DAY 3 | Sharing and reuse for research

July 3rd, 2023
11.00 - 12.00

Publishing and archiving data

Kim Ferguson & Cees Hof
Data Archiving and Networked Services (DANS-KNAW)

Learning outcomes

- Understand the relation between repositories and FAIR data.
- Be able to determine and prioritise criteria for selecting a repository.
- Practice in providing advice to researchers about data publishing and archiving.
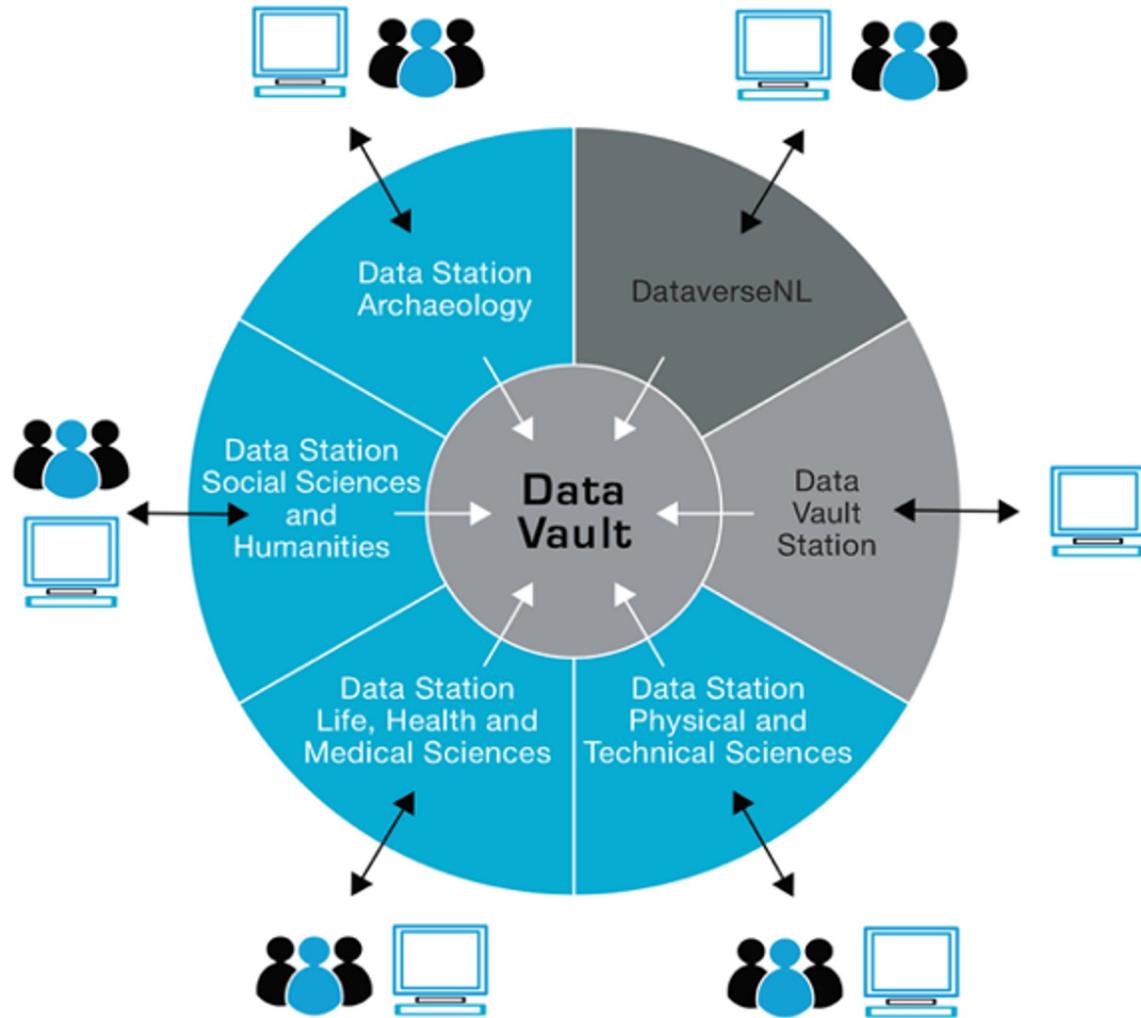
Phase in the Research Data Life Cycle
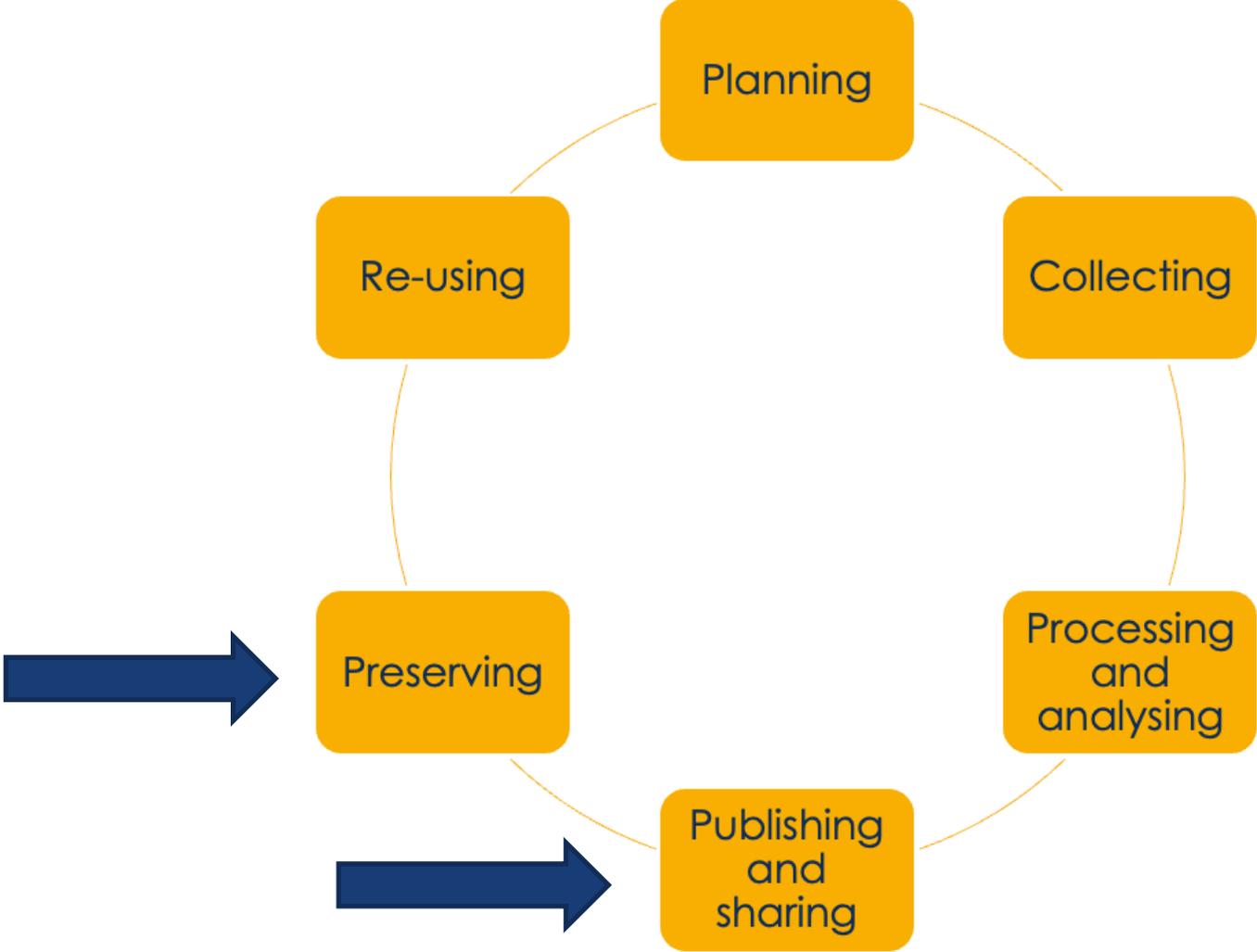
About DANS

Mission:
to enhance the reusability of research data and thus the quality of scientific research

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

The DANS Services

- Domain based Data Stations (DANS is custodian)
- DataverseNL (You are the custodian)
- Long term preservation still at the core (DANS Data Vault)

- Training and consultancy on RDM and FAIR Data
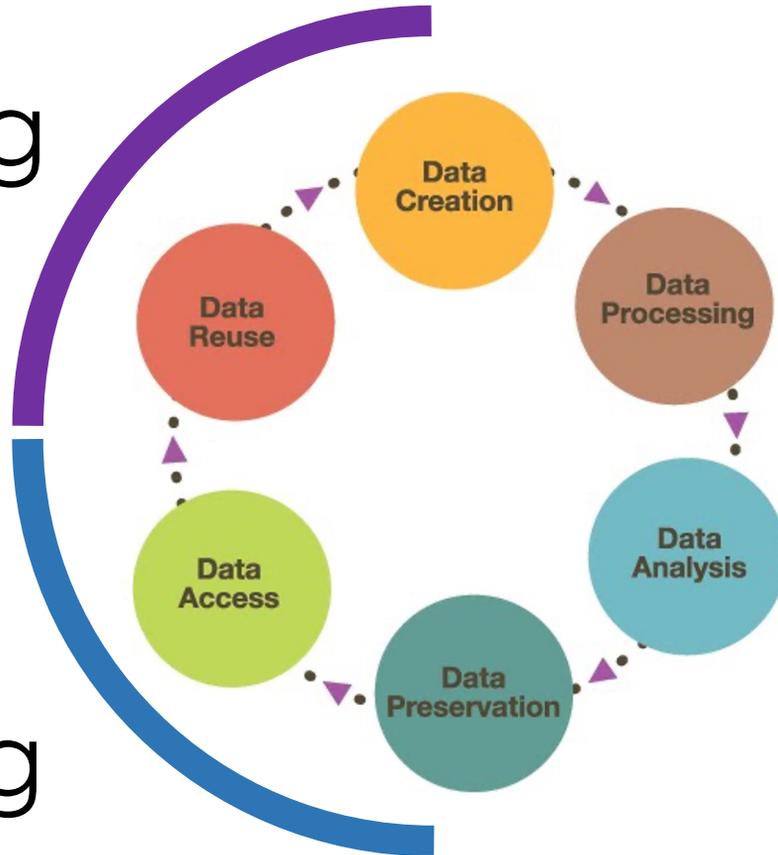
Terminology

Publishing

Sharing

Archiving

The difference between…

Publishing

Archiving

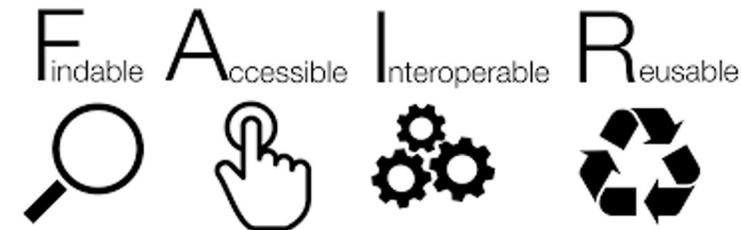Focus on **visibility** and **accessibility** of data and information.

Focus on **long term preservation** and retrievable data and information.

About FAIR and repositories...

Repositories preserve, manage, and provide access to many types of digital materials in a variety of formats. Materials in online repositories are curated to enable search, discovery, and reuse. There must be sufficient control for the digital material to be authentic, reliable, accessible and usable on a continuing basis.
(From previous CASRAI dictionary)

Findable Accessible Interoperable Reusable

- Provides a Persistent Identifier (PID) to your dataset = Findable
- Enables the inclusion of rich metadata = Reusable
- Allows for human and machine findable and readable metadata = Findable + Accessible
- Uses Knowledge Organisation Systems (KOS), such as vocabularies, to improve the quality of metadata = Reusable
- Provides long term and easy access to your data = Accessible + Reusable
- ....

Data Archiving and Networked Services
DANS

health RI
enabling data driven health & life sciences

Questions to ask with regards to data archiving

Ask them to consider:

- For how long must the data be archived?
- Are you bound to a certain discipline or domain?
- Are you dealing with sensitive data or GDPR related issues?
- Are specific (meta)data standards or data formats important for you?
- May your data leave the country?
- What does your funder or organization want you to do….
- What is the size of your data?
- Do you have a "preservation budget"?
- …..

Questions to ask with regards to data publishing

Which way of data publication is your preferred one?

- Supplementary material to paper...
- Own webserver/your institute's system...
- Certified data repository with review...
- Self-publishing through data repository...
- Peer-reviewed data publication (Data in Brief, F1000Research, …)

For what purpose?

Which kind of data publication would you trust?
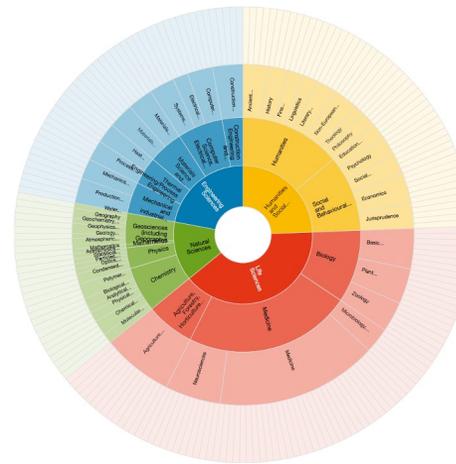
With which method would you reach the biggest audience?

What repositories are out there? For depositing and reuse?

Many Repositories available, see: re3data.org
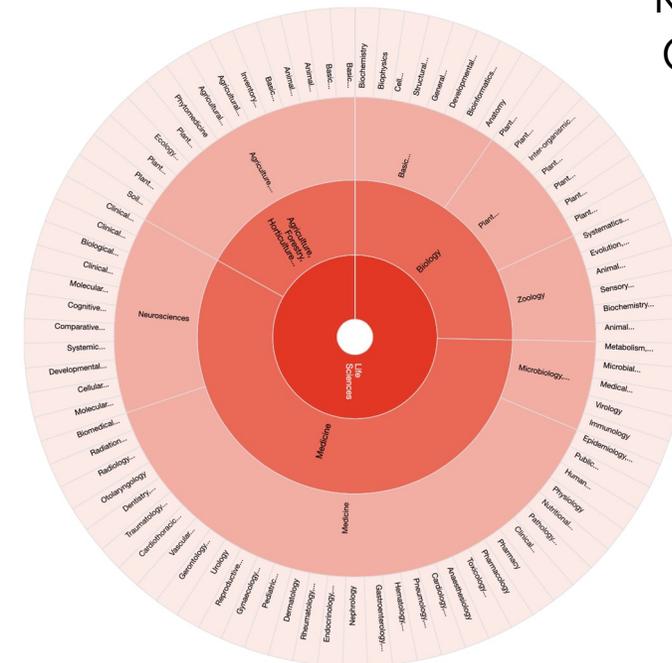
https://www.re3data.org

Over 2600 international repositories for a wide spread of domains
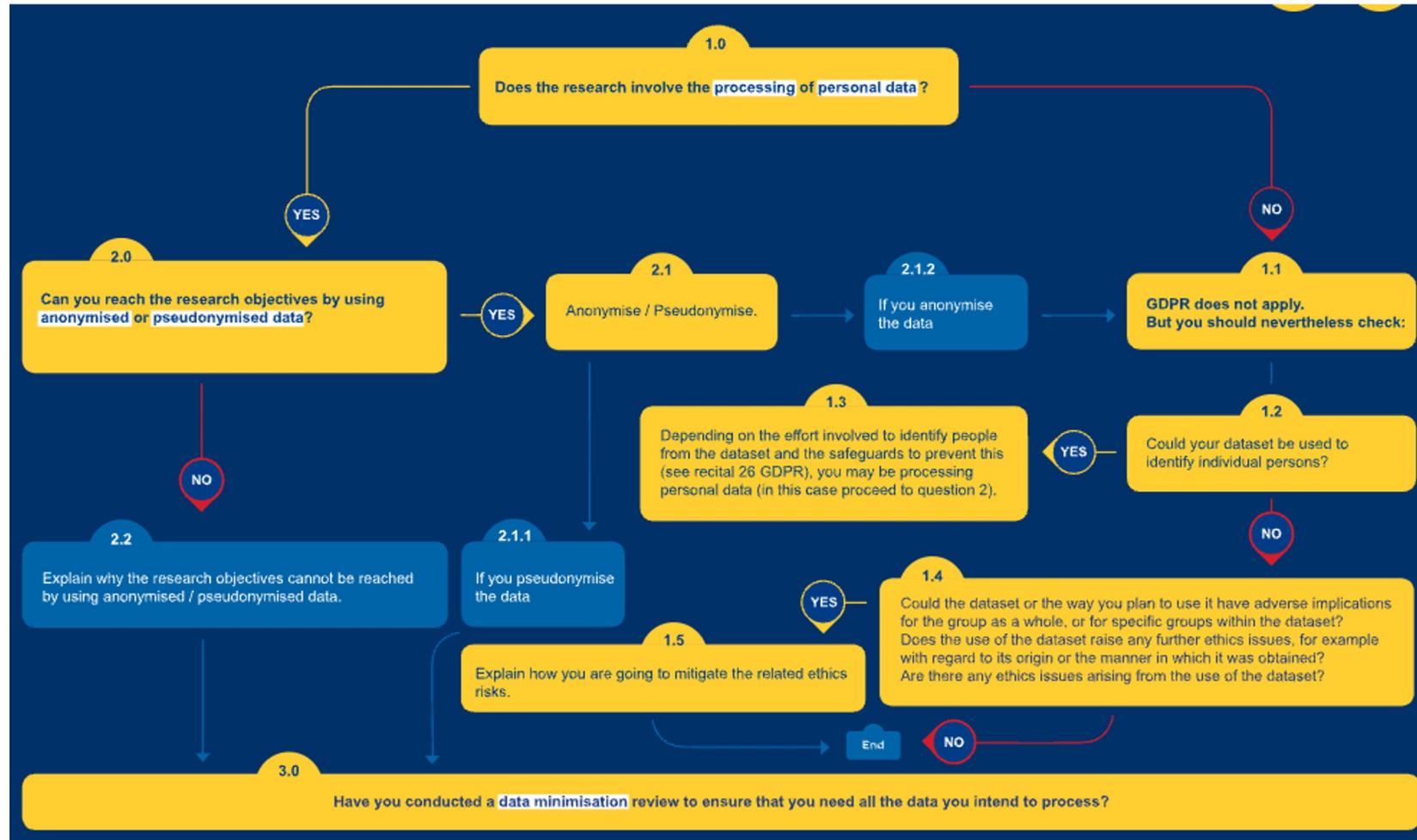
Plentiful filters to find appropriate repositories

All Categories

Life Science
Repositories
Categories

About sensitive data...
The Ethics & Data Protection Decision Tree from the EC

Exercise: Providing advice on archiving and publishing

Timeline:
- 40 minutes, split into two halves (20 min per use case)
- recap in the final 10 minutes of the session

Tools:
- flipcharts
- sticky notes
- your laptops (maybe re3data or the GDPR decision tree?)
- the use cases & the "researchers in need"

Use case 1-a: MRI research on brain damage

Project coordinator: Prof. dr. Erik Swaab, Maastricht UMC+

100 patients with light to severe brain damage, caused by cerebral haemorrhage, were monitored during their first 6 months of their recovery.

Measurements included:
- MRI imaging on a weekly base
- Continues measurements of general physical conditions (full blood picture, etc.)
- Weekly video interviews, monitoring the patient's speech and facial expressions
- Information on family history in relation to cardiovascular diseases

MRI data were all processed using proprietary software from the companies providing the MRI scanners.

The results were captured in a major publication in The Lancet.

All patients have signed an informed consent declaration considering the use of their personal data for research purposes.

Use case 1-b: Scandinavian Socio-economic Health research

Project coordinator: Ingrid Klungland [Kim], University of Oslo, Medical Faculty
(Project is an EU collaboration with researchers at VUmc, who have offered RDM expertise)

About 5000 randomly selected Scandinavian people with a known history of experiencing social problems (domestic violence, alcohol and drug addiction, structural unemployment, etc), are monitored for over 20 years.

Measurements included:
- Elaborate research on family history
- Continues (monthly) measurements of general physical conditions (full blood picture, etc.)
- Yearly video interviews, monitoring the general well being of the participant
- Monitoring of financial situation (income, tax, debts, etc.)
- Environmental data (air, soil, quality) of their near surroundings, based on existing environmental monitoring networks.

The results were captured in a major publication of the joint Scandinavian ministries of health and welfare.
All patients have signed an informed consent declaration considering the use of their personal data for research purposes.

Prompts to consider for both use cases

- How much data is there?
- Would one repository be appropriate for all types of data?
- For how long should the data/output be stored?
- Who should be able to find and re-use the data?
- How much budget do you have for preservation?
- What should be preserved and what not?
- Is this data considered sensitive data?
- Are there funder requirements or project requirements (DMP)?
- Any software and/or code involved?
- What is the publication strategy?
- How FAIR is the data now?

Take aways?

Contact details
kim.ferguson@dans.knaw.nl | Cees.hof@dans.knaw.nl

All emojis designed by OpenMoji – the open-source emoji and icon project. License: CC BY-SA 4.0

Possible answers to prompts, use case 1-a: MRI study

- **How much data is there?** Over 2500 MRI sessions and video interviews, plus continuous monitoring, patient background, documentation, >20 TB?
- **Would one repository be appropriate for all types of data?** No
- **Who should be able to find and re-use the data?** Restricted
- **Is this data considered sensitive data?** Yes, see if anonymisation of MRIs is possible, probably no hope for video?
- **Any software and/or code involved?** Yes, but proprietary software
- **What is the publication strategy?** The Lancet already, which is not usually open access (gold access >€6000 without APC agreement)
- **How FAIR is the data now?** F (sort of) but not AIR
- *Possible re3data repositories:* Donders(NL), OASIS, or studyforrest

Possible answers to prompts, use case 1-b: Scandinavian socio-economic health research

- **How much data is there?** Definitely in the TB

- **Would one repository be appropriate for all types of data?** No, also, might need to be country-by-country rather than a single country

- **Who should be able to find and re-use the data?** Restricted

- **Is this data considered sensitive data?** Yes, see if anonymisation of some data is possible, probably no hope for video?

- **Any software and/or code involved?** No

- **What is the publication strategy?** A joint ministry publication, which is "grey literature," so a scientific publication would be beneficial

- **How FAIR is the data now?** Maybe I, but not F, A, or R

- **Possible re3data repositories:** [Sikt Research Data](NO) for metadata