PID Infrastructure for Europe

Klaas Wierenga, Mark van de Sanden, Ulrich Schwardmann, Peter Wittenburg, Christophe Blanchi

Minimal Viable European PID[1] Organisation

This chapter is a summary of what is being described more broadly in the following chapter to describe a Minimal Viable PID Organisation for Europe. Such MV PID Organisation would be characterised by

- a EU Registration Authority that organises the European PID domain in full responsibility according to the emerging PID requirements and the needs from national, scientific, industrial communities
- implementing feasible delegation methods to be scalable
- integrating already existing European communities and service provider structures
- aligning their activities with global communities such as IDF, Crossref and Datacite which are also active in Europe
- interacting with and contributing to the DONA Foundation and its root node infrastructure.

Currently, a PID policy for EOSC is being developed by the PID Policy Task Force which consists of members of the two EOSC WGs on FAIR and Architecture[2]. The FAIR PID group collected various use cases and papers from different initiatives to have a broad coverage of contributions and ideas. A recent Amsterdam meeting on PIDs organised by RDA GEDE referred to different practices and expectations. The requirements of this meeting are available[3] and were also introduced to the PID Policy Task Force discussions. Therefore, we refer to these documents which describe the requirements for a European PID ecosystem[4].

1. Pre-Remarks

Preventing the Dark Digital Age

During the last two decades the landscape of PIDs, which is a widely agreed acronym for **globally unique, persistent and resolvable identifiers**, evolved rapidly since the relevant actors started understanding that mechanisms are required in the digital domain that have the potential to stay for very long times to prevent the "digital dark age" as V. Cerf stated it[5]. The threat of losing our digital memory is so serious that we need to overcome the phase of testing and turn to serious, stable and well-governed infrastructure components. EOSC is an attempt to lead to a change, thus it is time to look at the evolved landscape, compare it with the requirements and draw conclusions. We agree with Karel Luyben[6] who pointed out that looking at the evolved PID landscape is important, but not sufficient.

---

1 The term PID is used in this paper as being used in the Research Data Alliance as "globally unique, persistent and resolvable identifier.

2 https://www.eoscsecretariat.eu/eosc-working-groups

3 https://github.com/GEDE-RDA-Europe/GEDE/tree/master/PID

4 It should be noted that such an MV PID System would fulfil the major policy requirements currently being discussed.

5 https://www.youtube.com/watch?v=sfA2fwAL_W4

6 K. Luyben is currently chair of the EOSC Executive Board.

Making plans for EOSC requires to lean back and to describe the landscape as it should be ideally. We will not build any ultimate solution, but we need to understand which steps need to be taken so that long life-times of digital objects can be guaranteed - objects can include any bit sequence (data, metadata, software, assertions, references, etc.) - as intended by FAIR. Technologies will come and go; software tools will change fast replacing older tools using state-of-the-art technologies. What needs to be available over long-time periods are: (1) A **Basic Data Model** for describing, managing and accessing the bit sequences used to represent valuable content. (2) **Repositories** are pillars that need to store and maintain most of the valuable bit sequences. (3) **PID registry systems** to maintain the references to and between digital objects in a stable way so that they survive all technology changes.

Referenced Entities

There is an ongoing discussion what PIDs point to. Different terms and concepts are being discussed. In this paper we will take a simple and hopefully clarifying view.

- Every **digital entity** that has a PID and metadata is called a **Digital Object** using the definition as made by RDA DFT.
- There are **digital entities** that live on discs, smart cards, and all sorts of different devices and that have not been associated with a PID and are therefore not discoverable and accessible.
- **Physical objects** need to have a digital representation to be part of operations in the digital domain. If these digital representations are registered with a PID we call them DO.
- The term DO **abstracts** away from the type of content the bit sequences are encoding, i.e. all types of content can be made a DO at that moment where its encoded content is described and has been registered with a PID.
- If DOs are extended to fulfil all FAIR principles, we call the DOs **FAIR Digital Objects**.

Basic PID-based and added-value Services

Basis for all stable references is the availability of **basic PID services** that would allow authorized users to create, manage, and resolve PIDs independently of the evolving underlying technologies, i.e. the references themselves need to be independent of current technology[7]. Using information available at the level of PIDs[8] can be used, for example, to detect duplicates based on checksums. More such added-value services that are only be based on PID information will evolve over time.

We separate this domain of basic services that need to be available at low costs for every researcher in Europe (and beyond) from **added-value services** that are dependent on other information sources. Much effort is spent in initiatives such as euroCris, Datacite and Crossref and projects such as FREYA on linking between different types of digital objects such as publications, funders, projects, organisations, etc. Such important added value services are highly dependent on metadata as they are produced and maintained by scientific communities in very different flavours. Based on a functioning and integrated PID domain and FAIR compliant metadata these services and the linking can be re-produced by using state-of-the-art technology. We assume that this domain of added-value services will be a highly dynamic scenario of high interest to different stakeholders, with different business models.

To support a lean PID infrastructure we argue for a separation of these two aspects.

---

7 We assume more detailed specification on PID schemas and resolution behavior to improve interoperability that go beyond the ITU X.1255 interoperability guidelines.

8 We assume here that PIDs are resolved to useful and machine actionable "state" information as defined by the RDA Kernel Information Group.

Independence of Commercial Solutions

Specific registries are essential for managing a stable and trustworthy infrastructure landscape such as cadastral registries for land ownership for our societies. In this paper, with respect to registries maintaining scientific knowledge in form of referential structures we express the following positions:

- The backbone infrastructure for PIDs needs to be independent of commercial interests, since it manages essential scientific knowledge that needs to be available for long-term, it will be an increasingly important part of our digital heritage.
- The backbone infrastructure needs to be flexible enough to support all kinds of application scenarios evolving in state-of-the-art data science and the business/governance models need to be defined by the application community.
- There will be specific PID-related added value services that will be managed by services related to commercial providers such as given by Crossref under the envelop of the DOI Foundation.

2. Current PID Service Landscape

We currently see the following major players in the landscape of PID service providers.

- The **web community**, which includes most scientist, is by far the largest user of references in the digital domain. URIs are the basis for references to "resources" on the Web. In general, those references are URIs in the form of URLs. These references generally include hosting organization and ownership semantics and use DNS as a resolution system in combination with the HTTP protocol request passing on the service part of the URI to the server to access the referenced digital information. URLs are unstable because the DNS name can change, and the service request part of the URL is directly dependent on the service implementation to provide access to the digital information. To compensate for this inherent instability of URLs "Cool URIs" and the "PURL redirection service" were invented. But what worked well for the domain of general information where "link rot" is accepted, is not sufficient for the data and research community.
- **URNs** are another form of URIs are being used by several national libraries to give access to their digital holding. URNs where not designed to provide any global resolution solutions and because of this limitation no global resolvers were developed. Currently resolution occurs via dedicated web service and is not guaranteed to be resolvable nor unique.
- The **ARK system**, developed and maintained by the Californian Digital Library, is being used by some repositories, but it lacks broad uptake and will hardly persist over long time periods.
- The **Handle System**, developed by CNRI, has been operational and in constant use since the late 90s. The Global Handle Registry (GHR) is a global distributed service similar to the DNS root node in function that provides global handle resolution services. Operations of the GHR were originally administered by CNRI until they were fully handed over to the Swiss based, not for profit DONA Foundation at the end of 2015 and is operated now by DONA in collaboration with its Multi-Primary Administrators (MPAs). The Handle System has a broad global uptake in science, industry, and publishers. Over 50.000 prefixes have been assigned to various communities and organisations running their local services based on own specifications and business models.
- The **DOI community** is a large community of handle users. The DOI Foundation is an MPA at the DONA Foundation and it manages its allotted prefix 10. As result all DOIs currently start with a “10.” The DOI Foundation allots prefixes derived from 10 to entities called registration agencies according to its policies and business model. The registration agencies can have their own policies and business models. Some well know service providers are CrossRef and DataCite. These service providers also define technical specifications and business models of their own. Registration agencies typically provide added value to digital object system by providing services such as reverse lookup and linking to name just a few.

- The **ePIC community** is another large community of handle users mainly based in Europe. It is associated with the GWDG MPA who manages prefix 21. ePIC as service provider is offering a redundant network of Handle providers located at a number of large European data centres organised within EUDAT to address the needs of the labs that produce very large and complex data collections at high rate and need to process them increasingly using automatic workflows.
- The **National Chinese Handle Coalition** serves a very large industrial base in China. Its use of Handles is mostly targeted especially for large industrial applications[9]. Their operational requirements are very specific to Chinese regulations and the industrial needs of their customers.
- A completely different approach has been taken by **ORCID** which is now a global standard to assign PIDs to persons which started from unifying the identification of publishing researchers.

The two major ID systems being used are the domain of URIs which is guided by the W3C organisation and the domain of Handles (incl. the global DOI community). In the following we will focus on in this paper.

2.1 Authority and Responsibility Domains in the Handle System

A Handle is an identifier consisting of a prefix and a suffix separated by a "/". A prefix typically consists of numerical value separated by 0 or more "." delimiters. The suffix has no such limitations and can accommodate any sort of identifier scheme. It is important to note that the "." in a prefix does not per se represent an organizational hierarchy. Each new delimiter may be used to represent another organizational and prefix delegation. This discreet "." delimiter-based delegation is sometime called an authority domain. The Handle syntax and the handle system allow for a practically unlimited number of delimiters in a prefix, but in practice there is rarely no need for more than two delimiters:

<Credential-Prefix>.<1-DelimiterPrefix>.<2-DelimiterPrefix>/<suffix>#<fragment>

This delimiter scheme allows doing delegation to authorities that take full responsibility for a name space. The credential-prefixes are typically reserved to regions/countries, public/private coalitions, and global communities. The International DOI Foundation[10] (IDF) and International Telecom Union (ITU), for example, are globally organised communities and have Credential-Prefixes. The DONA Foundation (see below) is responsible for allotting credential-prefixes. Currently, all 10 so-called Multiple Primary Agencies (MPA) each operating a node within the distributed Global Handle Resolver have a credential-prefix. These MPAs act in two roles: (1) They are acting as GHR service providers that are involved in scaling and maintaining the integrity and availability of the GHR and its prefix information. (2) They act as registration authorities to assign 1 or 2-delimiter prefixes.

The Handle System (HS) is a global service that provides the functionality to register, resolve, and administer handles being able to accommodate many existing identifier syntax and requirements. Each handle can consistently be resolved into a handle record which consists of a set of typed attribute value pairs, also called Kernel Information Types. Profiles of such attribute-value pairs are defined in general by registration authorities and/or repositories. To foster interoperability these Kernel Information Types need to be defined in open type registries or ontologies to make them machine actionable.

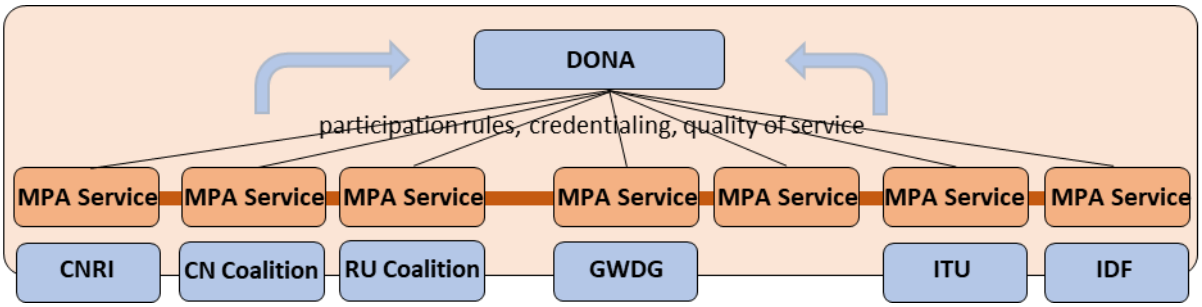
---

9 A production system in the area of food supply chain control has been set up for China leading to the creation and management of billions of Handles.

10 The DOI prefix is 10.

2.2 Embedding in DONA GHR Domain

The DONA Foundation only takes responsibility of the operations, performance, and integrity of the GHR. It enforces strict participation rules of all MPA GHR Services and asks each MPA for an annual fee. The DONA Foundation is neither involved in the technical specifications nor in the business models of any of the MPAs and their user communities. The DONA Foundation is based in Switzerland for neutrality reasons. It is responsible for evolving the Handle System standards and implementations along with its MPAs. The DONA Foundation is run by a board of committed experts acting in their personal capacity. The board takes care that major regional actors and communities and that MPAs are represented. It currently consists of 11 members including experts from the US, China, Russia, Africa and Europe. European members, nominated from the beginning, currently are Stefan Eberhard (legal advisor), Antoine Geißbühler (Uni Geneva) and Peter Wittenburg (Max Planck Society). Clear rules describe the activities of the board and its processes including the nomination/election of new board members (blue arrows).



DONA does not take any liability for the activities of the different institutions. Institutions could also be those that offer services to industrial consortia. Also, DONA does not make any suggestions about the delegation of responsibilities to registration authorities (RA) with respect to sub-namespaces and the subcontracting with the PID service administrators. The DONA Foundation is self-sustaining, thanks to yearly dues from its MPAs as well as through donation

3. European PID Infrastructure

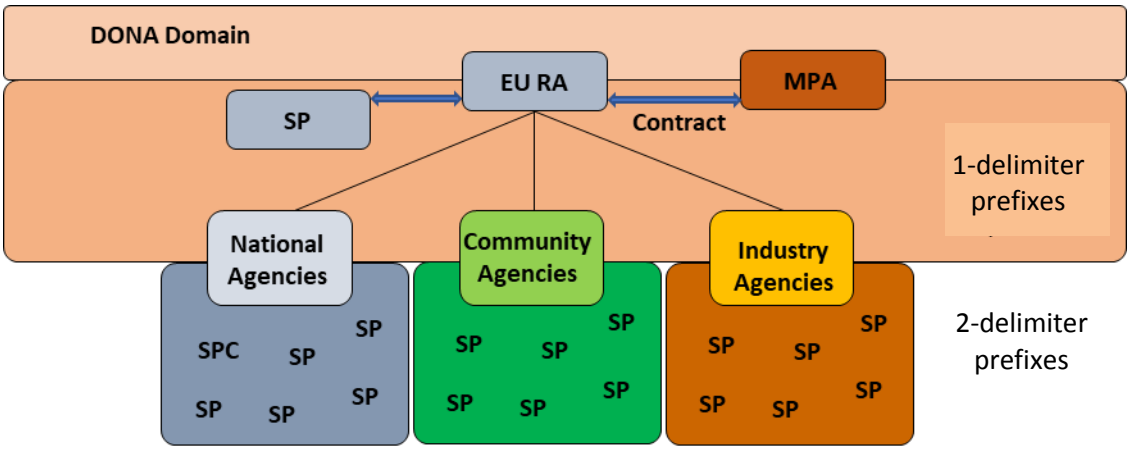
3.1 European PID Organisation

In this suggestion for a possible European PID Organisation we distinguish between organisational structures and service providers. We suggest to first describing a possible organisational structure for Europe.

Embedded in the global Handle Network there should be a European organisation that

- organises and coordinates the European PID landscape,
- overcomes possible conflicts in its domain,
- deals with requests for 1-delimiter prefixes
- making contracts with MPA(s) to allocate and maintain 1-delimiter prefixes
- establishes processes incl. those that lead to nominations for the DONA board
- eventually makes technical specifications
- does certification of all its service providers in their realm
- defines a business model
- offers training courses and carries out dissemination
- organises user support

The organisational structure is therefore fairly simple as shown in the diagram.



A European RA in collaboration with the corresponding MPAs would control a Credential-Prefix. There will be reasons that European countries may decide to establish a nation-wide service according to their business/governance level, nevertheless being part of a global resolution service, and thus request a 1-delimiter prefix as well as some communities. A few examples for such communities currently existing in Europe are:

- the ePIC community including the GWDG, EUDAT and other service providers in Europe offering also national PID services
- the DISSCO biodiversity community with an own service
- the Climate Modelling community with an own service

All these agencies in Europe are free to define their own technical specifications and business models as long as they adhere to the Handle specifications. Due to the distributed nature of data science we cannot prevent that there will be some competition between agencies. But some competition will prevent that monopolies will emerge and will keep the costs at a low level. It would be the task of the EU-RA to prevent conflicts that could lead to losing trustworthiness and to define minimal overall Quality of Service requirements.

3.2 Service Providers

National, community and industrial agencies do not have to operate their own handle services. They can outsource operations for example to a centre that offers a multitude of VMs for this purpose. An example is the ePIC community that acts as registrar, currently also as registration authority and as service provider with a few national data centres as nodes in the EUDAT network maintaining copies of all PIDs. All service providers could in principle develop their own local Handle software as long as they adhere to the specifications.

The landscape of services providers will evolve over time along the changing requirements defined by the developments in data science. The EU RA needs to define rules and processes to guide the developments.

Prefix Requesting and Allocation

The procedure for requesting a 1-delimiter prefix would include the following steps:

- request a 1-delimiter prefix from the European Registration Authority.
- this will be granted after having checked the application and a prefix is being assigned
- the prefix will be sent to the MPA as registrar to do proper allocation and maintenance

3.3 2-delimiter Prefixes

The national, community and industrial agencies that were assigned a 1-delimiter prefix can now address their respective community in optimal ways, i.e. assign 2-delimiter prefixes or offering

Handle registration and resolution services according to some technical specifications and a business model.

3.4 Overall European Landscape

If a European system is being set up as described above the overall Handle landscape is defined by the following services:

- The **European Registration Authority** would offer 2-delimiter services to interested European countries, communities and industry. Their services can be characterised by
    - control and guidance by their corresponding user communities
    - independent choice of their governance model
    - independent choice of technical specifications and business models
    - independent choice of their service providers
    - choice of Quality of Service and reliability constraints, as long as they conform to the minimal overall requirements

- The **IDF foundation** which specified the DOIs is acting globally and supports currently in particular 2 agencies issuing DOIs: **CrossRef** and **DataCite** - both acting also globally. Their original dedication was to serve electronic publications, which was extended to data publications. Their services can be characterised by:
    - agreements on technical specifications with some restrictions
    - agreements on business models
    - implementation of a specific service provider model
    - specifying generic metadata schemas (DC like) for different objects
    - offering added-value services based on metadata searches such as linking between publications and data, etc.

As mentioned earlier one cannot and should not set hard borders between the different service providers, which implies that there will be some competition. The future evolution needs to clarify which structures will survive.

4. Interoperability

Some comments are made on interoperability aspects in the domain of PIDs.

Basic Interoperability

Basic interoperability is given for the whole landscape due to the use of Handle specifications including the Handle syntax. Any Handle resolver will find the appropriate local server easily to resolve a Handle or DOI, i.e. Handle with prefix 10. The specific syntax chosen for the suffix is transparent to higher levels and will be resolved by the local service. Proxies to make Handles web-actionable should also work for all Handles independent of their prefix.

Kernel Information Types

The next level of interoperability is defined by the kernel information the Handle is resolved to. The result of the resolution is a structured set of attributes which can be defined and selected by the local Handle service providers. At this level we need to offer maximal freedom, since the Handles may be used for example by institutions or industry for very specific purposes. At the level of **FAIR Digital Objects** it is obvious that we need to find agreed definitions of some core attributes, register them and request its usage. Now even machines will be able to interpret the attributes. Within RDA some Kernel Information Types have been identified as crucial and been defined. The FDO Framework will further push the definition of such types, request their usage to be interoperable and

FAIR compliant and will request to use an accepted type registry to register them. The ePIC (GWDG) is already maintaining such a registry.

It will be the specification of the FAIR Digital Object Framework (FDOF) that will lead to a harmonisation of approaches with the Linked Data Platform community. Other initiatives might follow the FDOF specifications to increase interoperability and machine actionability.

Easy Transfer
Data science oriented institutions creating many digital objects often use Local Handle Systems to cope with the mass of the PIDs being registered and resolved at high granularity to enable workflows to work on specific collections. Some of them aggregate such digital objects to large collections which are assigned DOIs to support citation and linking. Most of them simply register Handles and store their data in trustworthy repositories and see this as a "publication step". Whatever strategy is applied, it should be made easy for users to make this transition to register DOIs for final curated collections.

Added Value Services
Much effort is currently put into added-value services such as linking a variety of different Digital Object types such as publications, data, projects, organisations, etc. Metadata schemas are being developed and associated with DOIs such as by DataCite. Different actors are engaged in this area which led to a fragmentation.

- CrossRef and DataCite rely on DOIs and its close combination with minimal metadata for building the linkages. This approach is mainly based on published objects where DOIs are being established. Minimal metadata is available for publications and data. Some groups are working on metadata schemes for organisations and instruments (FREYA).
- euroCris is relying on the CERIF metadata approach that offers a rich framework for contextual semantics, i.e. allowing to describe projects, funding, organisations and researchers. This approach is supported by many universities and countries in Europe.
- The scientific community where most of the data is being stored increasingly often uses Handles to assign to all kinds of digital objects. They are using rich metadata schemes to describe properties of their digital objects which includes for example instruments and instrument configurations. But this community is less organised.
- Some funding organisations such as the EC put a considerable effort in maintaining a proper database about organisations that can participate in European funding actions.

A European strategy to overcome this fragmentation is urgently requested, but should be independent of the PID ecosystem. Linking operations needs to include all digital objects registered in trustworthy repositories and include the rich information which is already available. There should be no barriers, information registries need to be independent of economic interests and double work should be prevented. But this paper will not elaborate on these aspects since we believe that continuously new technologies will lead to improved linking operations and smart services as long as the information registries are open.