KE Scoping Document[1] (Version 5.5)

Risks and Trust in Pursuit of a Well-functioning Persistent Identifier Infrastructure for Research

[1] This document is intended:
- To serve as KE internal consensus builder for activity raison d'être.
- To outline and structure the problem, justifying the activity.
- To scope the activity tasked to an Expert Consultant.

Contents

ABSTRACT AND KEYWORDS 3

KEYWORDS 3

INTRODUCTION 3

THE THIRD PILLAR OF SCIENTIFIC INVESTIGATION 3

THE HOURGLASS – PID PROVIDERS AND USER 3

PID AT THE HOURGLASS WAIST 4

THE STRATEGIC IMPACT OF NOT HAVING A WELL-FUNCTIONING PID INFRASTRUCTURE 4

A WELL-FUNCTIONING PID INFRASTRUCTURE 4

VISION AND AMBITION 4

VISION 5

AMBITION 5

THE PID LANDSCAPE - ECOSYSTEMS AND STAKEHOLDER ROLES 5

THE PID ECOSYSTEMS 5

Intrinsic Identifiers 5

Self-Sovereign Identity and DIDs 6

STAKEHOLDERS AND ROLES IN THE PID ECOSYSTEM 6

PID Authority 6

PID Service Provider 7

PID Manager 7

PID Owner 7

PID End-User 7

METHOD, STEERING, BUDGET AND TIME PLAN 7

THE CONCEPTS OF RISK AND TRUST IN A PID CONTEXT 7

CRITERIA FOR ASSESSING RISK AND TRUST 8

PID Authority 8

PID Service Provider 8

PID Manager 8

PID Owner 9

PID End User 9

METHOD – USING THE OPEN SCHOLARSHIP FRAMEWORK 9

PROJECT STEERING 10

BUDGET AND TIME PLAN 10

INVESTIGATION OF KE PID CASES STORIES 10

THEMES OF INTEREST **Error! Bookmark not defined.**

INTRODUCING THE KE MEMBER NATIONAL PID LANDSCAPE OF PROVIDERS AND USERS 11

Finland (CSC) 11

France (CNRS) 11

Denmark (DeiC) 13

Germany (DFG) 13

United Kingdom (Jisc) 14

Netherlands (SURF) 15

SUMMARY OF KE PID CASES **15**

RISK 15

TRUST 16

CONCLUDING PID RISK AND TRUST ISSUES 16

RECOMMENDATIONS 16

1 Abstract and Keywords

1.1 Abstract

Persistent Identifiers (PIDs) and their infrastructures are argued to be of significant strategic importance, to the increasingly digital reality of modern-day research.

With this investigation, we aim to better understand what is needed to build and exploit a well-functioning PID infrastructure for research. Our ambition is to identify, through investigation, analysis and recommendations, what could be the best possible strategic and operational paths to achieve a well-functioning PID infrastructure for Knowledge Exchange (KE) member states and beyond.

This scoping document serves as a starting point, by providing an overview of KE current PID ecosystems, focused on pinpointing issues that need to be addressed through the investigation. We propose to map these issues with KE's *Open Scholarship Framework*, aiding identification and structuring of potential solutions.

1.2 Keywords

Persistent identifiers; PID; Research organisations; Europe; Risk assessment; Trust-building; Open Science infrastructure; Open Scholarship Framework; Knowledge Exchange

2 Introduction

The age-old interchange of scientific theory and empirical evidence is still thriving as these remain the traditional two pillars of scientific investigation. Transparency in methods and scientific output, identification and referencing are all part of what constitute modern-day Open Science and Open Scholarship. However, with the advent of computing and its impact on the scientific method, as well as an ever-growing corpus of digital data, the situation is evolving.

For some years now, many have argued that computing and digital data have evolved to become the third pillar of scientific investigation.

We argue that reliable pointers to various digital objects, i.e., Persistent Identifiers (PIDs), are of paramount importance to said change. Of equal importance is the assumption that PIDs can only work in an efficient way if they point to repositories where digital objects are archived and preserved adequately in the long term. In this paper, we define an identifier as a sequence of characters that uniquely denotes a referent. This sequence is deemed persistent when the identifier, its binding to the referent and the related metadata survives over time and technical evolutions.

We aim to investigate what the main risks are when pursuing a well-functioning PID infrastructure for research. We want to better understand the most important elements of trust in creating said infrastructure.

2.1 The Third Pillar of Scientific Investigation

This *Third Pillar* is about how computer hardware, software and digital data together make data analytics, modelling and simulation possible on the larger and global scale. Recently the need has become apparent to seriously develop better ways in which we handle digital data. Consequently, professional language has evolved, aiming at defining elements of and developing that third pillar. Concepts like *Data Management*, *Data Stewardship* and the *FAIR Data Principles* (as a way to improve the ***F***indability, ***A***ccessibility, ***I***nteroperability, and ***R***euse of digital data) have all gained traction, as important means to further progress science, which has become reliant on infrastructure and is truly global. Furthermore, it is widely recognised that data needs to be machine accessible and understandable. Only then can the full potential of the globally scattered data and computer resources be tapped, in numerous likewise scattered and diverse ways, for data analytics, modelling and simulation.

To make this happen it has been suggested that an initial and very minimal set of community-agreed guiding principles and practices are needed.

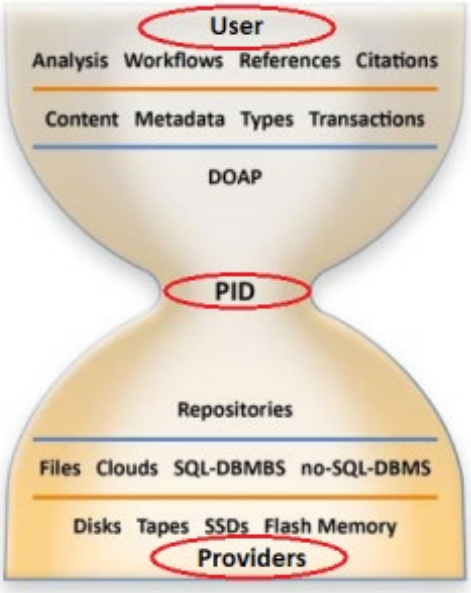
2.2 The Hourglass – PID Providers and User

In scoping a minimal set of community-agreed guiding principles and practices, it is argued that PIDs constitutes the waist in an hourglass model, with providers and users at the respective ends:

2.3 PID at the hourglass Waist

The importance of PID infrastructure, is argued as the need to remedy a single point of failure, or widen a bottleneck:

> "*These simple principles and practices should enable a broad range of integrative and exploratory behaviours and support a wide range of technology choices and implementations, just as the Internet Protocol (IP) provided a minimal layer - **the "waist" of an hourglass** - that enabled the creation of a vast array of data provision, consumption, and visualization tools on the Internet*".



It is thereby argued that the ability to uniquely identify any digital object constitutes the single most critical point – i.e., the waist of the hourglass. Therefore, Persistent Identifiers (PIDs) are introduced as a crucial infrastructure concept. A well-functioning PID Infrastructure has become one of the most essential generic scientific digital infrastructure elements to pursue.

2.4 The strategic impact of *not* having a well-functioning PID infrastructure

As a consequence of the evolution of the third pillar of scientific investigation, many items directly related to the research process itself now should and can bear globally accepted ***identifiers***. The items are manyfold e.g., publications, citations, conferences, researchers, organizations, clinical trials, patents, datasets, data repositories, grants, projects, samples, funding organisations.

Building on existing PID infrastructures and notably on DOI as an identifier spanning across categories of referents and contexts, various companies and publishers have developed cohorts of services for research organisations and agencies, e.g., journal impact assessment tools, scientometric studies, web impact of scholarly articles, whose added value derives from big data processing by artificial intelligence. These big data services aim to become indispensable in scientific workflows and in informing all sorts of decisions made by research institutions for e.g., recruitment, research funding, and the dissemination of results in a globally competitive environment.

It is crucial for research organizations to understand the challenges posed by the use of PIDs to turn research outputs into FAIR items without favouring any particular region or system.

We therefore conclude that a well-functioning PID infrastructure must be built and that potential mismanagement in bringing about such infrastructure will be a big loss to science.

2.5 A well-functioning PID infrastructure

By *well-functioning* we mean that it is:

- **technically** user-friendly and capable of uniquely and persistently identifying any digital object, deemed worthy of preservation.
- globally **accepted** (interoperable in its core design and technology) such that it independently of technology and geography always points to the data owners account and related metadata (i.e., resolves to an explanatory landing page), if not also the actual data.
- **organisationally** and **economically sustainable**, i.e., that the PID can still resolve even in the case of organisational change or economic turmoil - in principle for ever.
- **politically** trustworthy – in that there is minimal risk of sudden non-interoperability, legal obstacles or exploitive vendor lock-in.

We note that the above elements are necessary for any identification system for research output of any kind. Hence our PID scope covers all resolving systems, i.e., all types of PIDs.

3 Vision and Ambition

The PID concept has become crucial for achieving a globally agreed way of organising data, such that humans and machines can *find* any preserved digital object, anywhere, by any means at any time; and if deemed worthwhile, ethics and legal issues permitting, also to *process* it scientifically and computationally, according to the FAIR data principles.

3.1 Vision

Our vision is:

> ***A well-functioning PID infrastructure for research encapsulates and implements the PID concept globally across all scientific areas.***

We aim not only to investigate what it takes to reach such a state-of-affairs for PIDs, but also how one might go about it so that it is adopted and becomes ubiquitous. While the target beneficiary is KE membership, we expect outputs from this activity would also be relevant to broader concerns of open infrastructure.

3.2 Ambition

Our ambition is to:

> ***Identify, through investigation, analysis and conclusion, what might be the best possible strategic and operational recommendations to achieve a well-functioning PID infrastructure for KE member states and beyond.***

In so doing, an understanding of especially the two concepts of "**risk**" and "**trust**" are believed to be fundamental to the de facto adoption of a well-functioning PID infrastructure.

The study outcome could hopefully be of general use to the wider community of actors that work on PID implementations - globally and/or locally.

4 The PID landscape - Ecosystems and Stakeholder Roles

4.1 The PID ecosystems

Traditionally PIDs are considered to be identifiers that are:

- unique, i.e., never used again to represent another object,
- functional, i.e., there are services attached to the identifier,
- persistent, i.e., the same object is never given a new identifier within the same system.

It should be noted that the meaning of "Persistence" of PIDs is a grey area: Some stakeholders expect content behind PIDs to be immutable on a bit-level, in practice such a strong requirement creates new problems in PID Management. There is a need for clear policies on what is considered persistent.

4.1.1 Well-established Identifiers

A number of identifiers are already being used by research organisations worldwide, e.g., ARKs, DOIs, Handles, URNs that resolve to data and publications, ISNIs, ORCIDs that point to valid information about individuals[2].

Each of these identifiers have their own governance system usually managed by centralised registration authorities.

While established Persistent Identifier Systems have not yet solved all issues around Risk and Trust, new systems are emerging, i.e., Intrinsic Identifiers, which identify objects based on their bit-level content. The concept of Self-Sovereign Identity attempts to decentralize the minting and handling of identifiers by registration authorities and PID providers, while at the same time introducing mechanisms to build trust in a decentralized approach. Both of these new systems are detailed below.

4.1.2 Intrinsic Identifiers

As already mentioned, it is important to note that various types of PIDs do exist. Some are managed by PID Service Providers, others can be managed directly by the PID Owners. This is notably the case for cryptographically strong hashes that can compute from any file a short "signature" thus providing an intrinsic identifier for a given digital object. These digital objects, their intrinsic identifiers and associated metadata can be managed by the PID Owner and transferred via blockchains and distributed file systems thus bypassing PID Service Providers. "These intrinsic identifiers are quite powerful, as they allow not only to uniquely identify an object, but also to verify that the designated object has not been modified: it suffices to recompute the intrinsic identifier from the object itself to spot any alteration." (Allen A. et al., 2020).

[2]There are also PID ecosystems for citizens, like personal identification numbers in the Nordic Countries. Personal numbers are used for example in banking, for health and tax purposes.

The Software Heritage archive (https://www.softwareheritage.org/) has developed and implemented the SoftWare Heritage persistent IDentifier, i.e., SWHID, whose specifications are freely available at https://docs.softwareheritage.org/devel/swh-model/persistent-identifiers.html). SWHID is a versatile identifier enabling the user to select specific lines of code to be assigned the identifier. Within the archive, this unique hash enables anyone to retrieve the selected code. Through a browser, the permalink includes the resolver prefix and points to the right archive and lines of code.

Another worthy initiative is the freely available ISCC content code (https://iscc.codes/) that can be applied to any digital file of generic media types (text, image, audio, video). Processing the content with the algorithms defined by ISCC specification creates a unique composite code, consisting of four major elements, i.e., meta-ID, content-ID, data-ID and instance-ID. ISCC supports content versioning, related product identification, content variant detection, proof of data possession. ISCC is available under CC BY-NC-SA 4.0 License.

4.1.3 Self-Sovereign Identity and Decentralized Identifiers (DIDs)

Self-sovereign identity is "an identity management system which allows **individuals** to fully own and manage their digital identity" [Mühle et al., 2018] that exists independently of services, e.g., Facebook, Google or any other service requesting an individual to sign in. This identity management system is user-centric and relies on the distributed ledger of the blockchain that substitutes the registration authority in classic identity management systems. The principles of self-sovereign identity have been defined by [Allen, 2016]. A few services focused on individual identity have started operating, e.g., Sovrin, uPort [Naik & Jenkins, 2020], European Self-Sovereign Identity Framework (ESSIF).

Both Sovrin and uPort use Decentralized IDentifiers (DIDs) developed by the W3C to enable verifiable, decentralized digital identity. However, a DID can identify any referent (a "DID Subject", e.g., a person, organization, thing, data model, abstract entity, etc.) that the controller of the DID decides that it identifies. DID Subjects are defined using DID Documents which contain metadata about the DID Subject.

In contrast to typical, federated identifiers, DIDs have been designed so that they may be decoupled from centralized registries, identity providers, and certificate authorities. Specifically, while other parties might be used to help enable the discovery of information related to a DID, the design enables the controller of a DID to prove control over it without requiring permission from any other party. DIDs are URIs that associate a DID subject with a DID document allowing trustable interactions associated with that subject. Detailed information is available at https://www.w3.org/TR/did-core/. Investigations on how to use self-sovereign identity and DIDs to certify scientific datasets are under way [Barclay et al. 2020].

At the time of writing (Spring 2021) no working DID based implementations exist, however initiatives like GAIA-X are heavily promoting their adoption[3]. DIDs are machine readable by design and provide trust mechanisms for authenticity. It would be interesting to investigate if DIDs have the potential to complement or replace existing PID systems.

4.2 Stakeholders and Roles in the PID ecosystem

Simplified, the PID ecosystem consists of PID users and PID providers. I.e., at one end, there are the end users of the PID infrastructure, e.g., researchers, research communities and institutes. At the other end, we find the PID providers, e.g. PID Managers, like Repositories, Research Infrastructure Providers (like DEIC, CSC etc.), PID consortia (like those in EPIC, DataCite).

However, users can also be providers, and vice versa. I.e., research infrastructure providers can converge with users and research communities, thus working with more, if not all elements of the hourglass.

Moving forward as we aim to better understand the structure of and actors in the infrastructure landscape where PID must be embedded, we must investigate the current roles of the various stakeholders involved.,

Current PID stakeholders' roles (based on Hellström et al., 2020) can be analysed as follows:

4.2.1 PID Authority

A controller responsible for *maintaining the rules for defining the integrity and uniqueness 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 which enforce control over a PID infrastructure. But there

[3] See Chapter 3.4 in the GAIA-X Technical Architecture document: https://www.data-infrastructure.eu/GAIAX/Redaktion/EN/Publications/gaia-x-technical-architecture.pdf

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.

4.2.2 PID Service Provider

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

4.2.3 PID Manager

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 Authorit*y. 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 or publication repository, a data catalogue, or a research workflow system.

4.2.4 PID Owner

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 including location for the PID. The PID owner is the owner of the individual PID record.*

4.2.5 PID End-User

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 about these resources.*

5 Method, Steering, Budget and Time plan

Methodologically the concepts of Risk and Trust are governing, in that we aim to analyse the inherent risk to the stated vision, based on our perceived understanding of state-of-affairs for national, regional and global PID implementation.

The point of departure is an investigation of KE PID case studies. Being KE member centric, we gather knowledge from:

- Stakeholder interviews (researchers; e-Infrastructure providers; repositories)
- Documentation (reports, specs, websites, etc.)
- Recommendations for Good Practices, dos and don'ts
- Risks to be identified as far as these practices are concerned.

We will look at the extent that trust is needed, where, why and how, for a well-functioning PID infrastructure to emerge.

5.1 The concepts of Risk and Trust in a PID context

Organizations providing PIDs inherently create some level of *vendor lock-in*, i.e., into the PID scheme chosen. For example, even if a PID authority like *doi.org* can resolve URNs and another PID authority like *urn.fi* can resolve DOIs, the PID service providers are still bound to the respective management systems specific to a chosen schema. DOIs with metadata fields cannot easily be converted to URNs.

Researchers and organizations using PIDs, therefore, need to trust the governance of the underlying organization behind a PID schema, whichever organization it may be – foundation, commercial corporation, public body, non-profit organisation (e.g., ePIC, DataCite etc). Users need to ask if the PID scheme is sustainable; if the governance, strategy and tactics are transparent and trustworthy. Users must trust that PID providers in fact constantly and permanently resolve to the correct URL and that the referenced metadata is up to date.

Users need to assess their financial risks as well as potential risk to the future of their research agenda, should the PID infrastructure fail or become too burdensome. Consequently, PID service providers need to be well regulated to secure transparency, good governance, sound economy and solid technology, thereby guaranteeing sustainability of the PID's and the dependent research agendas. This holds true for publicly funded as well as commercial providers, for national providers as well as international providers.

The said need for regulation and governance is obviously quite challenging to do, for two very different reasons:

- Research institutions, indeed, the scientists themselves enjoy a large degree of autonomy, in organising science, choosing instrumentations and infrastructures. There are few governance

structures that aid agreement about scientific tool and infrastructure standardisation. While joint international science infrastructure endeavours are possible, they are so mostly in domain specific areas.

The risk here is that universities lack resolve and competence to be trusted to cooperate on building well-functional generic PID infrastructure.

- Commercial vendors are possibly better at orchestrating focus on enforcing or negotiating standards as well as developing specific infrastructure/software. However, for obvious reasons, they are only in it for the longer-term profits. Using a wide variety of PID systems at the same time for the same purpose might be confusing to users and administrators alike, but commercial vendors are inclined to seek schema dominance and create linkage and dependencies, which in turn are bad for scientific progress.

  The risk here is that the well-functional PID infrastructure comes at a very high cost, which we cannot trust to be only in monetary terms.

The risk for end users of using any given PID infrastructure or an already minted PID is more modest, as long as the data itself is somehow available. Non-resolving PIDs are a nuisance, certainly, but are not connected to the PID user's reputation. Nor is it critically disruptive to the research, like absent university/commercial regulation and governance could be.

5.2 Criteria for assessing risk and trust

Trust seems inevitable, and trust always entails acceptance of *some* risk. For each group of stakeholders, with a specific role, trust in a PID infrastructure could be based on a number of criteria, each entailing some degree of risk, big or small, potentially catastrophic or not. E.g.:

5.2.1 PID Authority

- placed under the control of an international standardization body that establishes and enforces transparent processes for creating, approving, maintaining and terminating PID standards,
- has signed a valid agreement with the international standardization body that defines its rights and obligations and provides for a succession plan,
- has a legal personality, public statutes, representative governance, governing bodies with clear voting rules.

5.2.2 PID Service Provider

- has an agreement with the PID Authority,
- has a clear business model which is reliable over the long term, accounts which are approved by a recognized auditing body/company,
- publishes an annual activity report,
- has a stable user base and cooperates on a regular basis with PID Managers implementing the PID within their own information systems,
- implements the financial, technical and human means to sustain the main PID information system and makes the necessary investments to this end,
- creates an information system under which the PID is persistently bound to its referent object notably through identification & description metadata for which creation, correction and updating rules and mechanisms have been defined, implemented and shared publicly,
- makes available PIDs in a timely manner to the PID Manager,
- PIDs managed by the PID Service Provider contain no semantics.

5.2.3 PID Manager

- manages sound processes to interact with the PID Service Provider to request PIDs for the local information system,
- uploads and updates referent metadata in the PID Service Provider's information systems,
- provides clear guidance to the PID Owner to create and update referent metadata regularly within the local information system,
- populates local information systems with PIDs.

5.2.4 PID Owner

- creates and updates PIDs' referent metadata regularly within the local system.

5.2.5 PID End User

- the identifier is unique for a given referent object and can be used as a proxy or as a link to the object,
- the identifier is recognisable as a persistent identifier and the type of the represented object is unambiguous (mere metadata, bit-level identity, identity of information content, dynamic object or collection etc)
- the identifier is granular enough to identify objects and their subparts along the requirements of the end user,
- the identifier is used at the international level by diverse content-producing and content-consuming organizations,
- the identifier resolves seamlessly to the referent object or to information about this object,
- the metadata linked to an identifier is available for reuse by authorized parties in a variety of formats,
- metadata relations and provenance are well documented,
- the PID is used widely across research production and consumption sectors.

Other criteria are expected to be identified during this study.

5.3 Method – Using the Open Scholarship Framework

When the concepts of trust and risk related to any given PID infrastructure are better understood, an analysis of different implementations and usage scenarios can be conducted from two perspectives: That of the PID users as well as that of PID providers, in a number of roles. Such analysis is conducted using KE member cases, where each KE members present PID implementations and where usage patterns are systematically analysed:

- **The present PID provider landscape**
  - Risks issues
  - Trust issues
- **The detected PID usage patterns across research disciplines and organisations**
  - Risks issues
  - Trust issues

When looking into and analysing various PID implementations and usage patterns, across the many KE cases, it is important to apply a systematic approach to identifying risk and trust issues.

The, arguably, best framework for such analysis is the **Open Scholarship Framework**, introducing a number of "points of analysis" along three dimensions – Research Phase; Arena; and Societal Level, as depicted below (Figure 1).

A lot of the emerging issues are expected to be 'collective action' challenges. With help of the Framework, we can identify and describe the issue, and cluster recommendations for each actor/stakeholder. The investigation can look at emerging problems and challenges to a well-functioning PID infrastructure through the lens of the Open Scholarship framework.

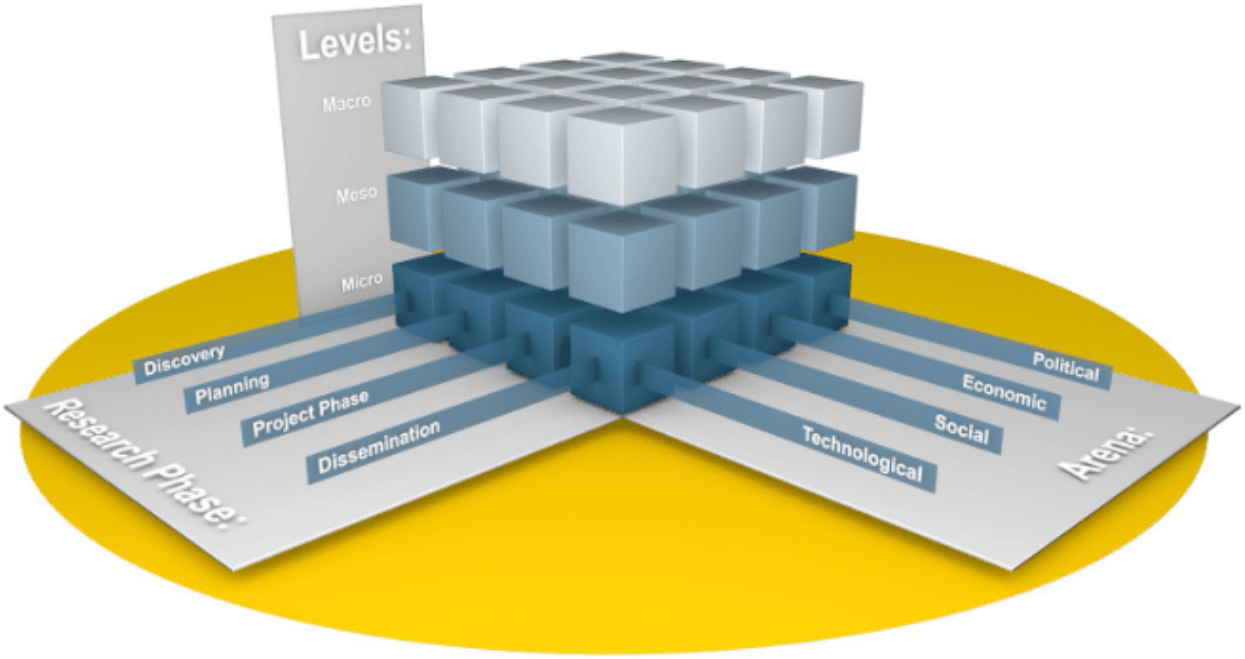


Figure 1: The Open Scholarship Framework

5.4 Project Steering

It is proposed that an Expert Consultant is contracted to draft a report – ***Risks and Trust in Pursuit of a Well-functioning Persistent Identifier Infrastructure for Research***. The report must be drafted in cooperation with appropriate KE groups.

5.5 Budget and Time Plan

A total budget of EUR xx is allocated for an Expert Consultant.

The activity – drafting and approving the report – is to be completed by summer 2022. There will be quarterly progress reports delivered to the Steering Group in connection with presentation and discussion meetings.

6 Investigation of KE PID Cases Stories

Case studies among KE members will be conducted analysing different PID implementations and usage scenarios. It is consciously KE member state centric. It is *not* a peer review of different PID systems, but specifically looking at current implementation and usages, and identifying *issues* that have bearing on a well-functioning PID infrastructure. In so doing risk and trust issues are identified from a user as well a provider perspective.

6.1 Themes of interest

Identified issues of importance, related to a well-functioning PID infrastructure, are mapped into the dimensions of interest – Research Phase; Arena; Societal Level (Figure 1) – and are described and analysed in terms of risk and trust. A number of questions are obvious to present others less so, and still others yet to be raised. An inconclusive number of questions could be gathered in themes:

- Degree of national coordination (policies, organisation, funding)
    - How are PID services/infrastructure organised, and how is it coordinated nationally/internationally?
- Awareness/Coverage
    - Do funders require the use of PIDs from researchers?
    - Is citation using PIDs common practice?
    - Are PIDs used in most of the large data providing organizations?
    - How fragmented is the PID landscape?
    - What PID systems are used in various KE member organizations?
- Implementations
    - Is PID technology self-maintained or outsourced?

    - How is metadata maintained and curated?
    - Presence policy and governance models. I.e., is management centralised or not; which responsibilities are allocated how (related to PIDs and exit strategies)?
    - Use of PIDs in research management, evaluation and/or analytics (i.e., collecting information about a person or project)
    - Which public providers and user relationships are present with any commercial partners?
    - Under which (risk/trust) terms are any commercial relationships?
- Ongoing developments
    - National and/or institutional strategy or planning
    - Which advanced PID features are presented and used, e.g., like part-identifiers?
    - New implementations (within e.g., CRIS, data, researchers, repositories, research funding, projects, etc.)
    - PIDs in the OS ecosystem, by looking into how to any given PID implementation and usage aligned with the current EU Open Access policies?
- Interoperability
    - To what extent are existing PID systems interoperable?
- Funding
    - How are the national PID ecosystems funded, how is their persistence guaranteed?
    - Which budget (EUR and Person Months) is consumed, on international, national and/or local levels?

6.2 Introducing the KE member national PID landscape of providers and users

Below we describe a basic set of attributes related to each KE member state, based on a standardised template. Obviously, when looking further into the specific implementations and usage scenarios for that KE member, the template must be seen as a point of departure.

6.2.1 Finland (CSC)

On a general level PID use is quite mature in Finland and a relatively centralized infrastructure. We have national consortia for ORCIDs and DataCite. For publications, journals.fi allocates CrossRef DOIs. URNs are actively supported by the National Library and widely used. Handles are also used, most notably by EUDAT Services National recommendations for PIDs in research exist. The national CRIS research.fi service makes extensive use of PIDs. Below we summarize potential issues and areas worth exploring.

Risks and trust:

- Risk: The FAIR principles are misused in assessment
- Risk: Use of URNs in a FAIR context is undefined
- Trust: Issues with dynamic information behind PIDs, where there can be a lack of trust when End-Users are not sure what elements are underlying -> persistency aspect becomes questionable. We need to define what the persistent element really consists of, and consequently, we also need to define what elements might be subject to change.
- Risk & Trust: Sustainability issues when projects use PIDs extensively and/or develop new PID approaches: Who maintains them after the project ends?
- Risk: PID interoperability (on EOSC level, e.g., PID Kernel Profiles vs. URNs vs. DOIs)
- Trust: Unresolved issues above have severe negative implications on PID usage

PID issues worth exploring:

- National PID Forum and PID landscape
    - DOI DataCite Consortium Finland
    - CSC a member of EPIC Consortium
    - Finnish National Library is active in providing an URN service
    - PIDs that are used and might be used more in the future in the national CRIS system research.fi

- PID Micro Service at CSC

6.2.2 France (Centre National de la Recherche Scientifique - CNRS)

French scientific organisations and authorities are convinced that the adoption of international identifiers, e.g., identifiers for individuals, organisations, publications, software, digital objects and research data, can only benefit the visibility and the impact of French scientific output. The national action plan conducted by the Ministry of Higher Education will focus on the following areas:

- Accelerate the adoption of identifiers by researchers and research institutions, and other players implied in research activities (funders, assessment and evaluation bodies, bibliographic data, open archives, publishers, etc)
- Identify viable business models to ensure sustainable deployment of identifiers,
- Improve interoperability and standardization of identifiers while ensuring the scientific community keeps control,
- Contribute to the monitoring and evolution of identification systems in order to guarantee openness and independence in the long term.

A detailed action plan shall be drafted by the special Open Science Committee reporting to the French Ministry of Higher education. This plan will focus on the coordination of governmental actors, organisations, researchers and service operators. It will foster the implementation of concrete actions to encourage the use of tools and identification systems and maintain an up-to-date inventory of the situation regarding their level of adoption. It must accompany the development of useful services based on use cases to demonstrate the benefits of PIDs. Finally, it must develop appropriate communication actions towards researchers to guide them in their daily use of identifiers.

Four distinct actions will be carried out regarding:
1. the identification of organizations,
2. the identification of individuals,
3. the identification of publications,
4. the identification of digital data and objects.

Action 1 - Identifiers for organizations
The identification of research structures and their affiliations is still under development on a global scale. Even if the international standards are not yet fully defined, it is important to make the national systems compatible with each other and with the rest of the world.

To date, the RNSR (Répertoire National des Structures de Recherche) appears to be the most successful national tool for identifying research organisations. The AureHAL directory built by the CCSD (Centre pour la communication scientifique directe) is more granular than RNSR as it identifies research teams. The national search engine for research activities, named ScanR, already uses these identifiers. The connection of national identifiers for structures with international ones like ISNI or ROR or Q Wikidata is a work in progress held by the higher education bibliographic agency (Abes) in its highly curated global registry IdRef.fr, and by the Bibliothèque nationale de France as national agency for ISNI, among other players. The French Ministry of Higher education is to coordinate all national stakeholders.

Action 2 - Identifiers for individuals
The growing international adoption of ORCID as the identifier for researchers should ground the French policy towards the systematic use of standard identifiers for authors and scientific contributors across the various scientific disciplines.

At the operational level, this will consist in:
- Enabling the widest possible adoption of ORCID by French researchers,
- Providing evidence of the usefulness of ORCID through use cases in universities and institutions and their information systems,
- Producing communication and training tools and materials aimed at researchers,
- Ensuring that French researchers alive and dead have an IdRef, an ISNI, and a Q Wikidata, aligned with their ORCID, IdHAL or Researcher ID(s)when available.

In 2019, the French Ministry of Higher Education, Research and Innovation, within the framework of the Committee for Open Science, has mandated the Couperin consortium to set up a joint membership to ORCID on behalf of higher education and research institutions. ABES (Agence bibliographique de l'enseignement supérieur) is in charge of the technical coordination of the French ORCID Consortium.

Action 3 - Identifiers for publications
Scientific articles are nowadays mainly - but not only - identified through DOIs, whose main assignment agency is Crossref. CrossRef has developed additional services, e.g., citation crosslinking. However, DOIs can also be assigned to publications by other agencies, e.g., DataCite, that are members of the International DOI Foundation based in the United States. In addition, other identifiers are assigned by French research agencies for specific uses, such as HAL ID (HAL/CCSD) or Handle ID (Isidore/CNRS).

Most French publications have already been assigned identifiers, but it would help to assess the importance of each identifier to inform the national strategy. In addition, alignment between these various identifiers is only partially realized and should be generalized by cooperation between the various stakeholders (CCSD, Abes, ScanR, HumaNum). Last but not least, open citations are highly desirable to monitor scientific progress and a report on this topic should be commissioned shortly by the French authorities.

The French Ministry of Higher Education supports the ISSN International Centre, an intergovernmental organisation which is the ISO Registration Authority for the management of standard ISO 3297:2020 - ISSN that identifies serial publications and continuing resources. This organisation receives funding from the French authorities and more than 90 member countries. ISSN is widely used for the identification of print and digital serial publications and other continuing resources by libraries and publishers. The ISSN Portal (portal.issn.org) provides free identification data and URLs for serial publications. Additional services provide data on title transfers and digital archiving. The ISSN IC is currently investigating the implementation of a URN resolver (urn.issn.org).

Action 4 - Identifiers for research data and objects
Because of the complexity and the variety of the data to be managed and its usage, the identification of research data is still immature. Various systems coexist, mainly based on DOI and Handle, but the needs of the communities and the way to meet them, are not yet fully defined.

Reflection within various groups has led to the development of the Digital Object concept developed in particular by the C2CAMP international expert group. The range of entities to be identified is practically infinite. An international initiative in which France participates has developed a promising software identification system (https://www.softwareheritage.org). The French node of the Research Data Alliance is involved in this project. Reflection is triggered at the European level by the current implementation of EOSC and the new Horizon Europe directives (https://oaamu.hypotheses.org/2722?s=09 in French). Concrete action plans are supposed to emerge shortly.

6.2.3 Denmark (DeiC)

Denmark has formed a national consortium under DataCite, where all DOI costs are covered nationally by the Danish e-Infrastructure Cooperation (DeiC) which is the consortium lead, representing all consortium members internationally.

In Denmark the first national ORCID consortium (http://orcid.dk/) was formed back in 2014. The consortium is led by Aalborg University. A national consortium approach to both ORCID and DataCite is widely accepted by Danish research organisations.

The Danish national Royal Library uses DOI from Crossref in their publishing platform for e-journals. It uses the widely used publishing platform Open Journal Systems (OJS), which is a program that allows for publishing online quickly and easily.

Royal Library also has a service publishing e-books based on an Open Monograph Press server (http://ebooks.au.dk/index.php/aul), using DOI from Crossref.

Aalborg University also uses Open Journal Systems (OJS) but with DOI's from DataCite.

Early on Denmark formed Danpid, a Danish PID service, that is used to create lasting links to objects at, for example, the Danish national libraries and museums or objects published by the government, municipal organizations and educational institutions. "Handle" is the name of the infrastructure on which Danpid is based (the same infrastructure that is also used by DOI).

Denmark has a long history within digital conservation initiatives. One prominent initiative was Digital Conservation (https://digitalbevaring.dk, a cooperation project by the Danish National Archives together with the Royal Library. It has now been closed, restructuring archiving and PID infrastructures to have a stronger linkage to universities. However, Library Open Access Repository (LOAR), under the Royal Library, is an open data repository established in 2016 as a service for storing and providing access to Danish research data. The service was formed with the following policy motivation and key goals:

- Make data accessible to review for publications
- Enable researchers to meet requirements for Danish and European grants
- Ensure data privacy and removal of data as appropriate
- Enable reuse of data where appropriate

Researchers who upload data are expected to share the data using Creative Commons licenses.

6.2.4 Germany (DFG)

Main Risk: Adoption of PIDs and agreement upon international standards
The German research infrastructure landscape is very fragmented in relation to repository system solutions, data management policies, metadata standards and PIDs. Another challenge is the adoption of

digital PIDs e.g., DOIs, ORCIDs, ROR in formal bibliographic indexing processes. Hereby URNs, administered and assigned by the German National Library, play an important role. Another key element in the collection process is the German Integrated Authority File (GND), which describes works, geographic entities, conferences and people. The German ORCID Consortium founded in 2016, led by Technical Informationsbibliothek (TIB) and supported by the DFG funded project ORCID DE, is advancing the ORCID implementation in Germany. In the process ORCID has been integrated in the GND. PIDs are recommended by several German policy makers but there is no mandate.

Germany is in the process of building a national research data infrastructure (NFDI), where consortia are tackling discipline specific issues by following an overarching strategy. The adaptation of PIDs based on international standards is one of the consortia objectives (cross-cutting topic). Therefore, it would be a good use case to analyse the arising risk and trust issues along the three dimensions.

The adoption of the Research Organization Registry (ROR) as an organizational identifier in Germany as a new PID system will be another good use case and the outcome of a recently conducted survey across the German research landscape can be a great starting point. Particularly interesting are the risk and trust issues associated with the challenge to overcome reliable but limited regional structures.

DataCite DOIs were provided free of charge in Germany from 2012 until the end of 2020, when the new DataCite fee and member model was implemented. TIB formed a consortium and many of the former clients became paying consortium members, joined DataCite directly or moved to Crossref. But there were some organizations that were no longer interested in DOIs. It will be interesting to analyse the risk and trust issues of these organizations based on the three dimensions.

The German National Library (DNB) assigns URNs to its digital, mostly legal-deposit publications and also offers the usage of its URN-infrastructure to registered partners (research institutions, libraries and publishers). The service is free of charge for partners and end-users alike. URN-partners usually administer and assign URNs themselves and are also responsible for ensuring their persistence. For URNs assigned by the DNB directly, the library stores a copy of the publication in its own digital archive. An analysis of risk and trust issues around URNs would be interesting in order to highlight differences between PID-infrastructures that target overlapping types of data.

6.2.5 United Kingdom (Jisc)

https://repository.jisc.ac.uk/8107/1/PIDs%20for%20OA%20project%20community%20survey%20report.pdf

Key recommendations based on the survey findings are:

- Interventions are needed to improve the scale and depth of PID integration in every day workflows
- Respondents want to see PIDs being used optimally in funding systems (both for grant application/award and for reporting), content platforms which host research outputs (including data and e-books as well as articles), and research information management tools within institutions
- Metadata associated with PIDs is vital. It needs to be predictably present, contain more consistent elements, and be reliably maintained and updated
- For new or emerging PIDs, there are lessons to be learned from ROR's engagement strategies, which have enabled it to make remarkable progress
- Barriers to adoption need to be lowered.

Another report, written by Josh Brown will be published later in 2021, based on the work of the UK PID stakeholder consortium, which was established at the encouragement of UKRI.

PIDs worth exploring: https://scholarlycommunications.jiscinvolve.org/wp/author/manistaf/

The OA/PID roadmap project has been passed over to an expert at Jisc. Jisc is leading follow-up work in collaboration with MoreBrains Cooperative, LTD. He will also be able to update on the internal resource and infrastructure strategic plan around PIDs.

The PID forum's new home with NISO after the end of the Freya project:
https://www.project-freya.eu/en/blogs/blogs/a-new-home-for-the-pid-forum-at-niso

ORCID iDs for people, Crossref and DataCite DOIs for outputs, Crossref grant DOIs, ROR identifiers for organisations, and RAiDs for projects.

They want to establish a national PID strategy based on the 2019 roadmap (discussed in those blog posts)

6.2.6 Netherlands (SURF)

In the Netherlands, there are several PID services embedded in specific use cases. Each (see below) operates independently from the others with a low level of coordination between them. SURF is presently developing a national framework to increase PID coordination and explore ways to increase interaction.

NL PID Landscape at a glance:

ORCID	Researcher, contributor ID (ORCID-NL consortium)
ISNI	Author ID, University Libraries (registrant)
DOI: HSS data	EASY for HSS datasets (DANS)
DOI: HSS data	DataverseNL, during research (DANS)
DOI: data	Datacite (41 repository accounts via 4TU)
ePIC: data	Handle for datasets during research (SURF)
varius: objects/data	Digital Cultural Heritage (pid guide)
PURL: objects/data	Biodiversity objects/collections
URN: NBN	Publications, National Library of the Netherlands

Additional PIDs of interest: RoR, RAiD, GrantID

Risk

Whereas community governed PID organizations support, and in some cases embrace, development of open research infrastructure, contributions from for-profit publishing/data companies complicate this effort. To be clear, this is not a tension between good and bad actors. Rather, it is a matter of principles. It's an ongoing tension between community interests and shareholder interests, a tension that produces uncertainty in the long-term sustainability of PID workflows that depend on enduring access to associated metadata.

A related concern is the commercial bundling of metadata (including PIDs) with a tightly integrated set of services. For example, the combination of CRIS, data subscription, and analytics software creates a form of vendor lock-in.

Mitigation of risks
Awareness at the national level of universities. Prevent vendor locking regarding knowledge systems. Special attention for the use of applications provided by commercial vendors and big tech. A set of principles is defined which should be part of negotiations with stakeholders who play a role in knowledge systems. At the local level Open-Source technologies are part of some of the Open Science programs of universities.

7 Summary of KE PID Cases

It is proposed that an Expert Consultant in close cooperation with the KE member state representative and the KE PID Task & Finish Group draft a summary of KE PID cases.

7.1 Risk

- Which are the perceived risks on the overall accumulated level?
- How can the risks in the current PID infrastructure be understood, in the light of various dimensions of analysis (The Open Scholarship Framework)?

- How can risk be mitigated (e.g., by building a stable foundation built on trust, or international organisation, agreements, standards; contract …)?

7.2 Trust

- How is the notion of trust relevant, for the well-functioning of a future internationally recognized, interoperable PID infrastructure?
- Trust entails acceptance of 'some' risk

8 Concluding PID Risk and Trust Issues

It is proposed that an Expert Consultant in close cooperation with the KE PID Task & Finish Group draft conclusions of the conducted analysis.

9 Recommendations

It is proposed that an Expert Consultant in close cooperation with the KE PID Task & Finish Group draft recommendations for short term and long term KE follow-up action aimed at the KE partner organisations, as well as KE member state research organisations.

~ # ~

10 Supporting documents

10.1 Related projects and reports

European projects

- CLARIN/DARIAH : several documents available re PIDs (https://www.clarin.eu/sites/default/files/pid-CLARIN-ShortGuide.pdf, https://zenodo.org/record/3744091#.YAqsHOhKg2w)
  - notably P. Wittenburg. "Persistent end Unique Identifiers". CLARIN-2008-2.
- FREYA project (https://www.project-freya.eu/en/deliverables/freya_d3-1.pdf)
- FREYA PID Federation Scoping Study: Final report (https://zenodo.org/record/4059557)
- THOR project https://project-thor.readme.io/docs/project-deliverables
- ODIN project https://odin-project.eu/project-outputs/deliverables/
- DEFF Opera project: https://deffopera.dk/
- PID Graphs: https://doi.org/10.1016/j.patter.2020.100180

EOSC

- Hellström, Maggie, André Heughebaert, et al. 'Second Draft Persistent Identifier (PID) Policy for the European Open Science Cloud (EOSC)'. 2020, May 1. https://doi.org/10.5281/zenodo.3780423. Or https://www.eoscsecretariat.eu/eosc-liaison-platform/post/launch-initial-persistent-identifier-policy-eosc
- A Persistent Identifier (PID) policy for the European Open Science Cloud (EOSC): https://op.europa.eu/da/publication-detail/-/publication/35c5ca10-1417-11eb-b57e-01aa75ed71a1/language-en
- EOSC PID Architecture https://docs.google.com/document/d/1T-bpNsmuxQewsLq48XTyUJoe0lsV7poaXohpgDo9W34/edit

Research Data Alliance

- RDA PID Kernel WG (https://www.rd-alliance.org/groups/pid-kernel-information-profile-management-wg)
- Research Data Alliance/FORCE11 Software Source Code Identification WG; Allen, Alice; Bandrowski, Anita; Chan, Peter; di Cosmo, Roberto; Fenner, Martin; Garcia, Leyla; Gruenpeter, Morane; Jones, Catherine M.; Katz, Daniel S.; Kunze, John; Schubotz, Moritz; Todorov, Ilian T. "Software Source Code Identification Use cases and identifier schemes for persistent software source code identification". 2020. https://zenodo.org/record/4312464#.X9BuzGgza70
- Wittenburg, Peter, Hellström, Margareta, Zwölf, Carlo-Maria, Abroshan, Hossein, Asmi, Ari, Di Bernardo, Giuseppe, Couvreur, Danielle, Gaizer, Tamas, Holub, Petr, Hooft, Rob,

Häggström, Ingemar, Kohler, Manfred, Koureas, Dimitris, Kuchinke, Wolfgang, Milanesi, Luciano, Padfield, Joseph, Rosato, Antonio, Staiger, Christine, van Uytvanck, Dieter, and Weigel, Tobias (2017) "Persistent identifiers: Consolidated assertions. Status of November 2017." https://doi.org/10.5281/zenodo.1116189

FAIRsFAIR WP2+WP3

- FAIRsFAIR 2nd Report on FAIR requirements for persistence and interoperability https://doi.org/10.5281/zenodo.4001631

JISC

- Developing a persistent identifier roadmap for open access to UK research (https://repository.jisc.ac.uk/7840/2/PID_roadmap_for_open_access_to_UK_research.pdf)
- UK PID Consortia: UK Open Access to Research publications - Adam Tickell (2018) (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/774956/Open-access-to-research-publications-2018.pdf)

KE's report on the Openness Profile: https://www.knowledge-exchange.info/event/openness-profile

10.2 Other resources

- Pidforum.org https://www.pidforum.org/
- Pid service registry https://pidservices.org/
- Fair Principles: https://www.go-fair.org/fair-principles/
- Systematize information on journal policies and practices - A call to action https://leidenmadtrics.nl/articles/systematize-information-on-journal-policies-and-practices-a-call-to-action
- PID Commons (Datacite) https://blog.datacite.org/power-of-pids/
- Choosing and implementing persistent identifiers: Guide for research organisations https://doi.org/10.5281/zenodo.4395767
- The use of Persistent Identifiers for Research Datasets: Recommendation by the Finnish Scientific Community for Open Research https://doi.org/10.5281/zenodo.3560738
- GAIA-X: Technical Architecture: Release June 2020 https://www.data-infrastructure.eu/GAIAX/Redaktion/EN/Publications/gaia-x-technical-architecture.pdf

10.3 Bibliography

- C. Allen. Self-sovereign identity principles. 2016. https://github.com/ChristopherA/self-sovereign-identity/blob/master/self-sovereign-identity-principles.md
- I. Barclay, A. Preece, I. Taylor et al. Certifying provenance of scientific datasets with Self-sovereign identity and verifiable credentials. (2020). https://arxiv.org/pdf/2004.02796
- A. Dappert et al. "Connecting the persistent identifier ecosystem: Building the technical and human infrastructure for open research". 2017. https://datascience.codata.org/articles/10.5334/dsj-2017-028/
- P. Golodoniuc et al. "PID Service. An advanced persistent identifier management service for the semantic web". 2015. https://www.researchgate.net/profile/Pavel_Golodoniuc/publication/284087065_PID_Service_-_an_advanced_persistent_identifier_management_service_for_the_Semantic_Web/links/564bf3a908ae3374e5ddec58/PID-Service-an-advanced-persistent-identifier-management-service-for-the-Semantic-Web.pdf
- Martin Klein, Lyudmila Balakireva. "On the Persistence of Persistent Identifiers of the Scholarly Web". 2020. https://arxiv.org/pdf/2004.03011.pdf
- Jens Klump, Robert Huber. "20 years of persistent identifiers. Which systems are here to stay?" 2017. https://datascience.codata.org/articles/10.5334/dsj-2017-009/
- Alexander Mühle et al. "A Survey on Essential Components of a Self-Sovereign Identity", 2018. https://arxiv.org/abs/1807.06346

- N. Naik & P. Jenkins. Self-sovereign identity specifications: Govern your identity through your digital wallet using Blockchain technology. 2020. https://publications.aston.ac.uk/id/eprint/41998/1/SSI_Specifications_uPort_SovrinDrNitinNaik.pdf
- Aurelia Vasile et al. "Le DOI, une impérieuse nécessité? L'exemple de l'attribution de DOI à la collection Pangloss, archive ouverte de langues en danger". 2020. https://halshs.archives-ouvertes.fr/halshs-02870206/file/DOI_VersionAuteur_HAL.pdf