
# Advanced infrastructure for PIDs in Photon and Neutron RIs

**Document Control Information**

| Settings | Value |
|---|---|
| **Document Identifier:** | D2.5 |
| **Project Title:** | ExPaNDS |
| **Work Package:** | WP2 |
| **Document Author(s):** | Vasily Bunakov (UKRI STFC), Rolf Krahl (HZB), Brian Matthews (UKRI STFC), Noel Vizcaino (UKRI STFC), Andrey Vukolov (Elettra) |
| **Document Reviewer(s):** | Oliver Knodel (HZDR), Sophie Servan (DESY), Darren Spruce (MAX IV) |
| **Responsible Partner:** | UKRI STFC |
| **Doc. Issue:** | 1.2 |
| **Dissemination level:** | Public |
| **Date:** | 03/03/2022 |

**Abstract**

This deliverable accumulates results of ExPaNDS works on the topic of persistent identifiers (PIDs) in facilities research. It refers to the substantial effort of persistent identifiers adoption in facilities and in a larger research ecosystem, and provides a general guidance, as well as prioritised recommendations for research information practitioners who consider going beyond the current levels of persistent identifiers adoption by photon and neutron research infrastructures.



*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

# Executive Summary

This deliverable accumulates results of ExPaNDS works on the topic of persistent identifiers (PIDs) in facilities research. It relates to other works in ExPaNDS, as persistent identifiers management and applications are rightly considered an important aspect of FAIR principles implementation. PIDs can be a valuable element of metadata for data catalogues, also they can be introduced and used at the data analysis stage of a research data lifecycle, or in the later stages of research impact assessment.

PIDs are not only for documents and data. They can also be used to reference other entities or agents, including people contributing to the research, software, research organisations, physical objects such as samples and instruments, and even abstract concepts such as terms in a controlled vocabulary. As a general rule, whenever something needs to be referenced in a reliable and lasting manner, a persistent identifier should be used.

This deliverable is an outcome of T2.4 “Persistent Identifier Infrastructure” that involved desk research, communication across ExPaNDS and with other projects, engagement with facilities staff, experimentation with metadata standards, community recommendations and software solutions, as well as dissemination activities.

The entire deliverable can be used as a reference by information practitioners in facilities and by facility managers when they plan for the introduction of persistent identifiers in the facility research lifecycle, or for using more types of persistent identifiers compared to those that are currently in use. It can be of help for facility users (visitor scientists), too, who are interested in raising their research information management culture.

There is the “PIDs guidelines for facilities” section towards the end of the deliverable that refers to other sections and suggests certain priorities for facilities to consider when designing, implementing and exploiting the PIDs and information services for managing them. These guidelines consider more PID types beyond PIDs for publications or data, specifically PIDs for facility instruments and organisations, for which reasonable services already exist and it is just a matter of good practice for facilities to adopt such PIDs.

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.*

# Table of Contents



*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

# 1. Introduction

A persistent identifier (PID) is a long lasting and unique reference to an entity or resource. PIDs are essential to reliably identify and find the referenced resource. The most common use case for PIDs in the scientific community are digital object identifiers (DOIs) to reference scientific papers: most publishers of scientific journals nowadays assign a DOI to each article they publish and researchers are familiar to retrieve that article following the DOI, and to use the DOI in order to reference the paper in a citation. The FAIR data principles[1] stipulate the use of PIDs for data: the very first guiding principle on findability "F1" requires that data and metadata are assigned a globally unique and persistent identifier.

But PIDs are not only for documents and data. They can also be used to reference other entities, including people contributing to the research, software, research organisations, physical objects such as samples and instruments, and even abstract concepts such as terms in a controlled vocabulary. As a general rule, whenever something needs to be referenced in a reliable and lasting manner, a persistent identifier should be used.[2]

## 1.1. This deliverable in the context of WP2, in relation to other ExPaNDS WPs and in relation to other projects

This deliverable presents an in-depth consideration of the best practice and usages of Persistent Identifiers for resources to support Facilities in the management, reuse and publication of FAIR data arising from experiments. It follows the deliverables D2.1[3] and D2.3[4] which recommended best practises in data policy for facilities, including a recommendation that the use of PID should be part of policy, and D2.2[5] which considered the whole data lifecycle for experimental science within facilities with recommendations for minimal metadata; this minimal metadata recommended assigning a PID within the publication record of resources, and considered the role of PIDs in assembling FAIR Digital Objects originating in facilities experiments.

The ExPaNDS project has conducted surveys on the current usage of PIDs across the partner facilities. The initial baseline was reported in the initial ExPaNDS Data Landscaping Survey undertaken in December 2019, with an update in 2021.[6] This demonstrated that partner facilities were in different stages of readiness in their implementation of PIDs. We use this information from surveys here as a starting point for this report and do not repeat their findings here.

---

[1] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[2] McMurry JA, Juty N, Blomberg N, Burdett T, Conlin T, *et al.* Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLOS Biology 15(6): e2001414 (2017). https://doi.org/10.1371/journal.pbio.2001414

[3] B.Matthews *et al.* Draft extended data policy framework for Photon and Neutron RIs. https://doi.org/10.5281/zenodo.4014810

[4] A.McBirnie *et al.* Final data policy framework for Photon and Neutron RIs. https://doi.org/10.5281/zenodo.5205824

[5] D.Salvat *et al.* Draft recommendations for FAIR Photon and Neutron Data Management. https://doi.org/10.5281/zenodo.4312824

[6] ExPaNDS session for gap and issue assessment. https://indico.maxiv.lu.se/event/5012/

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

PIDs are supported in the common infrastructure being developed in WP3 and WP4. WP3 is developing a federated data catalogue, centred around existing reference implementations of the ICAT and SciCAT systems supported or under development in the partner facilities. The federation is enabled by a common metadata Search API[7] that is developed jointly between ExPaNDS and the PaNOSC project; this Search API accommodates PIDs where appropriate.

The importance of PIDs in well-structured data management plans was recognised by the earlier ExPaNDS deliverable D2.4.[8]

A discussion with WP4 indicated an opportunity for the use of PIDs in the data analysis pipeline, which would allow clear provenance of analysed data, with references to raw data identified by PIDs and the software involved in data reduction and data analysis, again identified by PIDs.[9] The resulting data provenance graph would be a particular example of what was coined as “PID graph” by FREYA project, indicating rich opportunities for building all kinds of knowledge graphs and machine interfaces over them in cases when various contributing information entities are designated by persistent identifiers.[10]

Future work in ExPaNDS will take into account the recommendations of this report in the ongoing activities of the project. This will include the use of PIDs in FAIR assessment, common metadata and in DMPs (in WP2), in the evolution of a federated data catalogue (WP3) and in representations of data analysis workflows (WP4). Training and outreach material on PIDs will also be provided in WP5 and WP6; specifically, the WP5 are working on a training catalogue[11] that could host PIDs-related material.

## 1.2. The deliverable structure

The structure of this deliverable is as follows: first, we introduce methodology and terminology and give a context for PIDs-related works in EOSC policy making and architecture discussions inspired by FAIR principles, that is followed by a short overview of the persistent identifiers provision.

We then proceed with a practical overview of PID types available for their adoption by facilities. This overview is used for referring to it from other sections and is essential for the acquisition of a good background by anyone who wants to know what PID types are out there that can be adopted by facilities.

In the section that follows, we introduce the alternative approaches to minting and managing PIDs, and more complex PIDs uses beyond citations. As an example, producing metadata

[7] https://www.panosc.eu/wp-content/uploads/2020/12/D3.1_API-definition.pdf
[8] H.Görzig *et al.* DMPs for Photon and Neutron RIs. https://doi.org/10.5281/zenodo.5636095
[9] The WP4 deliverable with five data analysis pipelines identified is now in preparation, and this deliverable can be used for the follow-up discussion with WP4 about PIDs.
[10] Cousijn, H. *et al*. Connected Research: The Potential of the PID Graph. Patterns, 2(1),100180 (2021). https://doi.org/10.1016/j.patter.2020.100180
[11] The training catalogue for photon & neutron science. http://pan-training.eu

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.*

profiles or building PID graphs can be supported by API requests to the existing research information infrastructures, or data retrieval can be supported by PIDs.

We conclude with Guidelines that prioritise certain practices of persistent identifiers adoption by research facilities. Those who do not need the background information with deeper insights in the PIDs services landscape can proceed directly to the Guidelines section; or one can start with Guidelines then follow references from it to the previous sections with more background information.

The Appendices give examples of metadata records, interfaces, standards and recommendations, and point to additional documentation.

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

# 2. Methodology and terminology

## 2.1. Methodology

The deliverable aims to capture the results of works in T2.4 “Persistent Identifier Infrastructure” and to promote the practical use of persistent identifiers across various stages of facilities research lifecycle, using existing and emerging elements of the research information infrastructure.

Main methods used for the deliverable preparation have been:

- Desk research, to accumulate and to filter results of other works in relevant projects, working and interest groups, and by research information infrastructure providers
- Contributing to discussions across other tasks in ExPaNDS, in research community groups and in workshops, and getting the content of these discussions as an input
- Engagement with stakeholders in facilities: impact managers and instrument scientists
- Machine interfaces investigation and proof-of-concept experimentation with them

The task had communication with ExPaNDS WP3 and WP4 about proliferation of PIDs in federated data catalogue and in data analysis pipelines. The same is applicable to ExPaNDS sustainability considerations, as this deliverable can inform and support certain strains of photon and neutron community works beyond the project lifespan.

The T2.4 works were reported in the ExPaNDS Symposium for Librarians and data managers[12] that gathered librarians and data managers who work with and support Photon and Neutron (PaN) science facilities, with the focus on the interface between publications and data.

The T2.4 organised a dedicated workshop devoted to persistent identifiers for research facilities[13] where presenters gave an update on the state-of-the-art and recent developments in the use of persistent identifiers that are popular elsewhere but are not yet widely adopted by research facilities. The workshop aimed to facilitate a productive discussion across facilities about the actual and desired levels of persistent identifiers adoption across facilities research lifecycle, from the approval of experiments through data collection, analysis and towards the publishing of experimental results. Six presentations were followed by a round table discussion that collected ideas for future communication.

[12] Isabelle Boscaro-Clarke *et al.* ExPaNDS Librarian Symposium Report and presentations. https://doi.org/10.5281/zenodo.5973069
Video recording of the workshop is shared on the ExPaNDS website at https://expands.eu/presentations/

[13] Persistent Identifiers (PIDs) for Facilities Research Workshop - Presentations. https://doi.org/10.5281/zenodo.6024623
Video recording of the workshop is shared on the ExPaNDS website at https://expands.eu/presentations/

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.*

## 2.2. Other activities relevant to T2.4 and the deliverable

The Persistent Identification of Instruments working group in the Research Data Alliance explored solutions for the globally unique identification of instruments. Rolf Krahl was actively contributing to the work of that group, and the group presented in the workshop organised by T2.4.

The T2.4 had a targeted communication with PaNOSC about PIDs for instruments and beamtime awards as “grants-in-kind” that deserve their own unambiguous references, too. Some of the results of this discussion contributed to the “PIDs guidelines for facilities” section of this deliverable.

Brian Matthews was active within the EOSC Executive’s Architecture Working Group (2019-20), where he co-led the task force on PIDs. This group produced the PID Policy for the EOSC[14] in conjunction with the FAIR Working Group, and a reference architecture for PID Services, published as expert group recommendations for the implementation of the EOSC.

In EOSC Symposium 2021, the presentation[15] was made about the PID reproducibility in the context of FAIRness. This presentation promotes the tasks of this deliverable and also the possible influence of openly reproducible PIDs to data sharing workflows for involved RIs.

## 2.3. Terminology

| | |
|---|---|
| API | Application Programming Interface |
| DMP | Data Management Plan |
| DOI | Digital Object Identifier |
| EOSC | European Open Science Cloud |
| ESRF | European Synchrotron Radiation Facility |
| FAIR | Findable, Accessible, Interoperable, Reusable |
| FDO | FAIR Digital Object |
| IGSN | International Geo Sample Number |
| ISIS | ISIS Neutron and Muon Source |
| ISO | International Organisation for Standardisation |
| PaNOSC | Photon and Neutron Open Science Cloud |
| PID | Persistent Identifier |
| PURL | Persistent Uniform Resource Locator |
| RAiD | Research Activity ID |
| RDA | Research Data Alliance |
| RI | Research Infrastructure |
| SPDX | Software Package Data Exchange |
| URI | Uniform Resource Identifier, RFC3986[16] |

[14] 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

[15] Andrey Vukolov. (2021, June 18). Openly reproducible Persistent Identifiers (PIDs) as a factor of FAIRness in data sharing practices. EOSC Symposium 2021. Zenodo. https://doi.org/10.5281/zenodo.4980522

[16] Uniform Resource Identifier (URI): Generic Syntax, https://datatracker.ietf.org/doc/html/rfc3986

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.*

# 3. Role of PIDs in FAIR data policies and best practices

It is important to relate PIDs-related works in facilities with a wider policy making context, specifically because persistent identifiers are central to the notion of FAIRness as set out in the FAIR principles.[17] In particular, the following criteria explicitly recommend the user of persistent identifiers.

> F1. (meta)data are assigned a globally unique and persistent identifier
>
> F3. metadata clearly and explicitly include the identifier of the data it describes
>
> A1. (meta)data are retrievable by their identifier using a standardized communications protocol

PIDs play such a central role as they offer the universal addressability, stability of reference, and long-term persistence to support accessible and reusable research for the long-term. Thus any Integrated Research Infrastructure which supports FAIR data must support provision for PIDs, and this role is reflected in strategies to implement FAIR data, notably within the *the Final Report and Action Plan from the European Commission Expert Group on FAIR Data "Turning FAIR into Reality",* part of the effort to define the EOSC. This proposes the use of the notion **Fair Digital Objects (FDOs)** as the representation of accessible resources within a FAIR ecosystem of services and infrastructure. Essentially, a FDO is a package of data and metadata identified by PIDs sufficient to support their reuse, as follows:

> *"These [FDOs] could represent data, software, protocols or other research resources. They need to be accompanied by **Persistent Identifiers (PIDs)** and metadata rich enough to enable them to be reliably found, used and cited. Data should, in addition, be represented in common – and ideally open – formats, and be richly documented using metadata standards and vocabularies adopted by the related research community to enable interoperability and reuse. Software and algorithms, when shared, should include not just the source itself but also appropriate documentation including machine actionable statements about dependencies and licencing"*

The report includes the following recommendations:

> *Recommendation 2: Implement a model for FAIR Digital Objects Implementing FAIR requires a model for FAIR Digital Objects. These, by definition, have a PID linked to different types of essential metadata including provenance and licencing. The use of community standards and sharing of rich documentation is fundamental for interoperability and reuse of all objects. Action 2.1: The universal use of appropriate PIDs for FAIR Digital Objects needs to be facilitated and implemented.…*

[17] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

> *Recommendation 3: Develop components of a FAIR ecosystem The realisation of FAIR data relies on, at minimum, the following essential components: policies, Data Management Plans, identifiers, standards and repositories. There need to be registries cataloguing each component of the ecosystem, and automated workflows between them.*

Thus as a baseline, digital objects should be associated with a PID, and to support the use of PIDs in the data infrastructure, there should be PID services provided within the infrastructure.

A recommendation of the EOSC Executive's FAIR and Architecture Working Group's joint task force on PIDs proposed a PID Policy for the EOSC.[18] This is aimed at senior decision makers within EOSC service and infrastructure providers and: "*defines a set of expectations about what persistent identifiers will be used to support a functioning environment of FAIR research. Requirements of providers and the basic services they offer are also outlined."* The Policy defines minimal properties a PID should have to serve its function to support FAIR research. It then goes on to consider what are the minimum criteria that a particular scheme for providing PIDS, and the service provider supporting that scheme, should satisfy in order to be a *trusted PID service provider,* suitable for providing PIDs in the EOSC[19].

These recommendations are reflected in the work on data policy in both the ExPanDS and the PaNOSC projects. The ExPaNDS final recommendations on data policy for photon and neutron RIs[20] proposes as part of Recommendation 11:

> *The RI's data policy should include commitments to enable the experimental data in scope to be FAIR. This may include the following commitments:*
>
> - *The RI should provide the globally unique identification of experimental data via the association of an appropriate globally unique PID that conforms to the EOSC PID Policy.*

[18] 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://doi.org/10.2777/926037

[19] 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://doi.org/10.2777/525581

[20] A.McBirnie *et al.* Final data policy framework for Photon and Neutron RIs. https://doi.org/10.5281/zenodo.5205824

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

While the PaNOSC data policy template[21] proposes that a RIs data policy should include the clauses:

> *3.3. Persistent identifiers*
>
> *3.3.1. Persistent identifiers, for example DOIs, shall be generated for raw data and metadata. [see Implementation Note 7]*
>
> *3.3.2. Persistent identifiers shall be generated for processed data that is generated by facility-maintained automated systems.*
>
> *3.3.3. The experimental team shall be able to create a DOI for one or more specific datasets to be cited in a publication.*
>
> *3.3.4. Users shall cite the persistent identifier in any publication that refers to the data (or to a subset of the data).*

While the Implementation note 7 state:

> *The metadata should include the persistent identifier to the raw data. The recommended persistent identifier for raw data is the Digital Object Identifier (DOI) system. Where possible include other persistent identifiers in use at the facility e.g. Orcid for authors*.

Thus the PaNOSC data policy template presents a more concrete statement as a possible means to implement the ExPaNDS recommendation.

The EOSC Architecture Working Group went on to produce a Technical Architecture for PID Services within the EOSC[22]. This architecture describes a high-level view on components and stakeholders relevant for an architecture for PIDs. This PID document is mainly aimed at PID service providers as guidelines on implementation of PIDs and related services compliant with PID Policy within the EOSC, and is of less relevance to this report.

The programme of work in ExPaNDS, and in particular in WP2 seeks to explore how recommendations on the provision of FAIR data within a FAIR ecosystem can be instantiated within the Photon and Neutron science environment of the ExPaNDS partners. In particular, in this report, we consider how PIDs compliant to the EOSC PID Policy can be provided to support FDOs encapsulating experimental data and related resources.

[21] A.Gotz *et al.* PaNOSC FAIR Research Data Policy framework. https://doi.org/10.5281/zenodo.3826039

[22] 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://doi.org/10.2777/525581

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

# 4. PIDs infrastructures and services landscape

PIDs infrastructures consist of many elements, some of them heavier on technology and other leaned to standardisation effort, governance and best practices. In this section, we focus on just two aspects of PIDs infrastructures that present an immediate practical interest for the research information practitioner in facilities who intend to introduce persistent identifiers or needs more knowledge about the PIDs use elsewhere, in order to transfer and apply this knowledge in facilities research context. These aspects are: the PID providers (to better understand who can provide a service for minting PIDs) and the PIDs community landscape (to see where to find more information or to make inquiries).

## 4.1. PID providers

The Table 1 lists PID providers (PID services) of a general interest for facilities and facility users. Some of these providers own the infrastructure for issuing PIDs and some of them use the PID infrastructure by others in order to deliver their services on top of those "borrowed" infrastructures. We do not discriminate between these two flavours of PIDs provision, as from the user perspective, all these services can mint certain types of PIDs.

| **Provider *)** | **What PIDs** | **For what entities** |
|---|---|---|
| DataCite<br>https://datacite.org/ | DOIs | "Data" in a broad sense; can be also used for software, instruments and other entities |
| Crossref<br>https://www.crossref.org/ | DOIs | Publications,<br>Funders |
| ePIC<br>https://www.pidconsortium.net/ | Handles | Data and other entities |
| IGSN<br>https://www.igsn.org/ | IGSNs | Samples |
| ORCID<br>https://orcid.org/ | Bespoke (ORCID specific) | People |
| Figshare<br>https://figshare.com/ | DOIs (supplied by DataCite) | Papers, data and other research outputs |
| Zenodo<br>https://zenodo.org | DOIs (supplied by DataCite) | Papers, presentations, data, software and other research outputs |
| EUDAT B2SHARE<br>https://b2share.eudat.eu/ | Handles (supplied by EPIC) and DOIs (supplied by DataCite) | Data |
| FAIRsharing<br>https://fairsharing.org/ | DOIs (supplied by DataCite) | Standards, policies and databases |

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

| | | |
|---|---|---|
| ROR (Research Organisation Registry) https://ror.org/ | Bespoke (ROR specific) | Organisations |
| RAiD (Research Activity Identifier) https://www.raid.org.au/ | RAiD Handle (ARDC specific, inherited from Handle, backward-compatible) | Funders, organisations, collaborators, tools and services, datasets |

*) Some of these providers are not "native" PID providers but can mint a PID on a user's behalf; such cases are explicitly commented on, with the indication of an underpinning PID provider.

Table 1. PID providers (PID services) of a general interest for facilities and facility users.

Note that IGSN and DataCite concluded a partnership agreement. As a result, the IGSN ID will transition to the DataCite infrastructure, services and APIs.[23] The situation when a certain PID service has another PIDs service as a foundation is quite common. Other examples of that we are giving are Figshare or Zenodo that rely on the foundational DOI minting capabilities by DataCite, or EUDAT B2SHARE that issue both an ePIC handle and a DataCite DOI for the same record. From the user point of view, these "overlay" PID providers are PID services in their own right, so we do not discriminate against them when putting them in a common list of PID providers.
The conditions under which PID providers offer their services may significantly vary and involve some paperwork, even if the services are free of charge. As an example, Australian Research Data Commons (ARDC) is providing a free-of-charge model that requires adoption of the API and user registration, and ARDC acts as the only trusted provider as it is described in the Service Agreement[24]. Also the two-sided agreement needs to be ratified by the the both sides for automated PID generation[25].

The above table lists PIDs providers of a general interest to facilities, yet there are also providers that can mint identifiers of a high interest to certain branches of research that are supported by facilities experiments. One example of such more specialised providers is CCDC (Cambridge Crystallographic Data Centre)[26] who use the DOIs (supplied by DataCite) to designate crystallographic datasets. Another example is Identifiers.Org who assign namespaces for identifiers used in life sciences, and maintain a catalogue of such namespaces, which is an overlay information service on top of both new (emerging) identifier namespaces in life sciences, as well as for well-established ones such as Protein Data Bank.[27] It may not be a direct job of facilities to promote such more specialised PIDs, but information practitioners in facilities should be well aware of their existence, as they are circulating in significant segments of the facility users community.

[23] Buys, M., and Lehnert, K. (2021). Partnership between IGSN and DataCite. DataCite. https://doi.org/10.5438/7Z70-1155
[24] ARDC Standard Terms of Service. https://ardc.edu.au/wp-content/uploads/2020/10/ARDC_Service_agreement.pdf
[25] RAiD Service Schedule. https://documentation.ardc.edu.au/download/attachments/68290477/ARDC%20RAiD%20Service%20Schedule.docx?version=1&modificationDate=1637106138000&api=v2
[26] CCDC website. https://www.ccdc.cam.ac.uk/
[27] Identifiers.Org catalogue record for Protein Data Bank. https://registry.identifiers.org/registry/pdb

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

## 4.2. PIDs organisational and community landscape

PID providers operate using various business models: some of them are commercial organisations, some of them are organisations with membership (and membership fees) and some of them are supported by large organisations backed by sustained funding.

There is other PIDs-related effort though beyond PID providers. There has been substantial publicly funded effort devoted to PIDs, for which we can point out a succession of EU-funded projects ODIN[28], THOR[29] and FREYA.[30] Other substantial EU-funded projects and sustained organisations that emerged from them, such as OpenAIRE[31] or EUDAT[32] heavily use PIDs in their services and can be looked into by facilities as examples of PIDs adoption at scale.

There are a few past and current Research Data Alliance groups where PIDs have been the main focus or played a prominent role: Persistent Identifiers Interest Group[33], Persistent Identification of Instruments Working Group[34], PID Information Types Working Group[35], Open Science Graphs for FAIR Data Interest Group.[36] There has been an effort by RDA, too, for the coordination of national PID strategies.[37]

There is a community forum to discuss all topics related to the use of persistent identifiers in research that contains a knowledge hub, too.[38] It is the right place to go if facilities have some burning question about PIDs best practices, recommendations or metadata.

Persistent identifiers is a prominent topic in a number of conferences devoted to research information management, and the global PID community holds a periodic gathering that is exclusively devoted to PIDs.[39]

[28] ODIN (ORCID and DATACITE Interoperability Network). https://cordis.europa.eu/project/id/312788
[29] THOR – Technical and Human Infrastructure for Open Research. https://cordis.europa.eu/project/id/654039
[30] FREYA: Connected Open Identifiers for Discovery, Access and Use of Research Resources. https://cordis.europa.eu/project/id/777523
FREYA project website. https://www.project-freya.eu/
[31] OpenAIRE services. https://www.openaire.eu/
[32] EUDAT Collaborative Data Infrastructure. https://eudat.eu/
[33] RDA PID IG. https://www.rd-alliance.org/groups/pid-interest-group.html
[34] RDA PIDINST WG. https://www.rd-alliance.org/groups/persistent-identification-instruments-wg
[35] RDA PID Information Types WG. https://www.rd-alliance.org/groups/pid-information-types-wg.html
[36] RDA Open Science Graphs for FAIR Data IG. https://www.rd-alliance.org/groups/open-science-graphs-fair-data-ig
[37] National PID strategies: opportunities for collaboration and alignment RDA meeting. https://www.rd-alliance.org/national-pid-strategies-opportunities-collaboration-and-alignment
[38] The PID Forum. https://pidforum.org/
[39] PIDapalooza: The Open Festival of Persistent Identifiers. https://www.pidapalooza.org/

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

# 5. Types of PIDs to consider by facilities

The relationships between the PID types and the entities that PIDs designate are multilateral and complex. Some PID types only work for certain object types, such as ORCIDs for persons, so that the distinction may seem artificial. Other PID types, such as DOIs can be used for many different object types.

It becomes even more challenging if we take into account that some objects may have different aspects and different PID types may be applicable to the same entity depending on which aspect we consider. The most complicated example may be a proposal for facility beamtime: it may be considered a research activity or a project, an experiment, a grant, an umbrella for the data gathered during the experiment (in fact, data DOIs created by some facilities for raw data actually refer to the proposal description), or all of that at the same time.

Also the PID types adoption is uneven across different PID types, also with different flavours of adoption even within the same PID type across facilities, so the levels of maturity in regards to a particular PID type and its particular use by particular facilities may vary.

The below iteration of PID types is therefore a simplification of a more complex landscape of PIDs use, to make it more observable and more practical for research information practitioners in photon and neutron research infrastructures.

## 5.1. Research paper PID

Scholarly articles are nowadays commonly cited using a DOI. This is perhaps the most common and the best established type of PID in use. A facility may ask for the list of publications that support / justify the application for beamtime, and request bibliography in a certain format including a PID such as a DOI. Surprisingly enough, not all facility User Offices are giving due consideration to the use of PIDs in proposals; the references requested are commonly supplied in arbitrary bibliographic formats with no DOIs or other PIDs, which makes it difficult to automate the contextualization of a beamtime application: finding literature that is seemingly provided, without DOIs, can become an unnecessary complicated information retrieval exercise.

Another place for research paper PIDs in facility research lifecycle is when a facility asks visitor scientists to submit records of research papers supported by a facility experiment, and tries using this information for tracking or profiling exercises that can be useful e.g. as a proof of a facility or a certain beamline long-term impact.

## 5.2. Instrument (beamline) PID

An early approach to provide PIDs for instruments was the creation of the Journal of large-scale research facilities (JLSRF)[40]. This journal publishes articles describing large-scale scientific equipment. Facility users would be asked to cite the article describing

[40] Journal of large-scale research facilities (JLSRF). https://jlsrf.org/

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

an instrument in scientific papers that are based on data collected from this instrument. Based on these citations, it would then be possible to track the scientific output of the instrument. The JLSRF article's DOI would thus in a sense be used as a substitute for a PID of the instrument itself.

The direct attribution of PIDs to instruments has been explored by the Persistent Identification of Instruments working group (PIDINST) in the Research Data Alliance (RDA).[41] The group developed a metadata schema for the description of instruments.[42] The approach has exemplarily been implemented with two different PID infrastructure providers and PID types: ePIC Handles and DataCite DOIs.

It should be noted that the direct attribution of PIDs to instruments as proposed by PIDINST does not obsolete the JLSRF approach. Both options may complement each other. The descriptive article in JLSRF may provide better insight to a human reader on the properties and capabilities of the instrument, whereas the direct instrument PID makes it easier to add links to external resources and thus may be better suited for the automated aggregation of information by machines. If a direct PID for an instrument is issued, an additional JLSRF article's DOI should be considered what it strictly speaking is: a PID for an article describing the instrument, not a PID for the instrument itself. This article's PID should be linked from the instrument's PID metadata as a related identifier with relation type isDescribedBy. Still, both PIDs can be used in conjunction to track the scientific output of the instrument.

Instrument PIDs can be used for:

- tracking the scientific output of the instrument and indirectly of the facility,
- enriching the metadata of datasets created at the instrument by putting it into a context,
- providing additional information about the instrument, such technical specs and linking that to the dataset,
- giving credit to a particular beamline and, indirectly, to the instrument scientists in case they are included in the instrument PID metadata.

An example of a facility instruments metadata formed in accordance with the RDA PIDINST WG recommendations is shared in the Appendix A of this deliverable.

## 5.3. Facility PID

The potential need of having a dedicated PID type to designate facilities as a whole, such as a synchrotron radiation source or a neutron source, was extensively discussed by FREYA project.[43] The decision was not to introduce a dedicated PID type for facilities but use instead, depending on the information context, Organisation PID, Instrument PID or Funder PID. This reflects on the actual multi-facet mode of facilities operation, as they can be

[41] Stocker, M *et al*. 2020. Persistent Identification of Instruments. Data Science Journal, 19: 18, pp. 1–12. DOI: https://doi.org/10.5334/dsj-2020-018
[42] Krahl, R *et al*. 2021. Metadata Schema for the Persistent Identification of Instruments. Research Data Alliance. Submitted as a RDA recommendation. DOI: https://doi.org/10.15497/RDA00070
[43] FREYA deliverable D3.2 Requirements for Selected New PID Services. https://doi.org/10.5281/zenodo.3554196

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

considered Organisations, Instruments or Funders depending on the purposes and the context of research information modelling. For connections across all three PIDs, facilities can consider maintaining them via self-developed controlled vocabulary or ontology, which can allow cross-walking over organisational, instrumental and funding aspects when required, with all three PIDs included in the same vocabulary or ontology.

## 5.4. Organisation PID

Existing registries such as GRID.AC or ROR can source organisation PIDs. The organisation PIDs can be requested by a facility in a beamtime application (research proposal). They can be used to unambiguously link the organisation from related entities, such as the affiliation of the author in a publication or the facility as the owner of an instrument.

The metadata stored in the DOI record for a scientific paper or a data publication may include the affiliations with the organisation PID of its contributors. If this is used systematically, it may be used to track the scientific output of a research institution.

For the extensive research about Organisation PIDs and their use we can recommend a substantial report produced by FREYA project.[44]

## 5.5. Person PID

Facilities can request person PIDs like ORCIDs in the beamtime application (research proposal). ORCID can also be used as an external identity provider connected to the facility's SSO, so that users may be allowed to log in with their ORCID ID. As a positive side effect, the facility would get authentication (e.g. verification of the correctness) of the user's ORCID ID.

Person PIDs can be used for tracking down "returning customers", also tracking down facility research outcomes not necessarily reported by facility users themselves, i.e. those that can be discovered by looking into the researchers' publications records.

Person PIDs can also be used for getting credit for a facility use, e.g. a facility use can be an element of an ORCID record pushed there by the facility.[45]

## 5.6. Sample PID

The accurate description of the sample being measured certainly belongs to the most important parts in the metadata of any dataset collected during the measurement. In most cases, this requires tracking the history of the individual sample through all preparation steps to the measurement.

[44] FREYA deliverable D4.4 Organizational IDs in Practice.
https://doi.org/10.5281/zenodo.3666255
[45] Suggested ORCID workflow for research resources
https://info.orcid.org/documentation/workflows/research-resources-workflow/

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

A sample PID allows to keep track of its history when the sample moves across facilities. This may be particularly important, if the sample is held in a collection outside the context of the experiment or if it has been prepared by a third party. The sample PID record helps to give proper credit to scientists that contributed to the sample preparation, if they are not involved in the measurement. If the same sample is investigated more than once in a distinct project, the results of these investigations can be combined by relating the respective datasets via the sample PID.

If a sample is only created exclusively for a single experiment and discarded after the measurement or even destroyed by it, it is less obvious whether the attribution of a sample PID is useful. The sample description may then as well be embedded in the description of the experiment instead, with the additional unique sample ID generated that can be then added or linked to the proposal metadata, logbook, etc. This approach addresses the situation when the identifier minted for the sample itself is useless but the documentation of the experiment is shared for audit or reusability purposes. This ID would indicate that the sample and the related experiment were unique, although this would not be a persistent identifier to cite it in other contexts but only by facility internally.

Sample PID for its wider use in citations and other references in a wider research discourse beyond a particular facility can be assigned upon approval of a beamtime application (research proposal) or requested from a facility visitor scientist as a part of a beamtime application (research proposal). Facilities should be prepared to accept samples that already have a PID assigned elsewhere and should prefer using that existing PID rather than assigning a new one.

Sample PIDs can be considered “emergent” PID type. One potential implementation is IGSN. It originates from the geosciences and the original name was the International Geo Sample Number. In the meanwhile the scope of IGSN has been expanded to any physical specimen.[46] It is thus also applicable to samples in the PaN context. The STARS project in Digital LEAPS will explore options for sample PIDs and also consider IGSN.[47]

## 5.7. Data PID

Data PID is possibly the most common PID type in use by facilities.

Raw (collected) data PID can be assigned by a facility upon data acquisition or even before the actual data acquisition (as a data placeholder), which is the actual practice in some facilities (ISIS), whilst some other facilities (ESRF) issue raw (collected) data PID automatically, also upon a visitor scientist’s request. Raw data PIDs, when used as references, help with unambiguous data citation and give credit to a facility.

The actual semantic flavour of what is considered “raw data PID” greatly varies across facilities, and in many cases “raw data PID” in fact represents the facility investigation (series

[46] IGSN Documentation. https://igsn.github.io/
[47] Digital LEAPS: A European Strategy on the Digital Transformation of Accelerator-based Photon Sources towards a resilient and sustainable European Research Area. https://leaps-initiative.eu/wp-content/uploads/2021/10/DIGITAL-LEAPS-August-2021.pdf

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.*

of experiments) or the user proposal (application for beamtime) which becomes clear by looking into the “raw data PIDs” landing pages.[48] [49] This observation has some practical implications for research information search, e.g. what facilities call “data PIDs”, and mark up the associated PID metadata as designating “data” in a metadata profile offered by DataCite or other PID provider, can affect search results categorisation in faceted search (as in the very same DataCite): what is fed in “data” facet from facilities is not necessarily records of data, but rather records of facility investigations or user proposals.

PID can be also assigned to reduced or analysed data. It can be auto-generated by a facility upon an explicit visitor scientist’s request, or implicitly upon the use of a data analysis software offered by a facility as a service. Analysed data PIDs can be a part of added-value information services for facility users, such as generation of citations / references in a format required by a particular publisher; in case of using DOIs as PIDs, this can be automated through the DOI resolver content negotiation.

PIDs can be also a part of a quality metadata record for a curated data publication, e.g. as a supplement to a journal paper to be submitted. In the publication workflow, such a dataset is enriched with the bibliographic metadata to form a proper publication on its own, hence PID can serve as an unambiguous pointer to it. PIDs can help to promote data publications as a first class citizen scientific product in its own right, as well as used in automated data staging to computation through advanced PID resolution, making a data publication a machine-actionable knowledge artefact.

Either DOIs or Handles, e.g. supplied by EPIC, can be used as data DOIs, and some services like EUDAT B2SHARE[50] use both DOIs and Handles for the same dataset. DOIs issued by some providers popular with facilities such as DataCite may have advantage of rich metadata schema associated with a DOI that allows to express context in which data is collected; also content negotiation mechanism for DOIs that we discuss later in this deliverable has advantages for the automation of data records processing, including the ability to generate data citation references in thousands of various formats, as required by publishers.

## 5.8. Software PID

When scientific data has been generated by software, it is particularly important to precisely reference the program and version being used, preferably using a PID. Software PIDs should be created by its authors. A facility offering data analysis as a service should reference the PID when linking raw data and analysed data. Software PIDs can be used by facilities, too, in automated data processing pipelines (such as for data reduction), to indicate a piece of software as a specific agent in the analysed or reduced data record. The latter case can be supported by the actual practice of minting PIDs for data analysis and data reduction software, MANTID being one example.[51]

[48] V.Bunakov. Investigation as a member of research discourse. http://purl.org/net/epubs/work/12302226
[49] V.Bunakov. “What data identifiers actually refer to” section in Service for Data Retrieval via Persistent Identifiers. http://purl.org/net/epubs/work/21652744
[50] EUDAT B2SHARE service. https://b2share.eudat.eu/
[51] MANTID project. https://www.mantidproject.org/

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.*

Quite straightforward way of assigning PIDs to software could be using the Zenodo repository for this. There are two added-value features that make Zenodo a potentially good choice: its community features when records can be maintained as a collection, e.g. originating in a certain facility, also Zenodo connector for GitHub that can help with making any software shared in GitHub easily recorded by Zenodo and citable by a DOI.[52]

Software Heritage[53] promote the use of SoftWare Heritage persistent IDentifiers (SWHIDs) that are used on “intrinsic” properties of software assets such as hash codes generated that are independent of any assigned software versions or other “extrinsic” properties.[54]

The actual practices of assigning PIDs to software and of using such PIDs greatly vary. To give one example, the earlier mentioned MANTID software that is quite popular across facilities can be cited as a journal article (with a certain DOI assigned to it) or as a DataCite record (with a different DOI assigned to it); either of the methods is fine by the MANTID project. Moreover, MANTID versions receive their own specific DataCite DOIs, so in principle, researchers can use them for citations (rather than “master” DataCite DOI or the journal article DOI). If anyone is interested in counting citations of MANTID software, then citations of “master” journal article AND “master” DataCite DOI AND version-specific DataCite DOIs have to be aggregated, although a “deep stick” exercise for MANTID citation practices performed by FREYA project showed that researchers only rarely cite specific MANTID versions, hence a lion’s share of citations come through the use of “master” DOIs for the journal article or the DataCite “master” record for MANTID. This is an illustration that minting PIDs for software is fairly easy and popular services like DataCite are already being used for it, but good practice and community agreements are required in addition, to actually exploit the PIDs infrastructure for software.

One of more advanced uses of PIDs could be using them to capture dependencies of the software, but this requires additional effort, such as preservation of the computational environment then referring to the environment (perhaps, again with some PID associated). This is a complex topic that requires not only technology considerations but also significant effort for the common adoption of certain best practices that facilities cannot promote in isolation from a wider research ecosystem.

## 5.9. Data management plan PID

Data management plan PID can be assigned by a facility in case the facility proposal system offers a DMP template for a visitor scientist to fill in and then auto-generates a DMP. This will help a visitor scientist with an obligation to have a DMP and a clear reference to it, so can be considered an additional “convenience” service offered by a facility.The practice of offering a DMP template to facility users, or requesting them having a DMP is not there yet but if and when this practice is introduced, having PIDs for DMPs will be reasonable.

[52] GitHub: Referencing and citing content.
https://docs.github.com/en/repositories/archiving-a-github-repository/referencing-and-citing-content
[53] Software Heritage Archive. https://www.softwareheritage.org/
[54] SoftWare Heritage persistent IDentifiers (SWHIDs).
https://docs.softwareheritage.org/devel/swh-model/persistent-identifiers.html

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.*

PIDs for Data Management Plans can be DataCite DOIs[55]. The spread of this practice is unclear, and what seems more important at the moment is “peppering” the DMPs with as much PIDs for organisations, people or policies (having clear references to them in the DMP) rather than striving for having a PID for the DMP as a whole.

## 5.10. Policy PID

Some facilities have data management policies published on their websites[56] and as with every document, they could be assigned with PIDs such as a DOIs, should a facility deem it necessary. This will have an advantage of referring to an exact version of policy, as the policy may evolve.

The actual practical means for minting a policy DOI could be Zenodo, as an example, or FAIRsharing service that maintains the section for policies of all kinds.[57] Alternatively, facilities can use their existing agreements with DataCite or Crossref for minting Policy PIDs.

## 5.11. PIDs for standards and community recommendations

Of course, each standardisation body such as ISO or national organisations maintain their own system of references to standards - that may not amount to the status of universally resolvable PIDs though.

One specific service that is targeting the “market” for standards and recommendations PIDs is FAIRsharing who maintain the whole section devoted to them.[58] One particular use for such PIDs could be using them as clear pointers for machine agents (software) to designate data formats, so that the agent can decide on what visualisation or data analysis algorithm to apply.[59]

There is a potential for the use of PIDs in ontologies, vocabularies, taxonomies and other semantic assets developed by facilities, but this practice is not wide-spread yet and ontologies being developed by ExPaNDS rely mostly on PURLs[60] as the means of ontology elements identification.[61]

[55] A Brave New PID: DMP-IDs. https://blog.datacite.org/announcing-dmp-ids/
[56] E.g. ISIS Data Management Policy https://www.isis.stfc.ac.uk/Pages/Data-Policy.aspx
[57] FAIRsharing Policies section. https://fairsharing.org/search?fairsharingRegistry=Policy
[58] FAIRsharing Standards section. https://fairsharing.org/search?fairsharingRegistry=Standard
[59] E.g. see FAIRsharing record for JCAMP-DX format at https://doi.org/10.25504/FAIRsharing.v8nve2 that can in principle guide the software for spectral data visualisation or analysis.
[60] Persistent uniform resource locator. https://en.wikipedia.org/wiki/Persistent_uniform_resource_locator
[61] S. Collins *et al.* ExPaNDS ontologies v1.0. https://doi.org/10.5281/zenodo.4806025

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.*

## 5.12. Research activity PID (such as project PID or experiment PID)

The idea of a research activity PID has picked up in Australia where ANDS (Australian National Data Service) issues so-called RAiDs[62] to identify research projects. If a researcher moves across different organisations whilst the project continues, she can carry the RAiD along, thus maintaining the continuity of attribution of results to a particular project. This ANDS practice has been extensively communicated via Research Data Alliance and has a potential for a wider use.

Research activity PID can be requested by a facility as an element of a beamtime application (research proposal) - to indicate the facility user projects that are relevant to a proposed facility experiment. Research activity PID can be also assigned by a facility itself to indicate a facility experiment. To actually mint a research activity PID, DataCite or EPIC identifiers could be used, so the infrastructure for such PIDs is readily available, should facilities decide they need them.

## 5.13. Funding reference

“Funding references” may mean IDs for grants and IDs for funders. For the latter, there is a well-established Crossref Funder Registry that assigns DOIs to funders.[63] For grants, there is no universally accepted service that assigns PIDs, yet other kinds of grants references (that cannot be considered PIDs) are quite popular, hence they deserve mentioning in this deliverable.

Funding references are common in research papers and on the websites as acknowledgement of financial support, but are less common in the information artefacts (such as proposals) produced by facilities. There is a good potential for a wider adoption of funding references but the need of their conversion in a “true PIDs” with all associated metadata is an open question that should not be left only for facilities to act on: this is something for a wider research ecosystem, including publishers, to decide upon.

Facility users may receive external funding and may want to acknowledge that the scientific outputs they produce by doing research on a facility have been supported by a certain grant.

Possible applications of funding references are:

- Compliance with the requirement for the user to acknowledge external funding. Registering the funding references once at proposal submission stage may simplify that task.
- Giving credit to the facility by the acknowledgement of beamtime grants.
- Displaying the award of beamtime as a grant in a research paper or CV .

The latter application when beamtime may by itself be considered a grant-in-kind seems particularly important. There is a good potential for the promotion of such a practice, with its

[62] RAiDS: A persistent identifier for research projects. https://ardc.edu.au/services/identifier/raid/
[63] Crossref Funder Registry. https://www.crossref.org/services/funder-registry/

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.*

benefits for facilities discussed in more detail in the “PIDs guidelines for facilities” section of this deliverable.

There have been propositions of advanced use of funding references in facilities research lifecycle, such as for pushing references to beamtime awards in the facility users ORCID records[64] but there is no strong evidence that this practice proliferates across facilities. There may be natural reasons why this practice has not been adopted; in discussions around this deliverable, two likely reasons have been identified: ORCID proposed to collect facility users’ individual API keys, so that a facility can push a record of beamtime in the ORCID records, and this might seem too much to request from users; also some users might dislike this practice as not every facility experiment results in a research paper, hence a disbalance in ORCID records between too many beamtime awards compared to fewer papers might seem dubious from some users’ point of view.

Also the ORCID record structure makes having records of beamtime of a limited value for tracking down research lifecycle or research impact measurements, as ORCID record is designed as “a star” with everything attributed to a researcher and with no possibility to make cross-attribution, such as connect a research paper to a facility experiment (or a few): records of facility experiments and records of papers published remain unconnected in a researcher’s ORCID profile.

Yet the core idea of having beamtime awards clearly recorded as “grants-in-kind” (not necessarily using ORCID) deserves facilities’ attention, and we discuss the reasons for that in the “PIDs guidelines for facilities” section.

## 5.14. PID types conclusion

All PID types listed above may have value when used individually, but also have a potential for their connection to each other resulting in all sorts of “PID graphs” that can provide:

- richer information context for facility experiments,
- better (more structured and less ambiguous) data provenance,
- role-based credits to various participants of a facility research lifecycle,
- advanced reasoning over a facility research impact: not necessarily reduced to one or two aggregated metrics but based on connections between various elements of a facility research discourse and, potentially, on some graph-based algorithms.

All PID types have a potential for their use in metadata, ontologies and in all sorts of semantically explicit machine-interpretable statements e.g. statements about facility experiments, their contributing factors and their outcomes. The PID graph that gives a context to the facility research would allow advanced reasoning over facility impact; to give one example, if someone cites an instrument and the link between the instrument and its host facility exists in the PID graph, the facility involvement can be easily implied using

[64] Suggested ORCID workflow for research resources
https://info.orcid.org/documentation/workflows/research-resources-workflow/

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.*

machine reasoning, even if facility was not cited directly.

Figure 1 illustrates a PID graph that could contextualise facility research if there were enough persistent identifiers and enough connections among them.

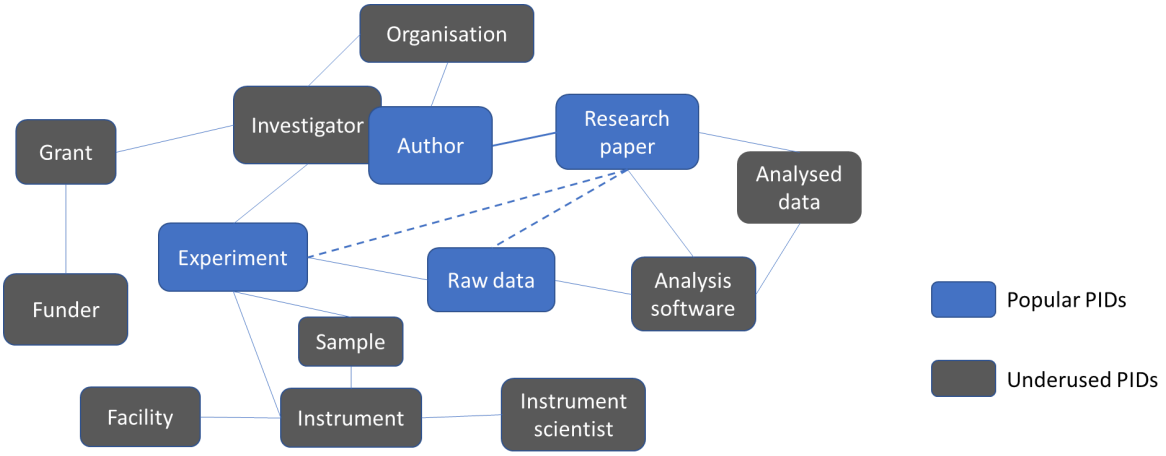



Figure 1. A principal structure of a PID graph for facility research contextualisation. Each node is an information entity that represents an element of a facility research discourse.

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

# 6. Alternative techniques and emerging applications of persistent identification

This section does some justice to alternative techniques of minting persistent identifiers that do not necessarily require a centralised external agency (authority) to produce them and to maintain all the resolving infrastructure. Additionally, this section introduces the topics of information structuring that rely on PIDs, and of information retrieval using PIDs in machine interfaces (APIs).

Therefore, this section could be of a less practical interest to facility users or management, but may be of interest to research information practitioners in facilities who are interested in "state of the art" elsewhere that is potentially applicable in a facilities context. The readers who find this section less relevant to their business-at-hand can skip it and proceed to the next "PIDs guidelines for facilities" section.

## 6.1. Hashed PIDs introduction and general aspects

The existing PID models and providers create PID representations suitable for issuing organisations. However, there are known issues of this centralised authority-based model that are enumerated below:

1. The resolving workflow implementation is proprietary-in-place as it is maintained by the public company, institution or research group. The resolvers are implemented on the PID provider side and this model creates the singular point of failure for the whole infrastructure. Also, the PID provider restricts the access by his own will and there is no mechanism to enforce actual PID resolution.
2. The PID itself is an atomic object to which the user has no access for provenance inquiries. Thus the absence of PID validation authority disallows the user to reproduce or reuse the results even in the case when the data is readily available.

To implement most important aspects of FAIR principles, the PID is required to be:

1. Self-describing. The PID resolution protocol should be as minimalistic as it could be, so that it can be implemented everywhere for interoperability. Then the PID could be resolved step-by-step from the string of symbols that represent a PID.
2. Recursive. It is necessary to have a PID that is able to also address the PID in the same or in a different format.
3. Reproducible. Any PID should have an algorithm to reproduce it from the given addressed data as a part of reusability of the metadata and also to avoid integrity control as a workflow stage.
4. Linkable. This point is especially important in a heterogeneous environment where centralised authority-based and hash-based PIDs coexist. The example of automatic server-side linking between IPFS CID and DOI for web application is presented in Appendix B.

Introduction of PIDs based on hash functions solves the described problems at least partially due to two-side reproducibility in the presence of actual data. In Elettra Sincrotrone Trieste,

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

the technical report[65] proposed the use of IPFS (InterPlanetary File System) as multi-node data storage engine. The hash-based PIDs are especially useful for identification of software because existing version control systems utilise hash functions to determine the attributes and versions of binary files, so that these functions' values can feed into the software metadata. Such generated metadata can be shared e.g. in the B2INST service[66]. The additional examples of hash-based PIDs and related technologies are presented in Appendix C.

## 6.2. Repository practices

Some existing repositories favour complex analytics by querying their underpinning graphs that connect various entities and elements of the research information landscape.

**OpenAIRE research graph** does this by enriching dataset metadata from other trusted sources (e.g. Crossref, Unpaywall, ORCID, Grid.ac) and scholarly sources by caching the dereferenced records and making the graph connections even deeper. OpenAIRE REST API allows dereferencing PIDs, i.e. obtaining further relevant data from PIDs, which aids the creation of custom views with richer metadata. This may be of interest to research information practitioners in facilities who can find facility-relevant records in OpenAIRE, and build richer connections around them.

The OpenAIRE collections available via API are:

- Publications metadata: **http://api.openaire.eu/search/publications**
- Research data metadata: **http://api.openaire.eu/search/datasets**
- Software metadata: **http://api.openaire.eu/search/software**
- Other research products metadata: **http://api.openaire.eu/search/other**
- Projects metadata: **http://api.openaire.eu/search/projects**

For more information about OpenAIRE API, there is documentation at https://graph.openaire.eu/develop/api.html

**GraphQL DataCite API** is another machine interface that information practitioners may want to utilize; it allows *custom views* of subsets of a PID graph. The client specifies the structure of the JSON to be retrieved, which is constrained by the structure of the DataCite PID graph.

[65] A.Vukolov. IPFS: Perspective for Application as the Dataset Sharing Infrastructure for Open Data projects. https://docs.google.com/document/d/1iRPrDyNQ8YOo8YylbTAB1uFG10NCNNhE-eLA86p40Ao/edit#heading=h.jjijmblkf9dx

[66] B2INST Database Documentation. EUDAT/EOSCHub/SurfSARA https://b2inst-poc2.eoschub-surfsara.surf-hosted.nl/help/docs

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

The resources that can be queried by PID are:

- Members
- Repositories
- Prefixes
- DOIs
- Researchers (using the ORCID API)
- Funders (using the Crossref Funder ID API)
- Organisations (using the ROR API)

The DataCite GraphQL API Guide is available at https://support.datacite.org/docs/datacite-graphql-api-guide

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

# 7. PIDs guidelines for facilities

This deliverable can serve as guidance for the research information practitioner on the use of persistent identifiers in a facilities research context, who can provide persistent identifiers, and where to look for more information and community advice.

This section summarises the discussions that T2.4 have had across and beyond ExPaNDS, and proposes recommendations of priorities for photon and neutron facilities to consider for the implementation of modern research information infrastructure with the appropriate use of persistent identifiers. We refer to other (earlier) parts of this deliverable when required, and do not repeat the arguments made there.

There are five priorities that are related, so a responsible research information practitioner should consider them as a whole. Yet the practical implementation can start from any of them individually, as a focal point from which other considerations and activities can spin out.

## 7.1. Minting PIDs for Raw Data

Minting PIDs for raw data acquired in facility experiments is well-established within a number of RIs, but there is a potential for it to be taken up in more RIs. This is reflected in the Data Policy recommendations quoted in Section 4.

Further, there are opportunities for sharing good practice within those RIs that already mint PIDs for raw data. The “Data PID” subsection indicates a variety of practices in respect to minting such PIDs, including the observation that what is considered “a data PID” can in fact be a representation of a facility investigation (a series of experiments) or of the beamtime application (proposal) in cases when “a data PID” is minted upon the proposal approval, even before the actual experiments.

The PID record for data should include rich metadata where possible. In particular, related objects or entities should be properly referenced using their respective PID. This includes the Person PID of the creators of the data,[67] the Organisation PID for those persons’ affiliations, the Sample PID of the sample being measured, and the Instrument PID of the instrument being used to collect the data. This presents a records curation challenge not only for facilities who consider minting new data PIDs with the inclusion of other PIDs in data PIDs metadata, but also for facilities with the established practice of minting data PIDs, as they may consider the value of updating legacy records of data PIDs, with the inclusion of Organisation, Person or Instrument PIDs in the existing data records.

The use of PIDs for data can help facilities with tracing the outcomes of facility experiments and with measuring facilities’ research impact; it is up to facilities rather than PID providers to implement and support such use of PIDs. There is no guarantee though that data PIDs can decisively help with impact studies, as publishers are de-facto a part of scholarly communication and set their own norms for citations, so even if such citations are accurately made by authors, how publishers handle the formal citation of facility data may vary between

[67] This can be facility users (visitor scientists) or / and instrument scientists.

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.*

journals. We would recommend that RIs track and participate in efforts in the community, including with publishers, to establish best practices in the use of PIDs within citation; the RDA or Crossref can be proper forums for such discussions. For example, there is an opportunity for the promotion of formal records of facility access awards as grants-in-kind; we discuss this opportunity later in this section.

## 7.2. Use of PIDs in data catalogues

Persistent identifiers can play a prominent role in data catalogues, and contribute to raising the metadata quality. FAIR data principles specify the use of persistent globally unique identifiers that point at data assets and at metadata for them, available via publicly available data catalogues. In practical terms, the RDA FAIR Data Maturity Model Working Group[68] outcomes are worth paying attention to, such as FAIR Data Maturity Model Specification and Guidelines.[69]

We recommend that records in data catalogues should have PIDs assigned to important metadata elements, such as personal identifiers for researchers, organisation identifiers for affiliations, instrument identifiers for experimental equipment where data was collected, or software identifiers for code that was used for data analysis or data generation (simulations).

Identifiers for publications associated with the experiment should also be recorded in facility data catalogues. Facilities' practice for the collection of bibliographic records should where possible include the collection of the publications' DOIs; we discuss this matter below in the "Proliferation of PIDs in facilities proposal systems" subsection.

DOIs may have advantages over other PID schemes for use within data catalogues as the DOI system has a body of tools and services associated with it to provide added value. For example, some DOI providers offer the mechanism of content negotiation, so that not only humans but machine agents (software) can be consumers of DOIs. One popular application of content negotiation is using it for the generation of references in a bibliographic format required[70], yet applications of content negotiation stretch beyond the references formation, e.g. this mechanism can be principally used for the automated data retrieval.[71]

The population of data catalogues records with PIDs for resource types beyond PIDs for data allows valuable insights in facilities research scope and research impact. Some PID providers may have added value for facilities and offer advanced machine interfaces that exploit the information retrieval benefit of connecting various entities via their PIDs ("the PID graph").[72]

[68] RDA FAIR Data Maturity Model Working Group. https://www.rd-alliance.org/groups/fair-data-maturity-model-wg
[69] FAIR Data Maturity Model. Specification and Guidelines. https://doi.org/10.15497/rda00050
[70] DOI Citation Formatter. https://citation.crosscite.org
[71] V. Bunakov. Service for data retrieval via persistent identifiers. In proceedings of: 4th International Conference on Data Management Technologies and Applications (DATA 2015), Colmar, France, 20-22 Jul 2015. http://purl.org/net/epubs/work/21652744
[72] DataCite GraphQL API Guide. https://support.datacite.org/docs/datacite-graphql-api-guide

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

## 7.3. Adoption of emerging PID types

We recommend that RIs develop strategies to adopt Organisation PIDs, Instrument PIDs and Person PIDs. These three PID resource types deserve special attention by facilities as the provision of them is sufficiently mature, and what is required is not building a new infrastructure for such PIDs, but adoption of existing best practices and standards for these PIDs. The respective subsections of this deliverable describe appropriate best practices of these PID types; some other emerging PID types such as for Samples should be on facilities horizon, but the infrastructure and community recommendations for Sample PIDs may require more maturity.

Instrument PIDs in particular can be considered “low-hanging fruits” for adoption, with more than one method of provisioning them, as indicated in the “Instrument PIDs” subsection of this deliverable. Appendix A contains an example of one particular implementation using DataCite metadata. The main purpose of instrument PID and PID-associated metadata is to indicate the instrument existence and make it possible to refer to the instrument in whatever information context is required, rather than giving full context of how the instrument is used; such a context can be given by using the instrument PIDs in references from publications, software and other elements of a facility research discourse.

## 7.4. Proliferation of PIDs in facilities proposal systems

An application for facilities access (proposal) is the first step in a facility research lifecycle, and the information that proposal systems collect has a potential to propagate down the facilities research lifecycle, especially taking into account that records for what facilities tend to call “data DOIs” are in fact proposal representations in cases when they are based on information excerpts from proposals.

We recommend that research paper PIDs are recorded for previous relevant publications listed in the user application for beamtime. This would allow easier identification and retrieval of relevant research, as well as measuring citations for it in cases when this practice is a part of the proposal evaluation.

Publication DOIs supplied in the proposals would allow making clear connections across multiple proposals, and would allow finding commonalities across them, even if experiments have been requested by different researchers from different organisations. Such an analysis through the commonality of references becomes an insurmountable effort-consuming task without bibliographic references assigned with DOIs.

Apart from DOIs for bibliographic references, of course proposal systems should collect other types of PIDs from researchers who apply for beamtime, such as Person / Researcher PIDs or Organisation PIDs for affiliations, which would again allow good contextualisation of research information and would support research impact studies that are in the facilities own long-term interest, e.g. when they have to justify next rounds of funding for upgrades. Good population of persistent identifiers in proposal systems, then propagating them down the facility research lifecycle would greatly help impact studies.

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.*

## 7.5. Promotion of facility beamtime awards as grants-in-kind

The motivation for having persistent identifiers for data was to an extent driven by hope that the practice of data citation would make research more reproducible and would give due credit to facilities staff beyond the authors of research papers resulting from experiments. Despite some progress, the practice of data citation is not as universal as it could be, also quite diverse, e.g. data can be cited in the acknowledgements section of a research paper, or in the list of references, or in the main body of the paper. This makes matching data to publications a not straightforward exercise that involves, as an example, the development of specialist software such as PUMA by ILL and ESRF.[73]

Moreover, even if well-formed references to data, with all clear attribution and PIDs, are made by the authors, thus fully fulfilling the facility requirement of having its support clearly acknowledged, then how publishers handle these records of science is beyond the authors' or the facilities' control. Some publishers distinguish between data citations and "regular" citations of research papers, and exclude data citations from their propagation in reference services such as Crossref. Or publishers may not share lists of references at all, hence the authors' doing a good job of formally citing data collected in a facility is not a guarantee that this valuable piece of a research discourse will be properly shared with the world. If something is not indexed by popular reference services, it becomes problematic to find even if (strictly speaking) it is present in the text of the paper.

On the other hand, there is a common universally adopted practice of referring to grants in the acknowledgements section of research papers, also in the records of software and on the websites. This practice is well understood and well adopted by researchers and by publishers, and it makes attribution of research-to-funders harvestable with a good level of automation, as the software agents such as the aforementioned PUMA would know where to look for such references: exactly in the acknowledgements section. Facilities will have to incentivise users so that they actually start using grants-in-kind references to facility beamtime, but this practice will be very similar to that widely adopted for monetary grants. From the policy making point of view, facilities' message to users can be very clear and strong: facility time costs money (which is public money in most cases), so it is fair to require that the users consistently admit receiving this funding-in-kind, with explicit well-formed references.

All facilities have their internal systems of research proposals (beamtime applications) identification, but these identifiers are not universal, with no global means of their resolution; they are not persistent identifiers. It is understandable that beamline applications cannot be shared immediately and in full, as they often contain information that may affect research priority, or is commercially sensitive. Having this said, some facilities such as ISIS neutron and muon source do use (selected) parts of the proposal description to mint a DOI even before the experiment is conducted – just upon the proposal approval, so there is a clear potential for the universal identification of beamtime awards.

[73] PUMA publications matching software.
https://streamline.esrf.fr/non-classe/puma-publications-matching-software/

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

What is missing is not the availability of suitable information infrastructure, but the appreciation by facilities of the importance of a beamtime award as a grant-in-kind. This is further blurred by a situation when facility call it “data PID” what is de-facto an excerpt from a proposal (with data that can be added or connected to the record later, when the experiment is actually conducted – which does not make it a record of data more than the record of a beamtime award).

Our recommendation is for facilities to consciously accept that a DOI (or other PID) and associated metadata that they produce upon the acceptance of a proposal is in fact a record of a beamtime award. Facilities can start treating such records – and promote this understanding across facility users and publishers – as records of grants-in-kind that should be handled in research papers and in other research artefacts (software descriptions, data publications, websites) in exactly the same way as monetary grants.

This practice, in our opinion, has a higher potential for its wide adoption than the decade-long attempts to promote formal citations of facilities data as “true citizens of research discourse”. Even from a policy making perspective, the message can be unequivocally clear: facility experiments are costly and supported by public money, so researchers should recognise receiving support by facilities in all their published research results.

The record of a facility instrument (potentially with clear indication of involvement of instrument scientists, too) can be naturally connected to the records of such grants-in-kind, or can be a part of the grant-in-kind metadata. References to relevant publications that are commonly requested by User Offices can be a part of the grant-in-kind metadata, too. The data, when actually collected in the facility experiment, can be attached to or referred from the grant-in-kind record as well. In short, it is up to a facility and to its perceived publicity needs, or impact study needs, what the facility would like to include in the grant-in-kind record, and share it as a landing page for the respective beamtime award PID.

The citation of a facility grant-in-kind record can be made as easy as possible for researchers, e.g. if it is a DOI from a renowned provider like DataCite, the formation of a clear citation is an out-of-the-box part of a service and does not present any technical problem whatsoever. Yet these meticulously formed references may not even be required, as using a DOI (or other PID) for a facility grant-in-kind in the acknowledgement section of a research paper, or respective sections of software descriptions, or data descriptions or websites, will be enough for the proper proliferation of such references, and for their inclusion in the automated pipelines for research information processing.

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.*

# References

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18

McMurry JA, Juty N, Blomberg N, Burdett T, Conlin T, *et al.* Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLOS Biology 15(6): e2001414 (2017). https://doi.org/10.1371/journal.pbio.2001414

Matthews, Brian, McBirnie, Abigail, Vukolov, Andrei, Ashton, Alun, Collins, Stephen, Da Graca Ramos, Sylvie, Gagey, Brigitte, Gonzalez-Beltran, Alejandra, Johnsson, Maria, Krahl, Rolf, Ounsy, Majid, & Van Daalen, Mirjam. (2020). Draft extended data policy framework for Photon and Neutron RIs. Zenodo. https://doi.org/10.5281/zenodo.4014810

McBirnie, Abigail, Matthews, Brian, Gagey, Brigitte, Minotti, Carlo, Salvat, Daniel, Schlünzen, Frank, & Vukolov, Andrei. (2021). Final data policy framework for Photon and Neutron RIs (1.0). Zenodo. https://doi.org/10.5281/zenodo.5205824

Salvat, Daniel, Gonzalez-Beltran, Alejandra, Görzig, Heike, Matthews, Brian, McBirnie, Abigail, Ounsy, Majid, Da Graca Ramos, Sylvie, & Vukolov, Andrei. (2020). Draft recommendations for FAIR Photon and Neutron Data Management. Zenodo. https://doi.org/10.5281/zenodo.4312824

Görzig, Heike, Matthews, Brian, McBirnie, Abigail, & Soler, Nicolas. (2021). DMPs for Photon and Neutron RIs (1.0). Zenodo. https://doi.org/10.5281/zenodo.5636095

Cousijn, H. *et al*. Connected Research: The Potential of the PID Graph. Patterns, 2(1),100180 (2021). https://doi.org/10.1016/j.patter.2020.100180

Isabelle Boscaro-Clarke, Kat Roarty, Rebecca Stredwick, Brian Matthews, Lauro Fava, Rebecca Grant, Vasily Bunakov, Mark Thorley, Robert Huber, Oliver Knodel, Renaud Duyme, & Abigail McBirnie. (2022). ExPaNDS Librarian Symposium Report and presentations (FINAL). Zenodo. https://doi.org/10.5281/zenodo.5973068

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://doi.org/10.2777/926037

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://doi.org/10.2777/525581

Gotz, Andy, Perrin, Jean-Francois, Fangohr, Hans, Salvat, Daniel, Gliksohn, Florian, Markvardsen, Anders, McBirnie, Abigail, Gonzalez-Beltran, Alejandra, Taylor, Jonathan, & Matthews, Brian. (2020). PaNOSC FAIR Research Data Policy framework (1.1). Zenodo. https://doi.org/10.5281/zenodo.3826039

Stocker, M., Darroch, L., Krahl, R., Habermann, T., Devaraju, A., Schwardmann, U., D'Onofrio, C. and Häggström, I., 2020. Persistent Identification of Instruments. Data Science Journal, 19(1), p.18. DOI: http://doi.org/10.5334/dsj-2020-018

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

Krahl, R., Darroch, L., Huber, R., Devaraju, A., Klump, J., Habermann, T., Stocker, M., & The Research Data Alliance Persistent Identification of Instruments Working Group members. (2021). Metadata Schema for the Persistent Identification of Instruments. Research Data Alliance. https://doi.org/10.15497/RDA00070

Christine Ferguson, Johanna McEntyre, Ginny Hendricks, Tina Dohna, Ketil Koop-Jakobsen, Frances Madden, Sunje Dallmeier-Tiessen, Stephanie van de Sandt, Artemis Lavasa, Simon Lambert, Vasily Bunakov, Robin Dasler, & Martin Fenner. (2019). Deliverable D3.2 Requirements for Selected New PID Services. Zenodo. https://doi.org/10.5281/zenodo.2649229

Christine Ferguson, Simon Lambert, Manuel Bernal Llinares, Frances Madden, Robin Dasler, Martin Fenner, Artemis Lavasa, Chris Baars, Tina Dohna, Ketil Koop-Jacobsen, & Dan Morgan. (2020). Deliverable 4.4 Organizational IDs in Practice. Zenodo. https://doi.org/10.5281/zenodo.3606059

Collins, Steve P., da Graça Ramos, Silvia, Iyayi, Daniel, Görzig, Heike, González Beltrán, Alejandra, Ashton, Alun, Egli, Stefan, & Minotti, Carlo. (2021). ExPaNDS ontologies v1.0. Zenodo. https://doi.org/10.5281/zenodo.4806025

FAIR Data Maturity Model Working Group. (2020). FAIR Data Maturity Model. Specification and Guidelines (1.0). https://doi.org/10.15497/rda00050

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

# Appendix A: Example of metadata record for a facility instrument

The record follows the recommendations of RDA PIDINST WG that are mapped on the DataCite metadata schema. This mapping may evolve along with changes in new versions of DataCite metadata schema, e.g. “`resourceTypeGeneral`” attribute value is "`Other`" at the moment, but DataCite consider the introduction of an “Instrument” value for it, in a controlled vocabulary that can be adopted in the future versions of schema.

```
<?xml version="1.0" encoding="UTF-8"?>
<resource xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://datacite.org/schema/kernel-4"
xsi:schemaLocation="http://datacite.org/schema/kernel-4
http://schema.datacite.org/meta/kernel-4.3/metadata.xsd">
  <identifier identifierType="DOI">10.5442/NI000001</identifier>
  <creators>
        <creator>
        <creatorName nameType="Organizational">Helmholtz-Zentrum Berlin für Materialien
und Energie</creatorName>
        <nameIdentifier nameIdentifierScheme="ROR"
schemeURI="https://ror.org/">02aj13c28</nameIdentifier>
        </creator>
  </creators>
  <titles>
        <title titleType="Other">E2 - Flat-Cone Diffractometer</title>
  </titles>
  <publisher>Helmholtz-Zentrum Berlin für Materialien und Energie</publisher>
  <publicationYear>2019</publicationYear>
  <subjects>
        <subject>BER II</subject>
        <subject>Diffraction</subject>
        <subject>Neutron source</subject>
  </subjects>
  <contributors>
        <contributor contributorType="HostingInstitution">
        <contributorName nameType="Organizational">Helmholtz-Zentrum Berlin für
Materialien und Energie</contributorName>
        <nameIdentifier nameIdentifierScheme="ROR"
schemeURI="https://ror.org/">02aj13c28</nameIdentifier>
        </contributor>
  </contributors>
  <dates>
        <date dateType="Available">2007-01-01/2019-12-31</date>
  </dates>
  <resourceType resourceTypeGeneral="Other">Instrument</resourceType>
  <relatedIdentifiers>
        <relatedIdentifier relatedIdentifierType="DOI"
relationType="IsDescribedBy">10.17815/jlsrf-4-110</relatedIdentifier>
  </relatedIdentifiers>
  <descriptions>
        <description xml:lang="en" descriptionType="Abstract">
```

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.*

A 3-dimensional part of the reciprocal space can be scanned in less then five steps by combining the "off-plane Bragg-scattering" and the flat-cone layer concept while using a new computer-controlled tilting axis of the detector bank. Parasitic scattering from cryostat or furnace walls is reduced by an oscillating "radial" collimator. The datasets and all connected information is stored in one independent NeXus file format for each measurement and can be easily archived. The software package TVneXus deals with the raw data sets, the transformed physical spaces and the usual data analysis tools (e.g. MatLab). TVneXus can convert to various data sets e.g. into powder diffractograms, linear detector projections, rotation crystal pictures or the 2D/3D reciprocal space.

</description>
</descriptions>
</resource>

# Appendix B: Example of linking hash-based to centralised PIDs

What is following is a simple “model” (illustrative) case for web-based connection between IPFS CID and DOI IDs that point to the same data - the Zenodo record referenced as a webpage.

Here we have URI:

https://ipfs.io/ipfs/QmPnVQTQo4nVkCnZFHxnAWWykCsTc2N4Nst35K8Ls5BVBM

This URI is resolved as the prefix to CID:

`QmPnVQTQo4nVkCnZFHxnAWWykCsTc2N4Nst35K8Ls5BVBM`

The resolver gateway returns a plaintext containing explicitly defined HTML redirect (here it is made for simplification):

```
<html><head><meta http-equiv="refresh" content="0;
url=https://doi.org/10.5281/zenodo.4980522"
/></head><body></body></html>
```

The HTML redirect wraps the Zenodo record `4980522`, pointing to the presentation dedicated to the FAIRness in context of PIDs. Here we have the following resolve chain:

`IPFS URI > WEB URI > CID > HTML PARSER > DOI.ORG URI > DOI RESOLVER > DOI.ORG DATABASE > DOI.ORG NAMETABLE > ZENODO URI > ZENODO NAMETABLE > ZENODO WEBPAGE > ZENODO RESOLVER > DIGITAL OBJECT`

In fact, this forces the IPFS CID issued for the HTML redirect to be the explicit uniform redirector for DOI `10.5281/zenodo.4980522`. As the CIDs are hash-based, they are definitely content-based and as the HTML wrapper string is uniform, the CID could be considered in this case the intrinsic but publicly exposed PID that is as persistent as DOI and HTML are. In addition, the CID in such a case could be cited equally with the connected DOI.

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.*

# Appendix C: PIDs in a distributed context

Distributed computation and distributed data management environments have raised their profile in recent years, so it is important for research information practitioners to appreciate these developments, and to see how they can be applied to managing persistent identifiers.

## Distributed Hash Table (DHT) based

An example of *hash based PIDs* commonly used for distributing large datasets like official Linux software distributions is the **BitTorrent Infohash**. Besides being an alternative, the main reason is understandably to avoid download bandwidth costs, so it is popular for free and "libre" software. A *magnet URI scheme* or just *magnet link*[74] [75] provides minimal data in a convenient way, and a .torrent metadata file is not necessary to download from the mainline BitTorrent DHT.

Now that the clients supporting the BitTorrent protocol are also the web browsers themselves (inc. via JavaScript library addons) this became a cost-effective way of distributing public datasets ( with no central point of failure). The good example here is Linux distribution dissemination model[76].

Integrity and Security measures:

1. This fixed length string is a **MD5** or **SHA-1** hash (digest) identifying the content, sufficient to guarantee both its integrity and uniqueness.
2. To ensure *no-tampering* (by rogue party) a **SHA-512** digest is required and therefore sometimes also provided.
3. Origin (**provenance**) is certified by the issuer using a (PKI) **cryptographic signature** (e.g. GPG .sig).

This also works in a variety of peer-to-peer[77] (p2p, decentralised and distributed[78]) data transfer scenarios. The (info)hash also identifies the data for sharing in a Distributed Hash Table (**DHT**) based solutions like **Mainline DHT** (BitTorrent[79]) or **IPFS**[80], or furthermore, even

[74] G. Mohr. MAGNET v0.1. Draft Technical Specs
http://magnet-uri.sourceforge.net/magnet-draft-overview.txt
[75] Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 8, pages 1031–1040 (2016)
http://dx.doi.org/10.15439/2016F87
[76] Debian Linux. Torrent downloads.
https://www.debian.org/CD/torrent-cd/
[77] E. K. Lua et al. A Survey and Comparison of Peer-to-Peer Overlay Network Schemes. IEEE Communication Syrvey and Tutorial,, March 2004
https://www.cl.cam.ac.uk/teaching/2005/AdvSysTop/survey.pdf
[78] Distributed Persistent Identifiers System Design
https://datascience.codata.org/articles/10.5334/dsj-2017-034/print/
[79]B. Cohen. The BitTorrent Protocol Specification
http://www.bittorrent.org/beps/bep_0003.html
[80] IPFS DAG/CID Protocols
https://docs.ipfs.io/concepts/how-ipfs-works/#directed-acyclic-graphs-dags

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

databases like **Apache Cassandra**[81] or search engines like **YaCy**[82]. These one-way cryptographic hash functions also solve the problem of having data in different nodes or instances (sharding), e.g. in a computing cluster. Thus, the global distributed solution is a generalised case of a similar problem.

Hash digests are added as part of Magnet URIs as the means of identification in the target DHT. The hashed PIDs could be easily linked with centralised solutions but transposal uniqueness is not guaranteed.

## Blockchain enabled PIDs

A blockchain or distributed ledger is a public decentralised and distributed database. The records form (cryptographic) trusted audit trails. The ultimate source of truth is established by a network consensus mechanism. Therefore, a *trusted third party* is not required to verify transactions *for the first time in history (*That is, so far. An issue has not been discovered yet*)*.

Any record entry can be independently verified and therefore has many uses like identity management, record keeping ( e.g. deeds), or any *digital notarisation* initiatives.
It is possible that the future of scientific (trusted) publishing will end up in public blockchains controlled by relevant stakeholders under some sensible governance model.

This technology underpins initiatives like Know Traveller Digital Identity (KTDI)[83], akin to a private passport for international travel. There are provisions for nuanced access to private data to account for legislation like the GDPR EU directive.
There are enormous possibilities to certify the provenance and lineage of scientific data using this technology in an IoT context. A token (related to a blockchain[84]), usually originating from a hash, will act as the PID or as part of a PID following a standardised scheme.

Blockchain-based distributed file systems are likely to emerge as global standards[85].

## Hash-based identifiers applications

For the use as content-based PIDs for applied software uses like identification and integrity.

[81] Apache Cassandra
https://cassandra.apache.org/_/index.html
[82] YaCy - P2P search engine software
https://yacy.net/faq/
[83] KTDI https://ktdi.org/
[84] Dunphy, P. and Petitcolas, F.A., 2018. A first look at identity management schemes on the blockchain. IEEE security & privacy, 16(4), pp.20-29
https://arxiv.org/abs/1801.03294
[85] H. Huang et al. When Blockchain Meets Distributed File Systems: An Overview, Challenges, and Open Issues. IEEE Open Access, March 2020
https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9031420

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

Highly notorious hash digest identifier for binary objects, particularly software:

- **Git** uses SHA-1 to identify binary blobs (and trees). On the filesystem level, it also can refer to IPFS CID[86]
- **Docker** uses SHA256 to identify internal entities (images, containers etc)[87].
- **Kafka**, **MINIO** and **S3** use MD5 for successful transfer notification (and as their etag).
- **HTTP ETag**[88] declared to be a *content-based* resource identifier.
- **IPFS DNSlink**[89] uses hash keys to associate directly with the data (CID) or to persistently associate the human-readable names with public keys specific for web location (analogue to WWW domains).

# Appendix D: *De Jure* standards for global PIDs

## W3C DIDs

Another distributed technology standard (with its own PID URI scheme) is **W3C Decentralized Identifiers (DIDs) v1.0**. with provisions for other global standards like e.g. **OpenID**.

Like in DIDs, lists of all known and/or equivalent PIDs are strongly promoted by various W3C standards. E.g.as used in *Google Datasets* metadata:

- **schema.org** *sameAs,*
- Or for **W3C DCAT profiles** *dct:identifier* list, (preferably IRIs).

## SPDX

Hash checksums are part of **W3C DCAT-AP** standard using **SPDX** vocabulary for which version 2.2.1 became an ISO standard[90] in 2021.

## OpenID

OpenID is an open initiative that promotes a standard allowing identification of the user and his actions alongside authorized URIs and linked APIs[91]. It aggregates also cryptographic-related standards (behind the scenes) to allow the use of authentication API URL as a cryptographically signed PID, including self-signing level of openness[92]. The main

[86] IPFS. Hosting Git-style repo. End-user howto. URL: https://docs.ipfs.io/how-to/host-git-style-repo/
[87] Docker. Official object reference. https://docs.docker.com/engine/reference/commandline/images/#list-the-full-length-image-ids
[88] Fielding, et al. RFC 2616. Hypertext Transfer Protocol -- HTTP/1.1. Chapter 14. Header Field Definitions. Section 14.19. ETag https://www.w3.org/Protocols/rfc2616/rfc2616-sec14.html#sec14.19
[89] DNSlink Standard. Official website. https://dnslink.io/
[90] ISO/IEC 5962:2021 - SPDX® Specification V2.2.1 https://www.iso.org/standard/81870.html
[91] OpenID specification v2.0. URL: https://openid.net/specs/openid-authentication-2_0.html
[92]K. Yasuda et al. Self-Issued OpenID Provider v2 https://openid.net/specs/openid-connect-self-issued-v2-1_0.html

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

advantages of OpenID standard are: its last generation API is built over OAuth 2.0 protocol which allowing the user to link many service providers to a single account[93], and the declaration of support exists in WWW Consortium FOAF Definition Vocabulary[94]. But the OpenID standard is declared as centralized on a provider level that can lead to vulnerabilities in the cases when the user changes OpenID identification provider[95], so the persistence of API identifier provided by this standard is still discussible.

# Appendix E: Semantic and graph technologies

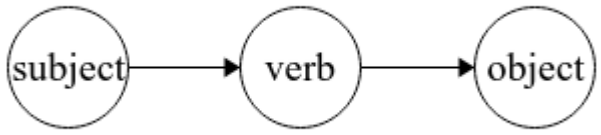
The topic of PID graphs and, more universally, knowledge graphs[96] is quite popular across the research information management discipline, so we reiterate here the key facts about popular graph-based information models and data structuring techniques, and give an example of one particular implementation of the PID graph.

## Semantic linked data: RDF model

The RDF model mimics Natural Language word order syntactic constructs (largely for indo-european languages, albeit word order may vary).

A collection of statements where in each statement:

- There is a subject or "**who** is doing the action".
- There is a **verb** (or **predicate**[97]) of "the action done".
- There is an **object** or "the target of the action"**.**




This gives us 3 (linked) elements (named a **triple**) which in RDF are in S-P-O (from S-V-O) sequential order, which only matters for vertex(node) identification purposes, otherwise any order is possible.

[93] OpenID Connect Core 1.0 incorporating errata set 1. https://openid.net/specs/openid-connect-core-1_0.html
[94] D. Brickley, L. Miller. FOAF Vocabulary Specification 0.99 http://xmlns.com/foaf/spec/#term_openid
[95] B. van Delft, M. Oostdijk. A Security Analysis of OpenID. In: de Leeuw E., Fischer-Hübner S., Fritsch L. (eds) Policies and Research in Identity Management. IDMAN 2010. IFIP Advances in Information and Communication Technology, vol 343. Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-642-17303-5_6
[96] M. Kejriwal et al. Knowledge Graphs: Fundamentals, Techniques, and Applications ISBN 978-0262045094
[97] Predicate is the sentence including the verb and not only the verb(linguistics). However, in RDF terminology due to the incidental need to include of other terms besides de verb the use of predicate is common

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*

The RDF identifies these 3 elements using IRIs (the internationalised version of URIs).
Using a collection of triples we can build a larger graph (or network) by aggregation of subgraphs. We will identify the vertices and edges using IRIs, which in the best case scenario are PIDs.

## (Labelled) Property Graph model (LPG)

Briefly, in the property graph model, both *vertices* and *edges* may have attached structured data in **key-value** pairs (aka *properties*).

The associative array of properties (also called dictionary) makes a great difference if chosen strategically for the right data. We are abandoning costly graph traversals for **O(n),** in a worst case *lookup*[98] , of a hash-based dictionary implementation (which is common).

This model appears to be more versatile and storing RDF data is possible, where the other way around is more convoluted, sometimes impossible, requiring new standards like RDF*.

This has been the default model for all sorts of graph implementations outside Web Technologies.

A PID should be *dereferenceable* and thus able to provide further metadata (to potentially grow the graph, even conceptually). Serialisation standards are largely Web centric. In querying terms LPG is concise and fast (but not too standardised yet). By software adoption (inc. databases) LPG has been more prominent.

The model of choice should be based one the nature of data to be stored and the practical use cases. In any case, PIDs are essential to build a “stable graph” to *persist* through the passage of time.

## Data structuring with graphs

Besides improving general metadata/data FAIRness, we should further structure the data using global standards, for numerous benefits including better API interfacing and storage[99] [100]. It is also a precondition to introduce AI techniques such as reasoning, enabled by semantic linked data in domains such as biomedical research.

Semantic linked data practices encourage the use of identifiers (ideally PIDs) in the RDF representations. Graphs implicitly allow for perfect partitioning (and parallel processing). This means that database sharding is not a problem for this metadata by design. In practice, different elements could be stored in different databases which gives maximum flexibility. Furthermore, the richness of the data view retrieved is only limited by the reach (depth) of the underlying PIDs-rich graph.

[98] Same for insertions and deletions.
[99] Guidelines for publishing structured metadata on the web v3.1: https://doi.org/10.15497/RDA00066
[100] Advanced SEO: https://developers.google.com/search/docs/advanced/structured-data/dataset

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.*

The semantic linked data relies to a good extent on various vocabularies and ontologies, and we are going to point out some of them that are most popular, so that information practitioners in facilities can consider their adoption.

As the result of international collaboration, DCAT was adopted as the vocabulary for datasets in governmental public institutions. DCAT-AP (pan-european) and DCAT-US (USA) are the resulting profiles based on the DCAT model.

There are many serialisations of RDF including the newer and developer-friendly JSON-LD. Contemporary REST APIs offer JSON serialisations, at least as an option. Nascent GraphQL APIs are JSON-native by design.

For managing semantic linked metadata, RDF and graph databases may be the best matches but there are multi-model databases to consider, too. This is to say that technology for semantic linked data is readily available and is not a limitation; what limits the adoption of semantic linked data is mostly best practices of metadata creation, such as infusion of as many PIDs in metadata definition as possible.

The main thread connecting these datasets is the practice of creating dataset metadata as separate documents representing the nodes of the graph. That results in a dataset series or collection as part of a larger graph or as part of a named dataset catalogue. This practice is relational database friendly and also allows for rudimentary testing using technologies like e.g. JSON-schema. Connected to these practices is the widespread use of dataset repository technologies like e.g. OKF CKAN and CERN Invenio.

## Knowledge Packages: an example of PID graph implementation

The presentation published on Zenodo[101] represents the concept of “Knowledge Packages” - the self-describing structures stored within the given infrastructure. It supports data reproducibility through persistent identification and integrity provenance not only for data but also for technologies, software, instruments and procedures involved. The package is declared as a single programmable entity to be deployed via an explicitly defined, mathematically reversible algorithm. From FREYA project[102] and Debian Policy Manual[103], it borrows the concepts of deployment subroutines that allow encapsulation of provenance procedures into the package, in a programmable way.

There are other examples of PID graph implementation, e.g. DataCite invested significant effort in building a GraphQL interface[104] underpinned by connections across PID-assigned entities. There is a Research Data Alliance Open Science Graphs for FAIR Data Interest

[101] A. Vukolov. Knowledge Packages. Draft Concept. Presentation. DOI 10.5281/zenodo.4737497. https://github.com/twdragon/knowledge-package-zenodo-demo/blob/v0.0.1/presentation.pdf
[102] M. Povey, C. A. Amaro. Using the FREYA PID Graph to help reproduce scientific research. https://zenodo.org/record/4277945
[103] Debian Policy Manual. Release 4.6.0.1 https://www.debian.org/doc/debian-policy/policy.pdf
[104] DataCite GraphQL API Guide. https://support.datacite.org/docs/datacite-graphql-api-guide

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.*

Group[105] that can serve as a forum for research information practitioners to discuss practical applications of graph modelling and graph technology.

[105] RDA Open Science Graphs for FAIR Data IG.
https://www.rd-alliance.org/groups/open-science-graphs-fair-data-ig

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.*