Second draft Persistent Identifier (PID) policy for the European Open Science Cloud (EOSC)

DOI for version 2: https://doi.org/10.5281/zenodo.3780423

This policy was authored by representatives of the EOSC FAIR Working Group and EOSC Architecture Working group. See Appendix 3 for details. This is version 2 of the policy and released in May 2020 for community feedback and comment.

We welcome further responses to and comments on this version. Please share them with the wider community on https://pidforum.org or email them to us pid-policy@eoscsecretariat.eu. We understand that some areas may require discussion and encourage you to have that discussion on https://pidforum.org. That is where we will also provide details of opportunities for face-to-face feedback and discussion.

We will develop a final version that will be delivered to the EOSC Governance Board in October 2020.

1. Context

1.1. This Persistent Identifier (PID) policy is written for senior decision makers within potential EOSC service and infrastructure providers, and will be of interest to all EOSC stakeholders. It 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 will be approved by EOSC governance, who will also oversee its implementation. The implementation will be guided through recommendations on the PID Technical Architecture which the EOSC Architecture Working Group will provide.

1.2. Many types of Persistent Identifiers are already in active use for a range of applications, for example Handles, DOIs, ARKs, ISSNs, URIs, URNs, ORCIDs. In addition, numerous technologies, databases or cloud systems use local identifiers. This PID policy will not make detailed statements about existing systems, but provides definitions, guidelines and usage requirements.

1.3. The GDPR regulations are guiding us in the use of PIDs that may store and provide information about people, such as ORCID. PID systems should support GDPR compliance as far as possible, but it should be noted that once identifiers have been used to exchange information across infrastructures, especially where names are part of citations and metadata of published digital objects, it is almost impossible to universally redact that information.

2. Principles

2.1. The PID Policy will concentrate on principles, desired results and governance which are designed to establish a viable, trusted PID infrastructure suitable for the long-term sustainability of the EOSC.

2.2. There are a range of current use cases for PIDs in research and scholarship. The full range of use cases covers a wide variety of direct human use and automated machine processing and actionability. This PID Policy should accommodate a wide range of use cases and not put in place barriers to the effective use of PIDs as needed by the research community.

2.3. The policy should result in a future where PIDs can be used as the preferred method of referring to its assigned entity, where appropriate, alongside human-readable means e.g. the common name. Multiple PIDs may identify any given entity and users should be able to use whichever they are most comfortable with.

2.4. The PID Policy should enable an environment of research practice, and services that satisfy the FAIR principles as appropriate for the particular domains of use. Central to the realisation of FAIR are FAIR Digital Objects and PIDs are core to the idea of FAIR Digital Objects, as highlighted in the Turning FAIR Into Reality report (FAIR Expert Group, 2018).

2.5. Technology independence of PIDs is required to allow for technological change. PID services will vary in maturity over time and the PID policy should identify the level of service maturity suitable for EOSC adoption.

2.6. The policy seeks to accommodate mature and established PID practice, schemes, technologies and providers, which have a global presence. The policy also needs to be balanced and not prefer one approach or technology over another.

2.7. The Policy should propose obligations on PID providers and users that raise the level of confidence that the PID Infrastructure is stable for the long term.

2.8. The Policy should encourage new and innovative services and tools, which use and build on the PID Infrastructure.

2.9. The policy should be maintained by the EOSC Legal Entity and reviewed by a dedicated group of experts and stakeholders on a regular basis and at least in 3 years time. All of those who perform Roles as defined in Section 4 should be consulted and involved in the review process.

3. Generic PID Definitions

3.1. For the purposes of this Policy, a Persistent Identifier that supports and enables research that is FAIR is one that is globally unique, persistent, and resolvable. Each of these three features may be delivered through various means and technologies, but some basic principles apply.

3.2. Globally Unique

3.2.1. To enable global uniqueness, a PID should comply with a controlled syntax to avoid clashes, for instance, by having namespaces that are governed by clearly defined authorities.

3.3. Persistent

3.3.1. Persistence relates to three aspects. The first two relate directly to the PID itself.

3.3.1.1. The PID should be managed and governed in such a way that the community can trust it to remain unique and resolvable for the long term. This may be beyond the lifetime of the entity it identifies, the creators of that entity or even the PID service provider itself. The latter point is dependent on fulfilling 8.4.

3.3.1.2. The syntax of the PID system should also be persistent.

3.3.1.3. Finally, persistence also relates to the object to which the PID resolves (the referent). The referent should also be stable, whether it is a born digital object or a digital representation of a physical or abstract entity. In some cases absolute fixity of the referent is required, e.g. for a dataset in support of research that should not change to ensure reproducibility; and in other cases the concept of the entity as it is currently understood is maintained, for example, an institution that has changed over time. This persistence should also be considered in conjunction with versioning (see section 5.5 and 5.6).

3.4. Resolvable

3.4.1. There can be three intentions of PID resolution. A PID is resolvable when it allows both human and machine users to access:

3.4.1.1. A digital object, a digital representation, or information on how the object can be accessed.

3.4.1.2. Kernel Information: A global resolution system should support access to Kernel Information from its PID. Kernel Information is a structured record with attributes whose semantics are retrievable in machine-interpretable form. In general, the Kernel Information should at least contain attributes that point to where the bit sequence of the referent can be found and a pointer to a type definition. Optionally, it may contain pointers to further contextual objects including metadata. Kernel Information profiles should be registered in open registries.

3.4.1.3. When an object or its representation are no longer available, resolution to Kernel Information must still be possible. In such cases, a standard practice is to include 'tombstone' information as part of kernel information (cf. 5.6).

3.4.2. To make it globally resolvable, the PID needs to be part of a namespace defined by a syntax that is controlled by an Authority (section 4.4).

3.5. In addition to these features of the PID itself, the use (or application) and management of the PID, and its related services and governance, should combine to enable research that is FAIR.

3.6. Note that to avoid doubts this policy does not cover the case of more than one PID referring to the same entity.

4. Roles and Responsibilities

4.1. A PID Infrastructure within the EOSC has several defined roles which actors can undertake. Each role is responsible for a particular component within the PID Infrastructure, with particular commitments to maintaining the integrity of that PID Infrastructure. In a particular scenario, one actor can play more than one role, but it is useful to separate them to add clarity.

4.2. A PID Infrastructure may also be formed of several components, either services, rules or standards, that enable the Infrastructure to operate according to the policies and expectations of its target communities.

4.3. PID Scheme (Component). A set of rules and standards defining the nature of a PID. This would include a set of lexical formatting rules for PIDs within a namespace. It could also define for example: associated PID Type; definition of associated metadata; quality assurance conditions; usage rights, terms and conditions, and algorithmic methods for generating PID names and enforcing PID properties.

4.4. PID Authority (Role). A controller responsible for maintaining the rules for defining the integrity of PIDs within a PID Scheme. These rules may include setting standards for lexical formats, algorithms and protocols to ensure global uniqueness, together with setting quality of service conditions to enforce compliance to the rules. PID Authorities may be organisations (e.g. DOI.org), which enforce control over a PID infrastructure. But there may also be Authorities which do not have a central control, but provide a community standardisation mechanism that specifies the conformance of PIDs to a PID Scheme.

4.5. PID Service Provider (Role). An organisation which provides PID services in conformance to a PID Scheme, subject to its PID Authority. PID Service Providers have responsibility for the provision, integrity, reliability and scalability of PID Services, in particular the issuing and resolution of PIDs, but also lookup and search services.

4.6. PID Service (Component)**.** Basic services are those that create, manage and resolve PIDs and their associated kernel information which conforms to a PID Scheme. Advanced, value-added services may also be provided, for example attribute search or metrics.

4.7. PID Manager (Role)**.** PID Managers have responsibilities to maintain the integrity of the relationship between entities and their PIDs, in conformance to a PID Scheme defined by a PID Authority. A PID Manager will typically subscribe to PID services to offer functionality to PID Owners within the PID Manager's services. One example is a Service Provider which uses PID Services as part of its own service delivery. For example, PID Managers may include a provider of a data repository, a data catalogue, or a research workflow system.

4.8. PID Owner (Role). An actor (an organisation or individual) who has the authority to create a PID, assign PID to an Entity, provide and maintain accurate Kernel Information for the PID.

4.9. End User (Role). The end user of PID services and PID User Services. These can be for example researchers, or software, or services produced to support researchers. End users will use PIDs to cite and access resources or Kernel Information.

5. PID applications

5.1. An ecosystem of PID Infrastructures is needed to support the wide variety of scientific applications and offer sufficient flexibility (service providers, scheme, attribute set) and capacity. There should be a common Application Programming Interface[1] to interact with PIDs (create, resolve and modify PID and PID Kernel Information) that should be offered by all PID Service providers. The details of this shall be further defined as part of implementation work (EOSC Architecture WG).

5.2. PID services for EOSC need to address a wide variety of applications including those that require secure mechanisms built into the PID Infrastructure. One example is the case of confidential data where access control and/or encryption of the Kernel Information might be appropriate

5.3. It must be clear at all times who is the PID Owner and thus who is allowed to make changes to the attributes. The owner is responsible for proper management of PIDs and to keep the attribute set up-to-date.

5.4. Granularity of PIDs is very much dependent on the communities and it will change over time. Multiple levels of granularity should be supported by the PID ecosystem and linking between levels of granularity should be fostered. Ever deeper levels of granularity and the subsequent increase in the volume of PIDs that would be created, managed and resolved, which may

[1] The RDA Working Groups on PID Information Types and PID Kernel Information have produced preliminary recommendations and prototypes for such an interface and principles for Kernel Information that can form the basis for further work.

increase the effort and cost of managing a PID Service and its functionality and features. And so levels of granularity appropriate to community best practice and use cases should be provided, while allowing for flexibility to respond to how those needs and practices will evolve.

5.5. PID services should support versioning. PID Service Providers and repositories must have clear policies and guidelines on how to manage versioning in case the FAIR Digital Object or entity changes.

5.6. PIDs should not be re-assigned or deleted. In case that the entity being identified is deleted or ceases to exist, a tombstone note needs to be in the PID attribute set. Moreover, every PID should provide information on the referenced object's fundamental type and management policy in a machine-actionable way, whether the referenced entity is long-term available or not and whether it can be expected to change.

6. PID types

6.1. PIDs can identify many different entities. These can be born digital (e.g. documents, data, software, services - otherwise known as digital objects - and collections made of them), physical (e.g. people, instruments, artefacts, samples), or conceptual (e.g. organisations, projects, vocabularies).

6.2. Physical and conceptual entities must be represented via a digital representation (e.g. Landing Page, metadata, attribute set, database index) to have a presence in the digital landscape. All digital representations should be FAIR Digital Objects.

6.3. Classes of digital objects may need different attribute sets a PID is resolved to. It is the responsibility of a community of practice to define and document these attribute sets (profiles).

6.4. The PID Kernel Information record is a non-authoritative source for metadata focused on facilitating automation of processes. If the information for an attribute duplicates metadata maintained elsewhere, the external source is the authority.

7. PID services and PID service providers

7.1. A PID Service Provider should offer services that integrate well with European Research Infrastructures, but not at the exclusion of the broader research community. The basic services of PID registration and resolution will have no cost to end users.

7.2. As with other EOSC services, basic PID Services should be at Technology Readiness Level (TRL) 8 (system complete and qualified) or 9 (actual system proven in operational environment). Added value PID Services may be provided at lower TRLs.

7.3. PID Service Providers need to ensure 24/7 availability of basic PID Services. Responsibilities for service maintenance and availability need to be documented clearly.

7.4. PID Service Providers need to have a clear sustainability and succession plan with an exit strategy that guarantees the continuity of the resolution (as per 3.4.1) of its PIDs.

7.5. For EOSC we need a set of trusted registration PID Authorities and PID Service Providers that are regularly certified based on agreed rule sets. Certification should cover both resolvability of PIDs to information from PID Service Providers and their management processes for maintenance of PIDs. It should clarify who is responsible for keeping the Kernel Information up-to-date, if necessary, by enabling third parties to modify it.

8. Governance and sustainability

8.1. PID Service Providers for EOSC should apply appropriate community governance to ensure that their PID Services and Systems adhere to these policies, and are agile and responsive to the needs of research, Open Science and EOSC. As such, the active EU research community needs to be represented in the governance structure to be able to influence the activities and business models.

8.2. PID Services should be provided at justifiable cost to EOSC services and resources.

8.3. The governance structure should be embedded in global governance.

8.4. PID Service Providers should have a public and independently verifiable exit plan that assures continuity of their PIDs and PID Services should they cease to operate. This plan will ensure other providers can replicate the PIDs and services should the original provider cease to exist.

8.5. The EOSC should encourage new and innovative usages, services and tools which use and build on the PID Infrastructure.

Appendix 1: List of sources consulted

This is a list of existing publications, policies, recommendations and other texts consulted in the authoring of this policy statement.

Berg-Cross G, Ritz R, Wittenburg P (2018). RDA DFT Core Terms and Model. B2SHARE. http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318

European Commission Expert Group on FAIR Data (2018). Turning FAIR into Reality. Publications Office of the European Union. https://doi.org/10.2777/1524

Fenner M, Demeranville T, Kotarski R et al (2016). THOR: Conceptual Model Of Persistent Identifier Linking. Zenodo. https://doi.org/10.5281/ZENODO.48705

FORCE11 (2015). The FAIR Principles (2015) https://www.force11.org/group/fairgroup/fairprinciples

Lannom L, Broeder D, Manepalli G (2015): Research Data Alliance: Data Type Registries working group output. https://doi.org/10.15497/A5BCD108-ECC4-41BE-91A7-20112FF77458

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

Meadows A, Krznarich L, Brown J, et al (2019). Recommendations for Supporting ORCID in Repositories. figshare. https://doi.org/10.23640/07243.7777274.v3

The ODIN Consortium (2013): D4.1 Conceptual model of interoperability. figshare. https://doi.org/10.6084/m9.figshare.824314.v1

Parland-von Essen J (2019). The Use of Persistent Identifiers for Research Datasets Recommendation by the Finnish Scientific Community for Open Research. Edunni wiki. https://wiki.eduuni.fi/display/csctuha/The+Use+of+Persistent+Identifiers+for+Research+Datasets

Rauber A, Asmi A, van Uytvanck D, Proell S (2015): Research Data Alliance Working Group output: Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC). https://doi.org/10.15497/RDA00016

RDA GEDE Digital Object Topic Group (2019). FAIR Digital Object Framework https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/FDOF

Reed D, Sporny M, Sabadello M, et al (2019). Decentralized Identifiers (DIDs) v1.0. Core Data Model and Syntaxes. W3C Working Draft 27 November 2019. w3.org. https://www.w3.org/TR/2019/WD-did-core-20191127/

TechLib libraries (DTU Copenhagen; ETH Zurich; TIB Hannover; TU Delft) (2019). Pitching for PIDs: European support for a sustainable PID infrastructure - Avoiding a PIDfall. Hannover : Institutionelles Repositorium der Leibniz Universität Hannover. https://doi.org/10.15488/8825

Weigel T, DiLauro T, Zastrow T (2015). Research Data Alliance: PID Information Types WG final deliverable. Zenodo. http://doi.org/10.15497/FDAA09D5-5ED0-403D-B97A-2675E1EBE786

Weigel T, Plale B, Parsons M et al (2018). RDA Recommendation on PID Kernel Information (Version 1). Research Data Alliance. https://doi.org/10.15497/RDA00031

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

Wittenburg P, Hellström M, Zwölf C-M et al. (2017). Persistent identifiers: Consolidated assertions. Status of November, 2017. Zenodo. http://doi.org/10.5281/zenodo.1116189

Appendix 2: Glossary

These terms are provided to aid understanding of the policy, and are in addition to definitions of roles and responsibilities defined in Section 4. All terms need to be seen in relation to PIDs.

Attribute	A value that describes a feature of an object or its representation, as part of PID Kernel Information or other metadata.
Digital Object	A Digital Object has a bit sequence that can be stored in multiple repositories and is associated with a Persistent Identifier (PID) and metadata.
Digital Object Identifier (DOI)	A digital object identifier (DOI) is a persistent identifier based on Handle used to identify objects uniquely, standardized by the International Organization for Standardization (ISO).
Digital Entity	A digital Entity denotes any sort of bit sequence that is being stored or transmitted without being registered to enable sharing.
European Open Science Cloud (EOSC)	An integrated infrastructure to create a web of FAIR data. The development of EOSC is a large and ongoing multi-stakeholder initiative.
FAIR Digital Object	A model proposed by the Turning FAIR into Reality report, denoting what elements are needed for a Digital Object to be FAIR and thus machine actionable.
FAIR Principles	FAIR stands for Findable, Accessible, Interoperable and Reusable. It refers to the FAIR Data Principles developed by the FORCE 11 community, that recommend data should be shared according to these four concepts.
Granularity	The varying levels of hierarchy or constituent parts that may form data or other research outputs. For example, the differing levels of granularity of a research publication, going from a whole Journal issue, to its constituent articles, to the article's constituent sections or figures, the levels in a complex scientific collection or the level of detail in a large scientific database.
Handle	A Handle is a globally resolvable, unique and persistent PID which is defined by RFCs 3650, 3651 and 3652 of the Internet Engineering Task Force (IETF). They are used by DOI, ePIC and many other service providers. The Handle System is governed by the independent Swiss DONA Foundation.

Kernel Information	A PID needs to be resolved to a structured record consisting of well-defined attributes to allow machine actions. This is the concept of PID Kernel Information as described in the matching RDA Recommendation (Weigel, T., et al. 2018).
Landing Page	A human-readable page, displayed in a browser, that provides human users with information on how to access and/or interpret the digital object or its representation that is identified by a PID.
Machine Actionable	Machine Actionable means that a formal statement is syntactically and semantically specified enabling computing systems to carry out automatic processing.
Metadata	With metadata in the context of this PID policy we mean all kinds of assertions about properties of the bit sequence of a digital object such as descriptive, deep scientific, contextual, provenance, access permissions, transactions etc. This kind of metadata is not stored in the PID record as kernel attributes, however the PID record in general should point to the metadata.
Namespace	A namespace ensures that all the identifiers within a given level have unique names so that they can be easily identified.
Persistent Identifier (PID)	A persistent, unique and globally resolvable identifier that is based on an openly specified PID Scheme.
PID Ecosystem	The set of PID Components and Services that may be federated or interoperable to support EOSC and research that is FAIR in an effective and sustainable way.
PID Profile	The PID profile describes the set of kernel attributes that are being selected by a repository to describe the DO. The profile is dependent on the type of object and the needs of the community. Only openly declared attributes with widely agreed semantic types will provide machine actionability.
Prefix	A PID syntax component denoting an authority or division of a namespace. A prefix needs to be globally unique associated with particular local authorities which are free in how they generate locally unique IDs.
Semantics	The meaning or interpretation of meaning attached to a given text string. It is recommended to not include semantics in an identifier string where meaning may change over time or may not be understood across linguistic and cultural divides.
Suffix	A “suffix” is the part of a unique and persistent identifier that is created by a local authority which also specifies the syntax rules for the suffix.

Appendix 3: Policy Authors

Authors of this policy are a subset of members of the EOSC FAIR and EOSC Architecture Working Groups, as well as PID experts invited to contribute to the Architecture PID Task Force.

Name	**EOSC Affiliation**	**Affiliation**
Maggie Hellström	Architecture PID TF	Lund University and ICOS RI
André Heughebaert	FAIR WG	Belgian Biodiversity Platform
Rachael Kotarski	FAIR WG	British Library
Paolo Manghi	Architecture WG	ISTI/CNR
Brian Matthews	Architecture WG	UKRI-STFC
Raphael Ritz	Architecture WG	MPCDF/MPG
Anders Sparre Conrad	FAIR WG	Danish e-Infrastructure Cooperation
Tobias Weigel	Architecture WG	DKRZ
Peter Wittenburg	FAIR WG	Max Planck Institute
Mario Valle	Architecture PID TF	Swiss National Supercomputing Centre (CSCS)

Appendix 4: Versions and tracked changes

Version	Release Date	Location	Changes from previous version
1.0	2019/12/13	https://doi.org/10.5281/zenodo.3574203	N/A
2.0 (This version)	2020/05/01		Numerous changes in response to feedback from: ● PIDapalooza 2020 ● PID Forum ● Emails and other personal feedback New DOI for Version 2.