

Madurez de sistemas de identificadores persistentes: oportunidades en el contexto español

Persistent identifier systems' maturity: opportunities in the Spanish context

Paloma Marín-Arraiza

Marín-Arraiza, Paloma (2021). "Madurez de sistemas de identificadores persistentes: oportunidades en el contexto español". *Anuario ThinkEPI*, v. 16, e16a06.

<https://doi.org/10.3145/thinkepi.2022.e16a06>

Publicado en *IweTel* el 2 de marzo de 2022

Paloma Marín-Arraiza

<https://orcid.org/0000-0001-7460-7794>

Engagement Lead (Consortia team) en ORCID

p.arraiza@orcid.org



Resumen: Los identificadores persistentes (PID) bien establecidos constituyen una base de confianza para la ciencia abierta. Ayudan a garantizar que las entidades a las que se refieren sean descubribles, accesibles, evaluables, inteligibles, utilizables e interoperables. Tras introducir el papel de los identificadores persistentes en la infraestructura de investigación y en el proceso de FAIRificación, se describen diferentes sistemas de PIDs y sus niveles de madurez. Por último, se comentan oportunidades en el contexto español.

Palabras clave: Identificadores persistentes; Infraestructura de la investigación; Sistemas de identificadores persistentes.

Abstract: Well-established Persistent Identifiers (PIDs) form a trusted basis for open science. They help to ensure that the entities they refer to are discoverable, accessible, Assessable, Intelligible, Useable, Interoperable. After introducing the role of persistent identifiers in the research infrastructure and in the FAIRification process, different PID systems and their maturity levels are described. Finally, opportunities in the Spanish context are discussed.

Keywords: Persistent identifiers; Research infrastructure; Persistent identifier systems.

1. Introducción

La idea de un sistema de identificadores persistentes (PID) es proporcionar una referencia duradera a una entidad (física, digital o abstracta), por ejemplo, un documento digital, un sitio web, una persona o una institución.

Un PID tiene una serie de metadatos asociados que son legibles por máquinas y deben ir estructurados en una URL (**Dappert et al.**, 2017). De hecho, sin una URL clickable, los PIDs serían inútiles (**Kunze**, 2018). Sin embargo, es importante destacar que la persistencia está relacionada con el servicio ofrecido por el sistema de PID y no con el identificador en sí. Esto significa que existe una organización que se compromete a mantener el identificador resoluble. El identificador conduce a los usuarios a los servicios que garantizan la referencia (**Kunze**, 2013). Por ejemplo:

- los *Archival Resource Key (ARK)* se mantienen y resuelven a través del servicio *EZID (Universidad de California)*:
<https://arks.org>
<https://ezid.cdlib.org>

- los DOI son gestionados por la *International DOI Foundation* y sus agencias de registro correspondientes, como *Crossref*, *DataCite* o *mEDRA*:
<https://www.doi.org>
<https://www.crossref.org>
<https://datacite.org>
<https://www.medra.org>
- los *Handles* están gestionados por la *Corporation for National Research Initiatives*(CNRI):
<https://www.handle.net>

De todas las formas, cabe señalar que no necesariamente los sistemas de identificación tienen que estar centralizados. También existen alternativas descentralizadas que pueden considerarse en determinados casos (**Bach**, 2021; **Kunze**, 2018), sin perjuicio de las opciones centralizadas.

Actualmente, el uso de PIDs en infraestructura de información científica se ha generalizado e, incluso, se los considera una de sus piezas fundamentales (**Meadows; Haak; Brown**, 2019), así como un elemento principal en la FAIRificación de datos y metadatos. Así podemos verlo en los principios FAIR relativos a la encontrabilidad:

- F1: A los (meta)datos se les asigna un identificador global único y persistente.
- F3: Los metadatos incluyen clara y explícitamente el identificador de los datos que describen.

Asimismo, el informe *Turning FAIR into a reality* propone un modelo de objetos de datos FAIR (**Hodson et al.**, 2018), cuyas capas consisten en metadatos, normas, identificadores y datos.

Para entender el objeto de datos FAIR los autores exponen que:

“Los datos deben ir acompañados de identificadores persistentes (PID) y de metadatos básicos de descubrimiento para que se puedan encontrar, utilizar y citar de forma fiable. Además, los datos deben representarse en formatos estandarizados —e idealmente abiertos— y estar ricamente documentados mediante normas y vocabularios de metadatos adoptados por las comunidades de investigación para permitir la interoperabilidad y la reutilización. El intercambio de código

Casi la totalidad de las entidades del proceso de investigación puede ser identificada por medio de un PID. De ahí que sea común escuchar el “PIDs para todo” (o PIDs for everything)

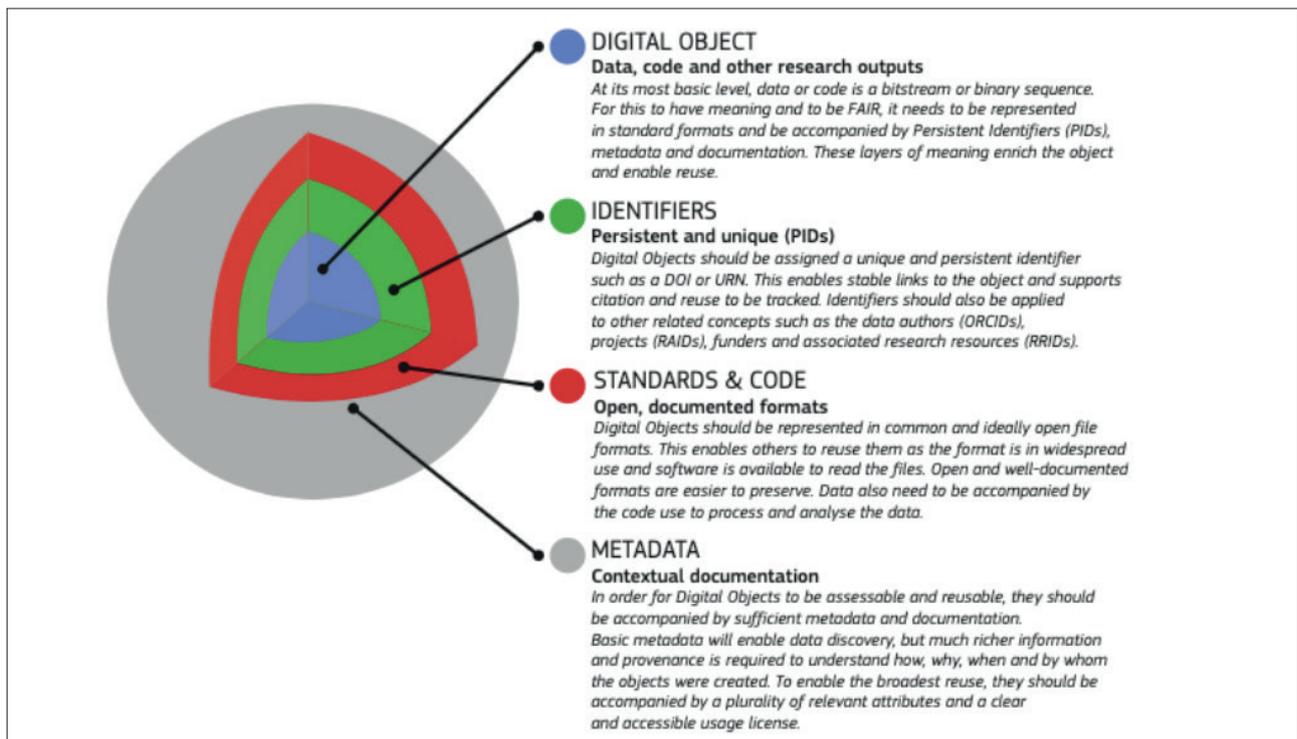


Figura 1. FAIR Data Object
Fuente: **Hodson et al.** (2018, p. 35)

también es fundamental y debe incluir no solo el propio código fuente, sino también la documentación adecuada, incluidas las declaraciones legibles por máquina sobre dependencias y las licencias” (Hodson et al., 2018, p. 39, traducción propia).

Además de para identificar y encontrar, los PID se utilizan para agregar recursos. Los resultados de investigación etiquetados con un PID son más fáciles de rastrear, lo que facilita las actividades de supervisión y monitorización. Sin embargo, como ya se ha mencionado, la persistencia no es una característica intrínseca de un PID, sino que está relacionada con el servicio y la organización subyacentes. También se utilizan como mecanismos de atribución al citar resultados de investigación (McMurry et al., 2017) y establecer vínculos entre entidades del ecosistema de investigación. De hecho, como afirman Wilkinson et al. (2016), las infraestructuras científicas —por ejemplo, los repositorios, los superordenadores o los equipos físicos— también pueden recibir un PID.

2. “PIDs para todo” y niveles de madurez en sistemas de PIDs

Casi la totalidad de las entidades del proceso de investigación puede ser identificada por medio de un PID. De ahí que sea común escuchar el “PIDs para todo” (o *PIDs for everything*). Dappert et al. (2017) agrupan estas entidades en cinco categorías:

- Agentes: individuos (investigadores o administradores); organizaciones (agencias de financiación, instituciones de investigación, centros de datos, editores e instituciones archivísticas); y otras entidades jurídicas (consorcios para colaborar en la investigación);
- Recursos: publicaciones, datos y otros artefactos de investigación, como cuadernos de laboratorio, programas informáticos, equipos o especímenes físicos;
- Declaraciones de derechos: acuerdos de concesión, licencias, patentes;
- Eventos: procesos relevantes para la procedencia de los recursos, como creación, curación, acceso, reclamación, actualización, citación, revisión;
- Entidades derivadas: como los proyectos, que pueden verse como un agregado de la entidad jurídica implicada, la agencia de financiación y los resultados obtenidos y las declaraciones de derechos que lo rigen. (Dappert et al., 2017, p. 6, traducción propia).

Los autores del texto anterior también visualizan las conexiones entre estas entidades mediante PIDs.

Para lo anterior, es importante señalar la madurez de la infraestructura que hay detrás de estos PID; es decir, la madurez del sistema de PID. Podemos considerar que una infraestructura ha alcanzado madurez cuando es de uso común en la comunidad investigadora y en todas las disciplinas del conocimiento.

En los estudios de Ferguson et al. (2019) y Lavasa et al. (2018) en el marco del proyecto Freya se analizan los niveles de madurez de diferentes sistemas de PIDs para varios tipos de entidades y su adopción en infraestructuras. <https://www.project-freya.eu/en>

La tabla 1 muestra, de todos los PIDs recogidos en el estudio, aquellos que son abiertos y que poseen una clasificación de “maduro” o “emergente”. El concepto de PID abierto hace referencia, en este caso, a aquellos PIDs

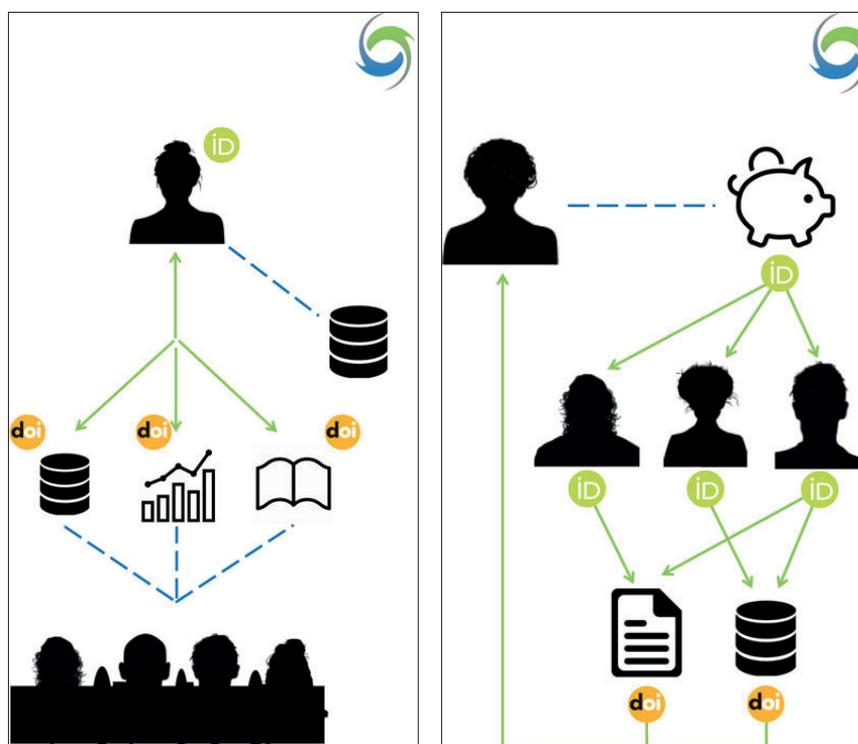


Figura 2. Relación desde el punto de vista de los investigadores que comparten datos (izquierda) y desde el punto de vista de las agencias de financiación (derecha).

Fuente: Dappert et al. (2017).

que no solo se pueden resolver, sino que también se pueden utilizar para descubrir metadatos abiertos, interoperables y bien definidos que contengan información de procedencia de manera predecible y pueden ser usados en cualquier tipo de sistema; es decir, no tienen ataduras a una infraestructura en concreto.

Tabla 1. Sistemas de PIDs abiertos y maduros o emergentes.

Entidad	Sistemas de PIDs usados	Nivel de madurez
Publicaciones	DOI, Handle, URN, PMID, arXiv identifier, ISSN, ISBN, PURL	Maduro
Citas	OpenCitation Identifier (OCI) https://opencitations.net/oci	Emergente
Conferencias	DOI	Emergente
Investigadores/as	ORCID iD, ISNI	Maduro
Organizaciones	Funder Registry, GRID, Research Organization Registry (ROR)	Emergente
Datos	DOI, Handle, PURL, URN, ARK	Maduro
Becas/Financiación	Grant DOI, PURL https://www.crossref.org/community/grants/	Emergente
Proyectos	Research Activity Identifier (RAiD) https://www.raid.org.au/	Emergente
Investigaciones	DOI	Emergente
Software	DOI, SHA-1 hash	Emergente
Simulaciones por computador	Universally Unique Identifier (UUID) https://www.uuidgenerator.net/	Emergente
Instrumentos, sensores, aparatos, plataformas, instalaciones de investigación	DOI, Research Resources Identifier (RRID) https://www.rrids.org/	Emergente
Archivo o instalación de almacenamiento	URI, DOI, UUID	Emergente
Muestras geológicas o biológicas	RRID, DOI, International GeoSample Number (IGSN) https://www.igsrn.org/	Emergente
Artefactos culturales	DOI, URN	Emergente

Fuente: Adaptado y completado de **Ferguson et al.** (2019)

Cabe señalar que algunos sistemas PIDs maduros como los DOIs, pueden ser considerados emergentes cuando se usan para determinadas entidades. Así, el uso de DOIs en publicaciones textuales y datos tiene una alta madurez y está aceptado por la comunidad científica, desde investigadores hasta instituciones. Sin embargo, los DOIs para becas o financiación o para software están aún emergiendo.

Una situación similar sucede con los identificadores para instituciones. Si bien el esquema Ringgold utilizado por grandes casas editoriales, o el esquema GRID gestionado por Digital Science, llevan tiempo en el ecosistema, no ha sido hasta la irrupción más fuerte del ROR ID que la identificación de instituciones no ha tomado mayor protagonismo.

<https://www.ringgold.com/ringgold-identifier/>

<https://www.grid.ac>

<https://ror.org>

Por ejemplo, en el análisis de 2019 citado en la tabla 1, ROR no aparecía. Desde que en julio de 2021 GRID anunció el cese de las actualizaciones de la base de datos y el paso a ROR, muchas instituciones lo han adoptado como estándar. De hecho, ya es el identificador por defecto en los esquemas de metadatos de DataCite, Crossref y ORCID.

<https://www.digital-science.com/press-release/grid-passes-torch-to-ror>

Queda el campo abierto para reflexionar sobre el uso de identificadores persistentes en España a nivel individual e institucional y profundizar desde instituciones y grupos de trabajo sobre buenas prácticas para su adopción e integración en flujos de trabajo. De esa manera se podrá contribuir al establecimiento de sistemas maduros y a la consolidación de los emergentes

3. Contexto europeo y oportunidades en España

En el entorno europeo hay una gran ventana de oportunidad abierta para la adopción de sistemas PIDs y la interoperabilidad de infraestructuras de la información.

Desde el proyecto supranacional de arquitecturas federadas y servicios de datos, *European Open Science Cloud (EOSC)*, en concreto desde el grupo de trabajo *FAIR y Architecture*, se publicó en 2020 una política de PIDs para *EOSC*. El objetivo principal:

“definir un conjunto de expectativas sobre los PIDs que se utilizarán para apoyar un entorno de investigación FAIR que funcione” (*European Commission*, 2020, p. 7).

También empiezan a emerger las estrategias nacionales de PIDs. Así se observa en el grupo de trabajo *National PID Strategies* promovido por la *Research Data Alliance*.

<https://www.rd-alliance.org/groups/national-pid-strategies-wg>

En Reino Unido se ha establecido un consorcio de PIDs (*UK PID consortium*) para ir un poco más allá en la adopción nacional de identificadores y la reutilización de metadatos (**Brown et al.**, 2021). Este consorcio funcionará en conjunto con el consorcio de *Orcid* (liderado por *Jisc*) y el de *DataCite* (liderado por la *British Library*).

Alemania ha presentado interés en realizar un proyecto de red de PIDs (*PID Network*) para continuar y ampliar el trabajo realizado en el consorcio de *Orcid*. Los dos proyectos predecesores, *Orcid DE I y II*, dieron lugar a la vinculación de *Orcid* al *Bielefeld Academic Search Engine (BASE)*, a la conexión de *Orcid* con el control de autoridades de la *Biblioteca Nacional Alemana*, así como a gran parte del desarrollo del plugin entre *OJS* y *Orcid*.

<https://info.orcid.org/announcing-the-base-orcid-search-link-wizard/>

<https://www.dnb.de/EN/Professionell/ProjekteKooperationen/Kooperationen/ORCID/orcidInfo.html>

<https://docs.pkp.sfu.ca/orcid/en>

A su vez, 13 de los 24 consorcios nacionales de *Orcid* están en Europa y contribuyen activamente a la creación de una comunidad de práctica en torno a los identificadores persistentes.

En ese contexto, a España se le presenta una ventana de oportunidad. Por un lado, legislativamente, la modificación de la *Ley 14/2011 de 1 de junio de la ciencia, la tecnología y la innovación*, en la que la ciencia abierta y las infraestructuras de investigación ganan protagonismo.

<https://t.co/UiYFvHlfkD>

Por otro lado, a nivel de comunidad cada vez hay más diálogo sobre la nueva generación de metadatos. Por ejemplo, en las mesas redondas organizadas por *OCLC* en las que los identificadores persistentes juegan también un papel fundamental.

<https://hangingtogether.org/spanish-round-table-on-next-generation-metadata-managing-researcher-identities-is-top-of-mind>

Queda el campo abierto para reflexionar sobre el uso de identificadores persistentes en España a nivel individual e institucional y profundizar desde instituciones y grupos de trabajo sobre buenas prácticas para su adopción e integración en flujos de trabajo. Se podrá, de esa manera, contribuir al establecimiento de sistemas maduros y a la consolidación de los emergentes.

4. Notas

1. Adaptación al castellano de FAIRification, término utilizado en inglés para referirse al proceso de aplicar los principios FAIR (*Findable, Accesible, Interoperable y Reusable*) a conjuntos de datos, metadatos e infraestructuras.

2. Los principios FAIR se listan y detallan en la página de la iniciativa global que los respalda “Go FAIR”

<https://www.go-fair.org/fair-principles/>

5. Referencias

Bach, Nicolas (2021). “Dezentrale Identifikatoren (DIDs)”. *o-bib. Das offene Bibliotheksjournal*, v. 8, n. 4.

<https://doi.org/10.5282/O-BIB/5755>

Brown, Josh; Jones, Phill; Meadows, Alice; Murphy, Fiona; Clayton, Paul (2021). *UK PID consortium: Cost-benefit analysis* (1.0). Zenodo.

<https://doi.org/10.5281/zenodo.4772627>

Dappert, Angela; Farquhar, Adam; Kotarski, Rachael; Hewlett, Kirstie (2017). “Connecting the persistent Identifier ecosystem: building the technical and human infrastructure for open research”. *Data science journal*, n. 16.

<https://doi.org/10.5334/dsj-2017-028>

European Commission (2020). *A Persistent Identifier (PID) policy for the European Open Science Cloud (EOSC)*. Publications Office. European Commission. Directorate General for Research and Innovation. EOSC Executive Board. <https://data.europa.eu/doi/10.2777/1926037>

Ferguson, Christine; McEntrye, Jo; Bunakov, Vasily; Lambert, Simon; Van-de-Sandt, Stephanie; Kotarski, Rachael; Stewart, Sarah; MacEwan, Andrew; Fenner, Martin; Cruse, Patricia; Van-Horik, René; Dohna, Tina; Koop-Jacobsen, Ketil; Schindler, Uwe; McCafferty, Slobhan (2019). *D3.1 Survey of current PID Services Landscape—Revised*. <https://doi.org/10.5281/zenodo.3554254>

Hodson, Simon; Jones, Sarah; Collins, Sandra; Genova, Françoise; Harrower, Natalie; Laaksonen, Leif; Mietchen, Daniel; Petrauskaitė, Rūta; Wittenburg, Peter (2018). *Turning FAIR data into reality: Interim report from the European Commission Expert Group on FAIR data*. <https://doi.org/10.5281/zenodo.1285271>

Kunze, J. (2013). *The ARK identifier scheme*. <https://tools.ietf.org/html/draft-kunze-ark-18>

Kunze, John (2018). "10 persistent myths about persistent identifiers (PIDs)". *Twitter*. @jakkbl, 24 agosto. <https://twitter.com/jakkbl/status/1033059109346586624>

Lavasa, Artemis; Dallmeier-Tiessen, Sunje; Van-de-Sandt, Stephanie; Tsanaktsidis, Ioannis; Trzcinska, Anna; Fokianos, Pamfilos; Dohna, Tina; Koop-Jacobsen, Ketil; Schindler, Uwe; Lemon, Barbara; Kotarski, Rachael; Ferguson, Christine; McEntrye, Jo; Lambert, Simon; Bunakov, Vasily; Baars, Chris; De-Jong, Maaike (2018). *Deliverable 4.1 integration of mature PID types*. <https://doi.org/10.5281/ZENODO.2414839>

McMurry, Julie A.; Juty, Nick; Blomberg, Niklas; Burdett, Tony; Conlin, Tom; Conte, Nathalie; Courtot, Mélanie; Deck, John; Dumontier, Michael; Fellows, Donal K.; Gonzalez-Beltran, Alejandra; Gormanns, Philipp; Grethe, Jeffrey; ... Parkinson, Helen (2017). Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLOS Biology*, v. 15, n. 6, e2001414. <https://doi.org/10.1371/journal.pbio.2001414>

Meadows, Alice; Haak, Laurel L.; Brown, Josh (2019). "Persistent identifiers: The building blocks of the research information infrastructure". *Insights the UKSG journal*, n. 32. <https://doi.org/10.1629/uksg.457>

Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, Ijsbrand-Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; Da-Silva-Santos, Luiz B.; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; ... Mons, Barend (2016). "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific data*, n. 3, 160018. <https://doi.org/10.1038/sdata.2016.18>

Dialnet | métricas

Nueva edición 2020

Fundación Dialnet

dialnet.unirioja.es/metricas