Making Social Science Research Transparent

Sonja Bezjak & Sergeja Masten, Slovenian Social Science Data Archives (ADP)
Marijana Glavica & Denis Vlašiček, Croatian Social Science Data Archive (CROSSDA)
Roland Ramthun, Leibniz-Institut für Psychologie (ZPID)
Peter Doorn, Data Archiving and Networked Services (DANS) & Leen Breure, Sciemedia
Jonas Recker, GESIS - Leibniz Institute for the Social Sciences
Simon Heuberger, Political Analysis
Matt Cannon, Taylor & Francis

11 November 2021, online *DOI:* 10.5281/zenodo.5719323

cessda.eu @CESSDA_Data

Making Social Science Research Transparent

Part 1: Introduction. Conceptual basis
Part 2: Transparency in practice

Online workshop

11 November 2021

cessda.eu

@CESSDA_Data

- The event will be recorded - the recording will be available on CESSDA Training Youtube channel
- Slides will be published on Zenodo and shared with participants.

Part 1: Introduction. Conceptual basis:

Introduction to Research Data Publishing

Online workshop

Sonja Bezjak and Sergeja Masten, Slovenian Social Science Data Archives, CESSDA

11 November 2021

cessda.eu

@CESSDA_Data

Content

- Journals & open data policies
- Sharing, publishing & archiving data
- Data repository services
- Consortium of Social Science Data Archives
- Data publication routes
- Smart planning

Journals & open data policies

More and more journals are demanding data availability statements

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

Are there data associated with the article you're submitting to a Taylor & Francis journal? Over 2,000 Taylor & Francis journals have **policies which state how these data should be shared.** The details below will help you get to grips with the policies and the steps you'll need to take. (https://authorservices.taylorandfrancis.com/data-sharing-policies/)

Research Data Policy Types at Springer: Data sharing and data citation is encouraged; Data sharing and evidence of data sharing encouraged; **Data sharing encouraged and statements of data availability required**; **Data sharing, evidence of data sharing and peer review of data required** (https://www.springernature.com/la/authors/research-data-policy/data-policy-types/12327096)

Experiences with researchers show that there is a confusion about different concepts and services available on the national or international level

"I will publish my data with Open Access Data Journal because I want recognition on the international level"
(researcher from Slovenia, 2020)

Availability of materials and data at the *Humanities and Social Science Communications*

An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims. **Therefore, a condition of publication is that authors are required to make materials, data and associated protocols promptly available to readers without undue qualifications. Any restrictions on the availability of materials or information must be disclosed to the publishing team at the time of submission.** Any restrictions must also be disclosed in the submitted manuscript, including details of how readers can obtain materials and information. If materials are to be distributed by a for-profit company, this must be stated in the paper. ***Humanities and Social Science Communications*** **(see Editorial policies)**

Sharing data “wheresoever”

Via USB, mail, on the project web page…

No guarantee for:

- long term preservation
- quality of data description
- data quality assessment

No credits for invested effort

Limited outreach of your data

Breakdown of Data Availability by Year of Publication (Vines et al. 2014)
Analysis of 516 articles, based on data, published between 1991 and 2011:

- strong effect of article age on the availability of data
- received only 19.5% of the requested data sets, and only 11% for articles published before 2000

The major cause of the reduced data availability for older papers

- data sets reported as either **lost or on inaccessible storage media**
- For papers where the authors gave the status of their data, the odds of a data set being extant fell by 17% per year.
- the odds that we could find **a working e-mail address for the first, last, or corresponding author** fell by 7% per year.

TIMOTHY H. VINES, ARIANNE Y.K. ALBERT, ROSE L. ANDREW, FLORENCE DE´BARRE, DAN G. BOCK, MICHELLE T. FRANKLIN, KIMBERLY J. GILBERT, JEAN-SEBASTIEN MOORE, SEBASTIEN RENAUT in DIANA J. RENNISON (2014): The Availability of Research Data Declines Rapidly with Article Age. *Current Biology* (24): 94–97. Dostopno na: http://dx.doi.org/10.1016/j.cub.2013.11.014 (10. november 2020).

Publishing data

Publishing data for reuse

To make your data reusable for purposes beyond the one for which you collected them, you should publish your data. Publishing your data is the act of **publicly disclosing the research data you have collected**, making them findable, accessible and reusable.

Archiving data for future reference

Research data archiving is about **storing and preserving research data for the long term**. When you archive your data, you make sure you can read and access the data later on. You can then also allow access to others for verification purposes when such a request arrives. In all cases, you should store your data safely, in a suitable file format, with adequate documentation.

CESSDA Training Team (2017 - 2020). *CESSDA Data Management Expert Guide.*
Bergen, Norway: CESSDA ERIC. Retrieved from https://www.cessda.eu/DMGuide

Describing data in different ways

Data description can be more or less structured:

1) within the scientific article
2) data paper in a data journal
3) standardized data description, using metadata format, recognized by data repositories

Have in mind:

- *For what purpose do you need a data description?*
- *Where will your data be accessible for reuse?*

Smart planning - record your research work (metadata)

https://brill.com/view/journals/rdj/rdj-overview.xml

Data repository services

*"A data repository is a **digital archive collecting, preserving and displaying datasets, related documentation and metadata**. Repositories and archives typically use terms like "preservation" and "curation" rather than "archiving" or "storage": long-term accessibility implies expertise and services to convert data to new formats and to add value to the data, for instance by new functionality to query the data."*

OPENAIRE (2017): Briefing Paper Research Data Management. Accessbile at: https://www.openaire.eu/briefpaper-rdm-infonoads (8th November 2021).

https://www.re3data.org/

Consortium of Social Science Data Archives - CESSDA

Provides a distributed and sustainable **research infrastructure,** that

- enables the research community to conduct high-quality research in the social sciences,
- offers **services to data producers** to easily describe and store their data,
- contributes to the production of effective solutions to the **major challenges facing society today**.

Facilitates teaching and learning in the social sciences.

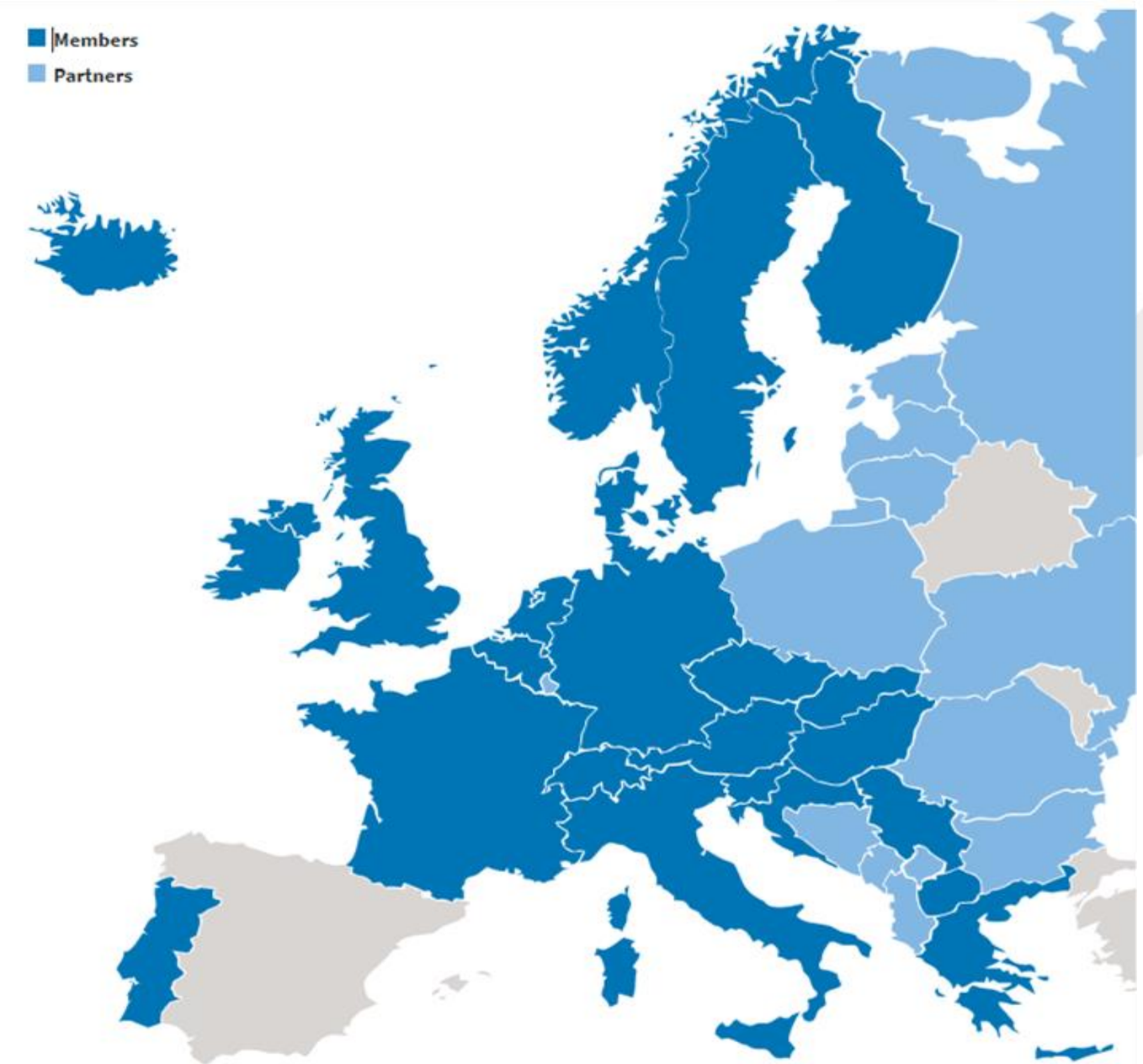
Consortium of Social Science Data Archives

Is a **consortium of trusted repositories with full European coverage**, offering a platform with tools and services to both data producers and data re-users.

CESSDA repositories ensure safe path to data publishing!

CESSDA Countries

Members	Partners
Austria	Albania
Belgium	Bosnia and Herzegovi
Croatia	Bulgaria
Czech Republic	Estonia
Denmark	Kosovo
Finland	Latvia
France	Lithuania
Germany	Luxembourg
Greece	Montenegro
Hungary	Poland
Iceland	Romania
Ireland	Russia
Italy	Ukraine
Netherlands	
North Macedonia	
Norway	
Portugal	
Serbia	
Slovakia	
Slovenia	
Sweden	
Switzerland (Observer)	
United Kingdom	



cessda

2 WGs under CESSDA Training WG

TRAINING EVENTS

- responsible for training of researchers (research data management) and data re-users (data discovery)

JOURNALS OUTREACH

- provides support for journals and editors in open data policy decision making

The Training Working Group maximises the potential of the trainings offered by each Service Provider and promotes harmonisation and knowledge transfer within CESSDA.

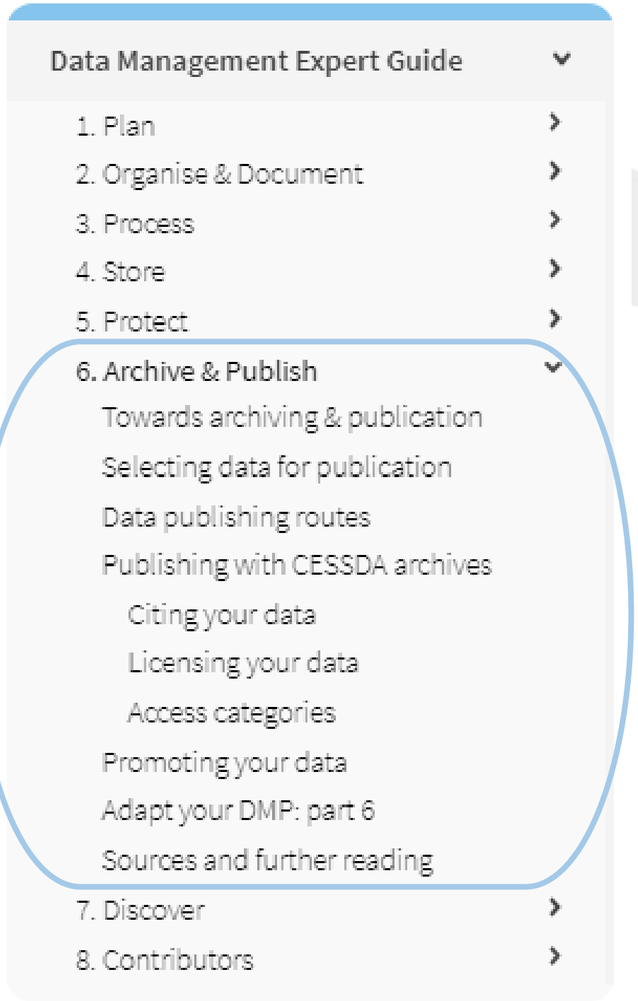
Data Management Expert Guide - DMEG

A **tool** for **social science researchers** who are in an early stage of practising **research data management**.

With this guide, CESSDA wants to **contribute to professionalism in data management** and **increase the value of research data**.

This guide is designed t**o help social science researchers make their research data** Findable, Accessible, Interoperable and Reusable (**FAIR**).

CESSDA Training Team (2017 - 2020). CESSDA Data Management Expert Guide.
Bergen, Norway: CESSDA ERIC. Retrieved from https://www.cessda.eu/DMGuide

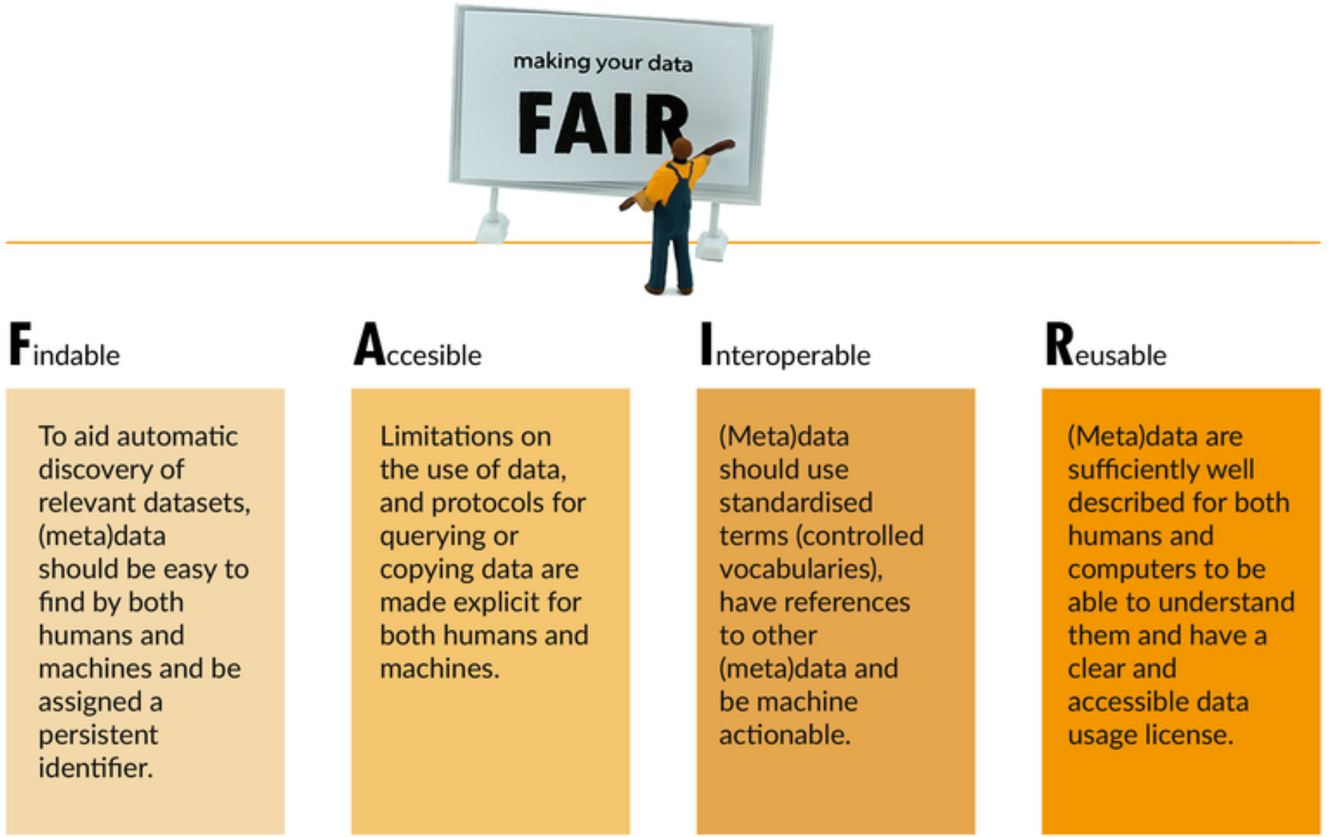


FAIR data

DMEG supports FAIR principles!

To achieve FAIRness, data objects should at least have:

- A **persistent identifier** (PID) for the data object as a whole
- A **sufficient** set of **metadata**
- A clear **licence**



CESSDA archives provide services in order to help researchers **PUBLISH FAIR DATA.**

Data Publication

It is expected that a Data publication will ensure that data will potentially be considered as a first class research output (Knowledge, 2013)

For a dataset to "count" as a publication, it should follow a similar publication process to an article (Brase et al., 2009) and should be:

- Properly documented with metadata;
- Reviewed for quality;
- Searchable and discoverable in catalogues (or databases);
- Citable in publications.

- Data is publicly accessible now and for the future
- Access to data is clearly determined and does not depend on author's caprice

PUBLICATIONS AND DATA

Data publication routes

- Institutional data repository
- General purpose repository
- Domain specific data repository
- Trusted domain specific data repository

(Trusted) Domain specific repositories

- data curators **for specific data** types/topics/disciplines
- build **specialized data catalogues**
- are connected with other research data archives in **archive community**
- archive and publish **data of higher quality** and potential for **reuse**
- provide **technical and content review**; some also scientific review
- (can) hold a certificate of being a trustworthy repository

Alternative routes

Institutional repositories

- meant for researchers from one institution; used when there is no domain specific repository available, but one should publish the data.

General purpose repositories

- recommended when there is no domain specific or institutional repository;
- publish data from various disciplines;
- services adapted to heterogenous and long-tail data;
- no guarantee for long term preservation;
- no technical and scientific review of data and documentation.

European Commission as funder about RDM

Research data management is **mandatory in Horizon Europe for projects generating or reusing data**. If you expect to generate or reuse data and/or other research outputs (except for publications), you are required to outline in a maximum of one page how these will be managed.

Responsible management of research data in line with the FAIR principles of 'Findability', 'Accessibility', 'Interoperability' and 'Reusability', notably through the generalised use of data management plans, and open access to research data under the principle 'as open as possible, as closed as necessary', under the conditions required by the grant agreement;

Important elements and resources for RDM useful already at proposal stage

- **Persistent identifiers (PIDs)** are key in ensuring the findability of research outputs, including data.
- To enhance the findability of research outputs, and their potential reuse, **standardised metadata frameworks** are essential, ensuring that data and other research outputs are accompanied by rich metadata that provides them with context.
- **Trusted repositories** assume a central role in the Horizon Europe for the deposition of and access to publications and research data.

Horizon Europe (HORIZON), Programme Guide, Version 1.2, 04 October 2021. Accessible at: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf

Be smart - start planning!

CESSDA Training Team (2017 - 2020). *CESSDA Data Management Expert Guide.* Bergen, Norway: CESSDA ERIC. Retrieved from https://www.cessda.eu/DMGuide

Questions!

Sonja Bezjak, sonja.bezjak@fdv.uni-lj.si
Sergeja Masten, sergeja.masten@fdv.uni-lj.si

cessda.eu

@CESSDA_Data

Making Social Science Research Transparent

Elements of Research Transparency

Marijana Glavica
Denis Vlašiček
Croatian Social Science Data Archive (CROSSDA)

11 November 2021

cessda.eu

@CESSDA_Data

What do we talk about when we talk about research transparency ?

[before] Transparent methods

SUBJECTIVITY and FREEDOM

questions

measures

methods

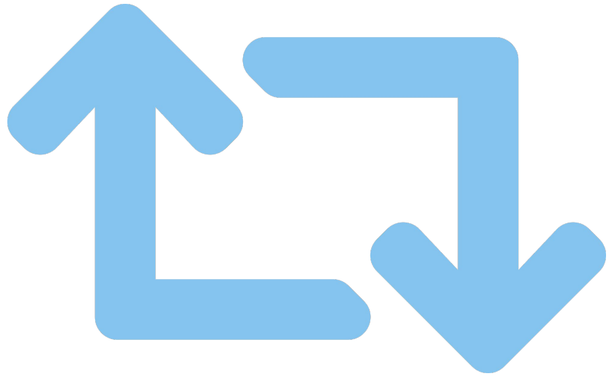
procedures

analyses

[before] Transparent methods

preregistration

- informal
- nonbinding
- not reviewed



registered reports

- ~~in~~formal
- ~~non~~binding
- ~~not~~ reviewed

[after] Preprints

institutional repositories

SocArXiv

PsyArXiv

[after] Open peer review

open identities

open reports

open participation

open interaction

open pre-review manuscripts

open final version commenting

open platforms ("decoupled review")

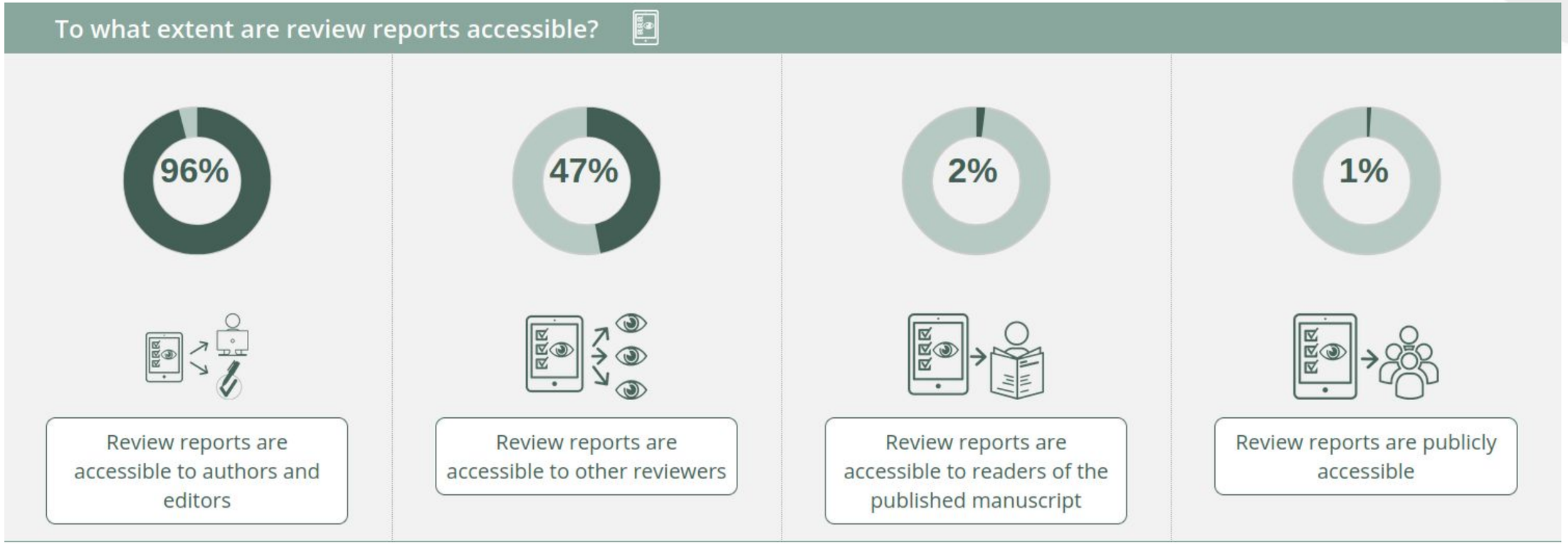
scienceOPEN.com
research+publishing network

PUBPEER
The online Journal club

[after] Open peer review

THE PLATFORM FOR RESPONSIBLE EDITORIAL POLICIES

https://www.responsiblejournals.org



cessda

[after] Responsible research assessment

San Francisco Declaration on Research Assessment (DORA)

Leiden Manifesto for Research Metrics

The Hong Kong Principles for assessing researchers

fair and transparent research assessment system

move away from publication metrics -- altmetrics

[beyond] Numerical reproducibility

short- and long-term

`numerical reproducibility = data`

openly available | trusted repository | well documented

[beyond] Numerical reproducibility

short- and long-term

```
numerical reproducibility = data +
                            code
```

openly available [| well written | well documented]

[beyond] Numerical reproducibility

short- and long-term

```
numerical reproducibility = data +
                            code +
                            software
```

```
R 1.0.0 (2000)
R 2.0.0 (2004)
R 3.0.0 (2013)
R 4.1.2 (2021)
```

Thank you for your attention!

Marijana Glavica <mglavica@ffzg.hr | @mglavica>
Denis Vlašiček <dvlasicek@ffzg.hr | @dvlasicek>
CROSSDA @crossda_data

cessda.eu

@CESSDA_Data

Making Social Science Research Transparent

Part 2: Transparency in practice - Implementation of an open-science research cycle model at ZPID

1. Rationales (in Psychology)

Replication crisis and trust in science

The New York Times

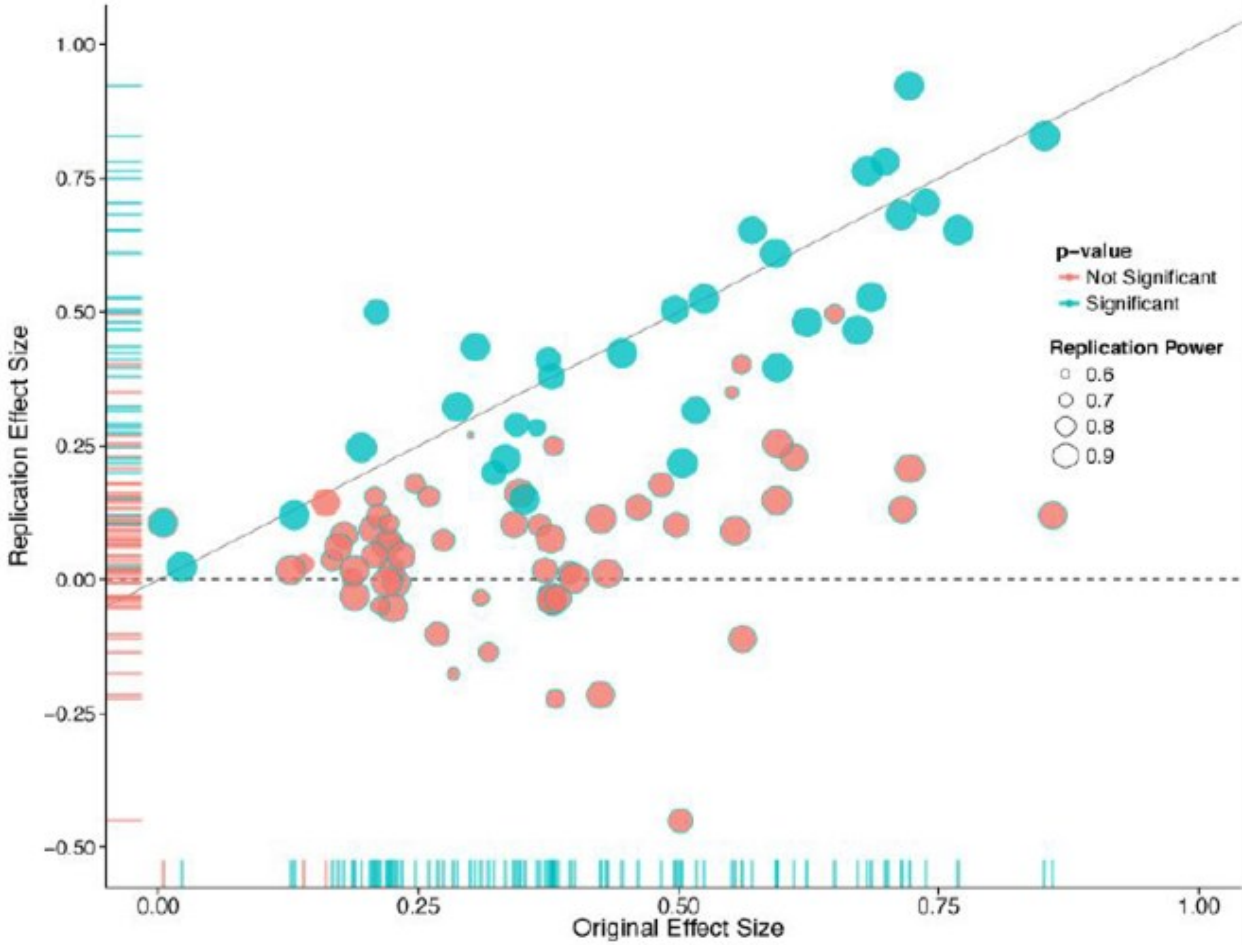
Opinion

LETTERS

How Reliable Are Psychology Studies?

Sept. 8, 2015

Replication, *noun:*
Purposeful repetition of a study to assess the reliability and generalizability of findings



Source: Open Science Collaboration (2015)

Questionable research practices


PLOS MEDICINE

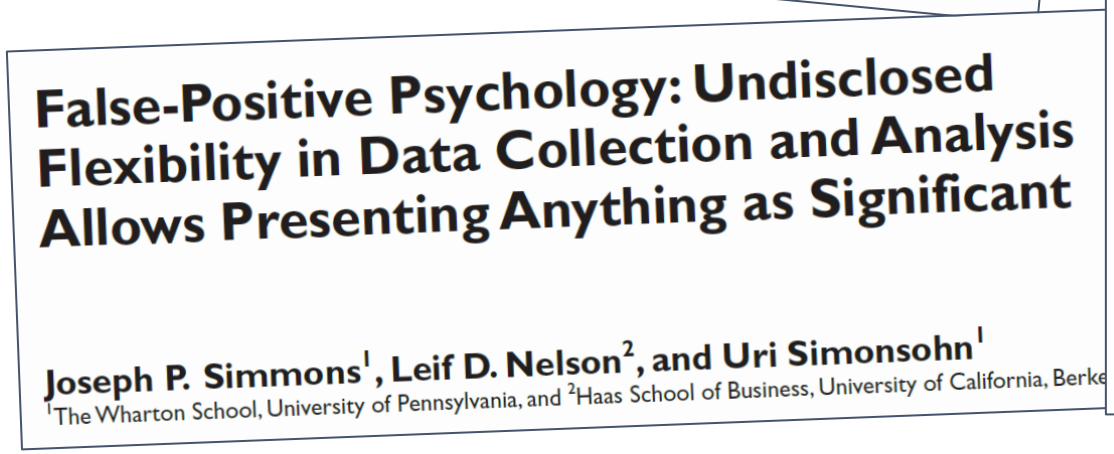
OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

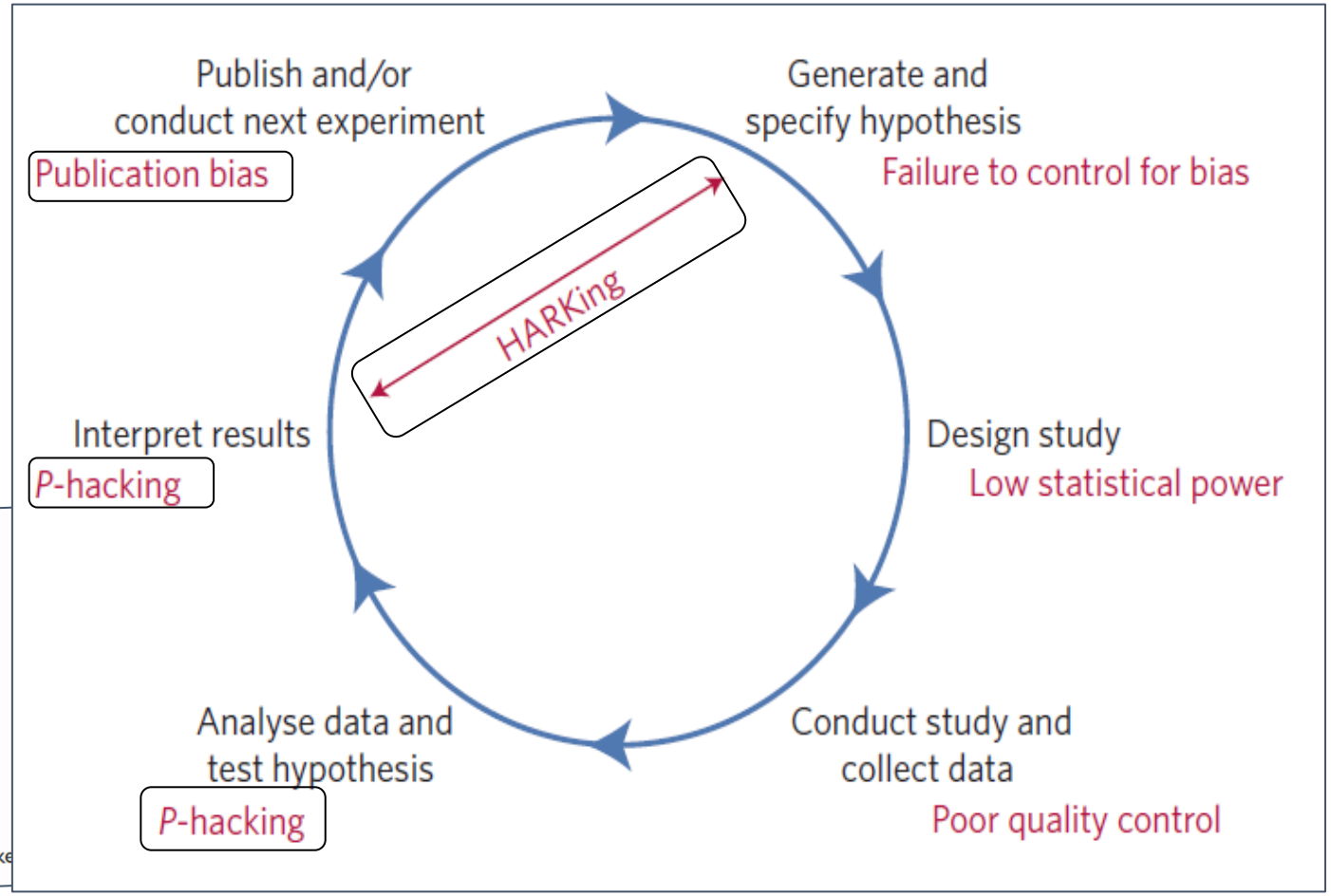
John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124


False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]

[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berke

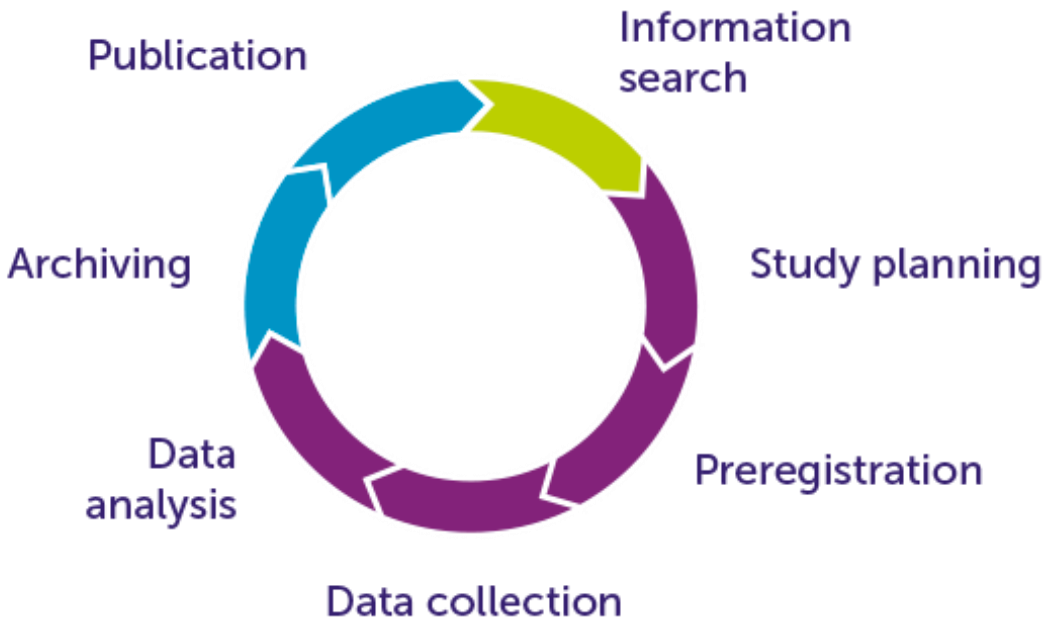


Source: Munafò et al. (2017)

2. ZPID Tools

Leibniz Institute for Psychology (ZPID)

- … is a Public Open Science Institute for Psychology
- … is in the process of strategic expansion towards a one-stop research support organization
- … aims to support the (scientific) community in psychology to make research accessible, transparent, reproducible, and replicable.

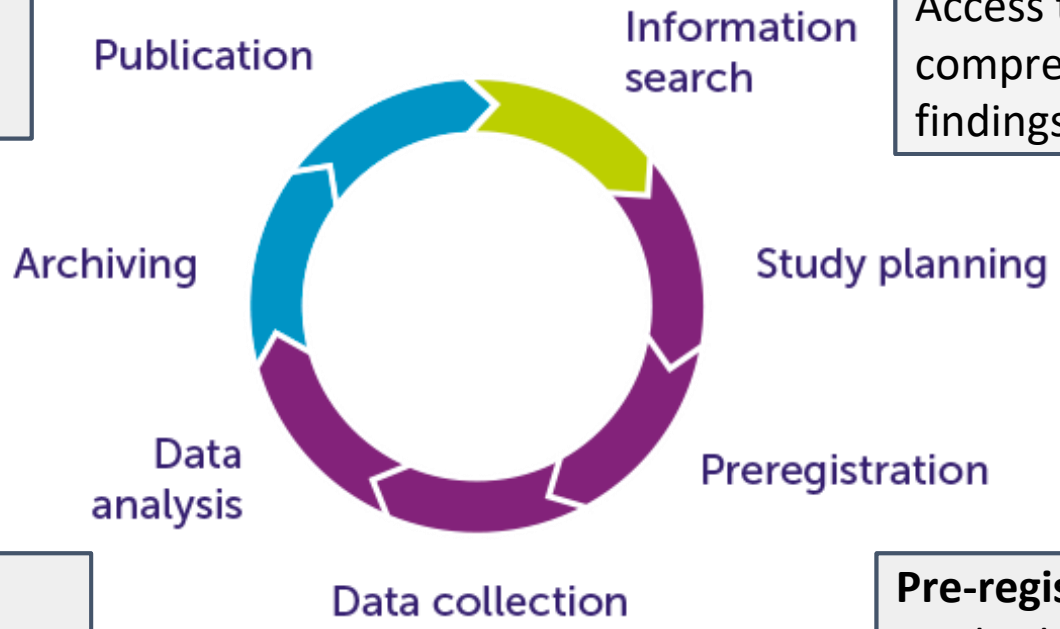


ZPID and Open Science Principles

Open Access
Open access publishing opportunity, accelerated synthesis

FAIR DROs
Licensing and support for re-usable resources

Open Source
Sharing of tools and code for reproducibility and collaboration



Open Access
Access to published works, comprehensible communication of findings to the public

Open Methods
Adherence to protocols for data collection and metadata

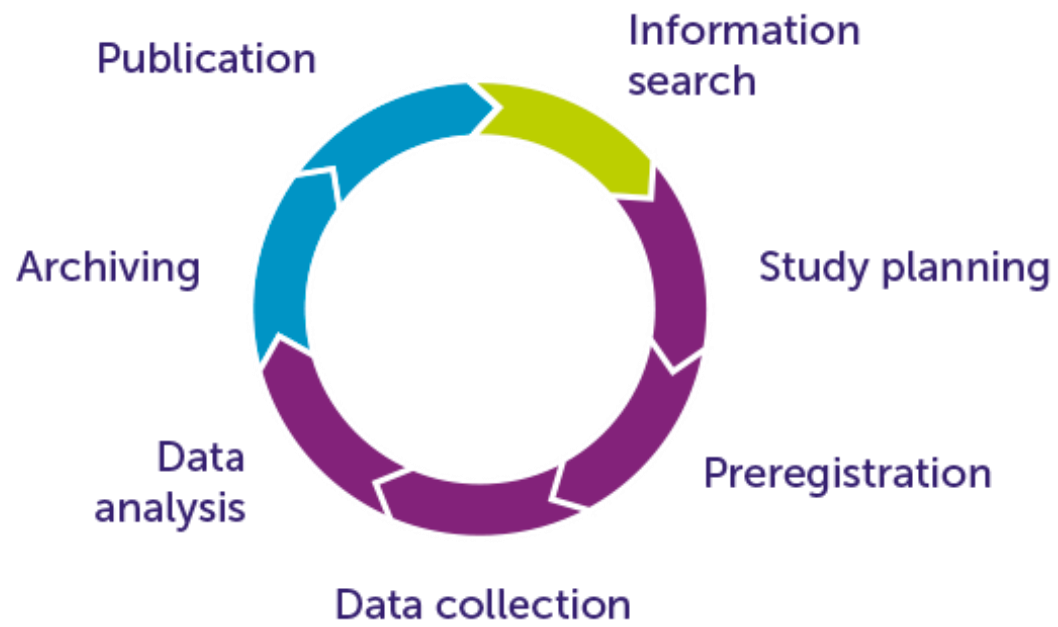
Pre-registration Assessment of methodological quality and feedback independent of results

ZPID and Open Science Tools

PsychOpen

PsychArchives

PsychNotebook



PubPsych

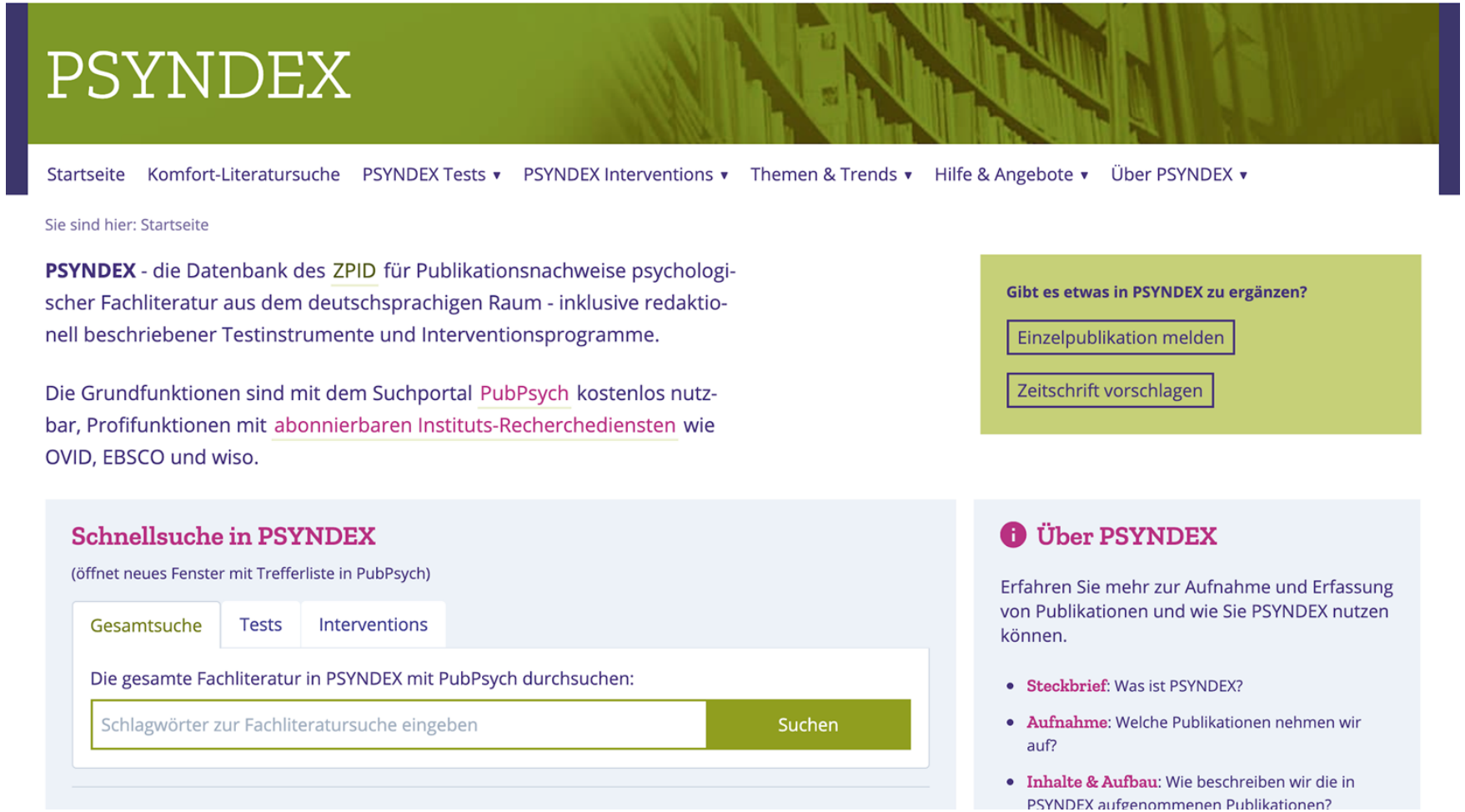
PsychNotebook

PreReg

PsychLab

PubPsych

https://www.psyndex.de/



PubPsych

https://www.pubpsych.de/

Study planning

Data analysis

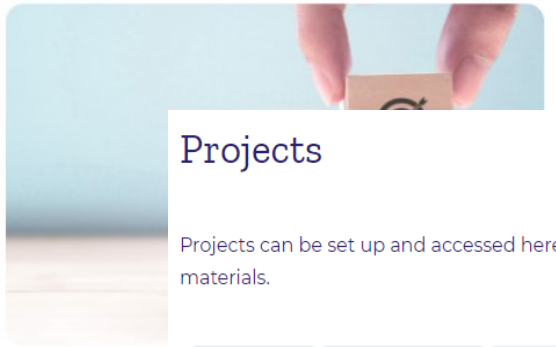
PsychNotebook

PsychNotebook

PsychNotebook

Projects FAQ

PsychNotebook is an online platform for planning and analyzing studies in psychology and related disciplines.



Our mission

Providing you with the tools to practice open science!

Projects

Projects can be set up and accessed here. Projects are collections of analysis scripts, data, and other materials.

Import project | New project

My projects | Shared projects | Public projects

Features

Log in and st
Share your pr
Create, copy
Export your p

Title	Description	Actions
2021 04 LJU		
Formatierungsbeispiele	Ein kurzer R Markdown Guide.	
Preregistration for Quantitative Rese...	This project contains the preregistration template PRP-QUANT, version 2, as...	
R_example_exercises	Interactive code exercises that were created with learnR. Run RMarkdown t...	

PreReg

Usage scenarios

- Notarized proof of authorship in the earliest possible stage
- A preregistration template is provided: Preregistration for Quantitative Research in Psychology (PRP_Quant)
- Provided in different formats
- Educational tool to train students in study planning.

PsychLab

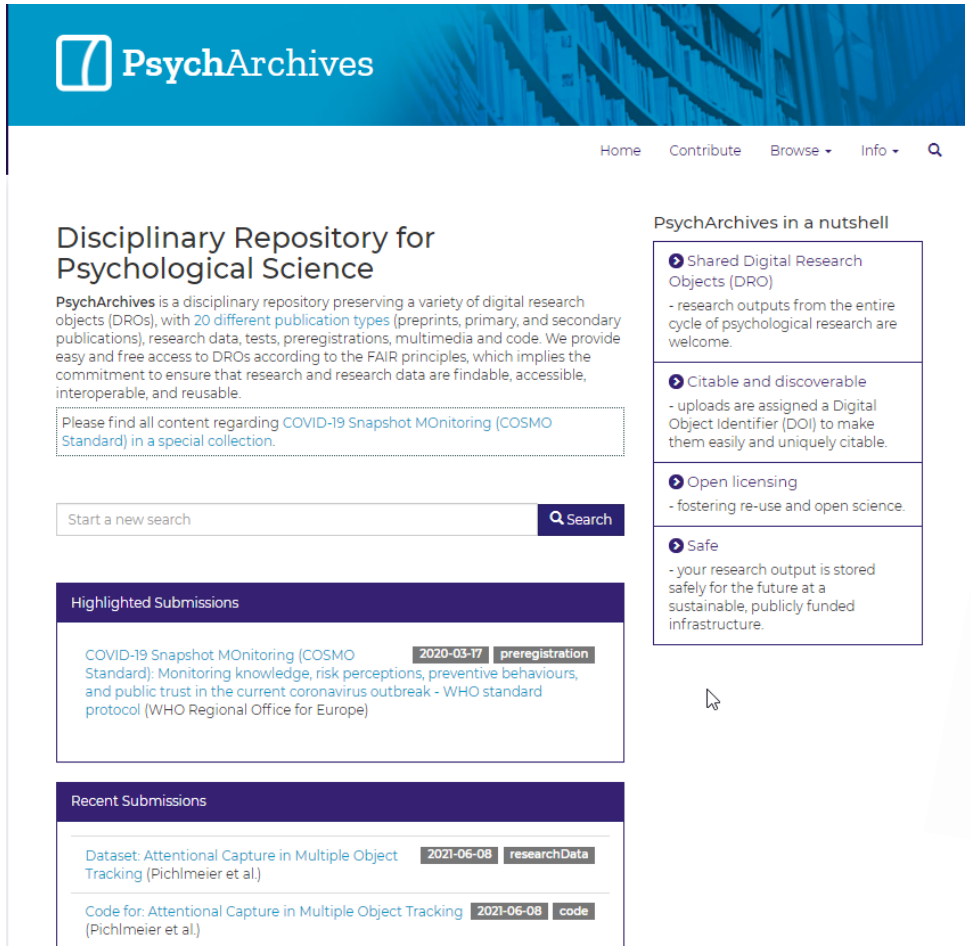
Scope Online Lab (surveys & online experiments)

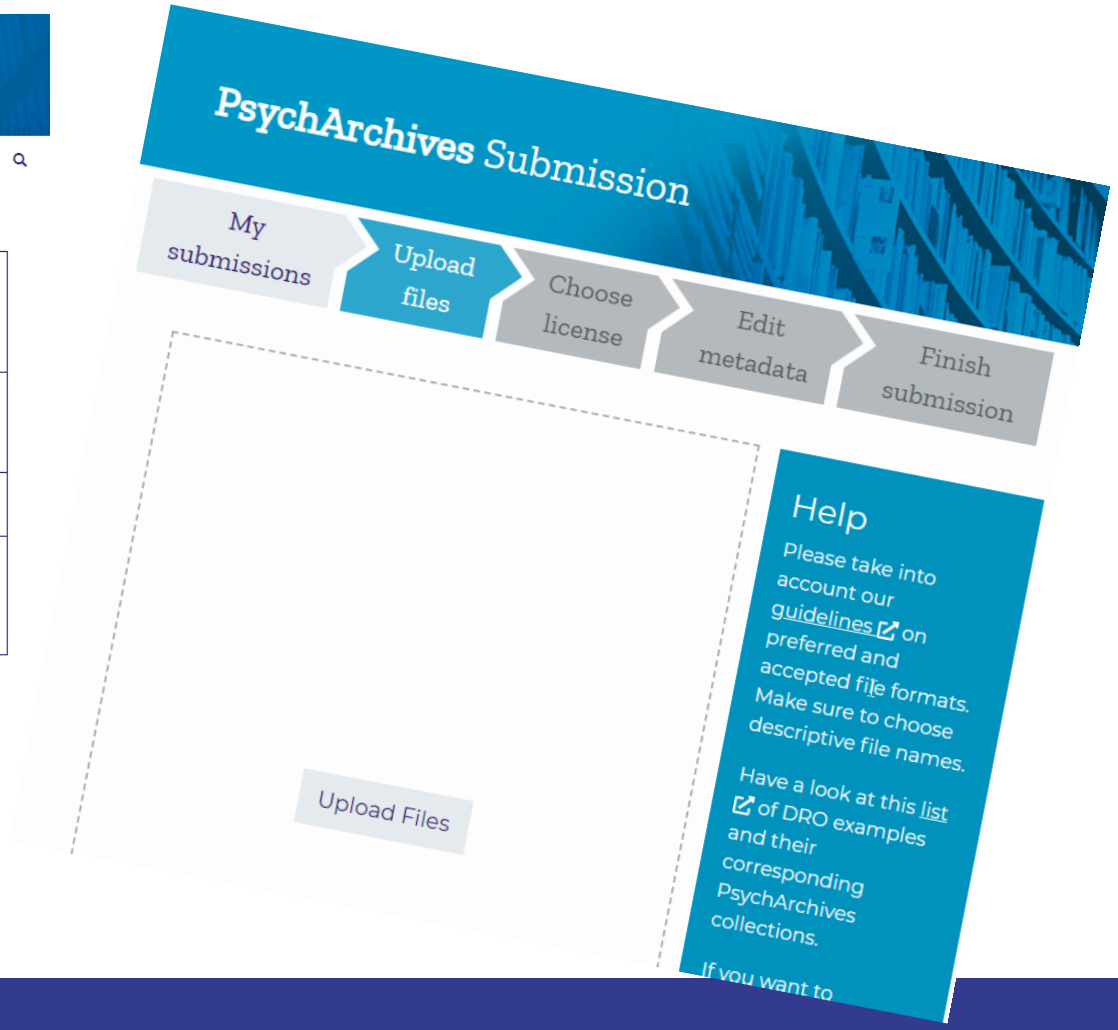
PsychLab

Scope Offline Lab (Eye tracking or any PC-based experiments)

PsychArchives

https://www.psycharchives.org/





PsychOpen

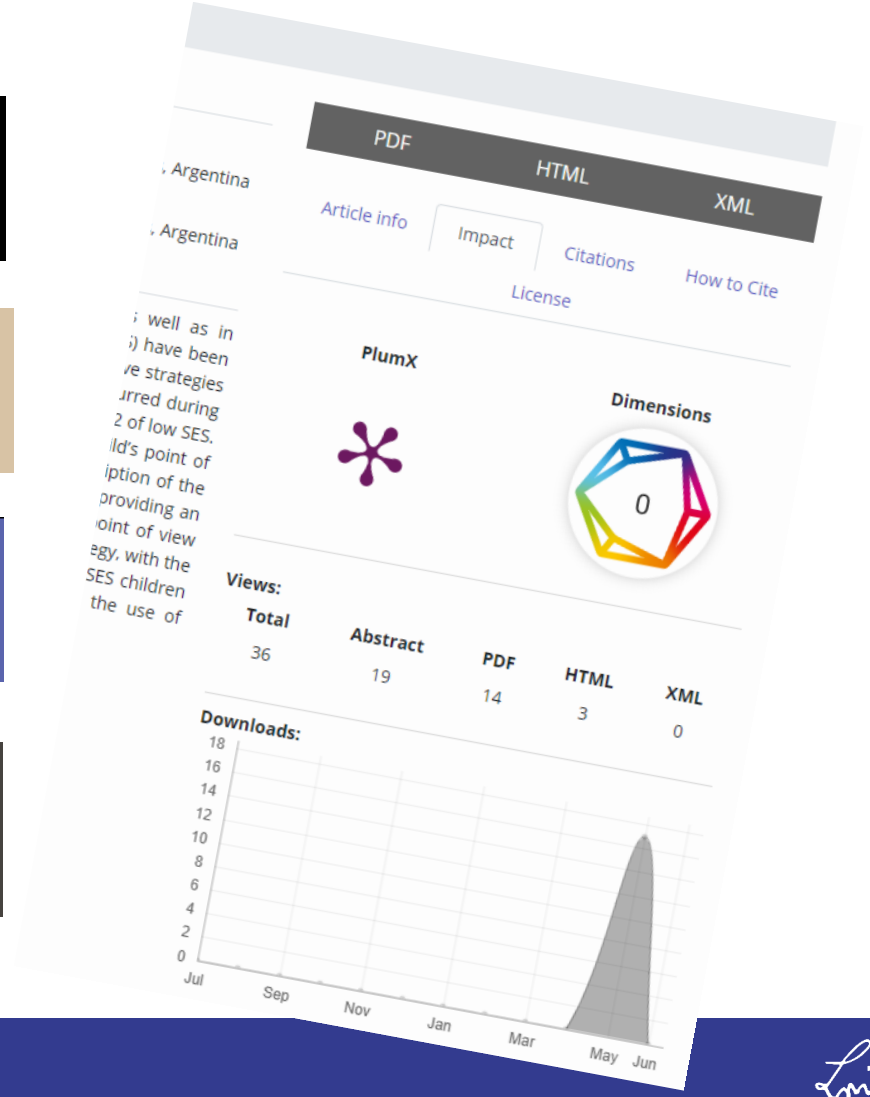
https://gold.psychopen.eu/

PERSONALITY SCIENCE

Quantitative and Computational Methods in Behavioral Sciences

METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

Sexual Offending
Theory, Research, and Prevention



PsychOpen

TOP guideline	ZPID offer
Citation Standards	-
Data Transparency	PsychArchives / RDC at ZPID
Analytic Methods (Code) Transparency	PsychNotebook / PsychArchives
Research Materials Transparency	PsychArchives
Design and Analysis Transparency	PsychNotebook
Study Preregistration	PreReg in Psychology / PsychArchives
Analysis Plan Preregistration	PreReg in Psychology / PsychArchives
Replication	PreReg in Psychology / PsychLab

https://www.cos.io/initiatives/top-guidelines

3. Lessons learned

Criticism on certain aspects of open science

- time required
- unclear questions w.r.t. credit of original authors
- material prone to misunderstanding
- material prone to misuse
- emerging privacy issues for participants

Criticism on the system of open science

- technology driven change
- science may transform into a “neo-liberal enterprise”
  - capture of publicly funded research value by commercial companies
  - just a different set of gatekeepers and more metrics
  - compare e.g. stoptrackingscience.eu

Time for your questions and thoughts

Data Archiving and Networked Services

DANS

Showcasing Research Data in Archives, Repositories and Data Journals

Why we need more than just catalogues, metadata and plain articles

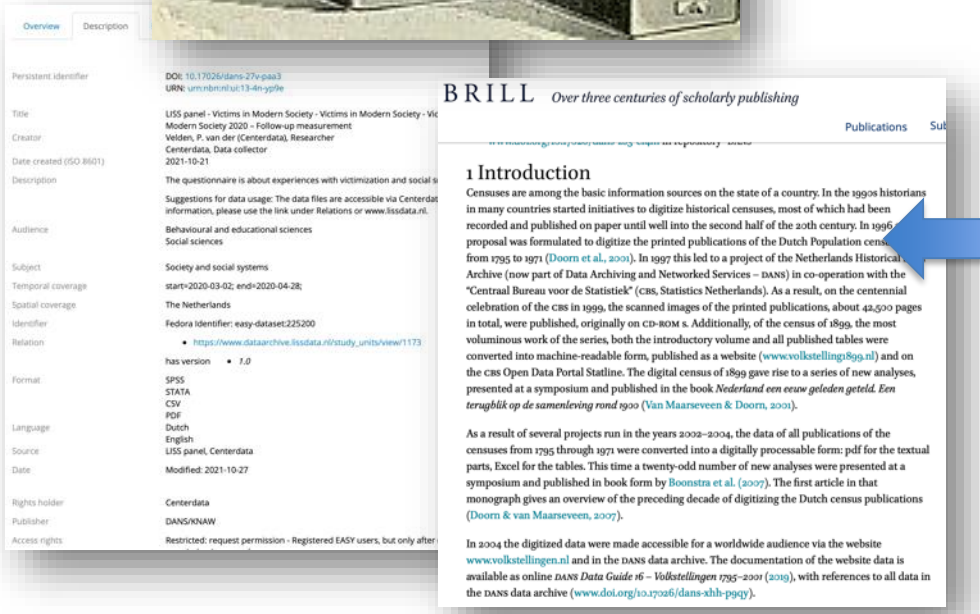
Peter Doorn, DANS

Leen Breure, Sciemedia

twitter @DANS_knaw_nwo @pkdoorn

CESSDA Workshop: Making Social Science Research Transparent.
Online, 11 Nov 2021

Data: boring metadata vs. user experience



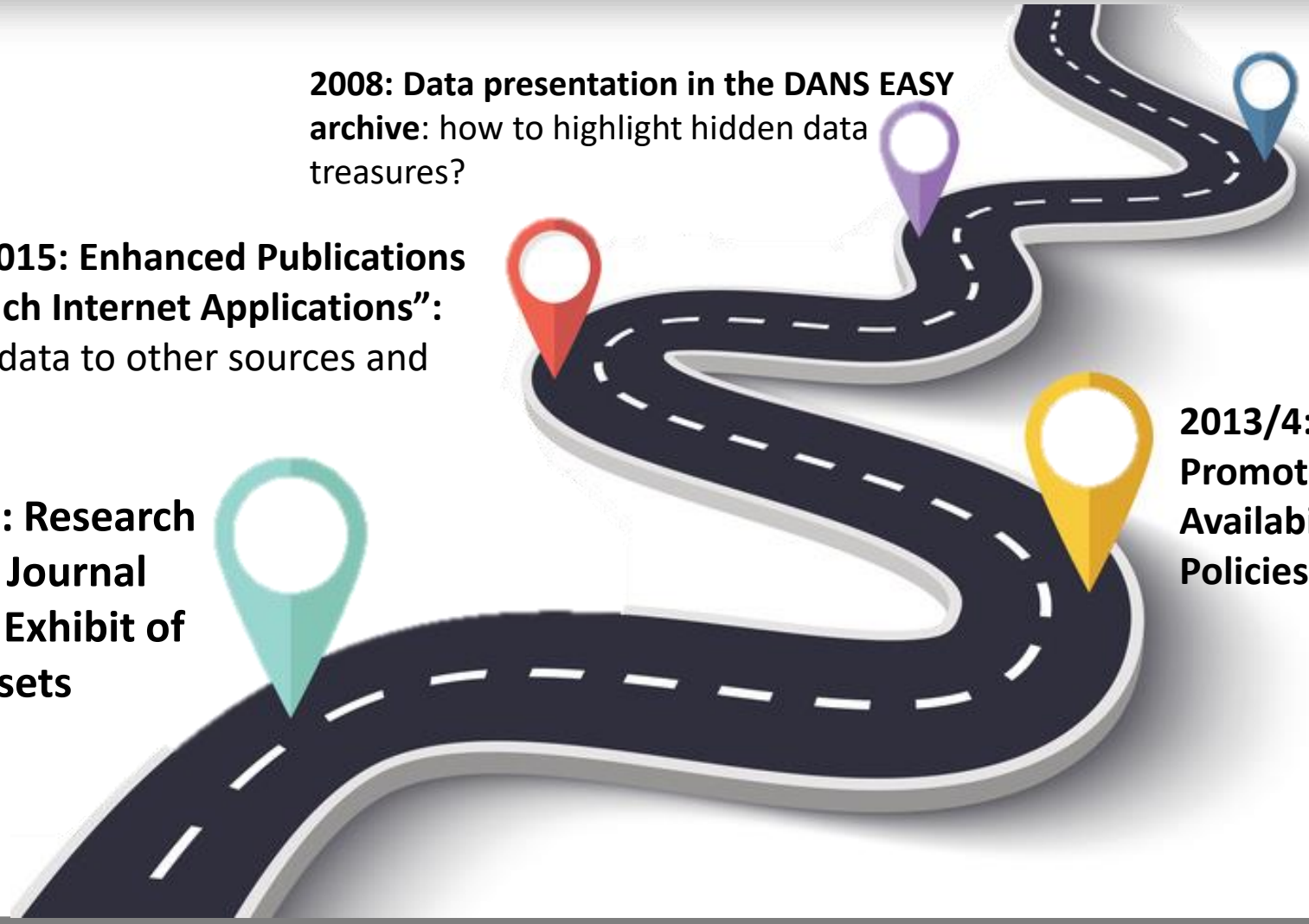
Catalogues, metadata, (data) articles
➔ only **information about** data

Provide users with data **exploration**
➔ experimental **usage**

Data Archiving and Networked Services
DANS

However, there is a long way to go…

On the road to experiencing data …



1990s: "Data Marketing" project: How to present data better?

2008: Data presentation in the DANS EASY archive: how to highlight hidden data treasures?

2005-2015: Enhanced Publications and "Rich Internet Applications": linking data to other sources and media

2013/4: Promote Data Availability Policies

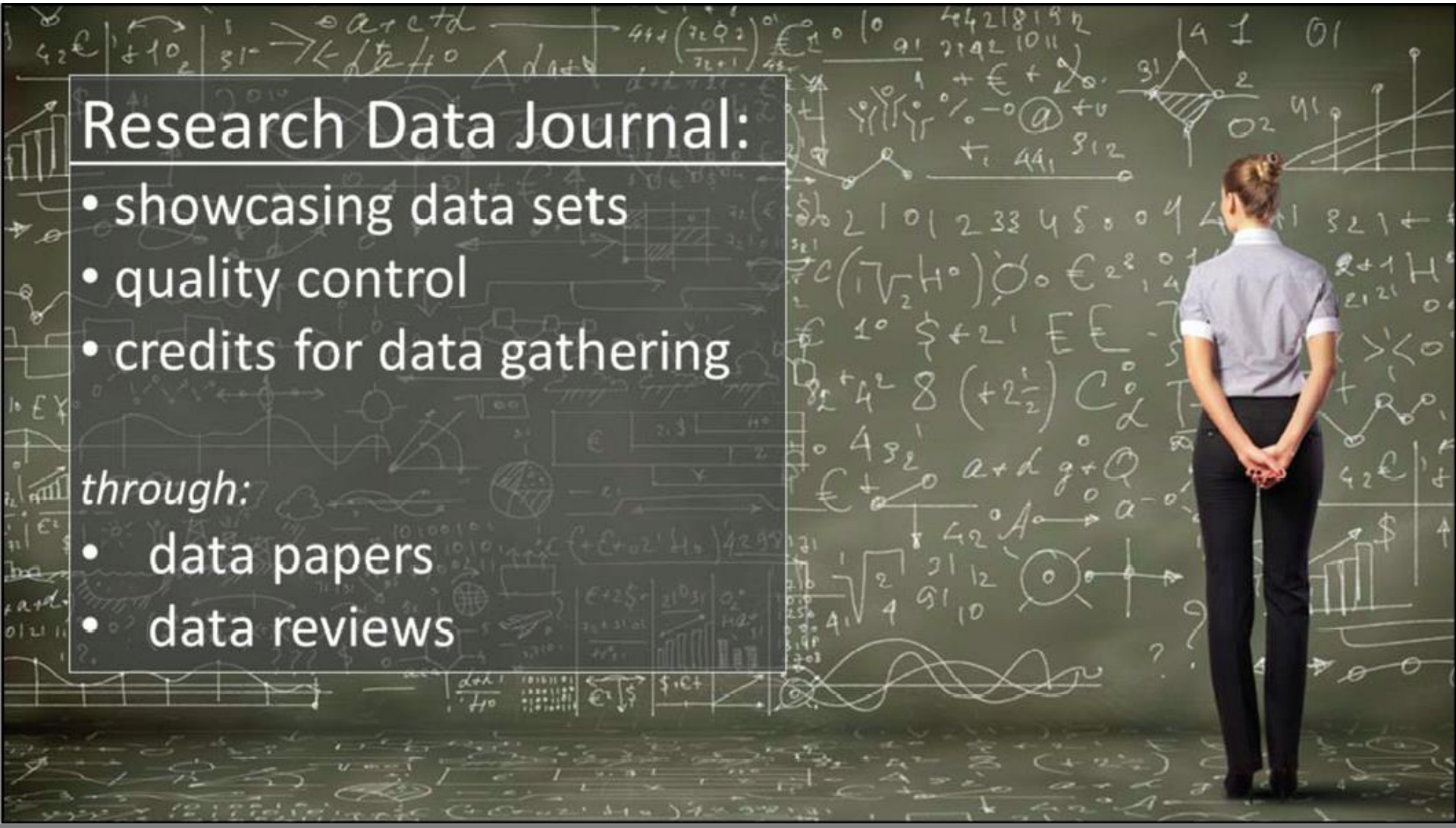
2016: Research Data Journal with Exhibit of Datasets

Gallery EP Features 3

Leen Breure: typology of 80 types of "Enhanced Publications" - www.xposre.nl/epfeatures/

DANS

Added value of Data Journal:

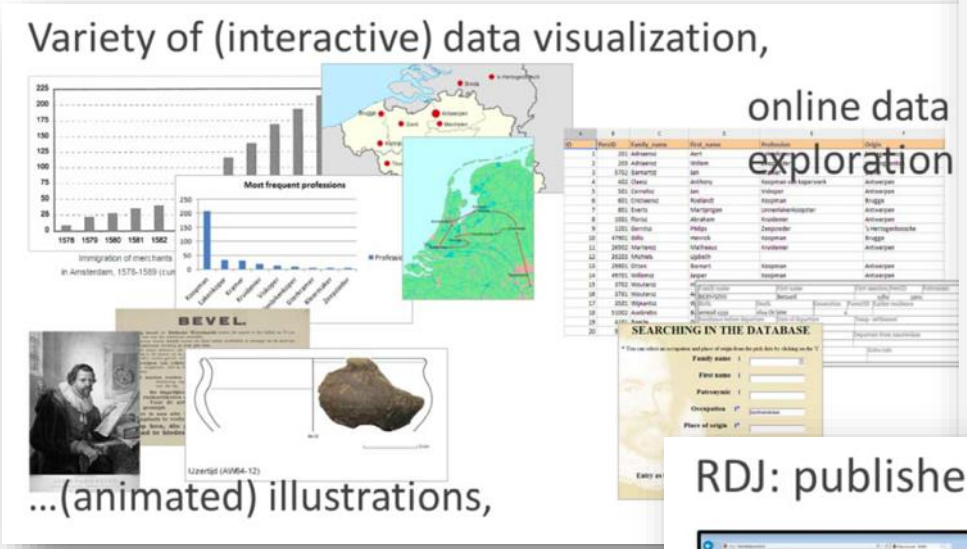


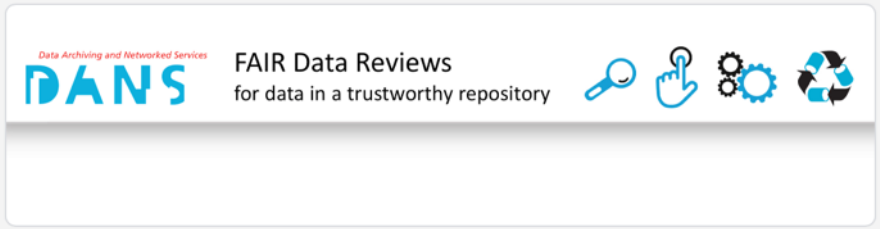
A data paper describes the research context of a data set, not the research question is the central problem

Data Archiving and Networked Services
DANS

Research Data Journal for the Humanities and Social Sciences: https://brill.com/view/journals/rdj/rdj-overview.xml

Added functionality we aimed for



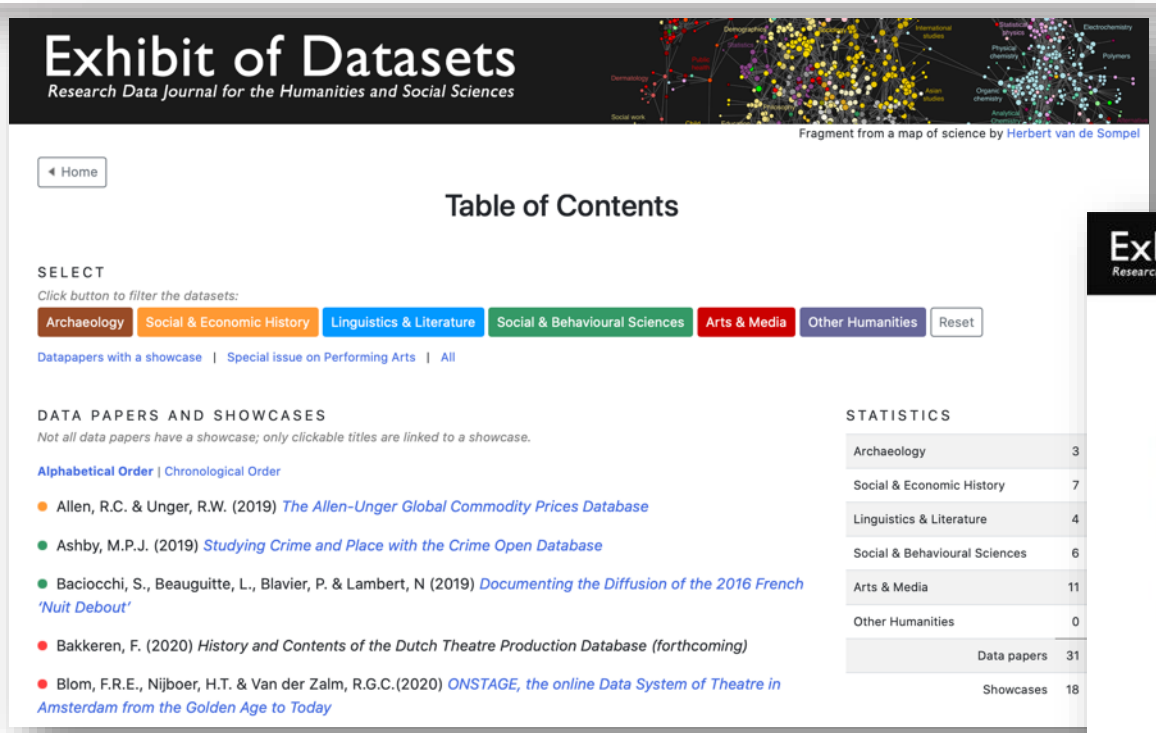


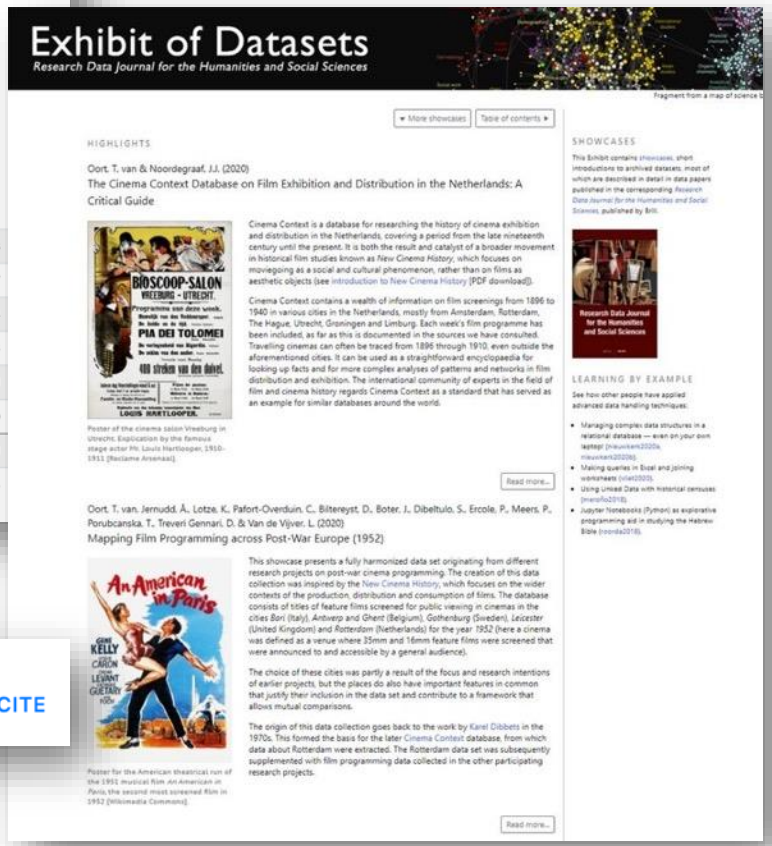
Research Data Journal - FAIR Data Review

This FAIR Data Review Form is aimed at editors and reviewers of data belonging to a data paper submitted to the Research Data Journal (RDJ). The data should be stored at a trustworthy repository, such as the DANS data archive: easy.dans.knaw.nl, which is certified with the Core Trust Seal (CTS, see https://www.coretrustseal.org).

However: the publishing platform, like many other online journals, supports only limited functionality in terms of data presentation and data reviewing

Data papers tended towards “normal” research articles; therefore:





Data Archiving and Networked Services
DANS

https://dansdatajournal.nl/rdp/toc.html

Data showcases in the “Exhibit” link to archived data sets and to data papers

UK Data Service

10.5255/UKDA-SN-6226-6

Citation:
Thompson, P. (2019). *Pioneers of Social Research, 1996-2012.* [data collection]. *4th Edition.* UK Data Service. SN: 6226, http://doi.org/10.5255/UKDA-SN-6226-6

RESEARCH DATA JOURNAL FOR THE H
AND SOCIAL SCIENCES (2019
BRILL
BRILL OPEN
om/rdj

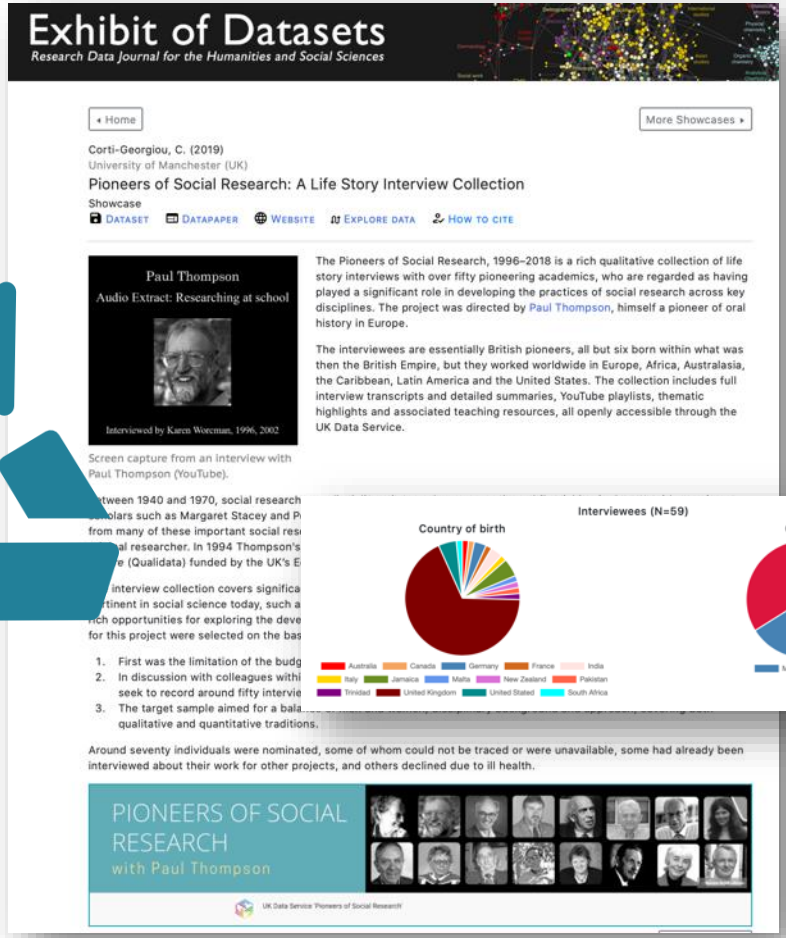
Pioneers of Social Research: A Life Story Interview Collection

Social and Behavioural Sciences

Camille Corti-Georgiou
University of Manchester, UK
camillegeorgiou@hotmail.co.uk

Abstract

The *Pioneers of Social Research, 1996–2018* is a rich qualitative collection of life story interviews with over fifty pioneering academics, who are regarded as having played a significant role in developing the practices of social research across key disciplines. The project was directed by Paul Thompson, himself a pioneer of oral history in Europe. The interviewees are essentially British pioneers, all but six born within what was then the British Empire, but they worked worldwide in Europe, Africa, Australasia, the Caribbean, Latin America and the United States. The collection includes full inter-view transcripts and detailed summaries, YouTube playlists, thematic highlights and



Objectives of Data Showcases

Attention

FAIRness

Coherence

Interactivity

Context

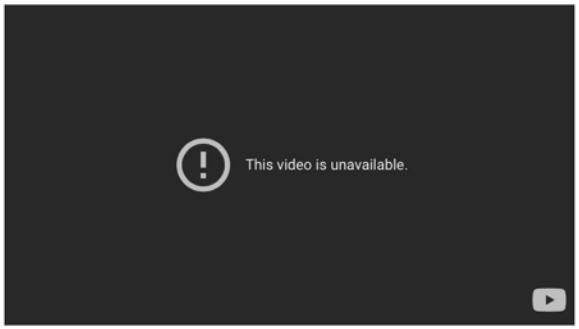
Multimodality

Trust

Transparency

Or is it a Dead End?

Elsevier: “The Article of the Future” (2012)



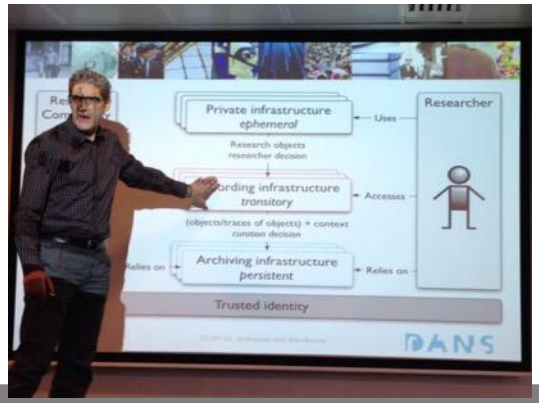
When will this future be?

The Future of Scholarly Skywriting

(-1999)

Vanishing into thin air...

2014: The future of scholarly communication and the archive:



Stef Scagliola’ desillusion (09/11/2021): *“10 years ago I managed to realise an ‘enhanced publication’ [...] I assumed that in the ten years that have past, there would be an easy way to again publish a PDF online with links to a variety of sources that are published elsewhere online [...] The article is published, but the creation of a space in which our private material could be published in order to be linkable, turned out to be impossible”.*

Core challenges

Scalability: make it easy and quick to produce data showcases

Templates

Instructions

Individual datasets and collections

Maintainability: can we preserve the functions?

Existing tools

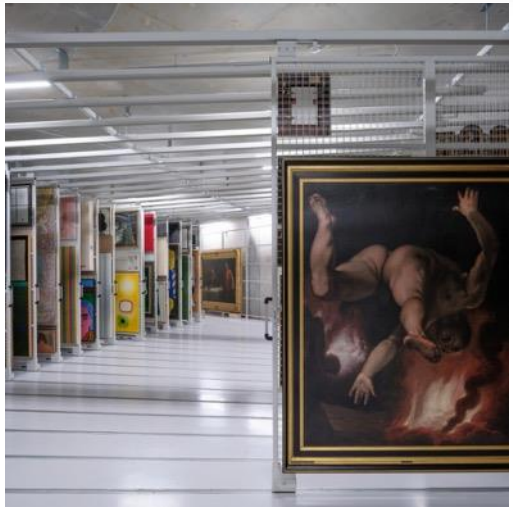
F-UJI

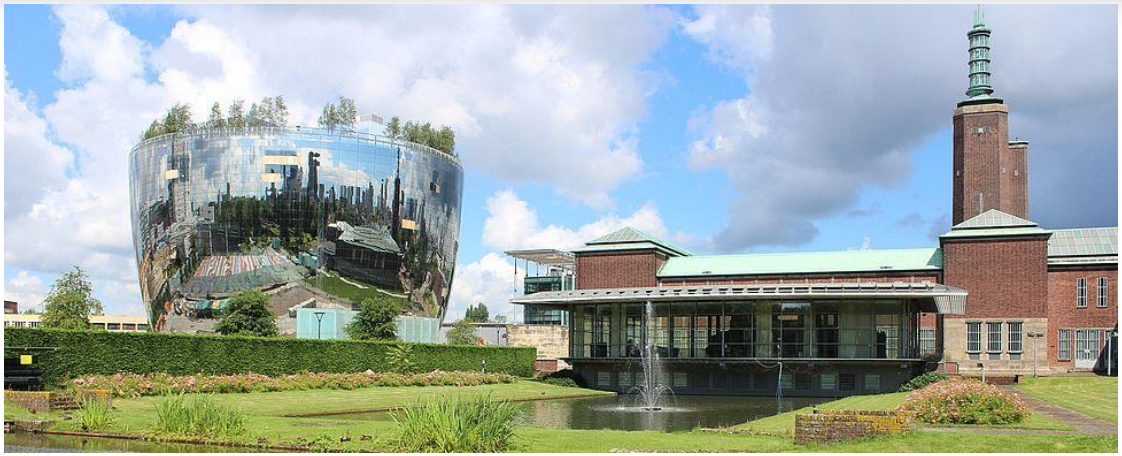
FAIR data assessment

open source

Acceptance...

An analogy from the museum world:





Depot Museum Boijmans van Beuningen, Rotterdam

The world's first publicly accessible art storage facility

https://www.boijmans.nl/en/depot

FAIR Data Review Form (manual): https://tinyurl.com/ehhdtw3u
Automated FAIR data assessment (F-UJI): https://www.f-uji.net

www.dans.knaw.nl
peter.doorn@dans.knaw.nl
l.breure@uu.nl

Some slides prepared by Leen Breure

gesis

Leibniz-Institut
für Sozialwissenschaften

Replikationsserver.de: A GESIS service for publishing replication packages

Workflows, lessons learned, and a look ahead

Dr. Jonas Recker, GESIS – Leibniz Institute for the Social Sciences

Leibniz Gemeinschaft

gesis Leibniz-Institut für Sozialwissenschaften

How it started vs How it's going

- Initiative Replikationsserver.de launched in 2015
  - Cooperation between Zeitschrift für Soziologie, Soziale Welt, and GESIS – Leibniz Institute
- Preceded by collaborative development of policies, (user) guidelines and workflows

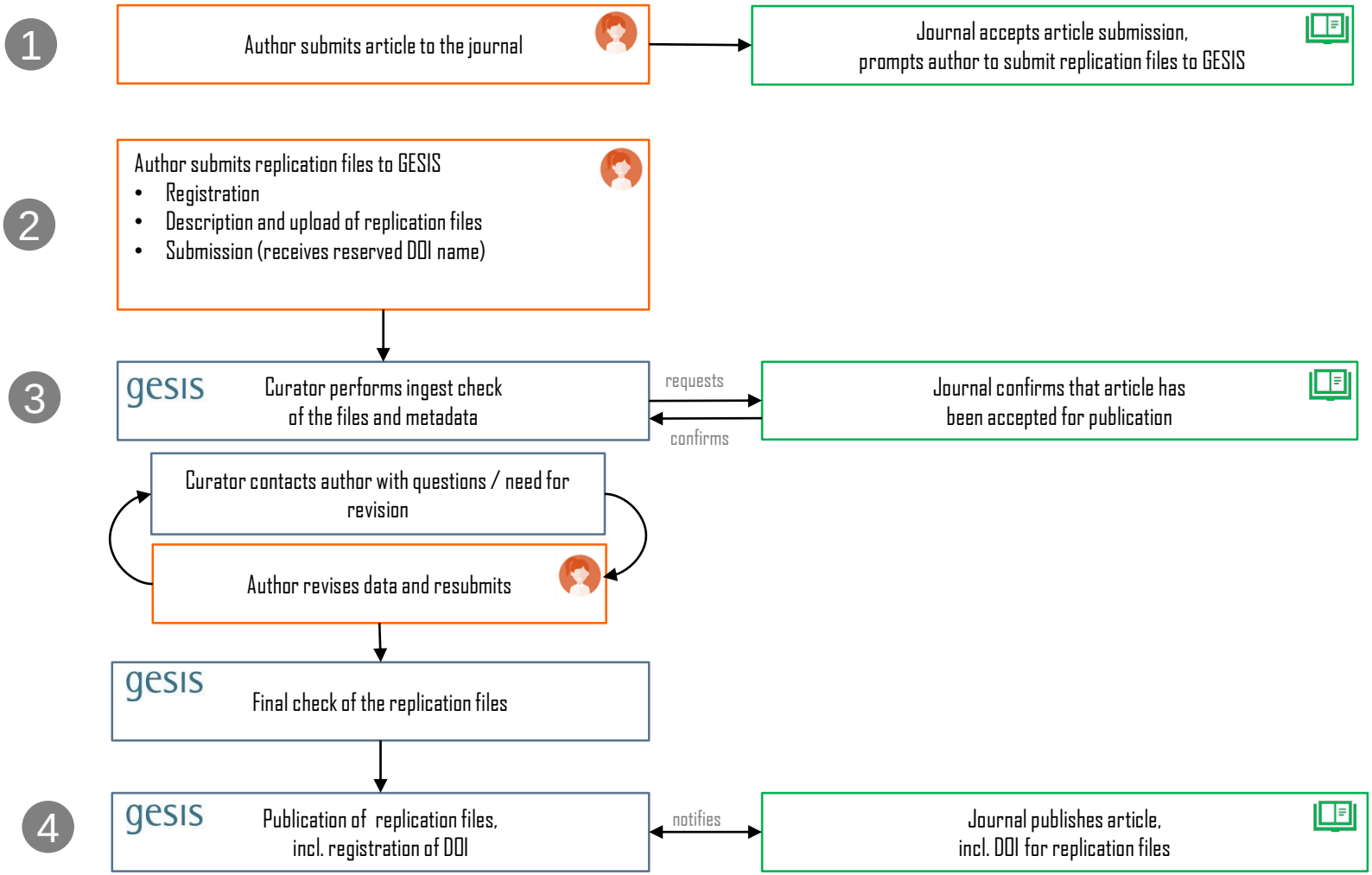
- Three active journals
- Three journals interested or in the process of implementation
- ~ 50 replication packages published (data, code, supplements)

GESIS provides technical infrastructure for upload and publication of data, incl. basic data curation and preservation measures

Information on the initiative: https://www.gesis.org/en/replikationsserver/home
Data upload and access: https://data.gesis.org/sharing/

gesis Leibniz-Institut für Sozialwissenschaften

Workflow



Example "Zeitschrift für Soziologie"

- 2015: Editorial announced new practice[1]
- 2016 onward: Mandatory submission of replication files to a repository for all manuscripts based on quantitative data (GESIS repository recommended)
- 2019: Evaluation of publication practice[2,3]
  - Increasingly good compliance on behalf of authors
  - Data and/or code per article downloaded >8 times on average (0 to 33 downloads)
  - No recognizable negative effect on submissions or journal impact

[1]Zeitschrift für Soziologie, Jg. 44, Heft 1, Februar 2015, S. 2–5

[2]Auspurg, Katrin & Jonas Recker, 2020: Mehr Offenheit in der Forschung? Eine Evaluation von Open Science Maßnahmen bei der Zeitschrift für Soziologie. Zeitschrift für Soziologie. 49(1): 1-9. https://doi.org/10.1515/zfsoz-2020-0001

[3]Auspurg, Katrin & Jonas Recker, 2020: Daten: Mehr Offenheit in der Forschung? Eine Evaluation von Open Science Maßnahmen bei der Zeitschrift für Soziologie. https://doi.org/10.7802/1992

Lessons learned

- It takes time!
  - Reaching consensus among editors, creating policies and guidelines, announcing new policies to authors may take several years
- It works!
  - Data/code are published and re-used
  - Many authors are motivated to publish well-documented data
  - But: Mandates more effective than recommendations
- ZfSoz evaluation
  - Restricted access creates problems for replication
  - "Data notes" to ensure that references to data is included

Outlook

- Challenges
  - Competing with commercial / generic data publication platforms
- Plans for 2022
  - Foster exchange and discussion on publication of replication packages with German journals in the social and economic sciences
  - Explore how we can make full(er) use of our technical platform's capabilities (e.g. embed a journal's data collection in the journal website, communication and review workflows, etc.)

Contact:

gesis
Leibniz-Institut
für Sozialwissenschaften

Leibniz
Gemeinschaft

Insufficiencies in Data Material or How (Not) to Replicate

—

A Walk-Through of a Real Replication Case for Political Analysis

Simon Heuberger
(with R. Michael Alvarez)

11 November 2021

Outline

- Replication crisis (*Nature*, *Science*)
  - Journals: Require and run code before publication
  - Authors: Provide organized, usable material
- Muchlinski, Siroky, He, and Kocher (2016). *Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data*
  - PA was contacted by researchers about irregularities in code
  - PA conducted in-house replication in 2018
  - PA published critique letters, in-house replication, MSHK response
- Increase in research transparency
  - Shift needed from data/code requirements to actual execution

Research Transparency Requirements in Top Political Science Journals

	Accessible policy*	Collective archive†	Data/Code required‡	Replicated #
AJPS	✓	✓	✓	✓
APSR	✓	✓	✓	✗
BJPS	✓	✓	✓	✗
EJPR	✓	✗	✓	✗
JOP	✓	✓	✓	✗
PA	✓	✓	✓	✓
POQ	✓	✗	✓	✗
PRQ	✓	✗	✓	✗
PS	✓	✓	✓	✗
QJPS	✓	✗	✓	✗

* Journal research transparency policy is easily accessible on the journal website

† Journal archive contains all data submissions and is permanently publicly available

‡ Provision of data and code is required prior to publication

\# Code is run during review to verify manuscript results

Author Requirements

- Our experience at *Political Analysis*: highly disorganized, unusable material still very much the norm
  - No basic documentation (README)
  - No master file
  - Local working directories
  - No saved output for figures/tables
  - No code comments
  - No computational requirements and running times
- Out of almost 100 replication data sets submitted to PA in the last two years, all except one suffered from at least one of these shortcomings

Timeline of Paper Replication

- MSHK publish paper in 2016 (PA does not run code)
- Two sets of researchers send critique letters to PA in 2018
- PA conducts in-house replication, notifies MSHK
- MSHK send updated code (March 2018), which is insufficient
- PA informs MSHK of publication of letters and replication
- MSHK send another updated code (June 2018), which is also insufficient
- PA publishes letters, replication, heavily redacted authors' response in 2019

Code Insufficiencies

- `RandomForest`: Machine learning to construct multiple decision trees to obtain more accurate predictions
- RF model needs to be trained on data sample to predict observations
- Training sample and prediction sample must not be the same

```
data = read.csv(file="SambnisImp.csv")
data.full <- data[,c("warstds", "ager", "agexp", ...)]

model.rf <- train(as.factor(warstds) ~ .,
                  metric = "ROC", method = "rf",
                  sampsize = c(30,90), importance = T,
                  proximity = F, ntree = 1000, trControl = tc,
                  data = data.full)

RF.out <- randomForest(as.factor(warstds)~., sampsize = c(30, 90),
                       importance = T, proximity = F,
                       ntree = 1000, confusion = T, err.rate = T,
                       data = data.full)

yhat.rf <- predict(RF.out, type = "prob")
Yhat.rf <- as.data.frame(yhat.rf[,2])
predictors.rf <- Yhat.rf[sample(nrow(Yhat.rf), 737), ]
pred.rf.africa <- prediction(predictors.rf, data3$warstds)
auc.rf.africa <- performance(pred.rf.africa, "auc")
```

Output for Main Evidence

- Table 1 as main evidence for claimed superiority of RandomForest
- Table lists predicted probabilities for civil war onset for 19 African countries
- MSHK provide `CompareCW_dat.csv` as output that forms Table 1

Table 1 Predicted probability of civil war onset: Logistic Regression and Random Forests

	Models and predicted probability of civil war onset			
Civil war onset	*Fearon and Laitin (2003)*	*Collier and Hoeffler (2004)*	*Hegre and Sambanis (2006)*	*Random Forests*
Afghanistan 2001	0.01	0.01	0.01	0.09
Angola 2001	0.04	0.01	0.01	0.13
Burundi 2001	0.00	0.00	0.00	0.05
Guinea 2001	0.00	0.00	0.01	0.22
Rwanda 2001	0.02	0.00	0.00	0.56
Uganda 2002	0.03	0.05	0.00	0.81
Liberia 2003	0.01	0.03	0.00	0.94
Iraq 2004	0.04	0.01	0.00	0.68
Uganda 2004	0.02	0.01	0.02	0.52
Afghanistan 2005	0.01	0.02	0.01	0.14
Chad 2006	0.01	0.07	0.02	0.21
Somalia 2007	0.00	0.00	0.00	0.52
Rwanda 2009	0.00	0.01	0.00	0.74
Libya 2011	0.00	0.01	0.00	0.34
Syria 2012	0.00	0.04	0.00	0.25
DR Congo 2013	0.00	0.00	0.00	0.76
Iraq 2013	0.01	0.00	0.00	0.25
Nigeria 2013	0.01	0.00	0.00	0.25
Somalia 2014	0.01	0.04	0.01	0.87

CompareCW_dat.csv

Filter

	data3.warstds	predictors.fl	predictors.ch	predictors.hs	predictors.rf
1	0	0.036354228	9.380534e-03	0.0209075119	0.29216867
2	0	0.009281192	3.714911e-03	0.0003546772	0.08926780
3	0	0.017708231	8.841568e-03	0.0204368389	0.53253253
4	0	0.034197142	3.136079e-03	0.0107442810	0.19759278
5	1	0.078753771	1.350169e-01	0.0597720247	0.06048387
6	0	0.025539793	1.478751e-02	0.0006688022	0.08625878
7	1	0.013429648	1.852500e-01	0.0073623881	0.71887550
8	0	0.027755230	5.513324e-03	0.0080306531	0.56493506
9	1	0.012395045	5.857035e-03	0.0311944315	0.63453815
10	0	0.006440910	1.213052e-03	0.0475583848	0.67439516
11	1	0.007812825	1.367884e-02	0.0077101553	0.51351351
12	0	0.002278077	5.956352e-03	0.0085521977	0.19719720
13	0	0.011487555	6.777899e-02	0.0045929770	0.84100000
14	0	0.021308705	2.346514e-03	0.0328846433	0.74274274
15	0	0.003904838	9.912065e-03	0.0326538764	0.27437186
16	0	0.008834318	3.621130e-02	0.0023257426	0.10531595
17	0	0.009155091	4.078427e-07	0.0013544994	0.31795386
18	0	0.005292905	6.455024e-02	0.0083213129	0.46339017
19	0	0.011793761	1.542060e-03	0.0121889713	0.27227227
20	0	0.004634562	9.332847e-03	0.0020158494	0.44567404
21	0	0.043097064	6.491715e-03	0.1029864243	0.03006012
22	1	0.005371223	2.583362e-03	0.0351988318	0.71385542
23	0	0.005423841	8.191679e-03	0.0511781647	0.18355065
24	0	0.009828007	1.048320e-02	0.0019674098	0.51055276
25	0	0.009263872	3.037736e-02	0.0149661627	0.06325301
26	0	0.008514706	3.958729e-02	0.0047936840	0.65361446
27	0	0.002212609	2.662163e-02	0.0152644593	0.04742684

Summary of MSHK Paper and Replication

- 2016 code:
  - In-sample predictions
  - Unusuable `.csv` output for main table
- March 2018 code:
  - Changed code for predictions but still in-sample
  - Loads different data files with differing dimensions
  - No output for main table
- June 2018 code:
  - Still in-sample predictions
  - Loads data files from 2016 version
  - Suspicious `.csv` output for main table

Looking Ahead

- Shift needed from data/code requirements to actual execution/evaluation
- Practices to adopt to help resolve the crisis
  - Journals need to run provided material
  - Authors need to start their work with replication in mind
- Dataverse-style full of potential problems. The future? Docker containers
  - Virtual, self-contained computer accessed through browser
  - Users install software, upload data, run code in remote container online
  - Eliminates environment mismatch
  - Ensures full replicability
  - Increases efficiency and effectiveness

Thank you!

Taylor & Francis Group
an **informa** business

TOP Guidelines at Taylor & Francis

Matt Cannon – Head of Open Research

11th November 2021

Summary

Introduction and background of TOP guidelines

Previous interactions

Journal case study

Next steps

Introduction to TOP Guidelines

- **TOP – Transparency and Openness Promotion**
- Developed by Centre for Open Science in 2015

- Eight modular standards, each with three levels. Guides journals how to share guidelines in instructions for authors.

- Have had over 5,000 signatories

- Signatories can be individuals, journals, organisations

Introduction to TOP Guidelines

	0	1	2	3
Data citation	No mention of data citation.	Journal describes citation of data in guidelines to authors with clear rules and examples.	Article requires appropriate citation for data and materials used consistent with the journal's author guidelines.	Article is not published until providing appropriate citation for data and materials following journal's author guidelines
Details	This section refers to already existing datasets. Rationale is to incentivize publishing of them and to treat them as citable contributions to scholarship.	"All data, program code and other methods should be appropriately cited. Such materials should be recognized as original intellectual contributions and afforded recognition through citation."	"All data, program code and other methods **must** be appropriately cited"	"**Articles will not be published until the citations conform to these standards.**"
Data transparency	Data sharing is encouraged, or not mentioned	Articles must state whether or not data are available.	Articles must have publicly available data, or an explanation why ethical or legal constraints prevent it.	Articles must have publicly available data and must be used to computationally reproduce or confirm results prior to publication
Details	Level 0 applies if the journal policy does not cover all of the underlying data reported in an article.	Requiring a data availability statement satisfies this level.	If the journal only requires some data to be preserved, e.g. proteomics, then this level is not reached. If the article must include an availability statement for all other data, then level 1 is reached, else level 0. If the policy strongly encourages data sharing, this level is not reached. Policies that require sharing with editors and reviewers only do not apply. Policies that require sharing only "upon request" do not apply. Words such as "should" or "expect" may be ambiguous. Typically, "should" implies an encouragement and not a requirement. "Expects" suggests a requirement, but clarification may be needed.	Policy must cover transparency and sharing requirements of level 2, plus include a computational reproducibility step.

TOP Factor

- In 2020 COS announced TOP Factor
- Scores journals based on the rubric

- **“The TOP Factor measures something that matters. It compares journals based on whether they require transparency and methods that help reveal the credibility of research findings.”**

- Evan Mayo-Wilson, Associate Professor in the Department of Epidemiology and Biostatistics at Indiana University School of Public Health-Bloomington.

- Can search publications and see how they score, filter by standards, publisher or disciplines

TOP Factor at Taylor & Francis

- 140+ journals included in TOP Factor rankings
- Top score of 23 (*Comparative Results in Social Psychology*)
- Majority are social sciences
  - Psychology
  - Behavioural science
  - Communication studies
  - Education
  - Also, criminology, politics, area studies, religion

- Taylor & Francis became an organisational signatory in 2019.

TOP Factor at Taylor & Francis

- In working to make our journals more transparent and reproducible
  - Adding open science badges
  - Offering registered reports
  - Data sharing policy framework

TOP Factor at Taylor & Francis – next steps

- Working with an education journal to encourage authors to make a full declaration of how they have met all TOP areas.
- Still in early stages and not all details are confirmed yet.
- Journal has shown its commitment by naming a reproducibility editor as part of the editorial team
- Authors will be asked to submit a form with a series of statements showing how they have complied with all the areas of TOP guidelines. Reproducibility editor will check and work with authors to finesse
- Statements will be included in the final version of the article, and outside the paywall so non-subscribers can access
- Launching in 2022

Questions?

Taylor & Francis Group
an **informa** business

Thank you

Informa.com