KonsortSWD

Consortium for the
Social, Behavioural, Educational
and Economic Sciences

Daten. Politik. Zukunft. 9| KSWD
Konferenz für Sozial- und Wirtschaftsdaten
27.–28. März 2023 in Berlin

9th Conference on Social and Economic
Data - KSWD 2023
Parallel Forums III (L)
28.03.2023

DOI: 10.5281/zenodo.7741688

nfdi

gesis
Leibniz Institute
for the Social Sciences

Use cases and benefits of persistent identifiers for dataset elements to foster reliable data citation

Janete Saldanha Bach
Claus-Peter Klas
Peter Mutschke

GESIS – Leibniz Institute for the Social Sciences

Authors

Janete Saldanha Bach

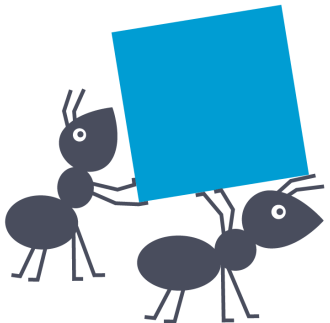
Dr. Janete Saldanha Bach, GESIS – Leibniz Institute for the Social Sciences. Postdoc in the NFDI consortium KonsortSWD in the department "Knowledge Technologies for the Social Sciences" , team FAIR Data and Human Information Interaction, working in the consortia KonsortSWD Project of the National Research 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.

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 lead 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.

Peter Mutschke

Peter Mutschke is deputy head of the department “Knowledge Technologies for the Social Sciences (KTS)" and leader of the team "FAIR Data and Human Information Interaction" 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). ORCID: https://orcid.org/0000-0003-3517-8071.



Agenda

- The PID Registration service for variables
  - Goal
  - The Research data granularity levels
  - Data formats: initial approach and future use
  - Hurdles of data citation current practices
  - PID registration service provider
  - Use cases
  - Takeaways for researchers, data provides and in terms of FAIRness

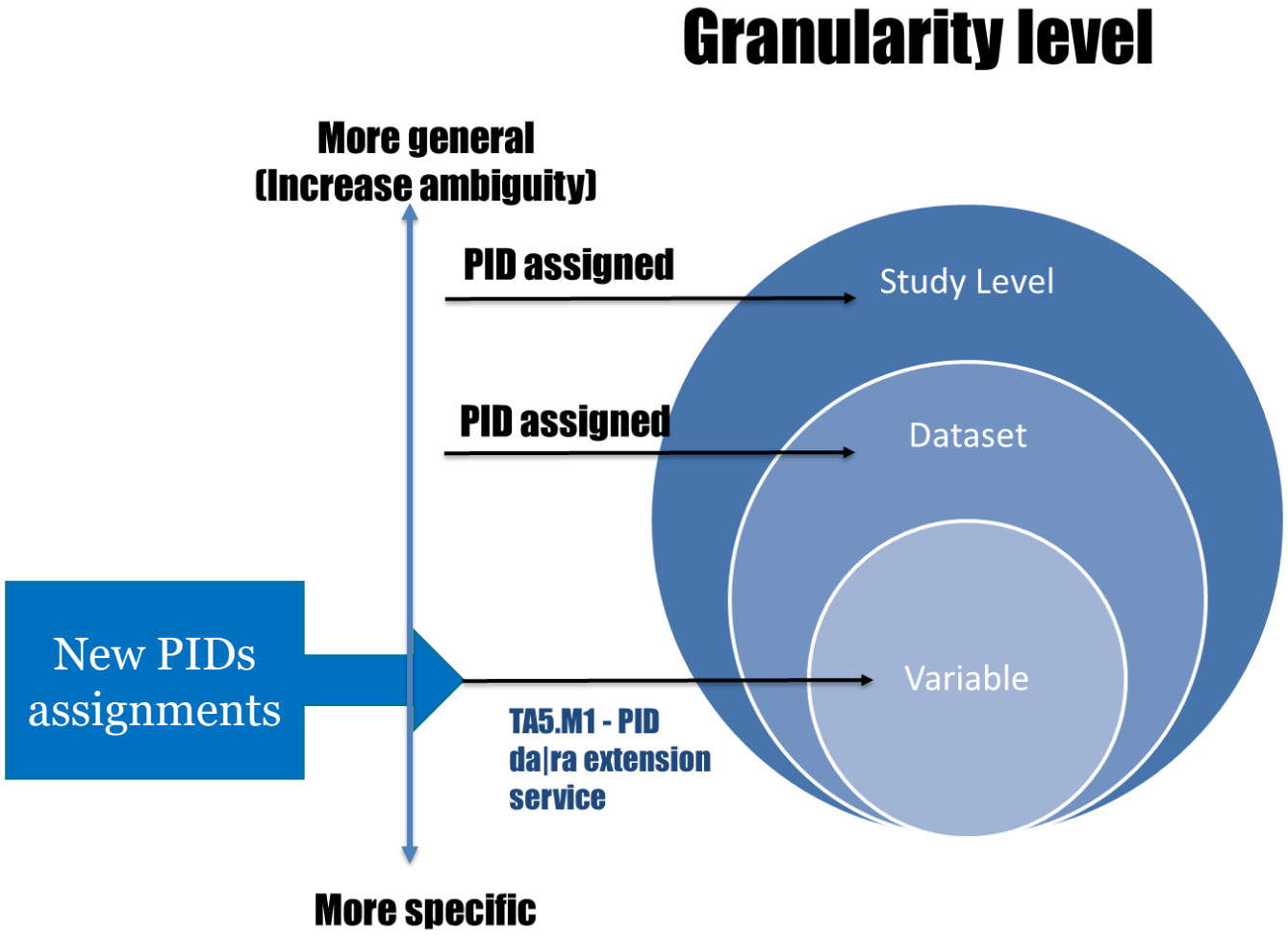
Goal

- Identify survey variables, **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

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

The Research data granularity levels

- Often, researchers do not use the **entire** dataset, rather a **set of variables**

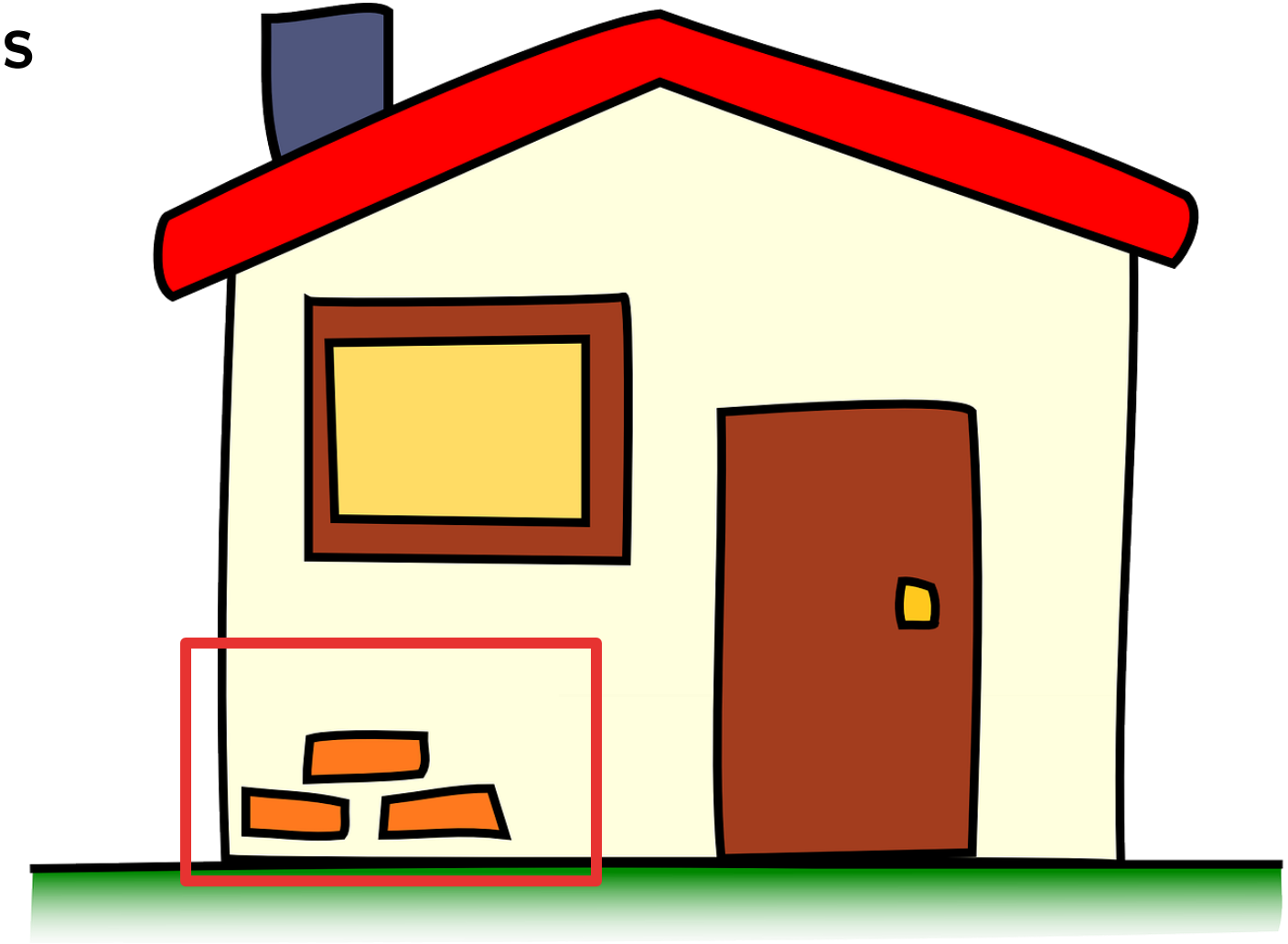
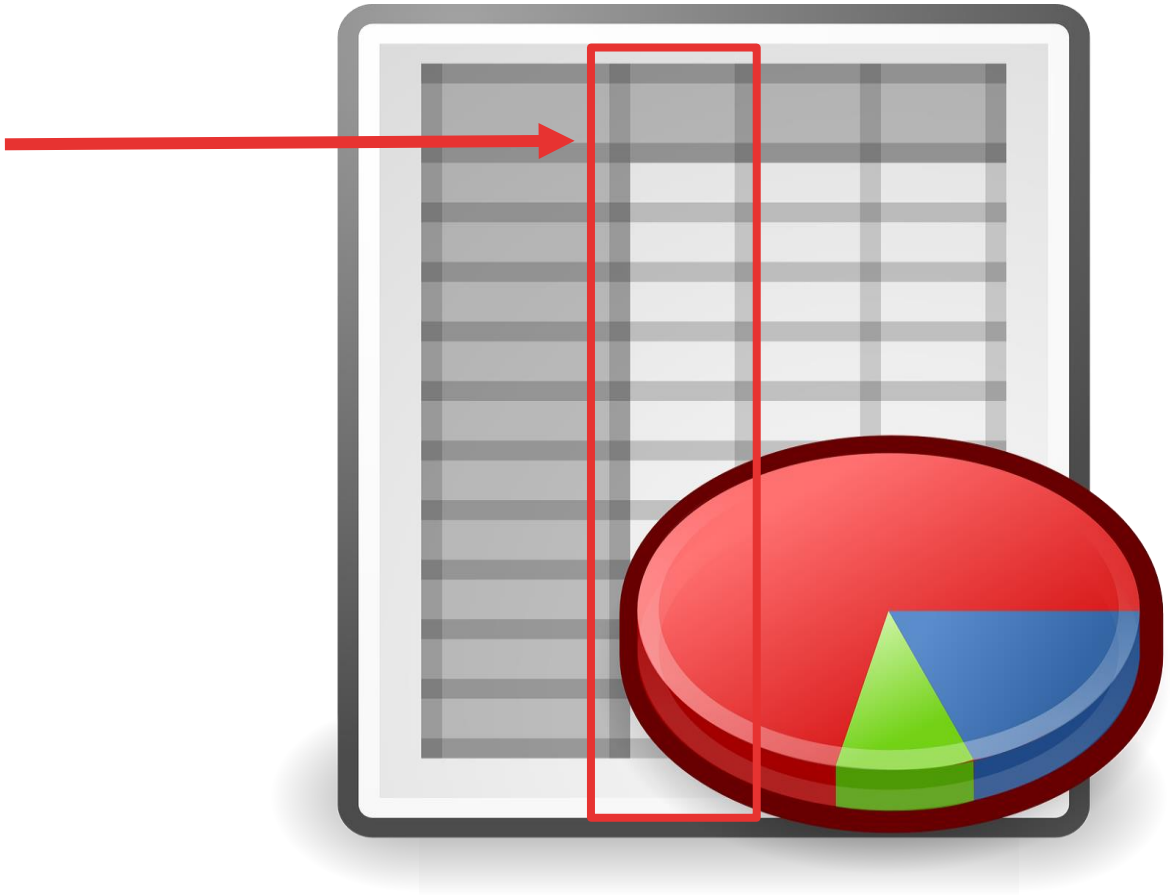


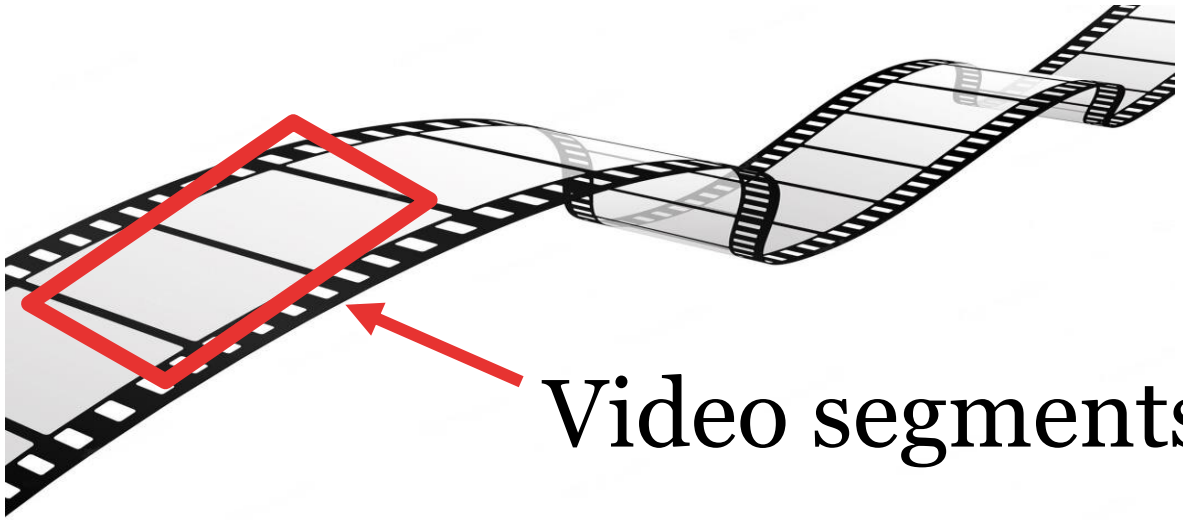
Image by Clker-Free-Vector-Images from Pixabay

Data formats: initial approach

Variables in tabular data format

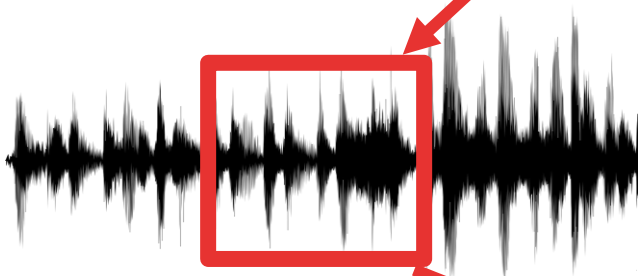


Data formats possible in the future



Video segments

Audio files



Transcripts

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]

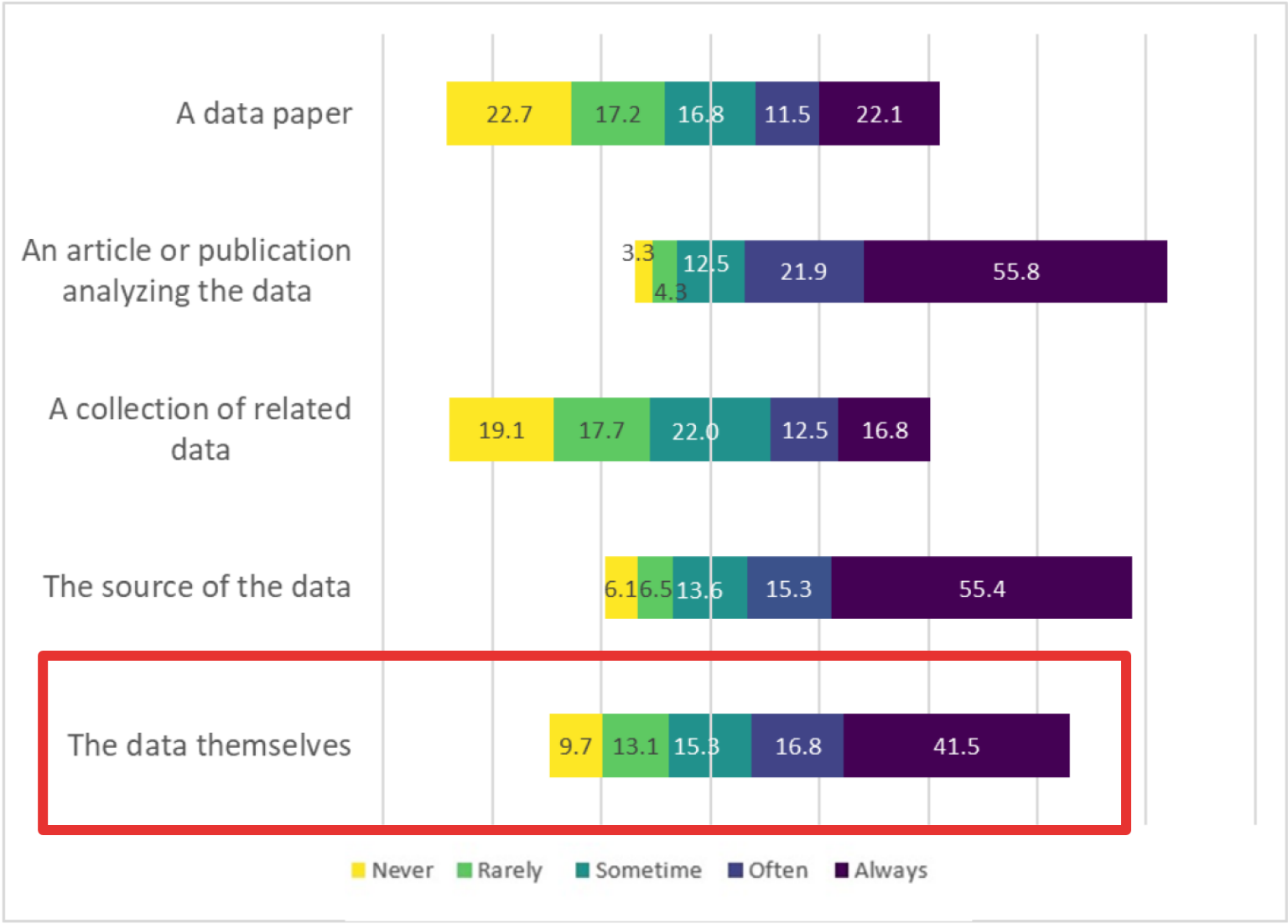
Ex. 2: Question cited in the text

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

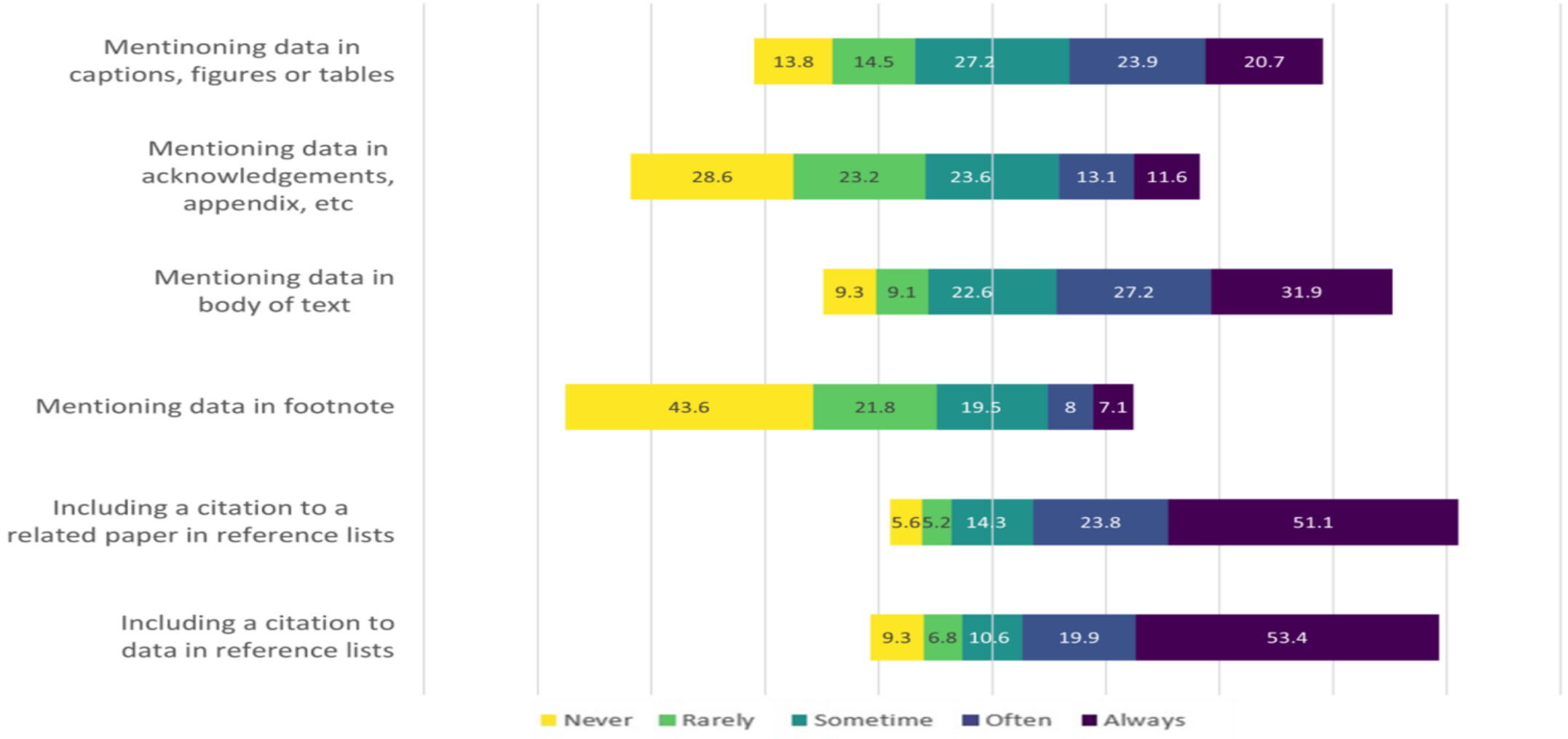
data re-users' data citation practices

- 7.7% cite another **publication** which analyzed data
- 70.7% cite the **source** of the data
- 58.3% refer to the **data themselves**



Preprint (10.5281/zenodo.7555266) under review at Quantitative Science Studies

Hurdles of data citation current practices



Preprint (10.5281/zenodo.7555266) under review at Quantitative Science Studies

Hurdles of data citation current practices

- Social Scientists prefer that other researchers cite their **data directly** instead of papers referring to the data [1]

- However, in contrast, Social Scientists, when citing data they reused, usually cite '**data studies' or other secondary sources** rather than the **datasets** (Robinson-García et al., 2016) [2]

[1] Preprint (10.5281/zenodo.7555266) under review at Quantitative Science Studies
[2] Robinson-García, N., Jiménez-Contreras, E., & Torres-Salinas, D. (2016). https://doi.org/10.1002/asi.23529

Data re-users' data citation practices

Variable: Q26_FOS
Label: OECD Fields of Research and Technology classification

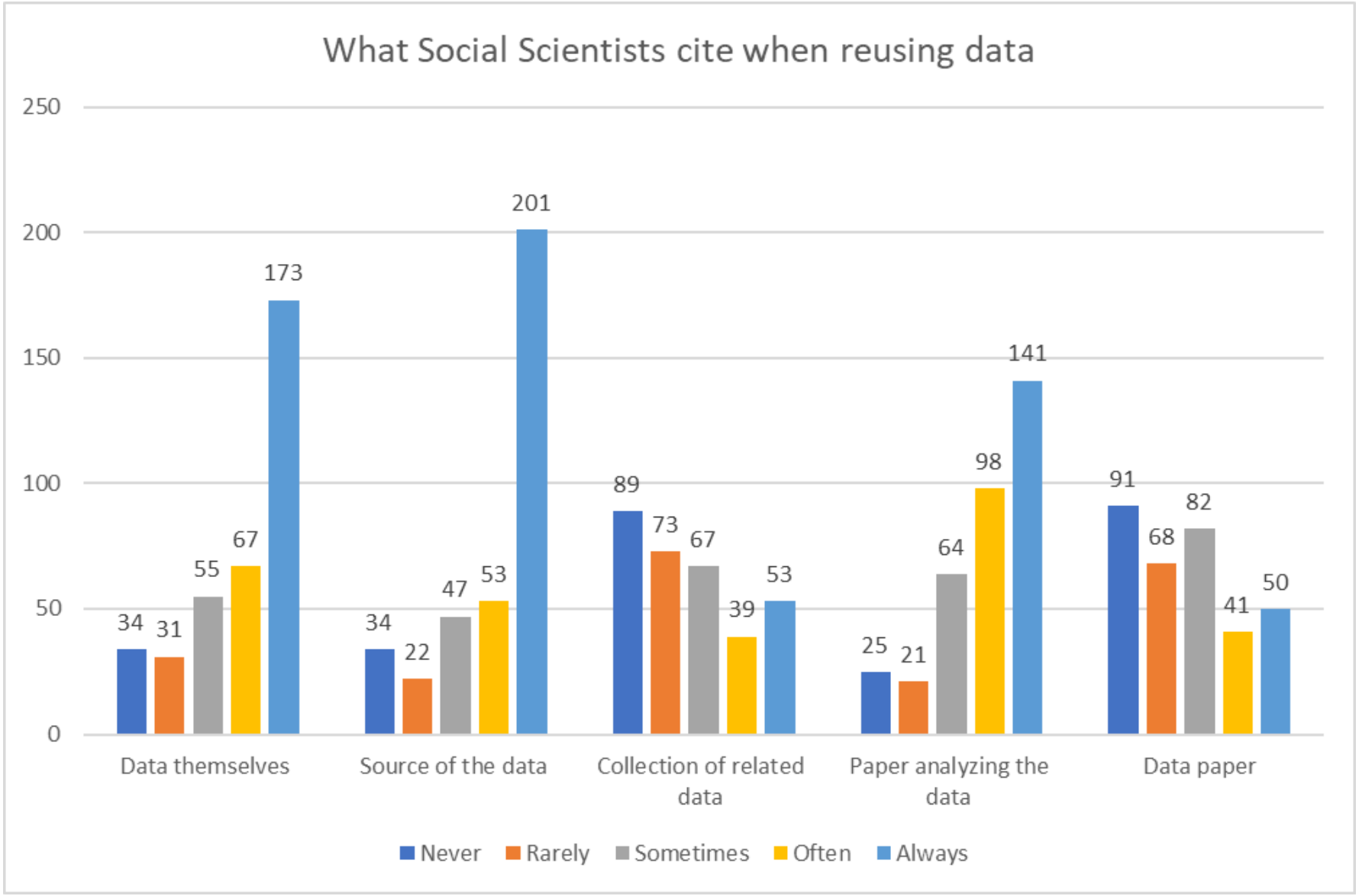
Question text: With which disciplinary domain do you most identify?

Universe: All respondents

Response categories	Code	Frequency		
Social sciences	5.00	3		
Psychology	5.01	98		
Economics and business	5.02	128		
Education sciences	5.03	47		
Sociology	5.04	51		
Law	5.05	5		
Political science	5.06	34		
Social and economic geography	5.07	18		
Media and communications	5.08	12		
Other social sciences	5.09	64		
N		**460**	2492	**(18,5 %)**

Ninkov, Anton Boudreau, Ripp, Chantal, Gregory, Kathleen, Peters, Isabella, & Haustein, Stefanie. (2023). A dataset from a survey investigating disciplinary differences in data citation (Version v1) [Data set]. Zenodo. **https://doi.org/10.5281/zenodo.7555363**

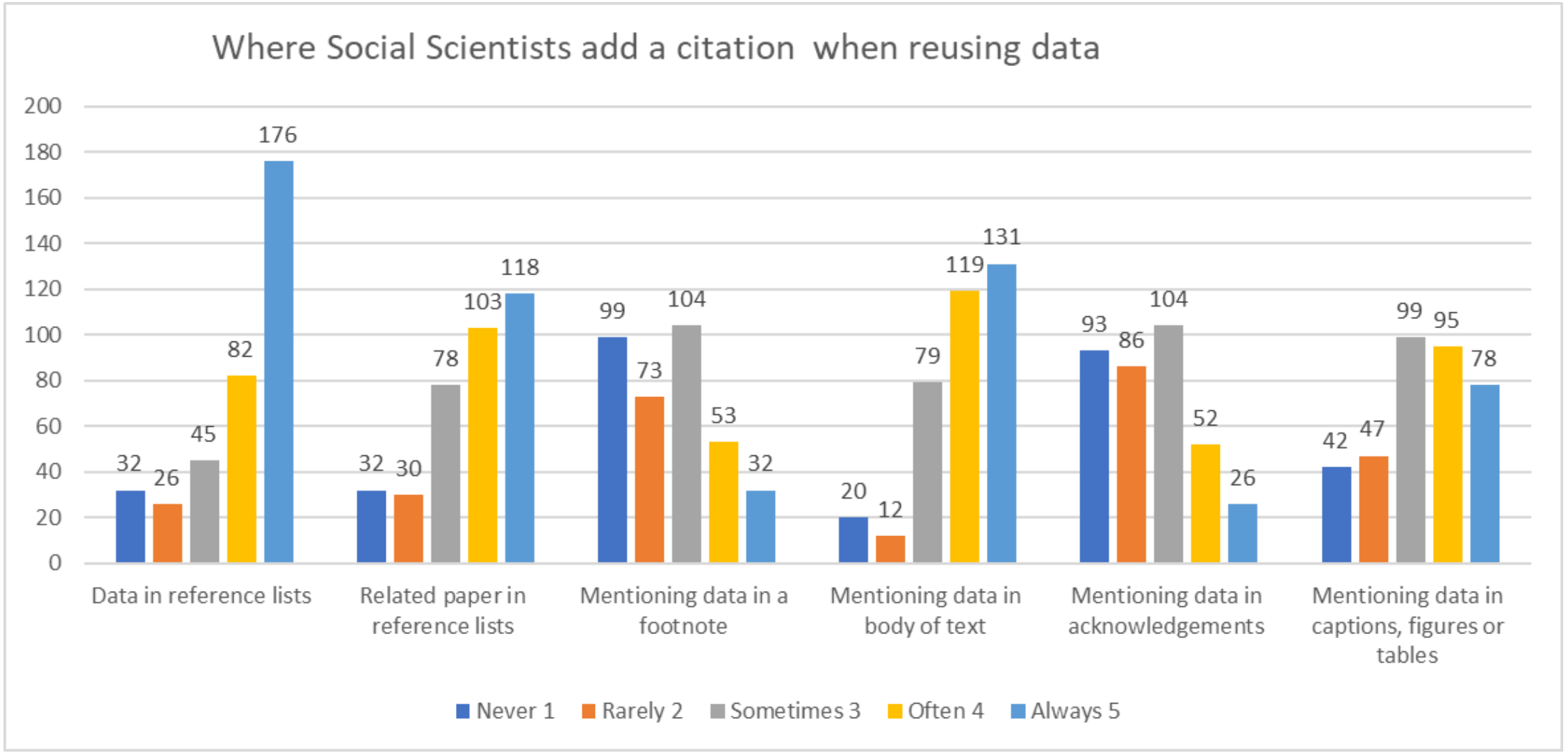
Data re-users' data citation practices



Ninkov, Anton Boudreau, Ripp, Chantal, Gregory, Kathleen, Peters, Isabella, & Haustein, Stefanie. (2023). A dataset from a survey investigating disciplinary differences in data citation (Version v1) [Data set]. Zenodo. **https://doi.org/10.5281/zenodo.7555363**

KonsortSWD

Data re-users' data citation practices



Ninkov, Anton Boudreau, Ripp, Chantal, Gregory, Kathleen, Peters, Isabella, & Haustein, Stefanie. (2023). A dataset from a survey investigating disciplinary differences in data citation (Version v1) [Data set]. Zenodo. **https://doi.org/10.5281/zenodo.7555363**

PID registration service provider

- This is a **da|ra** service **widening** and assigns a PID with **Handle** standard (ePIC);

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

da|ra

ePIC
Persistent Identifiers for eResearch

KonsortSWD

Use cases in a nutshell

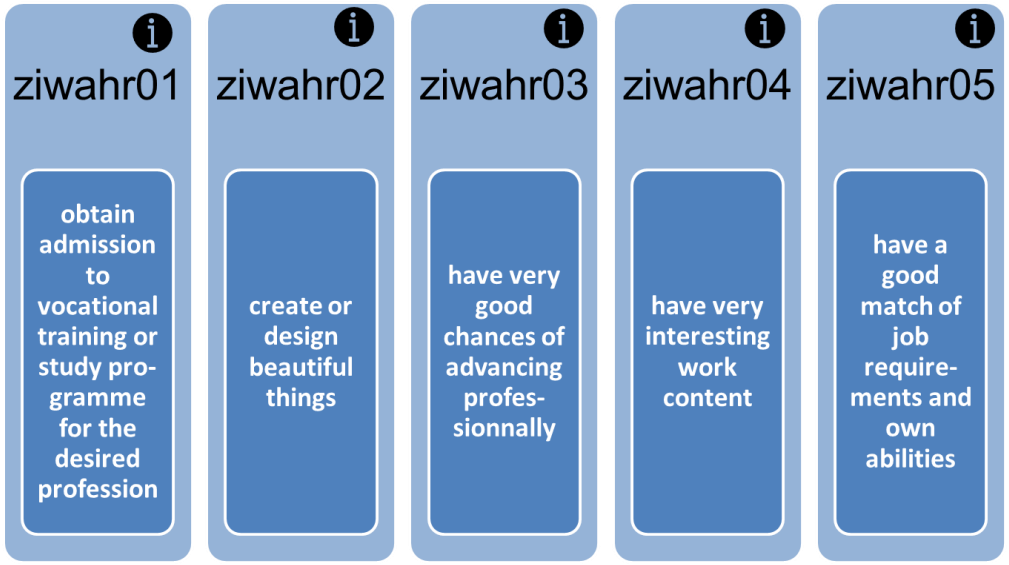
Institution name	DZHW | German Center for Higher Education Research and Science Research (DZHW)	GESIS | Leibniz Institute for the Social Sciences	GESIS harmonization tools			DIW | German Institute for Economic Research	Qualiservice | University of Bremen
			GESIS QuestionLink	ONBound - Old and new boundaries: 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

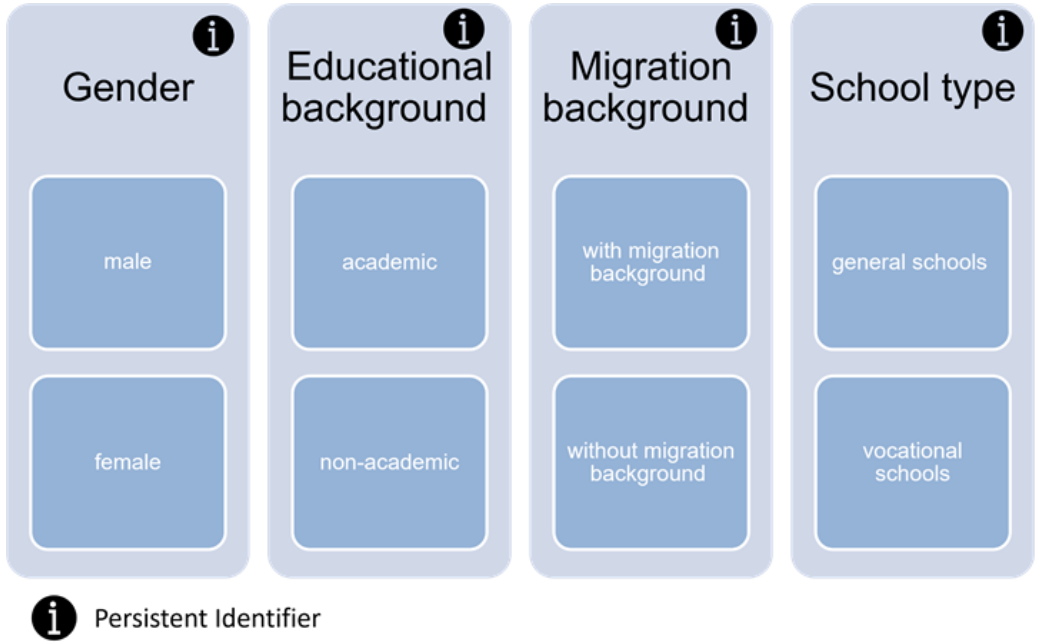
Use Case

DZHW
Deutsches Zentrum für Hochschul- und Wissenschaftsforschung

Higher Education Analytical Data System (HEADS) project at the DZHW needs a standard of data citation is to make its results widely usable and citable, particularly the entire information packages that comprise a central reporting variable (“indicator”) and the related multivariate analyses conducted in HEADS.

- PIDs for each variable, i.e., for each indicator or differentiation variables





Use Case

DZHW
Deutsches Zentrum für
Hochschul- und Wissenschaftsforschung

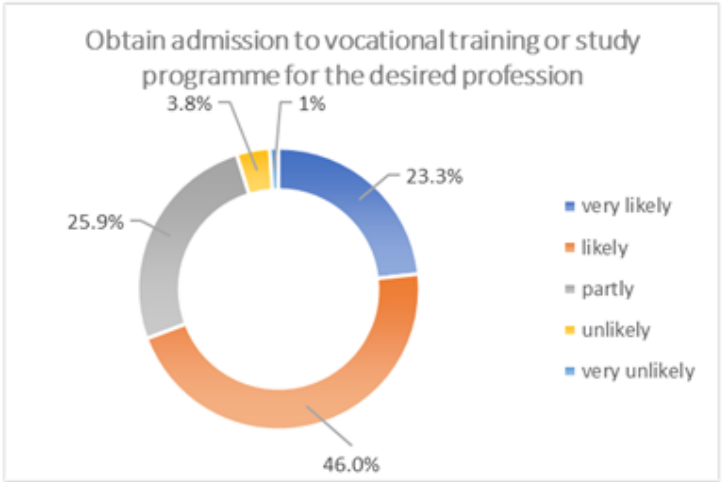
- PIDs for information packages

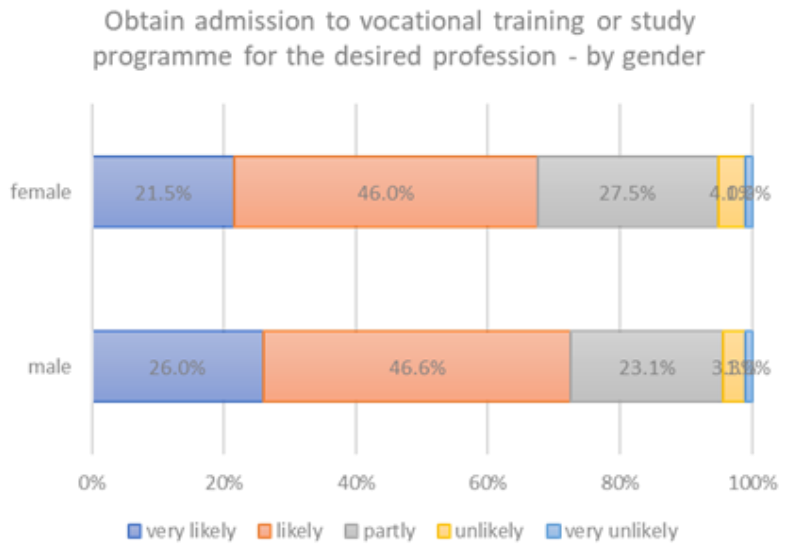
Students with university entrance qualifications in 2015 half a year before graduation

Indicator: Career goals - probability of realization
Question text: Everyone has goals. How likely do you think it is to achieve the following career goals?

Statement
obtain admission to vocational training or study programme for the desired profession (total value)

Feature
Gender





Options for differentiation variables
Gender
Educational background
Migration background
School type

Persistent Identifier

Use Case

gesis
Leibniz Institute
for the Social Sciences

- PIDs for more than 500,000 variables from 6,500 national and international studies covering various topics in the Social Sciences, Economics, and Behaviour Sciences.

Use Case

ONBound
Old And New Boundaries

HaSpaD

- PIDs for variables from harmonization tools and services
- Automatic access to variable data using scripts:
  - researchers are responsible for getting access to the datasets directly from the data providers;
  - With unique identifiers assigned, the data could be automatically accessed;
  - it makes it easier to use dozens of harmonised variables of the same topic from numerous diverse instruments.

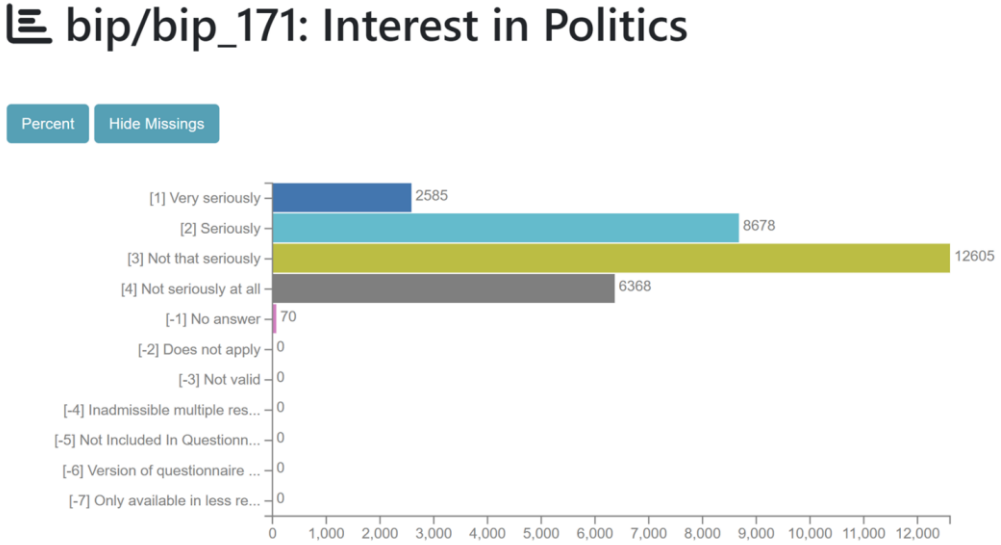
Use Case

SOEP The Socio-Economic Panel

Leben in Deutschland SOEP

Provide information on all household members: Germans living in the former eastern and western German states, foreign citizens, and immigrants residing in Germany. Some topics include household composition, occupational biographies, employment, earnings, health and satisfaction indicators.

- PIDs for 101.574 variables
  - available from 560 datasets,
  - distributed in 21.280 questions, and
  - 309 instruments



Variable graph: *bip/bip_171*: Interest in Politics

https://paneldata.org/soep-core/datasets/bip/bip_171

Use Case

Qualiservice consists primarily of qualitative interview transcripts and context data in text, videos, and description data.

- PIDs are assigned at the file level for disambiguating similar data types and file naming;
- Provides a direct way of citing, identifying, and getting the target file.

Takeaways for researchers

PIDs for **lower granularity level** simplify FAIR data usage because they:

- provide a **unique identifier for** data elements below study level, e.g. survey variables;
- **reference** and retrieve **individual** elements;
- **retrieve** metadata on data elements below study level;
- **disambiguate** data citation;
- enable safe and **accurate** data citation;
- increase **acknowledge** for produced **data**;
- foster **credibility** results and ensure the **sustainable reusability** of data;
- **reduce** documentation **complexity**;
- feasible identification **beyond rectangular data**, including other attributes, such as text, videos, and description data.

Takeaways for data providers

PIDs for **lower granularity level** simplify FAIR data management because RDCs can:

- get advantages of PIDs **machine-actionable** features, such as:
  - citation tracking and aggregating;
  - scientific production combination;
  - empowering authority;
- **track** and **monitor** the scientific outputs of a given variable;
- carry out a more **detailed evaluation** of the **dataset's** usage;
- **push** data **findability** and **accessibility** on the lower granularity level efficiently;
- improve **decisions making on services** based on data usefulness;
- **data governance** activities;
- potentially **explicit relations between variables** across studies and datasets (documenting *relation_types* metadata field), ground for **knowledge graphs** visualization;
- promote **digital connections** among researchers, organisations, and research outputs;
- **simplify harmonisation** processes, which are costly and time-consuming.

Takeaways in terms of FAIRness

Referencing research data and their inherited entities by PIDs **supports FAIR** data usage, since it:

- enhances data **findability**;
- facilitates easier (and automatic means, under conditions) **access** to data;
- boosts **interoperability** at a large scale by connecting variables and other individual elements;
- fosters data **reuse**;
- facilitates **reproducibility** of research.

Citation

Janete Saldanha Bach, Claus-Peter Klas and Peter Mutschke. 2023. Breaking down hurdles of current data citation practices: use cases and benefits of persistent identifiers for dataset elements. In 9th Conference on Social and Economic Data - KSWD 2023, 28 - 29 March. 2023. 28 slides. DOI: 10.5281/zenodo.7741688.

Acknowledgments

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