EOSC FAIR Executive Board Working Group

EOSC Architecture Executive Board Working Group

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

DOI for version 1: https://doi.org/10.5281/zenodo.3574203

This policy was authored by representatives of the EOSC FAIR Working Group and EOSC Architecture Working group. See Appendix 3 for details. This initial policy was released in December 2019 for community feedback and comment.

We welcome responses to and comments on this first 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 second version for March 2020 and a final policy will be delivered to the EOSC Governance Board in October 2020.

1. Rationale

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 in support of 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.

2. Principles

2.1. The PID Policy will concentrate on principles, desired results and governance which are designed to effect a sustainable, 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 the 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, and alongside any more obvious means e.g.

organisational 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 which satisfy the FAIR principles as appropriate for particular domains of use. Central to the realisation of FAIR are FAIR Digital Objects. These objects could represent data, software, protocols or other research resources. They need to be accompanied by PIDs and metadata rich enough for them to be reliably found, used by humans and machines, including automated data management processes, and cited.

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

2.6. There is a mature landscape of usage and practice. The policy seeks to accommodate mature and established PID practice, schemes, technologies and providers, which have global presence. The policy also needs to be balanced and not prefer one approach or technology over another as long as they fulfil the requirements and adhere to the policy.

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.

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 name should comply to a syntax that is controlled to avoid clashes, for instance by having namespaces that are governed by a single authority.

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 it can be trusted by the community 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 meeting 8.4.

3.3.1.2. The syntax of the PID should also consider persistence, and it is recommended to not include semantics in the identifier string where semantics may change over time or may not be understood across linguistic and cultural divides.

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 FAIR Digital Object or digital representation of a physical entity. In some cases absolute fixity of the referent is required, for example 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.

3.4. Resolvable

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

3.4.1.1. An object or its representation: This would either allow direct access to its assigned object or 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, further contextual objects including metadata referenced through their PID and a type.

3.4.1.3. When an object or its representation are no longer available resolution to Kernel Information must still be possible.

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 4.3).

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.

4. Roles and Responsibilities

4.1. A PID Infrastructure within the EOSC has a number of 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 for the purposes of clarity.

4.2. A PID Infrastructure may also be formed of a number of 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 Authority (Role). A controller who is 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.

4.4. PID Scheme (Component). A set of rules and standards defining the nature of a class of PIDs. 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.

4.5. 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 core PID Services, in particular the issuing and resolution of PIDs, but also lookup and search services.

4.6. PID Service (Component)**.** A service that creates and processes PIDs and their associated metadata which conforms to a PID Scheme. Service types can include: PID Issuing, PID Resolution, PID Search, PID Metadata, PID Linking, PID Graph, Citation services.

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. Adherence to the FAIR principles requires a highly scalable architecture with sufficiently large address space for a given application.

5.2. An ecosystem of PID Infrastructures is needed to support the wide variety of scientific applications and offer sufficient flexibility (service providers, scheme, attribute set). 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).

5.3. Applications require secure mechanisms built in PID Infrastructures and some applications require encryption of PIDs to protect activities.

5.4. It must be clear at all times who is the PID Owner and thus 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.

5.5. 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, will 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.

5.6. PIDs should support versioning. There should be clear guidelines for users assigning PIDs on how versioning is supported, and repositories need to have policies on how to manage versioning in case the FAIR Digital Object or entity changes.

5.7. 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 research 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).

[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.

6.2. Physical and conceptual research 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 research entities 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 should never be used as the single metadata storage point for any digital object.

7. PID services and PID service providers

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

7.2. A PID Service Provider can be characterised by providing basic services (e.g. PID registration, PID resolution, PID management, checks on PID correctness and consistency), within an identified PID Scheme.

7.3. PID Services should support interoperable Kernel Information profiles which should be registered in open registries.

7.4. As with other EOSC services, PID Services should be at Technology Readiness Level (TRL) 8 (system complete and qualified) or 9 (actual system proven in operational environment). The value of less mature technologies is recognised, and their maturity should be reassessed at defined intervals so that they can be brought into use as their maturity develops.

7.5. PID Service Providers need to ensure 24/7 availability. Responsibilities for service maintenance and availability need to be documented clearly.

7.6. PID Service Providers need to have a clear sustainability and succession plan with an exit strategy that guarantees the continuity of the resolution of its PIDs.

7.7. 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 Services Providers and responsibility for maintaining the link to the referenced entities. It should clarify who is responsible for keeping the Kernel Information up-to-date, if necessary, by enabling third parties to modify it. The initial scheme may be based on self-certification and not made mandatory yet. The existing certification schemes of DONA, ePIC and DOI could be used as a starting point.

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 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, needs to organise its representation and ensure that Open Science principles are upheld.

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.

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

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

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

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/

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.

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

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

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

The FAIR Principles (2015), https://www.force11.org/group/fairgroup/fairprinciples and Wilkinson M, Dumontier M, Aalbersberg I et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 https://doi.org/10.1038/sdata.2016.18

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

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

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

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

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

Appendix 2: GLOSSARY

These terms are provided to aid understanding of the policy.

Attribute	A value that describes a feature of an object or its representation, as part of PID Kernel Information or other 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)
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 of 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. 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
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.
Namespace	In computing, a namespace is a set of symbols that are used to organize objects of various kinds, so that these objects may be referred to by name. A namespace ensures that all the identifiers within it 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 schema.
PID Ecosystem	The PID Services that may be federated or interoperable to support EOSC and research that is FAIR in an effective and sustainable way.
Prefix	A namespace indicator has a syntax denoting different authorities such as <prefix>/<suffix>. 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.
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., Plale, B., Parsons, M., Zhou, G., Luo, Y., Schwardmann, U., Quick, R., Hellström, M., Kurakawa, K. (2018). RDA Recommendation on PID Kernel Information (Version 1). https://doi.org/10.15497/RDA00031
Suffix	A “suffix” is the part of a unique and persistent identifier that is created by a local authority.
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 quality metadata.
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.
Landing Page	A human-readble 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
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.

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

Appendix 4: Versions and tracked changes

Version Number	Release Date	Changes from previous version
1.0	2019/12/13	N/A