

# 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, Sciimedia*

*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](https://cessda.eu)

 [@CESSDA\\_Data](https://twitter.com/CESSDA_Data)

# Making Social Science Research Transparent

Part 1: Introduction. Conceptual basis

Part 2: Transparency in practice

---

Online workshop

*11 November 2021*

 [cessda.eu](https://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](https://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>)

### **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](#))

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)

# 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

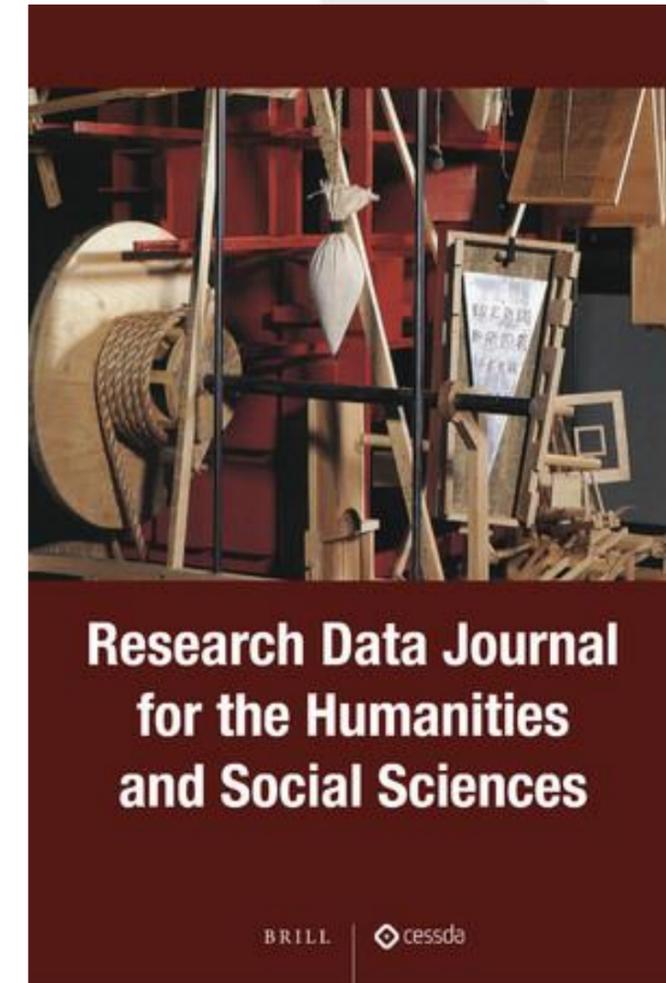
## Data description can be more or less structured:

- 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. Accessible at: <https://www.openaire.eu/briefpaper-rdm-infonoads> (8th November 2021).



<https://www.re3data.org/>

# Consortium of Social Science Data Archives - CESSDA

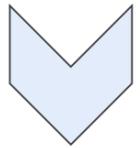
Provides a distributed and sustainable **research infrastructure**, that

- 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

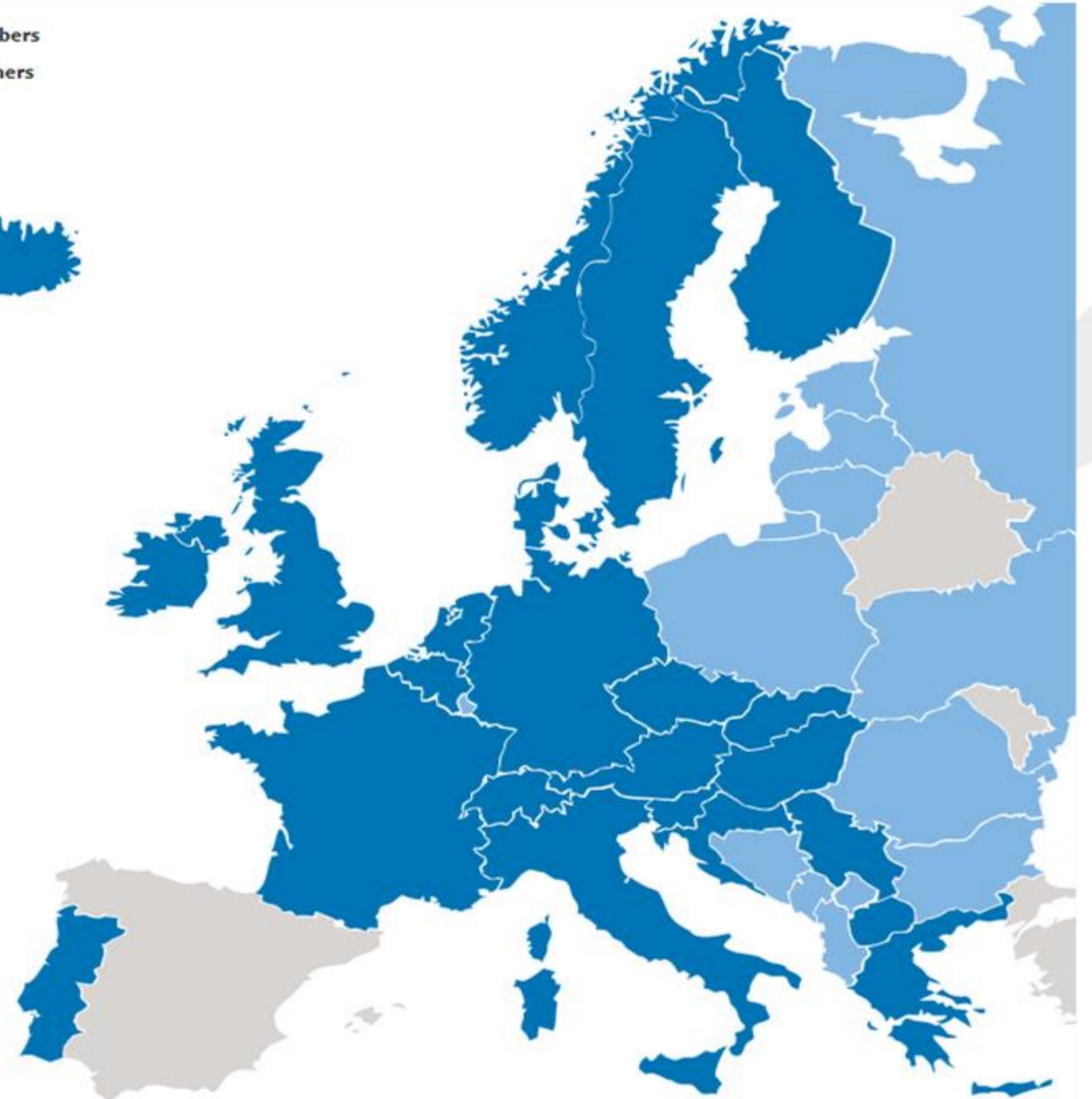
### Members

Austria  
Belgium  
Croatia  
Czech Republic  
Denmark  
Finland  
France  
Germany  
Greece  
Hungary  
Iceland  
Ireland  
Italy  
Netherlands  
North Macedonia  
Norway  
Portugal  
Serbia  
Slovakia  
Slovenia  
Sweden  
Switzerland (Observer)  
United Kingdom

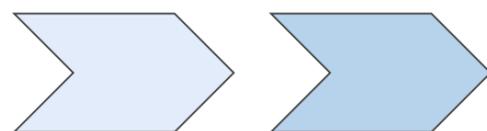
### Partners

Albania  
Bosnia and Herzegovii  
Bulgaria  
Estonia  
Kosovo  
Latvia  
Lithuania  
Luxembourg  
Montenegro  
Poland  
Romania  
Russia  
Ukraine

Members  
Partners

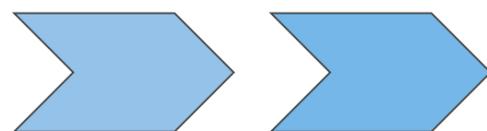


# 2 WGs under CESSDA Training WG



## 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 to **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>

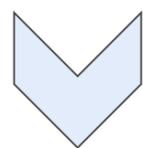
- Data Management Expert Guide ▾
  - 1. Plan >
  - 2. Organise & Document >
  - 3. Process >
  - 4. Store >
  - 5. Protect >
  - 6. Archive & Publish ▾
    - Towards archiving & publication
    - Selecting data for publication
    - Data publishing routes
    - Publishing with CESSDA archives
    - Citing your data
    - Licensing your data
    - Access categories
    - Promoting your data
    - Adapt your DMP: part 6
    - Sources and further reading
  - 7. Discover >
  - 8. Contributors >

# 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.**



## **F**indable

To aid automatic discovery of relevant datasets, (meta)data should be easy to find by both humans and machines and be assigned a persistent identifier.

## **A**ccessible

Limitations on the use of data, and protocols for querying or copying data are made explicit for both humans and machines.

## **I**nteroperable

(Meta)data should use standardised terms (controlled vocabularies), have references to other (meta)data and be machine actionable.

## **R**eusable

(Meta)data are sufficiently well described for both humans and computers to be able to understand them and have a clear and accessible data usage license.

# 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.

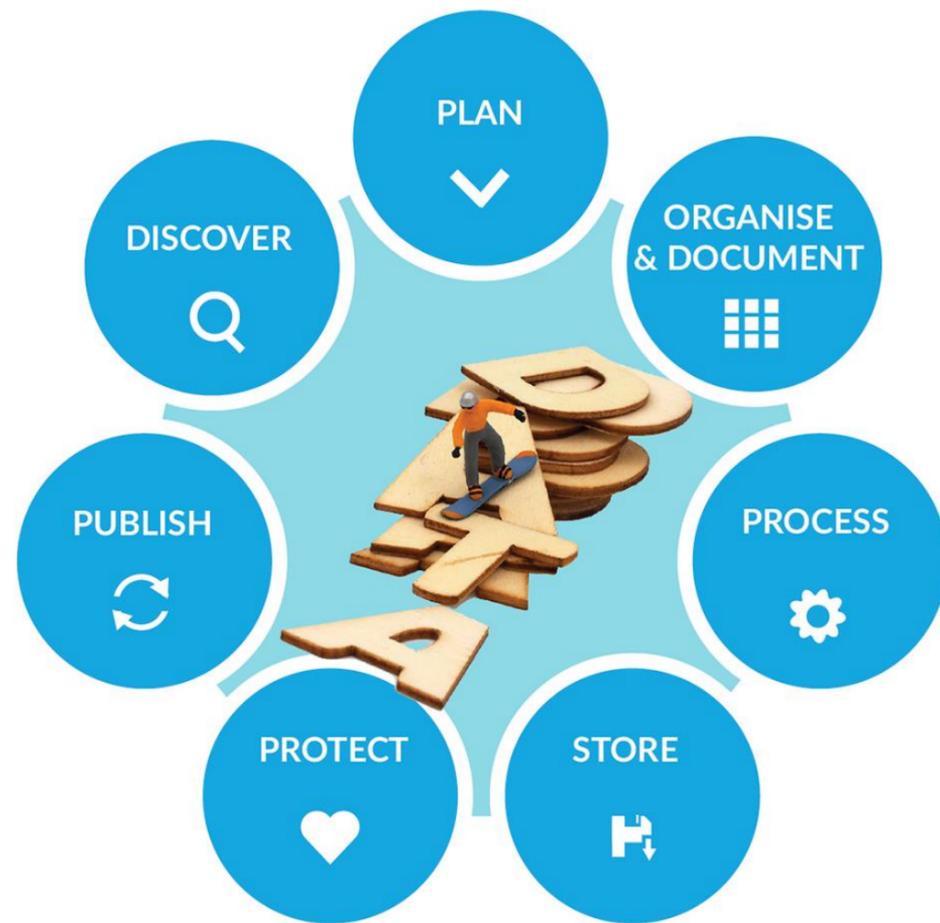
Responsible management of research data in line with the FAIR principles of 'Findability', 'Accessibility', 'Interoperability' and 'Reusability', notably through the generalised use of data management plans, and open access to research data under the principle 'as open as possible, as closed as necessary', 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](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](mailto:sonja.bezjak@fdv.uni-lj.si)*

*Sergeja Masten, [sergeja.masten@fdv.uni-lj.si](mailto:sergeja.masten@fdv.uni-lj.si)*

 [cessda.eu](http://cessda.eu)

 @CESSDA\_Data

  Licence: CC-BY 4.0

Making Social Science Research Transparent

# Elements of Research Transparency

---

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

*11 November 2021*

 [cessda.eu](https://cessda.eu)

 [@CESSDA\\_Data](https://twitter.com/CESSDA_Data)



Licence: CC-BY 4.0

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



# [before] Transparent methods

## **SUBJECTIVITY and FREEDOM**

questions

measures

methods

procedures

analyses



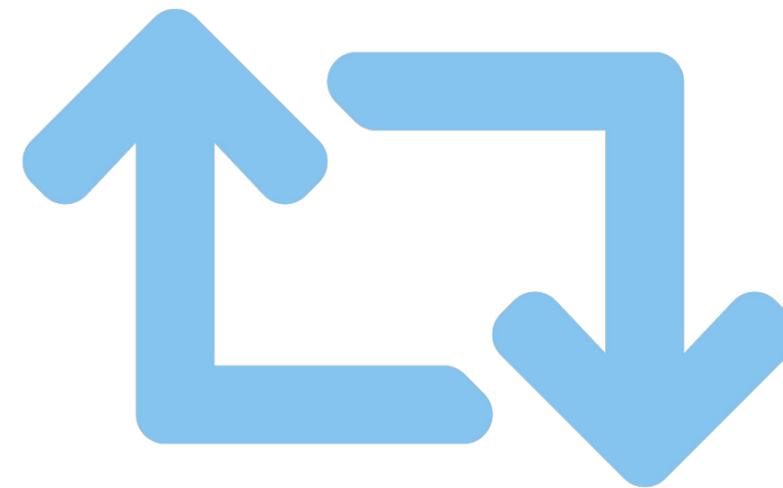
# [before] Transparent methods

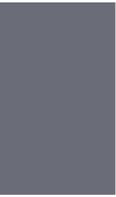
## preregistration

- 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

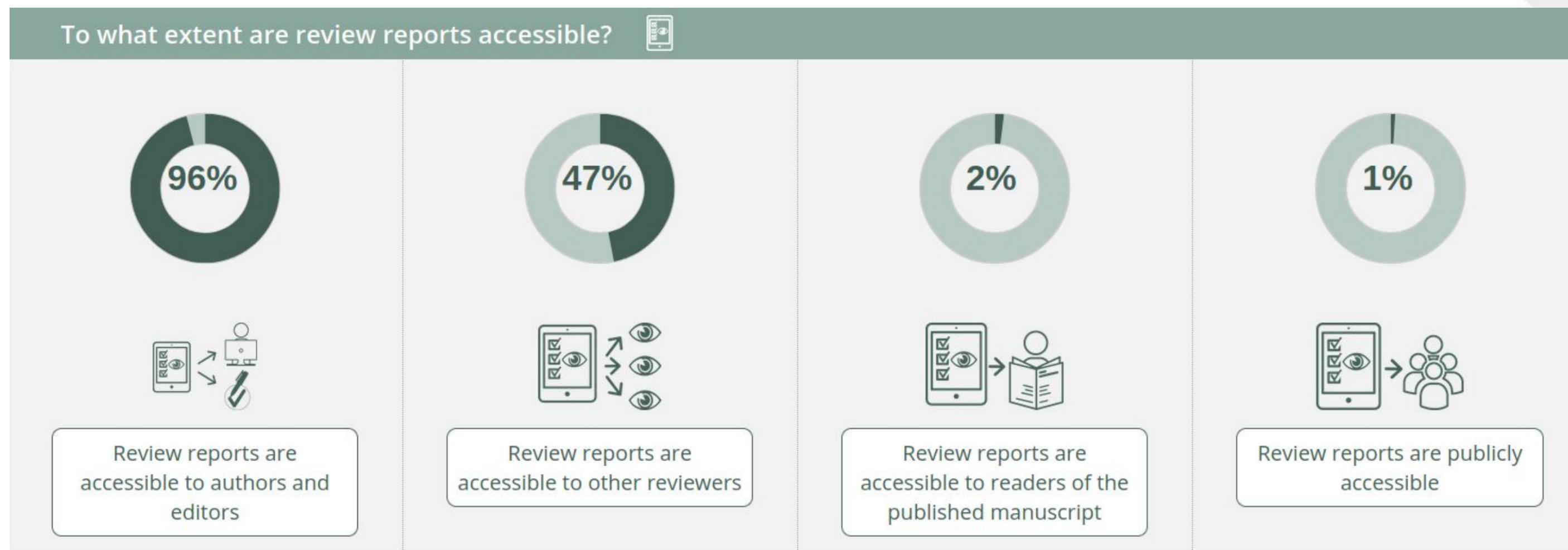


**PUBPEER**  
The online Journal club

# [after] Open peer review

## THE PLATFORM FOR RESPONSIBLE EDITORIAL POLICIES

<https://www.responsiblejournals.org>



# [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 <dvlasice@ffzg.hr | @dvlasicek>  
CROSSDA @crossda\_data

 [cessda.eu](https://cessda.eu)

 @CESSDA\_Data



Licence: CC-BY 4.0



# **Making Social Science Research Transparent**

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

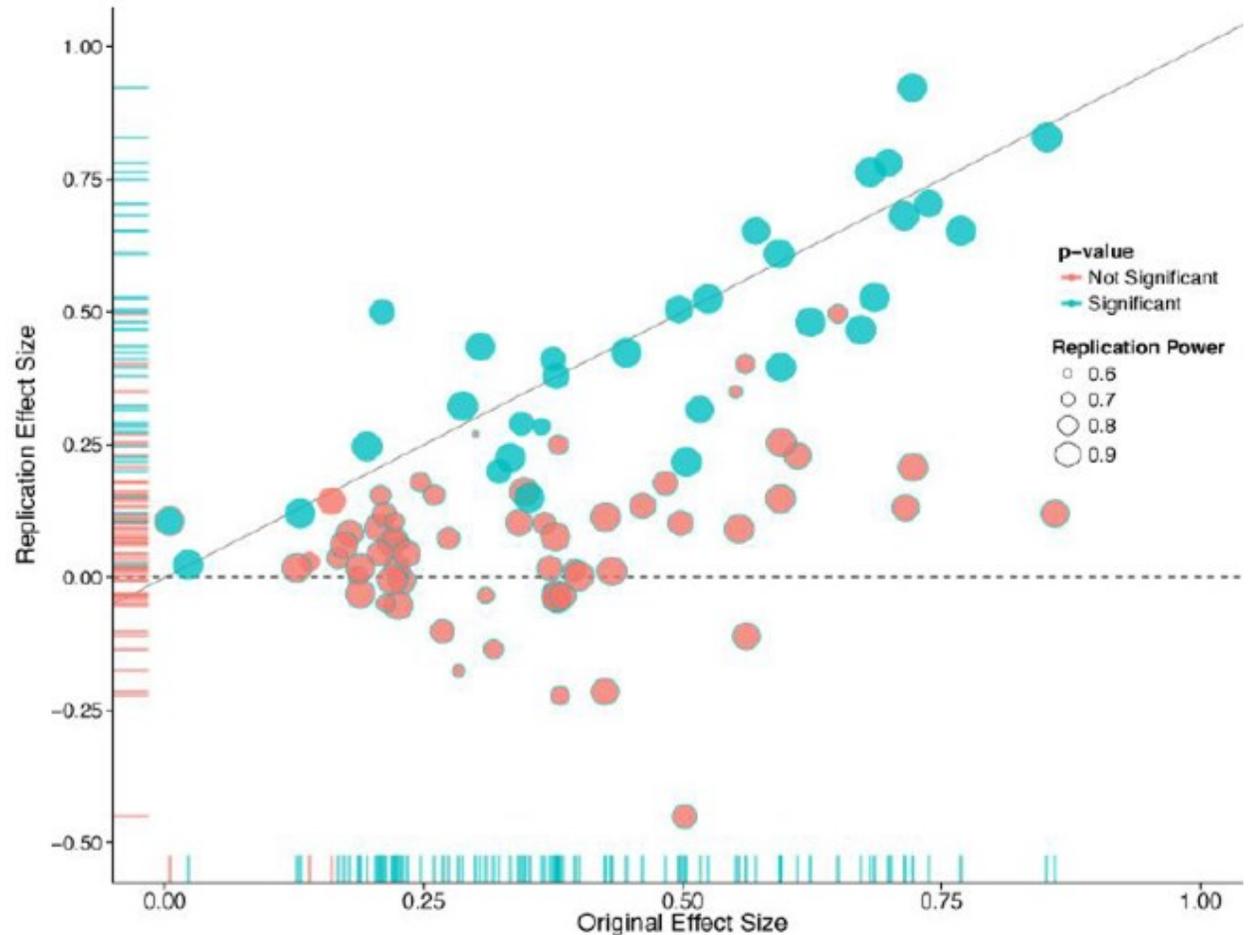
# 1. Rationales (in Psychology)

# Replication crisis and trust in science



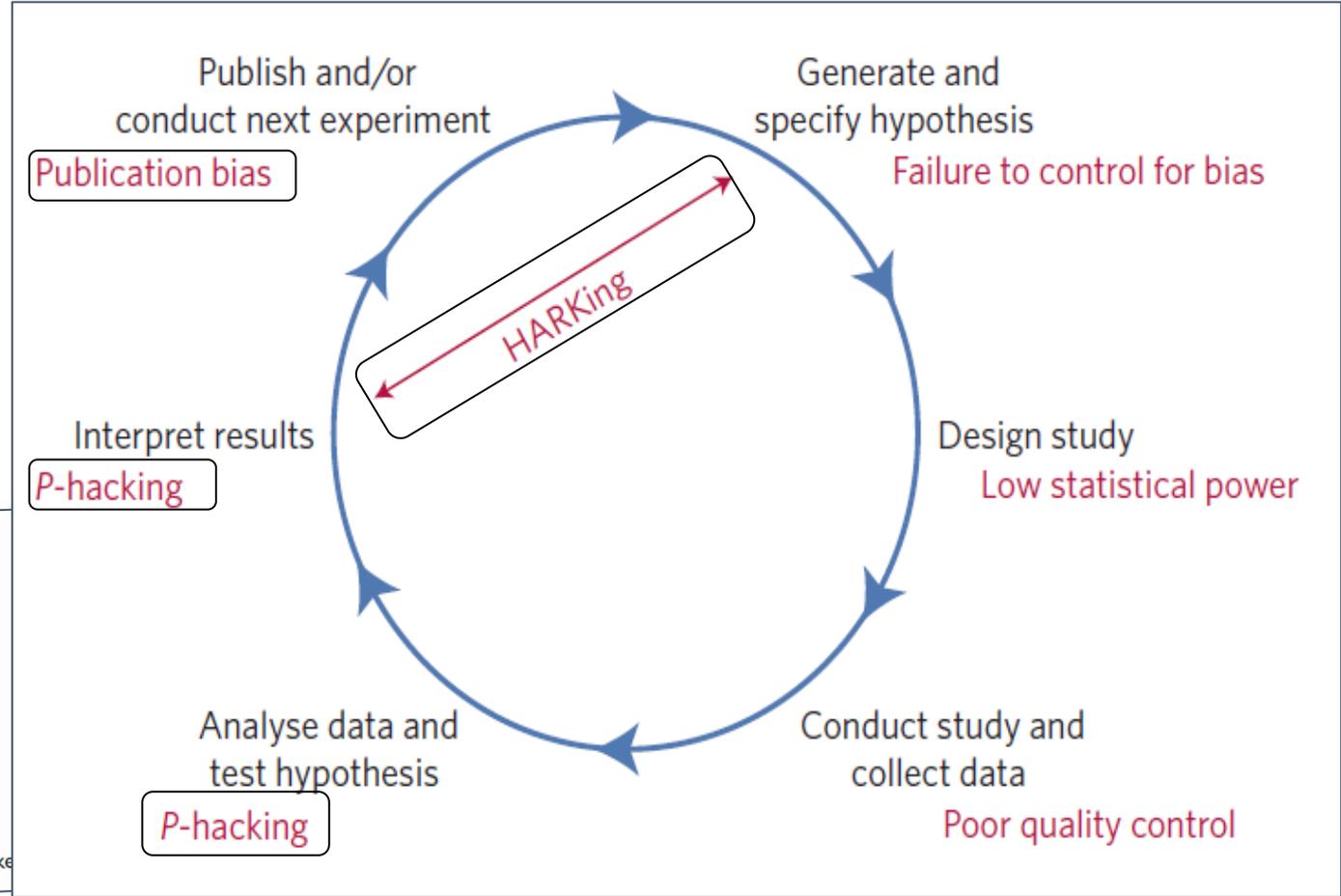
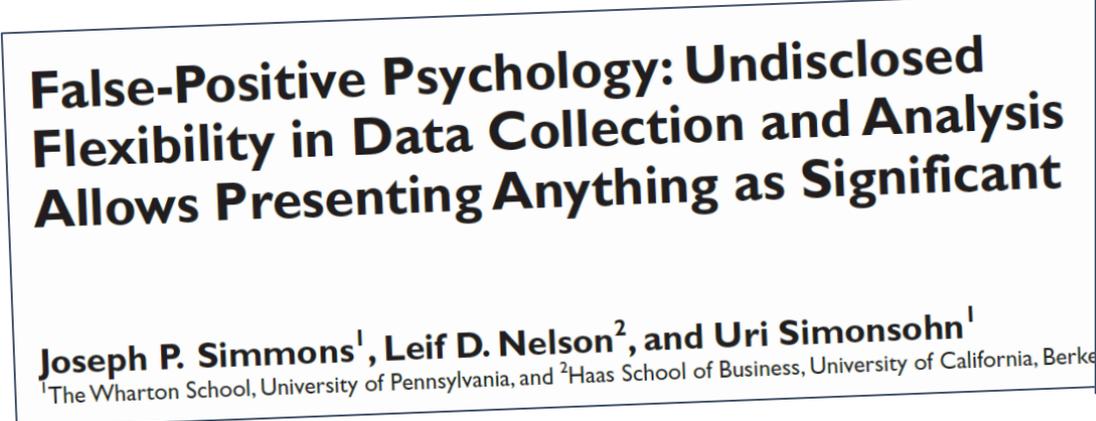
## **Replication, noun:**

Purposeful repetition of a study to assess the reliability and generalizability of findings



Source: Open Science Collaboration (2015)

# Questionable research practices

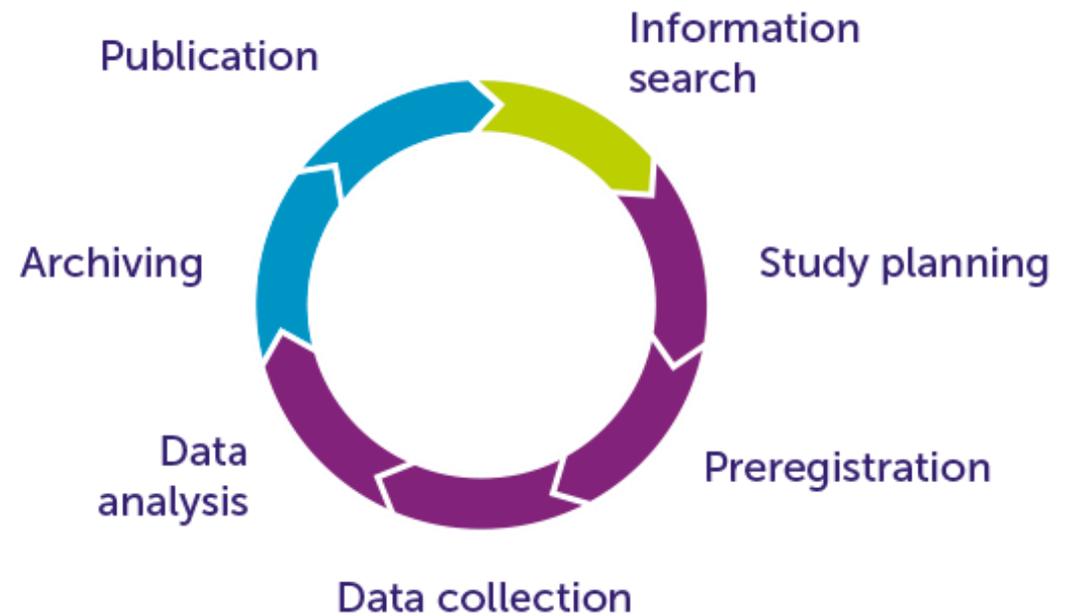


Source: Munafò et al. (2017)

# 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

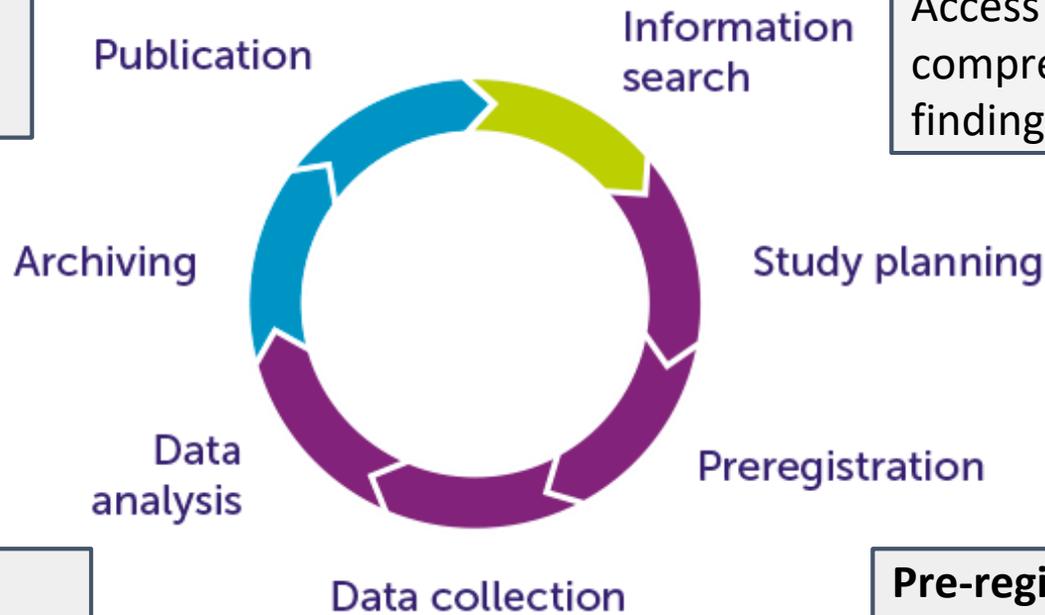
Open access publishing opportunity, accelerated synthesis

## FAIR DROs

Licensing and support for reusable 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





<https://www.psyndex.de/>

# PSYNDEX

[Startseite](#) [Komfort-Literatursuche](#) [PSYNDEX Tests](#) [PSYNDEX Interventions](#) [Themen & Trends](#) [Hilfe & Angebote](#) [Über PSYNDEX](#)

Sie sind hier: [Startseite](#)

**PSYNDEX** - die Datenbank des [ZPID](#) für Publikationsnachweise psychologischer Fachliteratur aus dem deutschsprachigen Raum - inklusive redaktionell beschriebener Testinstrumente und Interventionsprogramme.

Die Grundfunktionen sind mit dem Suchportal [PubPsych](#) kostenlos nutzbar, Profifunktionen mit [abonnierbaren Instituts-Recherchediensten](#) wie OVID, EBSCO und wiso.

Gibt es etwas in PSYNDEX zu ergänzen?

[Einzelpublikation melden](#)

[Zeitschrift vorschlagen](#)

## Schnellsuche in PSYNDEX

(öffnet neues Fenster mit Trefferliste in PubPsych)

[Gesamtsuche](#) [Tests](#) [Interventions](#)

Die gesamte Fachliteratur in PSYNDEX mit PubPsych durchsuchen:

Schlagwörter zur Fachliteratursuche eingeben

Suchen

## i Über PSYNDEX

Erfahren Sie mehr zur Aufnahme und Erfassung von Publikationen und wie Sie PSYNDEX nutzen können.

- **Steckbrief:** Was ist PSYNDEX?
- **Aufnahme:** Welche Publikationen nehmen wir auf?
- **Inhalte & Aufbau:** Wie beschreiben wir die in PSYNDEX aufgenommenen Publikationen?

## Information search



PubPsych

<https://www.pubpsych.de/>

En | Es | Fr | **De**

→ **Startseite**

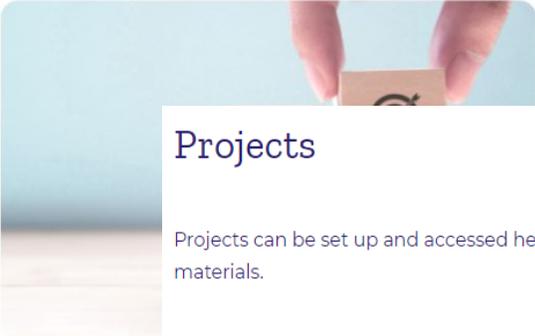
→ Hilfe



→ Über PubPsych → Nutzungsbedingungen → Kontakt → Impressum & Datenschutz



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



### Projects

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

My projects | Shared projects | Public projects

### Our mission

Providing you with the tools to practice open science!

[Import project](#) [New project](#)

### Features

- Log in and st...
- Share your pr...
- Create, copy i...
- 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...	





Preregistration & ZPID Track Types ▾ Submission Information ▾ FAQ

## Preregistration in Psychology

By using preregistration, researchers can verify that their studies have been conducted, analyzed, and reported as initially specified.

ZPID promotes preregistration via the **PreReg in Psychology** platform by offering a domain-specific repository ([Repository Track](#)) and, additionally, free-of-charge data collection for high-quality preregistrations in the field of psychology ([Lab Track](#)).

Preregistrations submitted to **PreReg in Psychology** will be archived in [PsychArchives](#), ZPID's repository for psychological science, and, to become citable, each will receive a timestamp and be assigned a DOI (digital object identifier).

All accepted [Lab Track](#) submissions will receive complimentary data collection via the [PsychLab](#) service, which provides either the funding for an online study sample or the opportunity to outsource the data collection of an eye tracking (or any PC-based) study to our on-site lab in Trier.

## 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.



# Scope Online Lab (surveys & online experiments)





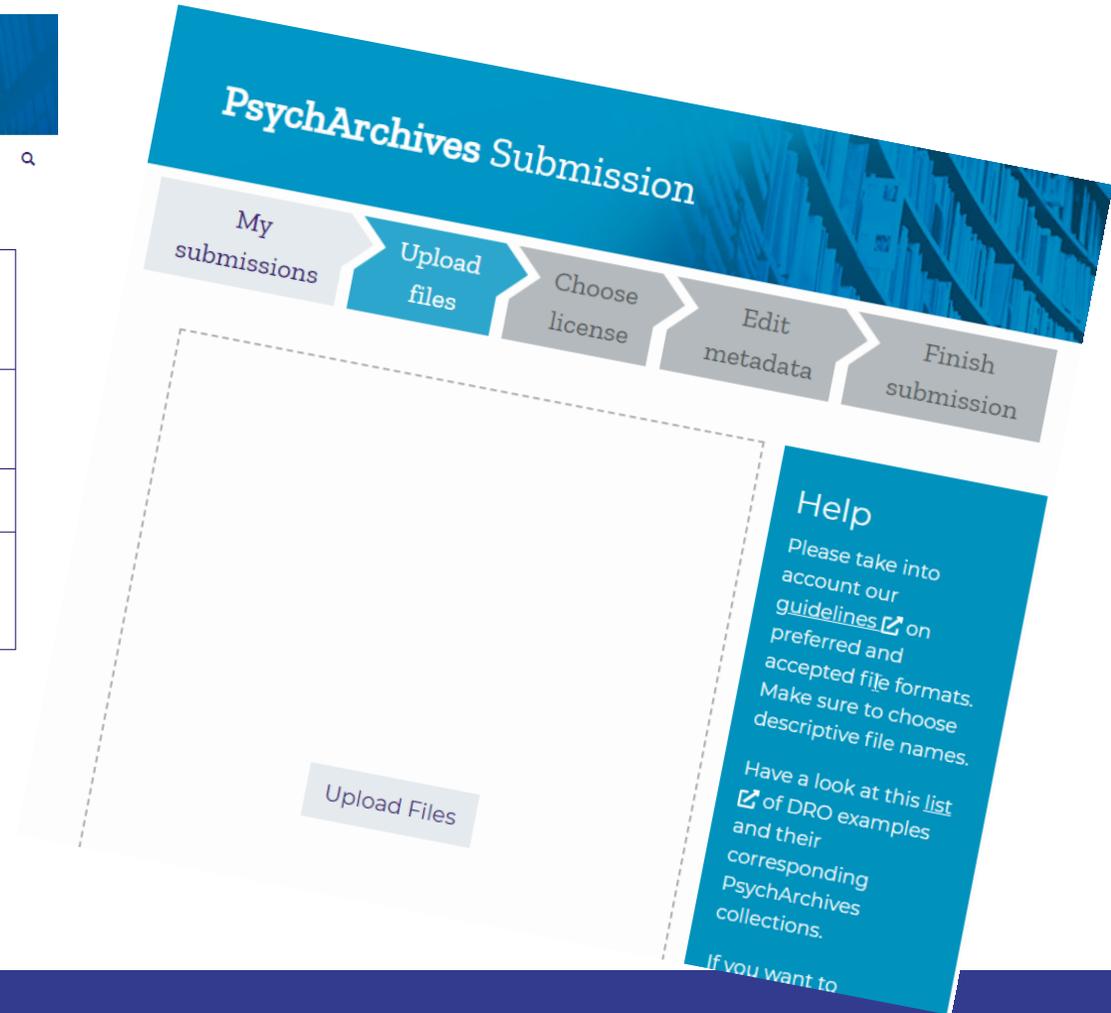
# Scope Offline Lab (Eye tracking or any PC-based experiments)



<https://www.psycharchives.org/>

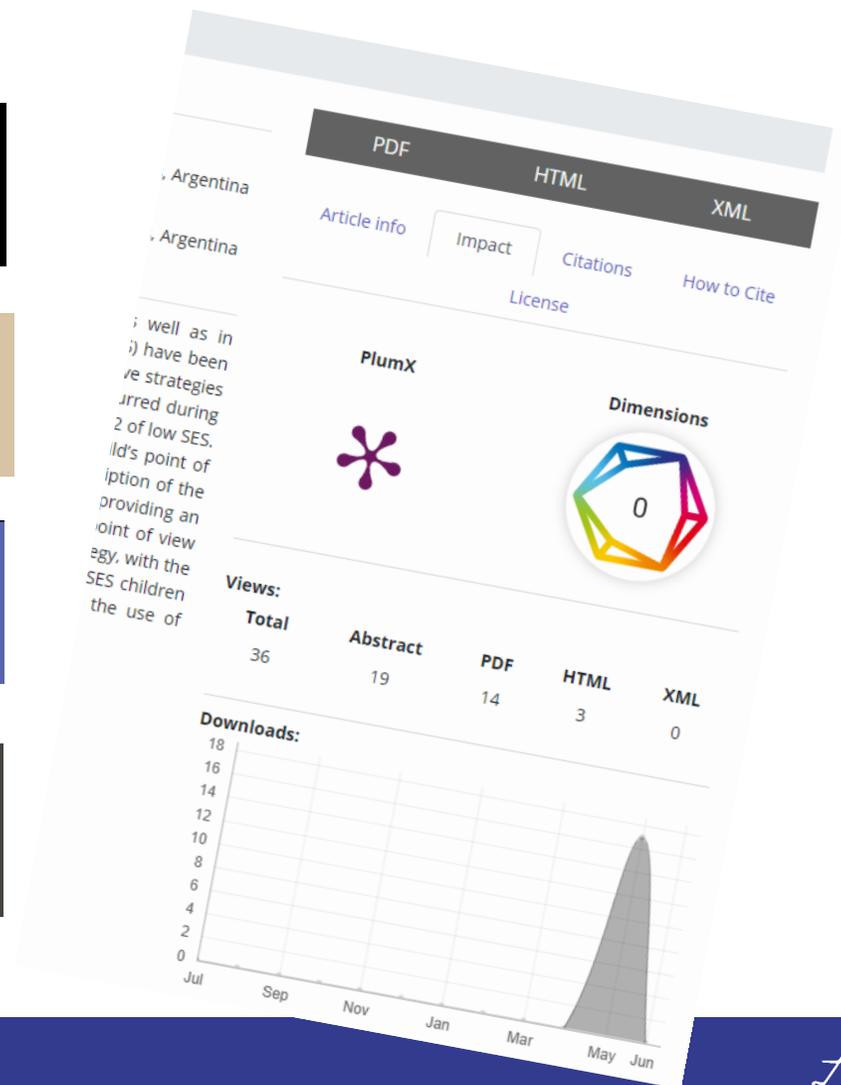
The screenshot shows the PsychArchives website homepage. At the top, there is a blue header with the PsychArchives logo and navigation links: Home, Contribute, Browse, and Info. Below the header, the main content area is divided into several sections:

- Disciplinary Repository for Psychological Science:** A section describing PsychArchives as a repository for digital research objects (DROs) with 20 different publication types. It includes a search bar with the text "Please find all content regarding COVID-19 Snapshot Monitoring (COSMO Standard) in a special collection." and a "Search" button.
- Highlighted Submissions:** A section featuring a submission titled "COVID-19 Snapshot Monitoring (COSMO Standard): Monitoring knowledge, risk perceptions, preventive behaviours, and public trust in the current coronavirus outbreak - WHO standard protocol (WHO Regional Office for Europe)" with a date of 2020-03-17 and a "preregistration" tag.
- Recent Submissions:** A section featuring two submissions: "Dataset: Attentional Capture in Multiple Object Tracking (Pichlmeier et al.)" with a date of 2021-06-08 and a "researchData" tag, and "Code for: Attentional Capture in Multiple Object Tracking (Pichlmeier et al.)" with a date of 2021-06-08 and a "code" tag.
- PsychArchives in a nutshell:** A sidebar section with four bullet points:
  - Shared Digital Research Objects (DRO):** - research outputs from the entire cycle of psychological research are welcome.
  - Citable and discoverable:** - uploads are assigned a Digital Object Identifier (DOI) to make them easily and uniquely citable.
  - Open licensing:** - fostering re-use and open science.
  - Safe:** - your research output is stored safely for the future at a sustainable, publicly funded infrastructure.





<https://gold.psychopen.eu/>





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](http://stoptrackingscience.eu)

Time for your questions and thoughts

# Showcasing Research Data in Archives, Repositories and Data Journals

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

Peter Doorn, DANS

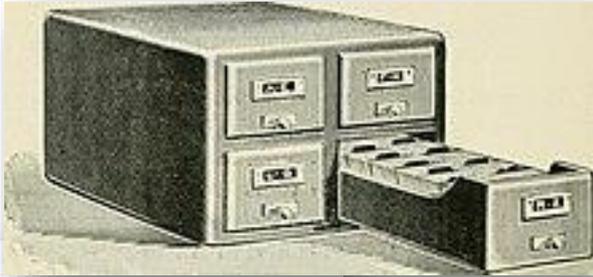
Leen Breure, Sciemedia



@DANS\_knaw\_nwo @pkdoorn

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

# Data: boring metadata vs. user experience



Overview	Description
Persistent Identifier	DOI: 10.17026/dans-27v-paa3 URN: urn:nbn:nl:ui:13-4n-yp9fe
Title	LISS panel - Victims in Modern Society - Victims in Modern Society 2020 - Follow-up measurement
Creator	Veldren, P. van der (Centerdata), Researcher Centerdata, Data collector
Date created (ISO 8601)	2021-10-21
Description	The questionnaire is about experiences with victimization and social s Suggestions for data usage: The data files are accessible via Centerdat information, please use the link under Relations or www.lissdata.nl.
Audience	Behavioural and educational sciences Social sciences
Subject	Society and social systems
Temporal coverage	start-2020-03-02; end-2020-04-28;
Spatial coverage	The Netherlands
Identifier	Fedora Identifier: easy-dataset:225200
Relation	<ul style="list-style-type: none"><li><a href="https://www.dataarchive.lissdata.nl/study_units/view/1173">https://www.dataarchive.lissdata.nl/study_units/view/1173</a></li></ul>
Format	has version <ul style="list-style-type: none"><li>1.0</li></ul> SPSS STATA CSV PDF
Language	English Dutch
Source	LISS panel, Centerdata
Date	Modified: 2021-10-27
Rights holder	Centerdata
Publisher	DANS/KNAW
Access rights	Restricted: request permission - Registered EASY users, but only after

## BRILL Over three centuries of scholarly publishing

### 1 Introduction

Censuses are among the basic information sources of the state of a country. In the 1990s historians in many countries started initiatives to digitize historical censuses, most of which had been recorded and published on paper until well into the second half of the 20th century. In 1996 a proposal was formulated to digitize the printed publications of the Dutch Population census from 1795 to 1971 (Doom et al., 2001). In 1997 this led to a project of the Netherlands Historical Archive (now part of Data Archiving and Networked Services - DANS) in co-operation with the "Centraal Bureau voor de Statistiek" (CBS, Statistics Netherlands). As a result, on the centennial celebration of the CBS in 1999, the scanned images of the printed publications, about 42,500 pages in total, were published, originally on CD-ROM s. Additionally, of the census of 1899, the most voluminous work of the series, both the introductory volume and all published tables were converted into machine-readable form, published as a website ([www.volkstelling899.nl](http://www.volkstelling899.nl)) and on the CBS Open Data Portal Statline. The digital census of 1899 gave rise to a series of new analyses, presented at a symposium and published in the book *Nederland een eeuw geleden geteld. Een terugblik op de samenleving rond 1900* (Van Maarseveen & Doorn, 2007).

As a result of several projects run in the years 2002-2004, the data of all publications of the censuses from 1795 through 1971 were converted into a digitally processable form: pdf for the textual parts, Excel for the tables. This time a twenty-odd number of new analyses were presented at a symposium and published in book form by Boonstra et al. (2007). The first article in that monograph gives an overview of the preceding decade of digitizing the Dutch census publications (Doom & van Maarseveen, 2007).

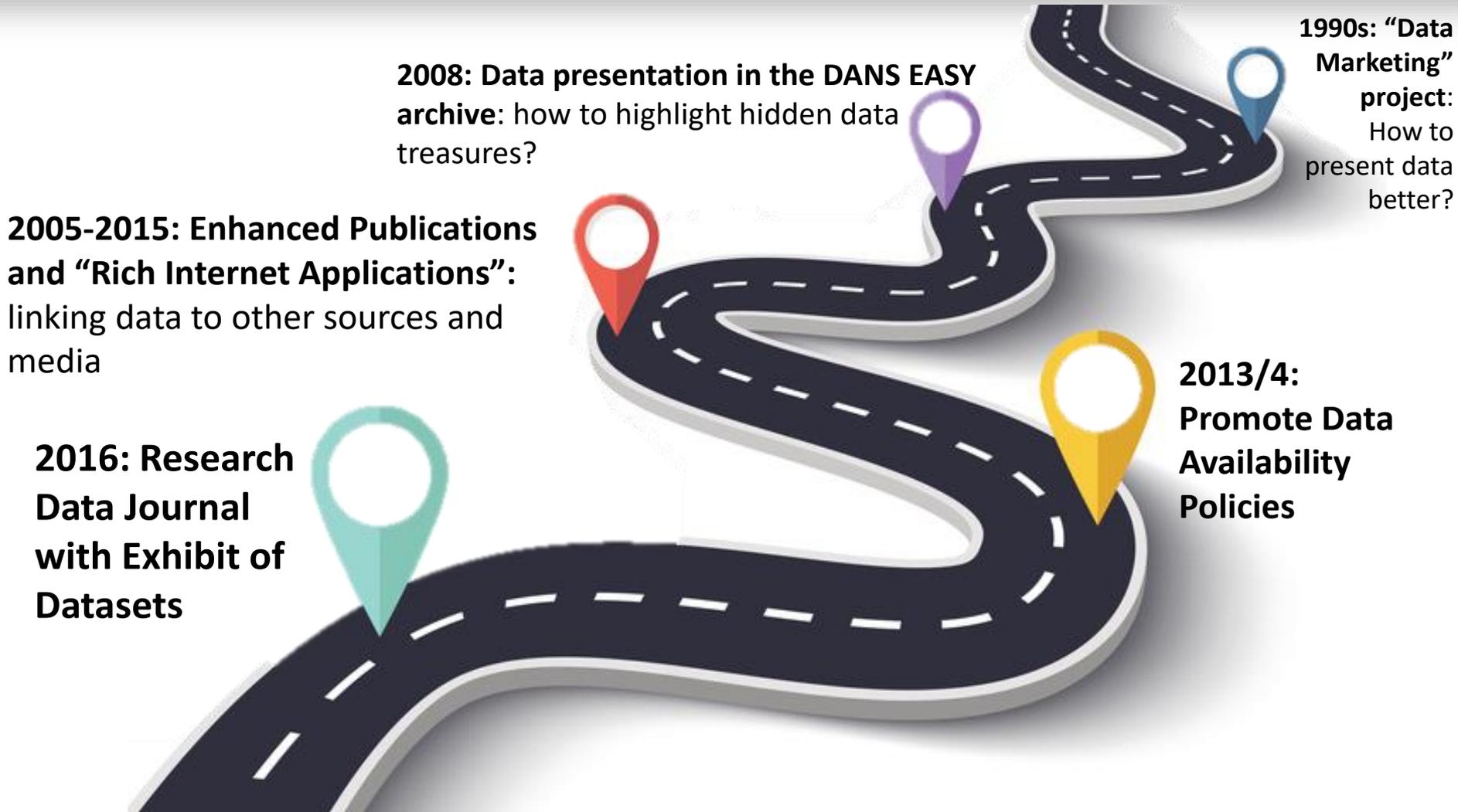
In 2004 the digitized data were made accessible for a worldwide audience via the website [www.volkstellingen.nl](http://www.volkstellingen.nl) and in the DANS data archive. The documentation of the website data is available as online *DANS Data Guide 16 - Volkstellingen 1795-2001* (2019), with references to all data in the DANS data archive ([www.doi.org/10.17026/dans-xhh-p9qy](http://www.doi.org/10.17026/dans-xhh-p9qy)).



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

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

# On the road to experiencing data ...





Gallery EP Features 3

# Leen Breure: typology of 80 types of “Enhanced Publications” - [www.xposre.nl/epfeatures/](http://www.xposre.nl/epfeatures/)

Page: 1 | 2 | 3



# Added value of Data Journal:

## Research Data Journal:

- showcasing data sets
- quality control
- credits for data gathering

through:

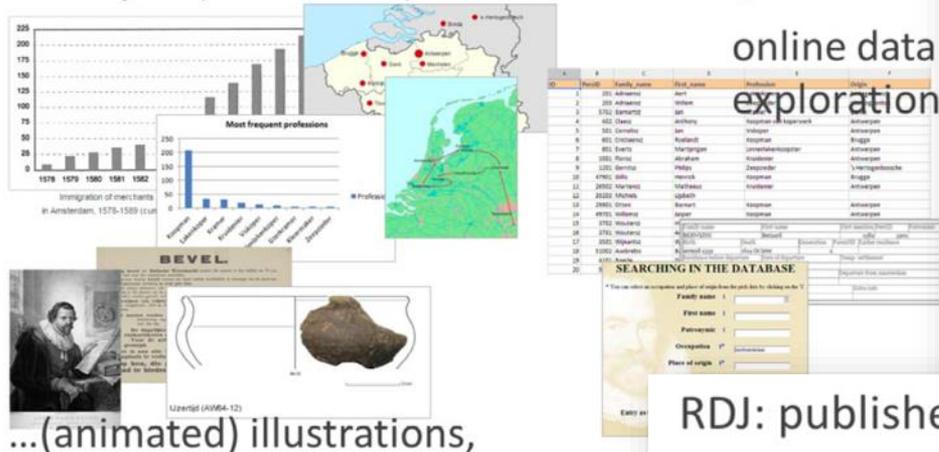
- data papers
- data reviews

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

The screenshot shows the Brill website interface for the article "Detailed Tables from the Dutch Census 1947: Experiences and Lessons Learned in Publishing a Large Dataset". The page includes a navigation bar with "Publications", "Subjects", "Services", "Open Access", "About", and "Contact". The article title is prominently displayed, along with its category "Social and Economic History" and its publication in "Research Data Journal for the Humanities and Social Sciences". The authors listed are Jan Jonker, Wouter Poot, and Peter Doorn. The online publication date is 26 Oct 2021. On the right side, there are options for "Open Access" (with a Creative Commons BY license icon), "Download PDF", "Download Citation", and "Get Permissions". A vertical sidebar on the right contains a journal cover image and social media icons. Below the article title, there are tabs for "Abstract", "Full Text", "PDF", "Metadata", "References", and "Figures". The "Abstract" tab is selected, showing the beginning of the text: "Since the end of the nineties, Dutch census publications have been digitized and made available for digital processing. New analyses of the data were presented at some fruitful conferences in the first decade of this century. In addition to the new publications, a mass of detailed census data was found in dossiers and so-called 'transcripts' in the archive of Statistics Netherlands. Most of that material was scanned into digital images, but not all of it was converted into numeric data. In the present article, the authors describe the process of digitizing the detailed tables of the Dutch Population and Occupational Censuses held in 1947, which is the first set of detailed census data that is made available in a machine-readable form. They give an example of historical analyses made possible by this data. Moreover, they take these census data as an example of preparing and publishing a large dataset. Experiences and lessons learned in the process lead to ample suggestions for further analysis of the data and for efficient ways to accomplish the content conversion of the many remaining images of census data." The keywords are: large dataset; census data; Netherlands; 1947; data-entry; versioning; documentation method; csv-text files. A yellow diagonal banner across the abstract reads "After initial support by DANS, now under CESSDA direction". A large orange arrow points from the abstract towards the right, containing the text "Data experience" still missing!".

# Added functionality we aimed for

Variety of (interactive) data visualization,



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



FAIR Data Reviews  
for data in a trustworthy repository



## 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](https://easy.dans.knaw.nl), which is certified with the Core Trust Seal (CTS, see <https://www.coretrustseal.org>).

RDJ: published for all devices



desktop



tablets



smartphones

and perhaps later push notifications on a smartwatch



...serving a generation of 'digital natives'

# Data papers tended towards "normal" research articles; therefore:

## Exhibit of Datasets

Research Data Journal for the Humanities and Social Sciences



Fragment from a map of science by Herbert van de Sompel

Home

### Table of Contents

#### SELECT

Click button to filter the datasets:

- [Archaeology](#)
[Social & Economic History](#)
[Linguistics & Literature](#)
[Social & Behavioural Sciences](#)
[Arts & Media](#)
[Other Humanities](#)
[Reset](#)

[Datapapers with a showcase](#) | [Special issue on Performing Arts](#) | [All](#)

#### DATA PAPERS AND SHOWCASES

Not all data papers have a showcase; only clickable titles are linked to a showcase.

[Alphabetical Order](#) | [Chronological Order](#)

- Allen, R.C. & Unger, R.W. (2019) *The Allen-Unger Global Commodity Prices Database*
- Ashby, M.P.J. (2019) *Studying Crime and Place with the Crime Open Database*
- Baciocchi, S., Beauguitte, L., Blavier, P. & Lambert, N. (2019) *Documenting the Diffusion of the 2016 French 'Nuit Debout'*
- Bakkeren, F. (2020) *History and Contents of the Dutch Theatre Production Database (forthcoming)*
- Blom, F.R.E., Nijboer, H.T. & Van der Zalm, R.G.C. (2020) *ONSTAGE, the online Data System of Theatre in Amsterdam from the Golden Age to Today*

#### STATISTICS

Archaeology	3
Social & Economic History	7
Linguistics & Literature	4
Social & Behavioural Sciences	6
Arts & Media	11
Other Humanities	0
<b>Data papers</b>	<b>31</b>
<b>Showcases</b>	<b>18</b>

#### Showcase

- [DATASET](#)
[DATAPAPER](#)
[WEBSITE](#)
[EXPLORE DATA](#)
[HOW TO CITE](#)

## Exhibit of Datasets

Research Data Journal for the Humanities and Social Sciences

Showcases | Table of contents

#### HIGHLIGHTS

Oort, T. van & Noordegraaf, J.J. (2020)

The Cinema Context Database on Film Exhibition and Distribution in the Netherlands: A Critical Guide



Poster of the cinema salon Wredstad in Utrecht. Expositions by the famous stage actor Pi. Louis Hartlooper, 1910-1911 (Stichting Anankai)

Cinema Context is a database for researching the history of cinema exhibition and distribution in the Netherlands, covering a period from the late nineteenth century until the present. It is both the result and catalyst of a broader movement in historical film studies known as *New Cinema History*, which focuses on moving away from a social and cultural phenomenon, rather than on films as aesthetic objects (see [introduction to New Cinema History](#) (PDF download)).

Cinema Context contains a wealth of information on film screenings from 1896 to 1940 in various cities in the Netherlands, mostly from Amsterdam, Rotterdam, The Hague, Utrecht, Groningen and Limburg. Each week's film programme has been included, as far as this is documented in the sources we have consulted. Travelling cinemas can often be traced from 1896 through 1910, even outside the aforementioned cities. It can be used as a straightforward encyclopaedia for looking up facts and for more complex analyses of patterns and networks in film distribution and exhibition. The international community of experts in the field of film and cinema history regards Cinema Context as a standard that has served as an example for similar databases around the world.

Read more...

Oort, T. van, Jermudd, A., Lotze, K., Pafort-Ovendiin, C., Bittereyst, D., Boter, J., Dibbelius, S., Ercole, P., Meers, P., Porubanska, T., Treveri Gemari, D. & Van de Vijver, L. (2020)

Mapping Film Programming across Post-War Europe (1952)



Poster for the American theatrical run of the 1951 musical film *An American in Paris*, the second most screened film in 1952 (Wikimedia Commons)

This showcase presents a fully harmonized data set originating from different research projects on post-war cinema programming. The creation of this data collection was inspired by the *New Cinema History*, which focuses on the wider contexts of the production, distribution and consumption of films. The database consists of titles of feature films screened for public viewing in cinemas in the cities Bari (Italy), Antwerp and Ghent (Belgium), Gothenburg (Sweden), Leicester (United Kingdom) and Rotterdam (Netherlands) for the year 1952 (here a cinema was defined as a venue where 35mm and 16mm feature films were screened that were announced to and accessible by a general audience).

The choice of these cities was partly a result of the focus and research intentions of earlier projects, but the places do also have important features in common that justify their inclusion in the data set and contribute to a framework that allows mutual comparisons.

The origin of this data collection goes back to the work by Karel Dibbets in the 1970s. This formed the basis for the later *Cinema Context* database, from which data about Rotterdam were extracted. The Rotterdam data set was subsequently supplemented with film programming data collected in the other participating research projects.

Read more...

#### SHOWCASES

This Exhibit contains showcases, short introductions to archival datasets, most of which are described in detail in data papers published in the corresponding Research Data Journal for the Humanities and Social Sciences, published by Brill.



#### LEARNING BY EXAMPLE

See how other people have applied advanced data handling techniques.

- Managing complex data structures in a relational database — even on your own laptop (Drauwiers2020a, reusand2020b)
- Making queries in Excel and joining workspaces (Van2020b)
- Using Linked Data with historical sources (Jansen2020)
- Jupyter Notebooks (Jupyter) as explorative programming aid in studying the Hebrew Bible (Schoor2020)

# Data showcases in the "Exhibit" link to archived data sets and to data papers

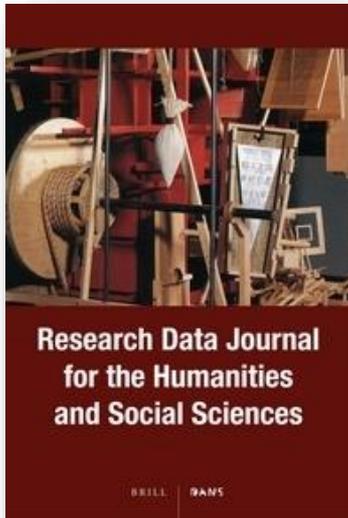


## UK Data Service

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>



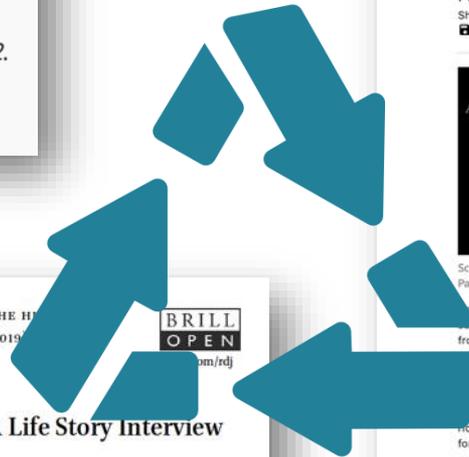
### Pioneers of Social Research: A Life Story Interview Collection

Social and Behavioural Sciences

Camille Corti-Georgiou  
University of Manchester, UK  
[camillegeorgiou@hotmail.co.uk](mailto: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 interview transcripts and detailed summaries, YouTube playlists, thematic highlights and



### Exhibit of Datasets

Research Data Journal for the Humanities and Social Sciences

Home More Showcases

Corti-Georgiou, C. (2019)  
University of Manchester (UK)

#### Pioneers of Social Research: A Life Story Interview Collection

Showcase

[DATASET](#) [DATAPAPER](#) [WEBSITE](#) [EXPLORE DATA](#) [HOW TO CITE](#)

Paul Thompson  
Audio Extract: Researching at school

Interviewed by Karen Wootman, 1996, 2002

Screen capture from an interview with Paul Thompson (YouTube).

Between 1940 and 1970, social researchers such as Margaret Stacey and Paul Thompson were pioneers of social research. In 1994 Thompson's *Pioneers of Social Research* (Qualidata) funded by the UK's Economic and Social Research Council (ESRC) was the first interview collection covers significant ground in social science today, such as research opportunities for exploring the development of this project were selected on the basis of:

1. First was the limitation of the budget.
2. In discussion with colleagues with a view to seek to record around fifty interviewees.
3. The target sample aimed for a balance of qualitative and quantitative traditions.

Around seventy individuals were nominated, some of whom could not be traced or were unavailable, some had already been interviewed about their work for other projects, and others declined due to ill health.

#### Country of birth

Country	Count
Australia	1
Canada	1
Germany	1
France	1
India	1
Italy	1
Japan	1
Norway	1
New Zealand	1
Pakistan	1
Tanzania	1
United Kingdom	44
United States	1
South Africa	1

#### Interviewees (N=59)

Gender	Count
Male	54
Female	5

#### PIONEERS OF SOCIAL RESEARCH with Paul Thompson

UK Data Service: Pioneers of Social Research

# Objectives of Data Showcases



Attention



FAIRness



Coherence

Interactivity



Context

Multimodality



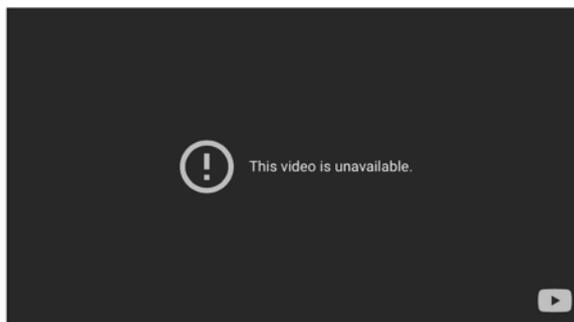
Trust



Transparency

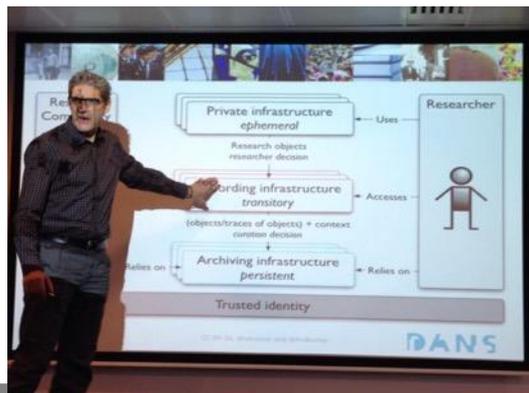
# Or is it a Dead End?

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

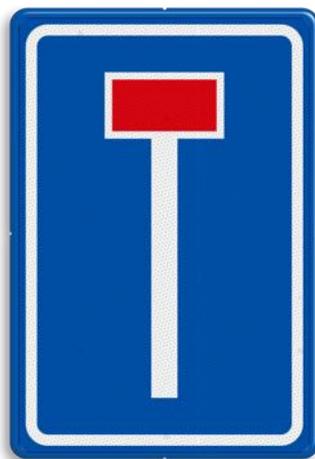


When will this future be?

2014: The future of scholarly communication and the archive:



Data Archiving and Networked Services



## The Future of Scholarly Skywriting

(C. Scagliola, 1990-1999)

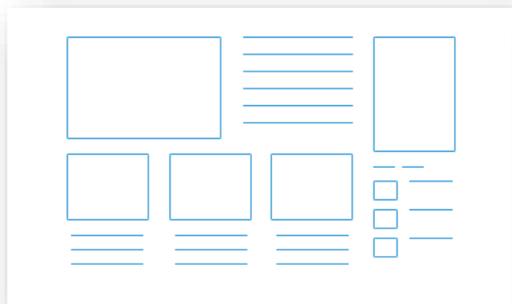


Vanishing into thin air...

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

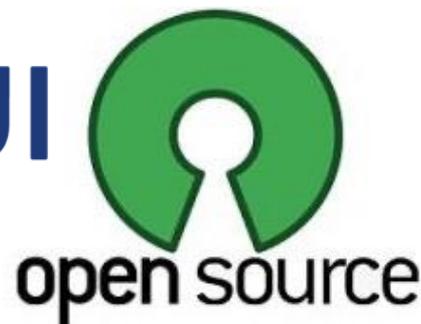
**Maintainability:** can we preserve the functions?



Existing tools



FAIR data  
assessment



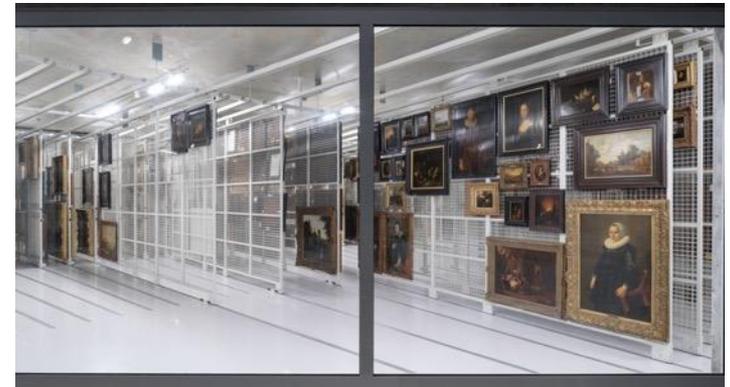
Acceptance...

# An analogy from the museum world:



Depot Museum Boijmans van Beuningen, Rotterdam

The world's first publicly accessible art storage facility



# Thank You!



FAIR Data Review Form (manual): <https://tinyurl.com/ehhdtw3u>

Automated FAIR data assessment (F-UJI): <https://www.f-uji.net>

[www.dans.knaw.nl](http://www.dans.knaw.nl)

[peter.doorn@dans.knaw.nl](mailto:peter.doorn@dans.knaw.nl)

[l.breure@uu.nl](mailto:l.breure@uu.nl)

# 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*

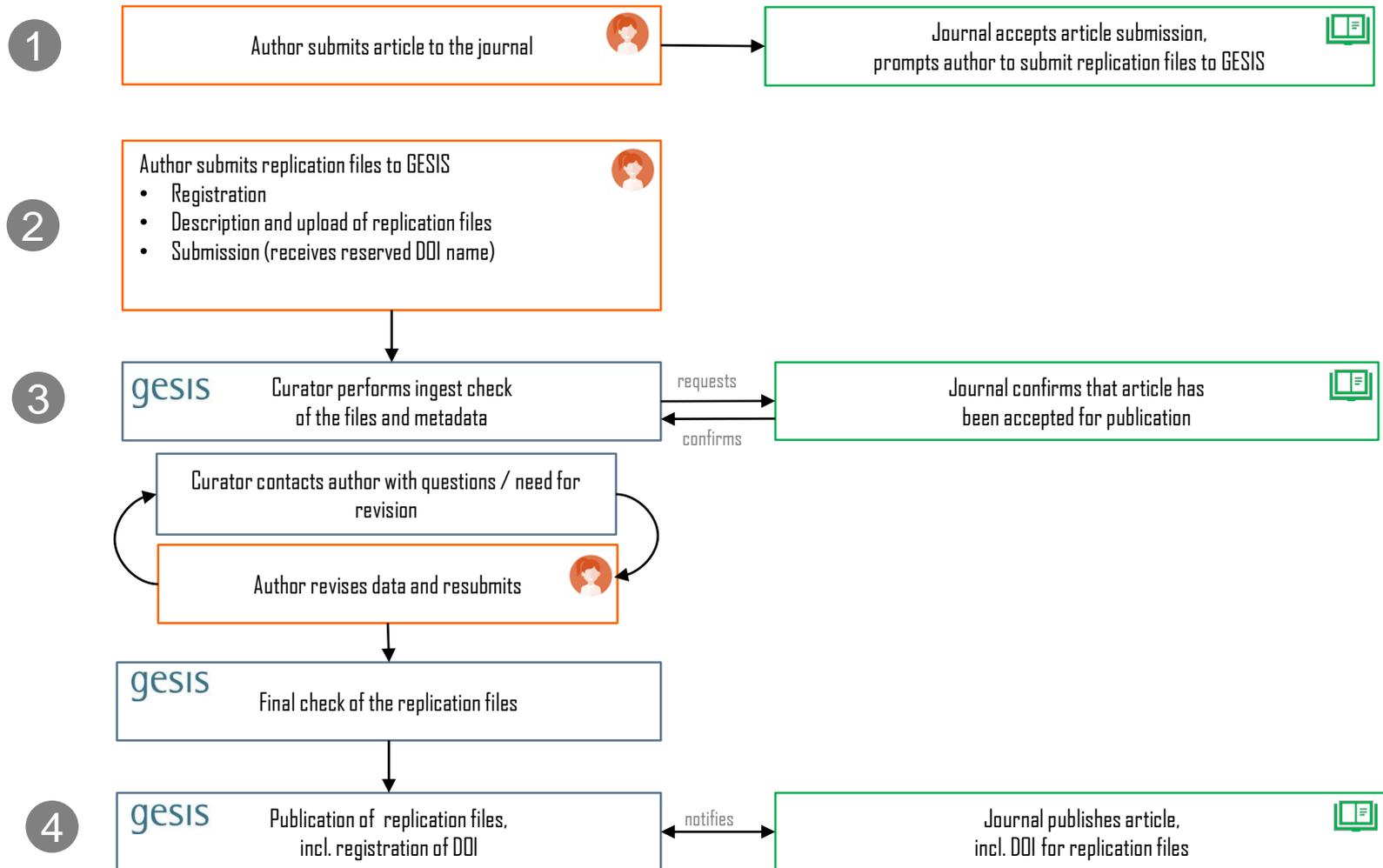
## How it started vs How it's going

- 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/>

# Workflow



## Example "Zeitschrift für Soziologie"

- 2015: Editorial announced new practice<sup>1</sup>
- 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<sup>2,3</sup>
  - ▶ 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

<sup>1</sup>Zeitschrift für Soziologie, Jg. 44, Heft 1, Februar 2015, S. 2–5

<sup>2</sup>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>

<sup>3</sup>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



# 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	✓	✓	✓	X
BJPS	✓	✓	✓	X
EJPR	✓	X	✓	X
JOP	✓	✓	✓	X
PA	✓	✓	✓	✓
POQ	✓	X	✓	X
PRQ	✓	X	✓	X
PS	✓	✓	✓	X
QJPS	✓	X	✓	X

\* 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

<i>Models and predicted probability of civil war onset</i>				
<i>Civil war onset</i>	<i>Fearon and Laitin (2003)</i>	<i>Collier and Hoeffler (2004)</i>	<i>Hegre and Sambanis (2006)</i>	<i>Random Forests</i>
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

	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	0.002212600	2.662163e-02	0.0157644503	0.04742684

# Summary of MSHK Paper and Replication

- 2016 code:
  - ▶ In-sample predictions
  - ▶ Unusable .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
<b>Data citation</b>	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
<b>Details</b>	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 <b>must</b> be appropriately cited"	<b>"Articles will not be published until the citations conform to these standards."</b>
<b>Data transparency</b>	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
<b>Details</b>	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.	<p>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.</p> <p>If the policy strongly encourages data sharing, this level is not reached.</p> <p>Policies that require sharing with editors and reviewers only do not apply.</p> <p>Policies that require sharing only "upon request" do not apply.</p> <p>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.</p>	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](http://Informa.com)

