RESEARCH DATA UNIT

Scientific AI
Hamprecht Lab, IWR,
Heidelberg University

SCIENTIFIC SOFTWARE CENTER

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

AI in research software: Best practices

Research Data Unit: Dr. Georg Schwesinger and Dr. Sebastian Zangerle
Scientific AI group: Peter Lippmann
Scientific Software Center: Dr. Inga Ulusoy

February 2025

1. Requirements of “ML-based science”

What this course is not

- An introduction to data science
- An introduction to machine learning
- A course about different ML algorithms
- A course about different ML training approaches and libraries
- …

What this course is

- A best practices guide to creating machine learning based research software (MLBRS)
- A recommendation on how to manage and prepare your data
- A recommendation on how to train your models
- An introduction to software engineering best practices for MLBRS
- A guideline on how to generate independently reproducible scientific results using data-based approaches
- A guideline on how to publish your data and your models

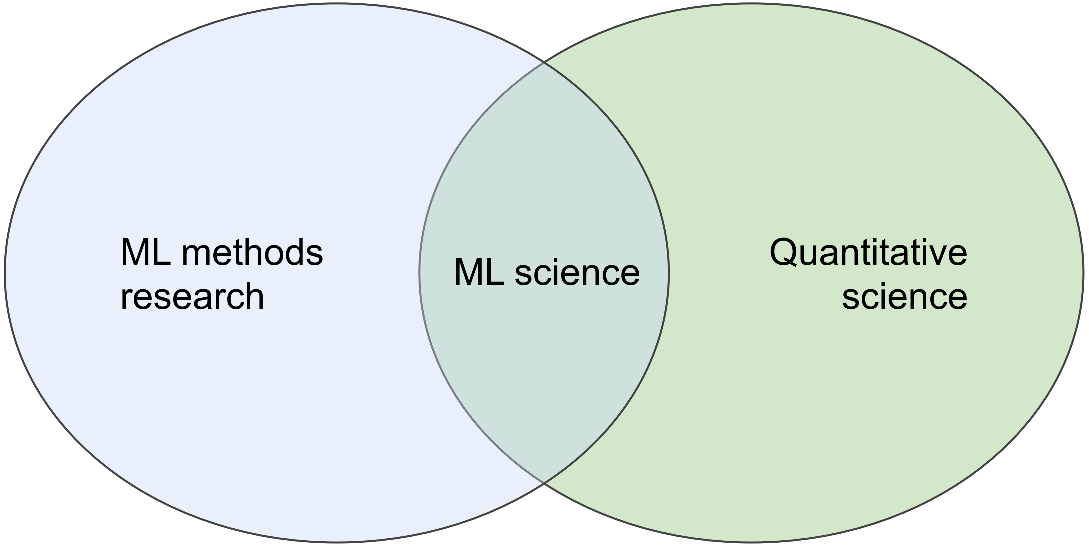
What is special about research software based on data? (“ML-based science”)

ML science

- Scientific research that uses machine learning models to extend scientific knowledge
- Answers a scientific question by using ML
- No restriction on algorithm, method, library, domain

Contrary to:

- ML methods research: Research on ML methods and algorithms with the goal to improve the field of ML



Sayash Kapoor *et al*, http://arxiv.org/abs/2308.07832

Research software

"... software that is developed and used in the context of research…"

Shifting requirements

A scientific question is answered using computation/simulation, but the way the problem is solved changes as part of the research process.

Passed along researchers

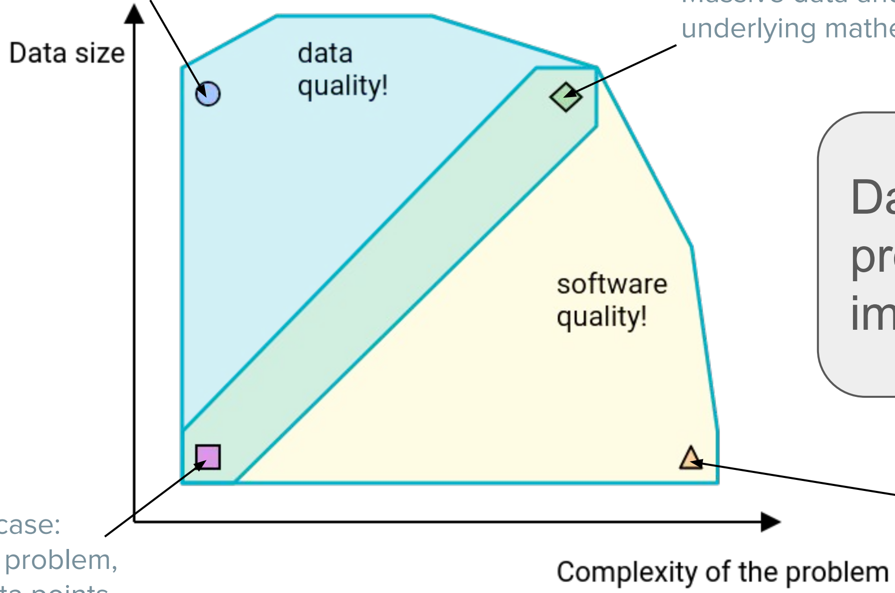
Initially developed for one purpose but then often organically extended depending on the researcher's needs.

Development Practices

Often created by researchers that have no fundamental training in software engineering and inherit practices from those around them.

ML-based research software

“... software that is developed and used in the context of research and predicts outcomes based on initial data…”



Mathematically simpler problem, but large amount of data

Massive data and complex underlying mathematical structure

Mathematically complex problem, but simple data

Trivial case: simple problem, few data points

Data ***and*** software quality provide limits to quality and impact of research!

MLBRS: Data

Data is foundation for..

…model training, decision making and/or predictions.

Different kinds of data

For example, numerical data, textual data, images, audio, video.

Metadata

What is relevant metadata and should be included on the data card?

Availability and licensing

Will the data be publicly available to the community? What license does/will the dataset have?

Legal considerations

Where does the data come from? Is it licensed? Is it public or private data? In what form is the data stored and processed?

Ethical considerations

Does the data exploit work of others? Does it break some sort of confidentiality? Will it impact in a possible harmful way or can it be misconstrued to do harm?

Bias

Is there an inherent bias in the data itself, due to the data collection approach, or other reasons?

MLBRS: Software

Purpose

Will the software be more widely used, be an in-house code, or one-person software?

Software engineering best practices

Does the software follow software engineering best practices (version control, testing, documentation, …)?

Usability and reproducibility

Does the software include documentation on how models can be trained, and keeps track of training parameters? Does the software help to generate model cards and provide models in transferable format?

Accuracy and reliability

Does the software create robust and consistent results, even though it is based on a non-deterministic process?

Legal considerations

Does the software incorporate third-party models and/or code?

Legal considerations

What license is the software published under?

What license are models published under?

Security

Is the software secure against data injection?

Reproducibility

- Provide data to enable others to reproduce findings
- Provide code to enable others to reproduce findings

➔ ***Computational reproducibility (i)***

- Make sure your findings are true findings, and do not arise from problems with your data/code

➔ ***Independent reproducibility (ii)***

Research software engineering generally targets (i), but with MLBRS we target (ii)

Why should you care?

Your research integrity, scientific best conduct (malpractice), can have long-lasting detrimental effect on science (impact on others and the field), affects society!

Key aspects

Legal aspects

Documentation on model training, hyperparameter tuning, model testing

Model bias

Software quality

Reproducibility of the model’s predictions

Robustness of the model(s)

Software security

Reproducibility of the model training

Ethical aspects

Data bias

Documentation on data collection, data cleaning, feature selection

Data leakage

2. Research Data Management

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

AI in research software
Part X: Research Data Management

RESEARCH
DATA UNIT
HEIDELBERG

Dr. Sebastian Zangerle
Heidelberg University Library
sebastian.zangerle@ub.uni-heidelberg.de

Dr. Georg Schwesinger
Heidelberg University Library
schwesinger@ub.uni-heidelberg.de

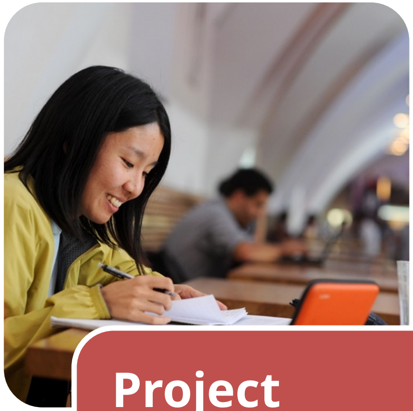
RESEARCH
DATA UNIT
HEIDELBERG

RESEARCH DATA UNIT

Research Data Unit at Heidelberg University

UNIVERSITÄT HEIDELBERG
ZUKUNFT
SEIT 1386

http://data.uni-heidelberg.de/



Project Planning

- Data Management Plans
- Courses & workshops
- Technical &

Data processing

- heiBOX
- heiCLOUD
- SDS@hd
- High Performance Computing

Data Archiving & Publication

- heiDATA
- heidICON
- Archive – your data preserved
- heiARCHIVE

Pictures: © Universität Heidelberg, Kommunikation und Marketing

RESEARCH
DATA UNIT
HEIDELBERG

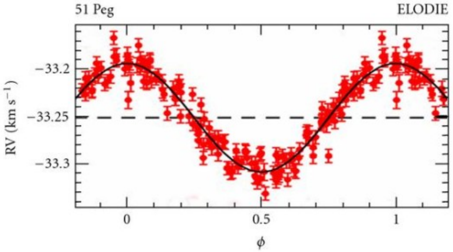
Newsletter Research Data Unit (RDU)

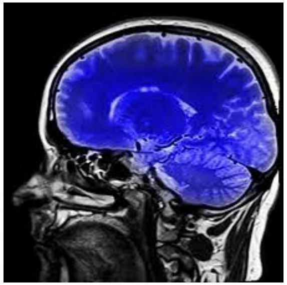
Update information regarding RDM at Heidelberg University:

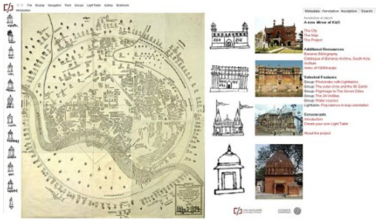
- new services
- current workshops
- Training courses

- Please subscribe to the mailing list: DATA-NEWS@LISTSERV.UNI-HEIDELBERG.DE
- https://listserv.uni-heidelberg.de/cgi-bin/wa?A0=DATA-NEWS

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386





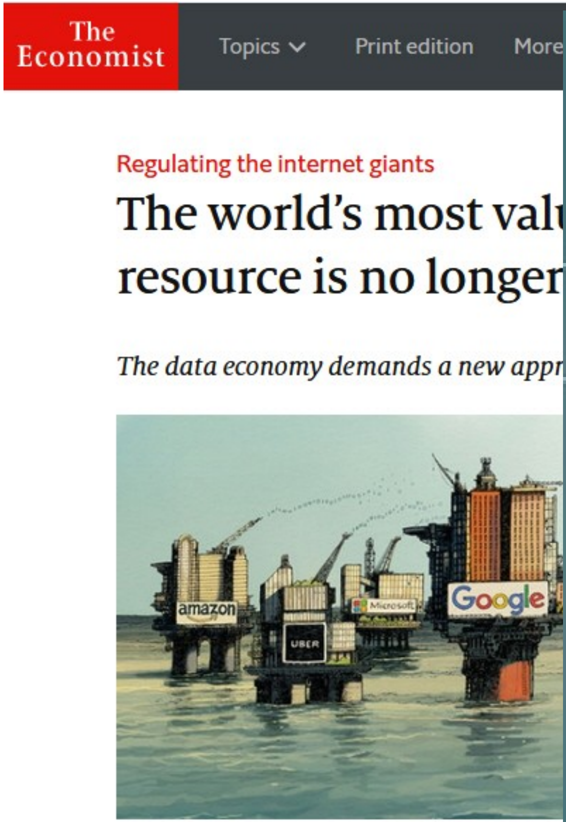


WHAT IS RDM ABOUT?

RESEARCH
DATA UNIT
HEIDELBERG

Data Driven Research

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386



Regulating the internet giants

The world's most valu
resource is no longer

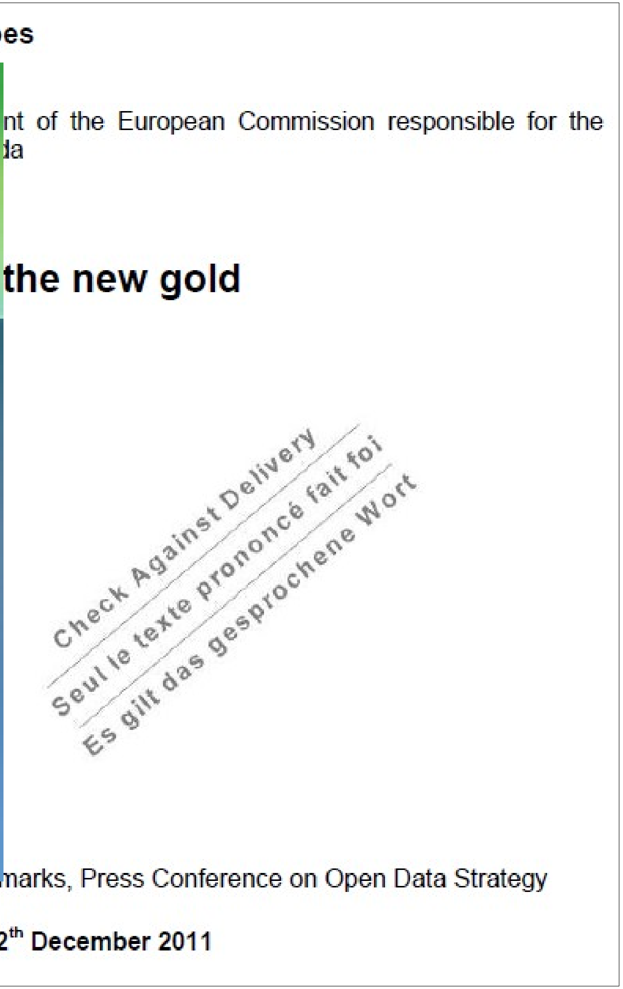
The data economy demands a new appr

Simon C. Lin
Eric Yen Editors

Data Driven e-Science

Use Cases and Successful Applications of Distributed Computing Infrastructures (ISGC 2010)

Springer



Neelie Kroes

nt of the European Commission responsible for the
da

the new gold

Check Against Delivery
Seul le texte prononcé fait foi
Es gilt das gesprochene Wort

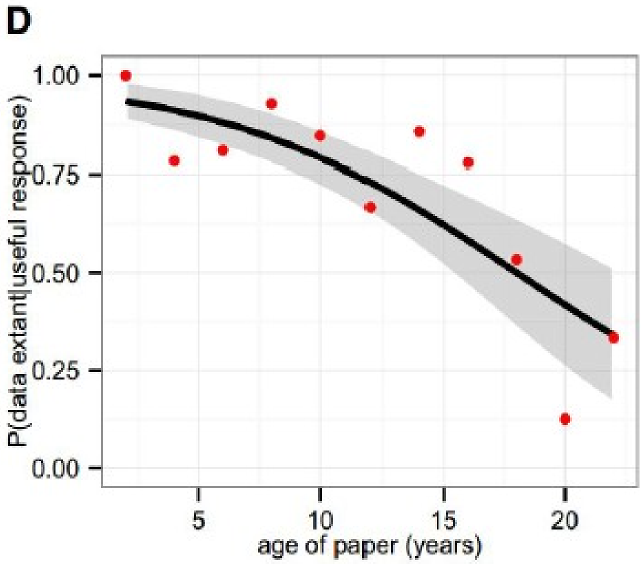
Opening Remarks, Press Conference on Open Data Strategy

Brussels, 12th December 2011

RESEARCH
DATA UNIT
HEIDELBERG

Data Driven Research?

"The underlying data researchers analyze to come to their published conclusions … **becomes less and less accessible to researchers over the years**." (Vines et al, 2014; Dehnhard, Weichselgartner & Krampen, 2013; Wicherts et al, 2006)

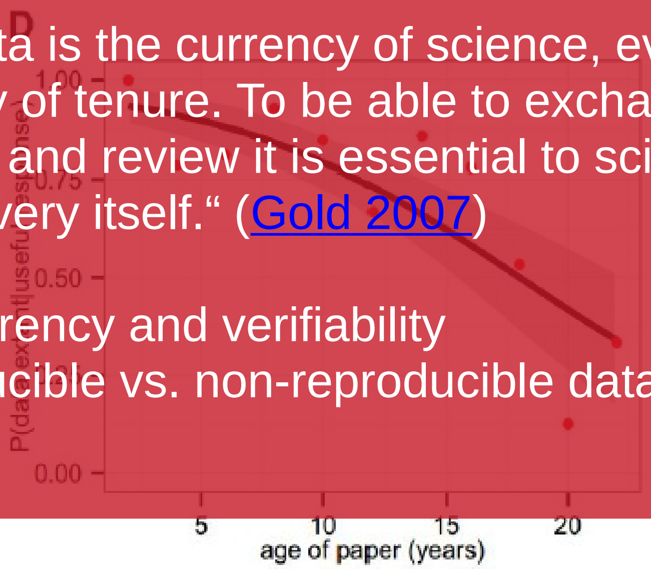


(D) Predicted probability that the data were extant (either "shared" or "exist but unwilling to share") given that we received a useful response.

Slide: Dehnhard 2014

Data Driven Research?

"The underlying data researchers analyze to come to their published conclusions … **becomes less and less accessible to researchers over the years.**" (Vines et al, 2014; Dehnhard, Weichselgartner & Krampen, 2013; Wicherts et al, 2006)



(D) Predicted probability that the data were extant (either "shared" or "exist but unwilling to share") given that we received a useful response.

Why is that disastrous?

- „ […] data is the currency of science, even if publications are still the currency of tenure. To be able to exchange data, communicate it, mine it, reuse it, and review it is essential to scientific productivity, collaboration, and to discovery itself.“ (Gold 2007)

- Transparency and verifiability
- Reproducible vs. non-reproducible data
- Re-use

RESEARCH
DATA UNIT
HEIDELBERG

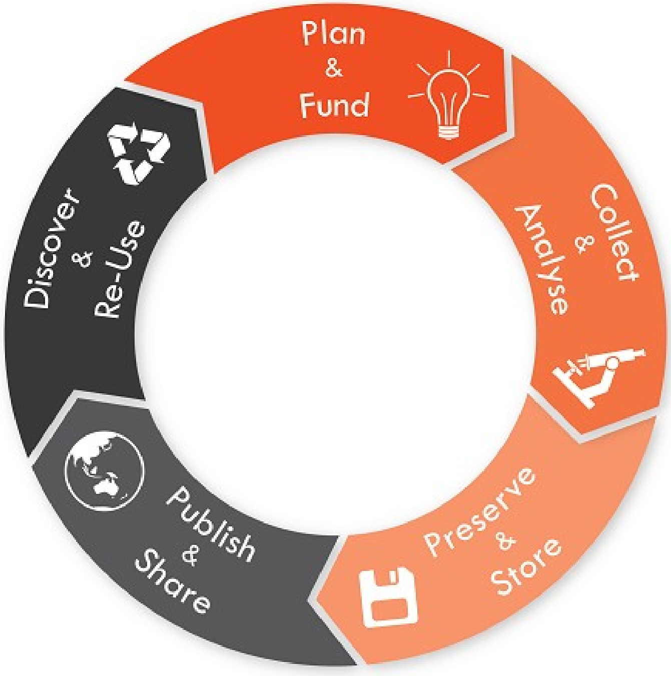
Slide: Dehnhard 2014

What is research data management?

Research data management

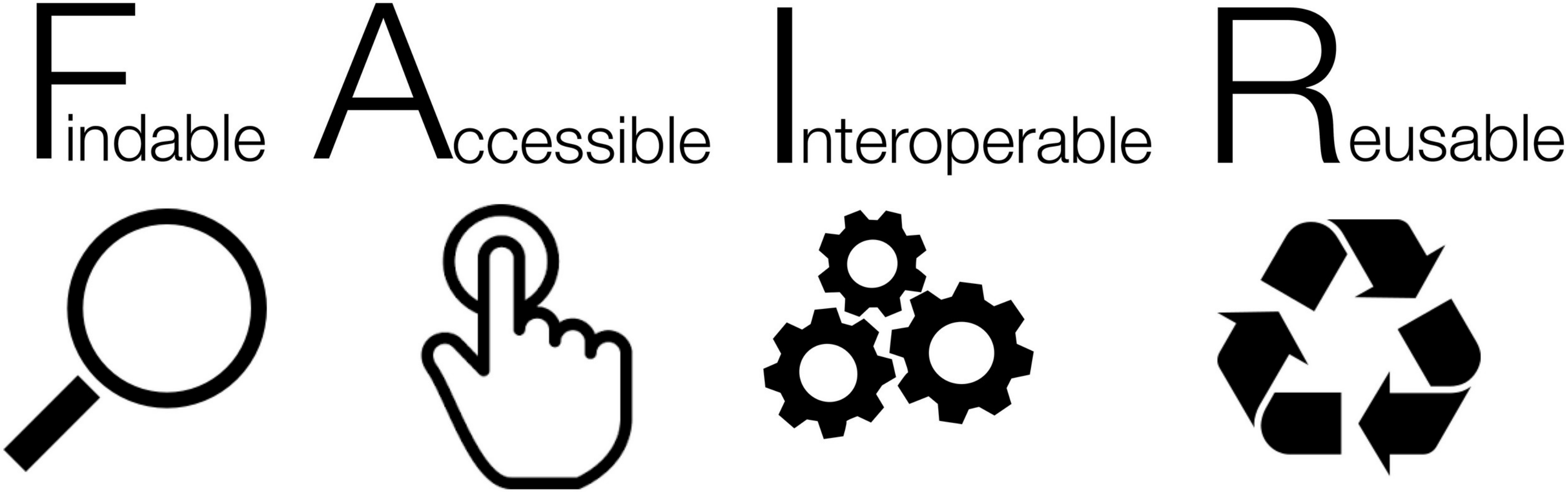
"Research data management concerns the organisation of data, from its entry to the research cycle through to the dissemination and archiving of valuable results. It aims to ensure reliable verification of results, and permits new and innovative research built on existing information."

(Whyte, A., Tedds, J. (2011). 'Making the Case for Research Data Management'. DCC Briefing Papers. Edinburgh: Digital Curation Centre.)



https://library.sydney.edu.au/research/data-management/research-data-management.html

FAIR Data Principles



- FAIR Data Principles
- Wilkinson et al. (2016), The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3, doi:10.1038/sdata.2016.18
- SNF: Explanation of the FAIR Data Principles

By SangyaPundir [CC BY-SA 4.0], from Wikimedia Commons

FAIR Data Principles

F = findable

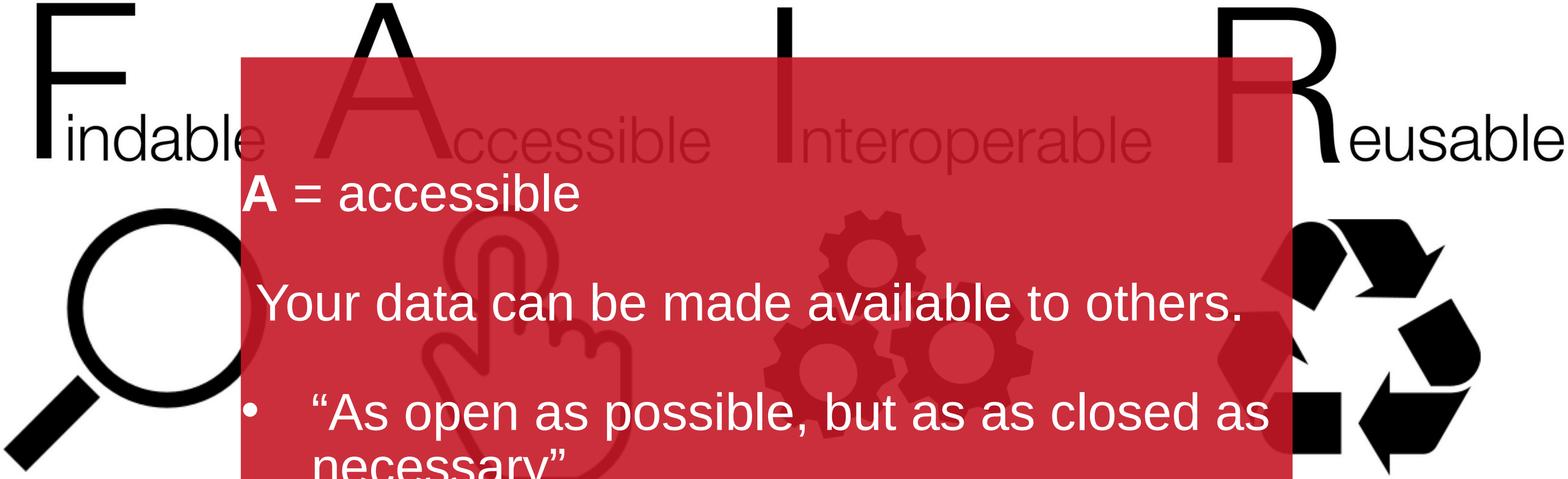
Others can discover your data.

- described via rich metadata,
- persistent identifiers (e.g. DOI),
- indexed in catalogues or databases

- FAIR Data Principles
- Wilkinson et al. (2016), The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3, doi:10.1038/sdata.2016.18
- SNF: Explanation of the FAIR Data Principles

By SangyaPundir [CC BY-SA 4.0], from Wikimedia Commons

FAIR Data Principles



A = accessible

Your data can be made available to others.

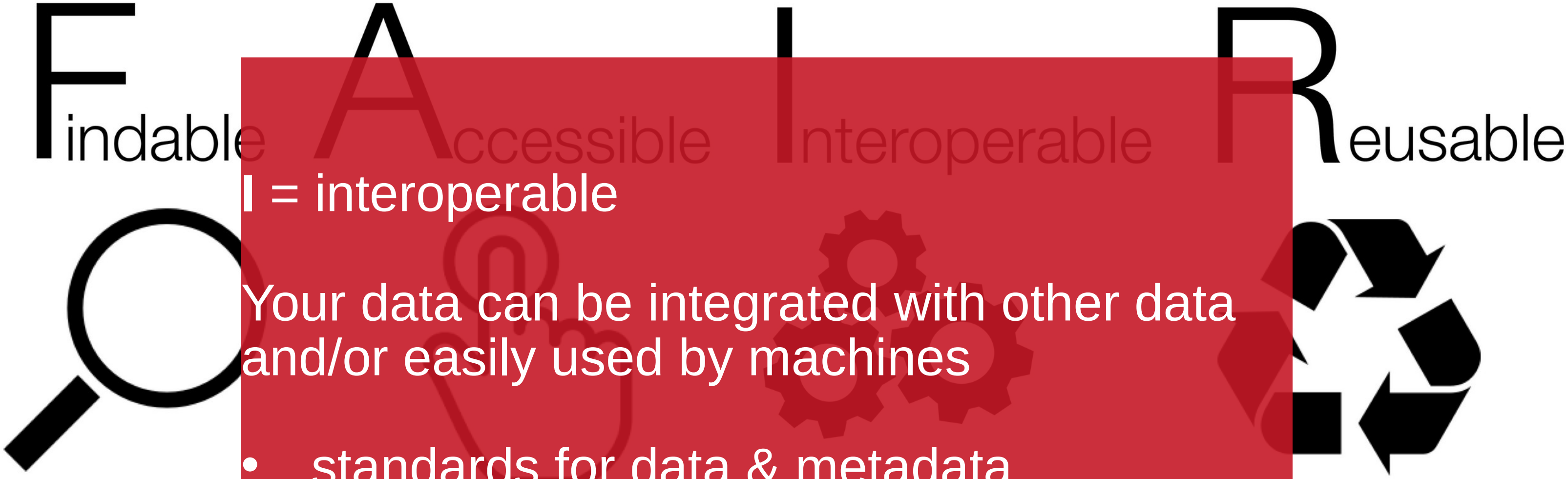
- “As open as possible, but as as closed as necessary”

- FAIR Data Principles
- Wilkinson et al. (2016), The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3, doi:10.1038/sdata.2016.18
- SNF: Explanation of the FAIR Data Principles

RESEARCH
DATA UNIT
HEIDELBERG

By SangyaPundir [CC BY-SA 4.0], from Wikimedia Commons

FAIR Data Principles



By SangyaPundir [CC BY-SA 4.0], from Wikimedia Commons

I = interoperable

Your data can be integrated with other data and/or easily used by machines

- standards for data & metadata
- non-proprietary file formats
- references to other (meta-)data

- FAIR Data Principles
- Wilkinson et al. (2016), The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3, doi:10.1038/sdata.2016.18
- SNF: Explanation of the FAIR Data Principles

FAIR Data Principles

Findable
Accessible
Interoperable
Reusable

R = re-usable

Your data can be used for new research as well as for replications.

- Data are comprehensibly described with relevant attributes,
- domain-relevant standards,
- open licenses,
- provenance

- FAIR Data Principles
- Wilkinson et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3, doi:10.1038/sdata.2016.18
- SNF: Explanation of the FAIR Data Principles

By SangyaPundir [CC BY-SA 4.0], from Wikimedia Commons

Picture: Ainsley Seago. doi:10.1371/journal.pbio.1001779.g001

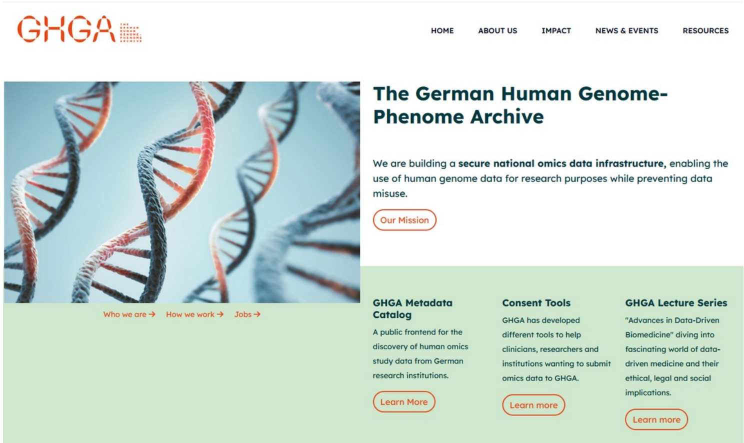
OPEN RESEARCH DATA

NFDI

- **National Research Data Infrastructure**
- The National Research Data Infrastructure (NFDI) has the objective to systematically index, edit, interconnect and make available the valuable stock of data from science and research.
- Funding for subject- and/or methods-specific consortia
- Overall budget: 85 Mio € per year for 10 years
- 27 subject-specific consortia
- https://www.nfdi.de/

NFDI

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386



The German Human Genome-Phenome Archive

We are building a **secure national omics data infrastructure**, enabling the use of human genome data for research purposes while preventing data misuse.

Our Mission

GHGA Metadata Catalog

A public frontend for the discovery of human omics study data from German research institutions.

Learn More

Consent Tools

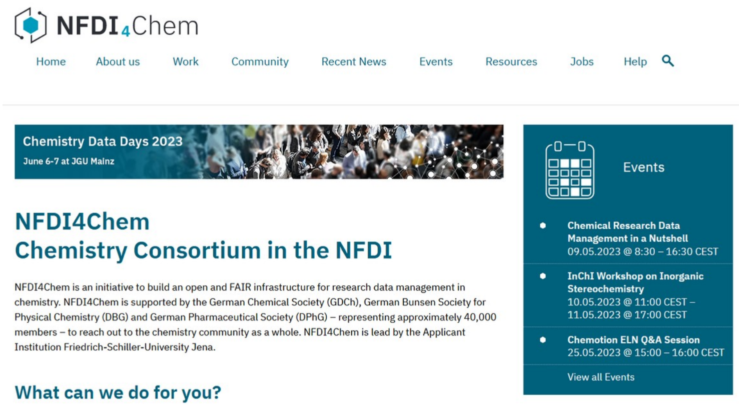
GHGA has developed different tools to help clinicians, researchers and institutions wanting to submit omics data to GHGA.

Learn more

GHGA Lecture Series

"Advances in Data-Driven Biomedicine" diving into fascinating world of data-driven medicine and their ethical, legal and social implications.

Learn more



Chemistry Data Days 2023
June 6-7 at JGU Mainz

NFDI4Chem Chemistry Consortium in the NFDI

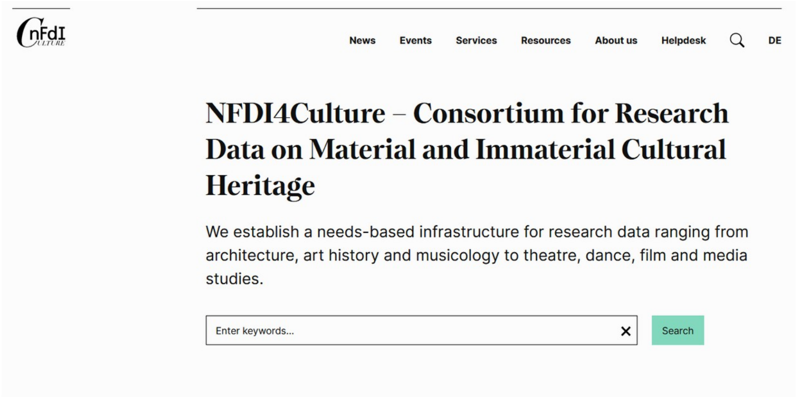
NFDI4Chem is an initiative to build an open and FAIR infrastructure for research data management in chemistry. NFDI4Chem is supported by the German Chemical Society (GDCh), German Bunsen Society for Physical Chemistry (DBG) and German Pharmaceutical Society (DPhG) – representing approximately 40,000 members – to reach out to the chemistry community as a whole. NFDI4Chem is lead by the Applicant Institution Friedrich-Schiller-University Jena.

What can we do for you?

Events

- **Chemical Research Data Management in a Nutshell**
  09.05.2023 @ 8:30 – 16:30 CEST
- **InChI Workshop on Inorganic Stereochemistry**
  10.05.2023 @ 11:00 CEST – 11.05.2023 @ 17:00 CEST
- **Chemotion ELN Q&A Session**
  25.05.2023 @ 15:00 – 16:00 CEST

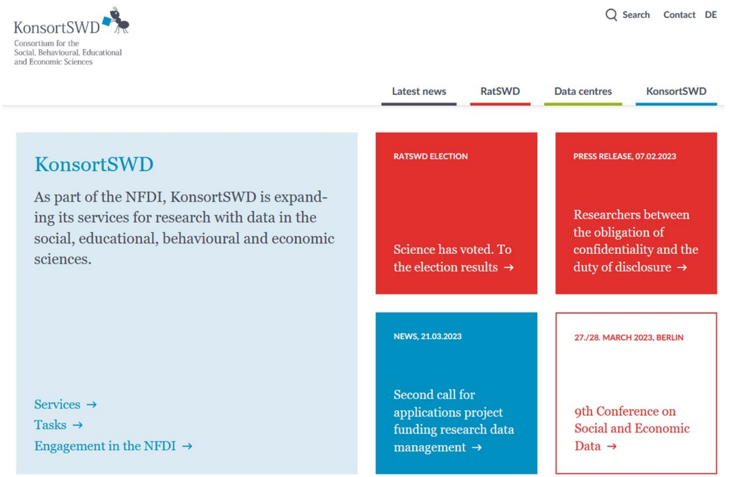
View all Events



NFDI4Culture – Consortium for Research Data on Material and Immaterial Cultural Heritage

We establish a needs-based infrastructure for research data ranging from architecture, art history and musicology to theatre, dance, film and media studies.

Enter keywords... ✕ Search


KonsortSWD
Consortium for the Social, Behavioural, Educational and Economic Sciences
Search
Contact
DE
Latest news
RatSWD
Data centres
KonsortSWD

KonsortSWD

As part of the NFDI, KonsortSWD is expanding its services for research with data in the social, educational, behavioural and economic sciences.

Services →
Tasks →
Engagement in the NFDI →

RATSWD ELECTION

Science has voted. To the election results →

PRESS RELEASE, 07.02.2023

Researchers between the obligation of confidentiality and the duty of disclosure →

NEWS, 21.03.2023

Second call for applications project funding research data management →

27./28. MARCH 2023, BERLIN

9th Conference on Social and Economic Data →

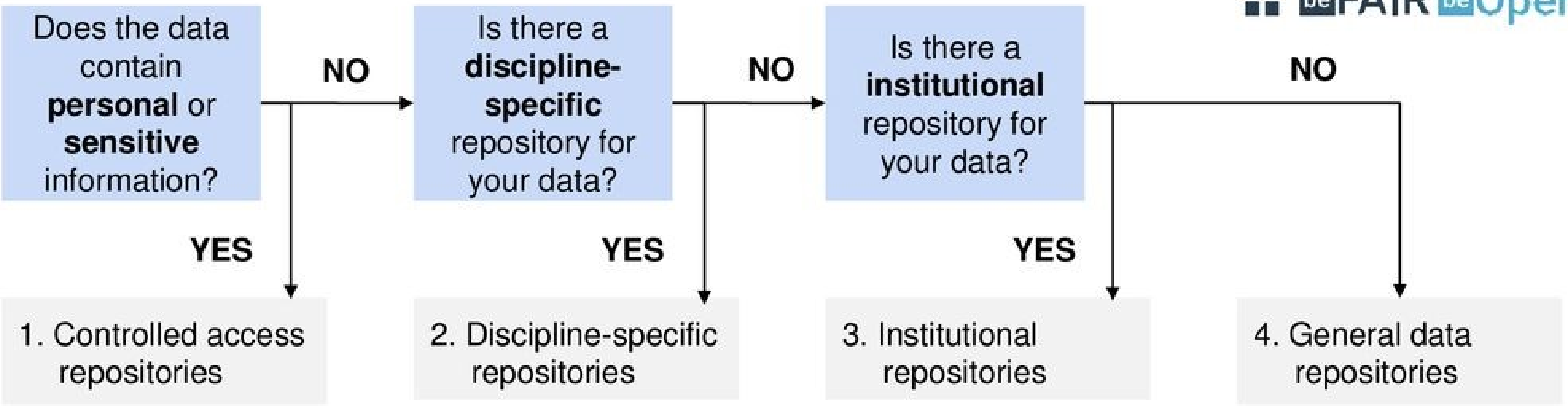
RESEARCH DATA UNIT HEIDELBERG

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Repositories

SciLifeLab

F1000 beFAIR beOpen



Does the data contain **personal** or **sensitive** information?

NO

Is there a **discipline-specific** repository for your data?

NO

Is there a **institutional** repository for your data?

NO

YES

1. Controlled access repositories

YES

2. Discipline-specific repositories

YES

3. Institutional repositories

4. General data repositories

EUROPEAN GENOME-PHENOME ARCHIVE

ENA European Nucleotide Archive

BioSamples

NCBI

dbSNP Short Genetic Variations

PRIDE

Protein Data Bank in Europe Bringing Structure to Biology

heiDATA

heiDATA Heidelberg Open Research Data

figshare

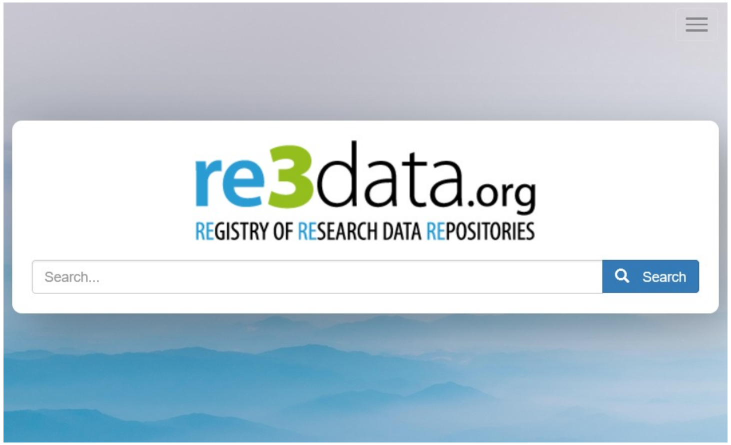
zenodo

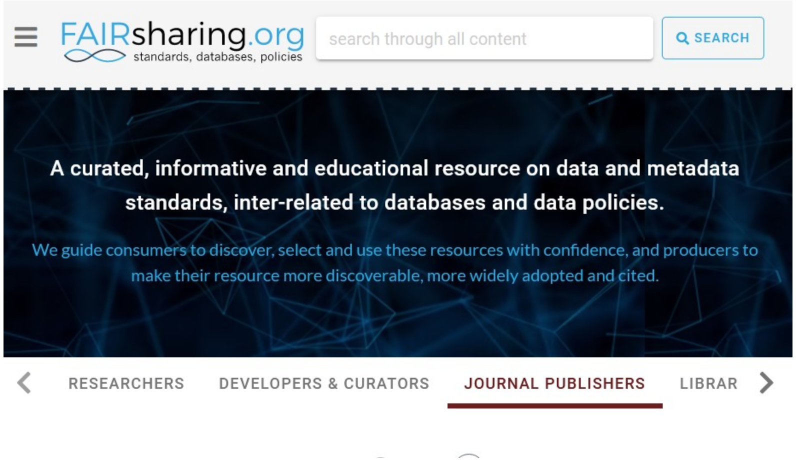
EUDAT

Etc...

Slide adapted from: N. Jareborg (2019), „Data management and repositories“, https://player.slideplayer.com/105/17629367/.

Finding repositories





A curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies.

We guide consumers to discover, select and use these resources with confidence, and producers to make their resource more discoverable, more widely adopted and cited.

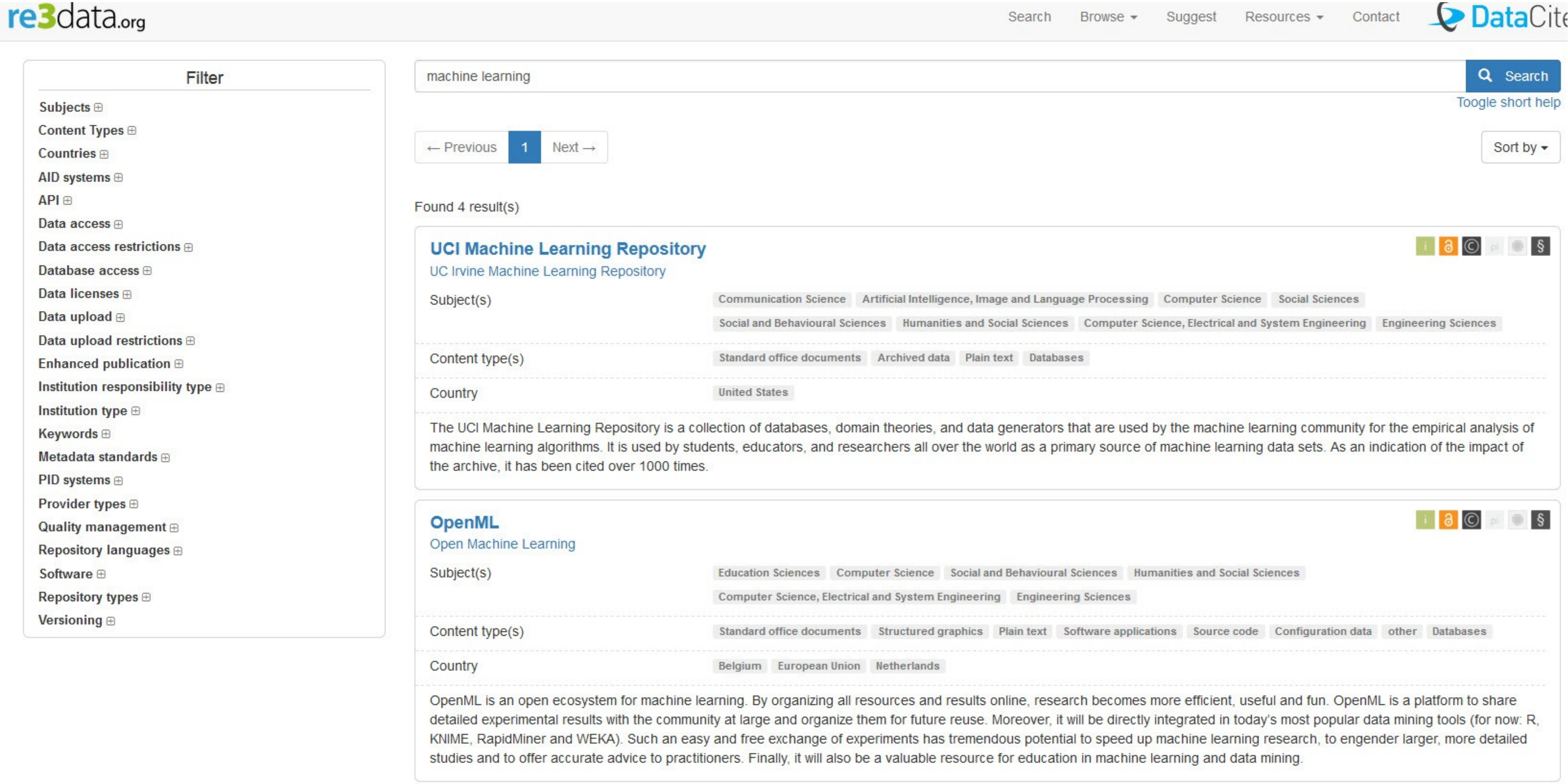
https://www.re3data.org/

https://fairsharing.org/

Publisher Guidelines

- https://www.nature.com/sdata/policies/repositories
- https://journals.plos.org/plosone/s/recommended-repositories
- https://www.springernature.com/gp/authors/research-data-policy/recommended-repositories

Data publication



Filter

- Subjects ⊞
- Content Types ⊞
- Countries ⊞
- AID systems ⊞
- API ⊞
- Data access ⊞
- Data access restrictions ⊞
- Database access ⊞
- Data 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 ⊞
- Repository types ⊞
- Versioning ⊞

machine learning

Search

Toogle short help

← Previous 1 Next →

Sort by

Found 4 result(s)

UCI Machine Learning Repository

UC Irvine Machine Learning Repository

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

Content type(s): Standard office documents, Archived data, Plain text, Databases

Country: United States

The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. It is used by students, educators, and researchers all over the world as a primary source of machine learning data sets. As an indication of the impact of the archive, it has been cited over 1000 times.

OpenML

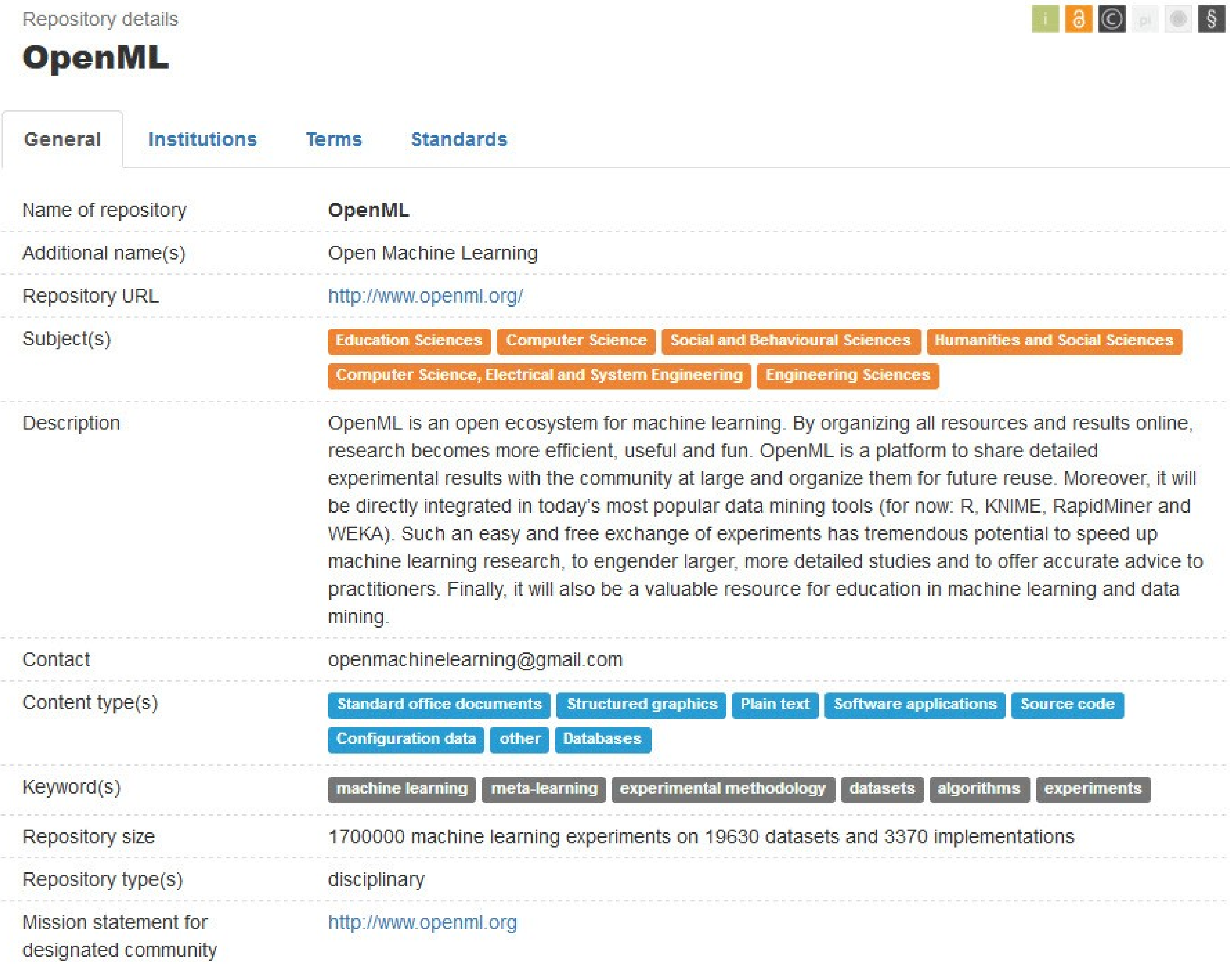
Open Machine Learning

Subject(s): Education Sciences, Computer Science, Social and Behavioural Sciences, Humanities and Social Sciences, Computer Science, Electrical and System Engineering, Engineering Sciences

Content type(s): Standard office documents, Structured graphics, Plain text, Software applications, Source code, Configuration data, other, Databases

Country: Belgium, European Union, Netherlands

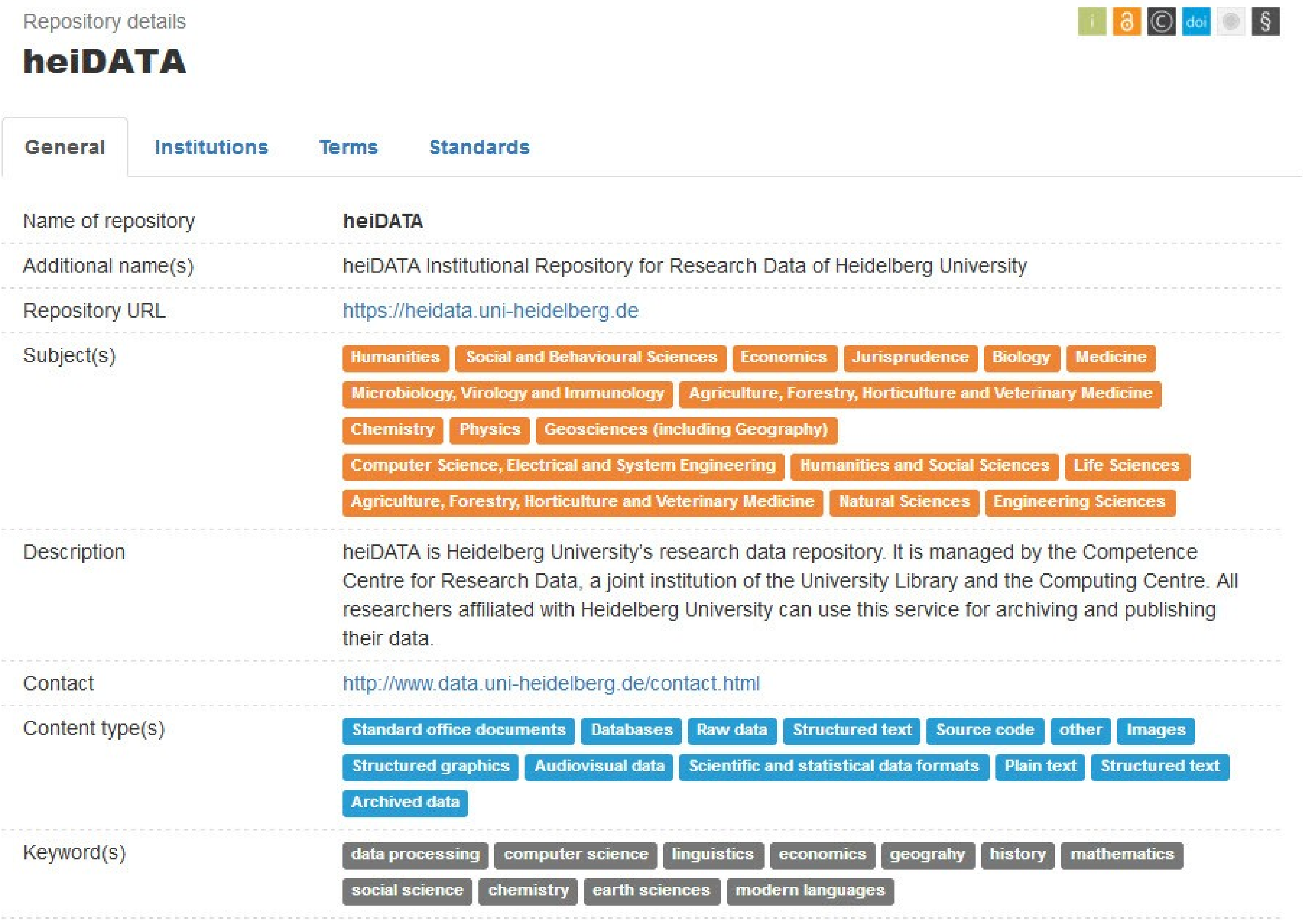
OpenML is an open ecosystem for machine learning. By organizing all resources and results online, research becomes more efficient, useful and fun. OpenML is a platform to share detailed experimental results with the community at large and organize them for future reuse. Moreover, it will be directly integrated in today's most popular data mining tools (for now: R, KNIME, RapidMiner and WEKA). Such an easy and free exchange of experiments has tremendous potential to speed up machine learning research, to engender larger, more detailed studies and to offer accurate advice to practitioners. Finally, it will also be a valuable resource for education in machine learning and data mining.



Repository details

OpenML

Name of repository	**OpenML**
Additional name(s)	Open Machine Learning
Repository URL	http://www.openml.org/
Subject(s)	Education Sciences Computer Science Social and Behavioural Sciences Humanities and Social Sciences Computer Science, Electrical and System Engineering Engineering Sciences
Description	OpenML is an open ecosystem for machine learning. By organizing all resources and results online, research becomes more efficient, useful and fun. OpenML is a platform to share detailed experimental results with the community at large and organize them for future reuse. Moreover, it will be directly integrated in today's most popular data mining tools (for now: R, KNIME, RapidMiner and WEKA). Such an easy and free exchange of experiments has tremendous potential to speed up machine learning research, to engender larger, more detailed studies and to offer accurate advice to practitioners. Finally, it will also be a valuable resource for education in machine learning and data mining.
Contact	openmachinelearning@gmail.com
Content type(s)	Standard office documents Structured graphics Plain text Software applications Source code Configuration data other Databases
Keyword(s)	machine learning meta-learning experimental methodology datasets algorithms experiments
Repository size	1700000 machine learning experiments on 19630 datasets and 3370 implementations
Repository type(s)	disciplinary
Mission statement for designated community	http://www.openml.org



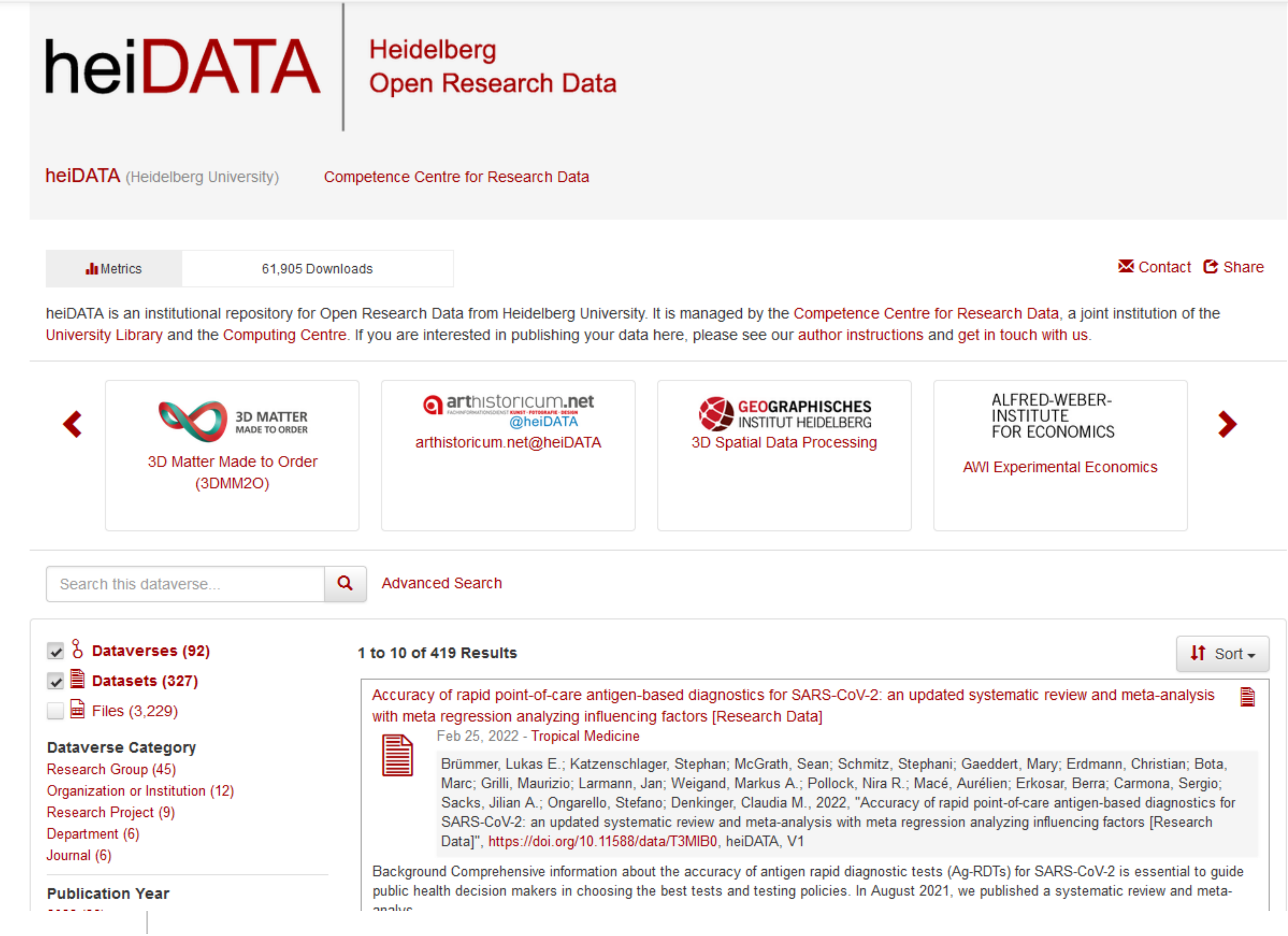
Repository details

heiDATA

Name of repository	**heiDATA**
Additional name(s)	heiDATA Institutional Repository for Research Data of Heidelberg University
Repository URL	https://heidata.uni-heidelberg.de
Subject(s)	Humanities Social and Behavioural Sciences Economics Jurisprudence Biology Medicine Microbiology, Virology and Immunology Agriculture, Forestry, Horticulture and Veterinary Medicine Chemistry Physics Geosciences (including Geography) Computer Science, Electrical and System Engineering Humanities and Social Sciences Life Sciences Agriculture, Forestry, Horticulture and Veterinary Medicine Natural Sciences Engineering Sciences
Description	heiDATA is Heidelberg University's research data repository. It is managed by the Competence Centre for Research Data, a joint institution of the University Library and the Computing Centre. All researchers affiliated with Heidelberg University can use this service for archiving and publishing their data.
Contact	http://www.data.uni-heidelberg.de/contact.html
Content type(s)	Standard office documents Databases Raw data Structured text Source code other Images Structured graphics Audiovisual data Scientific and statistical data formats Plain text Structured text Archived data
Keyword(s)	data processing computer science linguistics economics geograhy history mathematics social science chemistry earth sciences modern languages

heiDATA

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386



heiDATA | Heidelberg Open Research Data

heiDATA (Heidelberg University)   Competence Centre for Research Data

heiDATA is an institutional repository for Open Research Data from Heidelberg University. It is managed by the Competence Centre for Research Data, a joint institution of the University Library and the Computing Centre. If you are interested in publishing your data here, please see our author instructions and get in touch with us.

3D MATTER MADE TO ORDER
3D Matter Made to Order (3DMM2O)
arthistoricum.net @heiDATA
arthistoricum.net@heiDATA
GEOGRAPHISCHES INSTITUT HEIDELBERG
3D Spatial Data Processing
ALFRED-WEBER-INSTITUTE FOR ECONOMICS
AWI Experimental Economics
Search this dataverse...
Advanced Search

☑ Dataverses (92)
☑ Datasets (327)
□ Files (3,229)

Dataverse Category

- Research Group (45)
- Organization or Institution (12)
- Research Project (9)
- Department (6)
- Journal (6)

Publication Year

1 to 10 of 419 Results

Accuracy of rapid point-of-care antigen-based diagnostics for SARS-CoV-2: an updated systematic review and meta-analysis with meta regression analyzing influencing factors [Research Data]

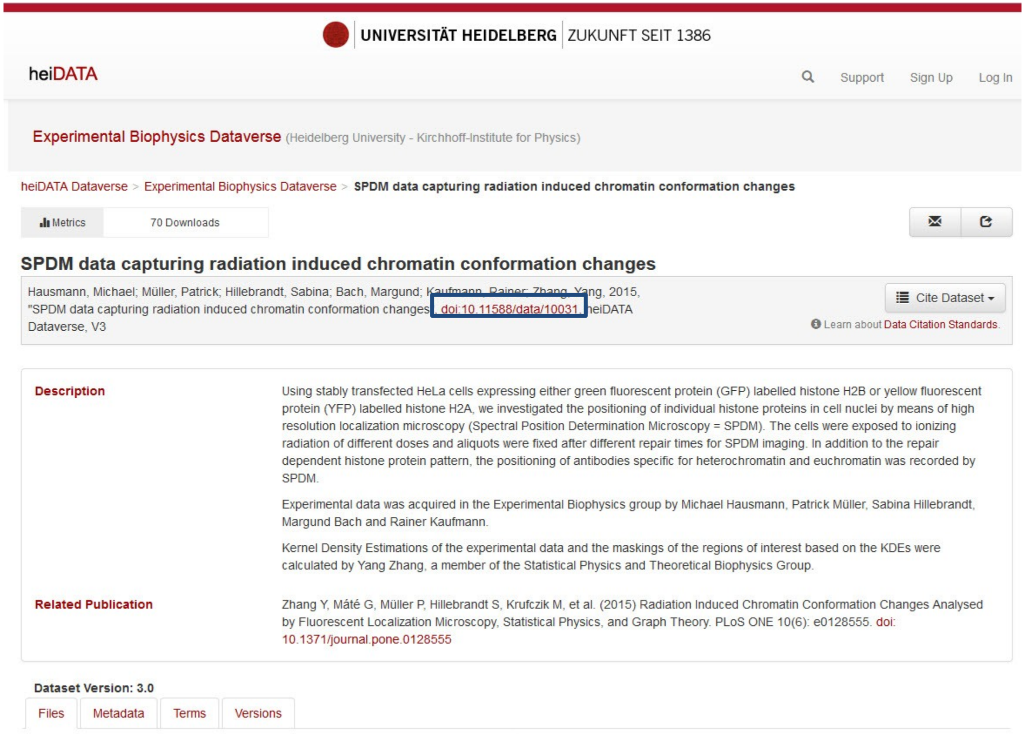
Feb 25, 2022 - Tropical Medicine

Brümmer, Lukas E.; Katzenschlager, Stephan; McGrath, Sean; Schmitz, Stephani; Gaeddert, Mary; Erdmann, Christian; Bota, Marc; Grilli, Maurizio; Larmann, Jan; Weigand, Markus A.; Pollock, Nira R.; Macé, Aurélien; Erkosar, Berra; Carmona, Sergio; Sacks, Jilian A.; Ongarello, Stefano; Denkinger, Claudia M., 2022, "Accuracy of rapid point-of-care antigen-based diagnostics for SARS-CoV-2: an updated systematic review and meta-analysis with meta regression analyzing influencing factors [Research Data]", https://doi.org/10.11588/data/T3MIB0, heiDATA, V1

Background Comprehensive information about the accuracy of antigen rapid diagnostic tests (Ag-RDTs) for SARS-CoV-2 is essential to guide public health decision makers in choosing the best tests and testing policies. In August 2021, we published a systematic review and meta-

heiDATA

Findability
- DOI’s



Experimental Biophysics Dataverse (Heidelberg University - Kirchhoff-Institute for Physics)

heiDATA Dataverse > Experimental Biophysics Dataverse > **SPDM data capturing radiation induced chromatin conformation changes**

SPDM data capturing radiation induced chromatin conformation changes

Hausmann, Michael; Müller, Patrick; Hillebrandt, Sabina; Bach, Margund; Kaufmann, Rainer; Zhang, Yang, 2015, "SPDM data capturing radiation induced chromatin conformation changes", doi:10.11588/data/10031, heiDATA Dataverse, V3

Learn about Data Citation Standards.

Description

Using stably transfected HeLa cells expressing either green fluorescent protein (GFP) labelled histone H2B or yellow fluorescent protein (YFP) labelled histone H2A, we investigated the positioning of individual histone proteins in cell nuclei by means of high resolution localization microscopy (Spectral Position Determination Microscopy = SPDM). The cells were exposed to ionizing radiation of different doses and aliquots were fixed after different repair times for SPDM imaging. In addition to the repair dependent histone protein pattern, the positioning of antibodies specific for heterochromatin and euchromatin was recorded by SPDM.

Experimental data was acquired in the Experimental Biophysics group by Michael Hausmann, Patrick Müller, Sabina Hillebrandt, Margund Bach and Rainer Kaufmann.

Kernel Density Estimations of the experimental data and the maskings of the regions of interest based on the KDEs were calculated by Yang Zhang, a member of the Statistical Physics and Theoretical Biophysics Group.

Related Publication

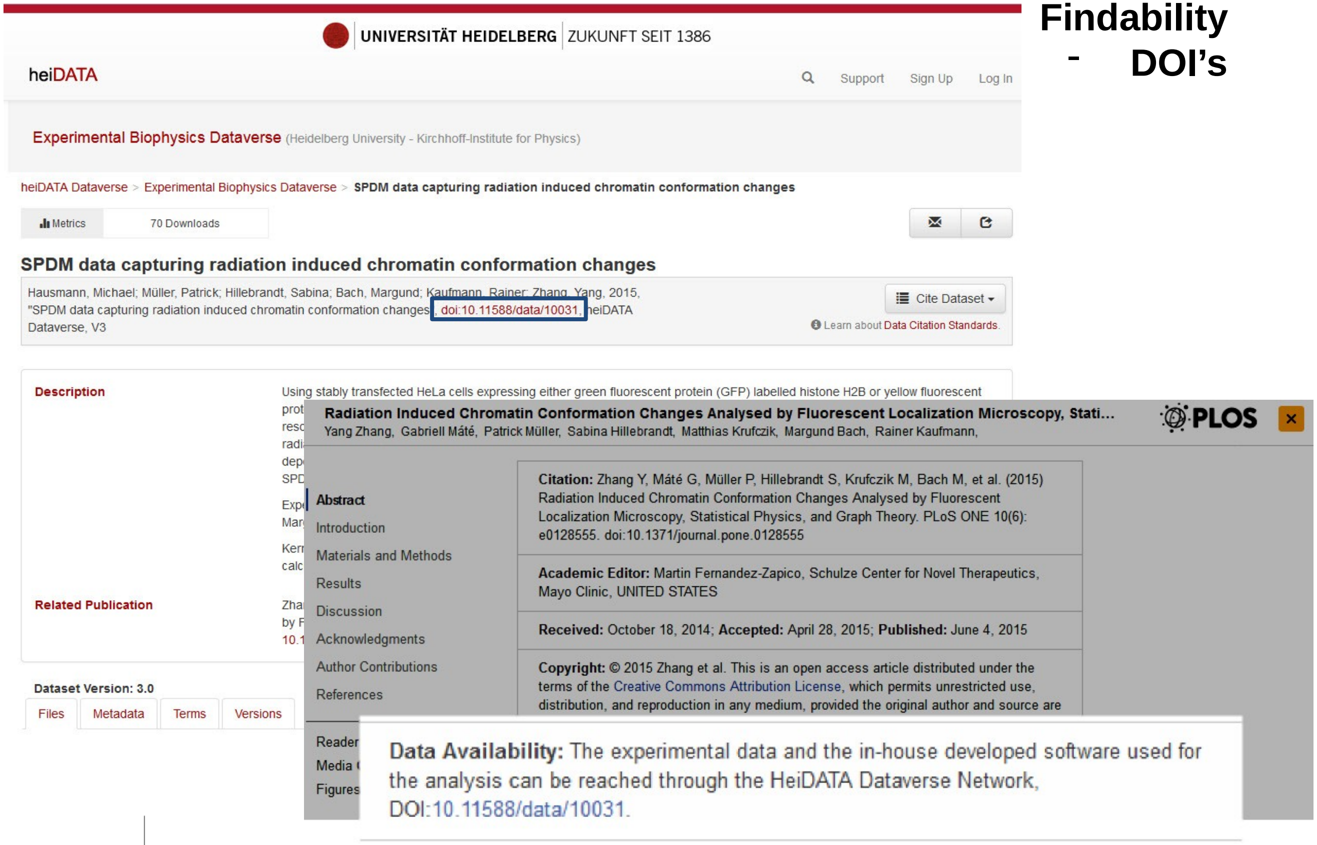
Zhang Y, Máté G, Müller P, Hillebrandt S, Krufczik M, et al. (2015) Radiation Induced Chromatin Conformation Changes Analysed by Fluorescent Localization Microscopy, Statistical Physics, and Graph Theory. PLoS ONE 10(6): e0128555. doi: 10.1371/journal.pone.0128555

Dataset Version: 3.0

heiDATA

UNIVERSITÄT HEIDELBERG
ZUKUNFT
SEIT 1386

Findability
- DOI’s



UNIVERSITÄT HEIDELBERG | ZUKUNFT SEIT 1386

Experimental Biophysics Dataverse (Heidelberg University - Kirchhoff-Institute for Physics)

heiDATA Dataverse > Experimental Biophysics Dataverse > SPDM data capturing radiation induced chromatin conformation changes

SPDM data capturing radiation induced chromatin conformation changes

Hausmann, Michael; Müller, Patrick; Hillebrandt, Sabina; Bach, Margund; Kaufmann, Rainer; Zhang, Yang, 2015, "SPDM data capturing radiation induced chromatin conformation changes", doi:10.11588/data/10031, heiDATA Dataverse, V3

Description: Using stably transfected HeLa cells expressing either green fluorescent protein (GFP) labelled histone H2B or yellow fluorescent prot… resc… radi… dep… SPD… Exp… Mar… Ker… calc…

Related Publication: Zha… by F… 10.1…

Dataset Version: 3.0

Radiation Induced Chromatin Conformation Changes Analysed by Fluorescent Localization Microscopy, Stati...

Yang Zhang, Gabriell Máté, Patrick Müller, Sabina Hillebrandt, Matthias Krufczik, Margund Bach, Rainer Kaufmann,

Citation: Zhang Y, Máté G, Müller P, Hillebrandt S, Krufczik M, Bach M, et al. (2015) Radiation Induced Chromatin Conformation Changes Analysed by Fluorescent Localization Microscopy, Statistical Physics, and Graph Theory. PLoS ONE 10(6): e0128555. doi:10.1371/journal.pone.0128555

Academic Editor: Martin Fernandez-Zapico, Schulze Center for Novel Therapeutics, Mayo Clinic, UNITED STATES

Received: October 18, 2014; **Accepted:** April 28, 2015; **Published:** June 4, 2015

Copyright: © 2015 Zhang et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are

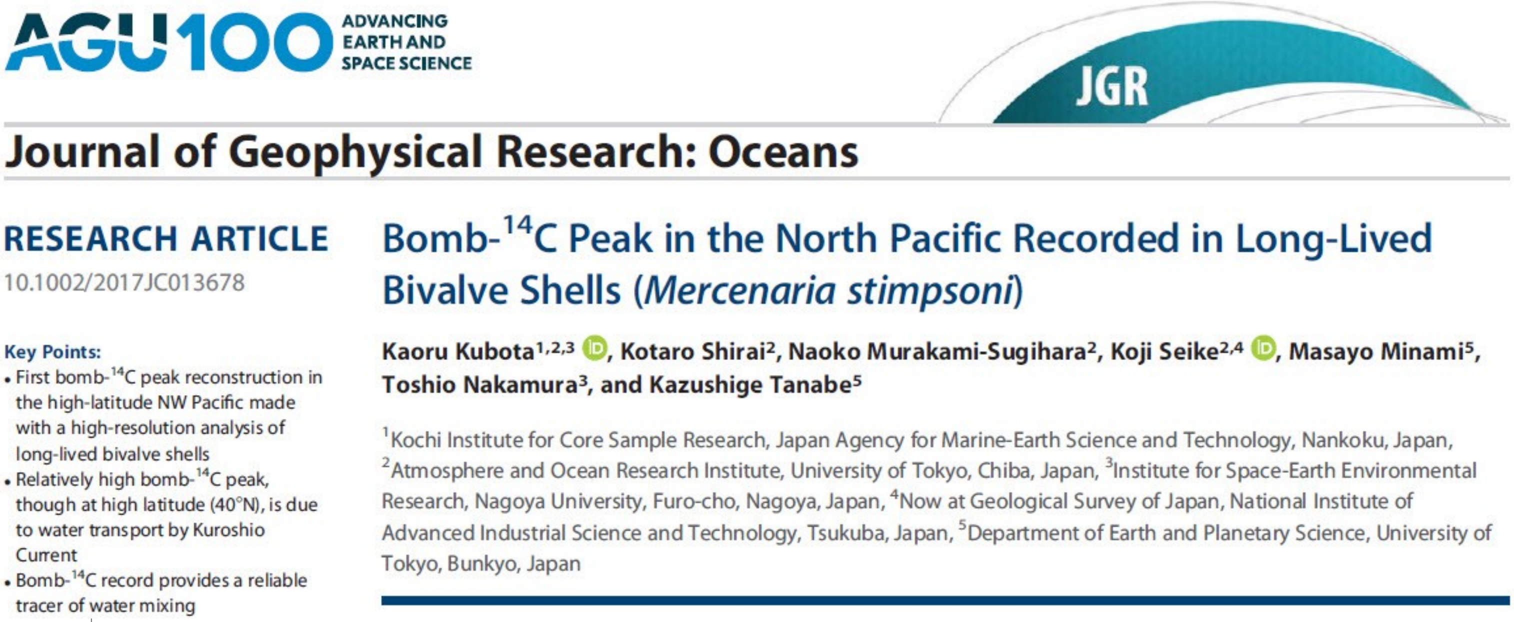
Data Availability: The experimental data and the in-house developed software used for the analysis can be reached through the HeiDATA Dataverse Network, DOI:10.11588/data/10031.

RESEARCH
DATA UNIT
HEIDELBERG

heiDATA

Findability

- **DOI’s**



AGU100 ADVANCING EARTH AND SPACE SCIENCE

JGR

Journal of Geophysical Research: Oceans

RESEARCH ARTICLE

10.1002/2017JC013678

Bomb-^{14}C Peak in the North Pacific Recorded in Long-Lived Bivalve Shells (*Mercenaria stimpsoni*)

Kaoru Kubota[1,2,3], Kotaro Shirai[2], Naoko Murakami-Sugihara[2], Koji Seike[2,4], Masayo Minami[5], Toshio Nakamura[3], and Kazushige Tanabe[5]

[1]Kochi Institute for Core Sample Research, Japan Agency for Marine-Earth Science and Technology, Nankoku, Japan, [2]Atmosphere and Ocean Research Institute, University of Tokyo, Chiba, Japan, [3]Institute for Space-Earth Environmental Research, Nagoya University, Furo-cho, Nagoya, Japan, [4]Now at Geological Survey of Japan, National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan, [5]Department of Earth and Planetary Science, University of Tokyo, Bunkyo, Japan

Key Points:

- First bomb-^{14}C peak reconstruction in the high-latitude NW Pacific made with a high-resolution analysis of long-lived bivalve shells
- Relatively high bomb-^{14}C peak, though at high latitude (40°N), is due to water transport by Kuroshio Current
- Bomb-^{14}C record provides a reliable tracer of water mixing

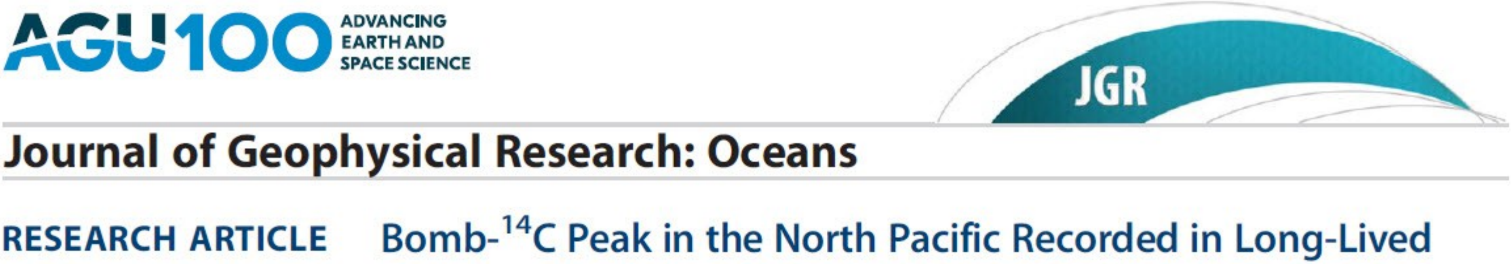
RESEARCH DATA UNIT HEIDELBERG

heiDATA

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Findability

- DOI's

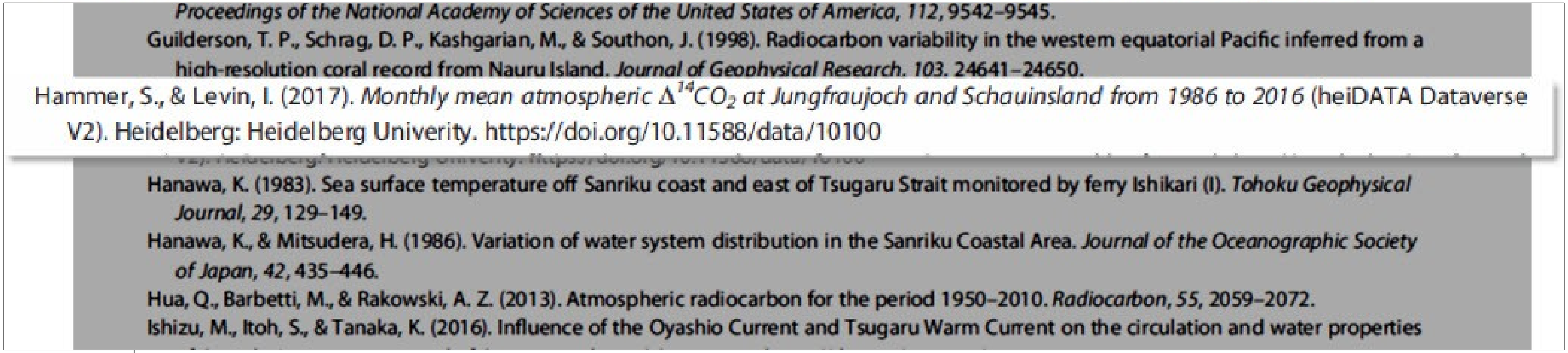


AGU100 ADVANCING EARTH AND SPACE SCIENCE

JGR

Journal of Geophysical Research: Oceans

RESEARCH ARTICLE Bomb-^{14}C Peak in the North Pacific Recorded in Long-Lived



Proceedings of the National Academy of Sciences of the United States of America, *112*, 9542–9545.

Guilderson, T. P., Schrag, D. P., Kashgarian, M., & Southon, J. (1998). Radiocarbon variability in the western equatorial Pacific inferred from a high-resolution coral record from Nauru Island. *Journal of Geophysical Research*, *103*, 24641–24650.

Hammer, S., & Levin, I. (2017). *Monthly mean atmospheric* $\Delta^{14}CO_2$ *at Jungfraujoch and Schauinsland from 1986 to 2016* (heiDATA Dataverse V2). Heidelberg: Heidelberg Univerity. https://doi.org/10.11588/data/10100

Hanawa, K. (1983). Sea surface temperature off Sanriku coast and east of Tsugaru Strait monitored by ferry Ishikari (I). *Tohoku Geophysical Journal*, *29*, 129–149.

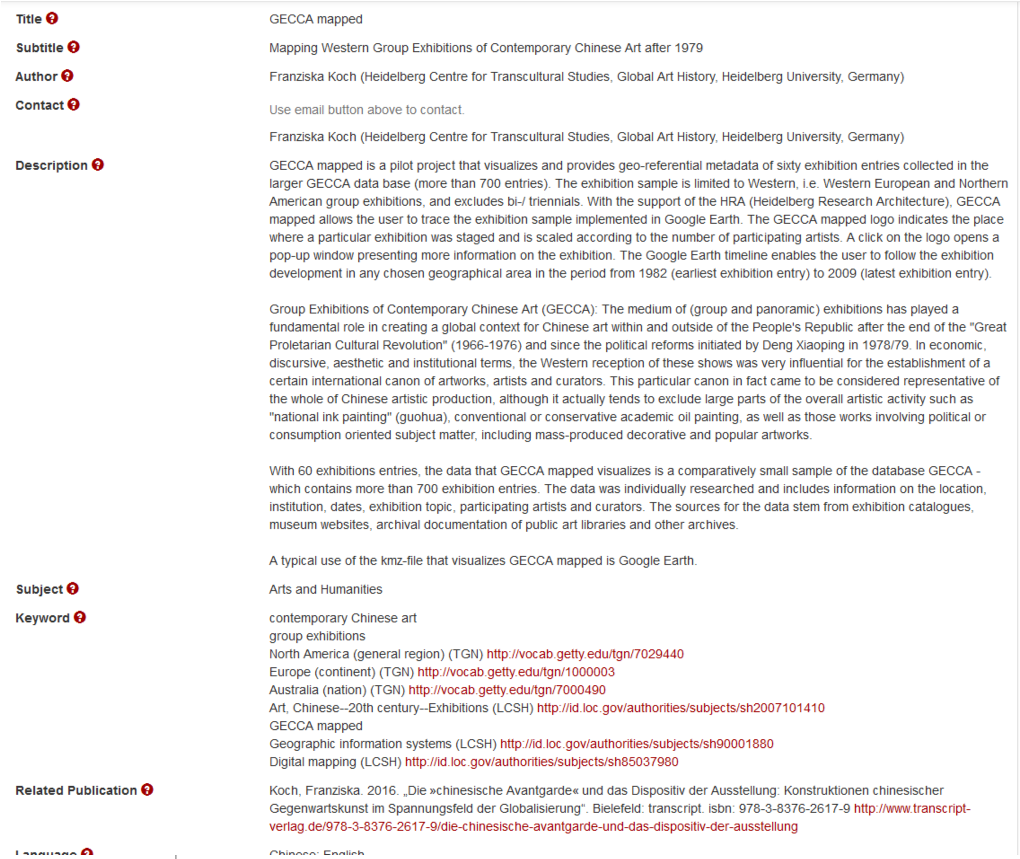
Hanawa, K., & Mitsudera, H. (1986). Variation of water system distribution in the Sanriku Coastal Area. *Journal of the Oceanographic Society of Japan*, *42*, 435–446.

Hua, Q., Barbetti, M., & Rakowski, A. Z. (2013). Atmospheric radiocarbon for the period 1950–2010. *Radiocarbon*, *55*, 2059–2072.

Ishizu, M., Itoh, S., & Tanaka, K. (2016). Influence of the Oyashio Current and Tsugaru Warm Current on the circulation and water properties

RESEARCH
DATA UNIT
HEIDELBERG

heiDATA



Title	GECCA mapped
Subtitle	Mapping Western Group Exhibitions of Contemporary Chinese Art after 1979
Author	Franziska Koch (Heidelberg Centre for Transcultural Studies, Global Art History, Heidelberg University, Germany)
Contact	Use email button above to contact. Franziska Koch (Heidelberg Centre for Transcultural Studies, Global Art History, Heidelberg University, Germany)
Description	GECCA mapped is a pilot project that visualizes and provides geo-referential metadata of sixty exhibition entries collected in the larger GECCA data base (more than 700 entries). The exhibition sample is limited to Western, i.e. Western European and Northern American group exhibitions, and excludes bi-/ triennials. With the support of the HRA (Heidelberg Research Architecture), GECCA mapped allows the user to trace the exhibition sample implemented in Google Earth. The GECCA mapped logo indicates the place where a particular exhibition was staged and is scaled according to the number of participating artists. A click on the logo opens a pop-up window presenting more information on the exhibition. The Google Earth timeline enables the user to follow the exhibition development in any chosen geographical area in the period from 1982 (earliest exhibition entry) to 2009 (latest exhibition entry). Group Exhibitions of Contemporary Chinese Art (GECCA): The medium of (group and panoramic) exhibitions has played a fundamental role in creating a global context for Chinese art within and outside of the People's Republic after the end of the "Great Proletarian Cultural Revolution" (1966-1976) and since the political reforms initiated by Deng Xiaoping in 1978/79. In economic, discursive, aesthetic and institutional terms, the Western reception of these shows was very influential for the establishment of a certain international canon of artworks, artists and curators. This particular canon in fact came to be considered representative of the whole of Chinese artistic production, although it actually tends to exclude large parts of the overall artistic activity such as "national ink painting" (guohua), conventional or conservative academic oil painting, as well as those works involving political or consumption oriented subject matter, including mass-produced decorative and popular artworks. With 60 exhibitions entries, the data that GECCA mapped visualizes is a comparatively small sample of the database GECCA - which contains more than 700 exhibition entries. The data was individually researched and includes information on the location, institution, dates, exhibition topic, participating artists and curators. The sources for the data stem from exhibition catalogues, museum websites, archival documentation of public art libraries and other archives. A typical use of the kmz-file that visualizes GECCA mapped is Google Earth.
Subject	Arts and Humanities
Keyword	contemporary Chinese art group exhibitions North America (general region) (TGN) http://vocab.getty.edu/tgn/7029440 Europe (continent) (TGN) http://vocab.getty.edu/tgn/1000003 Australia (nation) (TGN) http://vocab.getty.edu/tgn/7000490 Art, Chinese--20th century--Exhibitions (LCSH) http://id.loc.gov/authorities/subjects/sh2007101410 GECCA mapped Geographic information systems (LCSH) http://id.loc.gov/authorities/subjects/sh90001880 Digital mapping (LCSH) http://id.loc.gov/authorities/subjects/sh85037980
Related Publication	Koch, Franziska. 2016. „Die »chinesische Avantgarde« und das Dispositiv der Ausstellung: Konstruktionen chinesischer Gegenwartskunst im Spannungsfeld der Globalisierung". Bielefeld: transcript. isbn: 978-3-8376-2617-9 http://www.transcript-verlag.de/978-3-8376-2617-9/die-chinesische-avantgarde-und-das-dispositiv-der-ausstellung
Language	Chinese; English

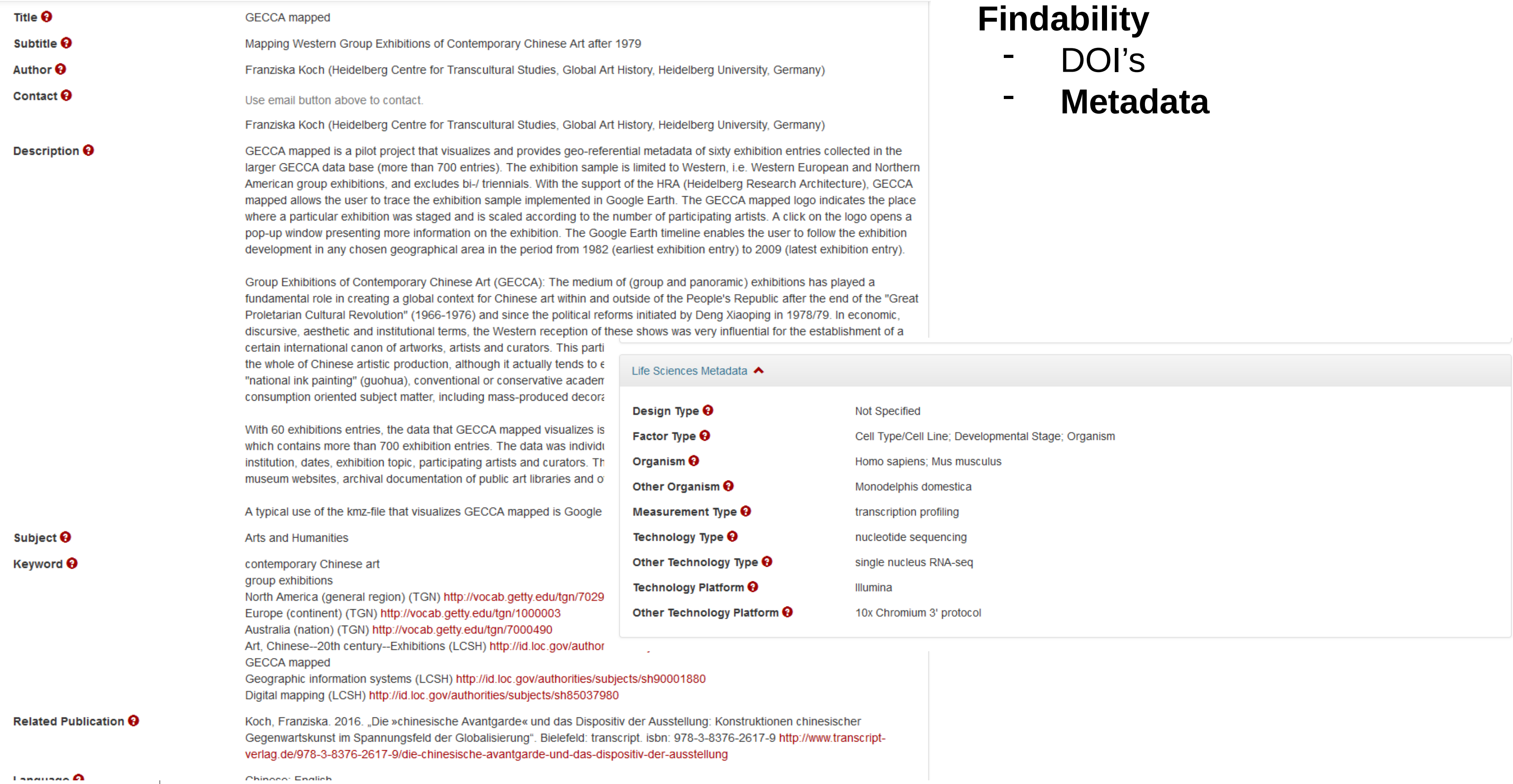
Findability

- DOI's
- **Metadata**

heiDATA

Findability

- DOI’s
- **Metadata**



Title ❓	GECCA mapped
Subtitle ❓	Mapping Western Group Exhibitions of Contemporary Chinese Art after 1979
Author ❓	Franziska Koch (Heidelberg Centre for Transcultural Studies, Global Art History, Heidelberg University, Germany)
Contact ❓	Use email button above to contact. Franziska Koch (Heidelberg Centre for Transcultural Studies, Global Art History, Heidelberg University, Germany)
Description ❓	GECCA mapped is a pilot project that visualizes and provides geo-referential metadata of sixty exhibition entries collected in the larger GECCA data base (more than 700 entries). The exhibition sample is limited to Western, i.e. Western European and Northern American group exhibitions, and excludes bi-/ triennials. With the support of the HRA (Heidelberg Research Architecture), GECCA mapped allows the user to trace the exhibition sample implemented in Google Earth. The GECCA mapped logo indicates the place where a particular exhibition was staged and is scaled according to the number of participating artists. A click on the logo opens a pop-up window presenting more information on the exhibition. The Google Earth timeline enables the user to follow the exhibition development in any chosen geographical area in the period from 1982 (earliest exhibition entry) to 2009 (latest exhibition entry). Group Exhibitions of Contemporary Chinese Art (GECCA): The medium of (group and panoramic) exhibitions has played a fundamental role in creating a global context for Chinese art within and outside of the People's Republic after the end of the "Great Proletarian Cultural Revolution" (1966-1976) and since the political reforms initiated by Deng Xiaoping in 1978/79. In economic, discursive, aesthetic and institutional terms, the Western reception of these shows was very influential for the establishment of a certain international canon of artworks, artists and curators. This parti the whole of Chinese artistic production, although it actually tends to e "national ink painting" (guohua), conventional or conservative academ consumption oriented subject matter, including mass-produced decora With 60 exhibitions entries, the data that GECCA mapped visualizes is which contains more than 700 exhibition entries. The data was individu institution, dates, exhibition topic, participating artists and curators. Th museum websites, archival documentation of public art libraries and o A typical use of the kmz-file that visualizes GECCA mapped is Google
Subject ❓	Arts and Humanities
Keyword ❓	contemporary Chinese art group exhibitions North America (general region) (TGN) http://vocab.getty.edu/tgn/7029 Europe (continent) (TGN) http://vocab.getty.edu/tgn/1000003 Australia (nation) (TGN) http://vocab.getty.edu/tgn/7000490 Art, Chinese--20th century--Exhibitions (LCSH) http://id.loc.gov/author GECCA mapped Geographic information systems (LCSH) http://id.loc.gov/authorities/subjects/sh90001880 Digital mapping (LCSH) http://id.loc.gov/authorities/subjects/sh85037980
Related Publication ❓	Koch, Franziska. 2016. „Die »chinesische Avantgarde« und das Dispositiv der Ausstellung: Konstruktionen chinesischer Gegenwartskunst im Spannungsfeld der Globalisierung“. Bielefeld: transcript. isbn: 978-3-8376-2617-9 http://www.transcript-verlag.de/978-3-8376-2617-9/die-chinesische-avantgarde-und-das-dispositiv-der-ausstellung
Language ❓	Chinese; English

Life Sciences Metadata ^

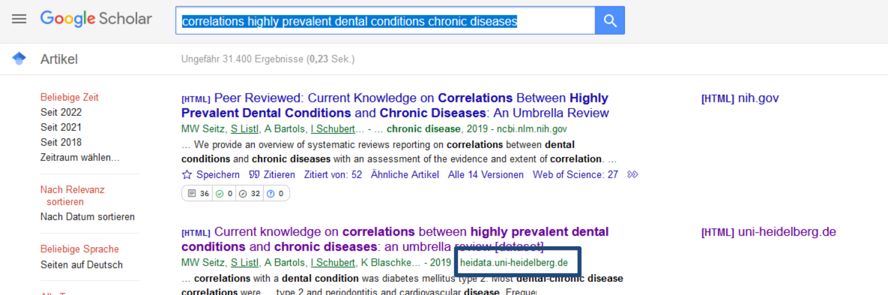
Design Type ❓	Not Specified
Factor Type ❓	Cell Type/Cell Line; Developmental Stage; Organism
Organism ❓	Homo sapiens; Mus musculus
Other Organism ❓	Monodelphis domestica
Measurement Type ❓	transcription profiling
Technology Type ❓	nucleotide sequencing
Other Technology Type ❓	single nucleus RNA-seq
Technology Platform ❓	Illumina
Other Technology Platform ❓	10x Chromium 3' protocol

heiDATA

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Findability

- DOI’s
- Metadata
- **Indexing in catalogs and databases (enabling automatic harvesting of metadata)**


Google Scholar

correlations highly prevalent dental conditions chronic diseases

Artikel

Ungefähr 31.400 Ergebnisse (**0,23** Sek.)

Beliebige Zeit
Seit 2022
Seit 2021
Seit 2018
Zeitraum wählen...

Nach Relevanz sortieren
Nach Datum sortieren

Beliebige Sprache
Seiten auf Deutsch

Alle Typen
Übersichtsarbeiten

☐ Patente

[HTML] Peer Reviewed: Current Knowledge on **Correlations** Between **Highly Prevalent Dental Conditions** and **Chronic Diseases**: An Umbrella Review

MW Seitz, S Listl, A Bartols, I Schubert... - ... **chronic disease**, 2019 - ncbi.nlm.nih.gov

... We provide an overview of systematic reviews reporting on **correlations** between **dental conditions** and **chronic diseases** with an assessment of the evidence and extent of **correlation**. ...

☆ Speichern 99 Zitieren Zitiert von: 52 Ähnliche Artikel Alle 14 Versionen Web of Science: 27 »

36 0 32 0

[HTML] nih.gov

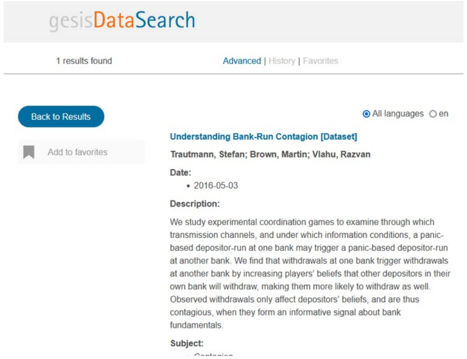
[HTML] Current knowledge on **correlations** between **highly prevalent dental conditions** and **chronic diseases**: an umbrella review [dataset]

MW Seitz, S Listl, A Bartols, I Schubert, K Blaschke... - 2019 heidata.uni-heidelberg.de

... **correlations** with a **dental condition** was diabetes mellitus type 2. Most **dental-chronic disease correlations** were ... type 2 and periodontitis and cardiovascular **disease**. Freque

☆ Speichern 99 Zitieren Zitiert von: 1 Ähnliche Artikel Alle 2 Versionen »

[HTML] uni-heidelberg.de


gesisDataSearch

1 results found

Advanced | History | Favorites

Back to Results

Add to favorites

● All languages ○ en

Understanding Bank-Run Contagion [Dataset]

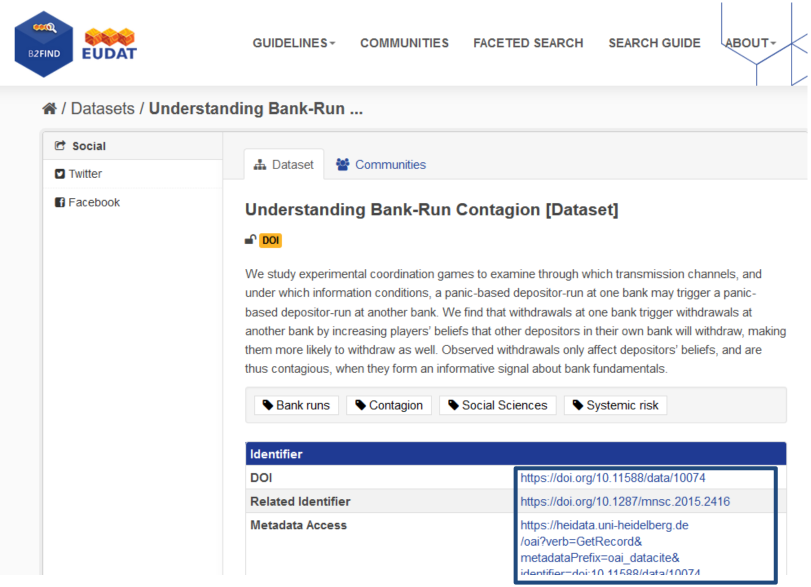
Trautmann, Stefan; Brown, Martin; Vlahu, Razvan

Date:
- 2016-05-03

Description:

We study experimental coordination games to examine through which transmission channels, and under which information conditions, a panic-based depositor-run at one bank may trigger a panic-based depositor-run at another bank. We find that withdrawals at one bank trigger withdrawals at another bank by increasing players’ beliefs that other depositors in their own bank will withdraw, making them more likely to withdraw as well. Observed withdrawals only affect depositors’ beliefs, and are thus contagious, when they form an informative signal about bank fundamentals.

Subject:


B2FIND EUDAT

GUIDELINES COMMUNITIES FACETED SEARCH SEARCH GUIDE ABOUT

/ Datasets / **Understanding Bank-Run ...**

Social
Twitter
Facebook

Dataset Communities

Understanding Bank-Run Contagion [Dataset]

DOI

We study experimental coordination games to examine through which transmission channels, and under which information conditions, a panic-based depositor-run at one bank may trigger a panic-based depositor-run at another bank. We find that withdrawals at one bank trigger withdrawals at another bank by increasing players’ beliefs that other depositors in their own bank will withdraw, making them more likely to withdraw as well. Observed withdrawals only affect depositors’ beliefs, and are thus contagious, when they form an informative signal about bank fundamentals.

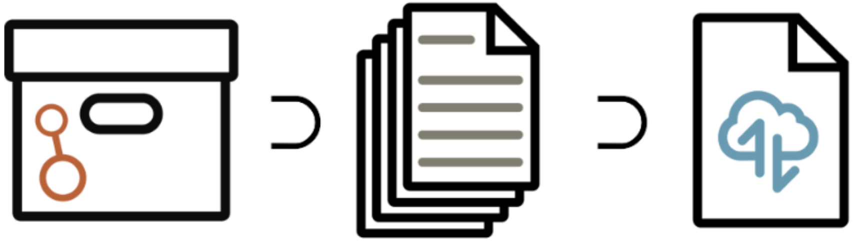
Bank runs · Contagion · Social Sciences · Systemic risk

Identifier	
DOI	https://doi.org/10.11588/data/10074
Related Identifier	https://doi.org/10.1287/mnsc.2015.2416
Metadata Access	https://heidata.uni-heidelberg.de/oai?verb=GetRecord&metadataPrefix=oai_datacite&identifier=doi:10.11588/data/10074

RESEARCH
DATA UNIT
HEIDELBERG

heiDATA

Organization of a Dataverse Repository



Dataverse

Collection of datasets
Own administration
Own branding (& can be embedded in your site)

dataset

Citation
Metadata
Versioning
Terms/permissions
Collection of Files

File

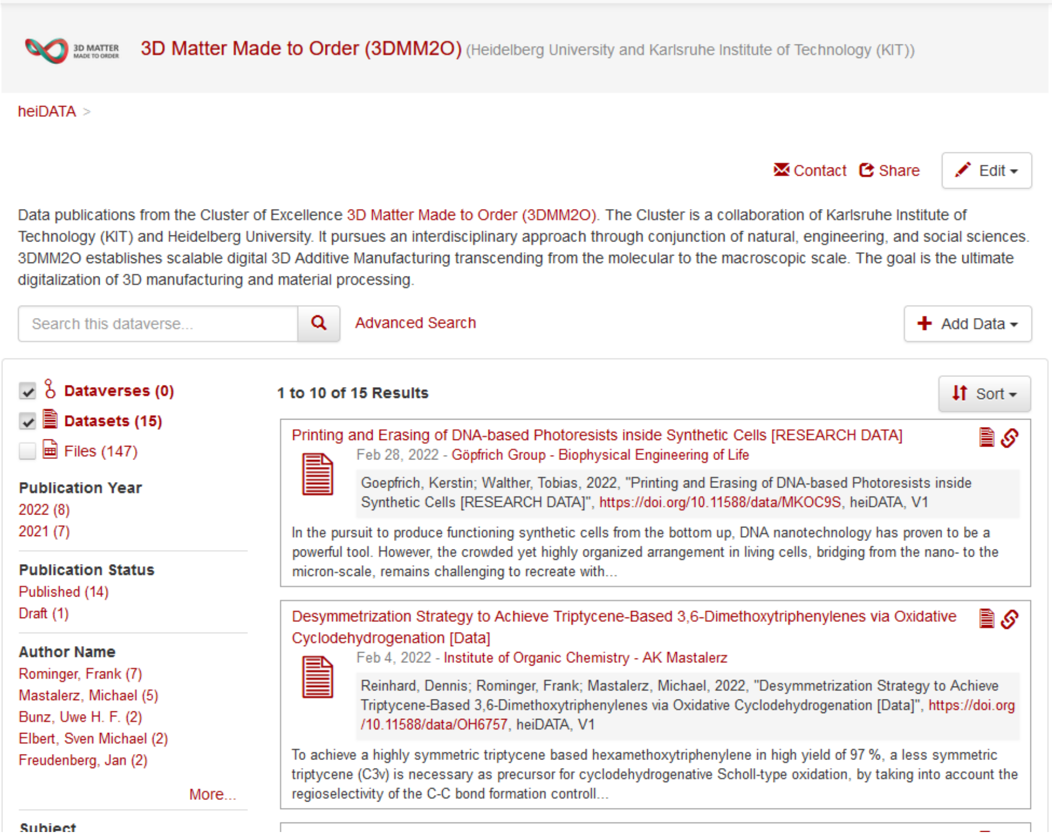
Citation
Preview/Explore
Metadata
Versioning
Permissions

Findability

- DOI’s
- Metadata
- Indexing in catalogs and databases (enabling automatic harvesting of metadata)
- **Dataverses: collection of datasets e.g. For research groups, projects,...**

heiDATA

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386



3D Matter Made to Order (3DMM2O) (Heidelberg University and Karlsruhe Institute of Technology (KIT))

heiDATA >

Data publications from the Cluster of Excellence 3D Matter Made to Order (3DMM2O). The Cluster is a collaboration of Karlsruhe Institute of Technology (KIT) and Heidelberg University. It pursues an interdisciplinary approach through conjunction of natural, engineering, and social sciences. 3DMM2O establishes scalable digital 3D Additive Manufacturing transcending from the molecular to the macroscopic scale. The goal is the ultimate digitalization of 3D manufacturing and material processing.

☑ Dataverses (0)
☑ Datasets (15)
□ Files (147)

Publication Year
2022 (8)
2021 (7)

Publication Status
Published (14)
Draft (1)

Author Name
Rominger, Frank (7)
Mastalerz, Michael (5)
Bunz, Uwe H. F. (2)
Elbert, Sven Michael (2)
Freudenberg, Jan (2)
More...

Subject

1 to 10 of 15 Results

Printing and Erasing of DNA-based Photoresists inside Synthetic Cells [RESEARCH DATA]
Feb 28, 2022 - Göpfrich Group - Biophysical Engineering of Life
Goepfrich, Kerstin; Walther, Tobias, 2022, "Printing and Erasing of DNA-based Photoresists inside Synthetic Cells [RESEARCH DATA]", https://doi.org/10.11588/data/MKOC9S, heiDATA, V1
In the pursuit to produce functioning synthetic cells from the bottom up, DNA nanotechnology has proven to be a powerful tool. However, the crowded yet highly organized arrangement in living cells, bridging from the nano- to the micron-scale, remains challenging to recreate with...

Desymmetrization Strategy to Achieve Triptycene-Based 3,6-Dimethoxytriphenylenes via Oxidative Cyclodehydrogenation [Data]
Feb 4, 2022 - Institute of Organic Chemistry - AK Mastalerz
Reinhard, Dennis; Rominger, Frank; Mastalerz, Michael, 2022, "Desymmetrization Strategy to Achieve Triptycene-Based 3,6-Dimethoxytriphenylenes via Oxidative Cyclodehydrogenation [Data]", https://doi.org/10.11588/data/OH6757, heiDATA, V1
To achieve a highly symmetric triptycene based hexamethoxytriphenylene in high yield of 97 %, a less symmetric triptycene (C3v) is necessary as precursor for cyclodehydrogenative Scholl-type oxidation, by taking into account the regioselectivity of the C-C bond formation controll...

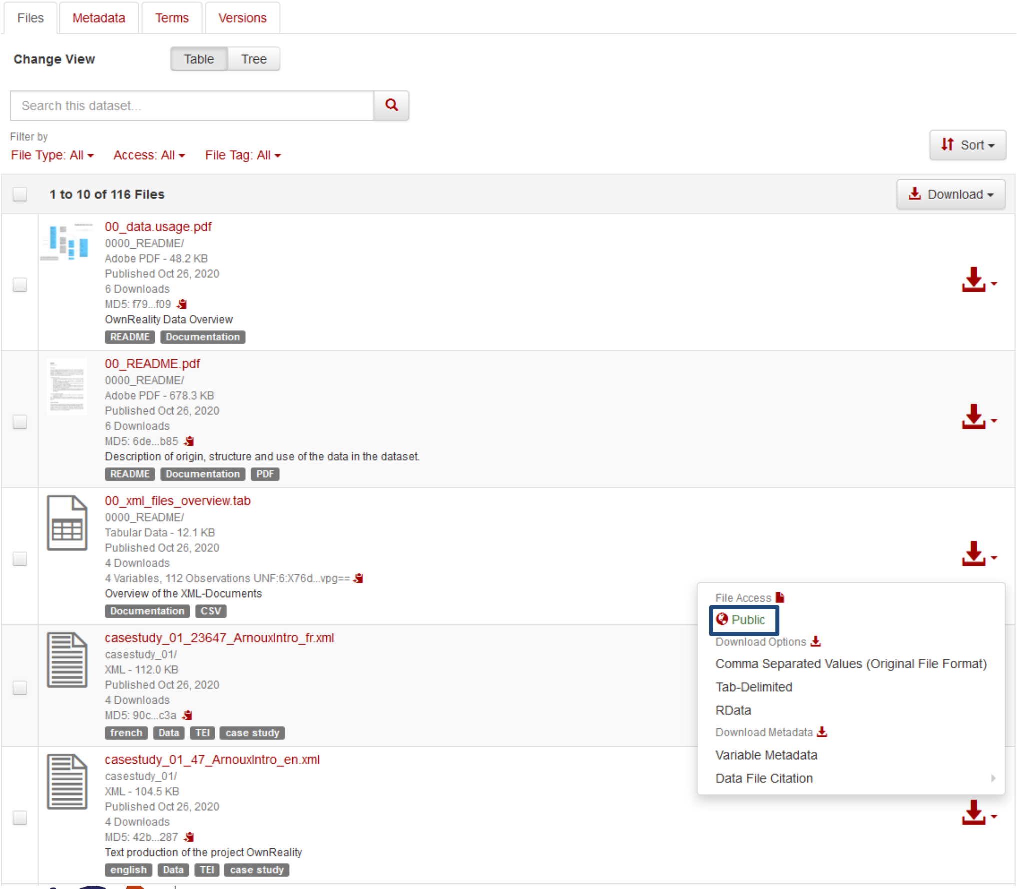
Findability

- DOI’s
- Metadata
- Indexing in catalogs and databases (enabling automatic harvesting of metadata
- **Dataverses: collection of datasets e.g. For research groups, projects,...**

RESEARCH
DATA UNIT
HEIDELBERG

heiDATA

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386



Files | Metadata | Terms | Versions

Change View: Table | Tree

Search this dataset...

Filter by
File Type: All ▾ Access: All ▾ File Tag: All ▾

Sort ▾

□ 1 to 10 of 116 Files — Download ▾

□ 00_data.usage.pdf
0000_README/
Adobe PDF - 48.2 KB
Published Oct 26, 2020
6 Downloads
MD5: f79...f09
OwnReality Data Overview
README Documentation

□ 00_README.pdf
0000_README/
Adobe PDF - 678.3 KB
Published Oct 26, 2020
6 Downloads
MD5: 6de...b85
Description of origin, structure and use of the data in the dataset.
README Documentation PDF

□ 00_xml_files_overview.tab
0000_README/
Tabular Data - 12.1 KB
Published Oct 26, 2020
4 Downloads
4 Variables, 112 Observations UNF:6:X76d...vpg==
Overview of the XML-Documents
Documentation CSV

File Access
Public
Download Options
Comma Separated Values (Original File Format)
Tab-Delimited
RData
Download Metadata
Variable Metadata
Data File Citation

□ casestudy_01_23647_ArnouxIntro_fr.xml
casestudy_01/
XML - 112.0 KB
Published Oct 26, 2020
4 Downloads
MD5: 90c...c3a
french Data TEI case study

□ casestudy_01_47_ArnouxIntro_en.xml
casestudy_01/
XML - 104.5 KB
Published Oct 26, 2020
4 Downloads
MD5: 42b...287
Text production of the project OwnReality
english Data TEI case study

Accessibility

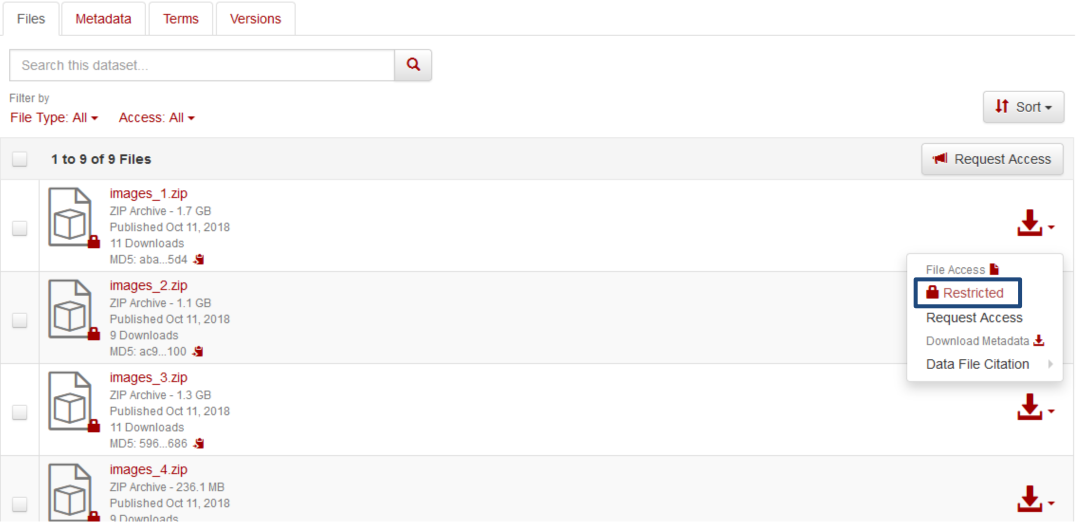
- Download of public files via browser or via API

RESEARCH
DATA UNIT
HEIDELBERG

heiDATA

Accessibility

- Download of public files via browser or via API
- **“As open as possible, but as closed as necessary”**



Files | Metadata | Terms | Versions

Search this dataset...

Filter by
File Type: All ▾ Access: All ▾

Sort ▾

□ 1 to 9 of 9 Files

Request Access

□ images_1.zip
ZIP Archive - 1.7 GB
Published Oct 11, 2018
11 Downloads
MD5: aba...5d4

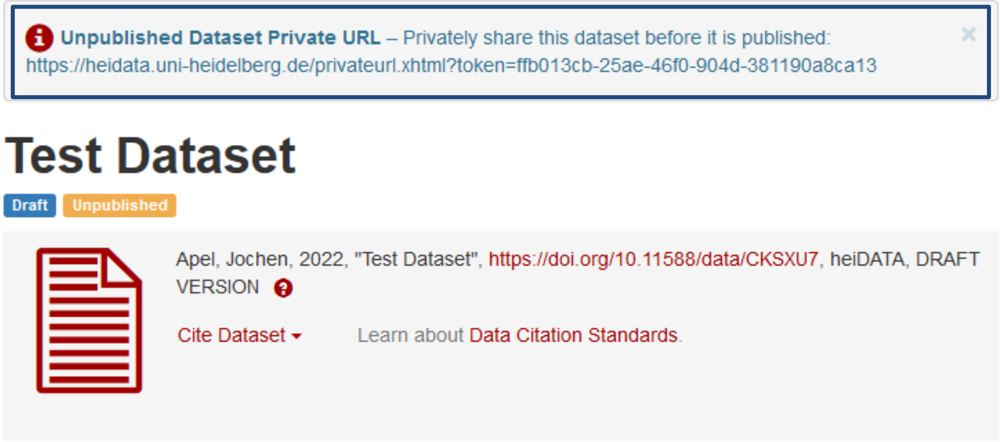
File Access
Restricted
Request Access
Download Metadata
Data File Citation

□ images_2.zip
ZIP Archive - 1.1 GB
Published Oct 11, 2018
9 Downloads
MD5: ac9...100

□ images_3.zip
ZIP Archive - 1.3 GB
Published Oct 11, 2018
11 Downloads
MD5: 596...686

□ images_4.zip
ZIP Archive - 236.1 MB
Published Oct 11, 2018
9 Downloads

heiDATA



Unpublished Dataset Private URL – Privately share this dataset before it is published: https://heidata.uni-heidelberg.de/privateurl.xhtml?token=ffb013cb-25ae-46f0-904d-381190a8ca13

Test Dataset

Draft Unpublished

Apel, Jochen, 2022, "Test Dataset", https://doi.org/10.11588/data/CKSXU7, heiDATA, DRAFT VERSION

Accessibility

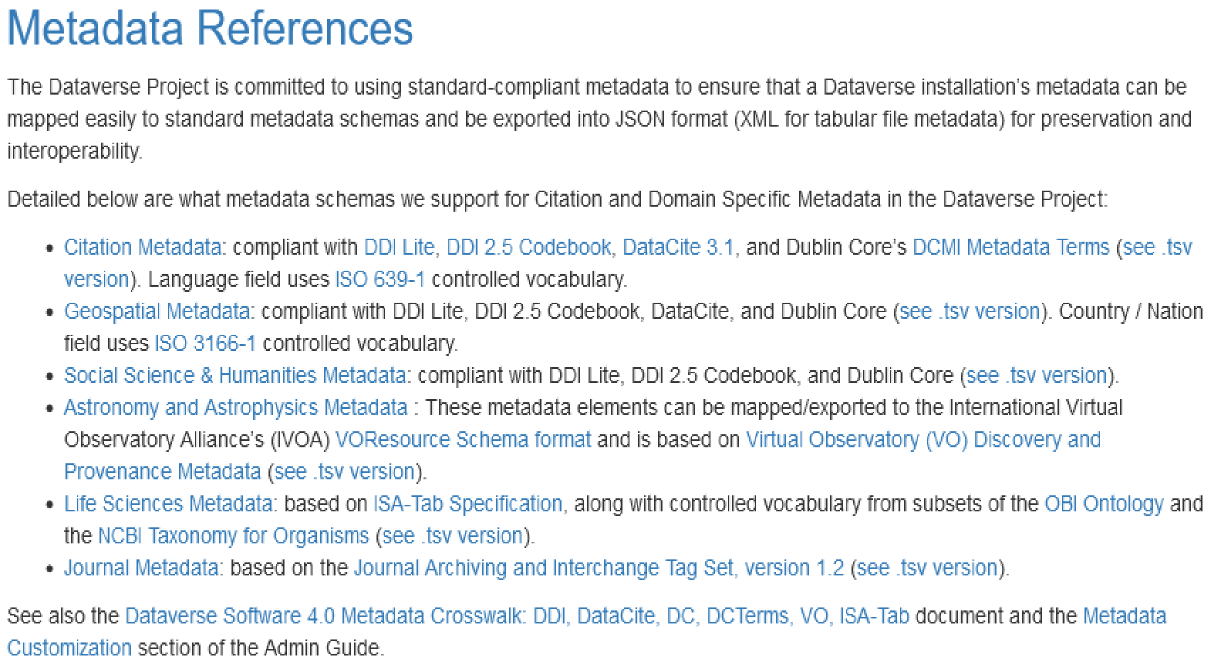
- Download of public files via browser or via API
- “As open as possible, but as closed as necessary”
- **Private URLs for pre-publication access (e.g. for reviewers)**

heiDATA

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Interoperability

- Metadata standards



Metadata References

The Dataverse Project is committed to using standard-compliant metadata to ensure that a Dataverse installation's metadata can be mapped easily to standard metadata schemas and be exported into JSON format (XML for tabular file metadata) for preservation and interoperability.

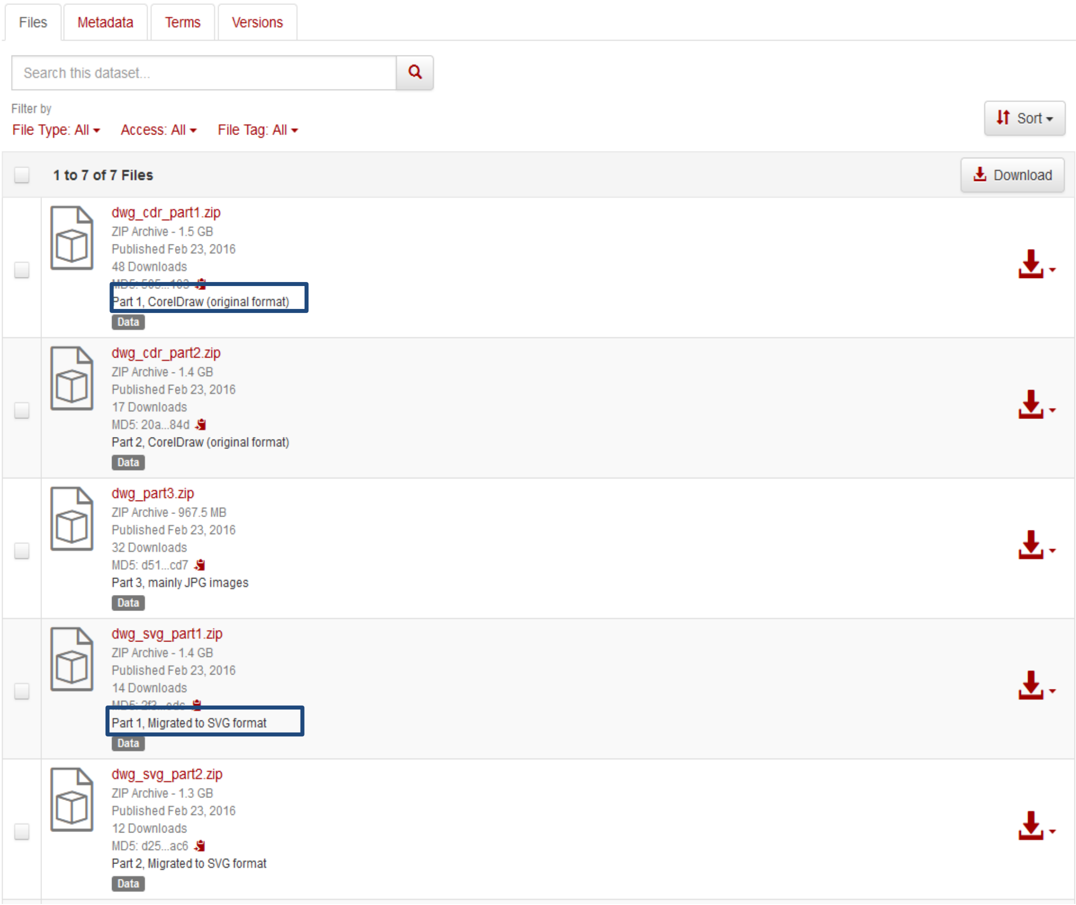
Detailed below are what metadata schemas we support for Citation and Domain Specific Metadata in the Dataverse Project:

- Citation Metadata: compliant with DDI Lite, DDI 2.5 Codebook, DataCite 3.1, and Dublin Core's DCMI Metadata Terms (see .tsv version). Language field uses ISO 639-1 controlled vocabulary.
- Geospatial Metadata: compliant with DDI Lite, DDI 2.5 Codebook, DataCite, and Dublin Core (see .tsv version). Country / Nation field uses ISO 3166-1 controlled vocabulary.
- Social Science & Humanities Metadata: compliant with DDI Lite, DDI 2.5 Codebook, and Dublin Core (see .tsv version).
- Astronomy and Astrophysics Metadata : These metadata elements can be mapped/exported to the International Virtual Observatory Alliance's (IVOA) VOResource Schema format and is based on Virtual Observatory (VO) Discovery and Provenance Metadata (see .tsv version).
- Life Sciences Metadata: based on ISA-Tab Specification, along with controlled vocabulary from subsets of the OBI Ontology and the NCBI Taxonomy for Organisms (see .tsv version).
- Journal Metadata: based on the Journal Archiving and Interchange Tag Set, version 1.2 (see .tsv version).

See also the Dataverse Software 4.0 Metadata Crosswalk: DDI, DataCite, DC, DCTerms, VO, ISA-Tab document and the Metadata Customization section of the Admin Guide.

RESEARCH
DATA UNIT
HEIDELBERG

heiDATA



□ 1 to 7 of 7 Files

□ dwg_cdr_part1.zip
ZIP Archive - 1.5 GB
Published Feb 23, 2016
48 Downloads
MD5: 505...103
Part 1, CorelDraw (original format)
Data

□ dwg_cdr_part2.zip
ZIP Archive - 1.4 GB
Published Feb 23, 2016
17 Downloads
MD5: 20a...84d
Part 2, CorelDraw (original format)
Data

□ dwg_part3.zip
ZIP Archive - 967.5 MB
Published Feb 23, 2016
32 Downloads
MD5: d51...cd7
Part 3, mainly JPG images
Data

□ dwg_svg_part1.zip
ZIP Archive - 1.4 GB
Published Feb 23, 2016
14 Downloads
MD5: 2f3...ede
Part 1, Migrated to SVG format
Data

□ dwg_svg_part2.zip
ZIP Archive - 1.3 GB
Published Feb 23, 2016
12 Downloads
MD5: d25...ac6
Part 2, Migrated to SVG format
Data

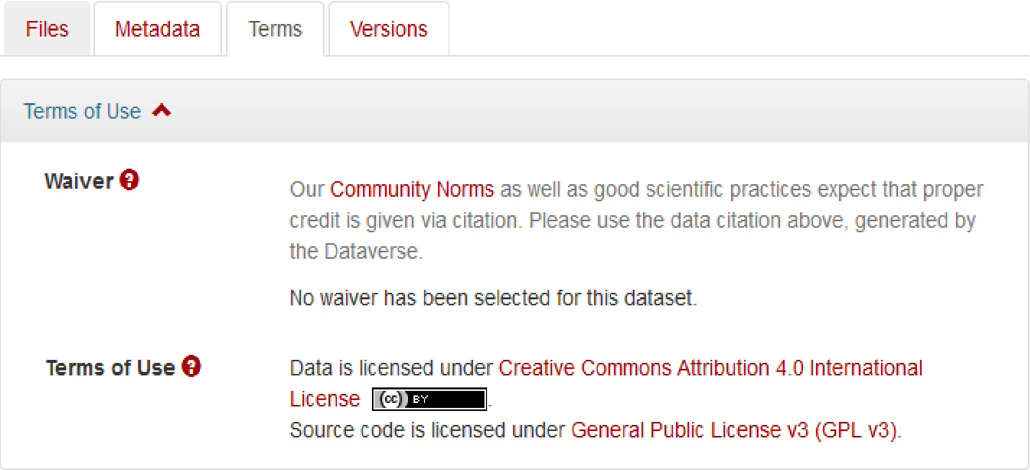
Interoperability

- Metadata standards
- **Advice on suitable file formats, support with format conversion**
- **Technical validity checks**

heiDATA

Reusability

- **Open content licenses**



Waiver

Our Community Norms as well as good scientific practices expect that proper credit is given via citation. Please use the data citation above, generated by the Dataverse.

No waiver has been selected for this dataset.

Terms of Use

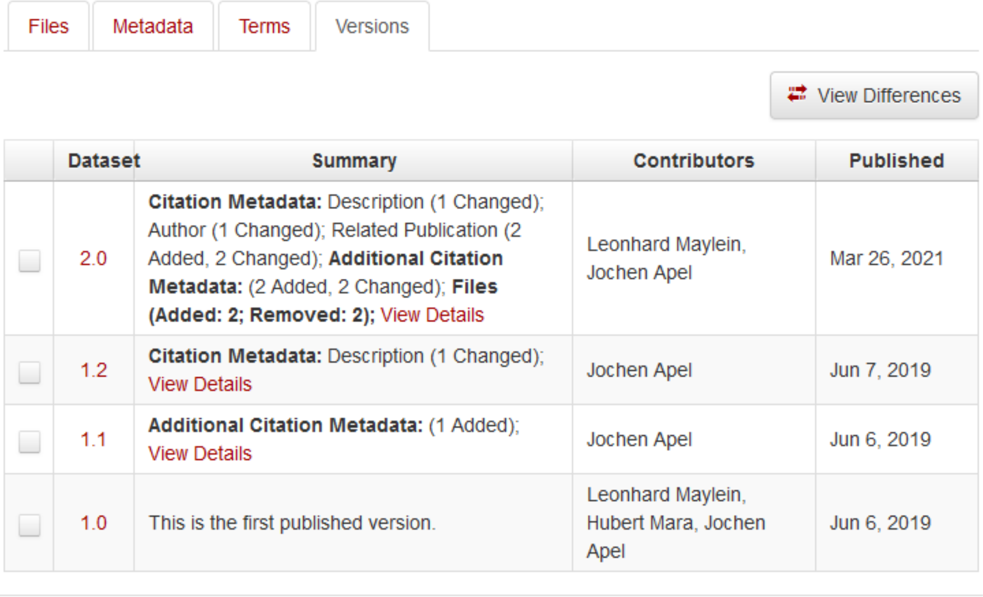
Data is licensed under Creative Commons Attribution 4.0 International License (cc) BY.
Source code is licensed under General Public License v3 (GPL v3).

heiDATA

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Reusability

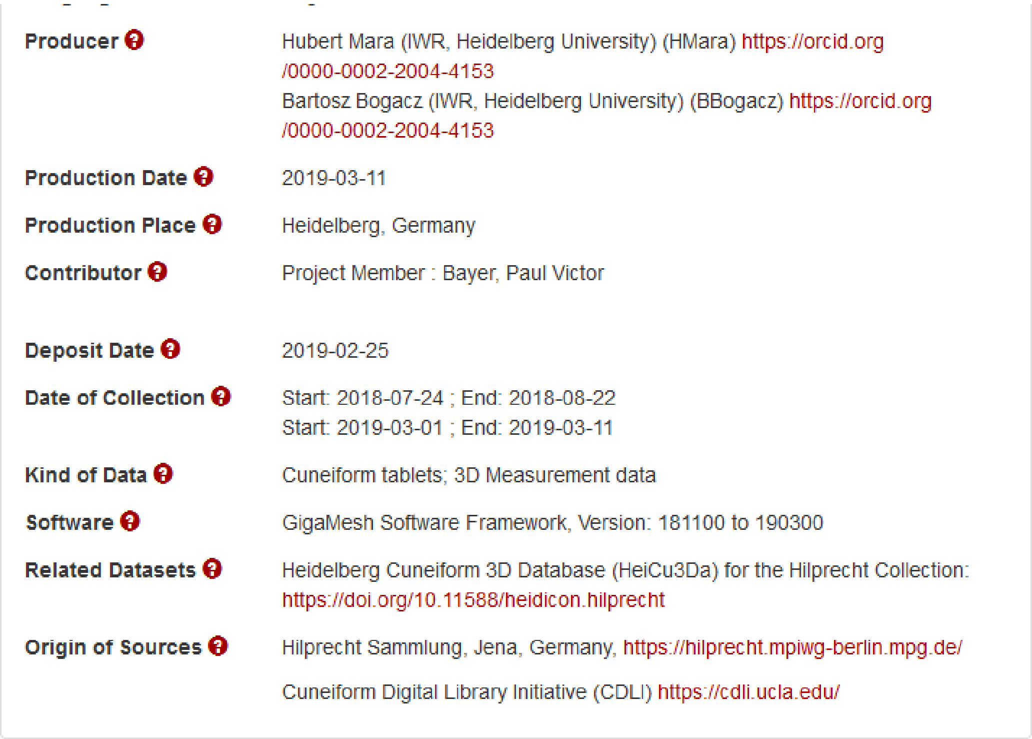
- Open content licenses
- **Transparent versioning**



	Dataset	Summary	Contributors	Published
☐	2.0	**Citation Metadata:** Description (1 Changed); Author (1 Changed); Related Publication (2 Added, 2 Changed); **Additional Citation Metadata:** (2 Added, 2 Changed); **Files (Added: 2; Removed: 2);** View Details	Leonhard Maylein, Jochen Apel	Mar 26, 2021
☐	1.2	**Citation Metadata:** Description (1 Changed); View Details	Jochen Apel	Jun 7, 2019
☐	1.1	**Additional Citation Metadata:** (1 Added); View Details	Jochen Apel	Jun 6, 2019
☐	1.0	This is the first published version.	Leonhard Maylein, Hubert Mara, Jochen Apel	Jun 6, 2019

RESEARCH
DATA UNIT
HEIDELBERG

heiDATA

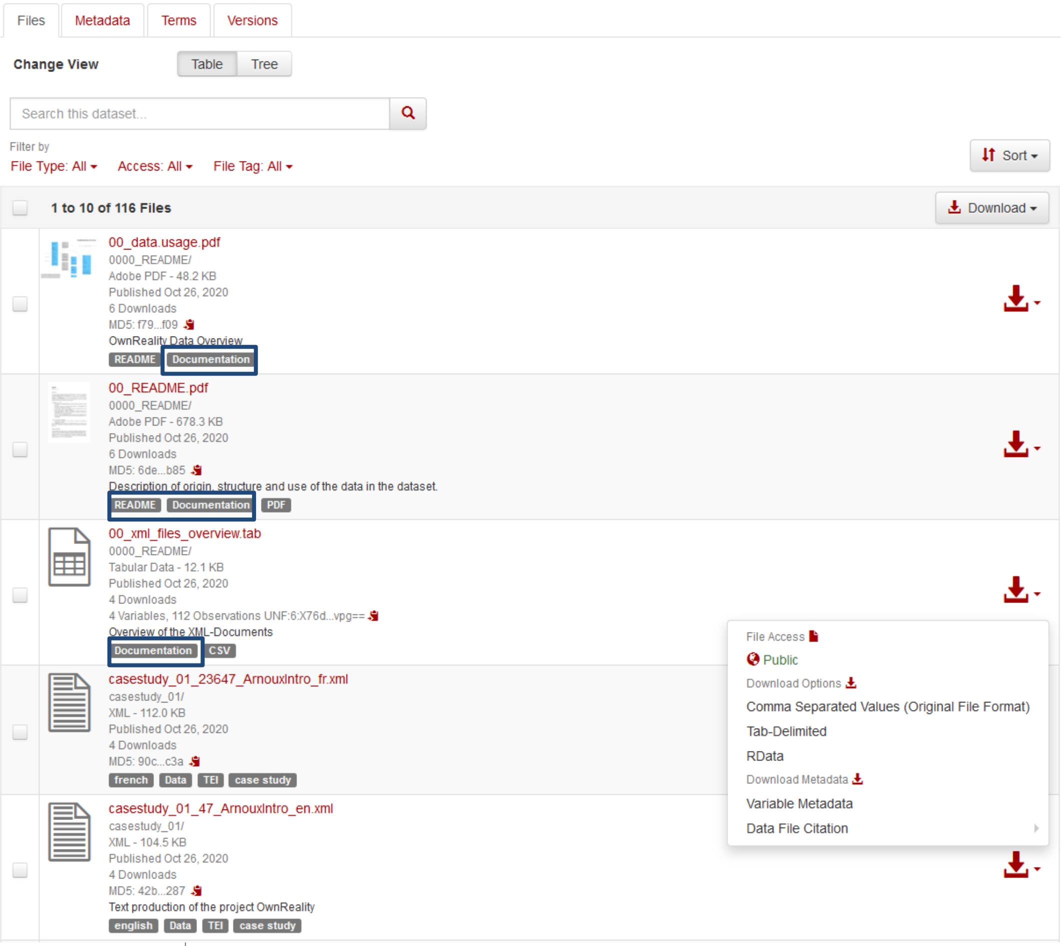


Producer ❓	Hubert Mara (IWR, Heidelberg University) (HMara) https://orcid.org/0000-0002-2004-4153 Bartosz Bogacz (IWR, Heidelberg University) (BBogacz) https://orcid.org/0000-0002-2004-4153
Production Date ❓	2019-03-11
Production Place ❓	Heidelberg, Germany
Contributor ❓	Project Member : Bayer, Paul Victor
Deposit Date ❓	2019-02-25
Date of Collection ❓	Start: 2018-07-24 ; End: 2018-08-22 Start: 2019-03-01 ; End: 2019-03-11
Kind of Data ❓	Cuneiform tablets; 3D Measurement data
Software ❓	GigaMesh Software Framework, Version: 181100 to 190300
Related Datasets ❓	Heidelberg Cuneiform 3D Database (HeiCu3Da) for the Hilprecht Collection: https://doi.org/10.11588/heidicon.hilprecht
Origin of Sources ❓	Hilprecht Sammlung, Jena, Germany, https://hilprecht.mpiwg-berlin.mpg.de/ Cuneiform Digital Library Initiative (CDLI) https://cdli.ucla.edu/

Reusability

- Open content licenses
- Transparent versioning
- **Provenance information**

heiDATA



1 to 10 of 116 Files

00_data.usage.pdf
0000_README/
Adobe PDF - 48.2 KB
Published Oct 26, 2020
6 Downloads
MD5: f79...f09
OwnReality Data Overview
README Documentation

00_README.pdf
0000_README/
Adobe PDF - 678.3 KB
Published Oct 26, 2020
6 Downloads
MD5: 6de...b85
Description of origin, structure and use of the data in the dataset.
README Documentation PDF

00_xml_files_overview.tab
0000_README/
Tabular Data - 12.1 KB
Published Oct 26, 2020
4 Downloads
4 Variables, 112 Observations UNF:6:X76d...vpg==
Overview of the XML-Documents
Documentation CSV

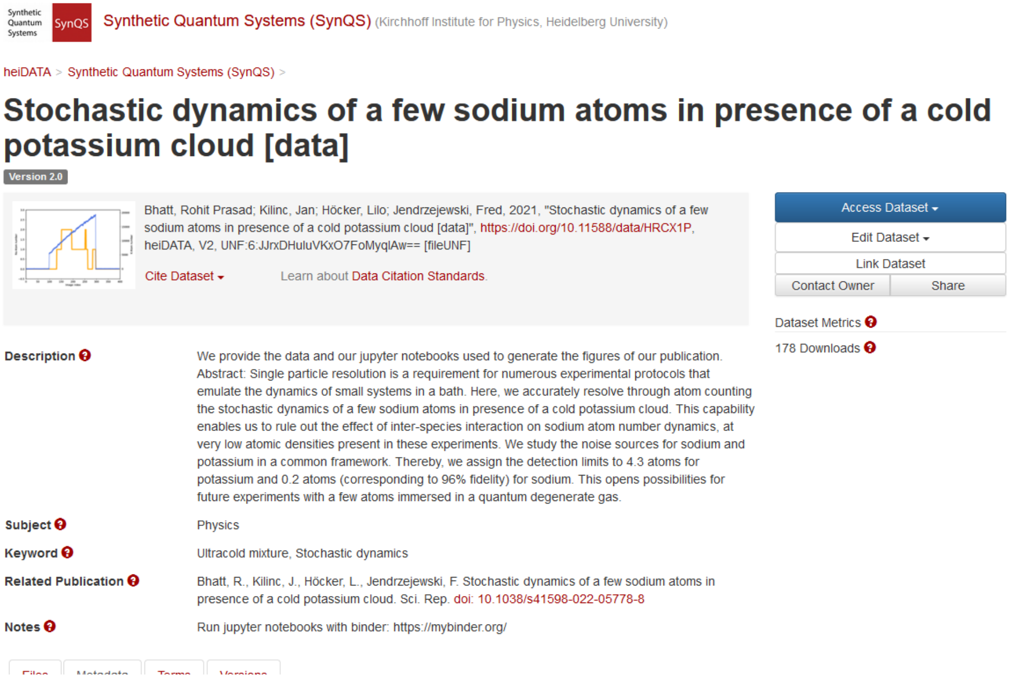
casestudy_01_23647_ArnouxIntro_fr.xml
casestudy_01/
XML - 112.0 KB
Published Oct 26, 2020
4 Downloads
MD5: 90c...c3a
french Data TEI case study

casestudy_01_47_ArnouxIntro_en.xml
casestudy_01/
XML - 104.5 KB
Published Oct 26, 2020
4 Downloads
MD5: 42b...287
Text production of the project OwnReality
english Data TEI case study

Reusability

- Open content licenses
- transparent versioning
- Provenance information
- **Documentation files**

heiDATA



Synthetic Quantum Systems (SynQS) (Kirchhoff Institute for Physics, Heidelberg University)

heiDATA > Synthetic Quantum Systems (SynQS) >

Stochastic dynamics of a few sodium atoms in presence of a cold potassium cloud [data]

Version 2.0

Bhatt, Rohit Prasad; Kilinc, Jan; Höcker, Lilo; Jendrzejewski, Fred, 2021, "Stochastic dynamics of a few sodium atoms in presence of a cold potassium cloud [data]", https://doi.org/10.11588/data/HRCX1P, heiDATA, V2, UNF:6:JJrxDHuluVKxO7FoMyqlAw== [fileUNF]

Description We provide the data and our jupyter notebooks used to generate the figures of our publication. Abstract: Single particle resolution is a requirement for numerous experimental protocols that emulate the dynamics of small systems in a bath. Here, we accurately resolve through atom counting the stochastic dynamics of a few sodium atoms in presence of a cold potassium cloud. This capability enables us to rule out the effect of inter-species interaction on sodium atom number dynamics, at very low atomic densities present in these experiments. We study the noise sources for sodium and potassium in a common framework. Thereby, we assign the detection limits to 4.3 atoms for potassium and 0.2 atoms (corresponding to 96% fidelity) for sodium. This opens possibilities for future experiments with a few atoms immersed in a quantum degenerate gas.

Subject Physics

Keyword Ultracold mixture, Stochastic dynamics

Related Publication Bhatt, R., Kilinc, J., Höcker, L., Jendrzejewski, F. Stochastic dynamics of a few sodium atoms in presence of a cold potassium cloud. Sci. Rep. doi: 10.1038/s41598-022-05778-8

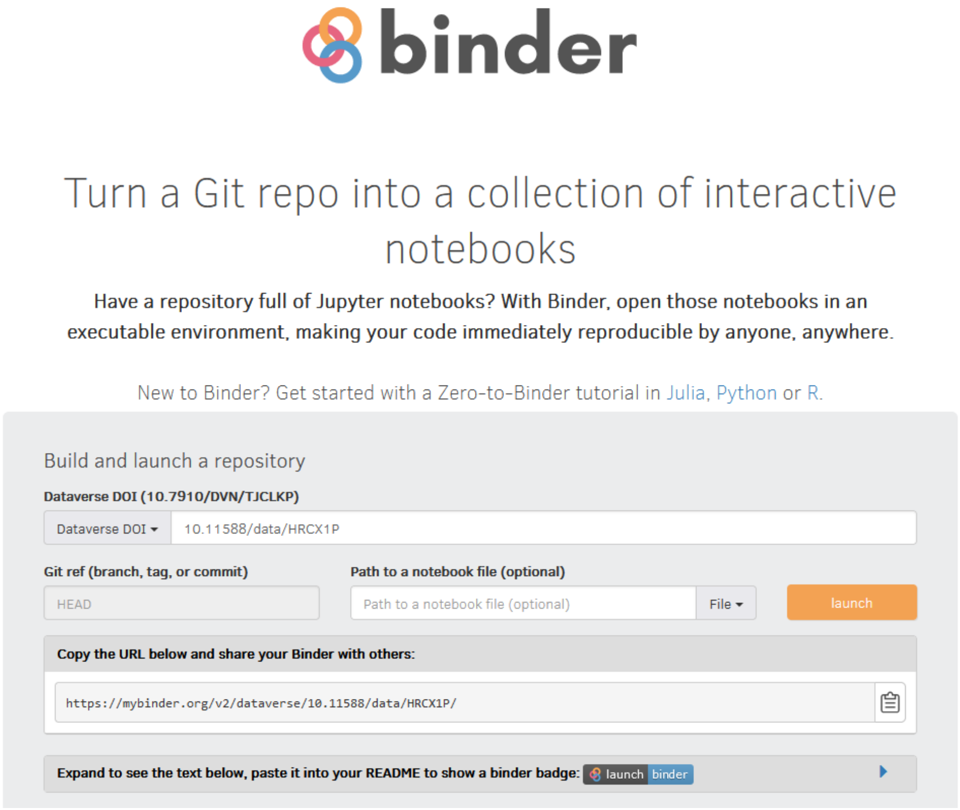
Notes Run jupyter notebooks with binder: https://mybinder.org/

Reusability

- Open content licenses
- transparent versioning
- Provenance information
- Documentation files
- **Integration with external services, e.g. binder (https://mybinder.org/)**

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

heiDATA



binder

Turn a Git repo into a collection of interactive notebooks

Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.

New to Binder? Get started with a Zero-to-Binder tutorial in Julia, Python or R.

Build and launch a repository

Dataverse DOI (10.7910/DVN/TJCLKP)

Dataverse DOI ▾ | 10.11588/data/HRCX1P

Git ref (branch, tag, or commit): HEAD

Path to a notebook file (optional): Path to a notebook file (optional) | File ▾

launch

Copy the URL below and share your Binder with others:

```
https://mybinder.org/v2/dataverse/10.11588/data/HRCX1P/
```

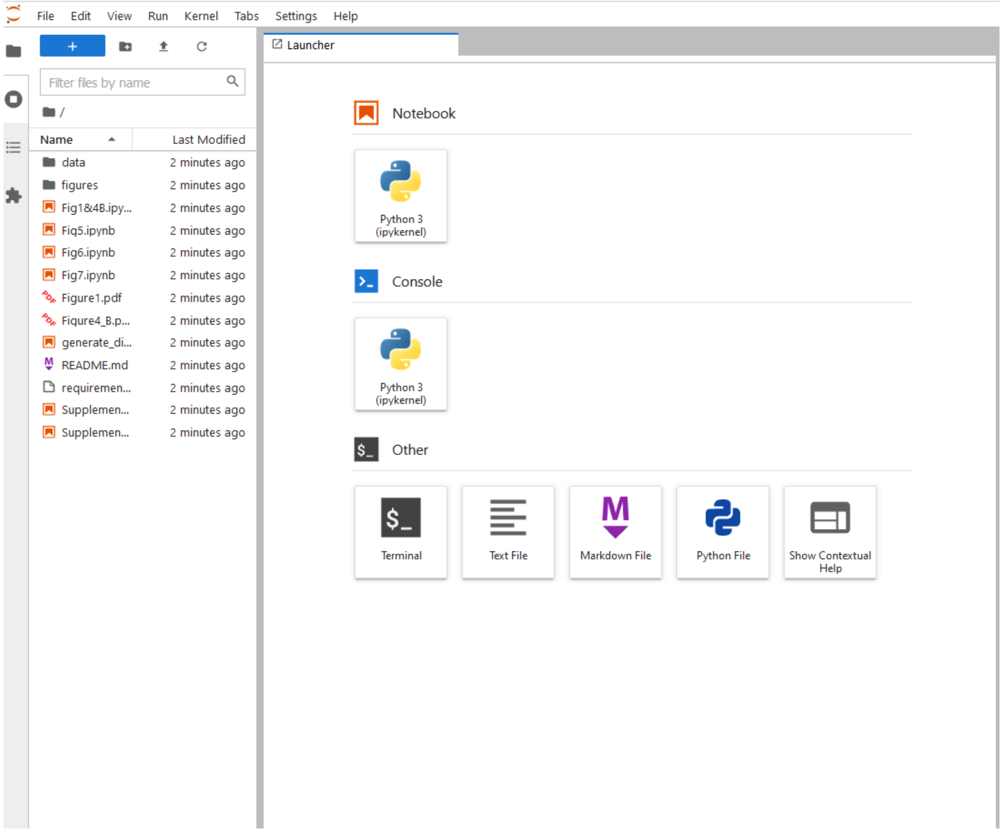
Expand to see the text below, paste it into your README to show a binder badge: launch binder ▸

Reusability

- Open content licenses
- transparent versioning
- Provenance information
- Documentation files
- **Integration with external services, e.g. binder (https://mybinder.org/)**

RESEARCH
DATA UNIT
HEIDELBERG

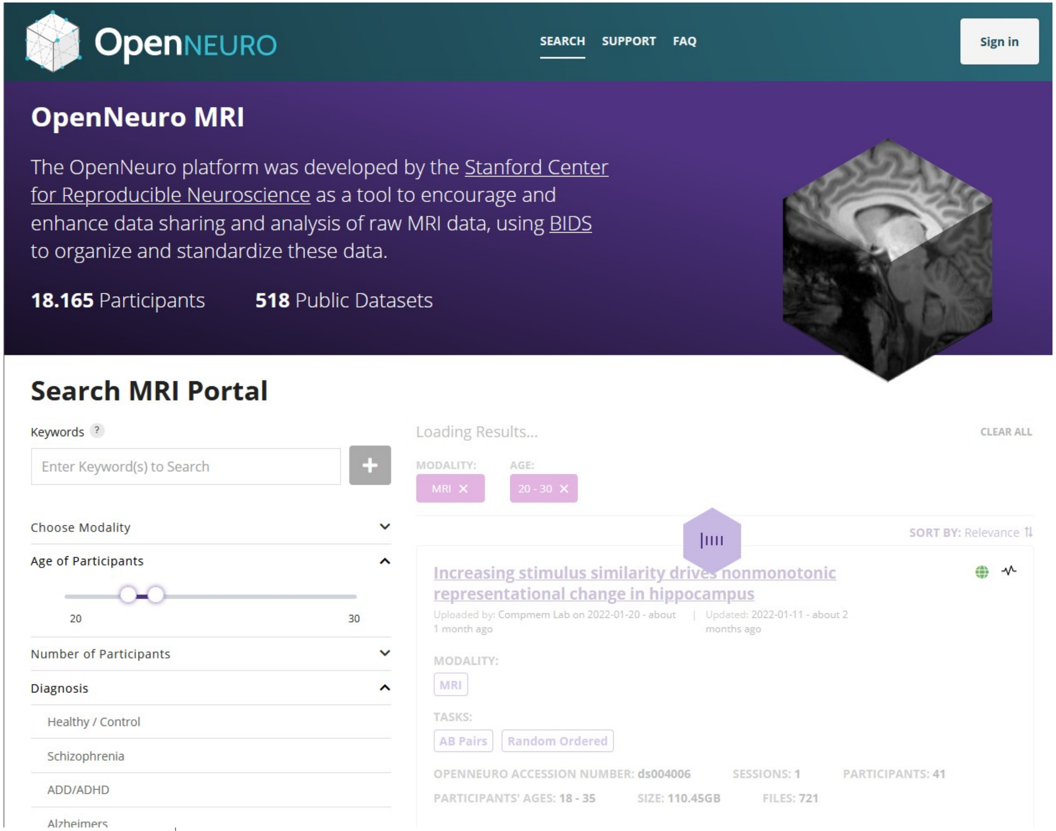
heiDATA



Reusability

- Open content licenses
- transparent versioning
- Provenance information
- Documentation files
- **Integration with external services, e.g. binder (https://mybinder.org/)**

Subject-specific repositories



OpenNEURO

SEARCH SUPPORT FAQ

Sign in

OpenNeuro MRI

The OpenNeuro platform was developed by the Stanford Center for Reproducible Neuroscience as a tool to encourage and enhance data sharing and analysis of raw MRI data, using BIDS to organize and standardize these data.

18.165 Participants **518** Public Datasets

Search MRI Portal

Keywords ?

Enter Keyword(s) to Search

Choose Modality

Age of Participants

20 30

Number of Participants

Diagnosis

Healthy / Control

Schizophrenia

ADD/ADHD

Alzheimers

Loading Results...

CLEAR ALL

MODALITY: MRI ✕

AGE: 20 - 30 ✕

SORT BY: Relevance

Increasing stimulus similarity drives nonmonotonic representational change in hippocampus

Uploaded by: Compmem Lab on 2022-01-20 - about 1 month ago | Updated: 2022-01-11 - about 2 months ago

MODALITY: MRI

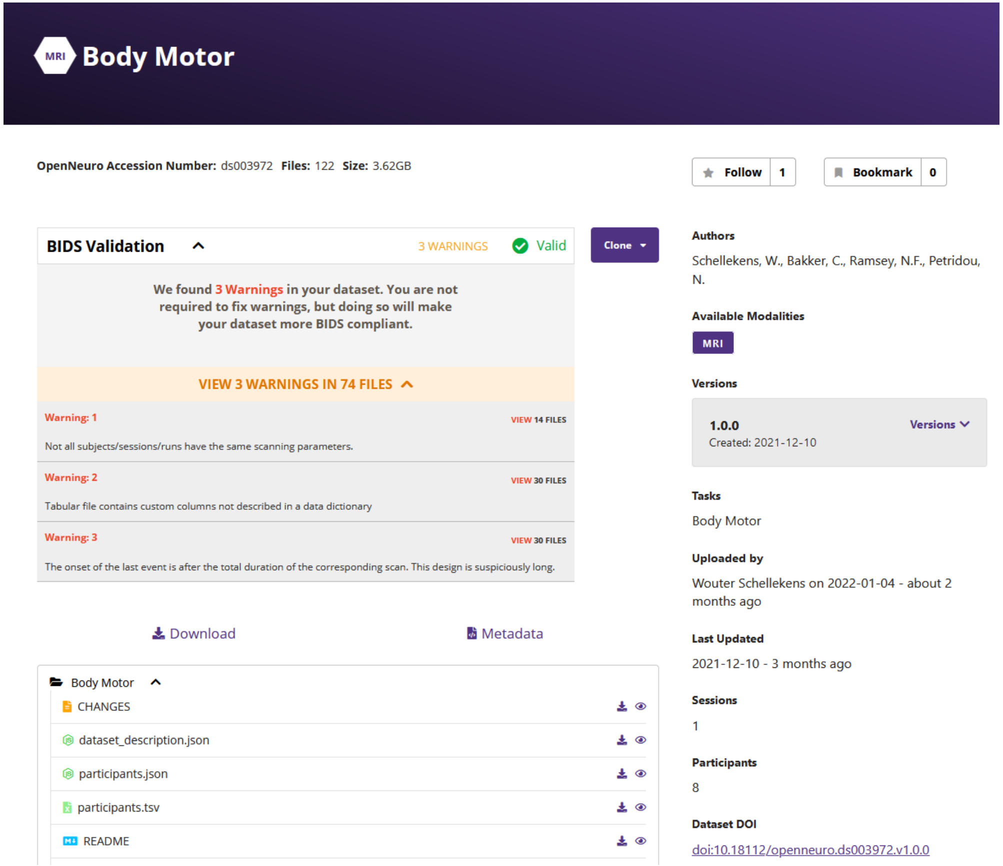
TASKS: AB Pairs Random Ordered

OPENNEURO ACCESSION NUMBER: ds004006 SESSIONS: 1 PARTICIPANTS: 41

PARTICIPANTS' AGES: 18 - 35 SIZE: 110.45GB FILES: 721

- **Domain-specific metadata enable specific functionalites and more effective retrieval.**

Subject-specific repositories



MRI **Body Motor**

OpenNeuro Accession Number: ds003972 **Files:** 122 **Size:** 3.62GB

BIDS Validation 3 WARNINGS ✓ Valid

We found 3 Warnings in your dataset. You are not required to fix warnings, but doing so will make your dataset more BIDS compliant.

VIEW 3 WARNINGS IN 74 FILES

Warning: 1 VIEW 14 FILES
Not all subjects/sessions/runs have the same scanning parameters.

Warning: 2 VIEW 30 FILES
Tabular file contains custom columns not described in a data dictionary

Warning: 3 VIEW 30 FILES
The onset of the last event is after the total duration of the corresponding scan. This design is suspiciously long.

Body Motor
- CHANGES
- dataset_description.json
- participants.json
- participants.tsv
- README

Authors
Schellekens, W., Bakker, C., Ramsey, N.F., Petridou, N.

Available Modalities
MRI

Versions
1.0.0 Versions
Created: 2021-12-10

Tasks
Body Motor

Uploaded by
Wouter Schellekens on 2022-01-04 - about 2 months ago

Last Updated
2021-12-10 - 3 months ago

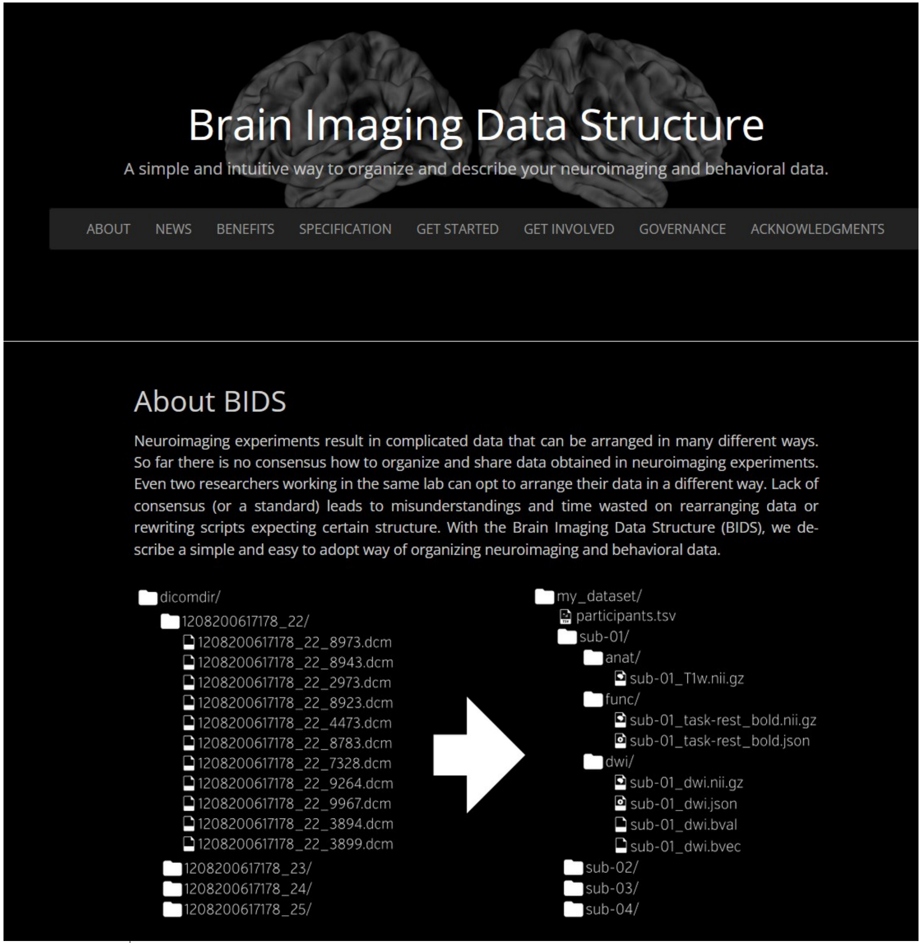
Sessions
1

Participants
8

Dataset DOI
doi:10.18112/openneuro.ds003972.v1.0.0

- Domain-specific metadata enable specific functionalites and more effective retrieval.
- **Data standards are implemented and data are validated against these standards.**

Subject-specific repositories



Brain Imaging Data Structure

A simple and intuitive way to organize and describe your neuroimaging and behavioral data.

About BIDS

Neuroimaging experiments result in complicated data that can be arranged in many different ways. So far there is no consensus how to organize and share data obtained in neuroimaging experiments. Even two researchers working in the same lab can opt to arrange their data in a different way. Lack of consensus (or a standard) leads to misunderstandings and time wasted on rearranging data or rewriting scripts expecting certain structure. With the Brain Imaging Data Structure (BIDS), we describe a simple and easy to adopt way of organizing neuroimaging and behavioral data.

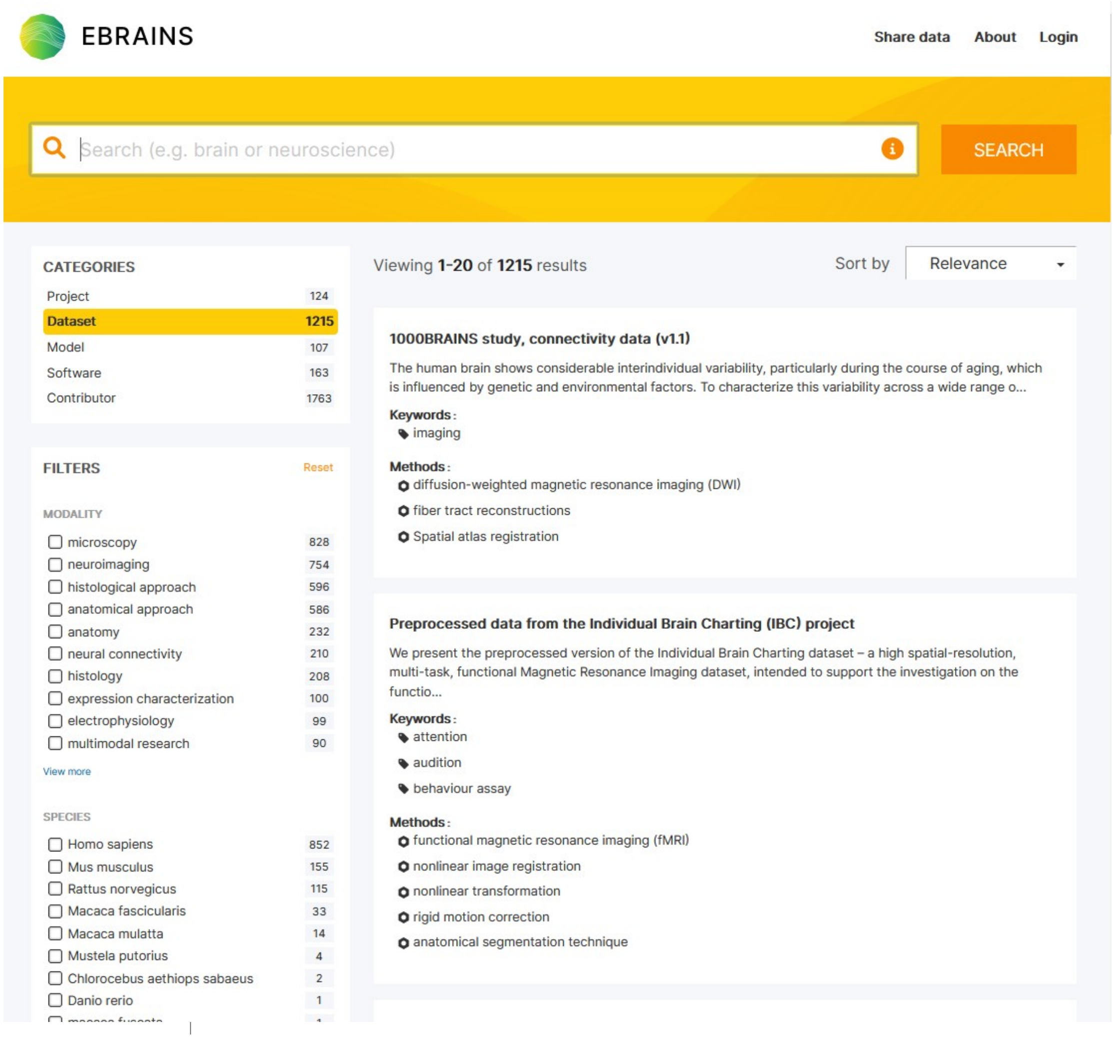
```
dicomdir/
    1208200617178_22/
        1208200617178_22_8973.dcm
        1208200617178_22_8943.dcm
        1208200617178_22_2973.dcm
        1208200617178_22_8923.dcm
        1208200617178_22_4473.dcm
        1208200617178_22_8783.dcm
        1208200617178_22_7328.dcm
        1208200617178_22_9264.dcm
        1208200617178_22_9967.dcm
        1208200617178_22_3894.dcm
        1208200617178_22_3899.dcm
    1208200617178_23/
    1208200617178_24/
    1208200617178_25/
```

```
my_dataset/
    participants.tsv
    sub-01/
        anat/
            sub-01_T1w.nii.gz
        func/
            sub-01_task-rest_bold.nii.gz
            sub-01_task-rest_bold.json
        dwi/
            sub-01_dwi.nii.gz
            sub-01_dwi.json
            sub-01_dwi.bval
            sub-01_dwi.bvec
    sub-02/
    sub-03/
    sub-04/
```

- Domain-specific metadata enable specific functionalites and more effective retrieval.
- **Data standards are implemented and data are validated against these standards.**

Subject-specific repositories

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386



EBRAINS

CATEGORIES

Project	124
Dataset	1215
Model	107
Software	163
Contributor	1763

FILTERS Reset

MODALITY

□ microscopy	828
□ neuroimaging	754
□ histological approach	596
□ anatomical approach	586
□ anatomy	232
□ neural connectivity	210
□ histology	208
□ expression characterization	100
□ electrophysiology	99
□ multimodal research	90

View more

SPECIES

□ Homo sapiens	852
□ Mus musculus	155
□ Rattus norvegicus	115
□ Macaca fascicularis	33
□ Macaca mulatta	14
□ Mustela putorius	4
□ Chlorocebus aethiops sabaeus	2
□ Danio rerio	1
□ macaca fuscata	1

Viewing **1-20** of **1215** results

Sort by: Relevance

1000BRAINS study, connectivity data (v1.1)

The human brain shows considerable interindividual variability, particularly during the course of aging, which is influenced by genetic and environmental factors. To characterize this variability across a wide range o...

Keywords:
- imaging

Methods:
- diffusion-weighted magnetic resonance imaging (DWI)
- fiber tract reconstructions
- Spatial atlas registration

Preprocessed data from the Individual Brain Charting (IBC) project

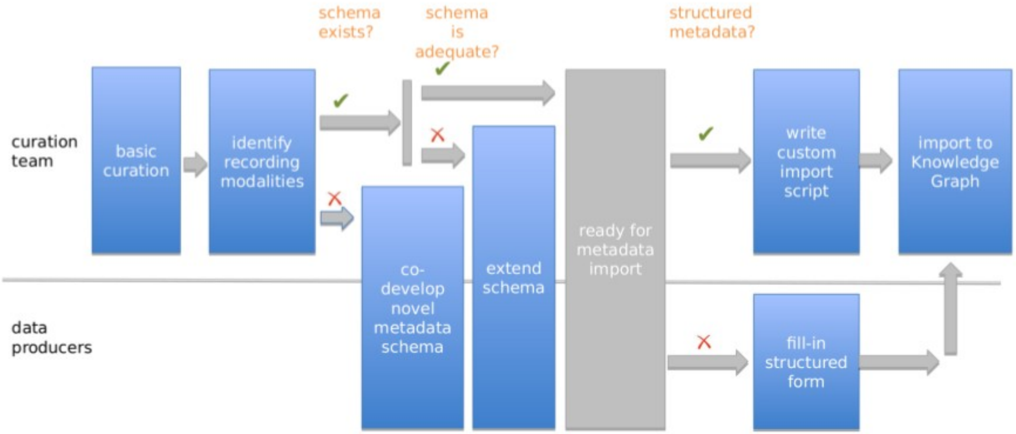
We present the preprocessed version of the Individual Brain Charting dataset – a high spatial-resolution, multi-task, functional Magnetic Resonance Imaging dataset, intended to support the investigation on the functio...

Keywords:
- attention
- audition
- behaviour assay

Methods:
- functional magnetic resonance imaging (fMRI)
- nonlinear image registration
- nonlinear transformation
- rigid motion correction
- anatomical segmentation technique

- Domain-specific metadata enable specific functionalites and more effective retrieval.
- Data standards are implemented and data are validated against these standards.
- **May be limited with regard to data types (e.g. OpenNeuro only accepts human-derived data)**

Subject-specific repositories



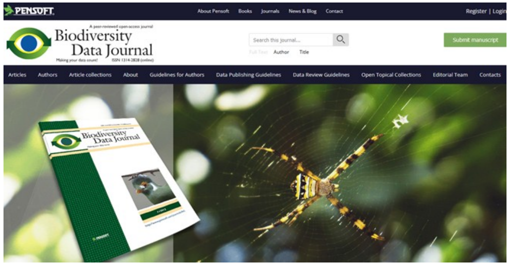
schema exists?
schema is adequate?
structured metadata?
curation team
data producers
basic curation
identify recording modalities
✓
✗
co-develop novel metadata schema
extend schema
ready for metadata import
write custom import script
fill-in structured form
import to Knowledge Graph

https://wiki.ebrains.eu/bin/view/Collabs/tier-3-data-curation/Data%20Curator%27s%20Handbook/

- Domain-specific metadata enable more effective retrieval or specific functionalites.
- Data standards are implemented and data are validated against these standards.
- May be limited with regard to data types (e.g. OpenNeuro only accepts human-derived data)
- **Optional, depending on the repository: Data curators with specific expertise supervise data publication and help preparing data for deposit.**

Finding data journals

https://www.forschungsdaten.org/index.php/Data_Journals



Dynamic World, Near real-time global 10m land use land cover mapping

Christopher F. Brown, Steven P. Brumby ... Alexander M. Tait

Data Descriptor | 09 June 2022

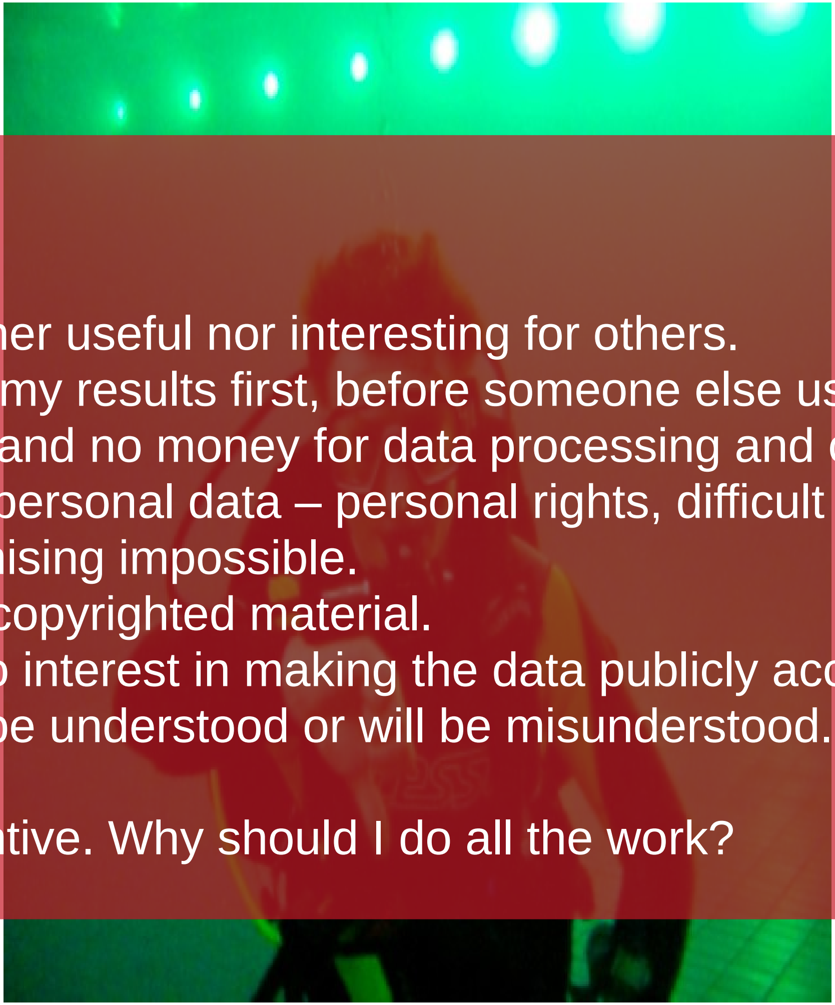
Software Sustainability Institute
Journal of open research software

Data publication

Data publication - pros and cons?

Picture:
Colin Kinner (Flickr: Question mark sign) [CC-BY-2.0], via Wikimedia Commons

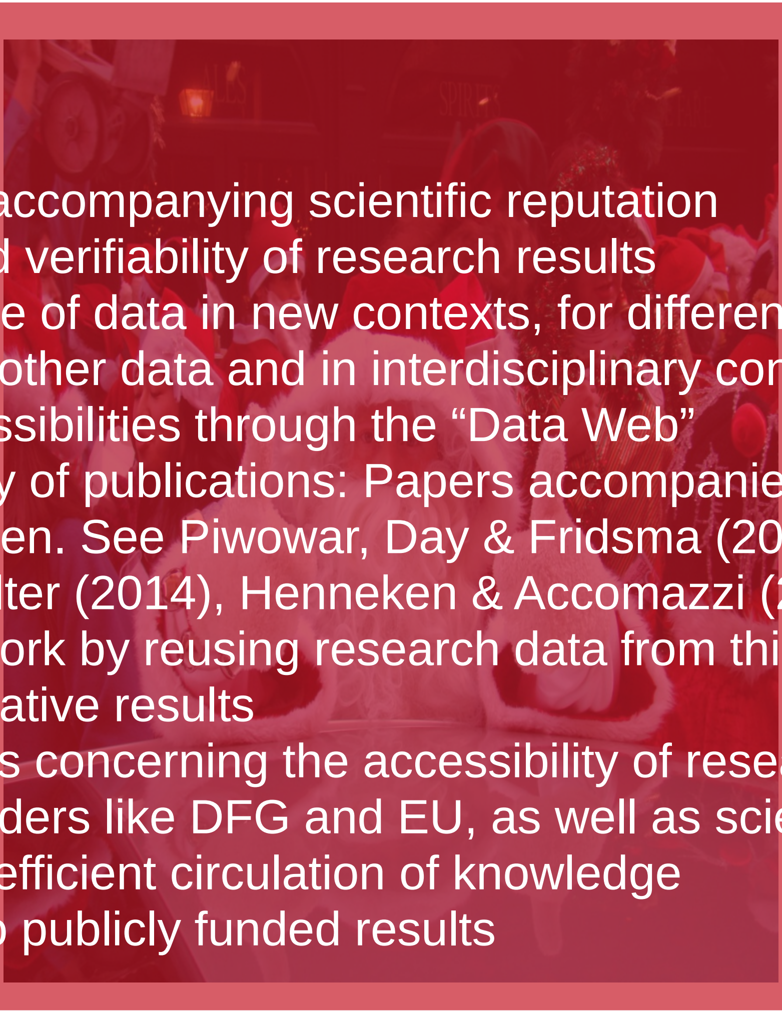
Data publication



Contra

- My data are neither useful nor interesting for others.
- I want to publish my results first, before someone else uses my data.
- There is no time and no money for data processing and curation.
- My data contain personal data – personal rights, difficult search for test persons, anonymising impossible.
- My data include copyrighted material.
- My funder has no interest in making the data publicly accessible.
- My data will not be understood or will be misunderstood. People will bother me with emails.
- There is no incentive. Why should I do all the work?

Data publication

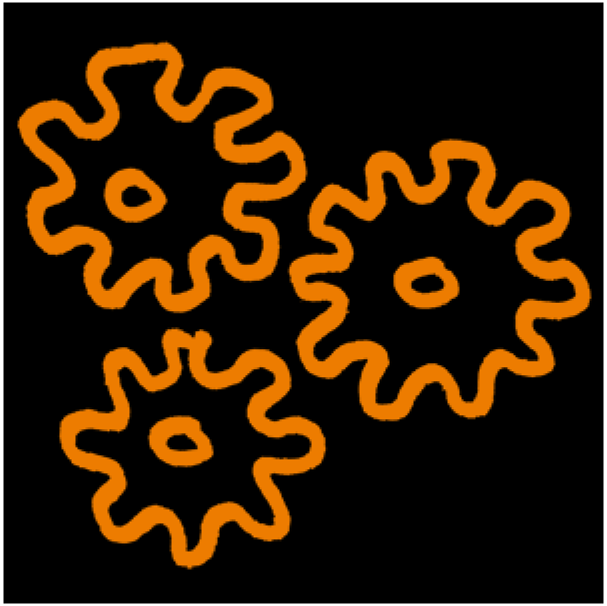


Pro

- Visibility and the accompanying scientific reputation
- Transparency and verifiability of research results
- Possibility of reuse of data in new contexts, for different problems in combination with other data and in interdisciplinary contexts
- New research possibilities through the “Data Web”
- Increased visibility of publications: Papers accompanied by research data are cited more often. See Piwowar, Day & Fridsma (2007), Piwowar & Vision (2013), Belter (2014), Henneken & Accomazzi (2011)
- Avoid duplicate work by reusing research data from third parties.
- Availability of negative results
- Fulfil requirements concerning the accessibility of research material as demanded by funders like DFG and EU, as well as scientific journals.
- Faster and more efficient circulation of knowledge
- Right of access to publicly funded results

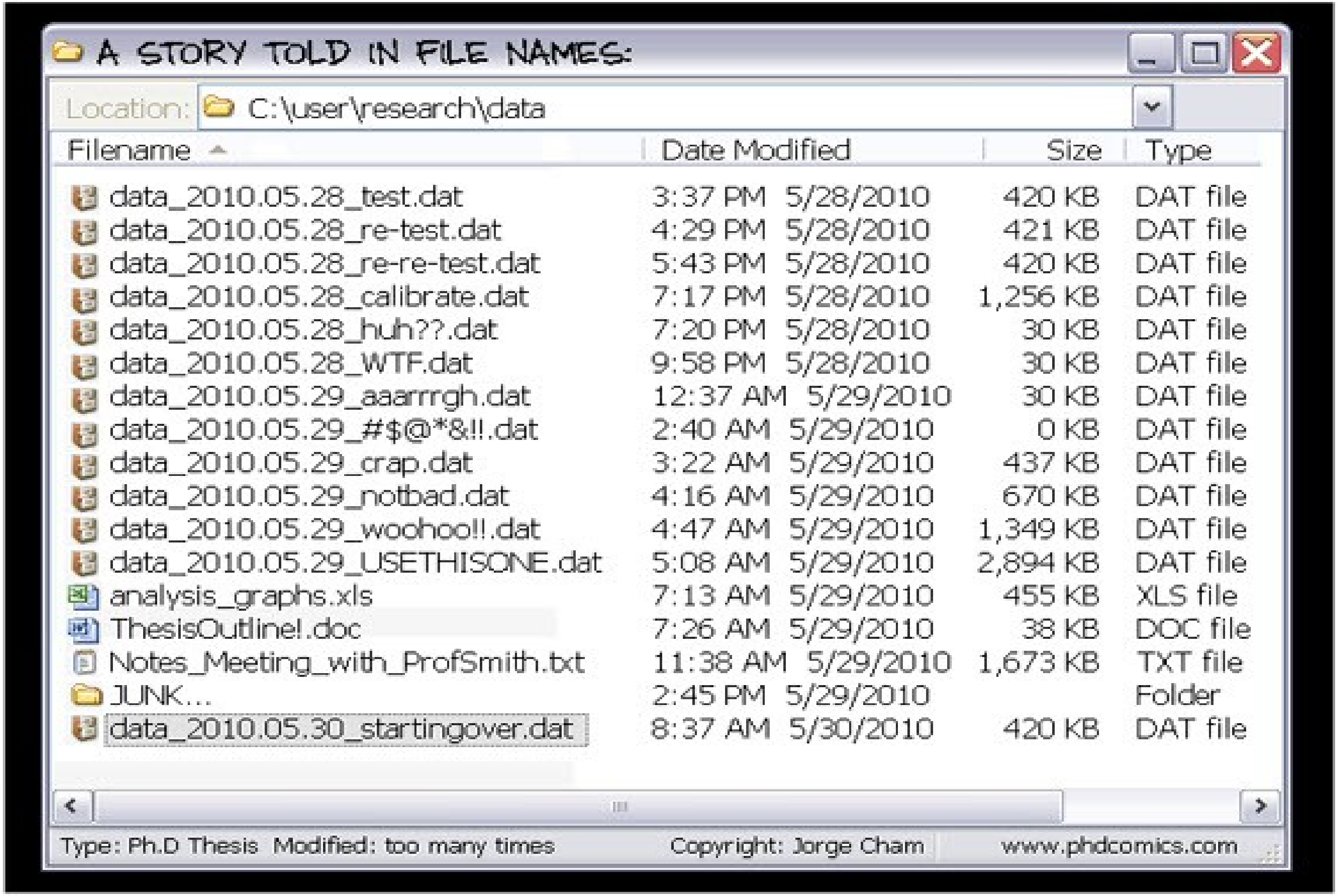
UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Data Handling Storage & Archiving –
some practical issues….



RESEARCH
DATA UNIT
HEIDELBERG

Piled Higher and Deeper ***by Jorge Cham*** **www.phdcomics.com**



A STORY TOLD IN FILE NAMES:

Location: C:\user\research\data

Filename	Date Modified		Size	Type
data_2010.05.28_test.dat	3:37 PM	5/28/2010	420 KB	DAT file
data_2010.05.28_re-test.dat	4:29 PM	5/28/2010	421 KB	DAT file
data_2010.05.28_re-re-test.dat	5:43 PM	5/28/2010	420 KB	DAT file
data_2010.05.28_calibrate.dat	7:17 PM	5/28/2010	1,256 KB	DAT file
data_2010.05.28_huh??.dat	7:20 PM	5/28/2010	30 KB	DAT file
data_2010.05.28_WTF.dat	9:58 PM	5/28/2010	30 KB	DAT file
data_2010.05.29_aaarrgh.dat	12:37 AM	5/29/2010	30 KB	DAT file
data_2010.05.29_#$@*&!!.dat	2:40 AM	5/29/2010	0 KB	DAT file
data_2010.05.29_crap.dat	3:22 AM	5/29/2010	437 KB	DAT file
data_2010.05.29_notbad.dat	4:16 AM	5/29/2010	670 KB	DAT file
data_2010.05.29_woohoo!!.dat	4:47 AM	5/29/2010	1,349 KB	DAT file
data_2010.05.29_USETHISONE.dat	5:08 AM	5/29/2010	2,894 KB	DAT file
analysis_graphs.xls	7:13 AM	5/29/2010	455 KB	XLS file
ThesisOutline!.doc	7:26 AM	5/29/2010	38 KB	DOC file
Notes_Meeting_with_ProfSmith.txt	11:38 AM	5/29/2010	1,673 KB	TXT file
JUNK...	2:45 PM	5/29/2010		Folder
data_2010.05.30_startingover.dat	8:37 AM	5/30/2010	420 KB	DAT file

Type: Ph.D Thesis Modified: too many times Copyright: Jorge Cham www.phdcomics.com

title: "A story in file names" - originally published 5/28/2010

Data Handling Storage & Archiving

File handling

- Data best practices (file naming, formats, versioning,…): https://guides.library.stanford.edu/data-best-practices/
- Make different versions of data distinguishable. Conventions for file naming – for you and in your research group.
- File names should deliver context. Distinguish a file from similar but different datasets and from different versions of the same dataset.
- Files may leave their folders. File names should be unique and descriptive without a directory structure.
- **Never delete your raw data!**
- But delete versions of processed data you do not need any longer.

Data Storage & Archiving – some practical issues….

Backup

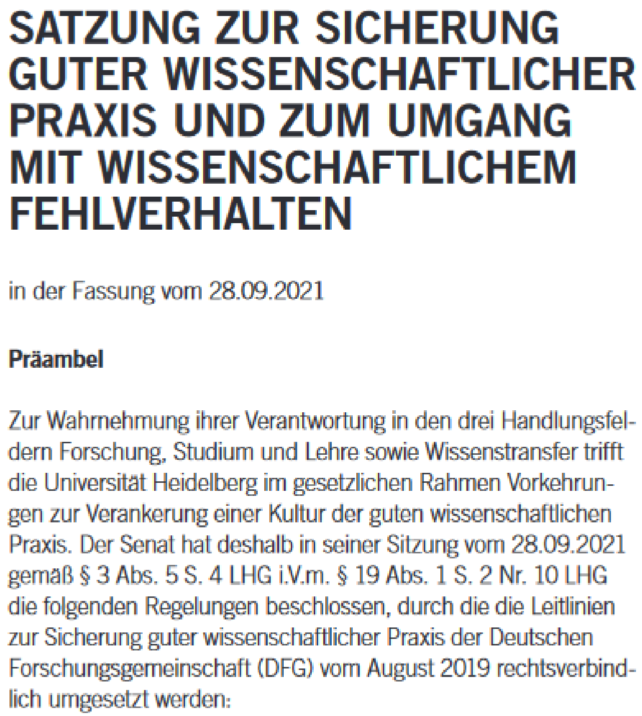
3 … 2 … 1 … Backup!

- At least 3 copies per file…
- …on at least 2 different media…
- and 1 at a different spatial location.

EXTERNAL REQUIREMENTS & POLICIES

Picture: https://www.flickr.com/photos/raummaschine/9214045295/

Policies & external requirements



SATZUNG ZUR SICHERUNG GUTER WISSENSCHAFTLICHER PRAXIS UND ZUM UMGANG MIT WISSENSCHAFTLICHEM FEHLVERHALTEN

in der Fassung vom 28.09.2021

Präambel

Zur Wahrnehmung ihrer Verantwortung in den drei Handlungsfeldern Forschung, Studium und Lehre sowie Wissenstransfer trifft die Universität Heidelberg im gesetzlichen Rahmen Vorkehrungen zur Verankerung einer Kultur der guten wissenschaftlichen Praxis. Der Senat hat deshalb in seiner Sitzung vom 28.09.2021 gemäß § 3 Abs. 5 S. 4 LHG i.V.m. § 19 Abs. 1 S. 2 Nr. 10 LHG die folgenden Regelungen beschlossen, durch die die Leitlinien zur Sicherung guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft (DFG) vom August 2019 rechtsverbindlich umgesetzt werden:

Policies & external requirements

Heidelberg University

Rules for safeguarding good academic practice and handling academic misconduct

§ 10 Documentation

(1) Researchers must document all information relevant to the establishment of a research result with the degree of transparency that is required and appropriate in the respective field. The same applies to individual results that do not support the research hypothesis. There must be no selection of results in such cases. Where research software is developed, the source code must be documented.

(2) The information required to understand the research, in particular research data and methodological, evaluation and analysis steps, is recorded. Third parties are to be given access to this information where this is possible.

Policies & external requirements

Heidelberg University

Rules for safeguarding good academic practice and handling academic misconduct

§ 11 Public access to research findings

“Researchers decide on their own responsibility whether, how and where to make their research findings publicly available. If they decide to publish their results, the data and principal materials upon which the published work is based must be stored in recognised archives and repositories where this is possible. The provisions of § 14 must be observed.”

Policies & external requirements

Heidelberg University

Rules for safeguarding good academic practice and handling academic misconduct

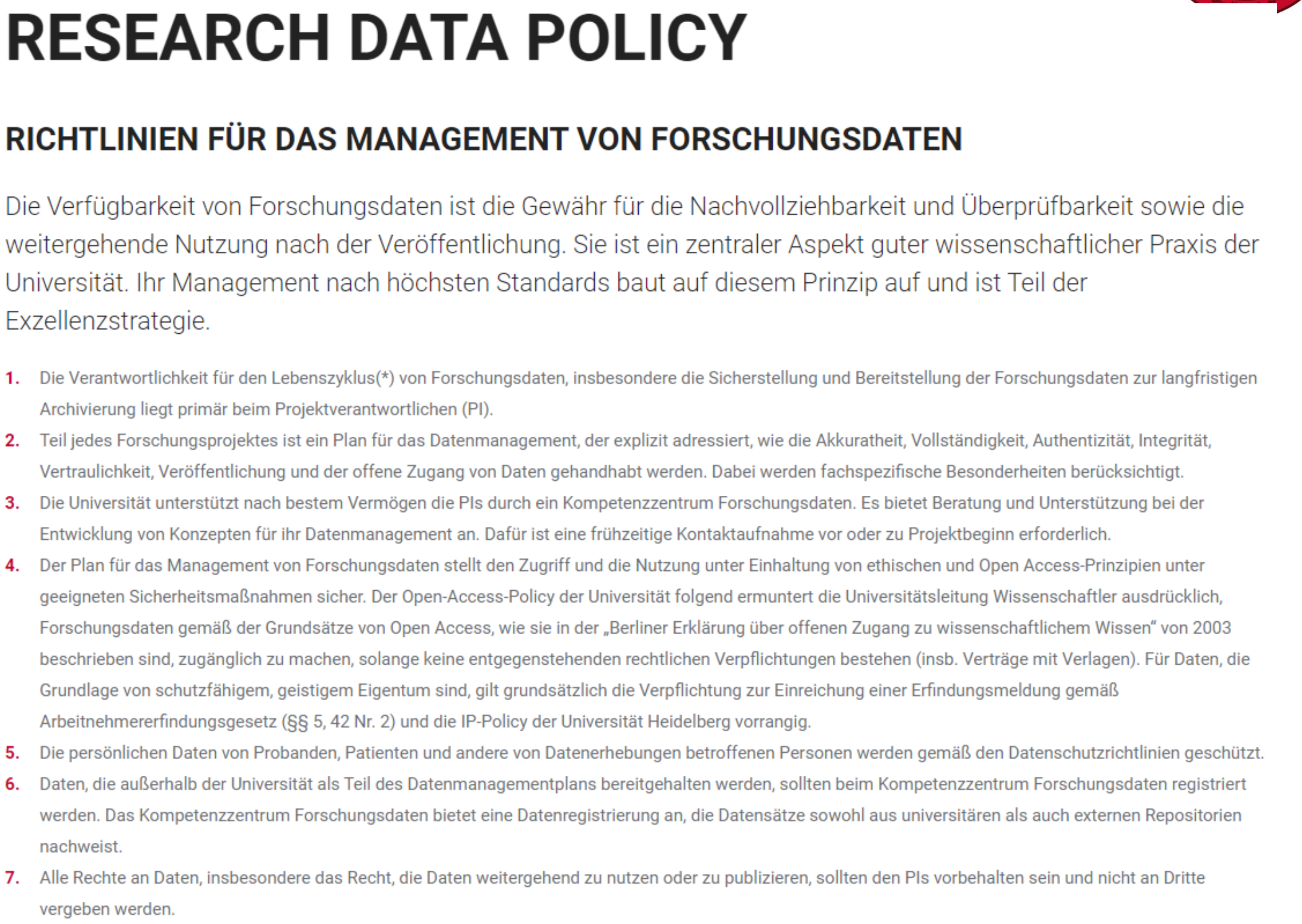
§ 16 Archiving

“(1) Once they have been made publicly available, research data and findings, and particularly the materials on which they are based, as well as the instruments and, where applicable, the research software used, must be backed up by adequate means according to the standards of the respective field and stored for the legally required time period (usually ten years). A shortening of this storage period must be justified. The storage period begins when the materials are first made publicly available.

(2) The materials are archived a) in the researchers’ home institution or b) in repositories serving several locations. In case a) the university will provide the necessary infrastructure for archiving. The selected publication medium must make reference to the archiving location in an appropriate manner.”

Policies & external requirements

UNIVERSITÄT HEIDELBERG
ZUKUNFT
SEIT 1386



RESEARCH DATA POLICY

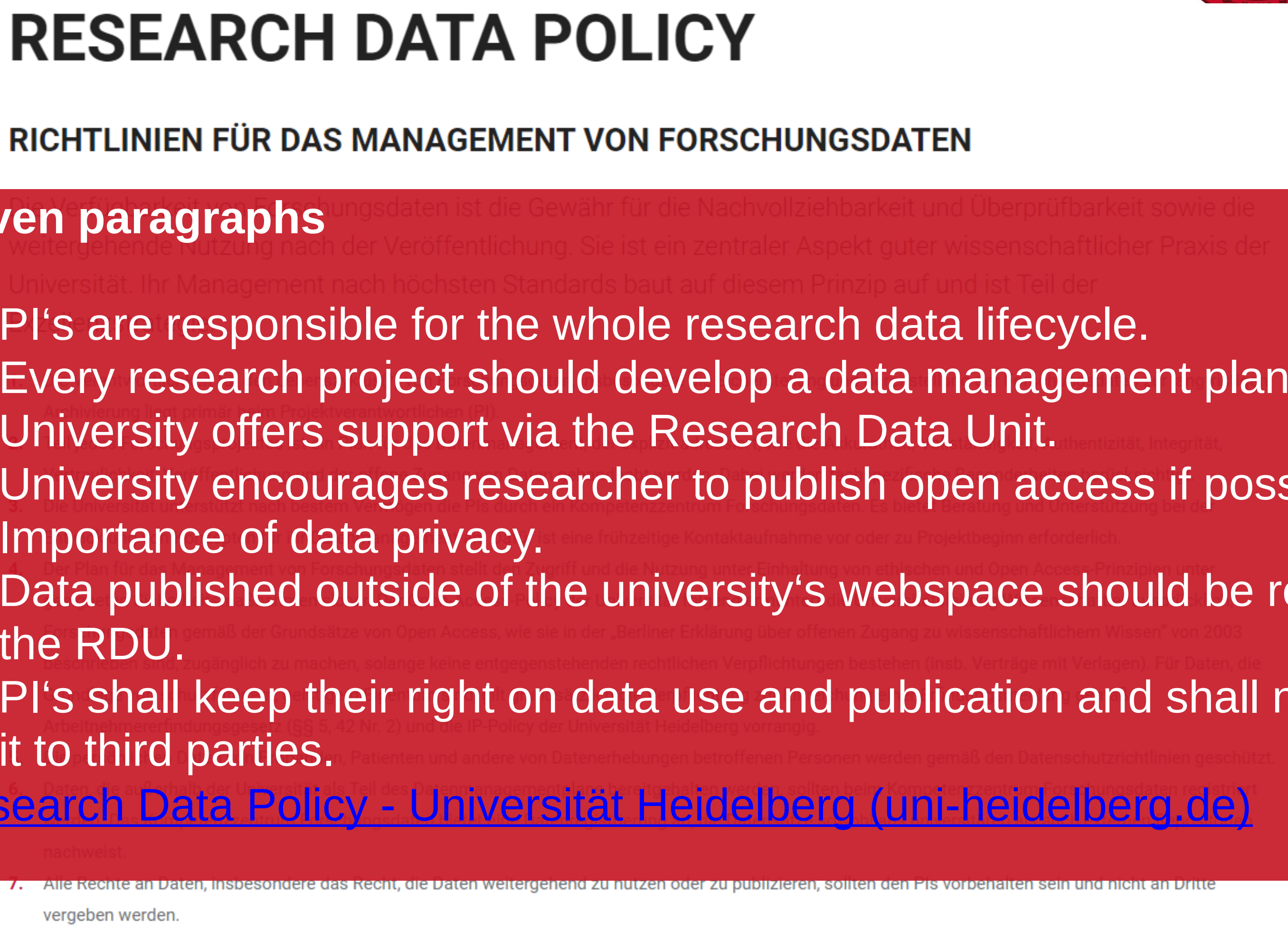
RICHTLINIEN FÜR DAS MANAGEMENT VON FORSCHUNGSDATEN

Die Verfügbarkeit von Forschungsdaten ist die Gewähr für die Nachvollziehbarkeit und Überprüfbarkeit sowie die weitergehende Nutzung nach der Veröffentlichung. Sie ist ein zentraler Aspekt guter wissenschaftlicher Praxis der Universität. Ihr Management nach höchsten Standards baut auf diesem Prinzip auf und ist Teil der Exzellenzstrategie.

1. Die Verantwortlichkeit für den Lebenszyklus(*) von Forschungsdaten, insbesondere die Sicherstellung und Bereitstellung der Forschungsdaten zur langfristigen Archivierung liegt primär beim Projektverantwortlichen (PI).
2. Teil jedes Forschungsprojektes ist ein Plan für das Datenmanagement, der explizit adressiert, wie die Akkuratheit, Vollständigkeit, Authentizität, Integrität, Vertraulichkeit, Veröffentlichung und der offene Zugang von Daten gehandhabt werden. Dabei werden fachspezifische Besonderheiten berücksichtigt.
3. Die Universität unterstützt nach bestem Vermögen die PIs durch ein Kompetenzzentrum Forschungsdaten. Es bietet Beratung und Unterstützung bei der Entwicklung von Konzepten für ihr Datenmanagement an. Dafür ist eine frühzeitige Kontaktaufnahme vor oder zu Projektbeginn erforderlich.
4. Der Plan für das Management von Forschungsdaten stellt den Zugriff und die Nutzung unter Einhaltung von ethischen und Open Access-Prinzipien unter geeigneten Sicherheitsmaßnahmen sicher. Der Open-Access-Policy der Universität folgend ermuntert die Universitätsleitung Wissenschaftler ausdrücklich, Forschungsdaten gemäß der Grundsätze von Open Access, wie sie in der „Berliner Erklärung über offenen Zugang zu wissenschaftlichem Wissen" von 2003 beschrieben sind, zugänglich zu machen, solange keine entgegenstehenden rechtlichen Verpflichtungen bestehen (insb. Verträge mit Verlagen). Für Daten, die Grundlage von schutzfähigem, geistigem Eigentum sind, gilt grundsätzlich die Verpflichtung zur Einreichung einer Erfindungsmeldung gemäß Arbeitnehmererfindungsgesetz (§§ 5, 42 Nr. 2) und die IP-Policy der Universität Heidelberg vorrangig.
5. Die persönlichen Daten von Probanden, Patienten und andere von Datenerhebungen betroffenen Personen werden gemäß den Datenschutzrichtlinien geschützt.
6. Daten, die außerhalb der Universität als Teil des Datenmanagementplans bereitgehalten werden, sollten beim Kompetenzzentrum Forschungsdaten registriert werden. Das Kompetenzzentrum Forschungsdaten bietet eine Datenregistrierung an, die Datensätze sowohl aus universitären als auch externen Repositorien nachweist.
7. Alle Rechte an Daten, insbesondere das Recht, die Daten weitergehend zu nutzen oder zu publizieren, sollten den PIs vorbehalten sein und nicht an Dritte vergeben werden.

RESEARCH
DATA UNIT
HEIDELBERG

Policies & external requirements



RESEARCH DATA POLICY

RICHTLINIEN FÜR DAS MANAGEMENT VON FORSCHUNGSDATEN

Seven paragraphs

1) PI‘s are responsible for the whole research data lifecycle.
2) Every research project should develop a data management plan.
3) University offers support via the Research Data Unit.
4) University encourages researcher to publish open access if possible.
5) Importance of data privacy.
6) Data published outside of the university‘s webspace should be registered at the RDU.
7) PI‘s shall keep their right on data use and publication and shall not transfer it to third parties.

Research Data Policy - Universität Heidelberg (uni-heidelberg.de)

7. Alle Rechte an Daten, insbesondere das Recht, die Daten weitergehend zu nutzen oder zu publizieren, sollten den PIs vorbehalten sein und nicht an Dritte vergeben werden.

Funders are pushing RDM & Open Data

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

DFG Deutsche Forschungsgemeinschaft

DFG Guidelines on the Handling of Research Data

"[…] For this reason, the handling of research data and the objects on which the data is based have to be carefully planned, documented and described. Wherever possible it is important to enable subsequent use of the research data and potentially also the objects by other users.
[…]
For this reason, the DFG expects research projects to include a description of how research data is handled. The description should be based on the checklist for handling research data
[…]
Costs incurred for the project-specific handling of research data should be requested in connection with the project.[…]"

RESEARCH DATA UNIT HEIDELBERG

Checklist Regarding the Handling of Research Data

1. Data description

How does your project generate new data? Is existing data reused? Which data types (in terms of data formats like image data, text data or measurement data) arise in your project and in what way are they further processed? To what extent do these arise or what is the anticipated data volume?

2. Documentation and data quality

What approaches are being taken to describe the data in a comprehensible manner (such as the use of available metadata, documentation standards or ontologies)? What measures are being adopted to ensure high data quality? Are quality controls in place and if so, how do they operate? Which digital methods and tools (e.g. software) are required to use the data?

3. Storage and technical archiving the project

How is the data to be stored and archived throughout the project duration? What is in

Funders are pushing RDM & Open Data

Horizon 2020 & Horizon Europe: FAIR Data Management

- Participating projects will be required to develop a **Data Management Plan** (DMP)
- Participating projects are **required to deposit research data**, preferably into a research data repository
- “[…]as far as possible, projects must then **take measures to enable for third parties to access**, mine, exploit, reproduce and disseminate (free of charge for any user) this research data.“
- http://www.dfg.de/foerderung/antrag_gutachter_gremien/antragstellende/nachnutzung_forschungsdaten/
- Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 | Guidelines on Data Management in Horizon 2020

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

What types and formats of data will the project generate/collect?

Will you re-use any existing data and how?

What is the origin of the data?

What is the expected size of the data?

To whom might it be useful ('data utility')?

2. FAIR data

2. 1. Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

What naming conventions do you follow?

Will search keywords be provided that optimize possibilities for re-use?

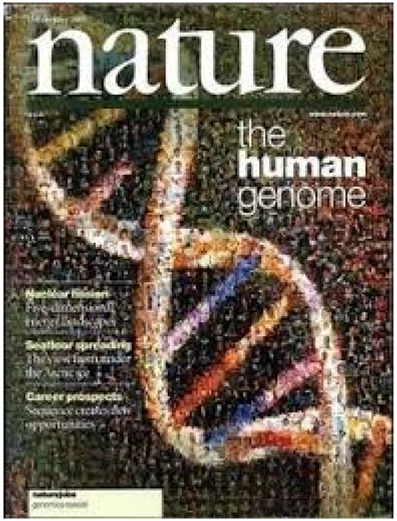
Journals: Nature

An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims. A condition of publication in a Nature Portfolio journal is that **authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications.**

[…]Providing large datasets in supplementary information is strongly discouraged and the preferred approach is to make data available in repositories.

https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards#availability-of-data

https://www.nature.com/sdata/policies/repositories



Journals: PLOS

Data Availability

PLOS journals require authors to make all data necessary to replicate their study's findings publicly available without restriction at the time of publication. When specific legal or ethical restrictions prohibit public sharing of a data set, authors must indicate how others may obtain access to the data.

[…]

Publication is conditional on compliance with this policy. If restrictions on access to data come to light after publication, we reserve the right to post a Correction, an Editorial Expression of Concern, contact the authors' institutions and funders, or, in extreme cases, retract the publication. […]

https://journals.plos.org/plosone/s/data-availability

LEGAL ISSUES

Legal issues

Research data and copyright

- Textual data typically are protected by copyright
- Copy right holder can grant simple or exclusive usage rights
- For publications in subscription journals: typically unlimited and irrevocable transfer of rights to the publishers

- Research data? Facts like measurements generally do not reach the threshold of originality, even though the data collection can be very sophisticated.
- Therefore: According to German copyright law, research data are in many cases not copyrighted.
- But many data are in databases and there is some kind of protection for these (EU directive 96/09/EG, UrhG §§ 87a-e). Virtually all data are useless without documentation. This documentation might very well be protected by copyright.

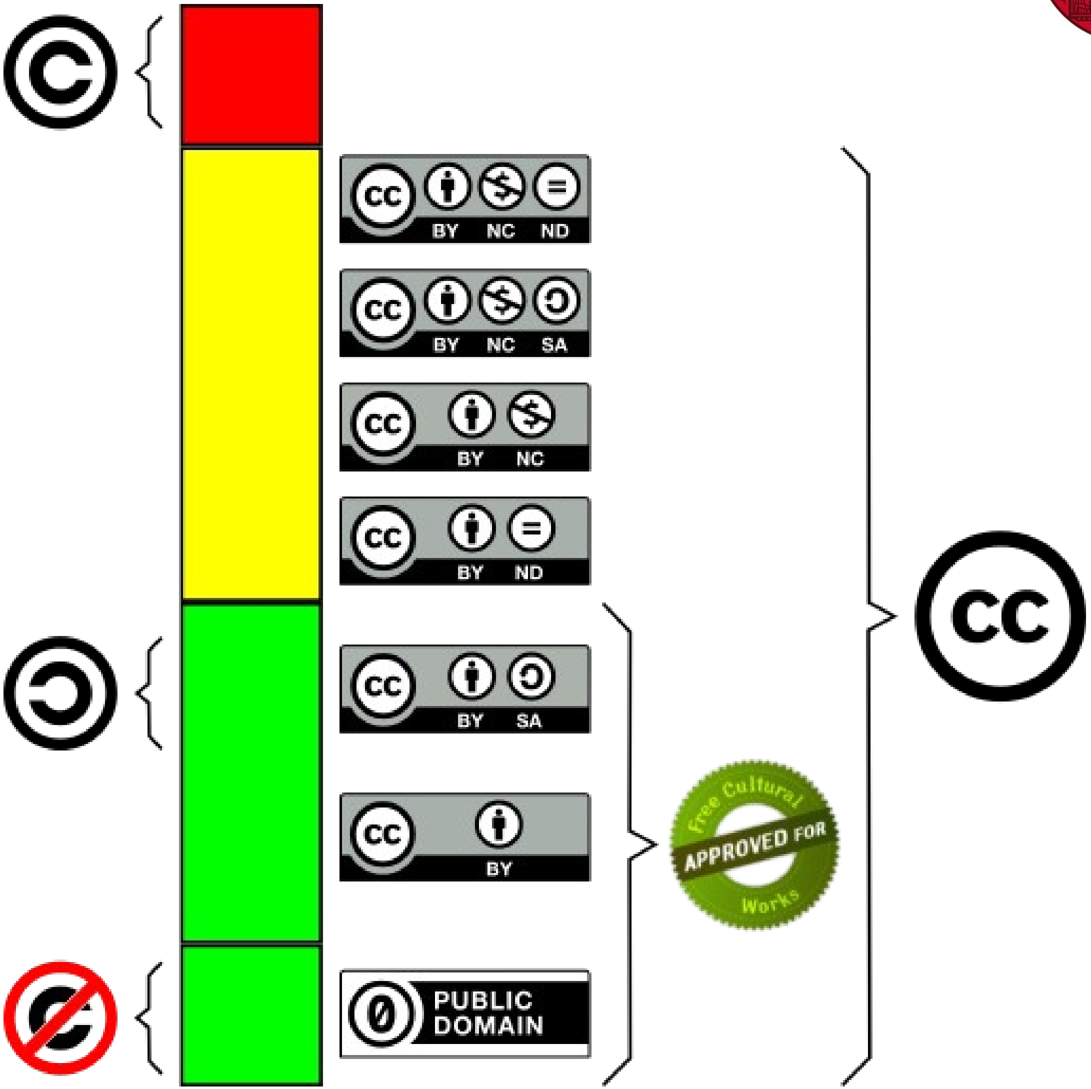
Legal issues

Creative Commons Licences

- Standard licences that determine the scope of use of a work
- Combination of layperson-friendly formulation and a legally proper license text adapted to the relevant national law.
- Licence content and metadata are available in machine readable form and can be added to a document. (□ TDM)
- Modular structure with differing “degrees of freedom”

- There are also alternatives, e.g. the Open Data Commons licenses.
- For Software there are specific software licenses

Legal issues



Picture by Marko Txopitea "Txopi" [CC0], via Wikimedia Commons, https://commons.wikimedia.org/wiki/File%3ACreative_Commons_Semaforoa.svg

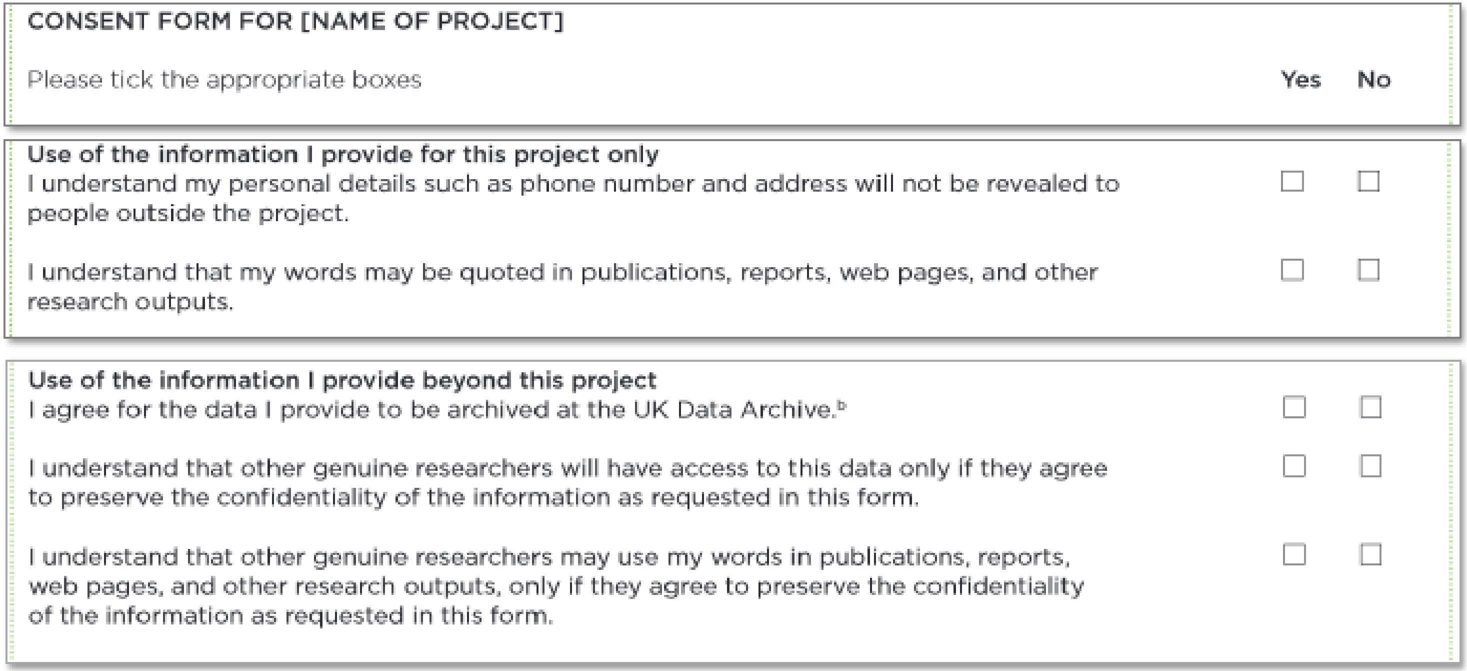
Legal issues

Data publication and data protection

- Informed consent: data sharing not excluded; information on whether and how data are disseminated

Beispiel UK Data Archive „Managing and Sharing Data“:

SAMPLE CONSENT FORM FOR INTERVIEWS



CONSENT FORM FOR [NAME OF PROJECT]		
Please tick the appropriate boxes	Yes	No
Use of the information I provide for this project only		
I understand my personal details such as phone number and address will not be revealed to people outside the project.	□	□
I understand that my words may be quoted in publications, reports, web pages, and other research outputs.	□	□
Use of the information I provide beyond this project		
I agree for the data I provide to be archived at the UK Data Archive.[b]	□	□
I understand that other genuine researchers will have access to this data only if they agree to preserve the confidentiality of the information as requested in this form.	□	□
I understand that other genuine researchers may use my words in publications, reports, web pages, and other research outputs, only if they agree to preserve the confidentiality of the information as requested in this form.	□	□

Thank you very much!

Dr. Georg Schwesinger
schwesinger@ub.uni-heidelberg.de

Dr. Sebastian Zangerle
sebastian.zangerle@ub.uni-heidelberg.de

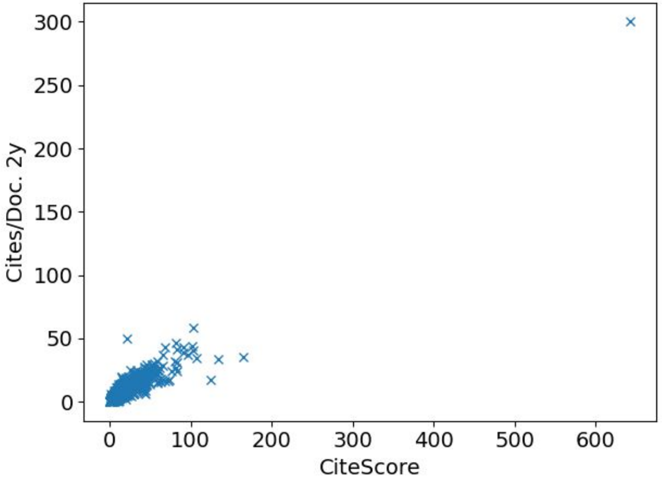
Research Data Unit
https://data.uni-heidelberg.de/

General Information on RDM
https://www.forschungsdaten.info/

3. Research Data Quality

Collecting data

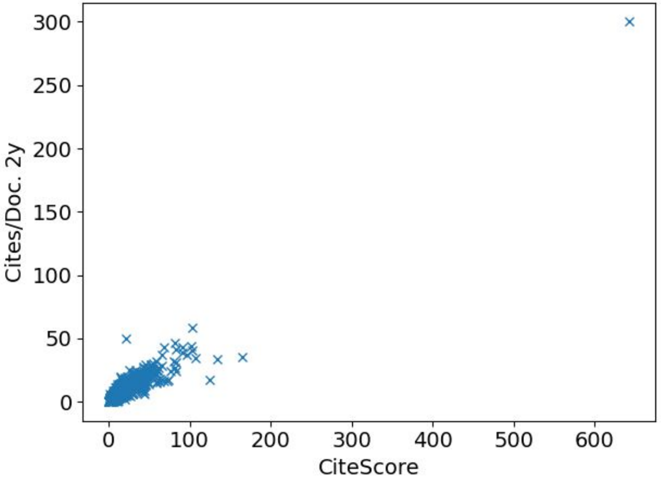
- Data must contain all ranges of the condition that is to be sampled
  - For example: To predict the impact of temperature on reactivity, all temperatures that are of interest need to be sampled (predictions only interpolate between data points but cannot extrapolate).
  - For example: CiteScore (Scopus citation index) vs. citations over all documents from last 2 years, for scientific journals.



Dataset: journal ranking dataset https://www.kaggle.com/datasets/xabirhasan/journal-ranking-dataset

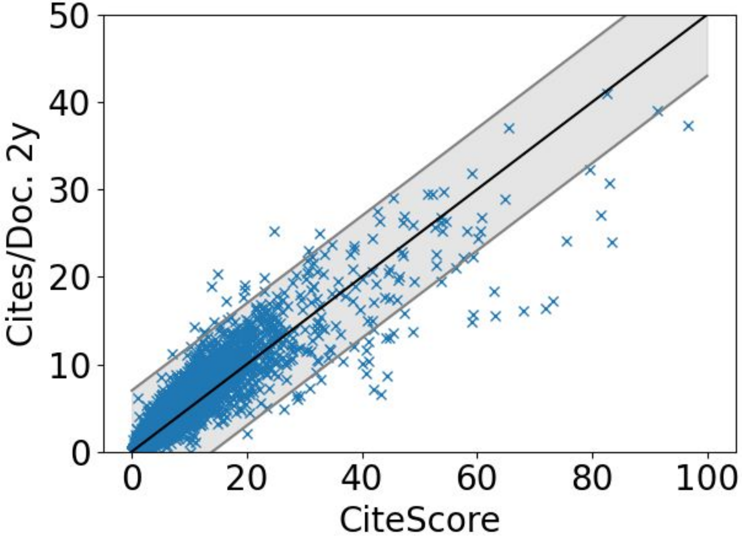
Collecting data

- Data must be homogeneous throughout feature space
  - For example: If temperature and pressure are both sampled, all combinations of features must be recorded for a homogeneous distribution of data points.
  - For example: CiteScore (Scopus citation index) vs. citations over all documents from last 2 years, for scientific journals.



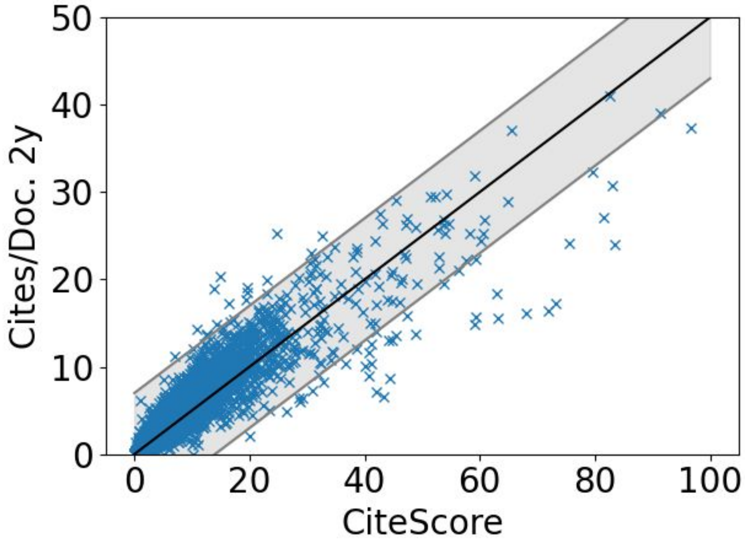
Collecting data

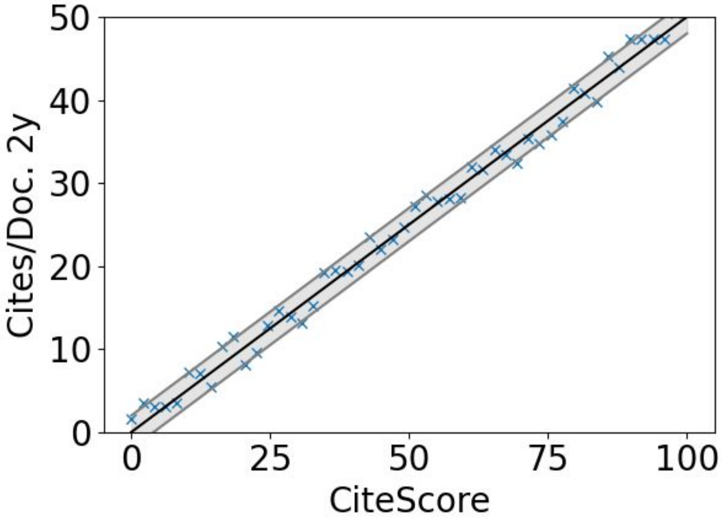
- Data must be of good quality
  - Whether it is real or synthetic data, the model can only make accurate predictions if the data itself is accurate.
  - For example: CiteScore (Scopus citation index) vs. citations over all documents from last 2 years, for scientific journals



Collecting data

- Data volume must be sufficient
  - Only with enough data can a model be trained to make accurate predictions.
  - For example: Complex data - more data points required; simpler data - fewer data points required





Collecting data

- Depending on the type of learning, data must be labeled and labeled correctly
  - Incorrect labelling interferes with the learning process.

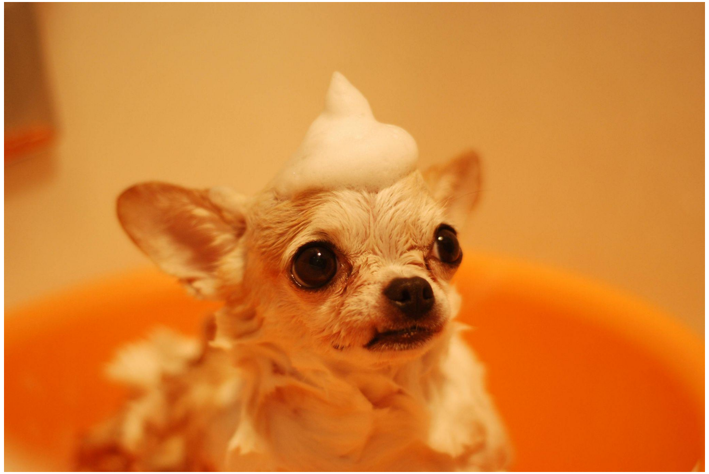


Photo by nishizuka: https://www.pexels.com/photo/brown-chihuahua-485294/

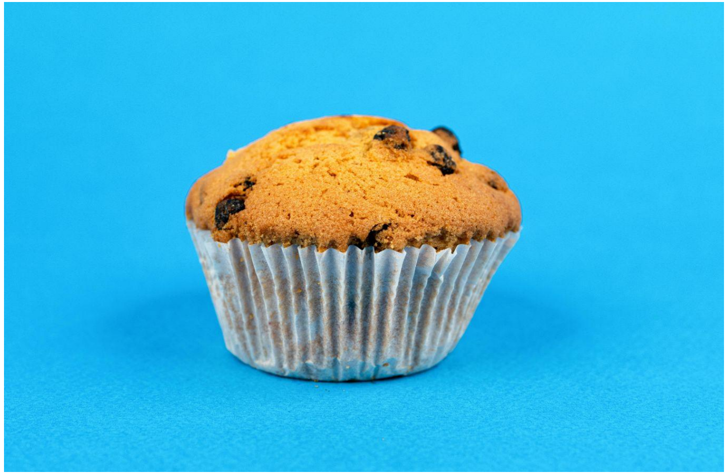


Photo by Maksim Goncharenok: https://www.pexels.com/photo/a-chocolate-muffin-on-blue-surface-5994864/

Data preparation

- Make sure data is clean.
    - Correct typos, misidentified data types

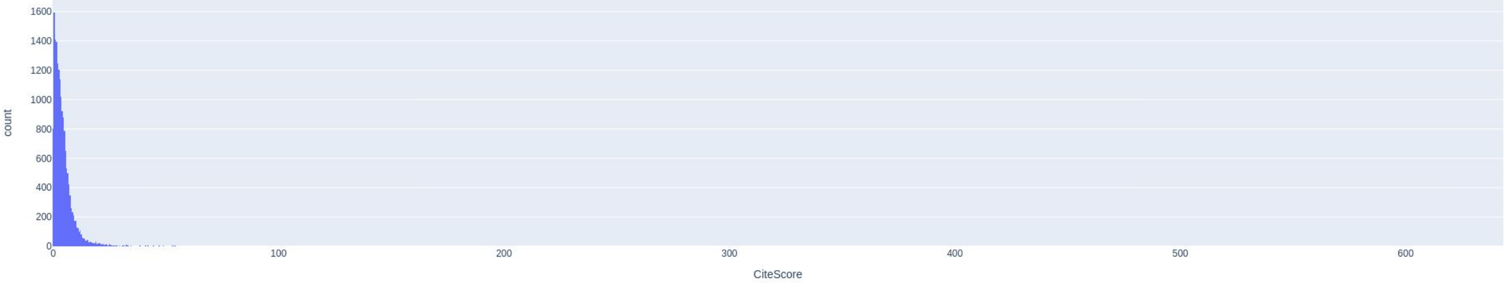
Chihuahuah →Chihuahua

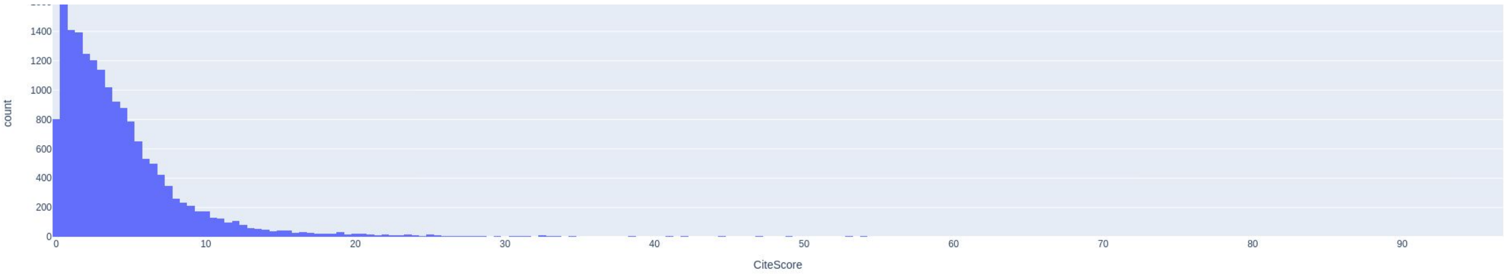
Photo by nishizuka:
https://www.pexels.com/photo/brown-chihuahua-485294/

“26-04-24” →`2024-04-26`

Data preparation

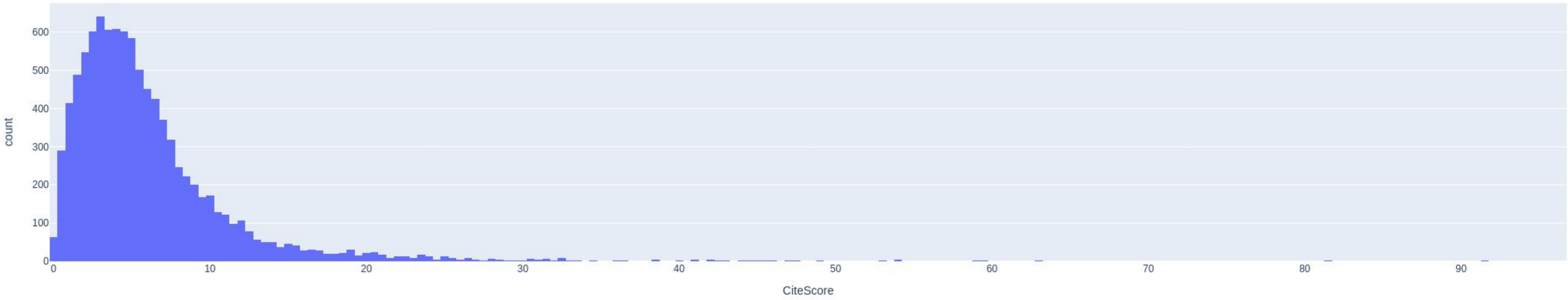
- Make sure data is homogeneous.
  - Visualize the data and use clustering analysis to identify outliers.
  - Use `df.describe()` and `plotly.express` to better understand your data





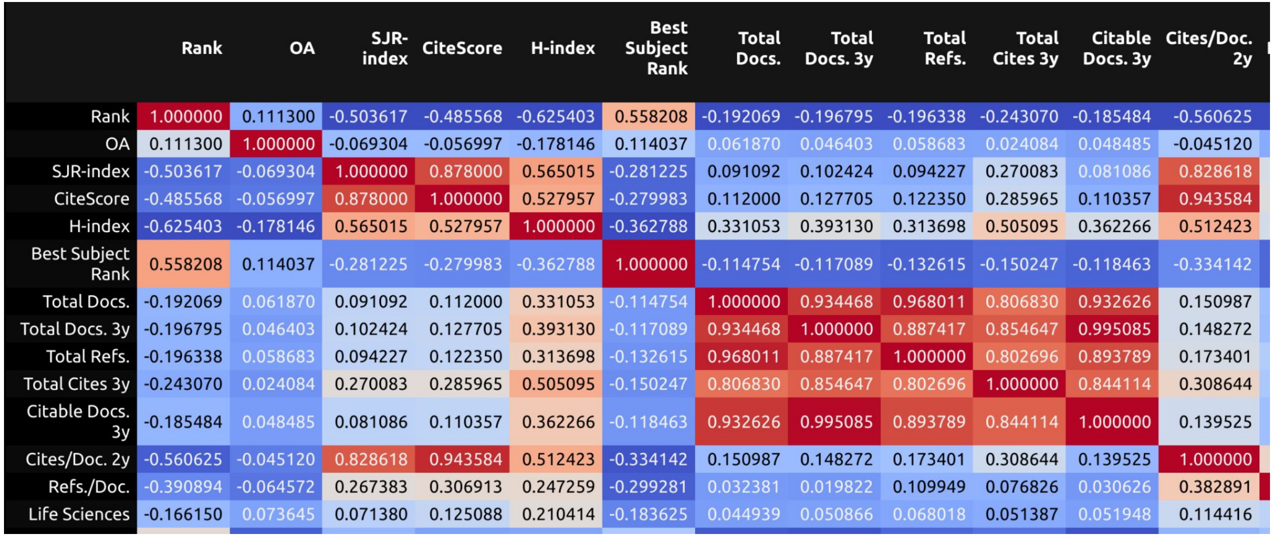
Data preparation

- Remove duplicates.
  - Duplicates introduce bias.
  - Use df.drop_duplicates()



Data preparation

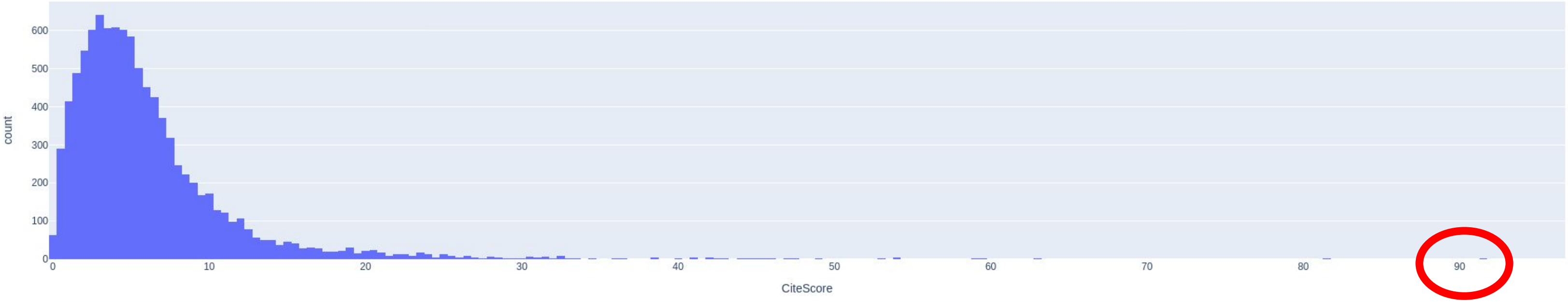
- Feature Engineering: Select influential features, remove unnecessary ones.
  - Unimportant features increase the complexity and reduce robustness.
  - For example: only choose features that are clearly correlated

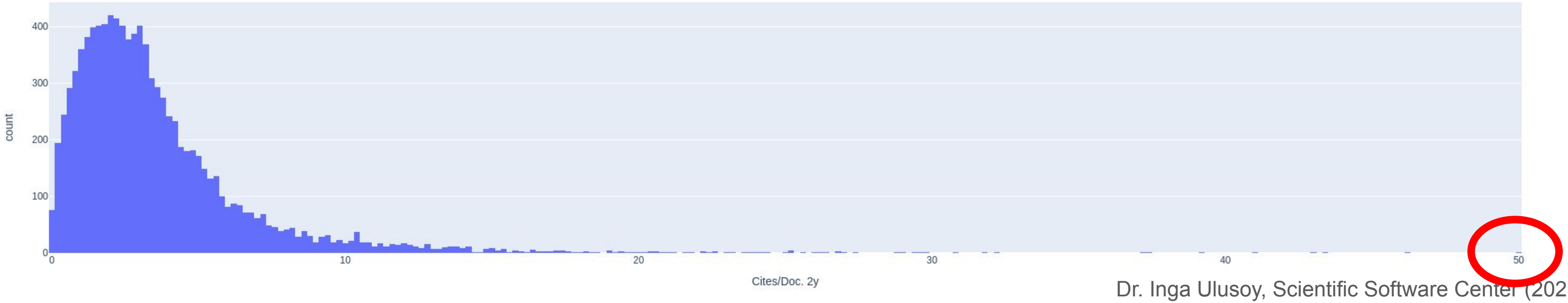


	Rank	OA	SJR-index	CiteScore	H-index	Best Subject Rank	Total Docs.	Total Docs. 3y	Total Refs.	Total Cites 3y	Citable Docs. 3y	Cites/Doc. 2y
Rank	1.000000	0.111300	-0.503617	-0.485568	-0.625403	0.558208	-0.192069	-0.196795	-0.196338	-0.243070	-0.185484	-0.560625
OA	0.111300	1.000000	-0.069304	-0.056997	-0.178146	0.114037	0.061870	0.046403	0.058683	0.024084	0.048485	-0.045120
SJR-index	-0.503617	-0.069304	1.000000	0.878000	0.565015	-0.281225	0.091092	0.102424	0.094227	0.270083	0.081086	0.828618
CiteScore	-0.485568	-0.056997	0.878000	1.000000	0.527957	-0.279983	0.112000	0.127705	0.122350	0.285965	0.110357	0.943584
H-index	-0.625403	-0.178146	0.565015	0.527957	1.000000	-0.362788	0.331053	0.393130	0.313698	0.505095	0.362266	0.512423
Best Subject Rank	0.558208	0.114037	-0.281225	-0.279983	-0.362788	1.000000	-0.114754	-0.117089	-0.132615	-0.150247	-0.118463	-0.334142
Total Docs.	-0.192069	0.061870	0.091092	0.112000	0.331053	-0.114754	1.000000	0.934468	0.968011	0.806830	0.932626	0.150987
Total Docs. 3y	-0.196795	0.046403	0.102424	0.127705	0.393130	-0.117089	0.934468	1.000000	0.887417	0.854647	0.995085	0.148272
Total Refs.	-0.196338	0.058683	0.094227	0.122350	0.313698	-0.132615	0.968011	0.887417	1.000000	0.802696	0.893789	0.173401
Total Cites 3y	-0.243070	0.024084	0.270083	0.285965	0.505095	-0.150247	0.806830	0.854647	0.802696	1.000000	0.844114	0.308644
Citable Docs. 3y	-0.185484	0.048485	0.081086	0.110357	0.362266	-0.118463	0.932626	0.995085	0.893789	0.844114	1.000000	0.139525
Cites/Doc. 2y	-0.560625	-0.045120	0.828618	0.943584	0.512423	-0.334142	0.150987	0.148272	0.173401	0.308644	0.139525	1.000000
Refs./Doc.	-0.390894	-0.064572	0.267383	0.306913	0.247259	-0.299281	0.032381	0.019822	0.109949	0.076826	0.030626	0.382891
Life Sciences	-0.166150	0.073645	0.071380	0.125088	0.210414	-0.183625	0.044939	0.050866	0.068018	0.051387	0.051948	0.114416

Data preparation

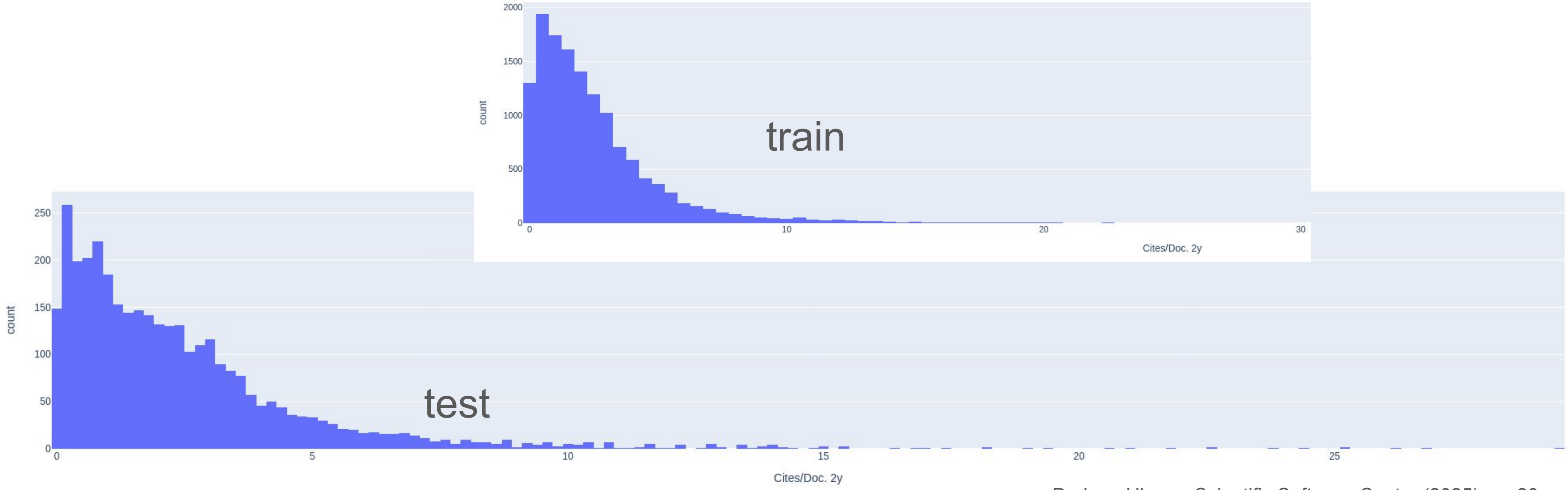
- Feature Engineering: Normalize features.
  - Features should have similar data ranges for the weights to be in similar ranges, and improved model robustness and faster training.





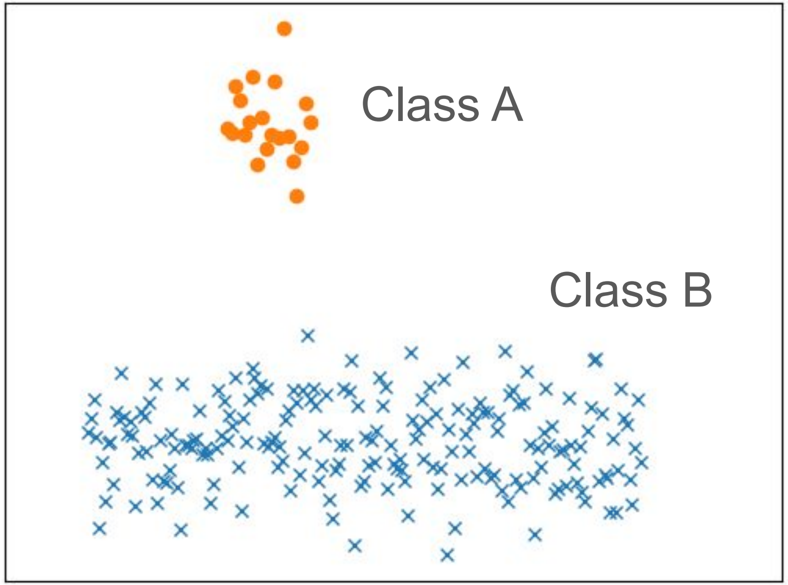
Data preparation

- Make sure to randomize your data.
    - Otherwise, your train and test data could contain more/less data of a certain kind (inhomogeneous data)



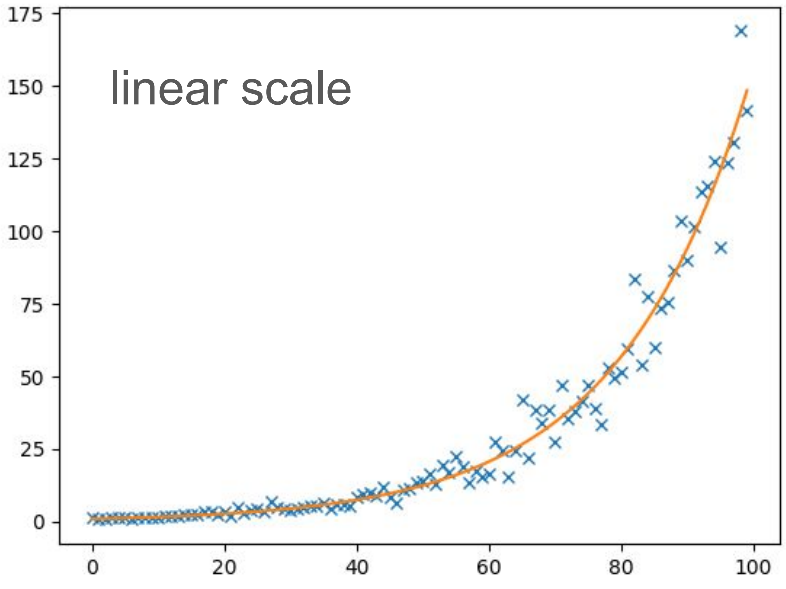
Data preparation

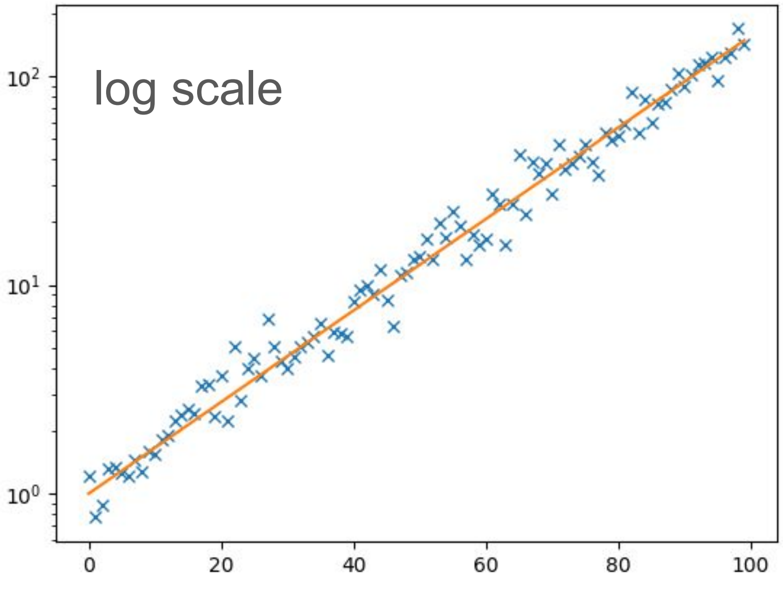
- Feature engineering: Make sure your dataset is balanced.
  - For classification tasks, all classes should have comparable sizes (similar numbers of examples).



Data preparation

- Feature engineering: Pick the right scale.
    - Visualize your data to see if you need to transform ie. onto a log scale.





4. Modeling of Research Data

Deep learning – Tools and Tricks

I wish I had known earlier

Peter Lippmann

13.02.2025

Scientific AI Lab, IWR, Heidelberg University

Deep learning overview

$$\min_{\theta} \; \mathbb{E}_{(x,y)\sim\mathcal{D}} \; L(y, f_\theta(x))$$

$\mathcal{D}$: data distribution, typically approximated by a set of finite set of input-target pairs $\{x_i, y_i\}$

f_θ: neural network with learnable parameters θ

L: differentiable loss function, typically minimal if $y = f_\theta(x)$

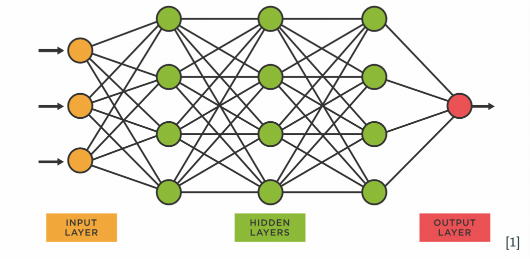
In practice:

$$\min_{\theta} \sum_{\substack{\text{sample } (x_i,y_i)\\ \text{in training set}}} L(y_i, f_\theta(x_i))$$

Optimization via (mini batch) gradient descent:

$$\theta^{(t+1)} = \theta^{(t)} \quad - \quad \alpha \sum_{\substack{\text{sample } (x_i,y_i)\\ \text{in mini batch}}} \nabla_\theta L(y_i, f_\theta(x_i))\Big)$$

The neural network



[1]

- Network types: CNNs for images, GNNs for graphs, sequence models for language, recurrent NNs for time series data, ... (often clear what to choose)
- BUT many representatives, e.g. many different CNN architectures (less clear)
- For given architecture, several hyperparameters (educated guess + trying out)

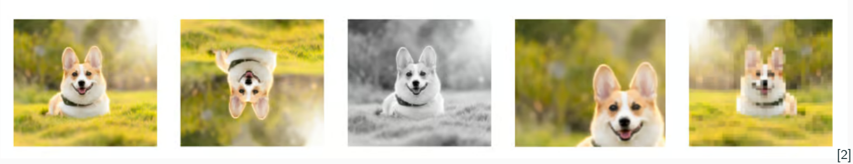
The loss function

$$\min_{\theta} \sum_{\substack{\text{sample } (x_i, y_i) \\ \text{in training set}}} L(y_i, f_\theta(x_i))$$

- heavily depends on your task, e.g.:
  - classification ↔ Cross Entropy loss,
  - regression ↔ MSE (a.k.a. L2-loss) $\|y_i - f_\theta(x_i)\|^2$
- Tip: for regression L1-loss $|y_i - f_\theta(x_i)|$ can be more robust to outliers
- sometimes a combination of losses is used (weighting them can be tricky)

The data

- more data is better, higher quality data is better
- visualize your data before: PCA, UMAP, t-SNE
- check your data is balanced (e.g. in instances per class)
- split your data into train, validation and test set
- standardize your data (both input and output)
- use data augmentation if possible



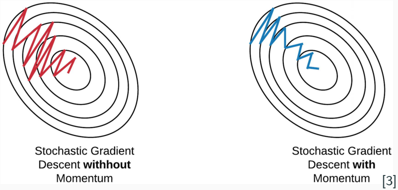
[2]

Data augmentation

The optimizer

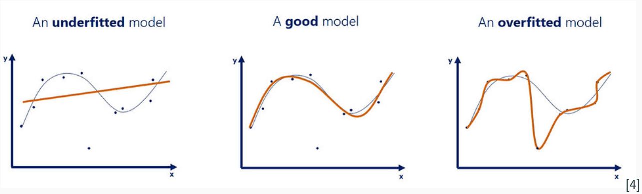
Use AdamW (Adaptive Moment Estimation + weight decay):

- adaptive learning rate helps against too small or too large gradients
- momentum stabalizes the gradient descent
- weight decay helps against overfitting


Stochastic Gradient Descent withhout Momentum
Stochastic Gradient Descent with Momentum

[3]

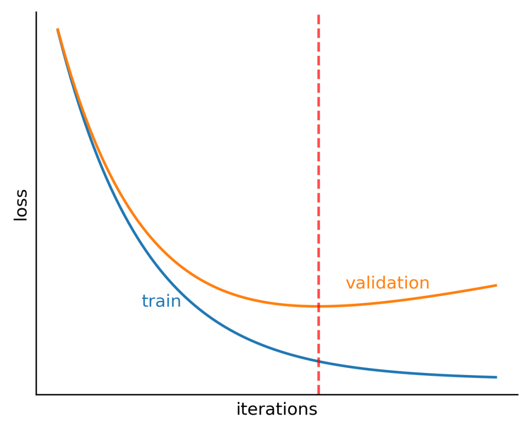
(a) Effect of momentum



[4]

(b) Overfitting model

More tricks against overfitting

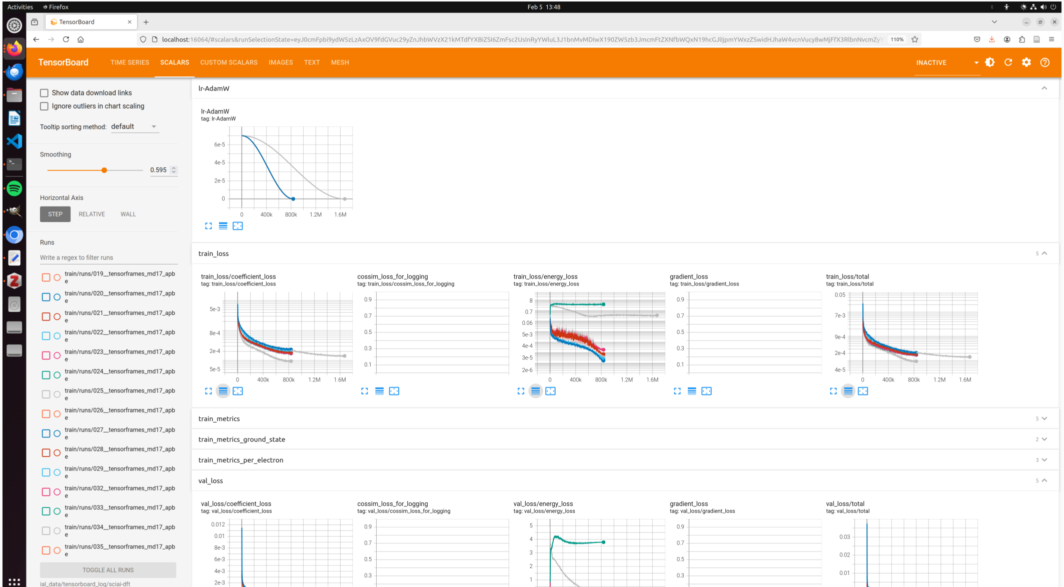


- Dropout: randomly drop/ignore neurons during training
- Save checkpoints of your model: last (if training crashes) & k-many best

More Training Tricks

- try to overfit on a single sample to debug your pipeline
- set a seed during training for reproducibility
- use enough CPU workers in dataloader to properly use GPU
- in dataloader use `shuffle=True` and `drop_last=True`
- try gradient clipping in the optimizer against instable training
- use a learning rate scheduler (e.g. cosine schedule)
- try learning rate warmup
- definitely try normalization layers: helps to standardize activations

Logging via Tensorboard



localhost:16064/#scalars&runSelectionState=eyJ0cmFpbi9ydW5zLzAxOV9fdGVuc29yZnJhbWVzX21kMTdfYXBiZSI6ZmFsc2UsInRyYWluL3J1bnMvMDIwX190ZW5zb3JmcmFtZXNfbWQxN19hcGJlIjpmYWxzZSwidHJhaW4vcnVucy8wMjFfX3RlbnNvcmZyY

□ Show data download links

□ Ignore outliers in chart scaling

Tooltip sorting method: default

Smoothing 0.595

Horizontal Axis: STEP RELATIVE WALL

Runs

Write a regex to filter runs

□ ○ train/runs/019__tensorframes_md17_apbe

□ ○ train/runs/020__tensorframes_md17_apbe

□ ○ train/runs/021__tensorframes_md17_apbe

□ ○ train/runs/022__tensorframes_md17_apbe

□ ○ train/runs/023__tensorframes_md17_apbe

□ ○ train/runs/024__tensorframes_md17_apbe

□ ○ train/runs/025__tensorframes_md17_apbe

□ ○ train/runs/026__tensorframes_md17_apbe

□ ○ train/runs/027__tensorframes_md17_apbe

□ ○ train/runs/028__tensorframes_md17_apbe

□ ○ train/runs/029__tensorframes_md17_apbe

□ ○ train/runs/032__tensorframes_md17_apbe

□ ○ train/runs/033__tensorframes_md17_apbe

□ ○ train/runs/034__tensorframes_md17_apbe

□ ○ train/runs/035__tensorframes_md17_apbe

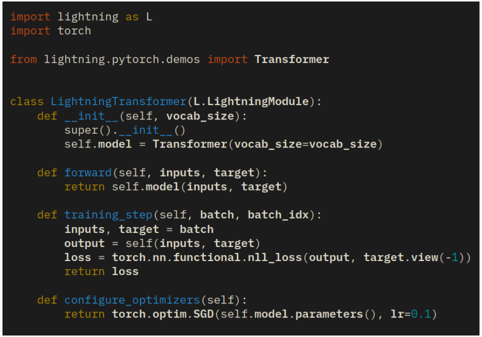
TOGGLE ALL RUNS

ial_data/tensorboard_log/sciai-dft

Pytorch Lightning

Pytorch lightning module combines pytorch model + optimizer + logging

- abstracts away `to("cuda")`, `loss.backward()`, `model.eval()` and much more
- makes complicated things which many people use easy, e.g. multi GPU support

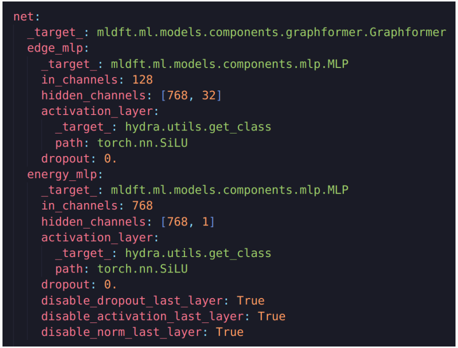


```python
import lightning as L
import torch

from lightning.pytorch.demos import Transformer


class LightningTransformer(L.LightningModule):
    def __init__(self, vocab_size):
        super().__init__()
        self.model = Transformer(vocab_size=vocab_size)

    def forward(self, inputs, target):
        return self.model(inputs, target)

    def training_step(self, batch, batch_idx):
        inputs, target = batch
        output = self(inputs, target)
        loss = torch.nn.functional.nll_loss(output, target.view(-1))
        return loss

    def configure_optimizers(self):
        return torch.optim.SGD(self.model.parameters(), lr=0.1)
```

Many great tutorials at
`https://lightning.ai/docs/pytorch/stable/starter/introduction.html`

Config management with Hydra

- save on boilerplate by "programming" in configs (customize models in config not in code)



```yaml
net:
  _target_: mldft.ml.models.components.graphformer.Graphformer
  edge_mlp:
    _target_: mldft.ml.models.components.mlp.MLP
    in_channels: 128
    hidden_channels: [768, 32]
    activation_layer:
      _target_: hydra.utils.get_class
      path: torch.nn.SiLU
    dropout: 0.
  energy_mlp:
    _target_: mldft.ml.models.components.mlp.MLP
    in_channels: 768
    hidden_channels: [768, 1]
    activation_layer:
      _target_: hydra.utils.get_class
      path: torch.nn.SiLU
    dropout: 0.
    disable_dropout_last_layer: True
    disable_activation_last_layer: True
    disable_norm_last_layer: True
```

- easy to add new models, datasets, tasks and experiments
- uses OmegaConf for configuration management

Great Hydra + Lightning template at `https://github.com/ashleve/lightning-hydra-template`

Thank You!

Any Questions?

Image References

[1] https://www.marktechpost.com/wp-content/uploads/2022/09/Screen-Shot-2022-09-23-at-10.46.58-PM.png

[2] https://media.datacamp.com/legacy/image/upload/v1669203370/Data_Augmentation_Header_f42227f2cb.png

[3] https://i.sstatic.net/epW89.jpg

[4] https://static.wixstatic.com/media/0ed3e8_a9b7d6d3dc6b4d5cbcb30c8b2fd4782b~mv2.jpg/v1/fill/w_1000,h_449,al_c,q_90,usm_0.66_1.00_0.01/0ed3e8_a9b7d6d3dc6b4d5cbcb30c8b2fd4782b~mv2.jpg

Model Metadata

Model card

- Model details
    - Architecture, parameters, citation information, license information
- Intended use
    - Use cases within the model’s scope
- Performance metrics
    - Intended performance on given data
- Training data
    - Description of training data and data distribution
- Quantitative analysis
    - Potential biases and limitations
- Ethical consideration
    - Privacy and fairness concerns, impact on society
- https://huggingface.co/spaces/huggingface/Model_Cards_Writing_Tool
- https://github.com/openai/gpt-3/blob/master/model-card.md

Where to share/publish/deploy your models

Model sharing platforms

You can make models available for others on model sharing platforms like

- Hugging face,
- OpenML,
- Kaggle.

Advantages: Public platform with version control and model cards, you can link the data into the repo, allows others to use your model for production or fine-tuning.

How to test your software that is based on ML models

Testing of non-deterministic processes

Try to make processes deterministic
For example: Use specified random seed.

Separate deterministic and non-deterministic processes and test separately
For example: Input processing can be tested separately from model prediction.

Test for output parameters and properties that remain constant
For example: Number of predictions, feature length, etc.

Include multiple valid outputs in your tests
For example: Three most probable classifications.

Robustness: A robust model is more likely to behave like a deterministic system
Make sure your model output is stable under a range of conditions.

Accuracy: The model accuracy will affect the testing strategy
A higher accuracy leads to more consistent predictions.

Distribution: You can also test for the distribution of results rather than the results themselves

Deploying machine-learning models and software

Model deployment

In addition to making models and software available for others to use in their own code, you can also directly deploy the model - together with your code - directly so that it can be used.

Examples:

- Diffusers: google colab https://colab.research.google.com/github/huggingface/notebooks/blob/main/diffusers/stable_diffusion.ipynb
- https://lightning.ai/ for paid service and deployable models
- See https://www.freecodecamp.org/news/deploy-your-machine-learning-models-for-free/ for tutorials and services

5. Software Engineering best practices

Version Control: git

What is this?

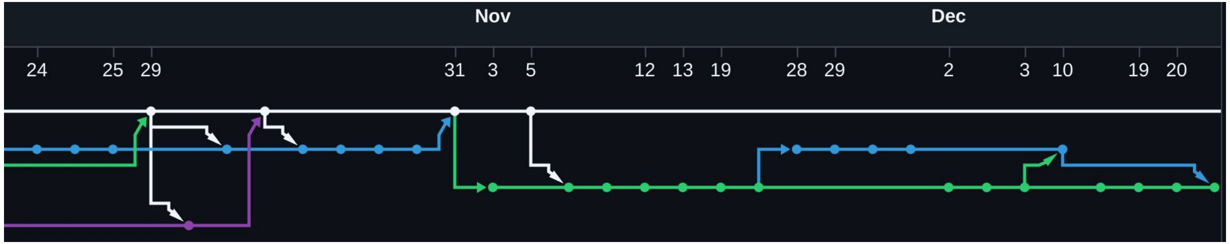
A tool to allow you to track and revert changes, and collaborate with others (change management).

Why is this important?

- Allow versioning of the code and continuing functionality.
- Allow simultaneous changes to the same files.
- Fundamental for reproducing historic states in the line of development.
- Development follows a story and allows other users to build confidence in your work.

Version Control in practice: git

- Create a repo on GitHub
- Clone the repo to your local machine
- Checkout a branch and make changes
- `git add`, `git commit` and `git push`: IDEs such as VSCode make it easy for you!
- Observe how your repo changes on GitHub
- GitHub offers great learning labs: GitHub skills https://skills.github.com/
- roadmap.sh offers a git roadmap: https://roadmap.sh/git-github



Development workflows: GitHub-flow

What is this?

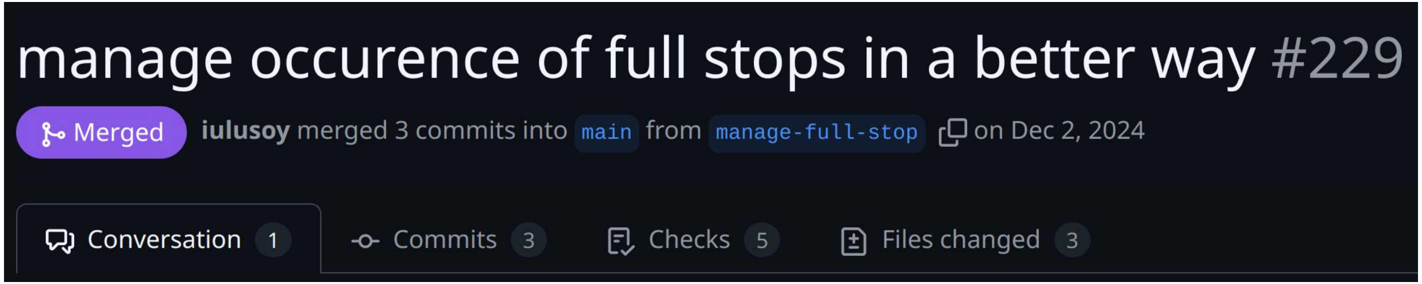
GitHub-flow is a lightweight workflow with creating branches, making changes, opening Pull Requests, running Continuous Integration, and requesting code review, together with Issues and Kanban project boards.

Why is this important?

- Manage how changes are incorporated in the software.
- Track progress in your project and highlight bottlenecks.
- Adhere to development guidelines, ensuring the implementation follows defined rules to ensure software quality.

Development workflows in practice: GitHub-flow

- After you made changes in your git branch, open a Pull Request on GitHub
- Observe how the PR highlights the changes in the line of development
- You can link issues, comment on the PR and run automated checks
- See GitHub skills https://skills.github.com/
- See roadmap.sh https://roadmap.sh/git-github



manage occurence of full stops in a better way #229

Merged iulusoy merged 3 commits into `main` from `manage-full-stop` on Dec 2, 2024

Requirements engineering and continuous delivery

What is this?

Requirements engineering is the process of translating stakeholder requirements on the research software into defined tasks. Early delivery and iteration over it allows refinement of the requirements and tasks.

Why is this important?

- Ensure that the software fulfills its purpose. In research software, requirements engineering is closely intertwined with the research process and subject to frequent changes.
- Understand the problem the software should solve and map this onto an efficient technically feasible solution considering all constraints.
- Allows prioritization of requirements/tasks and decision-making. Decisions should be documented together with the requirements.

Requirements engineering in practice

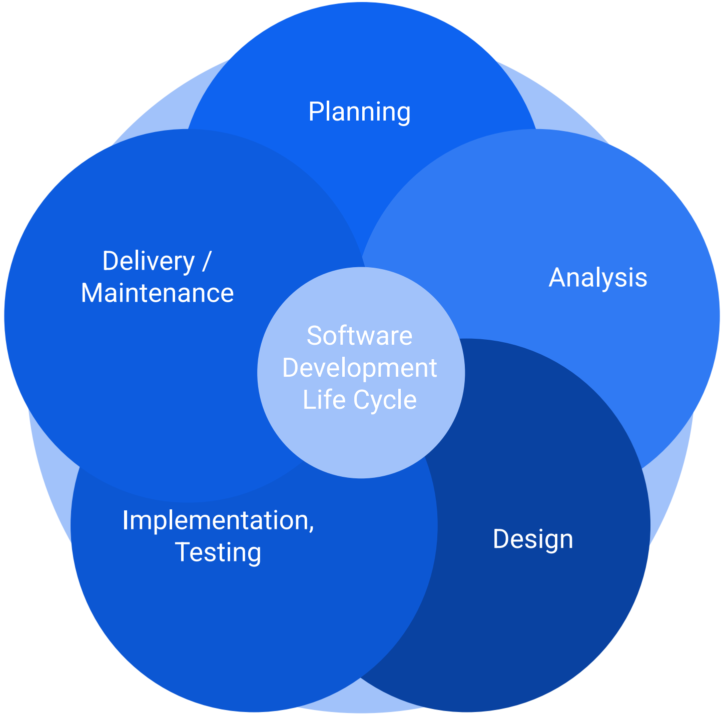
- Functional requirements (what the software should do: features)
- Non-functional requirements (how the software performs a task: ie performance, security)
- Domain requirements (specific to the domain: ie. Healthcare)
- Use tools such as draft.io or miro to gather requirements

As a <type of user>, I want <some goal> so that <some reason>.

As a <researcher>, I want <to obtain the probability of a class> so that <I can classify incoming images>.

Continuous delivery in practice

- Deliver early to find out if your software is fulfilling its purpose/moving in the right direction
- Use quality control to allow early delivery through the main branch in your GitHub-flow (*keep your main branch operational at all times*)
- Following agile / lean principles
- Use git to allow consistent usability of your software



Project management: Kanban boards

What is this?

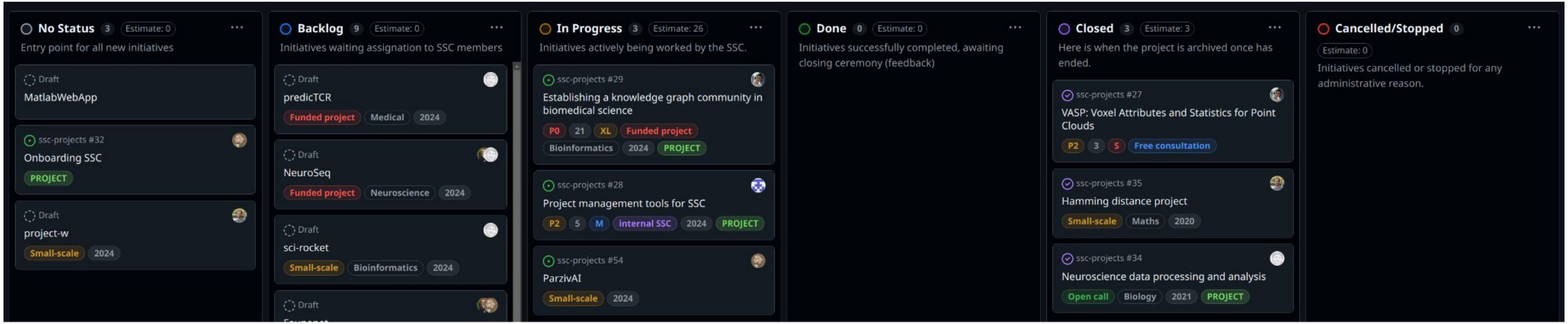
Project management is used to track progress, identify intertwined or dependent processes, and allows visual access to the project’s status.

Why is this important?

- Ensure that the software development moves in the right direction.
- Increase the flow of ongoing work.
- Allow prioritization of tasks and understand interdependencies and bottlenecks in the development.

Project management in practice

- Use a Kanban board to organize tasks
- Separate tasks into backlog/todo, in progress, done
- Prioritize tasks and assign contributors/identify necessities
- For example, GitHub projects



No Status 3 Estimate: 0

Entry point for all new initiatives

- Draft
  MatlabWebApp
- ssc-projects #32
  Onboarding SSC
  PROJECT
- Draft
  project-w
  Small-scale · 2024

Backlog 9 Estimate: 0

Initiatives waiting assignation to SSC members

- Draft
  predicTCR
  Funded project · Medical · 2024
- Draft
  NeuroSeq
  Funded project · Neuroscience · 2024
- Draft
  sci-rocket
  Small-scale · Bioinformatics · 2024
- Draft

In Progress 3 Estimate: 26

Initiatives actively being worked by the SSC.

- ssc-projects #29
  Establishing a knowledge graph community in biomedical science
  P0 · 21 · XL · Funded project · Bioinformatics · 2024 · PROJECT
- ssc-projects #28
  Project management tools for SSC
  P2 · 5 · M · internal SSC · 2024 · PROJECT
- ssc-projects #54
  ParzivAI
  Small-scale · 2024

Done 0 Estimate: 0

Initiatives successfully completed, awaiting closing ceremony (feedback)

Closed 3 Estimate: 3

Here is when the project is archived once has ended.

- ssc-projects #27
  VASP: Voxel Attributes and Statistics for Point Clouds
  P2 · 3 · S · Free consultation
- ssc-projects #35
  Hamming distance project
  Small-scale · Maths · 2020
- ssc-projects #34
  Neuroscience data processing and analysis
  Open call · Biology · 2021 · PROJECT

Cancelled/Stopped 0 Estimate: 0

Initiatives cancelled or stopped for any administrative reason.

Project planning: Architecture and design

What is this?

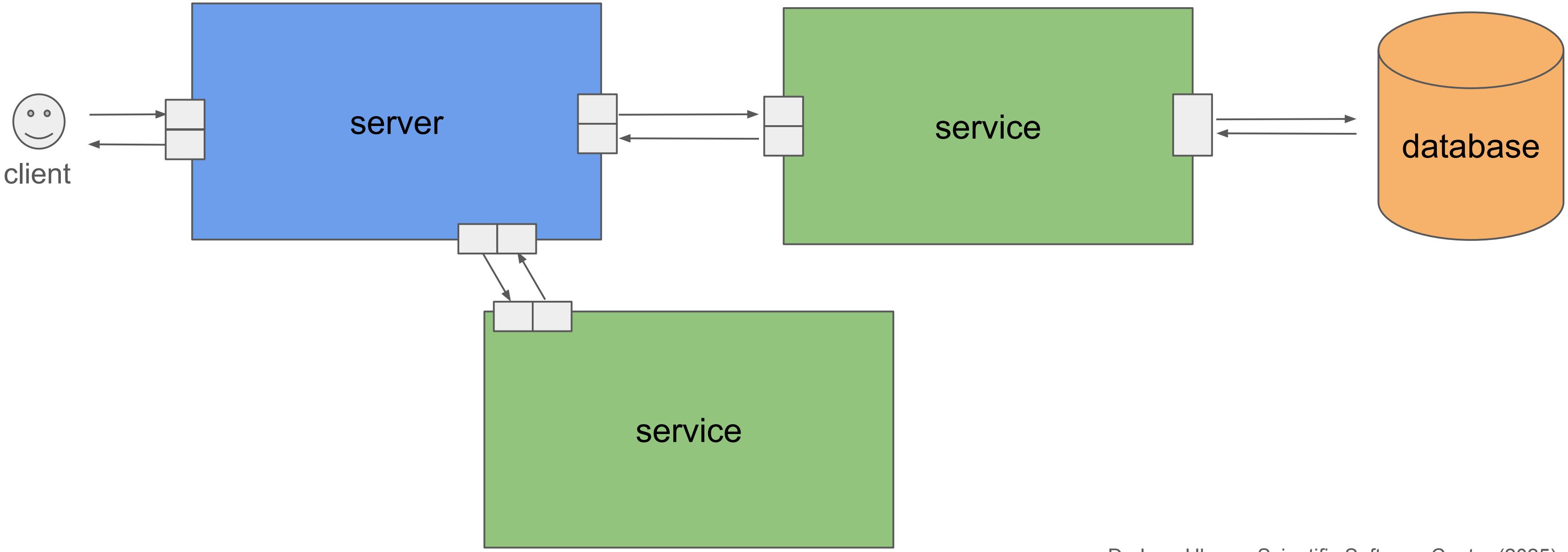
Software architecture describes how the system is composed of different pieces, and the interplay of the components. Design refers to the actual implementation of the requirements in the system as a whole and the different components.

Why is this important?

- Makes the software efficient and allow re-use of functionalities.
- Allows extensions and additions of features at a later stage without major refactoring.
- Makes the software maintainable.

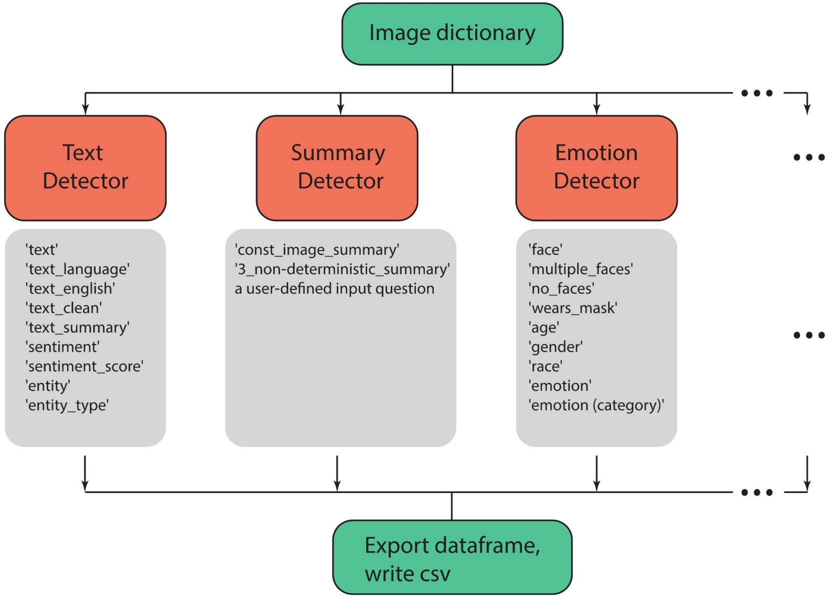
Architecture in practice

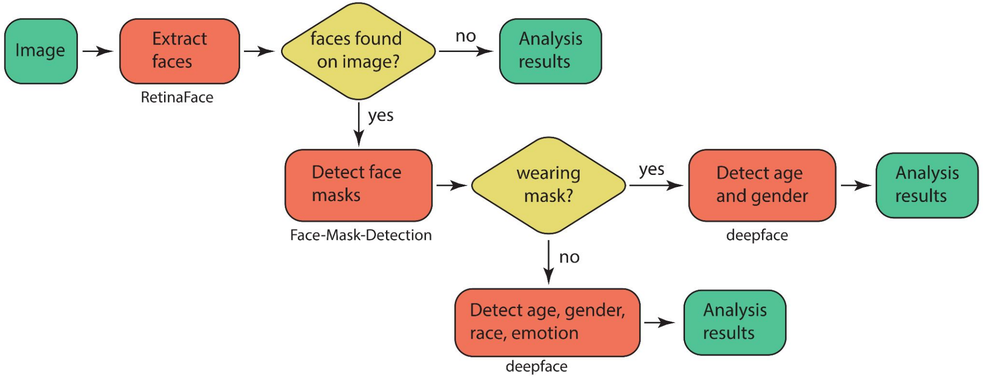
- Use (black/white) box diagrams to identify components and their interactions
- https://roadmap.sh/software-design-architecture / miro / draft.io / draw.io

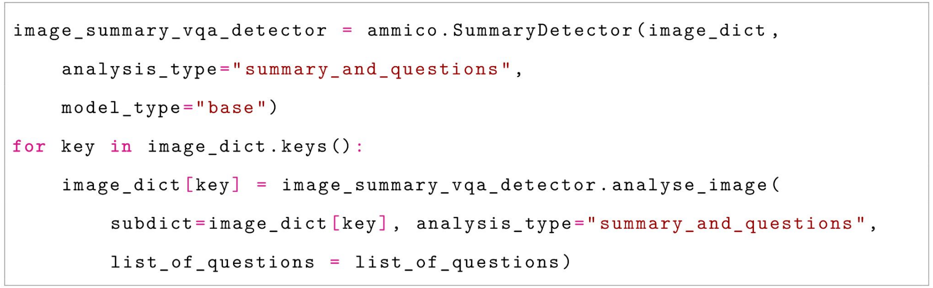


Design in practice

- Map the input/output, data formats, transformations / logic / processes
- https://roadmap.sh/software-design-architecture / miro / draft.io / draw.io







```
image_summary_vqa_detector = ammico.SummaryDetector(image_dict,
    analysis_type="summary_and_questions",
    model_type="base")
for key in image_dict.keys():
    image_dict[key] = image_summary_vqa_detector.analyse_image(
        subdict=image_dict[key], analysis_type="summary_and_questions",
        list_of_questions = list_of_questions)
```

Quality management: Testing and continuous Integration

What is this?

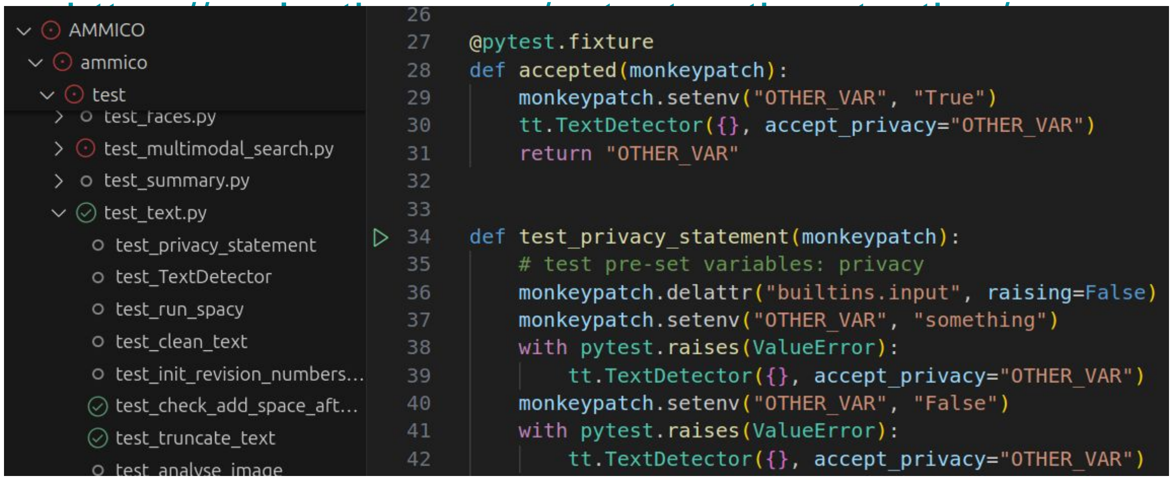
GitHub-flow is a lightweight workflow with creating branches, making changes, opening Pull Requests, running Continuous Integration, and requesting code review, together with Issues and Kanban project boards.

Why is this important?

- Manage how changes are incorporated in the software.
- Track progress in your project and highlight bottlenecks.
- Adhere to development guidelines, ensuring the implementation follows defined rules to ensure software quality.

Testing in practice

- Use testing frameworks such as `pytest`
- Write tests in a `tests/` folder: unit tests, integration tests, system tests, compatibility tests, …
- To learn how to use pytest: https://docs.pytest.org/en/stable/, https://realpython.com/pytest-python-testing/

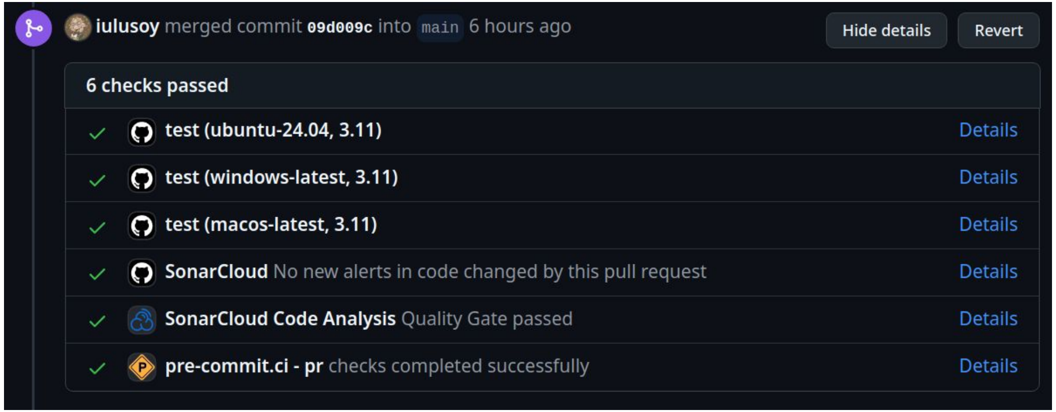


```
AMMICO
  ammico
    test
      test_faces.py
      test_multimodal_search.py
      test_summary.py
      test_text.py
        test_privacy_statement
        test_TextDetector
        test_run_spacy
        test_clean_text
        test_init_revision_numbers...
        test_check_add_space_aft...
        test_truncate_text
        test_analyse_image
```

```python

@pytest.fixture
def accepted(monkeypatch):
    monkeypatch.setenv("OTHER_VAR", "True")
    tt.TextDetector({}, accept_privacy="OTHER_VAR")
    return "OTHER_VAR"


def test_privacy_statement(monkeypatch):
    # test pre-set variables: privacy
    monkeypatch.delattr("builtins.input", raising=False)
    monkeypatch.setenv("OTHER_VAR", "something")
    with pytest.raises(ValueError):
        tt.TextDetector({}, accept_privacy="OTHER_VAR")
    monkeypatch.setenv("OTHER_VAR", "False")
    with pytest.raises(ValueError):
        tt.TextDetector({}, accept_privacy="OTHER_VAR")
```

Continuous integration in practice

- Set up your tests to be automatically run by GitHub actions
- Include code linter and quality control in your actions
- These should be set up to run automatically when you open a Pull Request
- GH actions, `codecov`, `sonarcloud`, `snyk`, `pre-commit`, code formatting (`black`), GitHub Guardian, `dependabot`



iulusoy merged commit 09d009c into main 6 hours ago

6 checks passed

✓ test (ubuntu-24.04, 3.11) Details

✓ test (windows-latest, 3.11) Details

✓ test (macos-latest, 3.11) Details

✓ SonarCloud No new alerts in code changed by this pull request Details

✓ SonarCloud Code Analysis Quality Gate passed Details

✓ pre-commit.ci - pr checks completed successfully Details

Software Management Plans

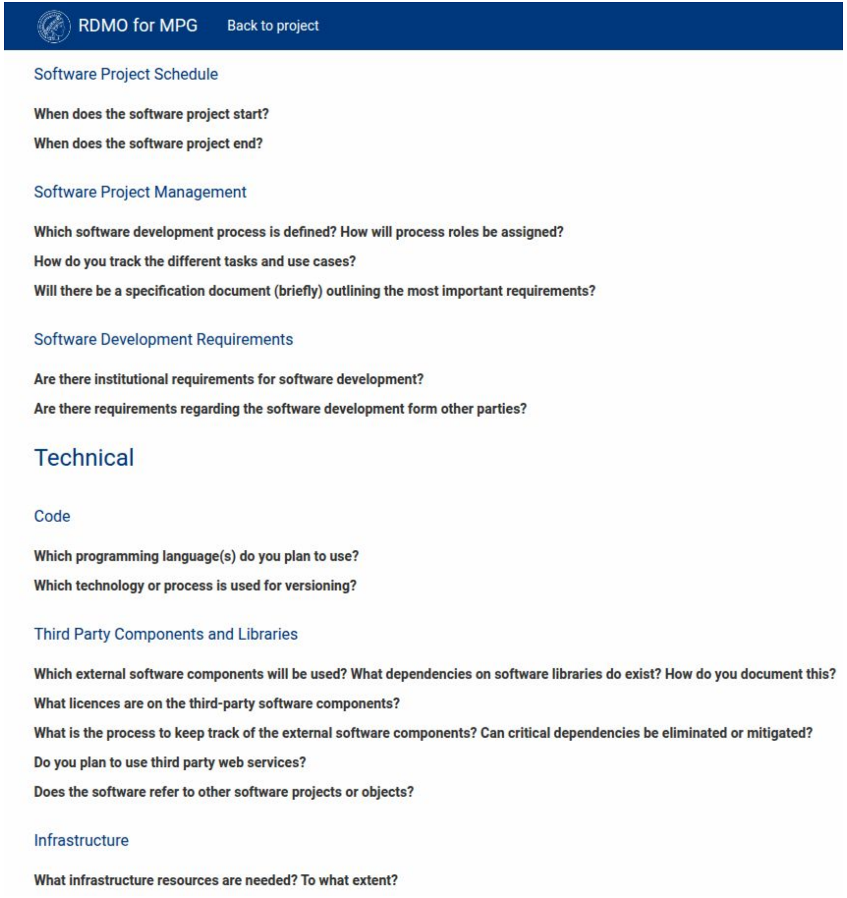
What is this?

Software Management Plans (SMPs) help to identify goals and the means required to pursue the goals in practice.

Why is this important?

- Identify criticality and required maturity of your software.
- Identify which measures are needed to ensure compliance of your software with the intended goals.
- Quantify milestones and tools for the intended purpose.

SMPs in practice



Software Project Schedule

When does the software project start?

When does the software project end?

Software Project Management

Which software development process is defined? How will process roles be assigned?

How do you track the different tasks and use cases?

Will there be a specification document (briefly) outlining the most important requirements?

Software Development Requirements

Are there institutional requirements for software development?

Are there requirements regarding the software development form other parties?

Technical

Code

Which programming language(s) do you plan to use?

Which technology or process is used for versioning?

Third Party Components and Libraries

Which external software components will be used? What dependencies on software libraries do exist? How do you document this?

What licences are on the third-party software components?

What is the process to keep track of the external software components? Can critical dependencies be eliminated or mitigated?

Do you plan to use third party web services?

Does the software refer to other software projects or objects?

Infrastructure

What infrastructure resources are needed? To what extent?

- Use the SMPs provided by the Max Planck digital library
- Helps you with your requirements and project management
- https://rdmo.mpdl.mpg.de/

Documentation

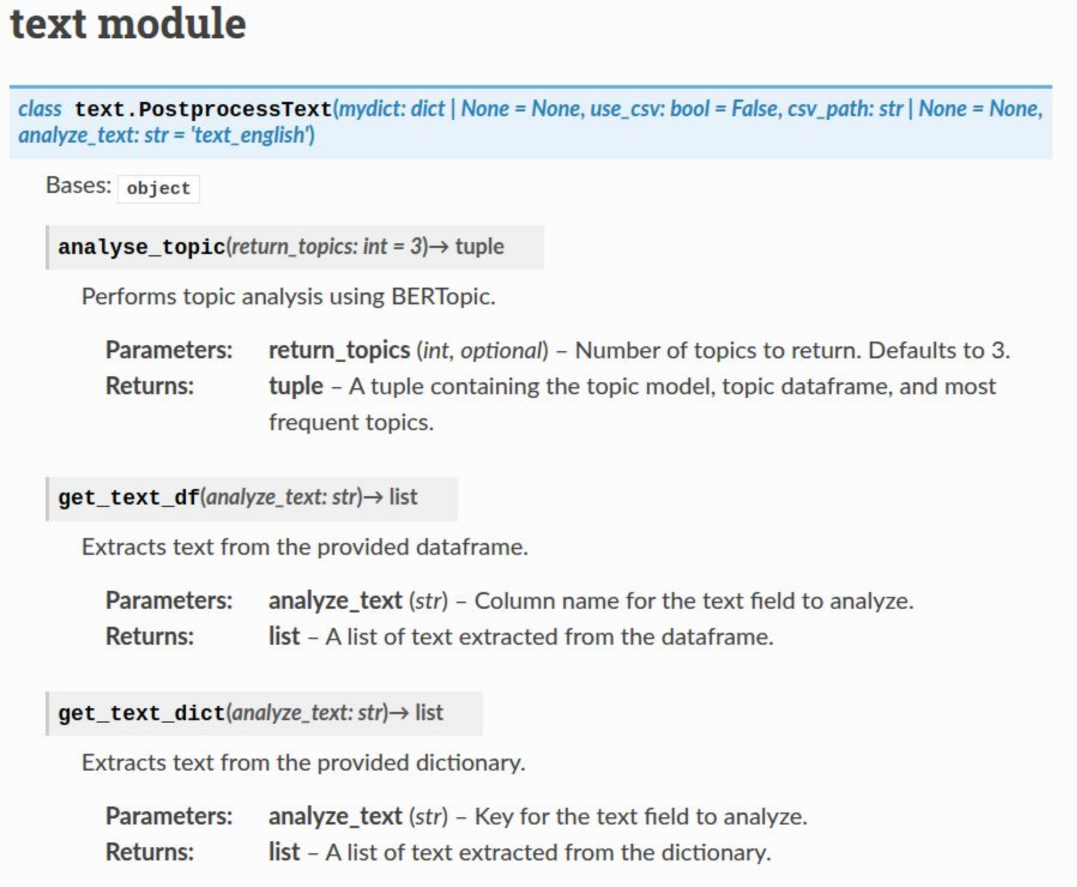
What is this?

Documentation can be comments, docstrings, readme's, tutorials, demonstration notebooks, and contains technical and domain-specific / application-specific descriptions of the software.

Why is this important?

- Document what the software can and cannot do, and parameter ranges.
- Allow others to install and use your software (or yourself, at a later time).
- Allow others to contribute to your software.

Documentation in practice



text module

class `text.PostprocessText`(*mydict: dict | None = None, use_csv: bool = False, csv_path: str | None = None, analyze_text: str = 'text_english'*)

Bases: `object`

`analyse_topic`(*return_topics: int = 3*)→ tuple

Performs topic analysis using BERTopic.

Parameters: **return_topics** (*int*, *optional*) – Number of topics to return. Defaults to 3.

Returns: **tuple** – A tuple containing the topic model, topic dataframe, and most frequent topics.

`get_text_df`(*analyze_text: str*)→ list

Extracts text from the provided dataframe.

Parameters: **analyze_text** (*str*) – Column name for the text field to analyze.

Returns: **list** – A list of text extracted from the dataframe.

`get_text_dict`(*analyze_text: str*)→ list

Extracts text from the provided dictionary.

Parameters: **analyze_text** (*str*) – Key for the text field to analyze.

Returns: **list** – A list of text extracted from the dictionary.

- Use tools like `sphinx` and `mkdocs` to render docstrings and markdown at minimal effort
- Include jupyter notebooks that showcase use of your software - these can be run on google colab
- Document dependencies in a requirements file and provide installation instructions

Deployment: Runtime environment / containerisation

What is this?

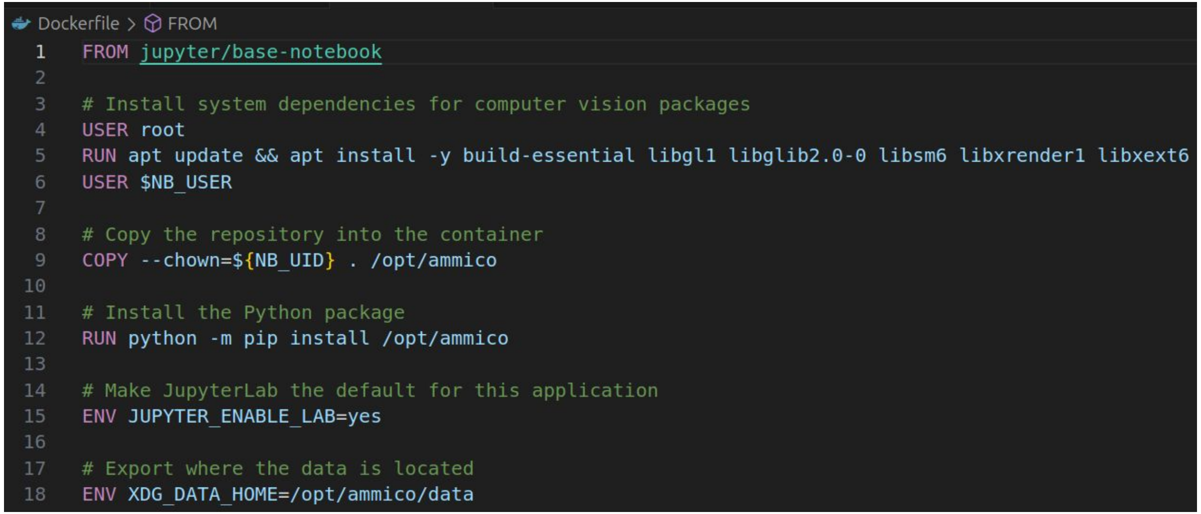
Deployment information such as runtime environments or containers allow easy adaption as they provide direct access to running the software without installation and dependency conflicts.

Why is this important?

- A big step towards reproducibility and transferability of your approach.
- The software ecosystem changes quickly, and this allows to preserve a snapshot that can be shared and run easily.

Containerisation in practice

- Use docker to provide build instructions for containers, and possibly deploy the containers on Dockerhub for anyone to download and use
- Docker roadmap https://roadmap.sh/docker, official tutorial https://docs.docker.com/



```
FROM jupyter/base-notebook

# Install system dependencies for computer vision packages
USER root
RUN apt update && apt install -y build-essential libgl1 libglib2.0-0 libsm6 libxrender1 libxext6
USER $NB_USER

# Copy the repository into the container
COPY --chown=${NB_UID} . /opt/ammico

# Install the Python package
RUN python -m pip install /opt/ammico

# Make JupyterLab the default for this application
ENV JUPYTER_ENABLE_LAB=yes

# Export where the data is located
ENV XDG_DATA_HOME=/opt/ammico/data
```

Software Licensing

What is this?

A software license states the terms of use, re-use and distribution, among others, without violating copyrights, and defines responsibilities.

Why is this important?

- So that others may use your code, and to prevent misuse.
- So that others may contribute to your code.
- So that the responsibilities for how the software is used are clear.
- Establishes the rights of all parties involved with the software.

Software licensing in practice

- Use the provided templates from GitHub: Either at repository creation or when adding a new file called LICENSE
- Permissive open-source license: BSD 2-Clause, MIT, Apache License 2.0
- Copyleft open-source license: GNU version 3, LGPL
- Proprietary licenses: Do not only keep your project close-source and potentially less visible, but also carry responsibilities for contract fulfillment
- https://opensource.org/licenses, https://choosealicense.com/
- **When using/incorporating third-party software**: ie. use of open-source libraries - is the third-party code distributed with your software? If so, compatibility needs to be confirmed!

6. Making your work public: Considerations of more general use and prominent failures

REFORMS checklist

To ensure reproducibility, include a checklist such as the REFORMS checklist into the publication of your results.

There are a couple of domain-specific checklists, especially in health and life sciences (STARD, CLAIM, see the EQUATOR* network), that may be more appropriate - REFORMS aims to be most general.

REFORMS checklist: DOI 10.48550/arXiv.2308.07832, https://reforms.cs.princeton.edu/

*EQUATOR Network, https://www.equator-network.org/reporting-guidelines/

REFORMS: Example

1. **Study goals**

1a) Population or distribution about which the scientific claim is made

The group to which the claim will be generalized

REFORMS: Example

1. **Study goals**

1b) Motivation for choosing this population or distribution

Choice of a particular population of interest - pure scientific interest, need for applied knowledge, …

REFORMS: Example

1. **Study goals**

1c) Motivation for the use of ML methods in the study

Why focus is on building a model that reliably maps input data to output data, instead of using traditional statistical methods?

REFORMS: Example

2. Computational reproducibility

2a) Dataset

Cite datasets with permanent links to clarify which version; if contains sensitive data: release synthetic datasets (`synthpop` library)

REFORMS: Example

2. Computational reproducibility

2b) Code

Cite exact version of code (DOI, version number, commit tag)

REFORMS: Example

2. Computational reproducibility

2c) Computing environment

Details about the hardware (CPU, RAM, disk space), software (operating system, programming language, version number for each package used), and computing resources (time taken to generate the results)

REFORMS: Example

2. Computational reproducibility

2d) Documentation

Document installation, requirements, running the code, usage examples, expected results

REFORMS: Example

2. Computational reproducibility

2e) Reproduction script

Reproduction scripts that prepare the environment, install the code, download datasets, and run code to reproduce the results in the paper

REFORMS: Example

3. Data quality

3a) Data source(s)

When, where, how data were collected, how ground-truth annotations were performed

REFORMS: Example

3. **Data quality**

3b) Sampling frame

List of people or units from which a sample is drawn

REFORMS: Example

3. Data quality

3c) Justification for why the dataset is useful for the modeling task

Why the dataset is a good representative to model the question/answer

REFORMS: Example

3. Data quality

3d) Outcome variable

How outcome variable is defined - this is usually not a perfect match for what it should represent

REFORMS: Example

3. **Data quality**

3e) Number of samples in the dataset

Total sample size, number of samples in each class, number of individual data in the dataset

REFORMS: Example

3. Data quality

3f) Missingness

Missing data

REFORMS: Example

3. **Data quality**

3g) Dataset for evaluation is representative

Justify why data is representative for target selected in 1a)

REFORMS: Example

4. Data preprocessing

4a) Excluded data and rationale

Justify why particular subset of data was chosen

REFORMS: Example

4. Data preprocessing

4b) How impossible or corrupt samples are dealt with

How erroneous or impossible data points are identified and amended

REFORMS: Example

4. Data preprocessing

4c) Data transformations

Normalizing, augmenting, imputing missing data, oversampling - the latter two must be done separately on each fold of the dataset!

REFORMS: Example

5. **Modeling**

5a) Model description

Input, output, type of model, loss function, algorithm

REFORMS: Example

5. Modeling

5b) Justification for the choice of model types implemented

Model needs to be interpretable, types of models considered

REFORMS: Example

5. **Modeling**

5c) Model evaluation method

Model needs to be evaluated on test data different from training data

REFORMS: Example

5. Modeling

5d) Model selection method

How final model(s) and hyperparameters were selected

REFORMS: Example

5. Modeling

5e) Hyperparameter selection

How hyperparameters were optimized

REFORMS: Example

5. Modeling

5f) Appropriate baselines

How baseline models were trained and optimized

REFORMS: Example

6. Data leakage

6a) Train-test separation is maintained

Use a hold-out test set that is also not used in synthetic data generation

REFORMS: Example

6. **Data leakage**

6b) Dependencies or duplicates between datasets

Could be more than one sample from same origin (patient); or time series is split randomly

REFORMS: Example

6. Data leakage

6c) Feature legitimacy

Any feature in the training that somewhat identifies predicted variable

REFORMS: Example

7. Metrics and uncertainty quantification

7a) Performance metrics used

Proper choice of metric depending on application

See Leist et al., Sci. Adv. 8, eabk1942 (2022) Table 4 for appropriate metrics depending on the data and algorithm used

REFORMS: Example

7. Metrics and uncertainty quantification

7b) Uncertainty estimates

Randomness in training or evaluation data, or training process

REFORMS: Example

7. Metrics and uncertainty quantification

7c) Appropriate statistical tests

Statistical testing of ML models

For an overview, see Sebastian Raschka, Model evaluation, model selection, and algorithm selection in machine learning, arXiv:1811.12808

REFORMS: Example

8. Generalizability and limitations

8a) Evidence of external validity

Report how claim is generalizable to target population.

REFORMS: Example

8. Generalizability and limitations

8b) Contexts in which the study’s findings will not hold

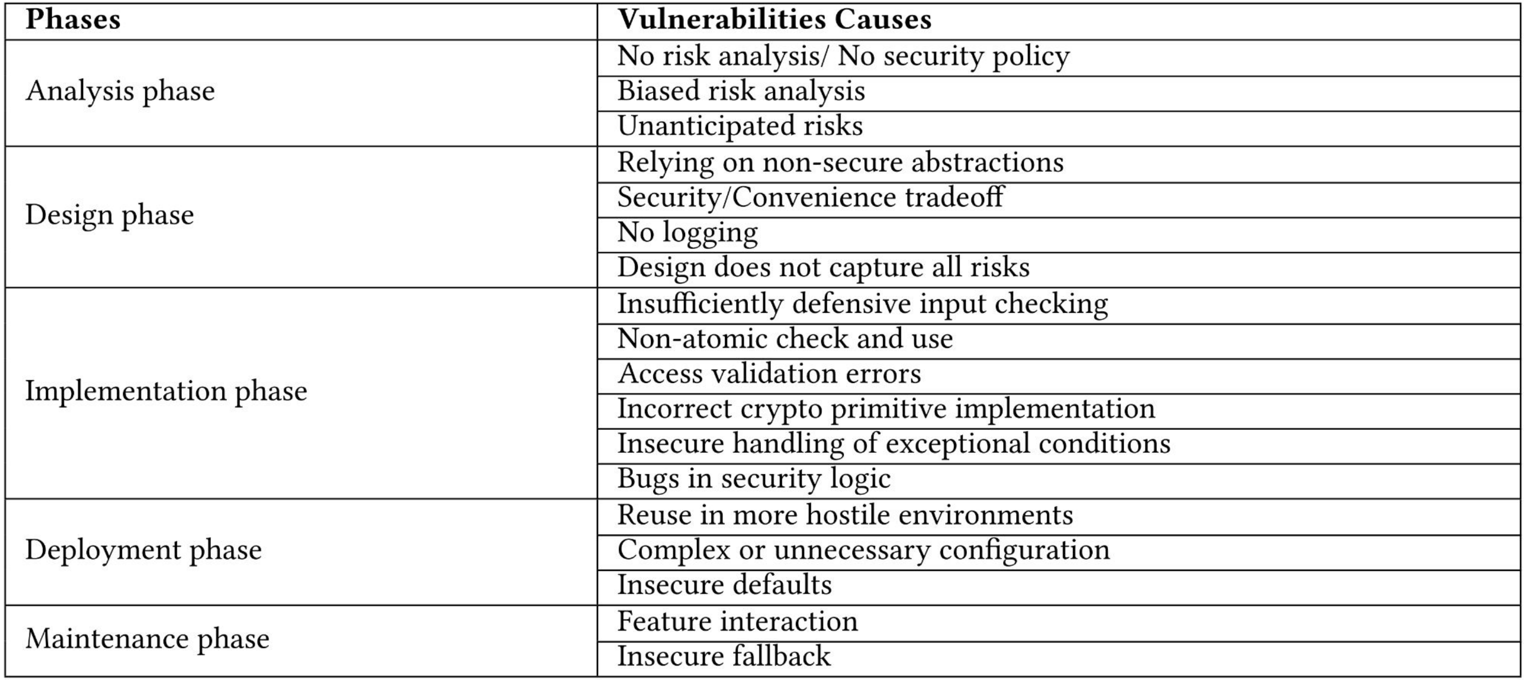
Clear expectations and unjustified hype.

Software security

Security

- Threat modelling: who, what, how
- Data-oriented attack: Access training data, poison data, inject trojan data
- Model-oriented attack: Modify training process (pre-trained malicious models), manipulate the deployed model (model patches, privacy information leakage, model inversion attacks)
- System-oriented attack: specialised hardware accelerators for ML software (SOC, trojan in GPU/TPU, for model corruption, backdoor insertion, model extraction, spoofing, information extraction, sybil attack)
- Possibility to carry out *pentests*

Security: best practices



Phases	Vulnerabilities Causes
Analysis phase	No risk analysis/ No security policy
	Biased risk analysis
	Unanticipated risks
Design phase	Relying on non-secure abstractions
	Security/Convenience tradeoff
	No logging
	Design does not capture all risks
Implementation phase	Insufficiently defensive input checking
	Non-atomic check and use
	Access validation errors
	Incorrect crypto primitive implementation
	Insecure handling of exceptional conditions
	Bugs in security logic
Deployment phase	Reuse in more hostile environments
	Complex or unnecessary configuration
	Insecure defaults
Maintenance phase	Feature interaction
	Insecure fallback

Chen, Barbar, Security for Machine Learning-based Software Systems, DOI 10.1145/3638531

Legal aspects

Legal aspects

- If you reuse data / models / code: Make sure the license terms allow this and that your license(s) is (are) compatible
- Make sure you do not violate the DSGVO / GDPR / European Data Act / Copyright / European AI Act
- Once a model is made available, it is impossible to restrict its use!
- Examples:
  - Models can put out near-exact copies of images/text in training data, ie Dall-E/Stable Diffusion generating images with Shutterstock/Getty Images watermarks, or reproducing artist’s work
  - Code-generating tools such as GitHub copilot allow recreation of code, that is already contained in other software, regardless of the license terms of that software

Ethical aspects

Ethical aspects

- Be aware of people's tendency of overreliance!
  - ELIZA effect: the tendency of users to project human traits onto interactive software
  - Trusting into predictions above one's own assessment
- Misuse of AI by bad actors
  - Face recognition used to detect Uyghur population (China), Clearview used to track people's movement and employment status by police officers (even though use was prohibited)
  - Facial analysis used to track people's attention on billboards, change advertisement based on their demographic
- Source of truth during training
  - Data that foundation models are trained on does also contain false statements, ie law professor incorrectly accused of sexual harassment
  - Data in foundation models contains toxicity of the internet; human labor is used to label / annotate toxic text and images ie subcontracting workers in Kenya (OpenAI)

Infamous AI mistakes

AI in general

- **Automatization at amazon:** Experimental hiring tool, developed by a team of five, used artificial intelligence to give job candidates scores ranging from one to five stars
- Tool was trained on all resumes of the last 10 years
- Tool preferably suggested male candidates, penalizing resumes that contained the word “women’s” or graduates from all-female colleges
- ***Why?***

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/ (2018)

Automated resume selection at amazon

- The dataset consisted of the resumes of the last 10 years: Predominantly male applicants
- Women are underrepresented in tech:

  https://fingfx.thomsonreuters.com/gfx/rngs/AMAZON.COM-JOBS-AUTOMATION/010080Q91F6/index.html
- Thus, the model learnt it was more often correct if it suggested male candidates: Unbalanced representation of the dataset

AI in general

- U.S. healthcare system uses commercial algorithms to guide health decisions
- Algorithm (Optum’s Impact Pro) to target patients for “high-risk care management” programs
- Identify patients who benefit the most: https://www.science.org/doi/10.1126/science.aax2342
- Model is trained on healthcare spendings to determine the healthcare need
- Algorithm was much more likely to recommend white patients for these programs than black patients, even though the black patients were evidently sicker
- ***Why?***

Bias in healthcare need estimation

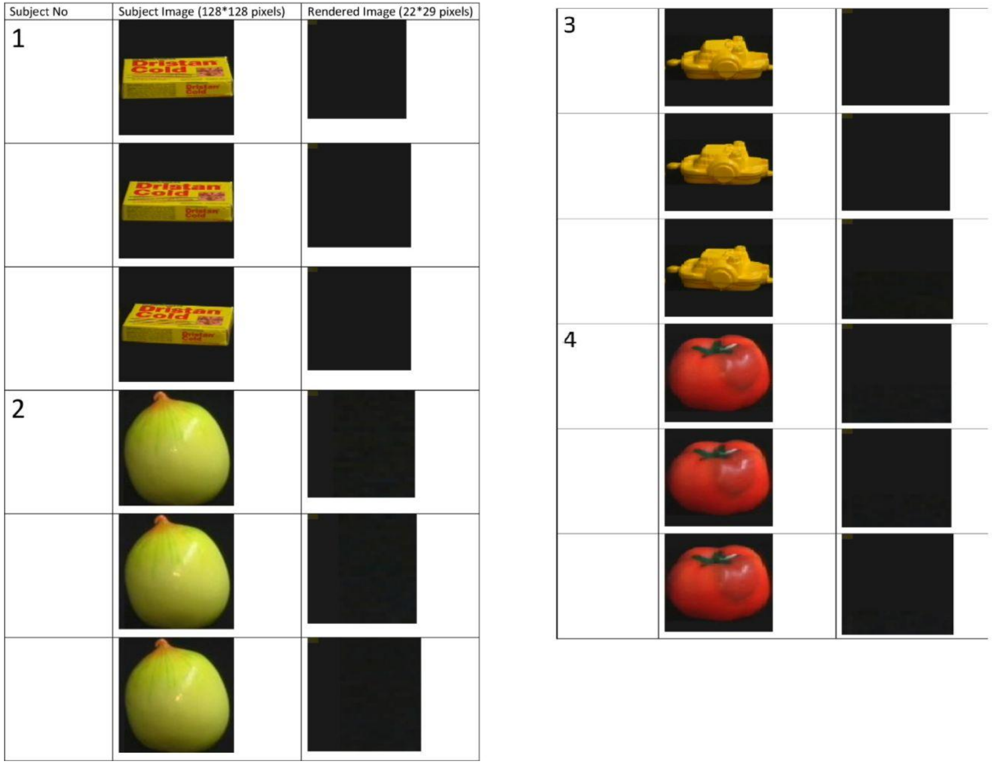
- Training based on healthcare spendings: But people of color are more likely to have lower incomes - making them less likely to access medical care even if they are insured
- Also, they may experience higher barriers to accessing health care (geography, transportation, work/childcare constraints), in addition to direct doctor-patient bias
- Data shows that race is correlated with substantial differences in health-care spendings: This results in a bias of the trained model

AI mistakes in research software

- Applying machine learning methods to COVID-19 radiological imaging for improving the accuracy of diagnosis
- Distinguish patients with COVID-19 from patients without COVID-19 but also bacterial pneumonia
- Tools were trained on public datasets with CT and CXR images
- Predictive tools failed practical tests: https://www.nature.com/articles/s42256-021-00307-0
- ***Why?***

Learning from the image background

- Image dataset was collected under controlled conditions: Does not represent the target distribution of interest
- CNN learnt to distinguish the image background rather than the image content
- Actually quite prevalent problem in computer vision



Subject No	Subject Image (128*128 pixels)	Rendered Image (22*29 pixels)
1		
2		
3		
4		

COIL-100 dataset https://www1.cs.columbia.edu/CAVE/publications/pdfs/Nene_TR96_2.pdf
https://doi.org/10.1016/j.visinf.2021.10.001

AI mistakes in research software

- Predict whether a country is likely to slide into civil war based on GDP, poverty rates, type of government structure, etc.
- Complex models using Random Forests and Adaboost outperform more standard statistical approaches like logistic regression by far
- Missing values in the dataset were constructed using imputation on the complete dataset
- Models proved to be over-optimistic and erroneous https://doi.org/10.1016/j.patter.2023.100804
- ***Why?***

Civil war predictions: Data leakage

- Data leakage: The data was imputed for missing values using the whole dataset
- Thus, the training dataset contained information about the test dataset
- This leads to an inflated estimate of the model performance

https://doi.org/10.1016/j.patter.2023.100804 supplemental material

Classification of failures/errors

Data leakage

Spurious relationship between independent variables and target variable

Artifact of collection, sampling, pre-processing

Leads to inflated estimates of model performance

Lack of clean separation training/test

- no test set
- pre-processing on training and test set (over/under sampling, imputation)
- feature selection on entire dataset
- duplicates in dataset

Kapoor, Narayanan, DOI: 10.1016/j.patter.2023.100804

Data leakage

Model uses features that are not legitimate

- for example, use of a certain drug when predicting illness (hypertensive drug, antibiotics)

Test set is not drawn from distribution of scientific interest

- temporal leakage (test set must not contain data from before the training set)
- non-independence between training and test samples (same people/units in both sets - use block crossvalidation)
- sampling bias in test distribution (spatial bias, age, image settings)

Kapoor, Narayanan, DOI: 10.1016/j.patter.2023.100804

Resources

- Good practices in machine learning (Mathieu Bauchy) (https://www.youtube.com/watch?v=WScUQnU-ozQ&t=3213s
- Roadmaps https://roadmap.sh/roadmaps
- Kaggle https://www.kaggle.com/learn
- Hugging Face: https://huggingface.co/learn
- REFORMS checklist https://reforms.cs.princeton.edu/
- More resources https://www.cs.princeton.edu/~arvindn/, https://www.aisnakeoil.com/p/introducing-the-ai-snake-oil-book
- Scikit-learn resources https://scikit-learn.org/stable/common_pitfalls.html
- Testing of non-deterministic software https://bssw.io/blog_posts/testing-non-deterministic-research-software