people who (want to) support researchers in storing, managing, archiving and sharing their research data.

Essentials 4 Data Support is a product of Research Data Netherlands.

https://datasupport.researchdata.nl/en/

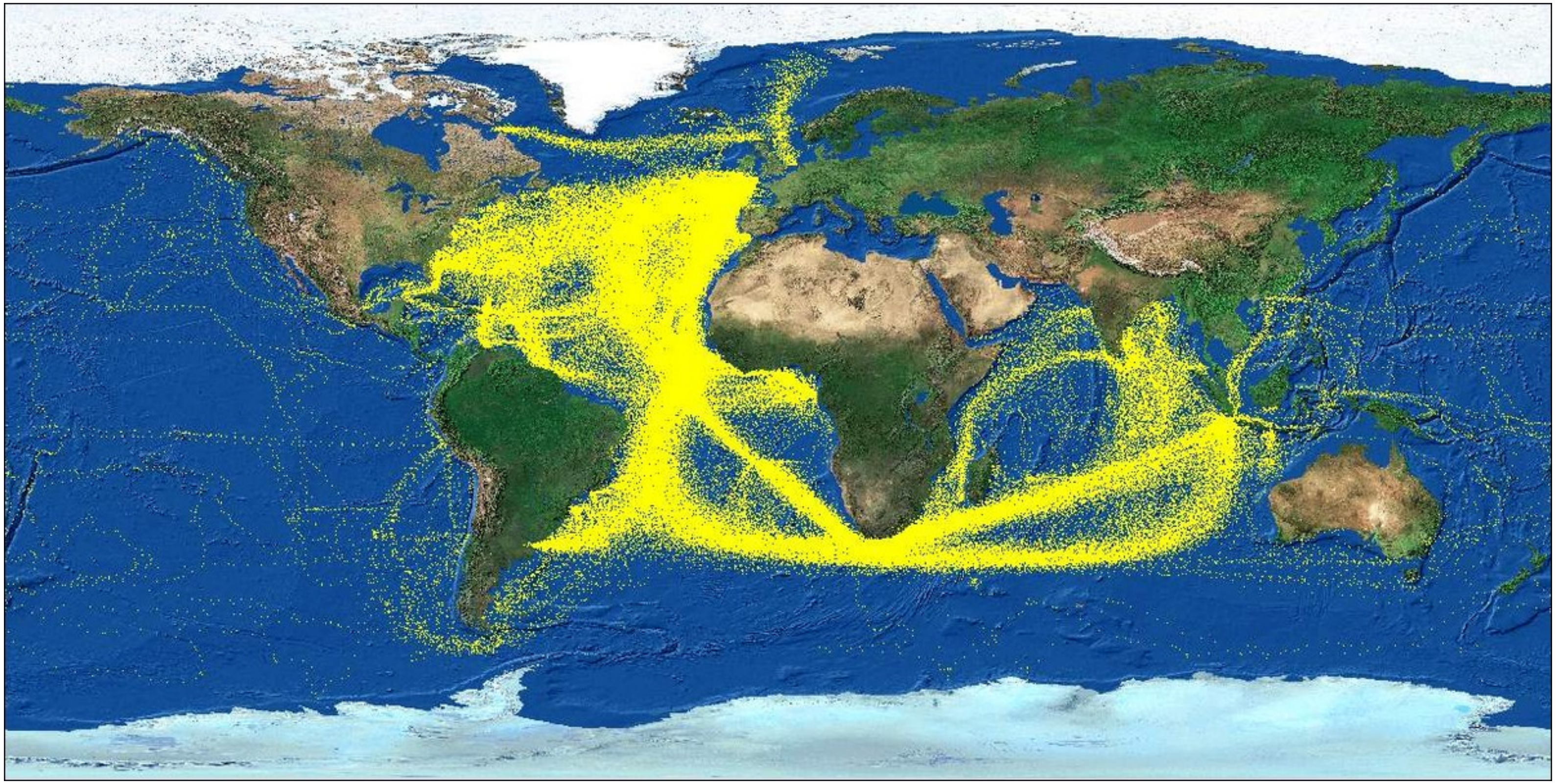
https://eudat.eu/
https://eoscpilot.eu/
https://www.eosc-hub.eu/
https://www.openaire.eu/
https://www.fairsfair.eu/

What is research data?

Research data is all FACTUAL EVIDENCE WHICH IS COLLECTED OR GENERATED BY RESEARCHERS in the course of their work, while they are STUDYING PEOPLE, THINGS OR EVENTS.
GoAnimate
Go Make Your Own

An introduction to the basics of research data
https://www.youtube.com/watch?v=q2aiDJzJPuw

From “real life” to research data: CLIWOC - climatological database for the world’s oceans



Every yellow dot represents a ship report.

Image copied from https://www.knmi.nl/kennis-en-datacentrum/achtergrond/cliwoc

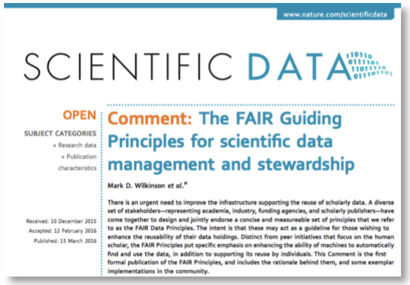
Project web site: http://pendientedemigracion.ucm.es/info/cliwoc/

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.nature.com/articles/sdata201618
www.force11.org/group/fairgroup/fairprinciples

https://librarycarpentry.org/Top-10-FAIR/ Top 10 FAIR Data and Software Things



www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

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

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

Mark D. Wilkinson *et al.*#

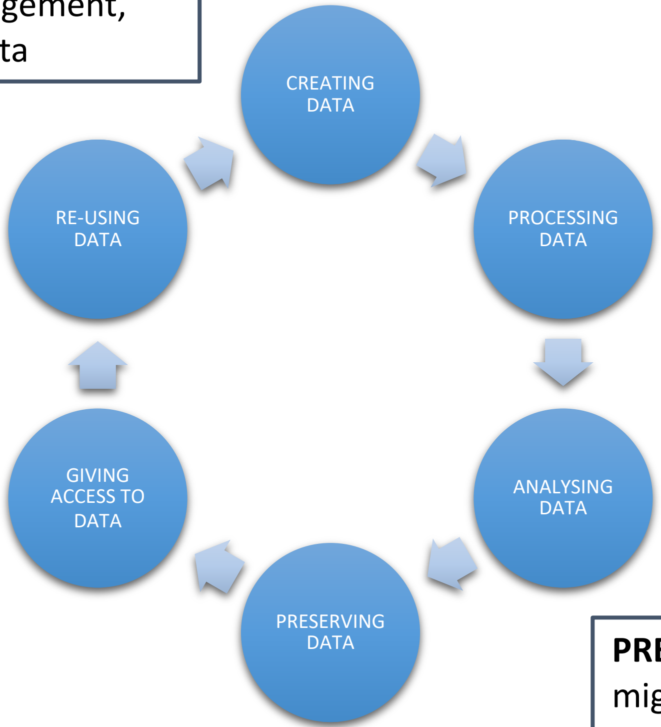
There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we 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.

Simplified research data lifecycle

CREATING DATA: designing research, DMPs, planning consent, locate existing data, data collection and management, capturing and creating metadata

PROCESSING DATA: entering, transcribing, checking, validating and cleaning data, anonymising data, describing data, manage, store, back-up data

RE-USING DATA: follow-up research, new research, undertake research reviews, scrutinising findings, teaching & learning



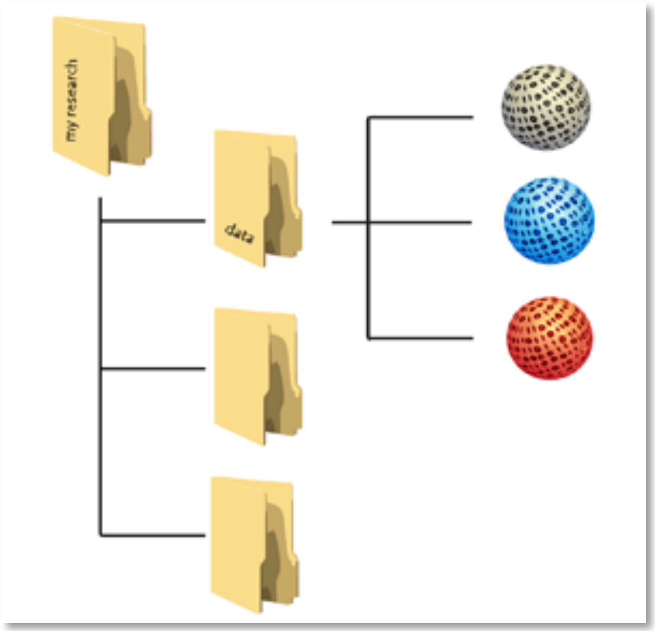
ANALYSING DATA: interpreting, & deriving data, producing outputs, authoring publications, preparing for sharing

ACCESS TO DATA: distributing data, sharing data, controlling access, establishing copyright, promoting data

PRESERVING DATA: data archiving, migrating to best format & medium for long term, creating metadata and documentation

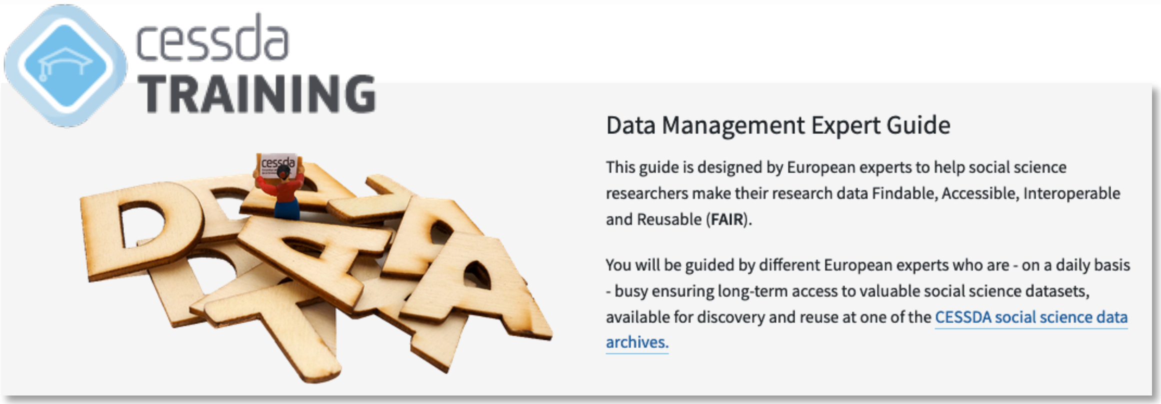
Based on UK Data Archive lifecycle: https://www.ukdataservice.ac.uk/manage-data/lifecycle
Used in OpenAIRE RDM briefing paper: https://www.openaire.eu/briefpaper-rdm-infonoads

Exercise 1: Data organisation – 15 minutes



- Form a group of 4-5 people
- Read the *Veteran tapes* description
- Design a data organisation for this project:
  1. Folder structure
  2. File-naming convention
- Don’t drown yourself in the details

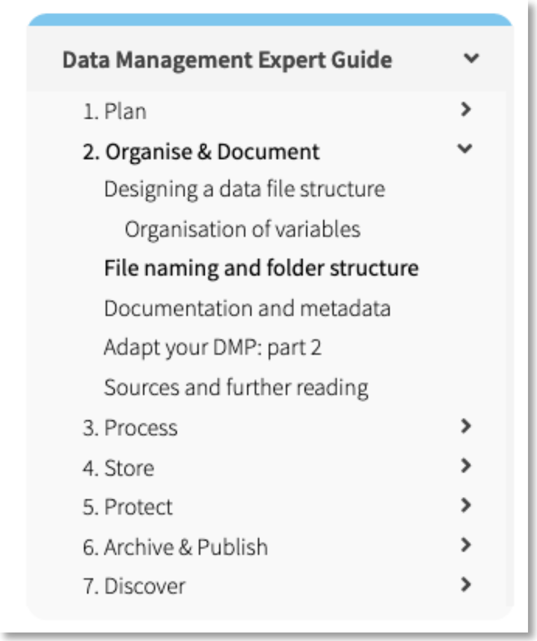
Folders and files: CESSDA DM Expert Guide



Data Management Expert Guide

This guide is designed by European experts to help social science researchers make their research data Findable, Accessible, Interoperable and Reusable (**FAIR**).

You will be guided by different European experts who are - on a daily basis - busy ensuring long-term access to valuable social science datasets, available for discovery and reuse at one of the CESSDA social science data archives.



Data Management Expert Guide

1. Plan
2. Organise & Document
   - Designing a data file structure
     - Organisation of variables
   - **File naming and folder structure**
   - Documentation and metadata
   - Adapt your DMP: part 2
   - Sources and further reading
3. Process
4. Store
5. Protect
6. Archive & Publish
7. Discover

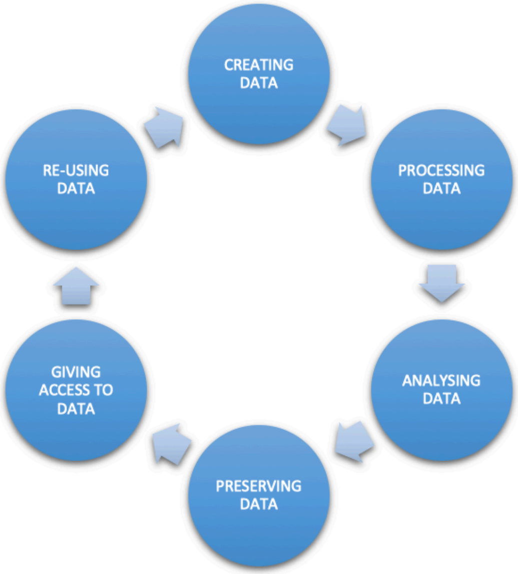
http://cessda.eu/dmeg Chapter 2

Plan data management for the full cycle

A Data Management Plan is a brief plan to define:

- how the data will be created
- how it will be documented
- who can access it
- where it will be stored
- **whether the data will be shared or "published"**
- **where it will be preserved**

Well, in a serious repository of course ;-)



How to select a repository?

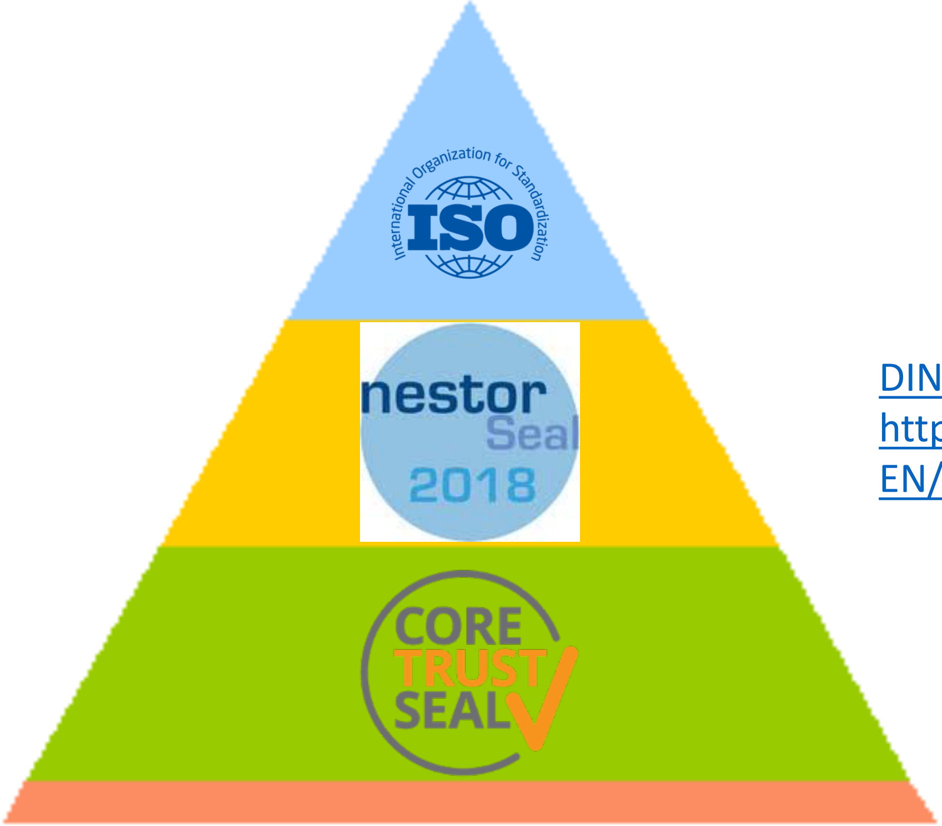
So…
WHERE WOULD YOU STORE THIS?
LIBRARIAN
AUKE
www.msgerry.com / www.aukeherrema.nl
cc BY

LEARNING HOW TO ARCHIVE DATA

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

- Certification as a 'Trustworthy Data 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;
  - usage licences
- 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

Standards of trust in repositories



Formal

http://www.iso16363.org/

Extended

DIN 31644
http://www.dnb.de/Subsites/nestor/EN/Siegel/siegel.html

Core

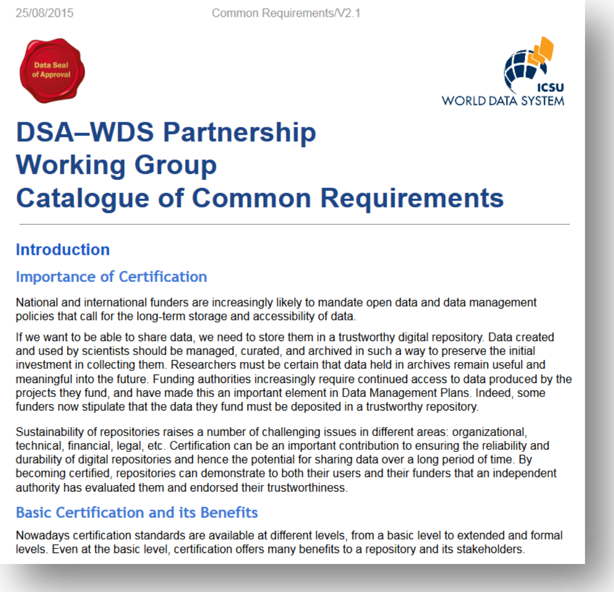
https://www.coretrustseal.org

CoreTrustSeal Data Repository Certification

16 Requirements:

- Context (R0)
- Organisational infrastructure (R1-6)
- Digital object management (R8-14)
- Technology (R15-16)

https://doi.org/10.17026/dans-22n-gk35



25/08/2015 Common Requirements/V2.1

Data Seal of Approval

ICSU WORLD DATA SYSTEM

DSA–WDS Partnership Working Group Catalogue of Common Requirements

Introduction

Importance of Certification

National and international funders are increasingly likely to mandate open data and data management policies that call for the long-term storage and accessibility of data.

If we want to be able to share data, we need to store them in a trustworthy digital repository. Data created and used by scientists should be managed, curated, and archived in such a way to preserve the initial investment in collecting them. Researchers must be certain that data held in archives remain useful and meaningful into the future. Funding authorities increasingly require continued access to data produced by the projects they fund, and have made this an important element in Data Management Plans. Indeed, some funders now stipulate that the data they fund must be deposited in a trustworthy repository.

Sustainability of repositories raises a number of challenging issues in different areas: organizational, technical, financial, legal, etc. Certification can be an important contribution to ensuring the reliability and durability of digital repositories and hence the potential for sharing data over a long period of time. By becoming certified, repositories can demonstrate to both their users and their funders that an independent authority has evaluated them and endorsed their trustworthiness.

Basic Certification and its Benefits

Nowadays certification standards are available at different levels, from a basic level to extended and formal levels. Even at the basic level, certification offers many benefits to a repository and its stakeholders.

Main CoreTrustSeal 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/

And other repositories?

Repositories without a trustworthy, long-term ambition may have a simpler process for depositing and preserving data:

- typically, they don’t ask for preferred file formats – because they won’t convert or migrate the data to new formats in future (mere “bit preservation”);
- they may be less demanding (or helpful!) regarding metadata, and
- they won’t remind data producers to add documentation – which probably diminishes the interpretability and reusability of the data;
- they may not have long-term budget, qualified staff, appropriate technical infrastructure nor a continuity plan, should the organisation or the budget fail.

Exercise 2: Use re3data to find a repository

http://www.re3data.org/

Read the *Veteran tapes* project brief and identify what should be kept for the long term (3 mins)

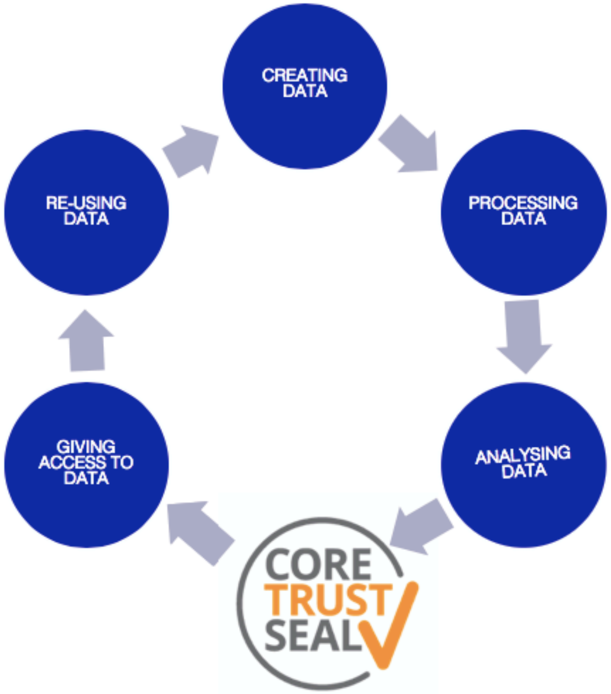
Search re3data.org for repositories (10 mins), considering:

1. Data type(s)
2. Discipline
3. Repository features

It’s all about trust

- All data needs to be properly managed.
- Decisions made early affect what you can do later. For instance:
  - Folder structure with authorisations
  - File-naming conventions
  - Domain metadata
  - Accompanying documentation
- Strong resemblance CoreTrustSeal requirements and FAIR principles: ongoing access, explicitness & clarity, metadata, persistent references, documentation, data discovery, understandability, reuse, …
- Depositing data in a certified repository makes life easier for researchers and keeps FAIR data FAIR.

All’s FAIR that ends FAIR – any questions?



Acknowledgements:

https://eoscpilot.eu/

https://eudat.eu/

https://www.fairsfair.eu/

https://www.eosc-hub.eu/

https://www.openaire.eu/

marjan.grootveld@dans.knaw.nl

https://dans.knaw.nl/en/projects/