Towards a National PID Roadmap

- National Roadmap for Persistent Identifiers
- Use case: Registering and reporting research output
- Use case: Reusability and reproducibility of research output
- General recommendations
- Colophon

Towards a National Roadmap for Persistent Identifiers

Introduction

Two complementary movements have energized the need for improved information about research: Open Science and Responsible Management of Research Information. In this context, Persistent Identifiers (PIDs) play an important role. PIDs are also an essential part of the FAIR data principles. With the help of PIDs, digital objects can be identified and reused in a more persistent and less ambiguous way.

International adoption of PIDs such as ORCID for researchers and RoR for organisations, coincides with key developments in The Netherlands. A sample of Dutch initiatives that could benefit from coordinated use of PIDs include open access, open data, data management plans, and responsible research assessment, as well as the possibility of a national Open Knowledge Base. PIDs provide additional structure to research information (metadata) while also enabling durable links between research objects, institutions, funding awards, and researchers.

> **“Enter once, reuse often”**
>
> - ORCID (2018)

Towards a National Roadmap for Persistent Identifiers

A national PID roadmap

To address the possibility of employing PIDs in a coordinated way, and to find alignment between present and future initiatives, the PID advisory board (NWO, DANS-KNAW, UKB, SURF and CWTS-Leiden University) requested the development of a national PID roadmap. This request led to the installment of a working group with representatives of eScience Center, Utrecht University, Leiden University, 4TU, Koninklijke Bibliotheken, DANS-KNAW, Saxion and SURF. Their work resulted in the present document, which provides a first step for engaging the broader community on the content and potential of a national PID roadmap.

Our approach

Since PIDs are a means to improve the quality and efficiency of research information, the working group approached this task by focusing on common use cases. The following use cases were considered:

1. Registration and reporting research
2. Reusability and reproducibility of research
3. Evaluation and recognition of research
4. Grant application
5. Researcher profiling
6. Journal rankings

Use cases 1 and 2 were selected for this task. Each is described, along with relevant entities and associated PIDs.

Towards a National Roadmap for Persistent Identifiers

Survey of recent developments

In order to obtain a broad understanding of the status of PIDs we surveyed recent developments that are of interest to the two use cases that we will describe.

International level

- Development and implementation of PID policy by EOSC.
- Development of National PID strategies, e.g. UK and AU, and the alignment of those national PID strategies via Research Data Alliance.
- Knowledge Exchange researching well-functioning PID infrastructure for research.

National level

- Action plan to lower the pressure of administration experienced by researchers, with attention to better flow of information between universities and research funders.
- Launch of the programme recognition and rewards.
- Development of Open Knowledge Base (OKB-NL) and the growing interest in PIDs for the findability of research (information).
- NWO’s PID strategy.

Towards a National Roadmap for Persistent Identifiers

Defining recommendations for PIDs

The value of employing PIDs depends to a large extent on the specific systems and needs of the local context. By focusing on specific use cases we aim to identify the general suitability of the identified PID services to a range of stakeholders.
Another important consideration is the fact that different PID systems are at different stages of maturity and adoption. This affects the recommended actions to support a well functioning use case.

Selection of the two use cases was guided by the following questions:
➢ In which research information workflows would PIDs provide the basis for improved accuracy and increased automation?
➢ In what ways can we exploit existing PID services to optimise, or improve, the efficiency of research information workflows?
➢ Which workflows are sufficiently common to provide an informative example across different stakeholders?

"By well-functioning we mean that it is:

- *technically user-friendly and capable of uniquely and persistently identifying any digital object, deemed worthy of preservation.*
- *globally accepted (interoperable in its core design and technology) such that it independently of technology and geography always points to the data owners account and related metadata (i.e., resolves to an explanatory landing page), if not also the actual data.*
- *organizationally and economically sustainable, i.e., that the PID can still resolve even in the case of organizational change or economic turmoil - in principle for ever.*
- *politically trustworthy – in that there is minimal risk of sudden non-interoperability, legal obstacles or exploitive vendor lock-in."*

Belsø, Rene, Matthiesen, Martin, Parland-von Essen, Jessica, Béquet, Gaëlle, & KE Task & Finish Group for PID Risk & Trust. (2021). Risks and Trust in Pursuit of a Well-functioning Persistent Identifier Infrastructure for Research. Zenodo. https://doi.org/10.5281/zenodo.5018216

Towards a National Roadmap for Persistent Identifiers

Maturity of PIDs for the types of digital objects that are identified in this document, according to survey in 2018

Mature	Emerging		Immature	Unknown
Publication	Organization	Software	DMP	Metadata Schema
Dataset	Grant	Instrument		Data Type
Person / Author	Project	Sample		Method
	Infrastructure			Data Format

Christine Ferguson, et al. (2019). D3.1Survey of Current PID Services Landscape - Revised (Version 2). Zenodo. https://doi.org/10.5281/zenodo.3554255

Use case:
Registering and reporting research output

Registering

Registration of research output is necessary to report to funders like NWO, ZonMW, SIA, etc. for monitoring and evaluation of research (e.g. according to SEP or BKO protocols). Persistent identifiers can be applied to ease the administrative burden. This results in better reporting, better information management and in the end better research information.

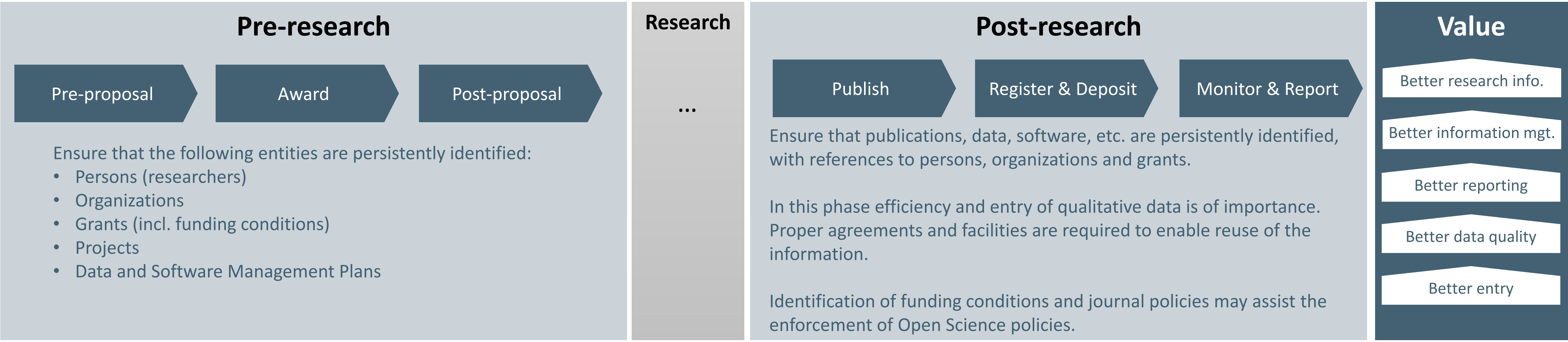
Differences between sectors

We have noticed that there is a difference in the application of PIDs between the Universities of Applied Sciences and Universities. PIDs are barely applied by non-scientific professional journals.

Use case:
Registering and reporting research output

General overview

Below we have included a general overview of whether usually PIDs are assigned before or after research:



Pre-research

Pre-proposal → Award → Post-proposal

Ensure that the following entities are persistently identified:
- Persons (researchers)
- Organizations
- Grants (incl. funding conditions)
- Projects
- Data and Software Management Plans

Research

...

Post-research

Publish → Register & Deposit → Monitor & Report

Ensure that publications, data, software, etc. are persistently identified, with references to persons, organizations and grants.

In this phase efficiency and entry of qualitative data is of importance. Proper agreements and facilities are required to enable reuse of the information.

Identification of funding conditions and journal policies may assist the enforcement of Open Science policies.

Value

- Better research info.
- Better information mgt.
- Better reporting
- Better data quality
- Better entry

Use case:
Registering and reporting research output

Relevant entities for the purpose of registration and reporting of research output

Relevant entities	Motivation (with examples)
Person	Person identifier is needed to connect research outputs with reseachers. Integration into an organization's systems requires mature and well-functioning PID infrastructure. **Example:** ORCID
Organization	Organization identifier is needed to identify to which institution, faculty or department a certain research output belongs to (might be more than one). The system should support the complex and dynamic structures of research institutions. **Examples:** ROR, Ringgold/ISNI, FunderID (Crossref)
Grant (incl. funding conditions)	Grant identifier is needed to identify how research output is funded. The identification of funder- and journal policies may support the reporting on compliance (e.g. with Open Science policies). The value of these identifiers should be investigated. **Examples:** DOI GrantID (Crossref), EC Funding ID
Output (Publication, Data, Software, DMP)	Output identifiers are needed to identify, refer and locate research outputs. The outputs should link to the related persons, organizations and grants. **Examples:** DOI, URN, ISSN, Handle, Etc.
Other (like project)	Other identifiers may be used to support the use case of registration and reporting of research output. In particular: Project identifiers support researchers to document their the ouputs during their project. It also helps organizations and funders to track which resources (people and infrastructure) and outputs are related to projects they fund or host. **Example:** RAiD

Use case:
Registering and reporting research output

Recommendation 1:

Improve adoption and integration of ORCID

ORCID is a well-functioning PID that is of added value for this particular use case if it is more widely adopted and integrated. To make sure that ORCID is embedded properly in the current infrastructure (funder's system like ISAAC, (institutional) repositories, CRIS systems, Portals, etc.) we recommend the following action:

- ➢ Agreements between stakeholders are necessary for assigning, updating and applying ORCID: every researchers must apply by themselves for an ORCID. Funders, institutions and repositories should implement ORCID in the systems used by researchers, and automate exchange between systems to ease administrative burden for researchers.

NB: This recommendation is in alignment with NWO PID strategy (2021)

Recommendation 2:

Support the development of Organization IDs

Organization identifiers are currently emerging in uptake and not easily applicable due to the university's dynamic organizational structure. Therefore, we recommend the following actions:

- ➢ SURF should support current developments of Organization identifiers, coordinate pilots with all stakeholders, and advise stakeholders in the responsible application and integration in their systems.
- ➢ Institutions should tend to the assignment and maintenance of their own organisation identifiers.

Recommendation 3:

Integrate the DOI GrantID

GrantIDs are emerging. To align with NWO PID strategy (2021), we recommend the following actions:

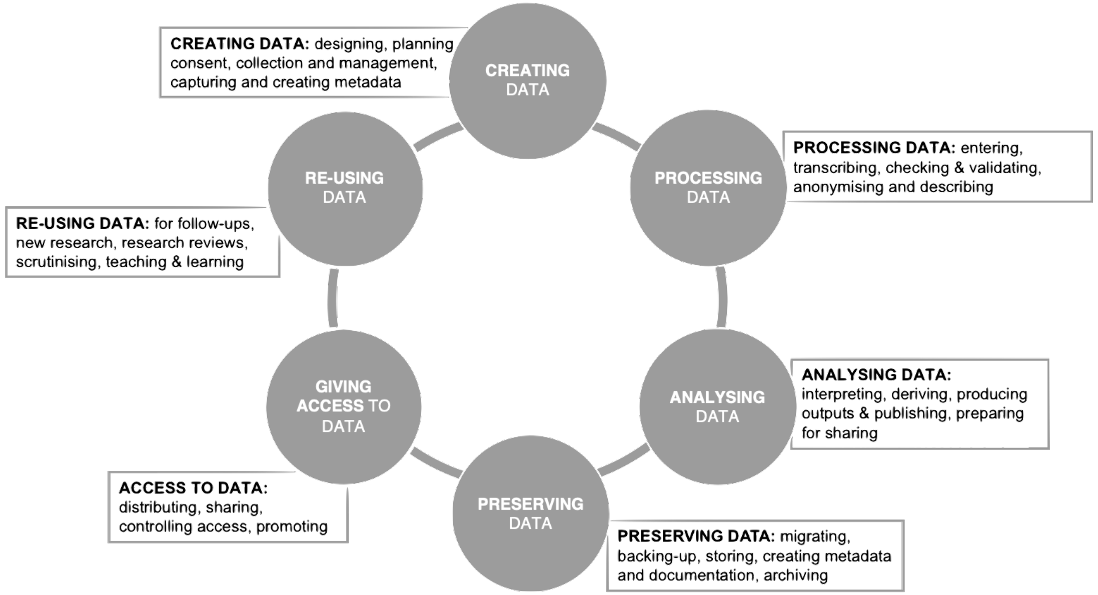
- ➢ Funders (NWO, ZonMW, SIA, EU, etc.) must align their PID strategies/interests and support the development of GrantID via Crossref Funder Advisory Board.
- ➢ Funders must assign GrantIDs to grants and make agreements with researchers and institutions for the integration of the GrantID into outputs.
- ➢ Funders must align their PID strategies with journals, repositories and CrossRef/DataCite to register and exchange GrantIDs in the systems

Use case:
Reusability and reproducibility of research output

To reuse and/or reproduce research it is desirable that research output be available with sufficient context and details for both humans and machines to be able to interpret the data as described in the FAIR principles ([FAIR data principles]).

For researchers to be able to reuse and/or reproduce research output, researchers need to understand the context in which the research output has been produced. Therefore, research output needs to be enriched with persistent contextual information about the dataset, owners, organisation, software, access level, license, communities, related publications, etc.
To be able to read and understand the content in a machine-actionable way, detailed information needs to be provided on how the data has been constructed (e.g. data format, data types and controlled vocabulary) and which software/tools (inc. version) and infrastructure has been used. Next, we will provide an example to showcase which information is currently available in a repository.

Data Lifecycle



CREATING DATA

CREATING DATA: designing, planning consent, collection and management, capturing and creating metadata

PROCESSING DATA

PROCESSING DATA: entering, transcribing, checking & validating, anonymising and describing

ANALYSING DATA

ANALYSING DATA: interpreting, deriving, producing outputs & publishing, preparing for sharing

PRESERVING DATA

PRESERVING DATA: migrating, backing-up, storing, creating metadata and documentation, archiving

GIVING ACCESS TO DATA

ACCESS TO DATA: distributing, sharing, controlling access, promoting

RE-USING DATA

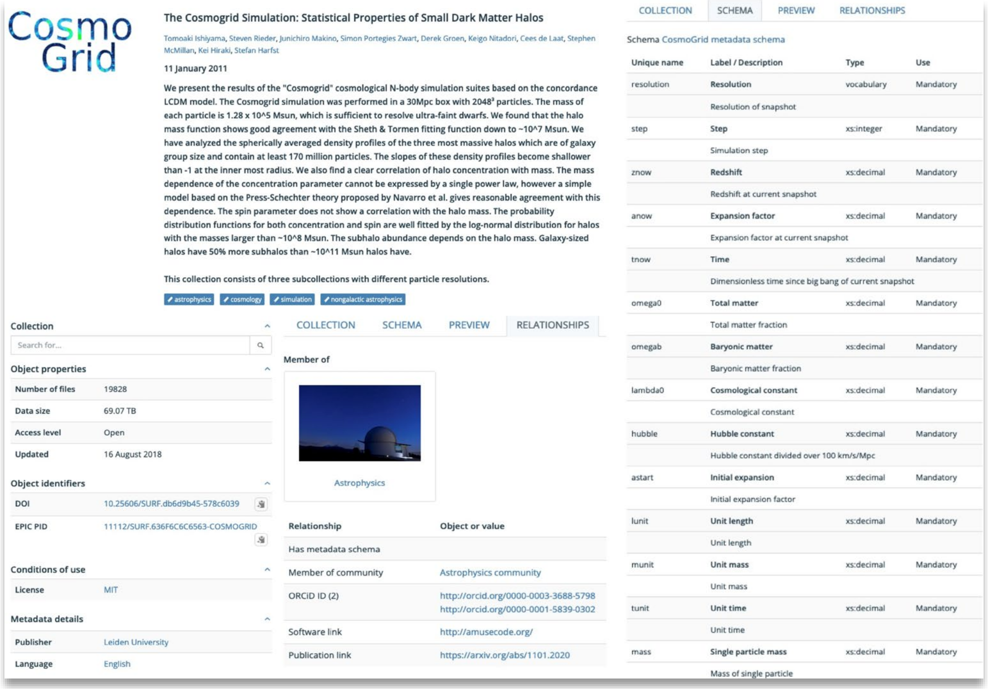
RE-USING DATA: for follow-ups, new research, research reviews, scrutinising, teaching & learning

Use case:
Reusability and reproducibility of research output

An example: CosmoGrid in SURFrepository

The CosmoGrid simulation (https://doi.org/10.25606/SURF.db6d9b45-578c6039) is just one example of how research data is currently stored in various repositories. Based on this example, our findings are:

- Dataset is referred by a DOI, and data objects are referred by an ePIC PID ('object identifier')
- Only two (2) individuals are identified by ORCID
- 'member of community' (Astrophysics community) has a PID but is locally defined
- 'Publisher' refers to a specific search within the repository on datasets published by the organization and not by e.g. a RoR ID
- 'License' does not refer to the license description
- 'Software' refers to the website instead of the software itself and/or the version used
- Reference to one (1) publication
- Metadata schema, data format and data types are defined, but not unambiguously in a machine-actional way



The Cosmogrid Simulation: Statistical Properties of Small Dark Matter Halos

Tomoaki Ishiyama, Steven Rieder, Junichiro Makino, Simon Portegies Zwart, Derek Groen, Keigo Nitadori, Cees de Laat, Stephen McMillan, Kei Hiraki, Stefan Harfst

11 January 2011

We present the results of the "Cosmogrid" cosmological N-body simulation suites based on the concordance LCDM model. The Cosmogrid simulation was performed in a 30Mpc box with 2048^3 particles. The mass of each particle is 1.28 x 10^5 Msun, which is sufficient to resolve ultra-faint dwarfs. We found that the halo mass function shows good agreement with the Sheth & Tormen fitting function down to ~10^7 Msun. We have analyzed the spherically averaged density profiles of the three most massive halos which are of galaxy group size and contain at least 170 million particles. The slopes of these density profiles become shallower than -1 at the inner most radius. We also find a clear correlation of halo concentration with mass. The mass dependence of the concentration parameter cannot be expressed by a single power law, however a simple model based on the Press-Schechter theory proposed by Navarro et al. gives reasonable agreement with this dependence. The spin parameter does not show a correlation with the halo mass. The probability distribution functions for both concentration and spin are well fitted by the log-normal distribution for halos with the masses larger than ~10^8 Msun. The subhalo abundance depends on the halo mass. Galaxy-sized halos have 50% more subhalos than ~10^11 Msun halos have.

This collection consists of three subcollections with different particle resolutions.

Collection

Object properties

Number of files	19828
Data size	69.07 TB
Access level	Open
Updated	16 August 2018

Object identifiers

DOI	10.25606/SURF.db6d9b45-578c6039
EPIC PID	11112/SURF.636F6C6C6563-COSMOGRID

Conditions of use

License	MIT

Metadata details

Publisher	Leiden University
Language	English

Member of

Relationship	Object or value
Has metadata schema	
Member of community	Astrophysics community
ORCiD ID (2)	http://orcid.org/0000-0003-3688-5798 http://orcid.org/0000-0001-5839-0302
Software link	http://amusecode.org/
Publication link	https://arxiv.org/abs/1101.2020

Schema CosmoGrid metadata schema

Unique name	Label / Description	Type	Use
resolution	**Resolution**	vocabulary	Mandatory
	Resolution of snapshot		
step	**Step**	xs:integer	Mandatory
	Simulation step		
znow	**Redshift**	xs:decimal	Mandatory
	Redshift at current snapshot		
anow	**Expansion factor**	xs:decimal	Mandatory
	Expansion factor at current snapshot		
tnow	**Time**	xs:decimal	Mandatory
	Dimensionless time since big bang of current snapshot		
omega0	**Total matter**	xs:decimal	Mandatory
	Total matter fraction		
omegab	**Baryonic matter**	xs:decimal	Mandatory
	Baryonic matter fraction		
lambda0	**Cosmological constant**	xs:decimal	Mandatory
	Cosmological constant		
hubble	**Hubble constant**	xs:decimal	Mandatory
	Hubble constant divided over 100 km/s/Mpc		
astart	**Initial expansion**	xs:decimal	Mandatory
	Initial expansion factor		
lunit	**Unit length**	xs:decimal	Mandatory
	Unit length		
munit	**Unit mass**	xs:decimal	Mandatory
	Unit mass		
tunit	**Unit time**	xs:decimal	Mandatory
	Unit time		
mass	**Single particle mass**	xs:decimal	Mandatory
	Mass of single particle		

Use case:
Reusability and reproducibility of research output

Relevant entities for the purpose of reproducibility and reusability of research output

Relevant entities	**Motivation (with examples)**
Data set	Identifier for data set is needed to locate and identify the source of the reused data. **Examples:** DOI, Handle
Person	A person identifier is needed to identify to whom the reused data belongs to, and possibly how to access the data (license, embargo, etc.). **Examples:** ORCID, ISNI, DAI
organization	An organization identifier is needed to identify to which institution the reused data belongs to, and how to access the data (license, embargo, etc.). **Examples:** ISNI/Ringgold, ROR
Research software	Identifiers for research software (including e.g. scripts, models, analysis pipelines and software packages) are important for reproducing research results, including reference to the specific version of the software. **Examples:** DOI and SWHID
Instrument, device, sensor, platform, research facility (infrastructure)	Identifying which instruments, devices, etc. are used by the initial researcher(s) is needed to reproduce the research since there might be differences in its application. Internationally the development of such a registry is supported by EUDAT, DataCite and ePIC. **Examples:** DOI, Handle, RRID, UID
Method	Identifying the research method for transparency and trust to reproduce research. This is (generally) described in the publication. **Example:** unknown
Publication	Because the research method is (generally) described in the publication it is important to ensure findability of publications. **Examples:** DOI, Handle, URN:NBN
Sample	Depending on the discipline, samples are collected and digitally represented in research. The identifiers used for samples differ because the type of sample may vary between disciplines. **Examples:** IGSN, ARK, URN, HTTP URI (CETAF URI), DOI, UUID, RRID, BioSample accession number
Metadata schema	A metadata schema identifier is needed to foster machine readability of the datatypes. There are developments for a metadata schema registry at an European level. **Example:** unknown
Data type	Identifying the combination of key value units is needed for machine readability purposes. EPIC currently develops a data type registry via Handle.
Data format	Data formats are standardized, but not yet identifiable.

Use case:
Reusability and reproducibility of research output

Recommendation 1:
Enrich context of research output

To foster trust and to be able to understand research, research output must be placed into context. This requires clear definitions on who owns and/or has contributed to the research output, which infrastructure and tools have been used to generate or analyse the dataset, and how data can be accessed and reused (e.g. access level and license). PIDs enable researchers to unambiguously refer and link to these entities. Therefore, we recommend the following action:

- ➢ Data repository platforms must allow researchers to include PIDs for contextual related information (e.g. PIDs for persons, organisations, software, algorithms/methods, infrastructure and/or instruments).

Recommendation 2:
Support FAIR and machine-actionability

Extending the reusability of research output, the FAIRness of research output must be increased. Using PIDs and describing digital objects via metadata in a machine-actionable way are among the main principles of FAIR. For machines to be able to interpret data and content unambiguously, IT systems require precise and persistent definitions on all related information describing the research output (e.g. metadata schema, data type and vocabularies). Therefore, we recommend the following action:

- ➢ We must have PIDs for metadata schemas, data types vocabularies within digital registration and repository systems.

Recommendation 3:
Apply DOI for research software

We recommend the use of DOIs for research software. For example, when a researcher develops software in a public repository such as GitHub, they can publish the software on Zenodo via the GitHub-Zenodo integration. There it is assigned a concept DOI, which is the general DOI for this software. When subsequent software versions are published on Zenodo in the same record, these versions are assigned version DOIs. Concept DOIs and version DOIs provide different levels of granularity when referring to the software: the concept DOI can be used to identify software as a project, while version DOIs can be used to identify a specific snapshot in time, which can be important for reproducibility. To encourage uptake of DOIs for software, we recommend the following actions:

- ➢ The assignment and registration of record-granularity-types, 'Concept DOI' and 'Version DOI', must be supported by data and software archives, research information systems such as Pure;
- ➢ Address DOI for software in data/software management plans, and policies of research organisations and funders.

General recommendations

During our PID working group meetings we have identified several concerns that can be translated into recommendations that are common across the use cases.

Recommendation 1:

Monitor emerging PIDs

There should be a team (or taskforce) to monitor emerging PIDs, and signal when a PID has reached a certain level of maturity based on the criteria of a well-functioning PID. A PID that is considered highly relevant for Dutch infrastructure could be purposefully contributed to to ensure it reaches the desired maturity level as soon as possible.

Maturity of PIDs for the types of digital objects that are identified in this document, according to survey in 2018

Mature	Emerging		Immature	Unknown
Publication	Organization	Software	DMP	Metadata Schema
Dataset	Grant	Instrument		Data Type
Person / Author	Project	Sample		Method
	Infrastructure			Data Format

Christine Ferguson, et al. (2019). D3.1Survey of Current PID Services Landscape - Revised (Version 2). Zenodo. https://doi.org/10.5281/zenodo.3554255

General recommendations

Recommendation 2:

Uptake and update of PIDs

In the Netherlands, several PIDs are being used and supported. An example is the national ORCID-NL campaign by the UKB (University libraries & Royal libraries). This campaign was organised to raise awareness and foster uptake amongst researchers. Further adoption (uptake and update) of well-functioning PIDs by the PID users in the Netherlands must happen.

Recommendation 3:

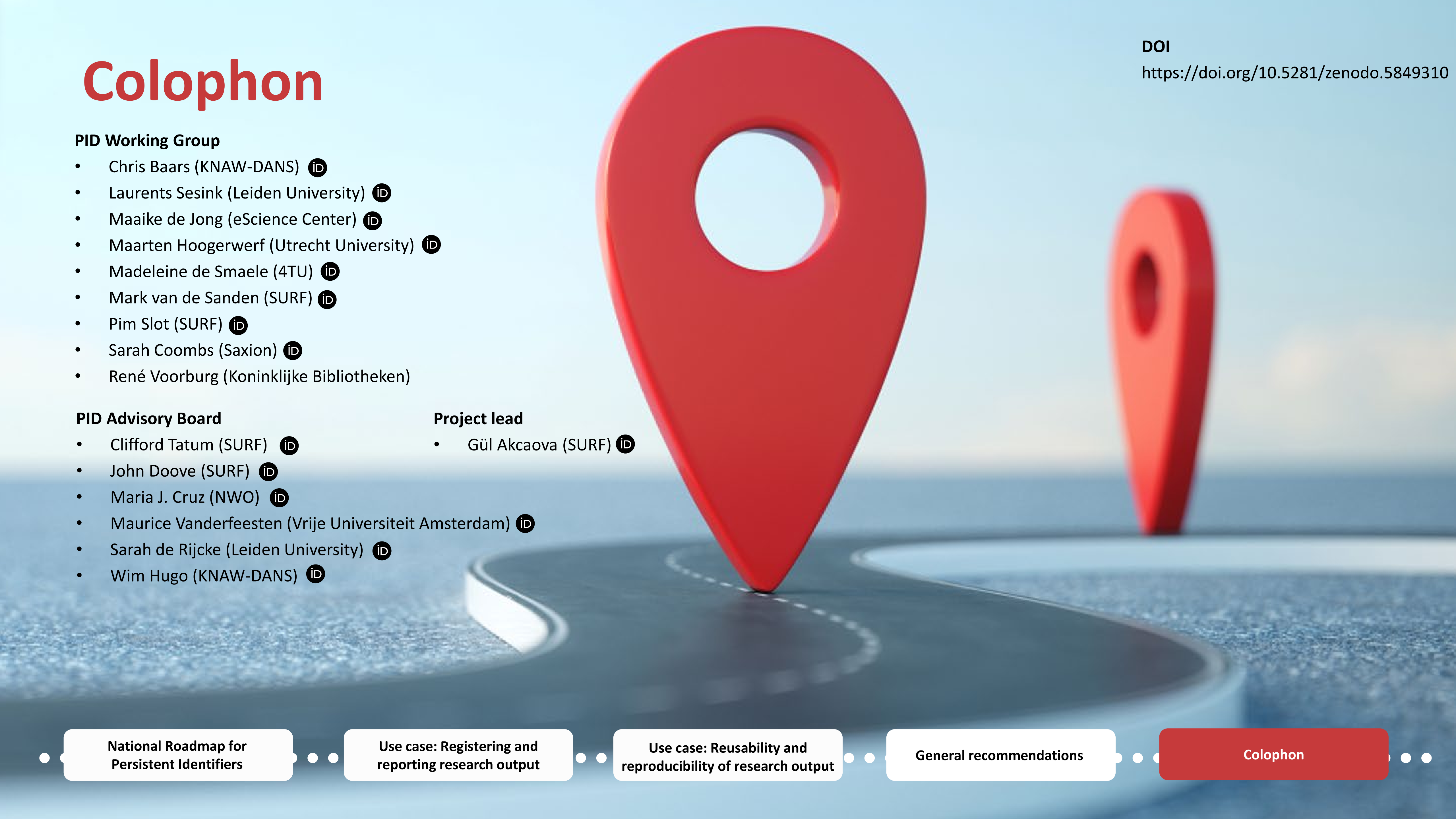
Agreement on vocabulary

For the interoperability of data, it is necessary to agree on a vocabulary. For example: It is not meaningful for a person to have an ORCID if it cannot be read by an application to identify if the person is author, creator or promotor. In the case of DOI for a publication, it should be possible to identify whether it is a peer-reviewed article or a book. Therefore, we must have a common language (controlled vocabulary).

Recommendation 4:

Develop and promote best-practices

The reusability of research output requires clear definitions on who owns a dataset or software, the access level, licence for reusage, just as details concerning how to interpret the data. Reproducibility has additional requirements for tools, software, software version, workflows, algorithms and which computing platform(s) have been used to produce the data. PIDs enable researchers to unambiguously refer and to link to these information. Therefore, we must promote and implement these PIDs into publication and data repository platforms.



Colophon

DOI
https://doi.org/10.5281/zenodo.5849310

PID Working Group

- Chris Baars (KNAW-DANS)
- Laurents Sesink (Leiden University)
- Maaike de Jong (eScience Center)
- Maarten Hoogerwerf (Utrecht University)
- Madeleine de Smaele (4TU)
- Mark van de Sanden (SURF)
- Pim Slot (SURF)
- Sarah Coombs (Saxion)
- René Voorburg (Koninklijke Bibliotheken)

PID Advisory Board

- Clifford Tatum (SURF)
- John Doove (SURF)
- Maria J. Cruz (NWO)
- Maurice Vanderfeesten (Vrije Universiteit Amsterdam)
- Sarah de Rijcke (Leiden University)
- Wim Hugo (KNAW-DANS)

Project lead

- Gül Akcaova (SURF)