Concept for Setting up the Persistent Identifier Services Working Group in the NFDI Section “Common Infrastructures”

Name of the working group
Persistent Identifier Services

Acronym
infra-pid

Contact (persons)
Philipp Wieder, GWDG, Text+
philipp.wieder@gwdg.de

Stephanie Hagemann-Wilholt, TIB, NFDI4Ing, NFDI4Chem
stephanie.hagemann@tib.eu

Authors
Sven Bingert (https://orcid.org/0000-0001-9547-1582),
Jan Brase (https://orcid.org/0000-0002-8250-6253),
Felix Burger (https://orcid.org/0000-0001-6836-1193),
Britta Dreyer (https://orcid.org/0000-0002-0687-5460),
Stephanie Hagemann-Wilholt (https://orcid.org/0000-0002-0474-2410),
Paul Vierkant (https://orcid.org/0000-0003-4448-3844),
Philipp Wieder (https://orcid.org/0000-0002-6992-1866)

Version 1.00
2022-04-28

Abstract

The working group consists of representatives of existing PID services and infrastructures who are already involved in various domain-specific consortia:

- DNB (Text+, NFDI4Culture): Provider of the URN-service for Germany (since 2001); shared lead of ORCID DE
- GESIS (KonsortSWD): da|ra: DOI registration for social sciences and economics
- GWDG (NFDI4Biodiversity, NFDI4Ing, Text+): Development, operation and provision of ePIC persistent identifiers; DONA MPA; partner in EOSC projects DICE & FAIRCORE4EOSC
- Helmholtz Open Science Office (./.): Project lead ORCID DE; project partner of re3data COREF, member re3data working group; strong connection to ROR; IGSN consortium & PIDs for instruments (RDA)
- IPK (FAIRAgro, NFDI4Biodiversity): PID registry for German crop plant genetic resources
- SUB Göttingen (Text+, NFDI4Biodiversity, NFDI4Culture): Lead of SUB DOI consortium in DataCite for DOI registration in the Humanities; member of ORCID
- TIB (NFID4Ing, NFID4Chem, NFDI4Culture, NFDI4DataScience): Lead of TIB DOI consortium as DataCite member, lead of ORCID DE consortium and metadata service as part of the German PID Competence Center
- DataCite (NFDI4Ing): DOI Registration Agency; Partner in ORCID DE and re3data COREF; partner in EOSC RDM and PID projects DICE, FAIRImpact, FAIRCORE4EOSC; NFDI4Ing seed funds project

The aim is to develop a common strategy for the implementation and extension of PID services that is closely aligned with the needs of NFDI consortia. Resulting solutions should enable FAIR research workflows balancing out generic metadata requirements for PIDs that maximise resource discoverability on the one hand and subject-specific needs on the other. At the technical level, the partners want to realise interoperability between PID types and established systems and build on a high level of maturity here; jointly developed services should be able to be rolled out for the entire NFDI.

Motivation and Objectives

In recent years, the use of persistent identifiers (PIDs) to identify data objects, general research outputs, or the researchers themselves has been widely accepted in the scientific community. Every existing and every proposed NFDI consortium uses one or more PID systems in its everyday handling of research objects. Furthermore, reliable PID systems are the backbone for many additional services, like knowledge graphs or portfolio analytics services. Thus, persistent identifiers are by definition a fundamental building block of research data management and a mandatory element of FAIR data infrastructures.

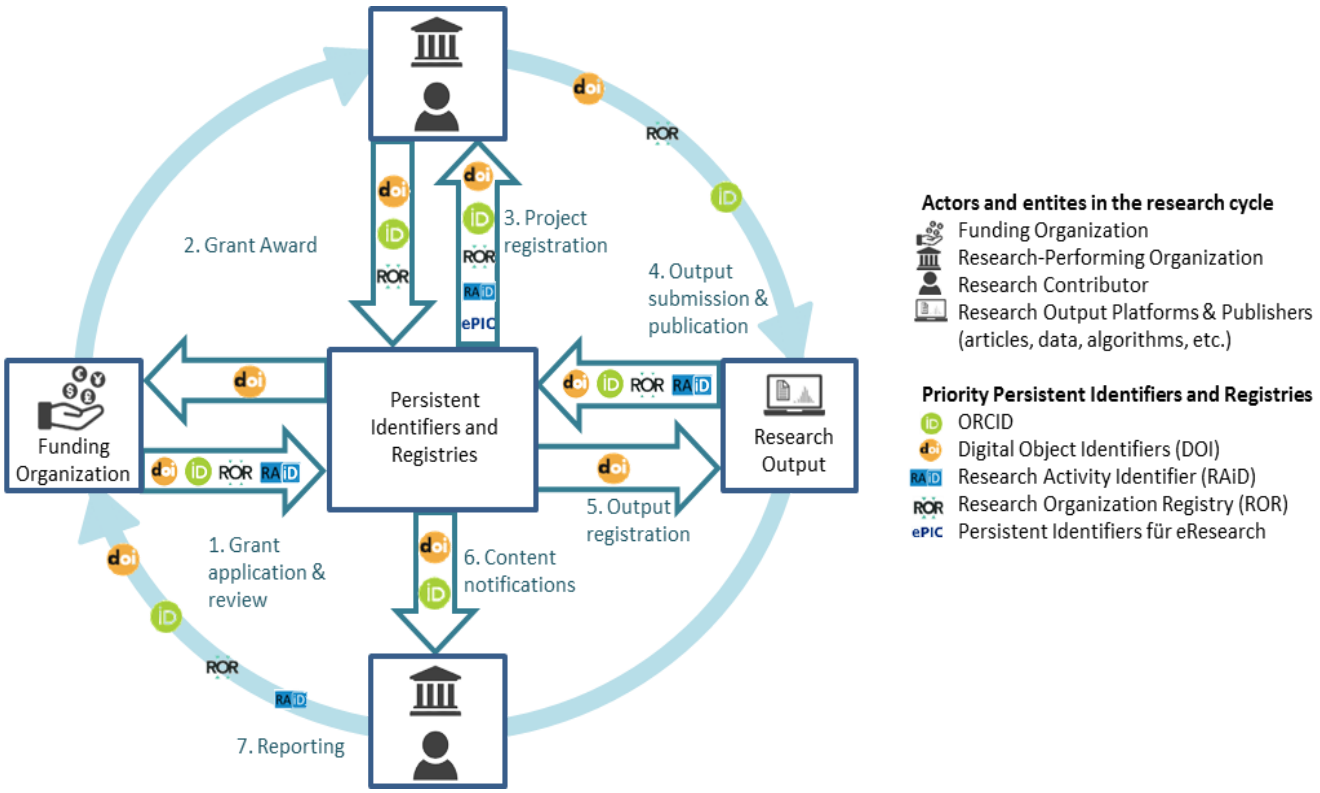
Globally operating organisations and consortia like DataCite, the DOI foundation, or the ePIC consortium already offer trustworthy, mature, and well established infrastructures that are used for research data identification by almost all NFDI consortia. Furthermore, systems for the persistent identification of individual entities such as persons, organisations, places, events, or for general terms (like ORCID, ROR, GND, or VIAF) are well established. Nevertheless, the assignment of identifiers in all NFDI consortia at the moment is scattered and heterogeneous in terms of actors, services, scope, quality, and costs involved. Having the core function of PIDs in RDM in mind, it is essential to analyse existing gaps and develop joint solutions in order to serve the needs of the individual communities and the NFDI as a whole. This implies finding answers to questions like: How can the existing infrastructure of PID services be optimally used and interoperability with global research

infrastructures like EOSC ensured? What interfaces and licences are needed? Which specific needs of consortia should be addressed? Where are central solutions - e.g. for support and training - needed?

Benefits to the NFDI

All stakeholders in the NFDI benefit from the use of PIDs in the research life cycle for unique identification of research resources, persons, institutions, projects, grants, outputs etc.. Their use enables more accurate, richer and more standardised metadata. Thus, researchers and research institutions benefit as producers and users of research output from improved discoverability and interoperability of research data. Funding organisations and publishers find support in selecting and recognizing reviewers. PIDs promote compliance with e.g. Open Access or Data Management Plan policies. They reduce the administrative overhead of metadata ingestion and the communication for all stakeholders by reducing the amount of manual metadata entries in communications systems (publication, reporting, project databases etc.) and provide more current metadata. Metadata entry errors are reduced through automated processes and metadata is validated through authorization and linking of metadata. Standardised and complete metadata is the key to the acceptance and success of PID services which can be improved through training and education on the benefits that fair and meaningful data management can provide. Services like the PID Graph - developed by the EU funded project FREYA - that aggregates information from PID metadata further increases the benefits and potentials described above by linking and contextualising information about research output, researchers, their institutions etc.. It can provide insights into the benefits of NFDI as a whole for the German research landscape and in international comparison.

Figure 1: Benefits of PIDs in the research life cycle



This Figure including description is based on Brown, J. et al. (2021). The PID-optimised Research Lifecycle. Zenodo. https://doi.org/10.5281/zenodo.4991733.

Processes of PID integration in research life cycle:

1. Grant Application and Review: Researchers and institutions pass PIDs for previous grants, outputs, organisations, people and projects to grant application systems. Funders ingest data about grants, outputs, organisations and projects from PID registries.
2. Grants Award: Funder register DOIs for new grants and associated metadata. Institutions ingest data about grants and associated people and organisations.
3. Project registration: Institutions register RAiDs and ePIC for projects and/update links to associated grants, equipment, people and organisations.
4. Output submission and publication: Researchers share their ORCID when submitting new outputs and connect ROR for institutional affiliations and grant DOIs for funding. Publisher/repository ingests data about grants, people, projects, and organisations linked to outputs.
5. Output registration: Publisher/repository register DOIs for new outputs and populates metadata.
6. Content notifications: PID registries send automatic updates of new publications etc. to institutions and funders.
7. Reporting: Researchers and institutions pass PIDs for outputs, organisations, people and projects to funders' reporting systems.

Objectives

- Objective 1: Understand NFDI PID use cases within domain specific consortia
- Objective 2: Develop NFDI-wide concepts for the practical implemention of identified PID Systems
- Objective 3: Disseminate knowledge to support implementation and usage of PID service(s)
- Objective 4: Develop NFDI PID governance and licensing models
- Objective 5: ollaborate with national & international stakeholder on PID related topics

Work Plan

WP1. Review (NFDI) PID landscape, collect requirements of domain specific consortia, analyse gaps

1.1. Analyse how PIDs are currently used and evaluate how this can be optimised to support research workflows, while keeping a project- and discipline-specific focus. Metadata quality tools and dashboards will be used to improve the completeness and quality of the metadata as key drivers for data FAIRness.

1.2. Conduct a PID use case-driven analysis across the NFDI consortia. Define, document and maintain PID requirements via structured requirement & engineering processes.

1.3. Map the identified PID requirements of the NDFI community to existing and established PID services.

WP2. Define technical & organisational measures to improve usage, compatibility & interoperability

2.1. Analyse which technical and organisational measures are needed to improve ease of use & integration of scalable PID services within the NFDI. Technical aspects (integration in infrastructure, software), organisational aspects (quality management of research data management: metadata interoperability, standardisation/harmonisation, workflow adaptation), as well as community aspects (identification of specific community demands, training for awareness & acceptance) need to be addressed.

2.2. Define standards for metadata interoperability and cross-walks within research disciplines, NFDI and internationally building on existing concepts and services. Focus initially on those that have advanced workflows/standards etc. and explore the extent

to which this can be made useful for the others at different maturity stages (scalability).

2.3. Develop concepts for NFDI infrastructure compatibility and interoperability with PID systems: Extending existing services - e.g. the PID Graph - to improve provenance of research data via connection with researchers, organisations, funders, enabling better documentation of research activities, data citation analysis.

2.4. Enable integration of PID systems in a functional NFDI architecture and highlight missing blocks (iterative process).

WP3. Support Base4NFDI and other potential service developers

3.1. Provide consulting and guidance on PID services. Develop training and communication concepts to raise awareness for PID potential and the importance of metadata.

3.2. Create cook books for simplified processes within the research workflows for PID registration (simple access) ensuring compliance with the FAIR principles

3.3. Create cook books/best practices for specific use cases & domains across NFDI, aligning with international standards.

3.4. Optimise the existing PID information and communication platforms to address NFDI needs.

3.5. Provide incentives for metadata curation (completeness): Both the DataCite Graph QL API and the web interface DataCite commons can be used to query the DOI, ORCID and ROR metadata and the connections, enabling metadata dashboards including usage metrics.

WP4. Establish PID-related governance & licensing within the NFDI

4.1. Evaluate the existing governance, business and licence models and sustainability of the PID solutions.

4.2. Adapt existing PID provider governance models and explore extensions based on NFDI requirements to fill the gaps.

4.3. Apply the governance, business and licence models for PID services in NFDI.

WP5. Collaborate with national & international networks & support for NFDI

5.1. Collaborate with NFDI Consortia, NFDI sections & (inter)national stakeholders and projects to ensure PID solution interoperability.

5.2. Contribute to & align with EOSC PID Policy.

5.3. Ensuring dialogues with international research communities and PID providers.

Collaboration Plan

The working group will collaborate with the following subgroups:

- *Identity and Access Management (AAI)*
- *Data Integration*
- *Data management planning (DMP)*
- *Long Term Archiving and Access (LTA)*
- *Fair Digital Object Architecture (FDO)*
- *Metadata Schemata and Application Profiles*
- *Knowledge Graphs*

- *Research Software Services*

Furthermore it is intended to coordinate efforts with the section “(Meta)data, Terminologies, Provenance” and the section “Training & Education”.

The European Open Science Cloud (EOSC[1]) is an integrated infrastructure to create a web of FAIR data[2]. The development of EOSC is a significant and ongoing multi-stakeholder initiative with a large number of associated projects that build services integrated into the overall EOSC landscape. The EOSC Persistent Identifier (PID) policy[1] establishes service and infrastructure requirements for potential services providers. Also, the PID technical architecture document[3] presents guidelines on the implementation of compliant PIDs and related services with PID EOSC Policy. It also identifies opportunities for interoperability between PID services and the EOSC framework. In this sense, these policy-related and technical guidelines shape the high-level aligning of PIDs4NFDI with EOSC. Example projects contributing to PID services for ESOC are FREYA, which built innovative PID-related services as a building block for EOSC, DICE, which offers i.a handle-based PID services for European researchers, and the upcoming FAIRCORE4EOSC project, which will develop i.a. a PID graph, metadata schemata, and a PID registry. Within all projects, PIDs4NFDI partners are contributing to the European service development as well as to international initiatives as the Research Data Alliance (RDA).

Initial Membership List

(Members from at least 6 institutions and at least 6 consortia)

Name	Institution	NFDI consortia	e-Mail
Michael Diepenbroek	GFBio	NFDI4Biodiversity	mdiepenbroek@pangaea.de
Jan Brase	SUB Göttingen	Text+	brase@sub.uni-goettingen.de
Jürgen Kett	DNB	Text+, 4Culture	j.kett@dnb.de
Janete Bach	GESIS	KonsortSWD	Janete.SaldanhaBach@gesis.org
Lorenz Reimer	DSMZ	NFDI4Microbiota	lorenz.reimer@dsmz.de
Philipp Wieder	GWDG	Text+, NFDI4Biodiversity	philipp.wieder@gwdg.de
Reinhard Altenhöner	SPK	4Culture	reinhard.altenhoener@sbb.spk-berlin.de
Britta Dreyer	TIB	4Chem, 4Culture, 4Ing	Britta.Dreyer@tib.eu
Stephanie Palek	DNB	Text+, 4Culture	s.palek@dnb.de
Sven Bingert	GWDG	NFDI4Biodiversity	sven.bingert@gwdg.de
Irina Sens	TIB	NFDI4Ing, 4Chem, 4DataScience	irina.sens@tib.eu

[1] Ferguson, Christine, McEntrye, Jo, Bunakov, Vasily, Lambert, Simon, Sandt, Stephanie van der, Kotarski, Rachael, Stewart, Sarah, MacEwan, Andrew, Fenner, Martin, Cruse, Patricia, Horik, René van, Dohna, Tina, Koop-Jacobsen, Ketil, Schindler, Uwe, & McCafferty, Siobhan. (2018). D3.1 Survey of Current PID Services Landscape (Version 1). Zenodo. https://doi.org/10.5281/zenodo.1324296.

[2] European Commission, Directorate-General for Research and Innovation, Schwardmann, U., Fenner, M., Hellström, M., et al., PID architecture for the EOSC: report from the EOSC Executive Board Working Group (WG) Architecture PID Task Force (TF), Publications Office, 2021, https://data.europa.eu/doi/10.2777/525581

Antonia Schrader	Helmholtz Open Science Office	Not directly involved in a consortia but coordinating the Helmholtz participations in the NFDI consortia	antonia.schrader@os.helmholtz.de
Bridget Murphy	CAU - Kiel	DAPHNE4NFDI	murphy@physik.uni-kiel.de
Matthias Lange	IPK	FAIRAgro	lange@ipk-gatersleben.de
Claus-Peter Klas	GESIS	KonsortSWD	claus-peter.klas@gesis.org
Peter Mutschke	GESIS	KonsortSWD, NFDI4DS, BERD@NFDI	peter.mutschke@gesis.org
Paul Vierkant	DataCite	NFDI4Ing	paul.vierkant@datacite.org
Felix Burger	TIB	NFDI4Ing	marleen.burger@tib.eu
Stephanie Hagemann-Wilholt	TIB		stephanie.hagemann@tib.eu