KonsortSWD

Consortium for the Social, Behavioural, Educational and Economic Sciences

OPEN SCIENCE FAIR

Open Science Fair 2023,
September 25-27, 2023,
Madrid, Spain

DOI: 10.5281/zenodo.8298743

nfdi

gesis
Leibniz Institute
for the Social Sciences

# Enhancing FAIR Compliance in Research Data Infrastructures: Insights from Applications of the RDA FAIR Data Maturity Model and the F-UJI Automated FAIR Data Assessment Tool

Janete Saldanha Bach
Brigitte Mathiak
Claus-Peter Klas
Yudong Zhang
Peter Mutschke

GESIS – Leibniz Institute for the Social Sciences

# Authors

## Janete Saldanha Bach

Dr. Janete Saldanha Bach, GESIS – Leibniz Institute for the Social Sciences. Postdoc in the NFDI consortium KonsortSWD in the department "Knowledge Technologies for the Social Sciences" , team FAIR Data, working in the consortia KonsortSWD Project of the National Research Data Infrastructure (NFDI). She holds a Ph.D. and a Master's degree in Science and Technology Studies (STS) and a bachelor's degree in Information Science. Her research expertise is in Open Science, especially in research data management and data reuse in the Social Sciences. She is currently involved in consortium KonsortSWD, Task Area 5 Measure 1 - developing the conceptual framework for the PID registration service at a variable level and Task Area 5 Measure 2 Enhancing data findability.

## Brigitte Mathiak

Brigitte Mathiak is a senior scientist at GESIS - Leibniz-institute for the Social Sciences with a PhD in Computer Science. Her research activities focus on data discovery, information extraction and research data management. She is the lead of the KonsortSWD measure on data findability. Deputy speaker of the NFDI section on (Meta)data, Terminologies and Provenance. Co-Speaker of the working group on Search & Harvesting in that section. Co-Chair of the GOFAIR Discovery Implementation Network and the RDA working group on data granularity.

# Authors

## Claus-Peter Klas

Dr. Claus-Peter Klas, GESIS – Leibniz Institute for the Social Sciences, Team Leader "Data & Service Engineering" and Measure Lead in the NFDI consortium KonsortSWD in the department "Knowledge Technologies for the Social Sciences" . He received his PhD in computer science at the University of Duisburg-Essen and was a postdoctoral researcher in the Department of Multimedia and Internet Applications, Faculty of Mathematics and Computer Science, University of Hagen, Germany. His research focuses on information retrieval, interactive information retrieval, information systems, databases, digital libraries, preservation and grid and cloud architectures. He developed the software Daffodil founded on a nation research project and worked in national and European research projects such as The European Film Gateway, SHAMAN (Sustaining Heritage Access through Multivalent ArchiviNg) and Smart Vortex (Scalable Semantic Product Data Stream Management for Collaboration and Decision Making in Engineering). He is currently responsible for several infrastructure projects within GESIS, such as da|ra, SowiDataNet or Missy, all concerned with providing information and data for social scientists. In addition, he leads the measure PID Services in the national research infrastructure project NFDI. In his team, they are developing an open source DDI suite to support getting DDI into operation.

## Yudong Zhang

Mr. Yudong Zhang is a Software Engineer in GESIS – Leibniz Institute for the Social Sciences, in the NFDI consortium KonsortSWD in the department "Knowledge Technologies for the Social Sciences", team FAIR Data, working in the consortia KonsortSWD Project of the National Research Data Infrastructure (NFDI). He holds a Master’s degree in Media Informatics, a bachelor's degree in Computer Science and a bachelor's degree in Accounting Banking and Finance. He is an expert in research data identify, management and exchange. He is currently involved in consortium KonsortSWD, Task Area 5 Measure 1 - developing the conceptual framework for the PID registration service at a variable level, Task Area 5 Measure 2 - Enhancing data findability, and applying FAIR principles assessment models and tools and Task Area 5 Measure 5 - Interface for Data Exchange.

# Authors

## Peter Mutschke

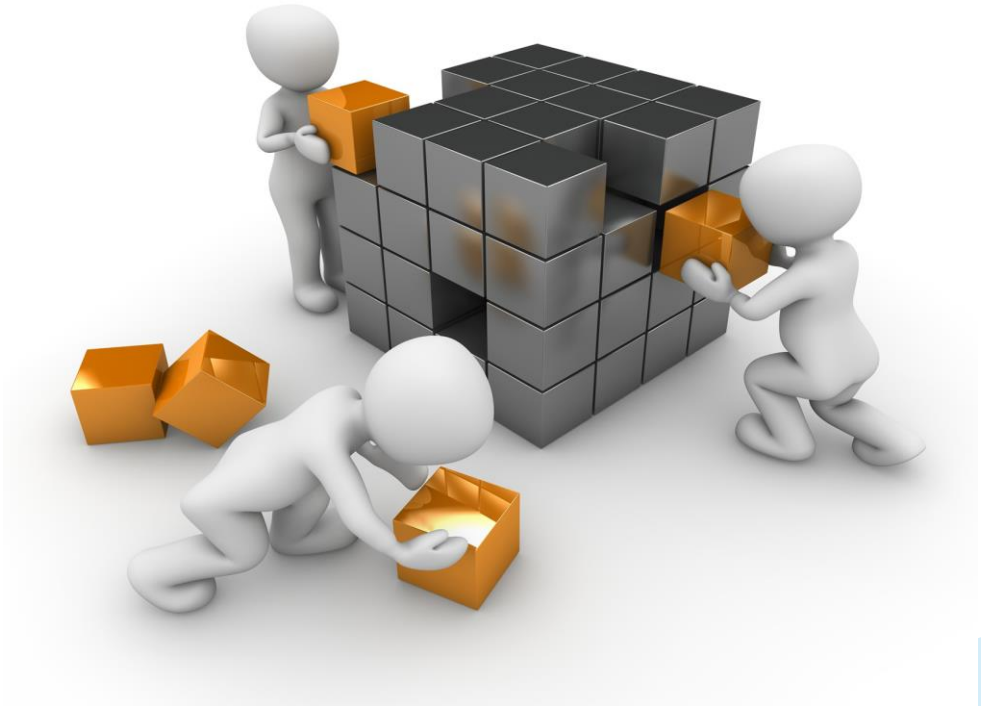

Peter Mutschke is deputy head of the department “Knowledge Technologies for the Social Sciences (KTS)" and leader of the team "FAIR Data" of KTS. His research interests include Information Retrieval, Network Analysis and Open Science. He worked in a number of national and international research projects, such as the DFG projects DAFFODIL and IRM and the EU projects WeGov, SENSE4US, OpenMinTeD and MOVING. Peter served as a member of the management committee of the Leibniz research alliance “Science 2.0/Open Science” from 2013-2021. He founded and coordinates the GO FAIR Implementation Network "Cross-Domain Interoperability of Heterogeneous Research Data (Go Inter)", and he is member of the steering committee of the FAIR Digital Objects Forum (fairdo.org) where he also co-chairs a working group on semantics. He is currently involved in consortia KonsortSWD, NFDI4DataScience and BERD@NFDI of the National Research Data Infrastructure (NFDI).

# Agenda

- **The PID Registration service**
  - General goal and claim
  - Hurdles of data citation current practices
  - The Research data granularity levels
  - Data citation using PIDs
- **The PID Registration service: FAIR assessment**
  - Criteria
  - Methodology
  - Results
- **The FUJI-Tool assessment**
  - Analysis
  - Results
  - Outcomes

# Goal


- Identify dataset elements, **using one identifier** - the PID - will simplify FAIR data management to:
  - to boost subsequent citation,
  - get direct (meta)-data access, and
  - promote data reuse.



FAIR: Findability, Accessibility, Interoperability, and Reusability

Image by Peggy und Marco Lachmann-Anke from Pixabay

# Hurdles of data citation current practices

## Ex. 1: Dastaset cited in the text

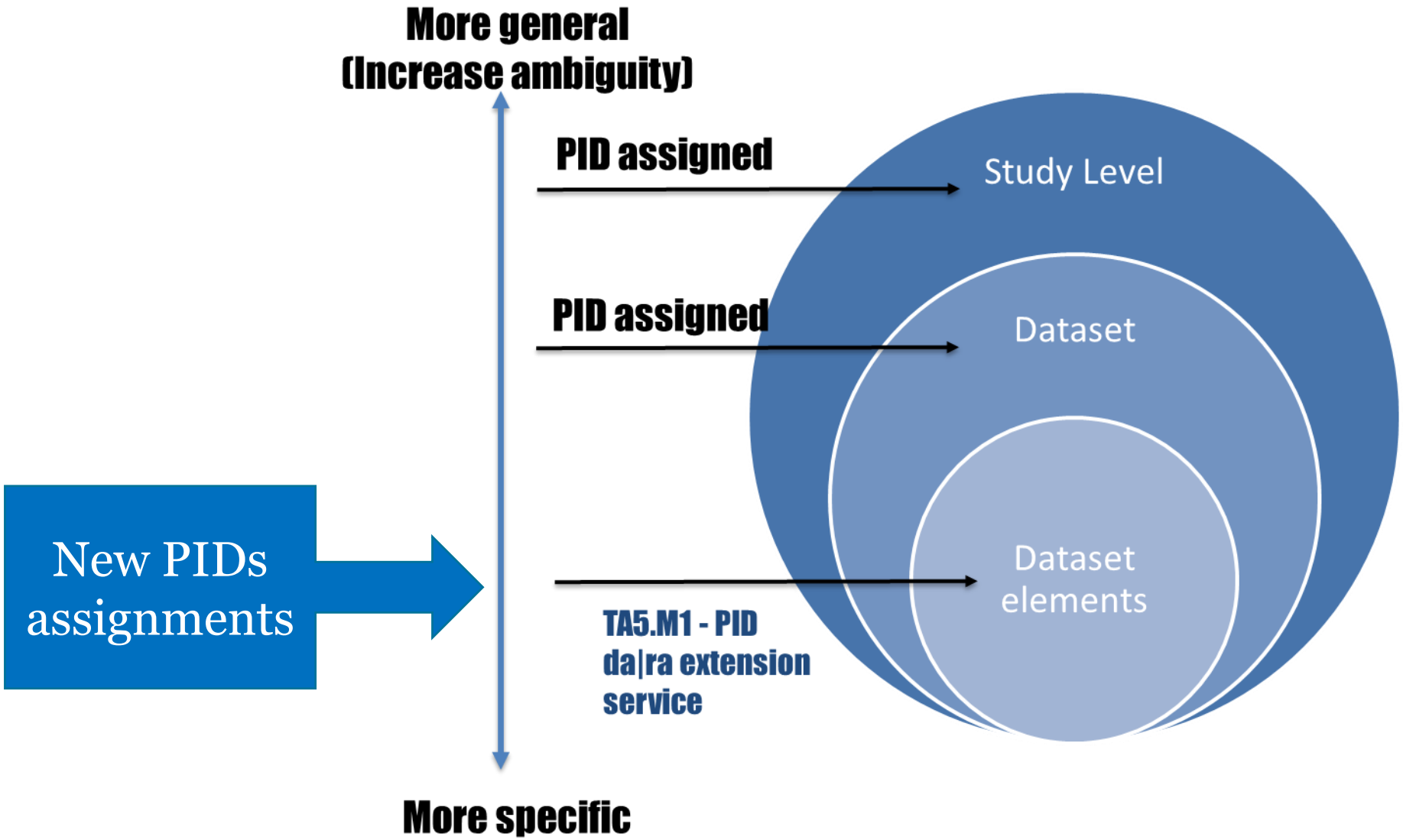
Religiousness. *General religiosity* was measured through the ISSP 2008 item: "Would you describe yourself as. . . . ?" (responses ranged from 1 = *extremely religious* to 7 = *extremely non-religious*). For the analyses, scores were reversed. *Religious practice* was measured through three ISSP 2008 items assessing frequency of prayer, religious attendance, and visitation to holy places (responses ranged from 1 = *never* to 11 = *once a day*; α = .61; αs across samples: .43-.64).[1]

## Ex. 2: Question cited in the text

participation rather than opinions and beliefs. The key variables concern attendance of religious services and several demographic and socioeconomic characteristics, such as age, work status, and income.

Several variables used below deserve a more precise definition. First, two levels of attendance are distinguished in the analysis based on the question: "How often do you attend religious services?" Weekly attendance means that a respondent claims to attend a religious service at least once a week; yearly attendance signifies participation at least once a year. Second, employment

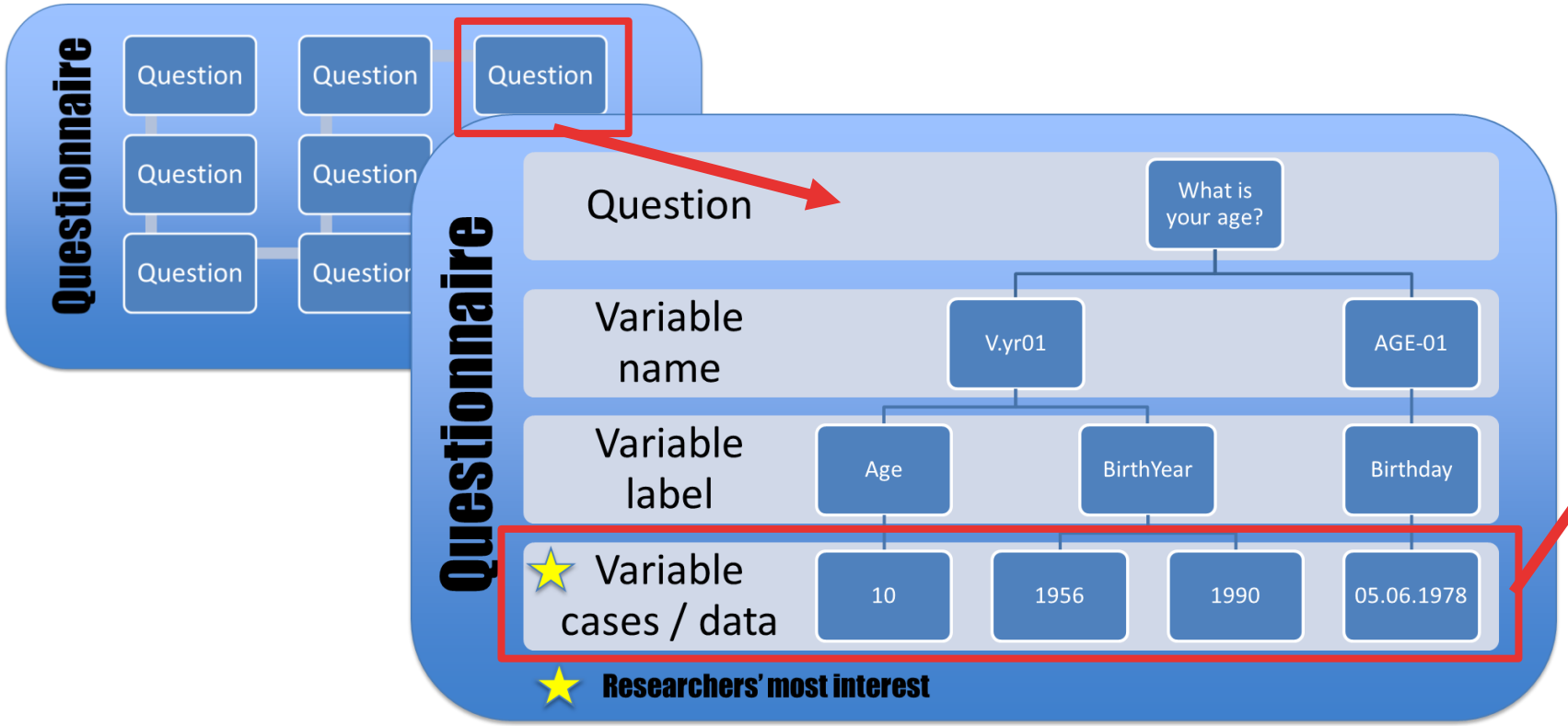
# The Research data granularity levels



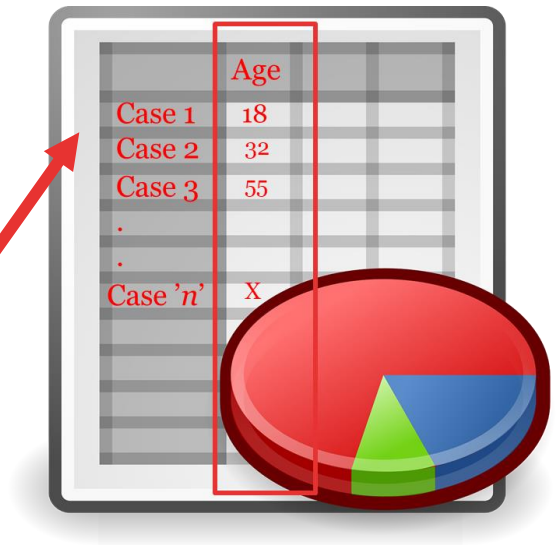

Klas, Claus-Peter, Zloch, Matthäus, Bach, Janete Saldanha, Baran, Erdal, & Mutschke, Peter. (2022). KonsortSWD Measure 5.1: PID Service for variables report (1.0). Zenodo. https://doi.org/10.5281/zenodo.6397367

# Data formats

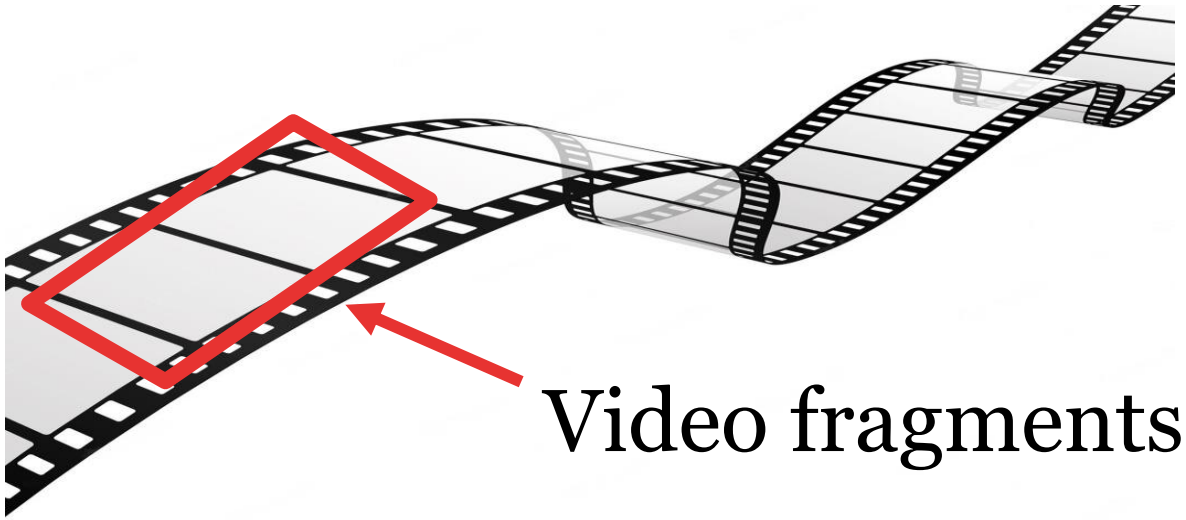
## Social Sciences' research data common structure (surveys)




Variables' data are stored in tabular data format



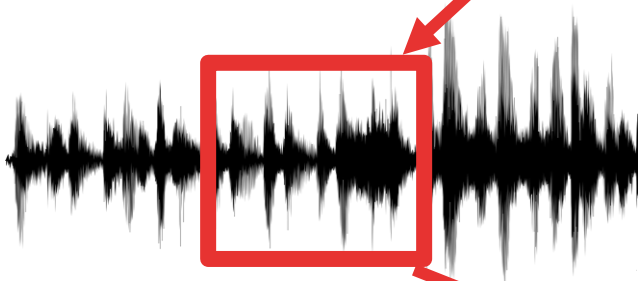
# Data formats possible in the future



Video fragments



Audio files



Transcripts

# PID registration service provider

- This service assigns a PID with **Handle** standard (ePIC);

- The service will be upgraded to **handle PIDs on variable** level;

ePIC
Persistent Identifiers for eResearch

KonsortSWD

# Use cases in a nutshell

| Institution name | DZHW \| German Center for Higher Education Research and Science Research (DZHW) | GESIS \| Leibniz Institute for the Social Sciences | GESIS harmonization tools | | | DIW \| German Institute for Economic Research | Qualiservice \| University of Bremen |
|---|---|---|---|---|---|---|---|
| | | | GESIS QuestionLink | ONBound - Old and new boundaries: National Identities and Religion | Harmonizing and synthesizing partnership histories | | |
| **Project / Study** | **HEADS - Higher Education Analytical Data System** | **Gesis Data Archive** | **QuestionLink Harmonisation tool** | **ONBound Harmonisation Wizard** | **HaSpaD - Harmonising and synthesising partnership histories** | **German Socio-Economic Panel Study (SOEP-Core)** | **Qualiservice as part of QualidataNet from KonsortSWD** |
| Attributes | **Survey variables** | **Survey variables** | **Survey variables** | **Survey variables** | **Survey variables** | **Survey variables** | **Qualitative data files** |
| PIDs uses cases | Variables in datasets; differentiation variables: complex system including one content (dependent variable) plus several independent variables to differentiate this dependent variable by subgroups; | Variables in large datasets from national and international studies | Variables from GESIS and third-parties collections (NEPS and SOEP) for harmonization purposes. Political interest pre-harmonized variables. Some surveys only have one variable name across several years, whereas other surveys have different variable names per wave. | Variables from third-parties collections for harmonization purposes; Religion and Nation in Constitutions Worldwide, Religion and State Project, Church Attendance and Religious; United Nations: Demographic Statistics Database | Variables from third-parties collections for harmonization purpose; The survey programs include Panel Analysis of Intimate Relationships and Family Dynamics s | Assign a PID for each variable from the SOEP-Core v37. SOEP-Core doi:10.5684/soep-core.v37o | Assign a PID for qualitative data, organized in files or dataset, regarding Transcripts, translation, audiovisual and context material for doctor-patient-interaction videos observed |
| **Variables #** | Depend on the user selection | 507.642 | 68 | 750 | Depend on the user selection | 101.574 | N/A |

# FAIR Assessment

## Assessment criteria: manual and automatic approaches


- FAIR Data Maturity Model (RDA-FDMM) [1]: 41 FAIR indicators, organized into three classes (essential, important, useful) and five levels of assessment
- F-UJI tool: automated assessment of 16 indicators (which can be assessed automatically) [2]

[1] https://doi.org/10.15497/RDA00050 | [2] https://doi.org/10.5334/dsj-2021-004

# FAIR Evaluation: Criteria


## The service FAIR maturity level assessment: Criteria

- We assessed the service under the **FAIR Data Maturity Model**

FAIR Data Maturity Model
Specification and Guidelines

RDA
RESEARCH DATA ALLIANCE

**DOI**: 10.15497/rda00050

**Co-chairs**: Edit Herczog, Keith Russell, Shelley Stall

**Published**: 25th June 2020

**Abstract**: Findability, Accessibility, Interoperability and Reusability – the FAIR principles – intend to define a minimal set of related but independent and separable guiding principles and practices that enable both machines and humans to find, access, interoperate and re-use data and metadata. The FAIR principles were defined in 2016 in an article by Mark Wilkinson et. al1. FORCE112 and GO FAIR3 provide further information on the principles. The principles have to be considered as

# FAIR Evaluation: Criteria

## The service FAIR maturity level assessment: Criteria

- The framework consists of **3 indicators classes**: Essential, Important, and Useful
- The sum of them is organized into **five levels**, according to the present indicator in each category
- When distributing the indicators per FAIR area, the principle of **Accessibility** and **interoperability** holds the majority of Essential and Important criteria for FAIRness

**3 indicators classes in five levels**

| FAIR Data Maturity Model: evaluation framework | | | Level 1 | Level 2 | Level 3 | Level 4 | Level 5 |
|---|---|---|---|---|---|---|---|
| Essential | 20 | | 20 | 20 | 20 | 20 | 20 |
| Important | 14 | | | 7 | 14 | 14 | 14 |
| Useful | 7 | | | | | 3 | 7 |
| **Total** | **41** | | **20** | **27** | **34** | **37** | **41** |

**Indicators according to the FAIR Principles**

| Distribution of priorities per FAIR area | | | | | |
|---|---|---|---|---|---|
| Principle | Findable | Accessible | Interoperable | Reusable | Total |
| Essential | 7 | **8** | 0 | 5 | 20 |
| Important | 0 | 3 | **7** | 4 | 14 |
| Useful | 0 | 1 | 5 | 1 | 7 |
| Grand Total | 7 | 12 | 12 | 10 | 41 |

# FAIR Evaluation: Methodology

## The service FAIR maturity level assessment: Methodology

- Applied the **stricter** evaluation method on each indicator, assessing them by passing or failing **(‘yes’ or ‘no’)** questions
- This approach was selected because the PID registration service is based on the data registration agency da|ra (www.da-ra.de)
- Link to assessment data [3]

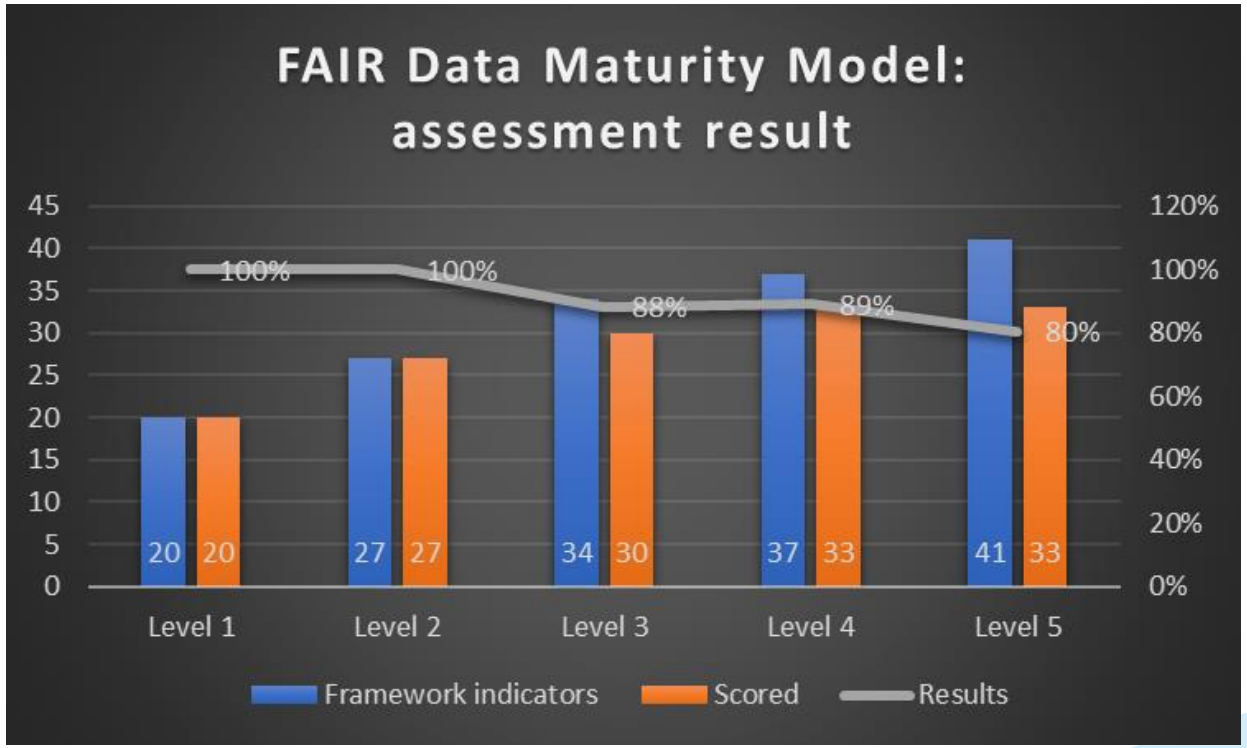

| Measure 5.1: PID Service for variables | Present | Not present | | |
|---|---|---|---|---|
| FAIR Data Maturity Model: criteria framework | Pass | Fail | Evidence | Comments |
| RDA-F1-01M Metadata is identified by a persistent identifier | Pass | | It has a variable PID assigned | Metadata and data is identified via DOI |
| RDA-F1-01D Data is identified by a persistent identifier | Pass | | It has a variable PID assigned | Metadata and data is identified via DOI |
| RDA-F1-02M Metadata is identified by a globally unique identifier | Pass | | Handle standard provides globally unique identifier | Metadata and data is identified via DOI |
| RDA-F1-02D Data is identified by a globally unique identifier | Pass | | Handle standard provides globally unique identifier | Metadata and data is identified via DOI |
| RDA-F2-01M Rich metadata is provided to allow discovery | Pass | | A metadata scheme is present to comply with the minimum metadata | Metadata is documented in DDI Lifecycle 3.2 |
| RDA-F3-01M Metadata includes the identifier for the data | Pass | | It includes the DOI of the study in which the variable to register appears | The DOI is part of the metadata |
| RDA-F4-01M Metadata is offered in such a way that it can be harvested and indexed | Pass | | The metadata is in fact harvested and indexed at Gesis Search and/or other institutional repository as a service user | Metadata can be harvested via OAI-PMH |

[3] https://doi.org/10.5281/zenodo.8339809

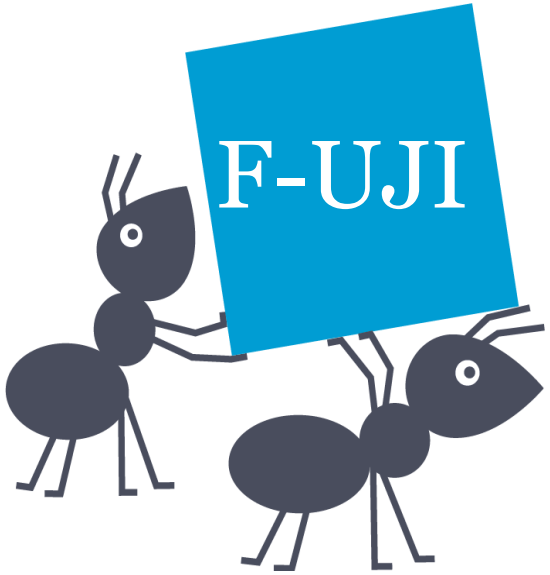
# FAIR Evaluation: results

## The service FAIR maturity level assessment: Results

- The results demonstrate **outstanding achievements at levels 1 and 2**, marking **100%** on the assessment measure
- The service achieves **88%** compliance at level 3 and **89%** at level 4. At level 5, the results show **80% o**f passed indicators
- The service meets **all** indicators classified as **essential**
- The failed indicators concerned with **automatic features**, including references and/or qualified references to other data, and data is accessed automatically (i.e., by a computer program)




PID registration service maturity model assessment results




# Automated FAIR Assessment with F-UJI

- automated FAIR assessment of the **GESIS Search** (search.gesis.org) as a **relevant repository** in the context of **KonsortSWD**

- to identify **how-to's for improving metadata** and/or metadata representation by automated means

- automated **improvements of metadata** led to a noticeable enhancement of the **FAIR maturity scores**

# Results

## Recommendations to improve automated FAIRness scores

- ensure that the **landing page is machine-readable**, avoiding JavaScript generated contents
- define **available metadata in JSON-LD**, both on the landing page and in the used PID registration system, e.g., DataCite
- **provide links to the content resources** (e.g., the PDF article, CSV datasets) on the landing page
- linked content resources of **long-term readability**, such as plain text, are preferred

# Results

## Recommendations to improve automated FAIRness scores

- ensure metadata for **linked data is correct and complete**
- use the **standards suggested by F-UJI** to complement free-form descriptions
- keep the re3data records **up to date** and define an OAI-PMI endpoint for it

# FAIR assessment: outcomes

- in-depth FAIR analysis using the RDA FAIR Data Maturity Model helps to **understand better where your services are** with regard to FAIR
- **automated tools** (like F-UJI) provide valuable hints on how to **improve your metadata** to achieve better automatically generated FAIRness scores
- some FAIR **indicators still require human mediation** and interpretation, as not all components of the research data ecosystem are machine-readable
- our experience underscores the importance of **assessing both machine-readable and non-machine-readable elements**, given the limitations of automated means

# FAIR assessment: outcomes

- to gain a **comprehensive view of research data infrastructure's** FAIR compliance, apply **broader standards** for manual FAIR assessment (like the RDA model) as well as automated assessment tools (like F-UJI)
- adopt a "**FAIR by design**" approach early in service development to ensure FAIR principles are **embedded in project outcomes** from the beginning
- include **regular FAIR assessments** throughout the project lifetime to evaluate how the ongoing improvement of research data infrastructures affects FAIR maturity score

# Acknowledgments


Saldanha Bach, Janete; Mathiak, Brigitte; Klas, Claus-Peter; Zhang, Yudong and Mutschke, Peter. 2023. Enhancing FAIR Compliance in Research Data Infrastructures: Insights from Applications of the RDA FAIR Data Maturity Model and the F-UJI Automated FAIR Data Assessment Tool. In *Open Science Fair 2023*. Madrid, Spain, September 25-27, 2023, 23 slides. DOI: 10.5281/zenodo.8092028.



PID Service report https://doi.org/10.5281/zenodo.6397367

PID use cases extended report https://doi.org/10.5281/zenodo.7588944

PID metadata schema extended report https://doi.org/10.5281/zenodo.7588902

The service is part of KonsortSWD project deliverable, NFDI funding number 442494171

nfdi

KonsortSWD

gesis
Leibniz Institute
for the Social Sciences