

# Data management

Andrea Scharnhorst, Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Sciences

[Andrea.scharnhorst@dans.knaw.nl](mailto:Andrea.scharnhorst@dans.knaw.nl)

Guest at Social Computing course, lecture May 11, 2021

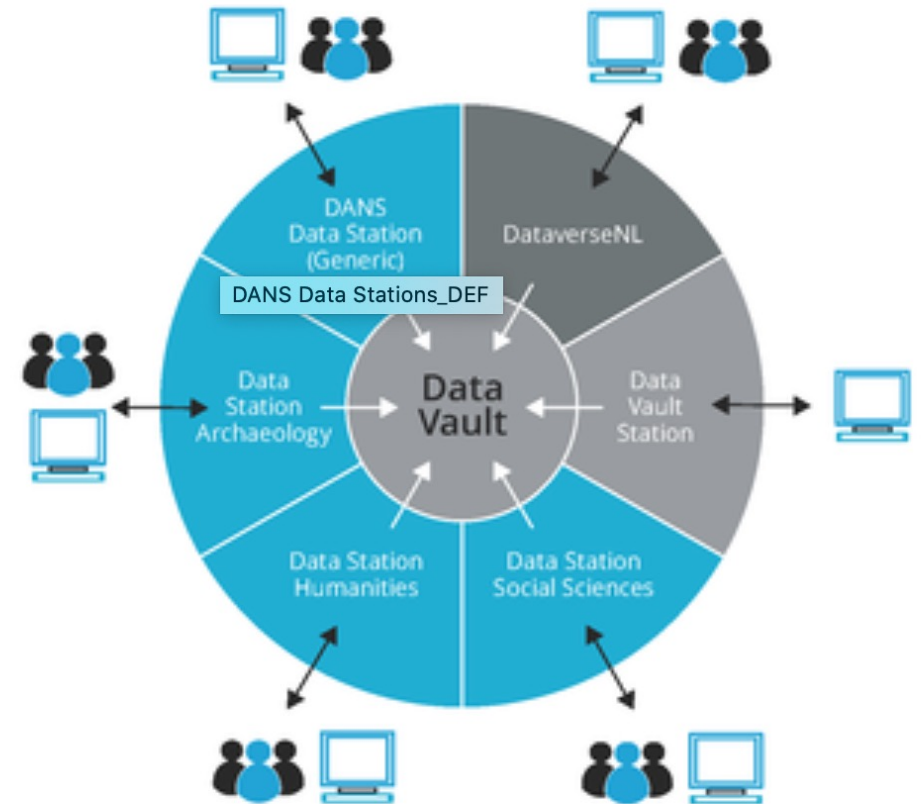
Contact me if you run into Data problems, put a pointer like **UU lecture 2021** in the mail subject header

# Introduction: Where I'm from?

DANS is the Dutch national centre of expertise and repository for research data. DANS is an institute of the Royal Netherlands Academy of Arts and Science (KNAW) and of the Dutch Research Council.

DANS service: Long-term certified archive  
<https://easy.dans.knaw.nl/ui/home>  
(hosting more than 150k datasets across SSH)

Expertise centre for research data and data stewardship (RDA, ERIC's, FAIRsFAIR Coordinator) – training/consultancy



# Who am I?

## Research interests

### Main research areas

Models, mathematical, non-linear, dynamic

- Social change as search in complex knowledge landscapes – framework G\_O\_E\_THE (Geometrically Oriented Evolutionary THEories) and interactive simulations (EVOLINO)
- Complex networks
- Science dynamics and science development
- Innovation dynamics and technological change
- Emergence of norms in social groups and learning strategies under uncertainty
- Emergence of new scientific fields, scientific careers and field mobility
- Evolution of knowledge organization systems (Wikipedia categories, UDC)
- Evolution of scholarly communication and the future of libraries and archives
- Modelling the use of competences for problem solving

Measurement, indicators

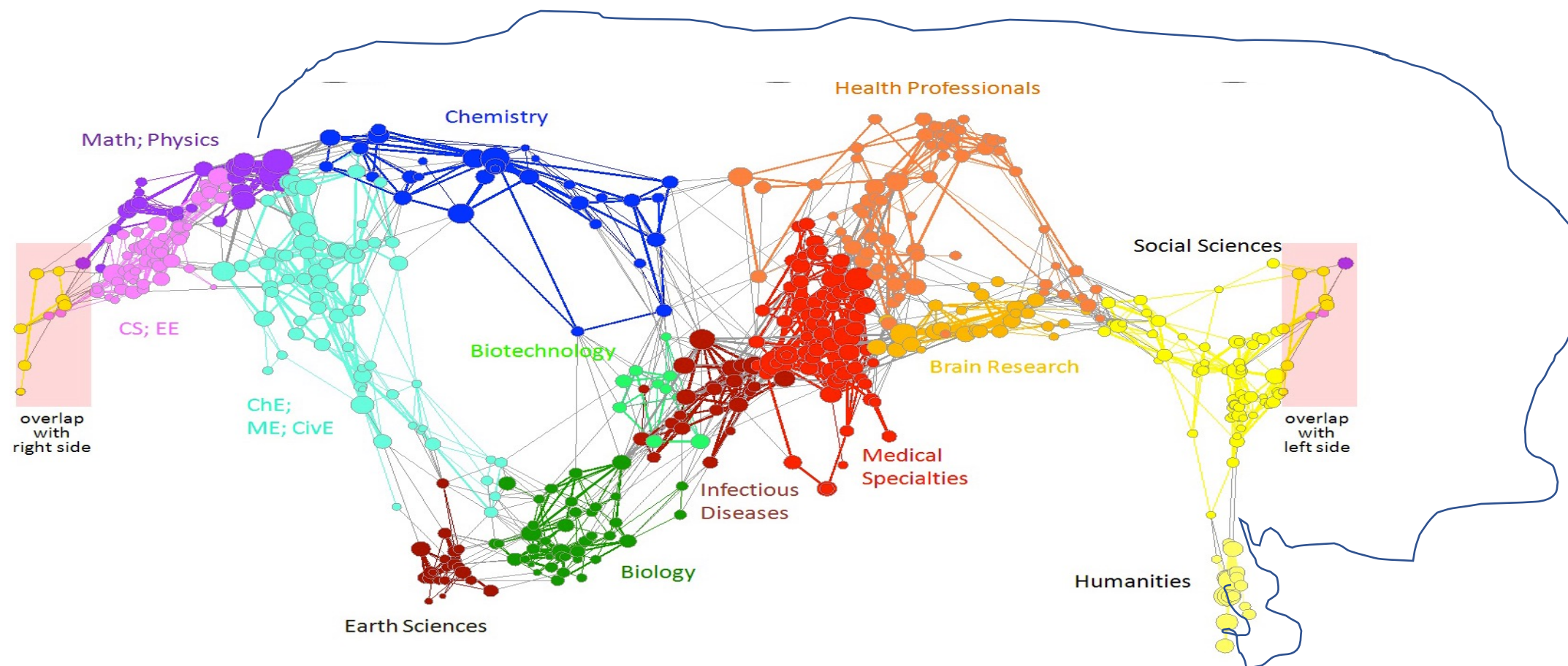
- Scientometrics; Science, technology and innovation indicators (including webindicators)
- Matthew effect of science
- Societal impact of research
- Baseline statistics applied to metadata of collections, user behaviour

### Visualisation

- Science maps
- Knowledge maps for collections
- Visual enhanced interfaces to collections of libraries and archives

### Specialties

- Theoretical physics; statistical physics; socio and econophysics, infophysics
- Philosophy of science, science history and science and technology studies(sts)
- Scientometrics
- Information Sciences



# Learning Topics

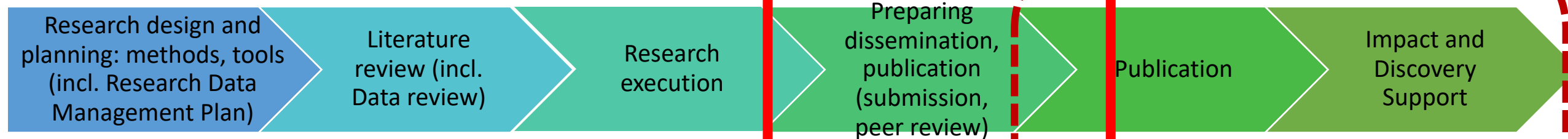
# Goals

- DATA&RESEARCH
  - Where are data in the research life cycle? What is a DataManagementPlan?
- SEARCH
  - What do we know about data search practices?
- RE-USE
  - What does it take to re-use data?
  - How easy do data travel?
  - Data reusability
- DATA STEWARD
  - Support by your library: What offers the UU for you?

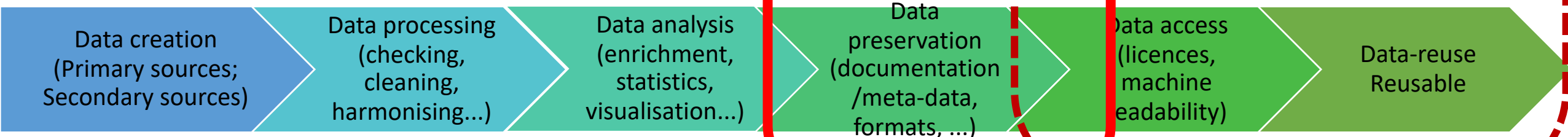
# From the data of others to your data

*Often overlooked:  
Data curation is a shared tasks by researchers and infrastructural support*

## Research cycle



## Data cycle

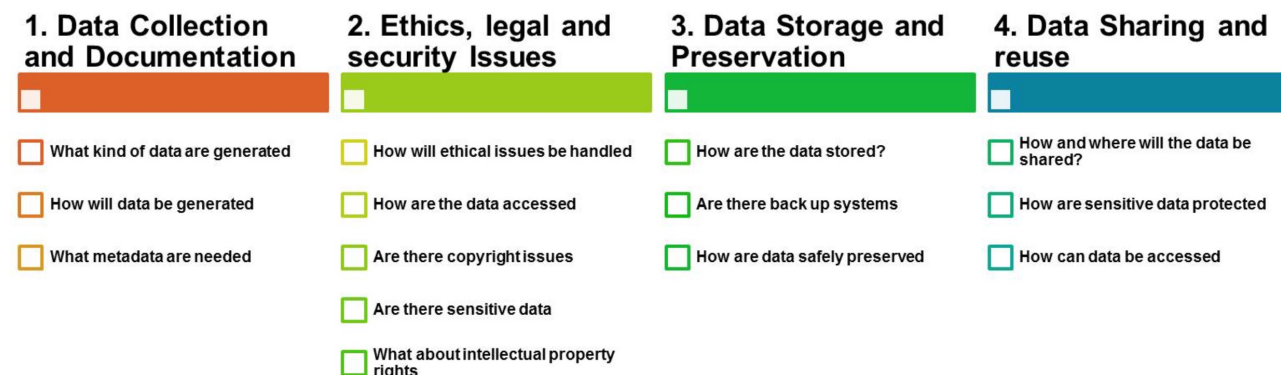


Scientific publishers  
Data repositories

# Data management is intrinsic to good research design

- As part of making research data findable, accessible, interoperable and re-usable (FAIR), a DMP (DataManagementPlan) should include information on:
- the handling of research data during & after the end of the project
- what data will be collected, processed and/or generated
- which methodology & standards will be applied
- whether data will be shared/made open access and
- how data will be curated & preserved (including after the end of the project).

A DMP is a formal document



# Learning Topics

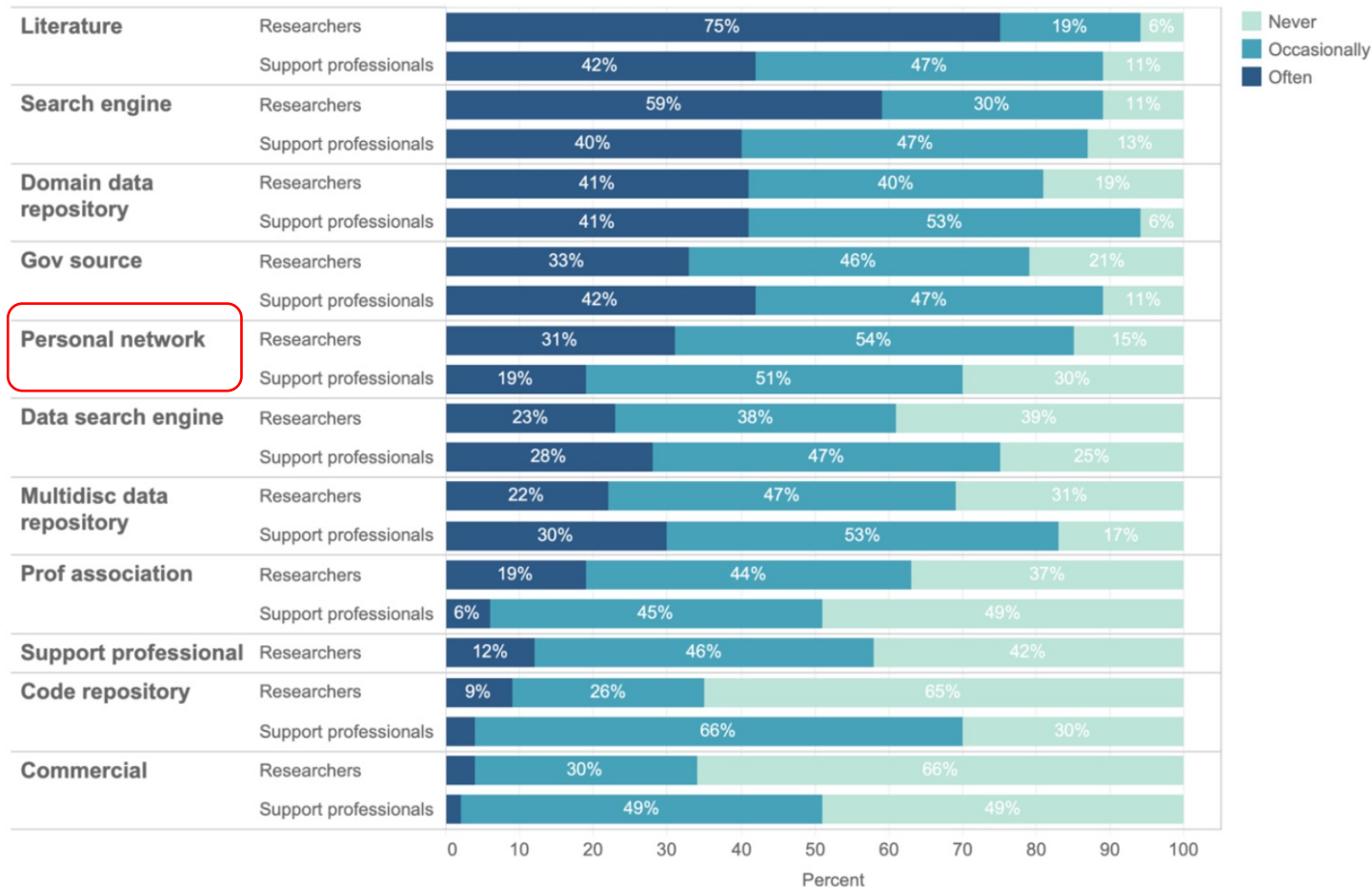
# Goals

- DATA&RESEARCH
  - Where are data in the research life cycle?  
What is a DataManagementPlan?
- SEARCH
  - What do we know about data search practices?
- RE-USE
  - What does it take to re-use data?
  - How easy do data travel?
  - Data reusability
- DATA STEWARD
  - Support by your library: What offers the UU for you?



# Where to find data for your research?

How frequently do you use the following to find data?



Data (re-) use is part of the epistemic culture of a community – hence social networks

“Because “found” data are designed by someone, I always recommend that you try to understand as much as possible about the people and processes that created your data.” Salganik

Figure source:

Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S. 2020, ‘Lost or found? Discovering data needed for research.’

Harvard Data Science Review (accepted)  
<https://arxiv.org/pdf/1909.00464v1.pdf>

Figure 8. Sources used to find data by researchers (including students, managers, and others, n = 1630) and research support professionals (n=47). Percents represent percent of respondents for each category. Listed in order of decreasing importance for researchers.


# Google – Google Scholar – Google Data search Documents – Bibliographic databases – Data registries

<https://datasetsearch.research.google.com>


Google

▼ Last updated ▼ Download format ▼ Usage rights ▼ Topic Free


100+ datasets found




**WHO Coronavirus disease (COVID-19) situation reports**  
www.who.int  
www.kaggle.com  
pdf



**Coronavirus Disease 2019 (COVID-19)**  
www.cdc.gov



**COVID-19 Coronavirus data**  
data.europa.eu  
Updated Apr 21, 2020



**WHO Coronavirus disease (COVID-19) situation reports**  
Explore at [www.who.int](http://www.who.int) Explore at [Kaggle](#)  
270 scholarly articles cite this dataset ([View in Google Scholar](#))  
pdf

**Dataset provided by**  
[World Health Organization](#)

**Area covered**  
Global

**Description**  
Daily situation updates and data regarding the COVID-19 outbreak

<https://datasetsearch.research.google.com>

DataSearch

Filter Results  542103 results for coronavirus covid-19

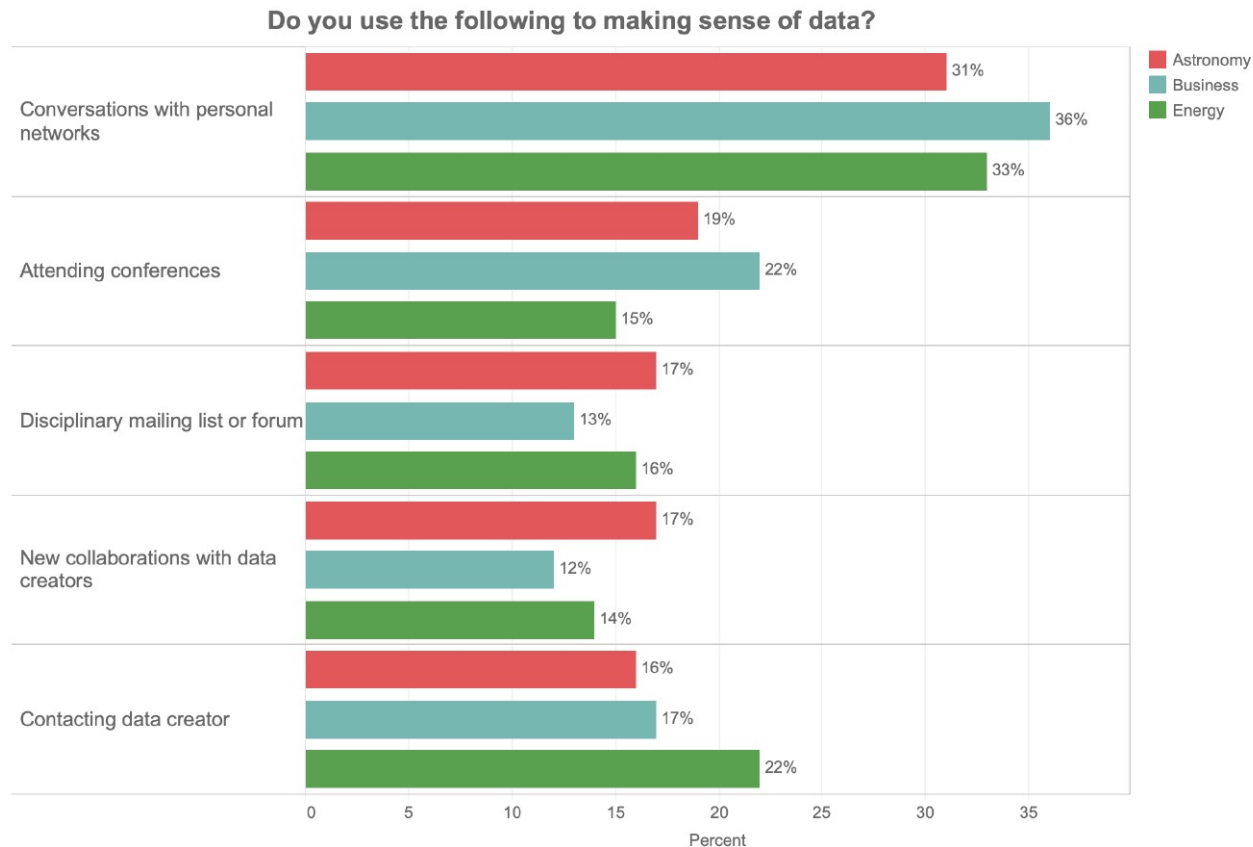
Data File Types	▼	zenodo <a href="#">Ecografía pulmonar en enfermos con coronavirus COVID-19   Lung ultrasound in patients with coronavirus COVID-19 disease</a> Martínez Buendía, Carmen & Martínez López, Félix - 2020-04-13 infecciones por coronavirus...2019 novel coronavirus disease...Una aproximación simple a la ecografía en la enfermedad por coronavirus COVID-19   A practical approach to lung ultrasound in coronavirus COVID-19 disease
Data Source Types	▼	
Data Sources	▼	
Date	▼	zenodo <a href="#">Coronavirus COVID-19 (2019-nCoV) Data Repository for Africa</a> Marivate, Vukosi, Nsoesie, Elaine, Esube Bekele & Africa open COVID-19 data working group - 2020-03-30 url:https://zenodo.org/communities/covid-19...coronavirus...The purpose of this repository is to collate data on the ongoing coronavirus pandemic in Africa. Our goal is to record detailed information on each reported case in every African country. We want to build a line list – a table summarizing information about people who are infected, dead,...
		zenodo <a href="#">Coronavirus COVID-19 (2019-nCoV) Data Repository for Africa</a> Marivate, Vukosi, Nsoesie, Elaine, Esube Bekele & Africa open COVID-19 data working group - 2020-03-30 url:https://zenodo.org/communities/covid-19...coronavirus...The purpose of this repository is to collate data on the ongoing coronavirus pandemic in Africa. Our goal is to record detailed information on each reported case in every African country. We want to build a line list – a table summarizing information about people who are infected, dead,...

▲ Top results from Data Repository sources. [Show only results like these.](#)

zenodo [Coronavirus \(COVID-19\): A new pandemic](#)  
Vijay Kumar, Siprali Priyadarshinee & Sujata Naik - 2020-04-05  
COVID-19...Coronavirus...

The search engines are commercial, but they rely in data gathering on public available information!

# How to make sense of secondary data?



Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: MIT Press.

Figure source:

Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S. 2020, 'Lost or found? Discovering data needed for research.'

Harvard Data Science Review (accepted)

<https://arxiv.org/pdf/1909.00464v1.pdf>

Figure 12. Social strategies of sensemaking in astronomy (n=77), business (n=136), and energy (n=115). Percents are percent responses; multiple responses were allowed.

# Learning Topics

# Goals

- DATA&RESEARCH
  - Where are data in the research life cycle?  
What is a DataManagementPlan?
- SEARCH
  - What do we know about data search practices?
- RE-USE
  - What does it take to re-use data?
  - How easy do data travel?
  - Data reusability
- DATA STEWARD
  - Support by your library: What offers the UU for you?

# What data are typically re-used?

Please select the options that describe the secondary data that you (might) need.

- ☐ **Observational or empirical** (e.g. sensor data, survey data, interview transcripts, sample data, neuroimages, ethnographic data, diaries)
- ☐ **Experimental** (e.g. gene sequences, chromatograms, toroid magnetic field data)
- ☐ **Simulation** (e.g. climate models, economic models)
- ☐ **Derived or compiled** (e.g. text and data mining, compiled database, 3D models)
- ☐ **Other**, Please specify \_\_\_\_\_

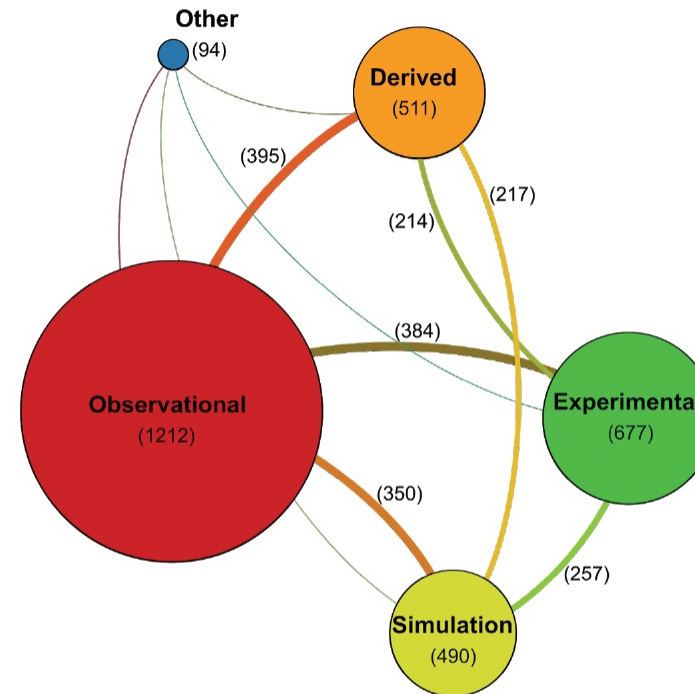


Figure 4. Question from survey with descriptions of data types. Node size in visualization represents number of respondents selecting a data type. Edges represent number of respondents selecting both of the connected data types. Color represents data type. Number of respondents selecting each data type and multiple data types shown in parentheses. (n= 1677).

“Roughly, observational data is any data that results from observing a social system without intervening in some way.”  
Salganik

In Information Science, observational data also come from observing natural systems.

Figure source:

Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S. 2020, ‘Lost or found? Discovering data needed for research.’  
Harvard Data Science Review (accepted)  
<https://arxiv.org/pdf/1909.00464v1.pdf>

# Do data travel across disciplines at all, if context is so important?

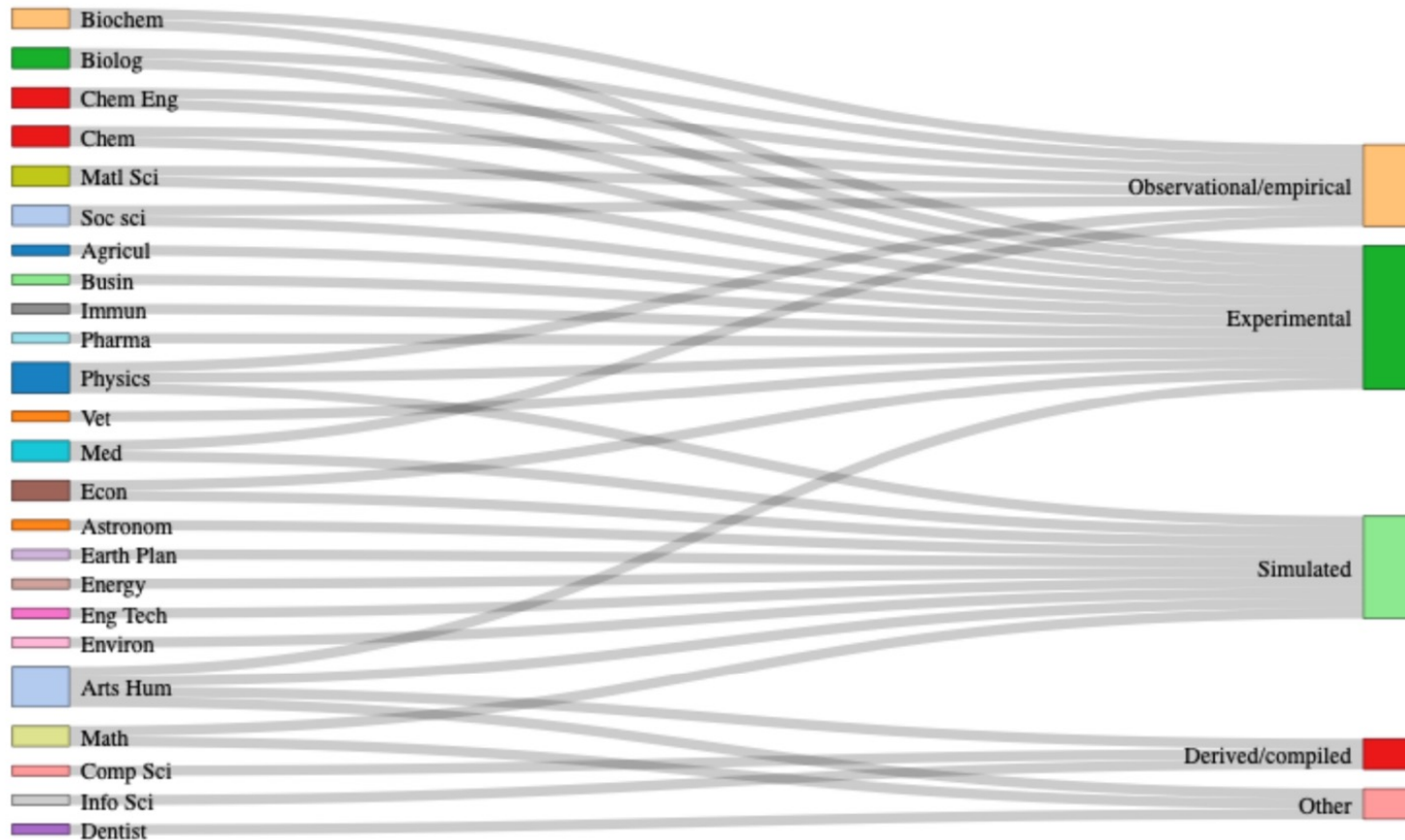


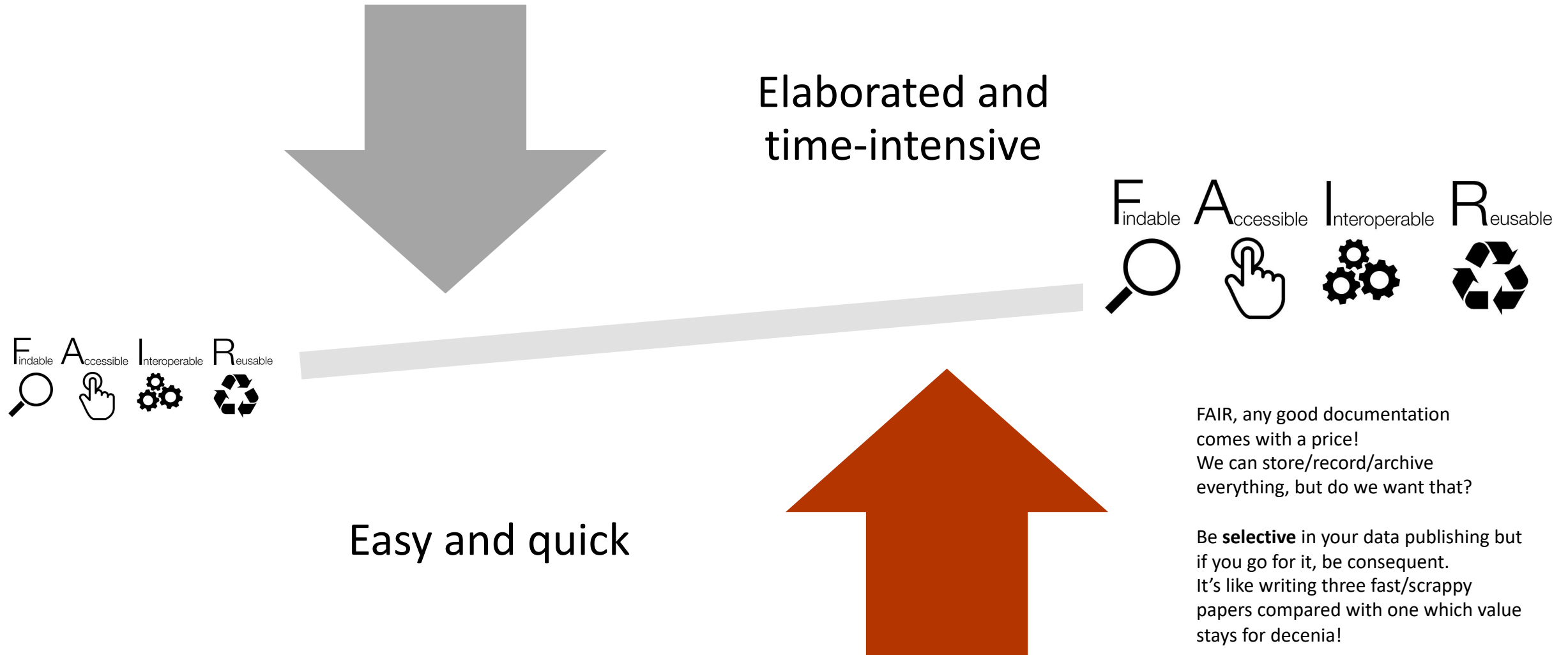
Figure 5. Significant associations between disciplinary domain and needed data. Associations detected using adjusted Bonferroni test for simultaneous pairwise marginal independence (significance level:  $p < 0.05$ ,  $n=1677$ ).

Yes, because the phenomena from which the data come is not confined to one discipline; you can always apply different perspectives on one and the same phenomena. Because disciplines share the data types you might be lucky to rely on data already collected, and re-use them.

Figure source:  
 Gregory, K., Groth, P., Scharnhorst, A.,  
 & Wyatt, S. 2020, 'Lost or found? Discovering  
 data needed for research.'  
 Harvard Data Science Review (accepted)  
<https://arxiv.org/pdf/1909.00464v1.pdf>



# What to do to make data fluid?



# Learning Topics

# Goals

- DATA&RESEARCH
  - Where are data in the research life cycle?  
What is a DataManagementPlan?
- SEARCH
  - What do we know about data search practices?
- RE-USE
  - What does it take to re-use data?
  - How easy do data travel?
  - Data reusability
- DATA-Stewards
  - Support by your library: What offers the UU for you?



# Seek support of data stewards



Utrecht University



Nederlands

[UU.nl](#) > [Research](#) > [Research Data Management Support](#)



**Services and solutions to make research data management work. Want to know how?**

[WATCH THE VIDEO](#) >

Research Data Management Support

[Home](#) [Guides](#) [Tools & Services](#) [Training & Workshops](#) [RDM Stories](#) [FAQ](#) [Contact us](#) [About](#) [Index](#)

# Where to go with my data?

re3data.org

Search Browse Suggest Resources Contact DataCite

Filter

Subjects Content Types Countries AID systems API Certificates Data access Database access Database access restrictions Database licenses Data upload Data upload restrictions Enhanced publication Institution responsibility type Institution type Keywords Metadata standards PID systems Provider types Quality management Repository languages Software Syndications Repository types Versioning

Utrecht

Search

Toggle short help

Sort by

Found 5 result(s)

**YODA**  
Universiteit Utrecht, Your Data - YODA

Subject(s) Humanities and Social Sciences Life Sciences Natural Sciences Engineering Sciences

Content type(s) Standard office documents Images Structured graphics Audiovisual data Scientific and statistical data formats Raw data Archived data Structured text

Country Netherlands

Yoda publishes research data on behalf of researchers that are affiliated with Utrecht University, its research institutes and consortia where it acts as a coordinating body. Data packages are not limited to a particular field of research or license. Yoda publishes data packages via Datacite. To find data publications use: <https://public.yoda.uu.nl/> , or the Datacite search engine : <https://search.datacite.org/data-centers/delft.uu>

**CLAPOP**  
The Dutch CLARIN Portal Pages

Subject(s) Humanities Linguistics Artificial Intelligence, Image and Language Processing Humanities and Social Sciences Computer Science Computer Science, Electrical and System Engineering Engineering Sciences

Content type(s) Standard office documents Audiovisual data Plain text Software applications

Country Netherlands European Union

CLAPOP is the portal of the Dutch CLARIN community. It brings together all relevant resources that were created within the CLARIN NL project and that now are part of the CLARIN NL infrastructure or that were created by other projects but are essential for the functioning of the CLARIN (NL) infrastructure. CLARIN-NL has closely cooperated with CLARIN Flanders in a number of projects. The common results of this cooperation and the results of this cooperation created by CLARIN Flanders are included here as well.

**DataverseNL**

Subject(s) Engineering Sciences Natural Sciences Life Sciences Humanities and Social Sciences

Content type(s) Databases Audiovisual data Raw data other Scientific and statistical data formats Archived data Plain text

Country Netherlands International

Online storage, sharing and registration of research data, during the research period and up to the prescribed term of ten years after its completion. DataverseNL is a shared service provided by participating institutions and DANS.

**TRAILS**  
Tracking Adolescents' Individual Lives Survey

Look into  
<https://www.re3data.org>  
Public enterprise

# Dataverse.nl

## UU



Utrecht University

Utrecht University (Utrecht University)

DataverseNL > Utrecht University

Contact Share



UU Social and  
Behavioural Sciences



UU Science



UU Medicine



UU Geosciences



Search this dataverse...

Find

Advanced Search

☒ Dataverses (56)

☒ Datasets (623)

☐ Files (2,545)

### Dataverse Category

Organization or Institution (12)

Research Project (5)

Research Group (3)

### Publication Year

2019 (438)

2021 (61)

2018 (41)

2017 (30)

2016 (21)

More...

### Author Name

Massen, Jorg (7)

Imperial, Mark (6)

Leferink, Esther (6)

Brown, Susan (5)

De Moor, Tine (4)

More...

### Subject

1 to 10 of 679 Results

Sort ▾

#### Canadian wildfire responder network

Apr 22, 2021 - Collaborative Governance Case Database

Baird, Julia; Summers, Robert; Plummer, Ryan, 2021, "Canadian wildfire responder network",  
<https://doi.org/10.34894/MWGGNQ>, DataverseNL, V1

Data for the Canadian wildfire responder network case. Contains qualitative and quantitative information on the conditions, processes, and outcomes of a specific instance of collaborative governance involving public, private, and/or community actors.

#### Independent Inquiry into Container Deposit Legislation in NSW

Apr 22, 2021 - Collaborative Governance Case Database

Hendriks, Carolyn, 2021, "Independent Inquiry into Container Deposit Legislation in NSW",  
<https://doi.org/10.34894/DXUXQ9>, DataverseNL, V1

Data for the Independent Inquiry into Container Deposit Legislation in NSW case. Contains qualitative and quantitative information on the conditions, processes, and outcomes of a specific instance of collaborative governance involving public, private, and/or community actors.

#### Blackfoot Challenge (Montana, USA)

Apr 22, 2021 - Collaborative Governance Case Database

Weber, Edward, 2021, "Blackfoot Challenge (Montana, USA)",  
<https://doi.org/10.34894/XGTF7D>, DataverseNL, V1

Data for the Blackfoot Challenge (Montana, USA) case. Contains qualitative and quantitative

# Long-term archiving - EASY

<https://easy.dans.knaw.nl/ui/home>

HOME

A. SCHARNHORST

MY DEPOSITS

MY DATASETS

MY REQUESTS

MY SETTINGS

LOG OUT

DANS

EASY

Can we ask you a few questions about EASY? [More information.](#)

We've had some problems with authentication. Currently you can log in with your EASY account. Federated login has been switched off for the moment.

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

wikipedia almila

SEARCH

> Search help

> Advanced search

> Browse

1 RESULT IN PUBLISHED DATASETS

List

Map

> Evolution of Wikipedia Categories

Date: 2012-05-30

Creators: Scharnhorst, A.; Gao, C.; Akdag Salah, A.; Suchecki, K.

Status: Status: published 2015-10-07

New R... 4

Reques... Submitted 2013-07-03

Releva... 100% relevant

Relation: title=Suchecki, K., Akdag Salah, A., Gao, C., & Scharnhorst, A. (2012). Evolution of Wikip...  
Source: Based on Wikipedia dump 2008  
Descri... of categories in the Wikipedia and the Universal Decimal Classification. 2009-2011. Background  
Subject: Wikipedia

Audien... Humanities

Paleography, bibliology, bibliography, library science

Behavioural and educational sciences

Social sciences

Life sciences, medicine and health care

Geodesy, physical geography

Access: Open (everyone)

Submit... 2012-05-29

REFINE

Audience

Behavioural and educational sciences 1

Humanities 2

Life sciences, medicine and health care 1

Science and technology 1

Social sciences 1

Collections

Access

Open (everyone) 1

# Learning Topics

- DATA& RESEARCH
  - Where are data in the research life cycle? What is a DataManagementPlan?
- SEARCH
  - What do we know about data search practices?
- RE-USE
  - What does it take to re-use data?
  - How easy do data travel?
  - Data reusability
- DATA STEWARD
  - Support by your library: What offers the UU for you?

## Goals

You are not alone! Data stewards are there to help you! Data Management Planning is important for your research design not just to fill a formal document.

Data search is agnostic; Learning from your peers is important!

Good documentation. You as your own data-reuser! Again learn from peers. Repurposing data is not easy; but also not impossible.

UU offers support by data stewards, data repositories to share during a project, links to long-term archives.