The National Roadmap for Persistent Identifiers for Finland

Traceable and distinct information (about objects) can be found and reliably linked now and in the future

PID Forum Finland

April 2023 (with minor updates in March 2024)

https://urn.fi/URN:NBN:fi-fe2024032512910

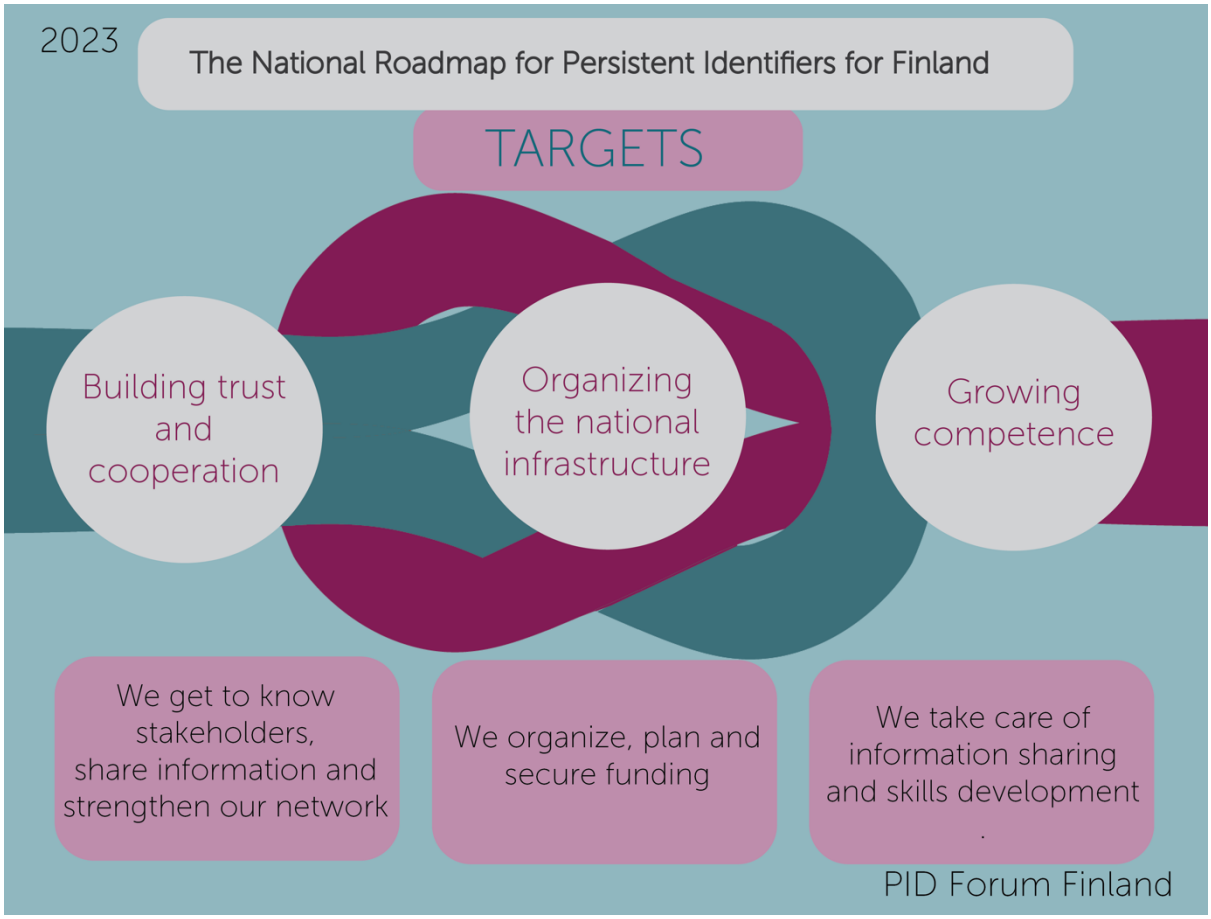
Slightly modified version from the original authoritative text in Finnish:
https://urn.fi/URN:NBN:fi-fe2023042138021

Summary

This document describes the field of persistent identifiers (PID) and the stakeholders in Finland. The first version of the roadmap is focused on research identifiers, but the actions it proposes can be extended to the wider field of data management as well.

The actions proposed in the roadmap aim for the following target state set by the expert network: "Traceable and distinct information (about objects) can be found and can be reliably linked now and in the future".

PID Forum Finland proposes the following actions to achieve the target state nationally:



This roadmap will be updated together with the stakeholders.

Content

Introduction

This roadmap has been produced by PID Forum Finland, which is an open expert network for stakeholders managing persistent identifiers.

With well-managed persistent identifiers (PIDs), an object is always discoverable, unambiguously identifiable, and traceable. Persistent identifiers enable reliable information to be searched for and found, even if the technological environment for providing information changes. Effective management and appropriate use of PIDs are beneficial to science and research as functions that support the reproducibility of research and accurate documentation. In addition, they support the recognition of merit for researchers and other stakeholders, as the utilization and reuse of research outputs and resources can be monitored with the help of persistent identifiers. Correctly used identifiers make referencing unambiguous and reliable.

PIDs can also be used for the documentation and repetition of repeatable, functional processes such as conversion or analysis. In addition, the management of research information is significantly improved.[1] The same can be assumed to be true regarding all administrative information management. Monitoring and documentation of the life cycle of information is a prerequisite to ensure long-term availability of information. Metadata related to identifiers enable information to be found, distributed, and integrated, as well as provisioning various services.

A persistent identifier can be assigned to different types of objects, such as:

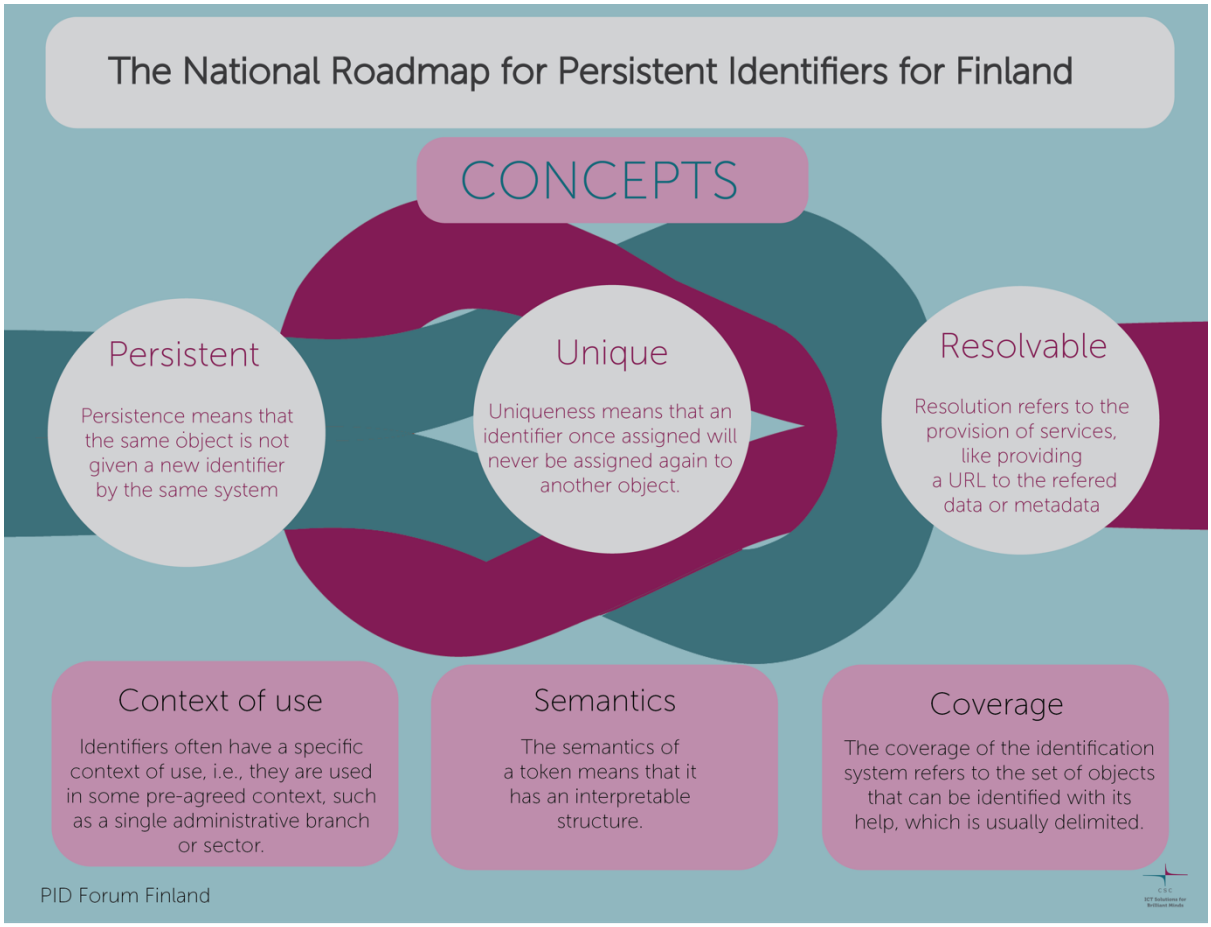
- For works or information content as different entities, for example administrative decisions or their parts.
- For digital objects, for example files, software, database queries or data; for the sub-objects of these data objects as well as the metadata concerning them as different entities.
- For physical objects, for example people, organisations, research equipment and their parts such as temperature sensors.
- For concepts and terms that can be stored in a machine-readable format so that the relationships between the concepts are also machine actionable.

This roadmap is a document prepared as a collaboration by Finnish experts. The work has been done within the framework of the open network PID Forum Finland in 2021 and 2022. We believe that persistent identifiers are at the core of sustainable and efficient data management and they require attention as part of the national data infrastructure - not least as the importance of researched and reliable data grows.

[1] Brown, Josh, Jones, Phill, Meadows, Alice, & Murphy, Fiona. (2022). Revised cost-benefit analysis for the UK PID Support Network (Version 1). Zenodo. https://doi.org/10.5281/zenodo.7356219

This document focuses specifically on persistent identifiers used within research in Finland. The aim is to extend what is shown here to other fields of information management as appropriate. There are concepts related to PIDs and their management, which we will discuss in the next chapter. After that, we describe the current situation in Finland - the PID systems in use and the actors operating in the field. Finally, we propose measures to achieve the target state we defined at the beginning of the document.

Basic concepts related to persistent identifiers



This chapter describes persistent identifiers using the basic concepts associated with them and defines how the terminology is used in this document. The order of concepts is therefore structured in such a way that the chapter forms as logical a whole as possible, where the meaning of the terms is explained as clearly as possible with examples.

Persistence

Persistence means that the same object is not given a new identifier by the same system.

Persistent identifiers are managed throughout the life cycle of the identified object and even after it has ended: the identified object can be destroyed or deactivated, but the persistent identifier is not destroyed or used again.

Persistence is an absolute requirement, and it requires documentation of both the identification system and the identifiers. The persistence of the identifier is in practice a promise that the identified object can be found and used throughout its life cycle, and some information about it even after that.

From the point of view of public administration information management, the persistence of identifiers is a key prerequisite for good administration, because the authority must take care of the correctness, accessibility, and traceability of the information it provides.

Uniqueness

Uniqueness means that an identifier once assigned will never be assigned again to another object. At the level of persistent identifier syntax, this requires management and control.

Context of use

Identifiers often have a specific context of use, i.e., they are used in some pre-agreed context, such as a single administrative branch or sector. The context can be based on an obligation given in legislation - for example, the European Union INSPIRE directive requires so-called Cool URI type identifiers for spatial data and in addition to this and in deviation from it, a national-level recommendation on a unique identifier for spatial data had been given that enabled broader contexts.[2]

Delimiting the context makes administration easier and enables more precise control, for example, regarding data models. However, delimiting the context does not guarantee interoperability, but it is necessary to take care, for example, of the uniformity of the metadata stored based on the identification system.

This document mainly focuses on the usage context of science and research.

Resolution and functionality

Resolution refers to the provision of services related to the identified object. The service can be, for example, the return of an Internet resource, a URL of descriptive metadata, or

[2] Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE): https://eur-lex.europa.eu/eli/dir/2007/2. Cool URIs don't change (https://www.w3.org/Provider/Style/URI /) is a pamphlet style guideline to design stabile URIs written by Tim Berners Lee in 1998. It has inspired also some further guidelines, for example Cool URIs for the Semantic Web (2008) https://www.w3.org/TR/cooluris/. JHS193 was a national recommendation for public administration on spatial data, currently archived here (in Finnish only): https://geoforum.fi/jhs-193-paikkatiedon-yksiloivat-tunnukset/.

metadata in response to an interface query. The returned URL is not necessarily the address of the landing page, but the user can be redirected to the correct URL.

A resolvable identifier is called functional. Traditional identifiers such as a book's ISBN are not functional as such, even though they can be used for e.g., Google searches. Many traditional identifiers such as ISBN and ISSN can also be presented as a functional identifier, i.e., as a URN or DOI identifier, for example. DOI and Handle identifiers are always functional, URN identifiers are either functional or non-functional depending on the namespace. For example, URN:ISBNs are functional, but URN:DEV - tokens (Internet of Things tokens) are not.

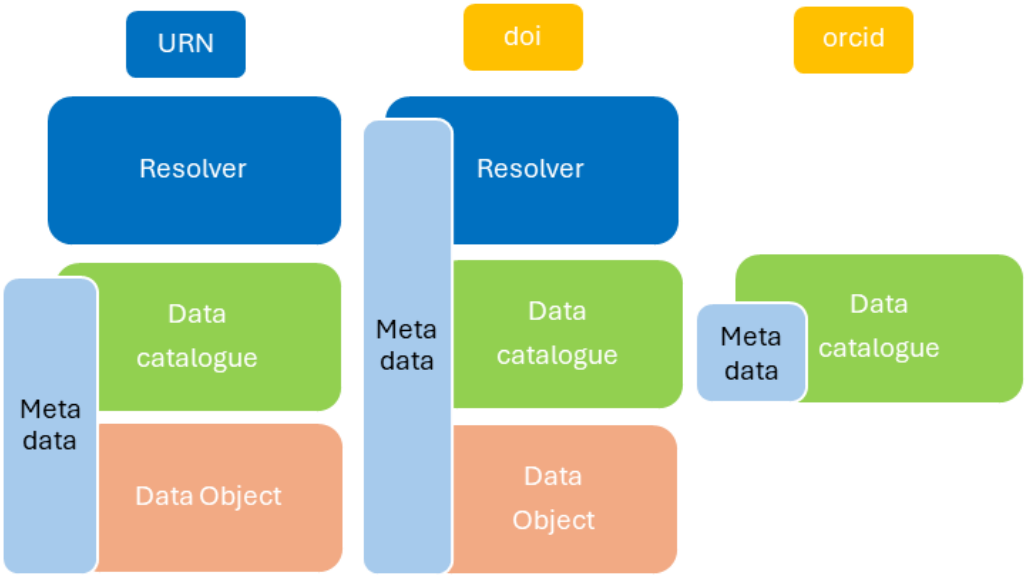
The resolution service (resolver) is an application responsible for resolving functional identifiers, i.e., redirection, such as Handle.Net related to Handle System, which is used in both Handle and DOI (Digital Object Identifier) systems. The URN (Uniform Resource Name) does not have one common application package, but in principle each entity that distributes URNs has its own resolver application. The resolution service can include many types of intelligence and additional services, so it is worth choosing a reliable service that best serves your needs.

Identifiers can be resolved in several different ways. The most common way is the normal network name service used for Cool URI type identifiers. The resolvability of these is therefore dependent on the namespace of the web address, and thus does not contain any other mechanism that provides permanence, if, for example, the web address of the organization offering the service changes.

In practice, the resolvers' service offering is based on the data stored in the resolvers[3]. The amount of data to be copied or created for a separate resolver (kernel metadata) is generally recommended to be kept to a minimum, because master metadata should be stored and maintained in the same place as the digital object or registry. Creating several copies is not only pointless, but also risky if errors occur when curating or updating the data. Copying is avoided by using information from its original source with a unique identifier. Metadata can be stored either in a separate catalogue or as part of the data object itself. Metadata should be structured so that it can also be interpreted programmatically. In this case, the metadata can also be easily indexed to be searched.

Metadata enables the provision of services related to PIDs, for example improving discoverability or machine utilization of information. Metadata and the chosen protocols and standards also greatly influence their technical interoperability and benefits. Since margins and usage contexts vary, there can and probably should be several levels of interoperability.

[3] It is possible to use the information stored in DOI and Handle resolvers to create a national guideline for each material type to harmonize the service offering. For the time being, it is only possible to save the URN and one or more URLs to which it links to the National Library's URN resolver. The application can be developed, for example, by enriching the data to be stored.



Different systems offer different metadata possibilities, which should be considered when planning data management.

Semantics

The semantics of a token means that it has an interpretable structure. For example, the ISBN (International Standard Bibliographic Number) of a book can be used to determine the country of publication and the publisher.

Organizations that develop and create identifiers do not agree on whether identifiers should be semantic or not. For example, publishers claim that the publisher information contained in the ISBN of the first edition of a book becomes obsolete when the publisher of subsequent editions of the book changes. But from the point of view of libraries, the publisher information included in the ISBN is permanently valid and enables the original publisher of the book to be checked from the publisher register of the national ISBN centre. Thanks to its semantics, the ISBN tells you where the resolution service can be found. The resolution of non-semantic identifiers such as ISSN must be handled centrally because the identifier does not provide any hint based on which the correct resolution service could be selected. Of course, finding the resolver is only challenging if the address of the resolver is not included in the ID.

Coverage

The coverage of the identification system refers to the set of objects that can be identified with its help, which is usually delimited. Only books can be identified with ISBNs, and researchers are identified with ORCIDs (Open Researcher and Contributor Identifier). In standardized identification systems, the cover is usually already defined in the standard itself. Limiting the margin makes it easier to manage the identification system.

Operators' identifiers are in a weaker position: researchers' identifier ORCID does not cover all researchers, nor does the operators' identifier ISNI (International Standard Name Identifier), the organizations' identifier ROR (Research Organization Registry) is still in the starting pits, and the RAiD (Research Activity Identifier), identifier for projects, is just coming into use. There are even weaker areas: there is an ISO standard identifier ISCI (International Standard Collection Identifier) for collections, but its use is minor and its application as a PID identifier is non-existent. Organizations' own identification systems are usually not based on official standards, except for the ISNI identifier.

When developing and implementing local identifier systems, the possibility of using standard identifiers should always be explored. Important identification systems whose coverage is basically unlimited are, for example, ISO's OID (object identifier), Handle, DOI, UUID (Universally Unique Identifier), URN and the so-called Cool URI. In practice, however, the application of PIDs is not free. The distribution of DOI identifiers is limited by agreements with Registration Agencies such as Crossref and DataCite. For example, if there are two versions of the data material with the same content in different file formats, they must be given a common DOI identifier. If, to enable long-term preservation, you also want your own PIDs for the versions, you must use some other PID.

In principle, any of the above-mentioned identification systems could be used as a public identifier for students, but in Finland the Board of Education has introduced an OID-based student number as a student identifier, and the student number can be made into a URN (URN:OID) or another functional identifier.

For example, datasets can be given a DataCite DOI, but metadata records describing them can be given a URN:NBN. For data material (or data object), DataCite's DOI works well at the object level. So, if there are CSV and Excel versions of the data object, they then have a common DOI, which is linked to the common landing page of the versions. For end users of the data object, this may be sufficient. But if the data object is stored long-term in, for example, CSC's preservation service, PID identifiers are also needed for the versions of the data object and the metadata describing them. In addition, in metadata, individual elements such as persons, organizations and terms describing the object and their content can have their own PID identifiers. How and what kind of PIDs are given, for example, to different parts of data (metadata, metadata elements, files or even to individual data objects or values) is context dependent. The most important thing is to create uniform and documented principles that serve the needs of the user community and support efficient and high-quality information management. The principles and practices should be published, for example, as part of the data policy, service documentation, material management plan or as a separate PID policy. It is important to remember that PIDs are not a guarantee of quality, there are many PIDs that do not have any kind of review process to issue them.

In the publishing industry, the common coverage of standard identifiers is quite good, but not perfect, and not even all standardized identification systems have been widely adopted. For example, a DOI identifier can be assigned to a scientific article, but if there are versions of the article in different file formats, they must be assigned a different identifier, i.e. for example URN:NBN.

The effective application of persistent identifiers and the avoidance of overlaps and gaps require careful planning and the parallel use of different identification systems.

Metadata model

Identifiers can be associated with a metadata model. When giving an ID, the identified object must be described either in the standard or in another way defined. For example, the ISSN standard requires that each Periodical that receives an identifier is described, and the information is sent to the ISSN database maintained by the international ISSN centre. The ISSN network has drawn up detailed instructions for storing metadata. The coverage of the metadata model varies depending on the identification system. To get an ORCID ID, the researcher only has to provide his/her first name and email address but getting an ISNI (International Standard Name Identifier) operator identifier requires that there is enough metadata to clearly distinguish the operators from each other (e.g., unambiguous name, year of birth and title of publication). If there is not enough information (e.g., the person's year of birth is missing), then confirmation is expected from another organization. The more comprehensive the metadata model, the more efficient the data management. The Albert Einstein found in the ISNI database can be reliably identified using the publication information contained in the database, but the identity of Albert Einsteins in the ORCID database is more uncertain. On the other hand, the requirement to store comprehensive metadata and complicated data management instructions make creating PIDs laborious. In addition, e.g. identifiers may include personal information such as date of birth, which cannot be shared freely because it is a personal register.[4]

[4] The definitions are based on a report on PIDs in public administration (in Finnish): Tunnuskäytänteet julkisessa hallinnossa-työryhmän loppuraportti. YTI-selvitys 2018. Version 1.01. pp. 2–5. https://vm.fi/hanke?tunnus=VM032:00/2017

The need for a national PID Policy

This chapter describes the importance of persistent identifiers for administration and information management in general.

Persistent identifiers as part of public information management

Authorities are obliged to provide information in machine-readable form via interfaces.[5] Also, the Ask Only Once principle and European interoperability requirements require management of data quality, provenance, and integrity data.[6] The key importance for the realization of these is the recognisability of the single explanations of the information and the reliable possibility of referring (linking). In practice, this requires managing persistent identifiers. However, currently the field and management of persistent identifiers is fragmented, and common practices are needed. Identifiers should be identified as part of the information management map, as common practices should be maintained and updated in cooperation with actors from different fields. The practices must be based on statutory reasons, so that there are no factors open to interpretation in the use of PIDs. The starting point for the management of identification systems must be that all identifications have a responsible party whose activities are sufficiently organized and financed. It is important to define which digital objects need PIDs and why these objects are important and which entity is responsible for the identifiers. Several different roles are associated with PIDs, which are explained in more detail in the subsection Actors and roles in the context of use of research. Regarding research data, it has been shown that the use of PIDs in data management also saves resources.[7]

The Ministry of Finance has highlighted the need for policies at different levels based on European interoperability work.[8] In order for interoperability to be implemented, interoperability at the operational, tactical, and strategic levels is needed. Based on the information management map of the central government, a more precise architecture should also be made for identifiers due to their central role.

---

[5] Act on Information Management in Public Administration, English translation: https://www.finlex.fi/en/laki/kaannokset/2019/en20190906 (e.g. §22, §24 §24a and §24b)

[6] Brown, Josh, Jones, Phill, Meadows, Alice, & Murphy, Fiona. (2022). Revised cost-benefit analysis for the UK PID Support Network (Version 1). Zenodo. https://doi.org/10.5281/zenodo.7356219

[7] More about EU interoperability https://joinup.ec.europa.eu/ . "Once only" is defined as a goal in the European Union (https://ec.europa.eu/digital-building-blocks/wikis/display/DIGITAL/Once+Only+Technical+System). The same principle is also related to the MyData principle, and it is mentioned in the proposed measure paper 12/22 published by the cooperation group of data policy actors of all parliamentary parties. https://tietopolitiikka.fi/. Definition of MyData according to SITRA: https://www.sitra.fi/en/dictionary/mydata/

[8] Guidelines for public administration interoperability (in Finnish) https://vm.fi/julkisen-hallinnon-yhteentoimivuuden-linjaukset, European Interoperability Framework https://joinup.ec.europa.eu/collection/nifo-national-interoperability-framework-observatory/european-interoperability-framework-detail, European Interoperability reference architecture https://joinup.ec.europa.eu/collection/european-interoperability-reference-architecture-eira/solution/egovera .

A national PID policy is needed, because clear guidelines and practices are needed to determine persistent identifiers, for example based on the Data Management Act and the data management model of the central government. Public Sector ICT (JulkICT) and the information management board play a key role. If we look at the public administration broadly as a producer of information (the state and municipalities), the key question is what form of (technical) identifier they give to the thing, document, etc. object, which should be technically referable in the later stages of the process and in other processes. For example, cities can have hundreds of information systems with different ways of identifying information data. At this point, resolution is also a challenging issue, because the availability of information is often not yet completely open in the environment in question, but perhaps sometime in the future this is how we want to act. It would be essential that the given identifiers would eventually be carried along with the information to the archives, and thus they could be used in research in the future as well.

The goal of the national PID policy would be to create practices that allow the creation of persistent identifiers to take place as part of the data production process and not as an afterthought, and PIDs are comprehensively utilized across organizational boundaries. The regulation of theses management created by the Ministry of Education and Culture is a good example of successful practice. The background must be a report that maps out usage needs and cases.

Basic register authorities and, for example, research institutes should be aware when the identifiers they use should be permanently identifiable and they themselves should act as reference information services. This becomes important at the latest when the materials are opened for use online through interfaces, but for the reasons mentioned above (e.g., archiving and long-term preservation), the management of identifiers at the organizational level is also useful in promoting the interoperability of administrative boundaries, for example. The management of identifiers should be coordinated as part of the public administration's information management map, which would improve information management. In the open data directive, thematic categories will be defined for valuable data, and owners should be named for these and, with that, the parties responsible for managing the PIDs of core data. If necessary, the authority can act in all the roles listed above. However, it is worth considering a sustainable and efficient implementation model and ensuring both interoperability and the entire life cycle of systems and identifiers.

Operators must be obliged to use existing identifiers, to apply PIDs also in their own services. This requires knowledge of the available PID identification systems and their usage possibilities.

Businesses should also be activated, on the one hand, to utilize PIDs to make their data management more efficient, and on the other hand, to open their own valuable information resources identified with PIDs. The European data spaces[9] created for the utilization of data

[9] Regulation (EU) 2021/694 of the European Parliament and of the Council of 29 April 2021 establishing the Digital Europe Program and repealing Decision (EU) 2015/2240. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L:2021:166:FULL&from=EN

materials increase the need for traceability, interoperability, and reliable referencing even on the commercial side.

It is important to describe the services offered by PIDs and their conditions, e.g., regarding metadata, and make the entire national field of actors available. It is essential to comprehensively describe the different identifiers already in use (DOI, Handle and different URN name ranges; operators' identifiers) and their management models and processes. The application of PIDs is guided by international practices such as Crossref's and DataCite's requirements for DOI distribution, as well as the standards behind the identifier systems. These limitations must be considered when thinking about the suitability of different systems for our circumstances. In Finland, it is not a good idea to introduce identifiers that are not suitable for the intended purpose of use. We must also avoid fading systems like the PURL tag, whose development is no longer active, and which are not based on any standard.

Identifiers of location data and identifiers of research actors (ISNI, ORCID and ROR) are, for example, URL addresses in practice. The persistence and functionality of operators' identifiers are based on global databases, where information about operators is stored centrally. The possibilities at the national level to influence how these codes are resolved are limited. Regarding ISNI identifiers, the National Library, as a Registration Agency organization, can create new identifiers, correct incorrect identifiers, and edit metadata related to identifiers. In the ORCID identifier system, the responsibility for maintaining the metadata rests with the researcher who created the identifier. The management of many identification systems in administration, such as Business ID (Y-tunnus), as PIDs, would be natural and beneficial for all stakeholders, considering the functionality, trustworthiness and persistence they offer.

National cooperation is important so that parallel PID systems can be systematically used together. The use of several parallel PIDs also enables preparation for future European Union requirements. Finland should agree on a uniform way of expressing different identification types and their relationship. However, Finland should not wait and prepare for the expanded use of PIDs.

Dependencies on international actors and systems are also a potential risk, which should be minimized, for example, by participating in international cooperation regarding persistent identifiers and, for example, mirroring critical PIDs in Finland. The careful and consistent use of PIDs must be promoted not only in Finland but also internationally. Guidelines related to resolution infrastructure must also be given. Who should own the resolution infrastructure? There are reasons why these should be under national control. Ownership of identifiers must also be considered when choosing PID services.

Persistent identifiers in research

This chapter describes the current state of the usage context of science and research, the identifiers in use, and the stakeholders.

Special features of research as context of use

In the field of science and research, the FAIR principles form an international, central guideline for the development of services and data infrastructure. The FAIR principles mean that various research materials, research methods and research publications must be discoverable, accessible, transferable, or combined, reusable, and reproducible. This requires the utilization and management of persistent identifiers. In addition, the material must also be able to be stored for a long time if necessary. Both the Finnish Research Council and EU research funders require consideration of the FAIR principles.[10] FAIR requires persistent identifiers to ensure research reproducibility, and their application should be as automated as possible.

Although the researcher is required to publish the results in accordance with the FAIR principles, he cannot practically do it alone, but needs the support and services of his framework organization. Universities, research institutes and other public administration organizations that use PID identifiers, on the other hand, receive support from national PID service providers. These organizations must refer their researchers to reliable PID services. Different levels of national and international services are available, such as Fairdata, Eudat and Zenodo services for research materials, scientific publishers' DOI services for scientific articles, for example, and Software Heritage for research software. These services make it possible to obtain a PID for research outputs.

A PID is given to the outputs published or planned to be published in research. During the research process, PIDs may already be needed within the organization and several organizations. Either internal identifiers are used for this, or, for example, data can be published as cumulative or dynamic materials. Common guidelines for the use of PIDs should be drawn up, for example, such that only one PID is given to an object, but different manifestations of the same content (versions with identical content published in different file formats) can be given their own PIDs.

International standards and best practices are developing rapidly. It is necessary to participate in this development work within the limits allowed by resources. Finland has participated, for example, in the development of the DOI and ISNI standards. The Research Data Alliance (RDA) also does a lot of work related to datasets and identifiers, but the mandate is unclear. DataCite has recently strengthened its position by expanding its data model and its coverage.

[10] EOSC strategy https://www.eosc.eu/sites/default/files/EOSC-SRIA-V1.0_15Feb2021.pdf, European Research Council guidelines https://erc.europa.eu/manage-your-project/open-science and Guidelines of the Academy of Finland https://www.aka.fi/en/research-funding/apply-for-funding/how-to-apply-for-funding/az-index-of-application-guidelines2/data-management-plan/data-management-plan/

With the help of identifiers used in research data, for example, organizations and financiers are able to follow the development of research. Information about the research is collected from different sources in databases, such as Research.fi [11] and OpenAIRE, which can be used to monitor the findability and effectiveness of the research. Nowadays, these data are used to produce entities of linked data where the information becomes even better structured and easier to find. Data graphs put data into context with the help of linking and semantic metadata and thus enable the integration, combination, analysis, visualization and sharing of data in an automated way.

In research, the documentation of the research process, file formats and reproducibility, as well as the management of rights and access can be important and at the same time particularly challenging.

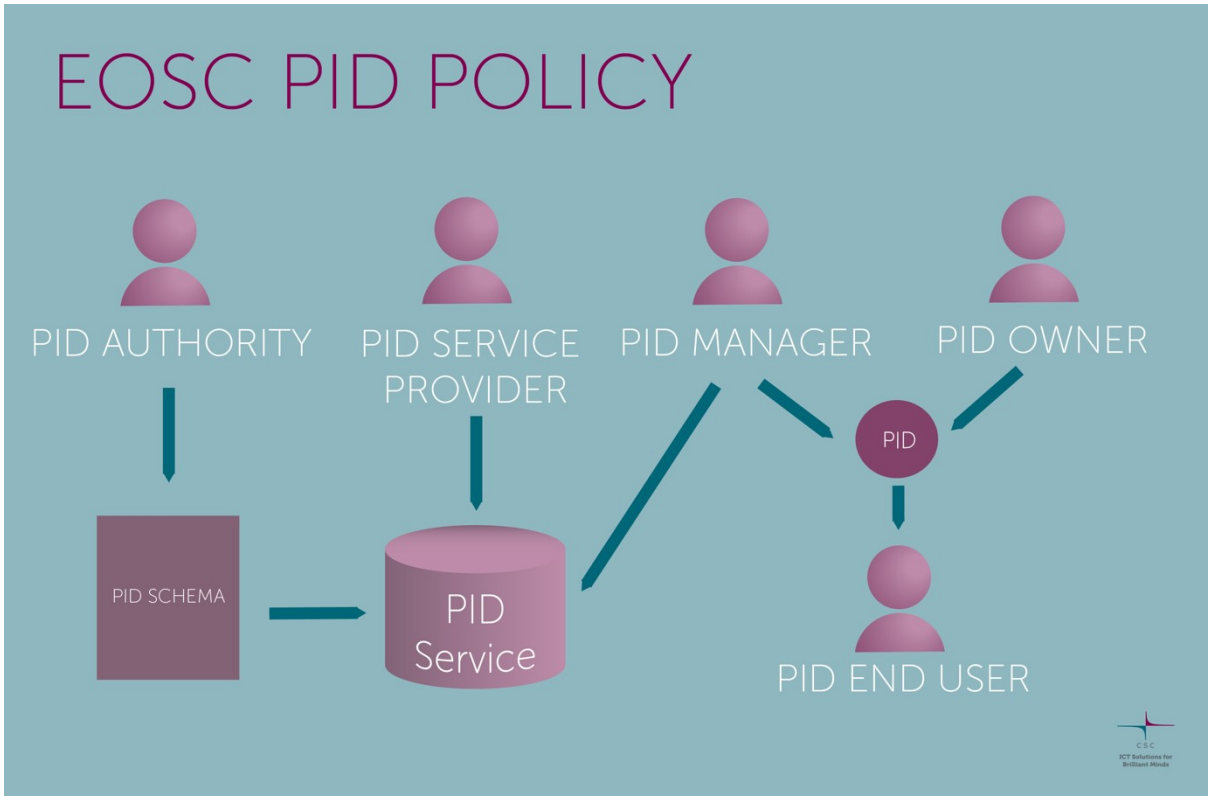
Actors and roles in the research context of use

On the research side, EOSC's PID policy[12] defines the following general operator roles:

- PID authority (responsible for maintaining the standard)
- PID service provider (responsible for PID resolver and namespace management)
- PID manager (responsible for managing core metadata)
- PID owner (creates and allocates the identifier, is responsible for maintaining the content of (meta)data
- PID user (end user)

[11] The PIDs identified by the research database (draft) can be found here: https://koodistot.suomi.fi/codescheme;registryCode=research;schemeCode=PID (18/03/2024)

[12] 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, 2020, https://data.europa.eu/doi/10.2777/525581



Rapid development is currently taking place in the field of PIDs also at the European level. In connection with the European Open Science Cloud projects, e.g., resolution service, PID policies are developed and a cooperation forum for PID service providers is formed.[13] In addition, interoperability between e.g. URNs and DOIs is being developed and the new RAiD identifier is being standardized.

In Finland, the coordination of cooperation is mainly dependent on the PID network and informal activities mainly by the National Library and CSC. If, in the creation of a common national infrastructure, funding came from more sectors, an actual PID center and a network could be created where several PID resolvers with decentralized maintenance would be financed. In this PID roadmap, the following national service providers in the research area have been identified:

[13] See in more detail https://fair-impact.eu/ and https://faircore4eosc.eu/ and https://www.eosc.eu/advisory-groups/pid-policy-implementation (18/03/2024)

Type	Definition	Service provider	Example coverage or use case
PID-service provider	An organization that provides PID services under a specific managed PID scheme. The service provider is responsible for providing a reliable service so that the identifiers issued by it remain intact and resolved. The PID service provider is responsible for ensuring that the service is scalable, interoperable and that new identifiers can be created.	National Library	Publications: URN, Actors: ISNI
		DataCite (CSC), CrossRef (Federation of Finnish Learned Societies)	DataCite DOI, CrossRef DOI: Research outputs like articles, papers or datasets
		ORCID	Researchers: ORCID
Repositories	The producers of the data services are responsible for the management of PIDs and the integrity of their objects, the correctness of metadata both in the resolver and elsewhere, as well as the correct usage margin and the way of use according to the PID scheme. Typically, the data service provider acquires identification services from a suitable PID service provider. The costs should not be passed on to end users.	National Library, Federation of Finnish Learned Societies (TSV)	Publications
		Higher education institutions, Research institutes	Research outputs in the organization's own services
		The Language Bank of Finland, Fairdata.fi services, Finnish Social Science Data Archive (FSD), Finnish Biodiversity Information Facility (Laji.fi)	Research materials available by discipline and in joint services
		Archives and museums	Cultural heritage materials
		Research Information hub	Research infrastructures, funding decisions

PID authorities

PID authority refers to the entity that owns the PID scheme and directs its management and development. Such entities include, for example, DONA (DOI) and IANA (URN), which,

among others, manage namespaces so that they are unique and can define their coverage. In some cases, PID service providers rely on the DNS system and then create their own schema, as for instance ORCID has done.

Consortia

There are a few PID consortia in Finland, which aim to facilitate and support the utilization of PID services by organizations in practice.

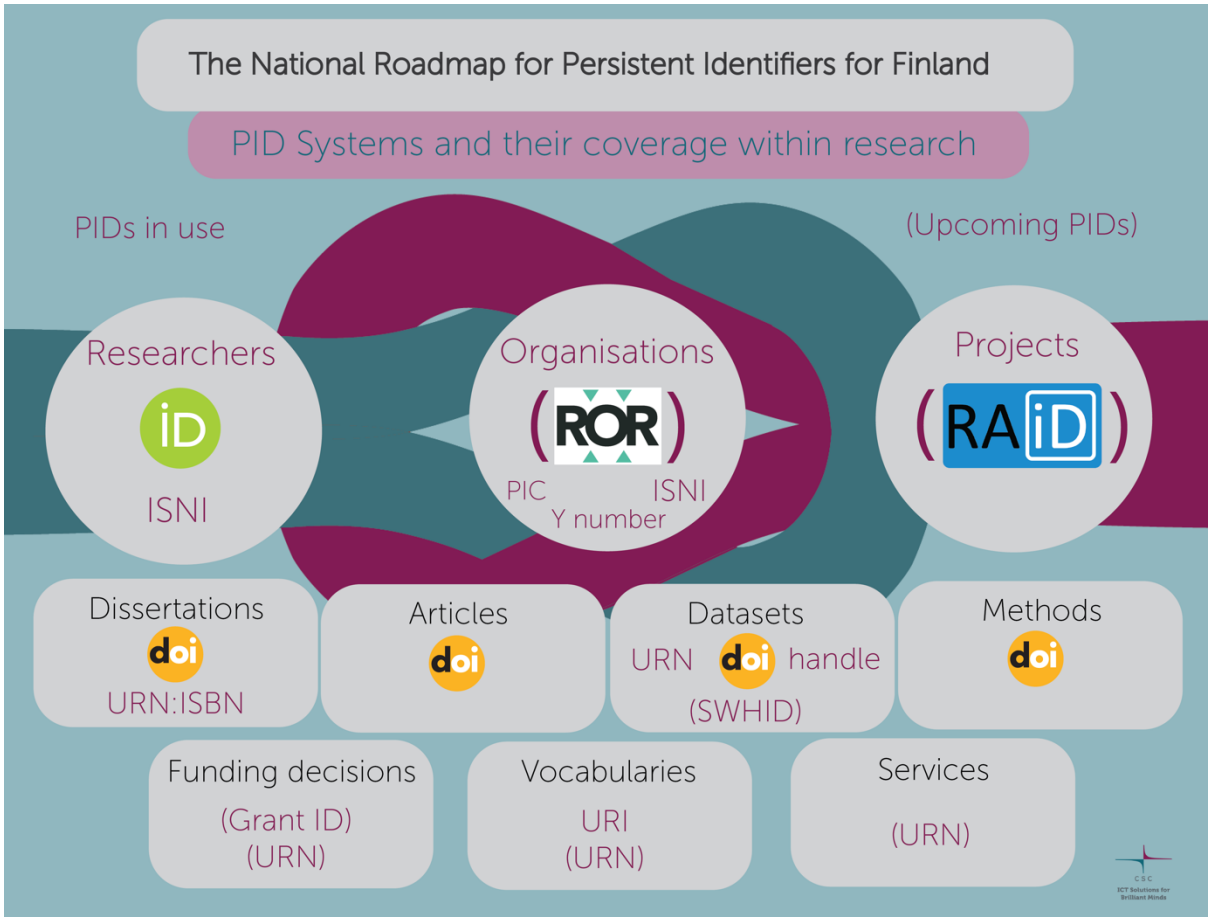
Organizations can use the PID service (DOI) suitable for DataCite's research data by becoming members of DataCite directly or through the DataCite Finland consortium managed by CSC. Membership allows you to use DataCite's Fabrica system and the DataCite REST API. In 2022, this consortium will have 6 members. CSC is also a member of the ePIC consortium, and handle identifiers can be obtained, for example, through this international consortium.

In the same way, even the national ORCID consortium does not directly manage the ORCID identifiers intended for researchers or their use. It takes care of the ORCID memberships of Finnish research organizations, which gives the organizations the right to use ORCID's most comprehensive two-way interface. Currently, this Finnish consortium managed by CSC is mainly an administrative actor and has 32 members.

Other

The PID Forum Finland network has been operating as an informal network for several years. The network has compiled information on joint wiki pages, and know-how has been shared in regular meetings, usually a couple of times a year. It would be good to strengthen the network's position as an expert network, but the promotion of interoperability and coordination is currently still an informal and voluntary activity alongside one's own activities. To achieve strong enough coordination, especially at the boundaries of the development of research data management, the need for national guidance is clear. The guidance should be at the alignment level and consider use cases and needs so that the implementations are successful. This requires investment, for example, in mapping the aforementioned.

PID Systems and their coverage within research



Researchers

The use of ORCID can be recommended and the ISNI identifier replaces the gaps in ORCID, i.e., deceased, and reluctant researchers who cannot/do not want to apply for an ID.

Dissertations

The preservation of theses is regulated by law, but the instructions of the Ministry of Education and Culture only apply to the file format (PDF/A) of the permanently stored material, not the PID code or other metadata. The (pdf) version of the dissertation published online will receive an ISBN and a URN based on it, see e.g. https://urn.fi/URN:ISBN:978-951-29-9043-6. In Finland, it is not customary to give dissertations a DOI identifier. If a dissertation is given a Crossref DOI, it is a work identifier that covers all current and also future manifestations of the dissertation (versions with identical content published in different file formats). Crossref's DOI is then the work level ID, URN:ISBN is the ID of the PDF and EPUB versions of the publication and, for example, URN:NBN can also be given as the metadata ID of the publication. Several PID systems may

therefore be needed to identify different parts or manifestations of one publication[14]. According to the rules of the ISBN system, the URN:ISBN given to the dissertation is specific to the manifestation, but if a single manifestation of the dissertation consists of several files, they all have the same URN:ISBN, but each file must also have its own URN:NBN, i.e. the national bibliographies ID code, to enable long-term preservation.

Scientific articles

Among other things, Federation of Finnish Learned Societies offers Crossref DOI codes for articles from scientific journals that are published in their services (Journal.fi for articles and Edition.fi for monographs). A few other actors (at least University of Helsinki, University of Jyväskylä, SKS - Finnish Literature Society) have independently acquired DOI IDs.

The National Library's URN ISBN codes are given to books and URN:NBN codes are given to all kinds of objects. Among the external actors, these can be used by those who have an agreement with the National Library. There are still some gaps in the system:

- The DOI ID required by the old journal article, in which case the administrator must enter into an agreement directly with CrossRef.
- ID required for a new domestic magazine article, if it is not published in the Journal.fi service.
- If there are two versions of the article in different file formats, the DOI is shared by them, and the articles cannot be transferred to the preservation service.

Research data

Research data is data that is used in research or is created as a product of it. Any object - map, video, database - can be research material. Data produced in research are published in many different services, of which Fairdata services and several others use DataCite's DOI codes to identify the data. Handle and URN:NBN identifiers are also used alongside them. There is no specially developed identification system for dynamic research data.

It is possible to refer to distributed and/or accumulated data by giving an identifier to both the data source and the search targeting it. For example, The Finnish Biodiversity Information Facility (FinBIF) offers a DOI for search results, which makes it possible to refer to the obtained search result, even if the same search at a later time produces a different result. The DOI given to the search result also enables referring to several data sources with one identifier, and at the same time referring to the original sources. In the Fairdata service, it is possible to create an accumulating data set, to which it is possible to add data without changing or deleting existing identifiers, and without breaking the material. The long-term preservation service, which is one part of the Fairdata services, also uses DOI identifiers.

Regarding DOI identifiers, we are dependent on the agreed international agreements on the use of identifiers. For the others (Handle and URN:NBN), use is more free. In Finland, CSC

[14] The 'Identification of electronic publications' guide (in Finnish) tells how PIDs can be used to identify electronic publications. https://urn.fi/URN:ISBN:978-951-51-7661-5

offers DataCite DOI codes for small operators. These identifiers are mainly intended for research data, but they can also be applied to data stored in publication archives.
You can get a DOI ID for all material that you publish on the Zenodo service. The reliability of this metadata depends on whether the storer itself maintains it. Therefore, Zenodo cannot be considered a viable alternative to, for example, the Federation of Finnish Learned Societies' and CSC's DOI identifiers.

There is probably a need for organizations' own data archives, where, however, managing data integrity and curating data are demanding activities. Issuing PIDs requires considerable investment in the organization of information management. Joining DataCite is possible through the Finnish consortium or directly but requires sufficient own services. Gaps at the moment include the following:

- There are two versions of the same data material with different file formats, these cannot have their own identifiers. If different file formats are put into long-term preservation, they should have their own identifiers.
- Identification policies and services for databases and other complex data require development

Source code, workflows and other methods

In the research, applications and various models and other code are also created and utilized. The documentation and reference ability of these are often a prerequisite to ensure the reproducibility of the research.

Currently, the code may be published in Zenodo with the help of the GitHub integration in the service (DataCite DOI) or in the Software Heritage archive (https://archive.softwareheritage.org/), which uses its own internal identifier, which is, however, strongly marketed in e.g. EOSC. Ensuring reusability requires special care from the researcher regarding documentation and dependencies.

Workflows are still a relatively uncharted margin, but for the sake of data quality and research reproducibility, their development would be important. Referring to research methods requires that they have their own repository, where they are managed, described and their integrity is taken care of.

Other research methods can be published in various services, but service offerings and practices are often still incomplete.

Vocabularies, variables and code sets

Especially regarding research data, parallel, copied code sets should be referenced in a machine-readable manner to the correct version of the source code set. The goal would be to have one maintained master version of the code sets, in which you can create additional local objects or your own views. A separate recommendation has been made for the ontologies of research data, which is also recommended to be followed for subject

vocabularies.[15] Standard description of the variables and maintenance of the descriptions would be of primary importance, but this should at least mainly be done at the international level.

Organizations

Organizations also have PIDs. Depending on the context of use, these include, for example, RoR, PIC, ISNI, Y-identifier and CrossRef Funders. Context means that as a recipient of funding, the organization may have a different identifier (PIC), than as a party to the agreement (Y identifier), than as a grantor of funding (CrossRef_Funders). The link can be found, for example, in Wikidata (QID). A special feature of organizations is their changes, which require the management of relationships, preferably with the help of a machine-readable ontology. In Finland, such an ontology for national operators is KANTO - National operator information (https://finto.fi/finaf/en/). Finnish National Agency for Education (OPH) maintains the organizational information of Finnish public actors in education and research with their subunits in the Organization Service[16].

Funding decisions

Funding decisions in the Research Information Hub.

At the end of 2022, the research data reserve contains information on both national and EU funding decisions.

European Union funding decisions

Information about the EU's decisions can be found on the funding decisions of the H2020 Framework Program, the Regional Development Fund (fi:EAKR/en:RDF) and the Social Fund (en:ESF). The decisions of the previous framework programs have not been imported so far, and the uploading of the decisions of the Horizon Europe framework program that started in 2021 is in progress. For a long time, the primary identification of EU funding decisions has been the projects' funding decision number, which is the EU's local identifier. Formerly RCN, now Grant ID is a different matter. This local identifier is standardized to the extent that you can land on the www page of all projects in Cordis with a standard URL, which is a fixed initial part + grant id (eg: https://cordis.europa.eu/project/id/101057264). However, during 2022, DOI identifiers have started to appear in Cordis for funding decisions, (e.g., https://doi.org/10.3030/101057264) which apparently use a standardized format, where the end of the DOI is the EU's Grant ID. The research database is preparing to download these new identifiers.

National funding decisions

We are not aware of any national research funders managing PIDs for their own decisions. Everyone uses their own local identifiers and is content with using these. This has already

[15] Yann Le Franc, Luiz Bonino, Hanna Koivula, Jessica Parland-von Essen, & Robert Pergl. (2022). D2.8 FAIR Semantics Recommendations Third Iteration (V1.0). Zenodo. https://doi.org/10.5281/zenodo.6675295

[16] Service package of the Board of Education. Service card: Organizational services. (18.3.2024, in Finnish) https://wiki.eduuni.fi/x/AolcCw

caused conflicts when, for example, the local ID of the Research Council of Finland has been identical to the ID of another funder.

A report recently published in Knowledge Exchange on the role of research funders in the PID field [17] highlights the most obvious benefits of a common, global, new ID. Here, too, it is stated that when the funder manages its own identification systems, there is always a risk that identifications are duplicated between different financiers. Human errors also easily occur when authors refer to their works, which means that the identification structure of funding should be as simple as possible and support machine readability. In addition, a uniform identification ID of the funders would enable the utilization of the PID Graph, which would also enable all publications and datasets to be linked directly to research projects in a machine-readable manner. This unity also affects funders in a favorable way when they can easily follow what has been achieved in the projects, they finance from the landing pages of the projects.

Research projects

RAiD[18] (Research Activity identifier) is most likely the future identifier for research projects. Technically, RAiDs are DOI, Handle or Cool URI identifiers, so it is recommended to use the DOI format RAiD, which will become an EOSC service in the next few years.

Services and infrastructures

In the Research Information Hub, URN identifiers are created for infrastructures. At least not yet identifiers are created for the services, which would, however, facilitate the utilization of their metadata in, for example, service catalogs or material management planning. In addition, identifiers for devices are coming into use internationally.[19]

[17] de Castro, Pablo, Herb, Ulrich, Rothfritz, Laura, & Schöpfel, Joachim. (2022). The role of research funders in the consolidation of the PID landscape. Zenodo. https://doi.org/10.5281/zenodo.7258210

[18] RAiD Research Activity Identifier Service. (18/03/2024) https://ardc.edu.au/services/ardc-identifier-services/raid-research-activity-identifier-service

[19] PIDINST White paper. Persistent Identification of Instruments (PIDINST) working group output report https://docs.pidinst.org/en/latest/white-paper/index.html

The target state and the necessary steps to achieve it

This chapter describes the necessary measures to develop the use and management of persistent identifiers in the context of science and research.

PID Forum Finland started this roadmap work 2021 by producing a description of the common target state, which is as follows:

Traceable and distinct information (about objects) can be found
and can be reliably linked now and in the future

Next, the steps that this expert group deems necessary to reach the target state are described concretely.

Target	Description	Actions

Building trust and cooperation	We get to know stakeholders, share information and strengthen our network	• we build cooperation between operators in relation to PID matters, e.g.:
		  ○ The research communities
		  ○ PID services
		  ○ LAM sector
		  ○ Digital and Population Data Services Agency, Ministry of Education, Ministry of Finance
		  ○ The linked data community
		• we work related to the quality of information
		• A PID policy is produced
		• we clarify needs and benefits

Organizing the national infrastructure	We organize, plan and secure funding	• we develop a platform for PID work • we identify development needs • we agree on responsibilities • we find out financing • we increase awareness and competence
Growing competence	We take care of information sharing and skills development	• we raise awareness about PIDs, e.g. through the Open Science training group agenda • we systematically develop international networking and activities

Building trust and cooperation

Finding out the state of data management and interoperability at the national level and drawing up a PID policy together would be a good basis for creating cooperation and joint policies. Research cannot do the entire field of PIDs alone, but a broader framework is

needed, for example with the state administration, so that interoperability, usability, and efficiency can be guaranteed as part data management.

When building trust among the PID providers and between them and the entities that use and manage PIDs, agreeing on common rules of the game is at the centre. They guarantee that all parties have the same understanding of the purpose of use of each identification service and its maintenance. Since we operate in an international arena, an international perspective such as the EOSC PID Policy must also be considered when drafting the shared rules.

The goal should be to build a way of working where you can act in a self-organizing manner following common guidelines. It is necessary to describe together what kind of IT services (i.e. systems) that are used by everyone are needed. Regarding the Ministry of Finance and the national Information management board, we also hope for guidance and identification of persistent identifiers as part of the knowledge management map and overall architecture work. The predictability and control of operations can also be organized through formal agreements. For example, the National Library is currently [2023] preparing a URN agreement that formalizes the use and management of URN identifiers.

Things that increase trust

Trust is built between different actors: between the researcher and the service provider, between the service provider and the producer of the PID service. National PID guidelines support trust between service providers and PID service producers, which also affects the relationship between researcher and service provider.

The trust between the end user and the service provider is based on the following factors:

- Reputation of the service and its provider
- Service documentation and transparency
- Quality of the content provided by the service (is there a statistical service, metadata model, visibility of content)

Trust between the service provider and the PID service producer

- Documentation
- Level of service (e.g., support, is it always available)
- Reliability of the service (does the promise match the received service)
- Contracts
- Reliability of operators
- Audits and certificates

Trust in a functioning PID infrastructure nationally

- Clear operating principles and instructions
- Clear responsibilities
- Permanent funding
- Demonstrating the benefits of PIDs

Necessary actions

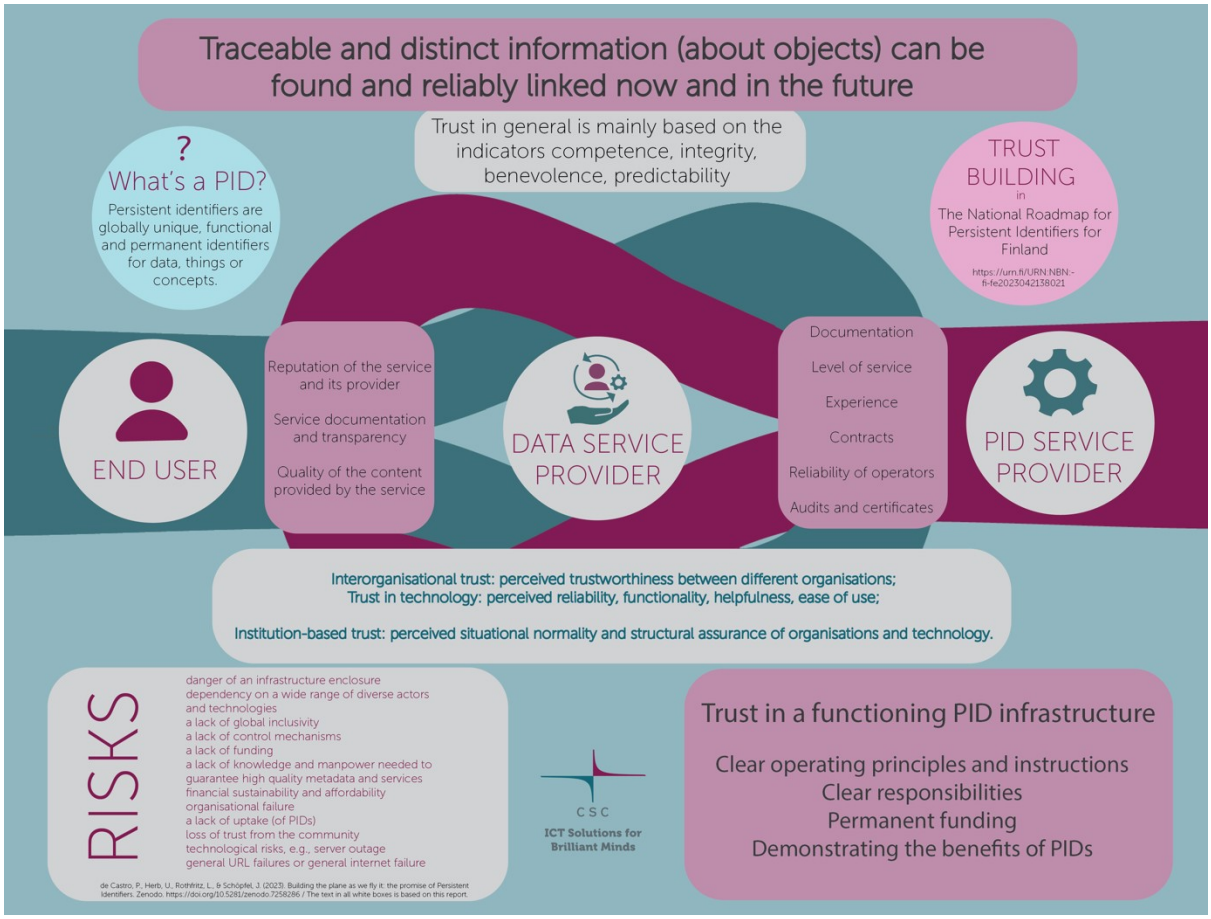
1) Establish cooperation and position it in relation to other actors (especially Ministry of Education and Culture, National Library, National Archives, Digital and Population Data Services Agency, and other ministries and actors that produce research data, e.g., Ministry of Social Affairs and Health, Finnish Social and Health Data Permit Authority Findata, Tulanet (cooperation body of Finnish government research institutes) and the Linked data community)

2) A new, more specific PID policy should be produced
   a) The PID network must prepare and maintain recommendations on which identifiers are used for different objects and in relation to usage contexts.
   b) Different actors and their roles are taken into account in the policy
   c) Let's agree nationally on which criteria PID schemes and services are selected.
   d) Let's nationally agree on how to fix the identified deficiencies in the PID offering

3) The necessary other documents are identified and, if necessary, created: other necessary PID policies, contracts, certificates

4) Ensuring the transparency of operations - from Knowledge Exchange's report, matters that promote trust
5) Demonstrate the benefits of using PIDs with use cases, effectiveness evaluation (cf. JISC, ARDC, RDA report)

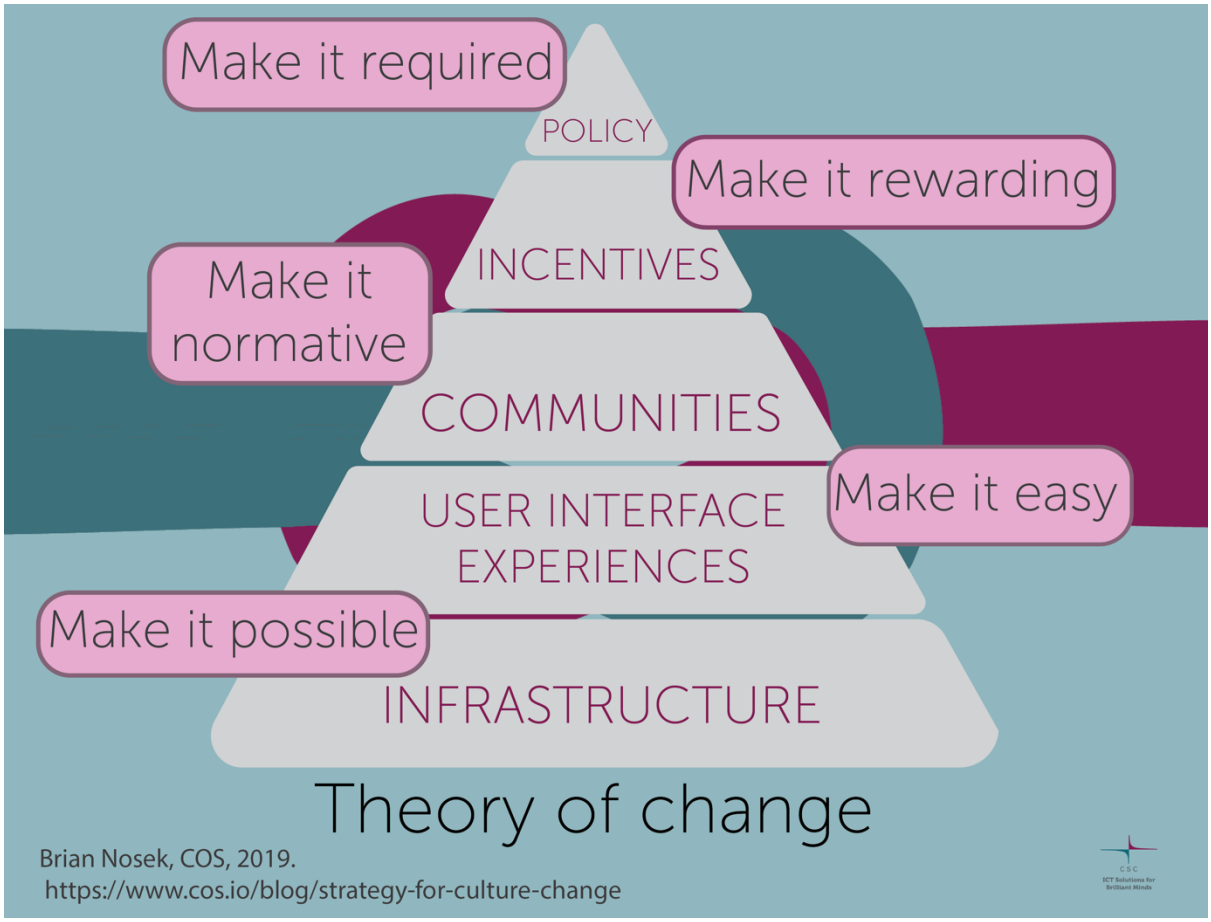
Cautionary examples should also be considered and presented.



Organizing the national infrastructure

The Finnish PID infrastructure has been formed organically based on needs and natural responsibilities, and artificial centralization is not necessary. The focus should be that the goals listed above are realized and that all identification needs are covered with sufficient resources. The infrastructure must be sustainable. What is important is interoperability and reliability, as well as operational cooperation. Management should be model-oriented and not actor-oriented.

Sufficient funding enables investment in services, which in turn enable a change in the operating culture and, with it, the creation of savings through the improvement and efficiency of information management.



With sufficient funding, it is possible to:

- secure the reliable operation of services and the resilience of systems with sufficient transparency and diversity
- ensures sustainable development (reduce overlaps, good data management, electricity is expensive)
- support organisations to identify their core data and provide it via interfaces with PIDs for use by others
- to include PIDs as part of general data management (data models and interoperability)
- maintains the linking of once-issued identifiers to newer systems (if parallel systems are created in the future)
- ensures international interoperability and contribution to international systems, standards and services

PIDs are often provided by a party other than the one responsible for metadata or content quality and integrity. In target mode, PID systems are an integral part of other data management. From the end user's point of view, PIDs are objectified 'automatically' in different services such as research infrastructures or data archives. The necessary PIDs will be defined through the actions of the end user, end users want easy, automated solutions. This means that the owners of the data or the providers of the material services are responsible for the selection and implementation of PIDs.

There are several PID systems available, which may be produced by private operators or produced by national organizations. In Finland, there is probably no need to start producing new PID systems for the needs of research, but we can choose from the available services. The aim is to use different PID systems for different objects, so there will be several systems.
The creation of national recommendations for the selection of systems can be done as part of joint guidelines. At the same time, justified policies are created as the basis of such a set of criteria. An example of the policy could be that when choosing a PID system, care must be taken to ensure the permanence of PIDs, i.e., the PID system and related PID service providers should be committed and prepared for changes, and one criterion could be e.g. an ISO 27001 certificate to ensure the organization of service production. Such criteria are being created in EOSC projects.

Recommendations help stakeholders choose suitable solutions. Identification systems are also part of the international infrastructure and need both administration and services. Some of the PID systems are strongly controlled by foreign organizations (e.g., ORCID). In this case, standardization is particularly important, as it brings transparency and predictability to operations. Some of the identifiers (e.g., CrossRef DOI) are, on the other hand, strongly held by commercial operators. On the public administration side, of course, European information management control.

At the European level, there is scope for the management of PID systems and the creation of joint guidelines. The guidelines will also affect the organization of infrastructure at the national level. The eArchiving Common Services Platform (EARK-CSP) is produced by the so-called synergy entities in the direction of the Digital Europe project. One work item is to support the utilization of PIDs and produce general recommendations on applicable standards and practices.

Regarding DOI, management is three-tiered: International DOI Foundation, Registration Agency (Crossref / DataCite) and national contracting party (CSC – IT Center for Science / Federation of Finnish Learned Societies). The most loosely managed are ARK, Handle and a few URN namespaces such as URN:UUID which do not have a centralized control organization, and users of the token are bound only by the standard and, in the case of ARK and Handle, the resolver applications.

International developments will probably clarify the situation in the next few years and thus facilitate the selection and integration or application of PID solutions to different systems.

Necessary actions

1. A common platform should be developed to maintain the PID policy and other information related to PIDs.
2. Let's map the international context, monitor the development and act proactively and coordinated also internationally
3. Let's find out how gaps and shortcomings can be solved.
4. Let's agree nationally how the services are financed and organized in the context of the use of the research for different covers.

Increasing awareness and competence

A common understanding of the current state of the PID field of activity and the related goals is a prerequisite for effective communication and raising awareness. Building cooperation has been discussed above, but communication of the results and benefits achieved in collaboration is central. Awareness of PIDs should be increased in a way that is meaningful to each target group. Successful communication clarifies policies and shows benefits for each stakeholder group.

It is good for every researcher to know which PID identifier can and should be used at any given time and where it can be received, but it would be best if everything happened automatically as part of the research process and by the services. The services should be comprehensive so that no object that needs a PID is left out without it. The processes of minting and allocating PIDs should be automated as far as possible and integrated into user interfaces and systems in such a way that their use does not hinder the utilization of services and that the information is linked in an optimal way. The use of PID systems selected in data management services should be instructed in a very user-oriented manner. When are they useful and when are they necessary?

The importance of PIDs for long-term preservation and national heritage must be opened to public administration actors, so that state actors understand why persistent identifiers should be financed and managed. The same also applies to representatives of the researcher's own organization.

Now, all the information needed for the implementation of PIDs has not been gathered in one place. Because of this, it is difficult to form an overall understanding. Basic information about PIDs can be found in the Electronic Publications Identification Guide published by the National Library, but identifier-specific user guides have not yet been prepared.

In addition to raising awareness, training is needed, which should be based on identified competence needs. The identification of competence needs in the following sectors should be the first steps to promote competence: service providers and consortia, organizations and support services, and researchers. Generally, at the national level, interoperability is developed at the Digital and Population Data Services Agency.

Increasing competence is best done by teaching through examples. Bringing up case studies in the training material is therefore particularly important. Knowledge and awareness of PIDs also accumulates when you must search and utilize open data from different sources, and at the same time you can easily observe the value of using PIDs. The development of the international field must also be closely followed, especially in terms of standards and technical solutions, and common practices must be agreed together across national borders. Research data support is key and the personnel should be trained so that they know how to recommend the right solutions to researchers. The researchers rarely choose the identifier themselves because they primarily choose the service. The aspects of the different identifiers, as well as the related benefits and possible challenges, should still be explained to them.

Extensive knowledge of data management is needed within the scientific fields, part of which is the knowledge of persistent identifiers. Training on PIDs is therefore also on the shoulders of home organizations, but information about the PID systems in use must be made easily available to trainers as well.

Necessary actions

1) In addition to/on top of the wiki site, a more common landing site will be created, for which user manuals and other up-to-date information on persistent identifiers will be collected. On this site about PIDs, the information should be organized according to the PID system according and to the reader's role, and the pages should have defined owners who keep them up to date.
2) Let's organize training. The following competence needs have been identified:
    a) actors and consortia
    b) knowledge of identification systems
    c) knowledge of the available PID services
    d) national and, where necessary, international services
3) The target groups of the training are at least:
    a) Organizations, services, and support services:
        i. Use cases for persistent identifiers
        ii. Persistent identifiers
        iii. PID services
    b) Researchers:
        i. ORCID, what it is and how it works
        ii. referencing, not only data but also other things
        iii. choosing services
4) The Digital and Population Data Services Agency should include PIDs as part of its educational activities

Guidelines should be produced for researchers regarding the different stages of the research process and how PIDs are used, created, and maintained. For this, a working group within the scope of Open Science and Research Coordination can be established.