Recommendations on PIDs extracted from Documents

Peter

In this note I simply collect statements on PIDs from some relevant documents. These aggregated set of recommendations may help us to define a set of general policy statements.

# FAIR Principles (2015)

Here I only mention those principles that refer to PIDs directly. It should be noted that the FAIR principles stress that everything needs to be machine actionable to enable automatic processing for example to assess FAIRness.

[**F1**. (Meta)data are assigned a globally unique and persistent identifier](https://www.go-fair.org/fair-principles/fair-data-principles-explained/f1-meta-data-assigned-globally-unique-persistent-identifiers/)

[**F3**. Metadata clearly and explicitly include the identifier of the data they describe](https://www.go-fair.org/fair-principles/f3-metadata-clearly-explicitly-include-identifier-data-describe/)

[**A1**. (Meta)data are retrievable by their identifier using a standardised communications protocol](https://www.go-fair.org/fair-principles/542-2/)

[**A1.1** The protocol is open, free, and universally implementable](https://www.go-fair.org/fair-principles/a1-1-protocol-open-free-universally-implementable/) (*here several protocols are meant including the protocol to resolve PIDs, for example)*

From the FAIR principals we can derive a few core issues. Using PIDs for ALL digital objects will become a MUST within a few years and scientists will need to adapt to have the chance to apply for funds. This will change the

* quantity of PIDs being generated dramatically and thus also is a new quality requiring proper organisational/administrational ways
* number of researchers and institutions that want resp. need to participate extremely including the HPC centres which currently believe they are far away from digital objects
* level of support extremely since currently one can assume that those who know already participate, but afterwards others not informed will join.

# Extended FAIR Digital Object Requirements (2019)

Intensive discussions between GEDE and GOFAIR resulted in extended specifications for FAIR Digital Objects that would meet the requirements from FAIR Maturity Indicators as they are emerging. Here only those requirements are mentioned that directly refer to PIDs. Focus is of course on enabling automatic processing. These requirements will now be subject of meetings in the coming two months in Washington, Beijing, Helsinki and Paris to achieve global consensus. In October we expect an official manifesto to express the agreements.

* A PID (GUPRI[[1]](#footnote-0)) is assumed to be the basis of the Internet of FAIR Data and Services (IFDS).
* Such a PID is resolved to a structured record with state information the attributes of which are semantically defined and registered in a FAIR type registry and/or include explicit semantics.
* The structured record includes at least a pointer to the locations where the bit-sequences encoding the content of the DO can be accessed, a PID of the metadata DO and the DO type(s).
* Each Digital Object identified by a PID can be accessed or operated on using an interface protocol by specifying the PID of an operation.
* The relations between FDO types and operations are maintained by a type ontology potentially held in a type registry.
* "Deletion" of an FDO leads to standardised and thus machine interpretable tombstone notes in the metadata and the PID, i.e. only the bit-sequences will have been deleted.

# GEDE PID Document (2017)

The GEDE group in which 47 ESFRI/ERICs collaborate worked out and endorsed a document on PID Usage which also has been endorsed by RDA. It can be found here: <https://github.com/GEDE-RDA-Europe/GEDE/tree/master/PID>. The major recommendations are as follows:

* A PID is uniquely identifying a specific Digital Object and is supported by a persistent resolution system that resolves PIDs to machine actionable state information[[2]](#footnote-1).
* A PID System must support an address space larger than the maximum number of identifiers required in the foreseeable future
* A PID System must support secure access mechanisms to prevent tampering.
* A PID System needs to employ a redundant and secure architecture and support open standards such as the ITU X.1255 interoperability guidelines.
* A PID needs to be actionable on the web, by extending it to a fully defined URI, if required[[3]](#footnote-2)
* A trustworthy and sustainable PID registration and resolution system is subject of regular quality assessments and needs to be managed by clearly specified registration authorities engaging a team of experts and being backed by organisations with long-term perspective[[4]](#footnote-3)
* A PID System's registration and resolution services must ensure 24/7 availability, and provide openly documented APIs, optimally supporting accepted data models.

# RDA Europe Conference on PIDs (2016)

To this conference stakeholders from many relevant initiatives who were active in the PID scenario were invited (CNRI, CrossRef, IDF, ITU, some ESFRI initiatives, RDA etc.) to come to statements everyone could finally agree upon. Here only those statements are mentioned that refer to PIDs.

* Proper PID usage and support will become key for competitiveness in science and industry.
* PIDs need to be used by all parties dealing with data professionally to make full use of advanced opportunities. A PID centric approach to data management, access and use will open the way towards new and comprehensive way of data handling and finally to a Global Digital Object Cloud[[5]](#footnote-4) as a generic, non-proprietary virtualisation layer.
* International and national steps need to be taken urgently to offer a sustainable, structured and mature PID service landscape based on quality assessed service providers to all interested parties.
* Only such a structured and massive approach will prevent ending up with unresolvable PID zombies.[[6]](#footnote-5)
* PIDs are becoming essential across sectors and communities for different application scenarios and efforts need to be taken to offer services across these sectors and communities.
* We need to design the required mechanisms and build the needed tools now with high urgency[[7]](#footnote-6).
* We urgently need to come to a structured and integrated domain of Handle Service Providers.[[8]](#footnote-7)
* Service providers need to ensure that these two interoperable domains are part of one integrated landscape of rich services[[9]](#footnote-8).
* The PID centric approaches that are key to manage the data Tsunami require simple and clear messages for the users.
* RDA needs to continue to play its integrative role in the data domain.

# PLOS Paper (2017)

The PLOS paper from colleagues mainly engaged in the biomed area describes lessons learned and some characteristics. According to PLOS a URI consists of a "URI pattern" and a "local ID"[[10]](#footnote-9). A "Compact URI (CURIE)" consists of a "prefix" and a "suffix".

**Lessons**

* Credit any derived content using its original identifier.
* Help local identifiers travel well: document prefix and patterns.
* Opt for simple, durable web resolution.
* Avoid embedding meaning, or relying on it for uniqueness.
* Design new identifiers for diverse uses by others.
* Implement a version-management policy.
* Do not reassign or delete identifiers.
* Make URIs clear and findable.
* Document the identifiers you issue and use.
* Reference and display responsibility.

**Characteristics**

* One Local ID must be associated to no more than one entity locally. One URI must be associated to no more than one entity globally.
* One entity should ideally be identified by no more than one URI.
* The URI, and by extension the local ID, should wherever possible stay the same over time.
* Identifier must NOT be reassigned to an altogether different entity, though the original entity may evolve provided a change history is documented.
* If the entity's definition or essential metadata changes substantially, (Lesson 7) the identifier should, wherever possible be versioned and/or change history documented.
* The identifier must NOT be deleted (but may be deprecated).
* The URI must be resolvable to a web address where the data or information about the entry can be accessed.
* The local ID and its URI counterpart must be inter-convertible by applying the URI pattern to the local ID. Note that in some communities (e.g., ontologies), the local ID is often a CURIE by default.
* The total set of assignable identifiers for the database must be describable through a formal pattern (regular expression).
* The local ID should wherever possible be of a format that does not need special handling when used in URL and common exchange formats (e.g., XML).
* The identifier should ideally be assigned at no cost to individuals depositing data in a repository.
* The identifier and its label should be able to be transparently referenced and actioned (e.g., in a public index or search) anywhere by anyone and for any reason. Restrictions on associated data may apply but are not recommended.
* The identifier scheme should be documented.

# IOT and Data Challenges (2017)

At the IoT week 2017 a workshop was done on "Globally Interoperable IoT Identification and Data Processing" bringing together people from big data and industry. Two major questions were discussed:

* How do data creation, organisation and management be changed in a way to make data intensive projects in science and/or industry much more efficient and thus enable re-use effectively? What is the right level for global intereoperability and which steps could stimulate re-thinking? Is the Digital Object Architecture which puts global Persistent Identifiers (PIDs) that are resolved to useful property information in the core a good kick-off point?
* How can an open data forum be created where data is offered to foster exchange/trading and how does it need to be structured? How can data collectors in particular in industry be convinced to respect the rights of all stakeholders and based on clarified rights became willing to trade data on such a forum.

The workshop showed that IoT industry is heterogeneous using many different ID schemes for various purposes and new efficient transport protocols. Data management did not play such a prominent role, since companies have their proprietary solutions. Recent discussions with large consortia (IDS, BDVA) and ITU indicate that the pressure for interoperable solutions for data management&reuse is increasing and that new solutions are required. While ITU decided to use Handles for big projects for example for a global ship tracing system, various schemes, such as GS1, are still being discussed.

# Turning FAIR into Reality Report

This report of EC's expert group[[11]](#footnote-10) is also of relevance, since it includes quite a number of statements on PIDs. Here, I only present a selection of statements associated with PIDs, although there is much more useful information. It is obvious that the report looks from the view point of FAIR DOs.

* Central to the realisation of FAIR are **FAIR Digital Objects**, which may represent data, software or other research resources. These digital objects must be accompanied by persistent identifiers, metadata and contextual documentation to enable discovery, citation and reuse. Data should also be accompanied by the code used to process and analyse the data.
* As observed above, the atomic entity for a FAIR ecosystem is a FAIR Digital Object, generally comprising data, a persistent identifier, metadata conformant to standards, and code when relevant. Openly-specified persistent identifiers and persistent resolution systems available at a global level can create a global domain of registered FAIR Digital Objects as a precondition for the Findability, Accessibility, Interoperability and Re-use of data. Using persistent identifiers introduces a step of indirection[[12]](#footnote-11) that requires maintenance, but is necessary to support stable references in a global virtual data domain in which data locations will change, in which copies and versions will be created and in which provenance information, attached to the persistent identifier, will clarify the versioning history of the data.
* Persistent Identifiers are assigned to many aspects of the ecosystem, including data, institutions, researchers, funders, projects and instruments. The PIDs are indexed and used by several components to interlink relevant information and provide context.
* Stable PIDs allow referencing to digital objects, for example in automatic workflows or citations in publications. State information associated with PIDs allows users to check (even after many years) whether the bit sequences have been changed since registration or whether the digital object is mutable or not.

# ODIN Project (2013)

The ODIN project proposal highlighted three main threats or “items of unfinished business” emanating from lack of recognition of the need for robust ways of identifying contributors and their data in e-Science:

* Inability to follow interconnections between datasets and contributors as a method of data discovery.
* Inability to share and connect identifiers of contributors and authors between different user communities.
* Inability to uniquely identify datasets attributed to a particular contributor and contributors to a particular dataset.

So ODIN addressed the issue of how to easily link between the domain of data and the domain of authors/creators both being characterised by identifiers and metadata descriptions. For persons the ORCID schema is widely agreed and for data DOIs and the DataCite metadata schema is assumed. It is the metadata interoperability that is addressed in particular. Also the relation to LD was addressed, i.e. exporting metadata assertions as RDF.

# FREYA Scoping Notes (2019)

The FREYA project recently submitted the "Scoping a PID Policy for the EOSC" paper which I try to summarise in assertions as well. They are addressing more abstract points than for example mentioned by other papers. Baseline statements taken from the "Turning FAIR into Reality" report are that (1) digital objects should be associated with PIDs and that PID services need to be available to use PIDs in EOSC. 6 assertions are made:

* Definition of terminology is required for EOSC to overcome heterogeneity.
* Classes of digital objects which are essential for EOSC and which should be associated with PIDs need to be defined.
* We need to recommend the properties of sets of PIDs that make them valid PIDs in EOSC.
* Recommended core PID services need to be defined that are essential for EOSC. There will also be desirable services that will enhance the functionality of EOSC.
* PID Authorities and PID Service Providers govern the PID services and criteria need to be worked out to make them trustworthy actors within EOSC.
* A discussion on governance of PID services needs to be started considering the global dimension and sustainability.

# DONA Foundation (since 2015)

For those who are not involved so much in this domain it might be helpful to sketch briefly the DONA Foundation which was established as an independent, non-profit Swiss Foundation in Geneva to guide the development of the Digital Object Architecture of which the Handle System as an essential part. It is guided by an International Board having members of many regions (US, CN, RU, EU, AF, etc.). It also includes experts from ITU, Int. DOI Found., and big data labs.

With respect to the Handle System the DONA Foundation takes care that the Global Handle Resolution System is running 24/7, reliably and stable. This system is self-funded and is based on a redundant network of currently 10 regularly checked Multiple Primary Agencies located worldwide. It also specifies the Handle Syntax, the resolution protocol and it offers software to run local handle resolvers (currently 4000 worldwide). People can build their own Local Handle Software as long as they adhere to the specifications. Every institution can request a prefix and run a local handle resolver connected to the GHR, but it has to take care about business models, persistence, redundancy, etc. The costs associated with a prefix request may vary between registration authorities (currently the MPAs). The costs for registering PIDs depend on the service provider such as DOI or ePIC, etc.

The Handle System is independent of a protocol system, i.e. it is not dependent on HTTP. However, it offers a proxy service to make it web-actionable. Using the provided local Handle system makes it powerful for registration as well as resolution. For some big data jobs, however, labs are using queuing systems to master the registration.

There is a pressure to have PID service providers per country and DONA intends to have some more MPAs in future.

# Relevant RDA Groups (2013-2019)

There were a couple of RDA working groups that are relevant in the context of PIDs and produced concrete outputs (here mentioning only those that produced concrete recommendations[[13]](#footnote-12)):

* **RDA Data Foundation & Terminology** that produced the Core Model which defines Digital Objects as basis for data management. <http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318>
* **RDA PID Information Types** that started a first discussion on formal attributes describing properties of DOs. <https://www.rd-alliance.org/groups/pid-information-types-wg.html>
* **RDA PID Kernel Information** that defined attributes that can be associated with PIDs. <https://www.rd-alliance.org/group/pid-kernel-information-wg/outcomes/recommendation-pid-kernel-information>
* **RDA Kernel Information Profile Management** that is defining how profiles of Kernel attributes can be managed. <https://www.rd-alliance.org/groups/pid-kernel-information-profile-management-wg>
* **RDA GEDE** that created recommendations for PID usage. <https://zenodo.org/record/1116189>
* **RDA GEDE DO** started a broad interaction to work out details of Digital Objects and recently FAIR Digital Objects. <https://github.com/GEDE-RDA-Europe/GEDE/tree/master/Digital-Objects>
* **RDA Data Citation** making statements of how to assign PIDs to dynamic data. <https://www.rd-alliance.org/groups/data-citation-wg.html>
* **RDA Data Type Registries** defining a model for defining types and relating them with operations. <https://www.rd-alliance.org/groups/data-type-registries-wg.html>
* **RDA Persistent Identification of Instruments** that wants to work out a model how to describe instruments b metadata. <https://www.rd-alliance.org/groups/persistent-identification-instruments-wg>

1. It should be noted that we mean "globally unique, persistent and resolvable identifier" when we say PID. There are so many different local ID schemes in any database and cloud system which are not of relevance here. This for example also excludes URNs as they are used by some national libraries. [↑](#footnote-ref-0)
2. This implies FAIR compliance. [↑](#footnote-ref-1)
3. This is admitting that PIDs should be protocol independent, since these will change over time, but nevertheless support current mainstream. [↑](#footnote-ref-2)
4. This directly addresses the quality of services that are expected. [↑](#footnote-ref-3)
5. This now evolved to the FAIR DO community which is globally organised. [↑](#footnote-ref-4)
6. This refers to the deprecation of LSIDs, the eminent problems of PURL and the disappearance of some other PID systems. [↑](#footnote-ref-5)
7. This refers to added-value services using PIDs where CrossRef, for example, is very active. [↑](#footnote-ref-6)
8. The DONA Foundation has been setup to organise this. [↑](#footnote-ref-7)
9. This refers to the separate worlds of DOIs and "other" Handle Services. Since the prefix "10" is hardcoded in the CrossRef software, there is an interoperability gap making this integration not easily possible. [↑](#footnote-ref-8)
10. This definition makes the PID dependent on the web-protocol. Proper identifiers need to be independent of a protocol since they will change over time. An ISBN number does not point to a shelf where a book copy is stored. [↑](#footnote-ref-9)
11. <https://publications.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en/format-PDF/source-80611283> [↑](#footnote-ref-10)
12. References do not specify a location, but an identifier that points to a location. When locations are being changed, only the location information associated with the identifier needs to be changed and not all the references, which would be impossible. [↑](#footnote-ref-11)
13. There may be some missing, please add short descriptions. [↑](#footnote-ref-12)