

Enhancing FAIR Research Data Management: advancing fine-grained data citation by PIDs for dataset elements.

Authors:

Janete Saldanha Bach^{1*} and Peter Mutschke²

*Lead presenter

¹ janete.saldanhabach@gesis.org, GESIS – Leibniz Institute for the Social Sciences

² GESIS – Leibniz Institute for the Social Sciences

Abstract:

Persistent Identifiers (PIDs) at the study or dataset level are insufficient for addressing the complexity of data management in research. The lack of granularity citation at the level of inline data objects, such as individual survey variables, qualitative data files, and even smaller data points, leads to ambiguities in data citation, inadequate metadata, and challenges in data discovery and reusability.

The PID registration service introduced by KonsortSWD [1], part of the German National Research Data Infrastructure (NFDI) [2], significantly advances the granularity of PIDs. This service supports the assignment of PIDs to these finer dataset elements and ensures accurate data citation since researchers usually use only a subset of the elements in a dataset. Data citation practices vary widely [3], and researchers often do not follow a standard such as the Data Citation Principles (Joint Declaration of Data Citation Principles [4]). The service boosts data referencing practices and improves research transparency and integrity. It also ensures adherence to FAIR principles by enabling precise referencing of individual data elements.

Beyond benefits for human access, such as how researchers can effortlessly find and cite data, Research Data Centres (RDCs) also benefit from PIDs in data governance activities and their services. PIDs' machine-actionable features have potential advantages, including citation tracking and impact measurement, linking articles using the same dataset elements, increasing data findability, and interoperability enhancement. It empowers the RDC's authority by demonstrating a commitment to best practices, extending its reputation within the research community by adopting PIDs at multiple granularity levels, such as recommended by the European Open Science Cloud (EOSC) PID policy [5]. The service facilitates streamlined FAIR research data management by integrating with established metadata standards [6] [7], fostering reproducibility. A single data point, the PID, enhances data citation, reuse, and direct access for automated access (i.e., by a computer program and under some requirements) [8], [9].

In terms of implementation, the technical solution [10] [11] employs the ePic API [12] and relies on the Handle standard [13]. Tests PIDs have been successfully applied to diverse datasets, including survey variables from GESIS and the German Institute for Economic Research (DIW). Tests are also currently taking place at the German Center for University and Science Research (DZHW) in 2024 and are planned at Qualiservice in 2025.

The PID service's design and operational framework align with broader NFDI initiatives such as PID4NFDI. PID4NFDI sets the foundation for a unified and FAIR-enabled PID harmonization towards NFDI-wide operation research data infrastructure. The PID registration service is necessary for harmonized services within the research data infrastructure across the NFDI consortia. It addresses critical gaps in data management by providing a stable and permanent way to reference and cite fine-grained data elements. This capability is crucial for ensuring that researchers can directly access specific data points without navigating entire datasets, facilitating more effective data reuse and verifying research findings. Moreover, by

serving as a base service under the umbrella of PID4NFDI, KonsortSWD's PID service provides a scalable framework that can adapt to many domains across the NFDI.

Keywords: Persistent Identifiers (PIDs), Granular Data Citation, Research Data Services - Technical Infrastructure, Social Sciences - Survey Variables, Dataset elements.

-
- [1] KonsortSWD (Consortium for the Social, Behavioural, Educational and Economic Sciences) is funded by the National Research Data Infrastructure (NFDI) <https://www.konsortswd.de/>
- [2] German National Research Data Infrastructure (NFDI): <https://www.nfdi.de/>
- [3] Gregory, K., Ninkov, A. B., Ripp, C., Roblin, E., Peters, I., & Haustein, S. (2023). Tracing data: A survey investigating disciplinary differences in data citation (Version v1). Zenodo. <https://doi.org/10.5281/zenodo.7555266>
- [4] Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 <https://doi.org/10.25490/a97f-egy>
- [5] European Commission, Directorate-General for Research and Innovation, Hellström, M., Heughebaert, A., Kotarski, R. et al., A Persistent Identifier (PID) policy for the European Open Science Cloud (EOSC), Publications Office, 2020, <https://data.europa.eu/doi/10.2777/926037>
- [6] Bach, Janete Saldanha, Klas, Claus-Peter, & Mutschke, Peter. (2023). *KonsortSWD Measure 5.1: Metadata schema extended report* (1.0). Zenodo. <https://doi.org/10.5281/ZENODO.7588902>
- [7] 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>
- [8] Klas, Claus-Peter, & Hopt, Oliver. (2022, December 7). DDI Variable Documentation and data access using R. EDDI - European User Conference on DDI, Paris, France. Zenodo. <https://doi.org/10.5281/zenodo.7408629>
- [9] Klas, Claus-Peter, Bach, Janete Saldanha, & Mutschke, Peter. (2023). *GESIS Use case Variable publication and citation & Fine granular access to research data*. <https://doi.org/10.5281/ZENODO.7750031>
- [10] Zhang, Y., Baran, E., Zloch, M., Mühlbauer, A., Klas, C.-P., & Mutschke, P. (2022). *KonsortSWD PID Registrator* (v0.1.0). Zenodo. <https://doi.org/10.5281/zenodo.5914219>
- [11] Baran, E., Zloch, M., Saldanha Bach Estevao, J., Klas, C.-P., Zhang, Y., & Mutschke, P. (2023). *KonsortSWD Measure 5.1: extended report on the architecture of the PID registration service*. In *KonsortSWD Measure 5.1: extended report on the architecture of the PID registration service* (1.0, pp. 1–18). Zenodo. <https://doi.org/10.5281/zenodo.10401641>
- [12] PIC is an international consortium that provides a reliable Handle-based PID infrastructure for research data. ePIC currently has nine members and is open to any centre that stores scientific/research data. ePIC Homepage: <http://www.pidconsortium.net/>.
- [13] Handle Homepage: <https://www.handle.net/>