GEDE

Group of European Data Experts

Persistent identifiers:

Consolidated assertions

Status of November, 2017

Editors:

Peter Wittenburg (RDA Europe), Margareta Hellström (ICOS), and Carlo-Maria Zwölf (VAMDC)

Co-authors:

Hossein Abroshan (CESSDA), Ari Asmi (ENVRIplus), Giuseppe Di Bernardo (MPA), Danielle Couvreur (MYRRHA), Tamas Gaizer (ELI), Petr Holub (BBMRI), Rob Hooft (ELIXIR), Ingemar Häggström (EISCAT), Manfred Kohler (EU-OPENSCREEN), Dimitris Koureas (NHM), Wolfgang Kuchinke (ECRIN), Luciano Milanesi (CNR), Joseph Padfield (NG), Antonio Rosato (INSTRUCT), Christine Staiger (SurfSARA), Dieter van Uytvanck (CLARIN) and Tobias Weigel (IS-ENES)

RDA

RESEARCH DATA ALLIANCE

The Research Data Alliance is supported by the European Commission, the National Science Foundation and other U.S. agencies, and the Australian Government.

Document revision history

2016-11-01	v 1.0 (extracted from RDA Data Fabric IG wiki)
2016-12-15	v 2.0, after comments & feedback from Bratislava 2016-11-14
2017-03-29	v 3.0, partially discussed at RDA P9 in Barcelona 2017-04-07
2017-05-10	v 4.0, discussed during GEDE telco on 2017-05-29
2017-06-21	v 5.0 for GEDE PID group review on 2017-06-28
2017-08-29	v 5.1 for GEDE PID group review on 2017-08-30
2017-11-13	v 6.0, for GEDE review on 2017-11-20
2017-12-13	V 6.1, circulated for RDA (Data Fabric IG) review

Abstract

Experts from 47 European research infrastructure initiatives and ERICs have agreed on a set of assertions about the nature, the creation and the usage of Persistent Identifiers (PIDs). This work was done in close synchronisation with the RDA Data Fabric Interest Group (DFIG) ensuring a global validation of the assertions. The intention of this cross-disciplinary report is to overcome still existing confusions about PIDs and the lack of detail knowledge in many disciplines. It is not meant to produce yet another comprehensive document on PIDs, but to identify agreements across documents that have been suggested to be included by experts. With this document GEDE is happy to help demystify PIDs, overcome confusion and create bridges between the various disciplines.

About GEDE

The aim of the Group of European Data Experts in RDA (GEDE-RDA) is to promote, foster and drive the discussions and consensus relating to the creation of guidelines, core components and concrete data fabric configurations, based on a bottom-up process. To achieve these goals GEDE-RDA is composed of a group European data professionals appointed by invitation from various research and e-Infrastructures and European co-chairs of Research Data Alliance (RDA) Groups. GEDE-RDA will operate within the global RDA framework, thereby guaranteeing that discussions are openly communicated and publicly accessible to the global community of experts – RDA members. For more information, see the group's web pages at https://www.rd-alliance.org/groups/gede-group-european-data-experts-rda.

Acronyms used in this report

Term	Definition
API	Application Programming Interface, a computing term. See e.g. https://en.wikipedia.org/wiki/Application_programming_interface.
CESSDA	Consortium of European Social Science Data Archives, a European research infrastructure. See https://www.cessda.eu/.
CLARIN	Common Language Resources and Technology Infrastructure, a European research infrastructure. See https://www.clarin.eu/.
CMIP	Climate Model Intercomparison Project, see Sources section.
CORBEL	Coordinated Research Infrastructures Building Enduring Life-science Services, a European research project. See http://www.corbel-project.eu/.
CURIE	Compact URI. See assertion PID-10.
DFIG	RDA Data Fabric Interest Group, see https://www.rd-alliance.org/group/data-fabric-ig.html.
DFT	RDA Data Foundation and Terminology group, see https://www.rd-alliance.org/groups/data-foundation-and-terminology-wg.html.
DO	Digital Object.
DOI	Digital Object Identifier, https://en.wikipedia.org/wiki/Digital_object_identifier.
DSA	Digital Seal of Approval – see https://www.datasealofapproval.org.
DTR	Data Type Registry. See http://typeregistry.org/registrar/.
ENES	European Network for Earth System modelling, see https://portal.enes.org/
EUDAT	A European collaboration of data service providers, see https://eudat.eu/.
FAIR	Findable, Accessible, Interoperable, Reusable – see FORCE11.
FORCE11	International group working on research data issues. See https://force11.org/.
GDOC	Global Digital Object Cloud, see https://www.rd-alliance.org/group/data-fabric-ig/wiki/global-digital-object-cloud.
GEDE	Group of European Data Experts (in RDA Europe). Authors of this report.
HTML	HyperText Markup Language, see e.g. https://en.wikipedia.org/wiki/HTML.
ITU	International Telecommunication Union, see http://www.itu.int/.
LSID	LifeScience Identifier, see Appendix 1.
MD5	A hash function producing a 128-bit hash value often used as a checksum, see https://en.wikipedia.org/wiki/MD5.
NBN	National Bibliography Number, https://wiki.surfnet.nl/display/standards/URN-NBN.
ORCID	Non-profit organization issuing persistent identifiers for individuals and organizations, see https://orcid.org/.
OWL	Web Ontology Language, https://en.wikipedia.org/wiki/Web_Ontology_Language.
PID	Persistent (digital) Identifier.
PIT	Persistent Identifier data Type, read more at https://www.rd-alliance.org/group/pid-information-types-wg.html.
RDA	Research Data Alliance, see https://www.rd-alliance.org/.
RDF	Resource Definition Framework, https://en.wikipedia.org/wiki/Resource_Description_Framework.
TDWG	Taxonomic Database Working Group, see http://www.tdwg.org/.
URI	Uniform Resource Identifier, see https://en.wikipedia.org/wiki/Uniform_Resource_Identifier
URL	Uniform Resource Locator, https://en.wikipedia.org/wiki/URL.
URN	Uniform Resource Name, https://en.wikipedia.org/wiki/Uniform_Resource_Name.
WDS	World Data System. See https://www.icsu-wds.org/.

Contents

1 Why and how this document came about

Despite 20 years of discussions about PIDs, there still exists much confusion and lack of knowledge about persistent identifiers (PIDs) in many research communities. In part, this is the consequence of the vast number of documents and reports on the PID topic in existence, many of which focus on very specific aspects, or use different and sometimes incorrect terminologies.

To help bring more clarity, it was therefore agreed in GEDE and the RDA Data Fabric Interest Group (RDA DFIG)[1] to form a focus area group[2] on PIDs, which would exist for a limited time

As output, the group should not produce yet another document with new ideas about PIDs, nor aim to write a comprehensive document about PIDs covering all aspects. Instead, the end result should be a compilation of relevant assertions (extracted statements from original sources), supplemented with context information and explanations as required, and leading up to a summary of agreements and disagreements reflecting the convergence of understanding.

Some of the milestones of the GEDE work on PIDs include:

1. The work was initiated in late 2016 with a bootstrap document summarizing some recent discussions and work called the Persistent Identifier Bundle which was spread within GEDE and DFIG[3].
2. At the GEDE fact to face meeting in Bratislava in November 2016, this document was discussed and commented on. Participants referred to other documents of relevance, from their communities or those with a broader relevance, and it was agreed to extend the bootstrap document with assertions collected from these new sources. The focus area group was formed to aid in this process.
3. At the RDA plenary P9 in Barcelona in April 2017, an updated and extended paper was presented and discussed in the realm of RDA DFIG.
4. Based on comments and suggestions from the P9 participants, the PID focus area group continued its work on refining the document during the summer and autumn of 2017, helped by a series of video conference calls. During this revision process, it was decided to annotate the collected assertions with comments and explanations in order to make the report more accessible also to non-experts. In addition, several appendices were added.
5. In November 2017, the near-finalized document was discussed in a video conference with the entire GEDE group. After taking into account their comments, the present version of the report (v 6.2) was produced by an editing team consisting of the GEDE co-chairs.

[1] https://www.rd-alliance.org/group/data-fabric-ig.html

[2] Similar to an Agile Task Force.

[3] http://doi.org/10.23728/b2share.8c6d56318ebd4914b359606051c128f6 and http://hdl.handle.net/11304/fcec9990-0df3-40b7-a4a0-f4853c3c36d3

2 Convergence of definitions and understanding

In this section we summarise the essentials of all the collected assertions, which are listed below in Chapter 3. To simplify intercomparisons, this chapter follows the same subdivision into subtopics as used for the assertion collection. **Note:** For more detailed definitions of the terms used here, please refer to the assertions and our comments to these, as well as Appendices 1-4. (As an example, "digital object" is defined in assertion PID-1.)

2.1 Nature of PIDs and PID systems

PIDs are increasingly important and are being applied almost everywhere across sectors and disciplines, and for all types of digital objects. (Here, the term "sectors" covers science, industry, governments, health care, etc.) Data management experts are becoming increasingly dependent on the availability of functioning persistent identifiers which:

- are uniquely identifying a specific Digital Object,
- in general consist of a name space indicator (prefix) and a local identifier (suffix),
- are actionable on the web, by extending it to a fully defined URI, if required
- can be persistently resolved to state information and/or a landing page (see appendix 3),
- are associated with a persistent resolution system (see appendix 4),
- are issued and managed by a clearly specified registration authority.

2.2 Relevance of PIDs and PID Systems

In order to be useful and reliable, PID registration and resolution systems need to be trustworthy and sustainable. Pre-requisites include the need for the systems to be authorised and maintained by dedicated and reliable teams, backed by organisations that have a long-term perspective and are based on transparent, sustainable business models. The organisations should be governed by boards that guarantee proper strategies for the future, and be the subject of regular quality assessments by external parties. The technical platforms supporting registration and resolution services should also employ a redundant and secure architecture, as well as support open standards. In addition, exit strategies should be thought of, as well as redundancies, and embedded into the system architecture at all levels.

The systems, as a whole, should support and encourage federations between multiple partners, as well as be both scalable and resilient to unforeseen abrupt changes in the organisational structure or funding. Experience shows that only cross-sector and cross-disciplinary systems such as Handle, DOI, etc. will have a chance to survive due to the enormous investments required including the support of layered services. For DONA, responsible for the future of the Handle System (incl. DOIs) it has been calculated that financial sustainability of the distributed Global Handle Resolver is given when at least 12 Multi Primary Administrators share the efforts and costs. Currently 10 have been identified and are globally distributed.

PID systems need to be configured to support vast numbers of objects, and must therefore support an address space larger than the maximum number of identifiers required into the foreseeable future.

Both registration and resolution services must ensure 24/7 availability, and provide openly documented APIs, optimally supporting accepted data models. The ITU X.1255 interoperability guidelines[4] should be followed, including for example adherence to agreed specifications of types being used in the PID registry (also known as the PID metadata kernel) to enable machine

[4] See https://www.itu.int/rec/T-REC-X.1255-201309-I/en.

processing. Access mechanisms and user interfaces must be secured against tampering. As an example, only PID owners should be able to change crucial information in the PID record. PID resolution systems should follow the protocols and procedures defined by the relevant PID policies, and be capable of tracking PID resolution requests.

2.3 Assigning PIDs

The assertions in this section discuss questions such as when and how to assign PIDs, and what the importance of the assignment is.

Note: Many of the collected assertions in section 3.3 refer to the Life Science ID (LSID) URN-based PID system (see appendix 1). As LSIDs are currently deprecated, this section mainly ignores these statements. However, many of these assertions (e.g., assertion PID-44) contain what is in themselves very useful and practical information with respect to their URN choice (in particular the specs for "namespace", "object" and "revision").

When assigning PIDs it is important to consider who has the responsibility and the administrative rights for first registering the PID and later to make updates if needed. It is important to adhere to policies and rules for PID creation that are valid for the particular PID service that is being used. Regarding the format of the identifier, the recommendation is to not use semantics in the PID string. It is also important to make it robust and stable, for example by avoiding the use of string characters that may have special meaning in the protocols used, or not be interpretable by browsers.

Details on best practices for building PIDs are discussed in appendix 1 whereas appendix 3 gives examples of metadata and state information (together with their storage strategies) that can be associated with a DO. The PID form that should be used for e.g. citations should, in order to be fully *machine actionable (cf. assertions PID-10 and PID-54), include its* actionable extension (full URI, see appendix 1).

In order to allow automatic metadata processing and data analysis, the state information retrieved when resolving a PID should follow a well-defined format and defined semantics. The PID registration service must provide clear and unambiguous documentation of this format. Optimally, only specific attribute types, well documented in public data-type registries, should be used. One of the standard attributes is the location of the DO's bit sequence. In addition, the PID record or the landing page (typically XML structures that have machine readable sections, see appendix 3) should include attributes informing the user about the expiration date of the DO and its mutability.

With respect to the timing of PID assignment and the issues of granularity and versioning, there are different practices across domains and PID service providers. Repositories (and other data providers) need to make clear which policies they apply, particularly with respect to the recommended timing of assignment, and the granularity needed. However, a few rules of thumb can be given, including 1) In general a DO should be assigned a PID as early as possible, at least at the time it is uploaded into a trustworthy repository; 2) In order to optimally support citations and later re-use, PIDs should in general be assigned to the smallest “chunks” of scientifically meaningful digital information (“data”) that is practical to refer to. This often translates into a high granularity, but there are many exceptions where it is desirable to assign a PID to e.g. large data sets or to collections.

Some repositories support specific options and functionalities. These include 1) version indicators[5] in the attribute list or on the landing page, including support to allow machines to find and retrieve specific versions; 2) fragment indicator extensions (which are appended after the PID itself) that point to specific subsets embedded in the DO referred to; 3) embedding strings that are currently not globally resolvable (such as specific URNs) in the Handles as a suffix and thus making them resolvable; and 4) support for allowing typical life-cycle actions such as deleting, updating and splitting content. In all cases, repositories need to state clearly what their policies are.

2.4 Using PIDs

This chapter is more related to the inclusion of PIDs in catalogue metadata and the resolving of PIDs into actionable state information. In data management, cataloguing systems are typically used to compile metadata about collections of DOs, including both administrative and descriptive information. However, the specifics of which metadata to associate with a DO, and where to store it (in the PID record, in a metadata description serving a landing page, or in other separate locations) will always be situation-specific, and cannot be covered here. For an example, see Appendix 3.

The metadata are typically used to support searching by both human and machines, so to make objects "findable"[6], the schemas used should clearly indicate which attribute contains the PID relating to each catalogued object. By resolving the PID, the DOs' bit sequences covering the DOs' content can then be retrieved. When resolving a PID associated to a Digital Object, a user (both human and/or computer system) must have access to sufficient information for retrieving the content of the digital object - either directly or via a landing page (see appendix 3).

PIDs are a key element for the citation and/or data citation mechanisms, as stated in point 4 of the Force 11 joint declaration (see Footnote 6) on data citation principles: "A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community". This or a similarly widely agreed citation principle should be used when referencing a DO (compare the assertion PID-54).

The systematic use of interoperable PID systems (not only for data but for all kinds of objects with a digital representation) opens up new application opportunities such as those presented by the Global Digital Object Cloud concept[7]. More information and examples demonstrating how PIDs can be used in practice will be made available in a future publication from GEDE.

2.5 Handles and DOIs

Together with their respective end users, providers of DOIs (DataCite, CrossRef and others) form a sub-community of Handle System users[8], with their own particular assignment practices, metadata schemata and business models. Other authorities are building their own communities, as

[5] Adding version information to the PID (the LSID option) is not recommended.

[6] See the FAIR (Findable, Accessible, Interoperable, Reusable) guidelines from FORCE11, https://www.force11.org/group/fairgroup/fairprinciples

[7] http://hdl.handle.net/11304/a8877a1a-9010-428f-b2ce-5863cec4aff3. See also for example the "Recommendations for Implementing a Virtual Layer for Management of the Complete Life Cycle of Scientific Data" document from the Data Fabric IG, currently under RDA community review (https://www.rd-alliance.org/system/files/DataFabric_SupportingOutput_recommendation-aug-2017-v10.pdf).

[8] In fact, DOI uses only one of the about 3700 prefixes currently available in Handle.

exemplified by ePIC (European Persistent Identifier Consortium) which operates its own registration and redundant resolution services.

While DOIs are mainly assigned to objects and collections that are ready for public dissemination, Handles from other service providers are in general used to persistently identify other categories of digital objects (e.g. those created in the labs) to make them referable by software, workflows, etc.. For example, ePIC PIDs are generally used for referencing DOs which are associated with the various steps of a scientific workflow before considering data publishing. Other Handle-based identifiers are used in research contexts to identify (the digital representation of) physical samples, instruments, measurement platforms and people.

More efforts need to be undertaken to better integrate the DOI and other Handle domains to enable complete interoperability of layered services.

2.6 Communications/discussions/networking with communities, funders, users

These assertions are not directly addressing PID issues, so we refer to the individual statements.

3 Collected assertions and statements

Ignoring fine semantic differences, and instead focussing on the core messages, more than 60 assertions were extracted from the documents listed below (detailed references are in the Sources section at the end). Neither the selection of sources, nor the list of assertions are in any way complete or comprehensive - but the work has been guided and informed by discussions in the RDA Data Fabric Interest Group (RDA DFIG) and in the GEDE initiative.

Table 1. Abbreviations of sources used in this section. Please refer to the Sources section below for detailed references, including links.

Abbrev	Source
RDA DFT	RDA Data Foundation & Terminology WG
RDA PIT	RDA PID Information Type WG
RDA DFIG	RDA Data Fabric IG Discussion
FAIR	FAIR principles developed by FORCE11
PID WS	RDA EU PID Workshop
USE	GEDE document on PID usage (in preparation)
DOI	DOI Foundation documentation
ITU	ITU (International Telecommunications Union) documentation
BIO	Life science community report by McMurry et al.
COR	CORBEL (Coordinated Research Infrastructures Building Enduring Life-science Services) consensus document on providing data on clinical trial participants
CES	CESSDA ERIC (Consortium of European Social Science Data Archives) documentation
CLA	CLARIN (Common Language Resources and Technology Infrastructure) ERIC PID policy summary
CMIP	Climate Model Intercomparison Project 6 (CMIP6) PID implementation plan
LSID	Taxonomic Database Working Group (TDWG) document on Globally Unique IDentifiers (GUID)

The assertions were then compared and summarised where possible to identify overlaps, agreements and disagreements[9]. Some assertions come from PID initiatives such as Life Science Identifiers (LSIDs; see e.g. Appendix 1) which have been discontinued, but they nevertheless include interesting recommendations.

The outcome, listed below, gives a clear indication that there is much agreement across disciplines on PIDs and PID systems and that nuances can be found in the way disciplines and repositories deal with issues such as granulation of PID assignment, using PIDs for versioning, using semantics in the PID strings, etc. Whatever is being decided by infrastructures and/or repositories, it is agreed that they need to make their policies explicit so that everyone, including machines, can interpret the information.

In the following subsections the consolidated assertions are listed, using the following format:

PID-X. **[Source]** {tags and labels}
Statement as extracted from the source document.

> *Interpreting comments by GEDE (as applicable).*

3.1 Nature of PIDs and PID systems

PID-1. **[RDA DFT 1.2]** {definition}

A persistent identifier is a long-lasting ID represented by a string that uniquely identifies a DO and that is intended to be persistently resolved to meaningful state information about the identified DO.

> *In this document we are using the definition as given by the RDA DFT Core Term and Model[10] document. Note that in our context DO stands for digital object, which may be either a document, dataset, piece of software, a service or other similar non-physical entities, or the digital representation of an analogue entity, such as a device, a sample, a person or an organisation. Also see the PID-3 definition below, and Appendix 3, for an elaboration on "state information".*

PID-2. **[RDA DFT 1.3]** {definition}

A PID record contains a set of attributes stored with a PID describing DO properties.

> *For a definition of "PID record", see the DFT source document and Appendix 3.*

PID-3. **[RDA DFT 1.4]** {definition}

A PID resolver (aka PID resolution system) is a globally available infrastructure system that has the capability to resolve a PID into useful, current state information describing the properties of a DO.

> *State information can be interpreted as administrative or systems metadata, including for example checksums, owner, access paths, and references to additional information. Since the set of properties used for describing the state of the digital object may be arbitrarily chosen by the data provider, we interpret this assertion as meaning that some current state*

[9] While scientists are in general trained to focus on differences, here the endeavour was made to focus on the agreements.
[10] DFT Core Terms and Model: http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318

> *information describing properties of a DO are returned by the resolver. Examples of useful state information include an MD5 checksum to proof identity, a Uniform Resource Locator pointing to either the bit sequence belonging to the DO, or to a "landing page" with more information. Also see Appendix 3 for an elaboration about "state information".*

PID-4. [RDA DFIG] {sustainability, trust}

A trustworthy PID system must

- be maintained by a dedicated and reliable team,
- be based on a transparent sustainable business model,
- be provided by an organisation that has a long-term perspective[11],
- be subject of regular quality assessments by external parties,
- be governed by international boards,
- be based on open standards,
- be based on a redundant and secure architecture,
- support a huge address space[12] and
- support an openly documented API optimally supporting accepted data models.

PID-5. [ITU] {interoperability}

The platform for interoperability of heterogeneous identity management systems is an open architecture model which is described in ITU-T X.1255, based on the Digital Object Architecture[13] [...] and is capable to offer interoperability at global level.

> *The formulation of this statement refers to Identity Management Systems in general and is related with digital identities. The term "identity" refers to identification of digital objects in general and not only to the digital information associated to a person or organisation. There will be legacy systems that cannot easily be made compliant to X.1255. X.1255 has currently high impact in industry discussions where different systems such as Barcode systems are compared with the Handle System and where interoperability is a need to reduce costs.*

PID-6. [ITU] {interoperability, security}

The top 5 benefits from the X.1255 platform are: 1) framework to enable interconnection of objects, data, devices and processes, 2) in-built security regime (PKI) and data privilege delegation, 3) multilingual support and access to a variety of type value pairs, 4) enable defining new type value pairs for increased flexibility for new types of services and applications, 5) interoperable with existing identity management systems.

> *The term "platform" in this assertion is defined by the ITU standard, which is concerned about interoperability between PID systems. X.1255 is one consequence of the ongoing discussions about architectures for digital objects.*

[11] The exact formulation is still subject of discussion.

[12] Comparable to, or even larger than, the address space of IPv6.

[13] https://www.internetsociety.org/sites/default/files/ISOC-DOA-Overview-20161025-A4-3_0.pdf

PID-7. **[USE]** {sustainability}

The sustainability of a PID system mainly depends on its relevance and its investments, ensuring sufficient financial support.

PID-8. **[USE]** {sustainability}

To prevent PID Zombies it seems to be wise to define an exit strategy at all levels.

> *"Zombies" refers to a disruption in the service chain, breaking the routing between the PID and its digital object. The main cause of this kind of disruption is a failure of the sustainability model for the relevant resolving systems, leading to errors and failures when systems try to resolves PIDs.*

PID-9. **[BIO 1]** {sustainability, best practices}

If you [are going to] create identifiers, do not do identifiers by yourself.

> *In general, adopt widely used and documented standards and/or services for assigning PID, since your private solutions may not survive and can only be resolved locally. Explicitly describe or refer to the standard you are adopting in your documentation.*

PID-10. **[BIO]** {definition, community-specific}

- An Identifier is a sequence of characters that identifies an entity.
- A Local Resource Identifier (LRI) is an identifier that is only guaranteed to be unique within a single database.
- A Uniform Resource Identifier (URI) is an identifier that is guaranteed to be both uniform and globally unique.
- A CURIE is a compact URI comprised of <prefix>:<LRI>. A full URI is an identifier that also resolves to a webpage containing information about the identified entity.

> *The exact formulation on LRI, URI and CURIE are specific to the context of the document from the bio/life sciences community. In general one can state that a globally unique "prefix" given by an authorised institution to a client defines a name space allowing the client to create local resource identifiers according to its own policies. Making a PID actionable will require making it a full URI that can be resolved in the web. In the elaboration in Appendix 1 it will be indicated that there are only minor principal differences between this statement and how for example the Handle/DOI system is organised.*

PID-11. **[LSID1]** {community-specific, URN-structure}

The authority identification within LSID is used to identify the authoritative source of a set of LSIDs. The following criteria are defined:

- A provider should use a domain name registered to it as authority identification.
- A provider should plan to control the domain names it uses as authority identifications for as long as possible.
- A provider should transfer control of domain names to a successor if the names are forgone.
- Organisations susceptible to name changes should use domain names that will remain effective as authority identifications through reorganisation changes.

- If a suitable domain name is not available or likely to be unstable, request an authority identification from the LSID top instance.

> *LSIDs which were based on an URN scheme are not supported anymore (urn:lsid:authority:namespace:object:revision) making some of their statements obsolete. It was impossible to develop a common resolver that could turn URNs into actions. However this particular assertion discusses sustainability and strategical aspects which may still be valid. For more comments see Appendix 4 and http://dev.mygrid.org.uk/blog/2016/02/what-exactly-happened-to-lsid/.*

PID-12. [CLA] {community-specific, implementation}

- Handle is a system which is performing, scalable and robust enough [for CLARIN]. It offers enough flexibility.
- Centres using for example URNs are suggested two options to make their service compliant: a) a Handle was created that points to the URN:NBN resolver URL, b) URNs are transformed into Handles (example: urn:nbn:fi:lb-20140421 will become hdl:11113.1/20140421).

> *The CLARIN community makes concrete suggestions to their data centres. For details on Handles look here:* https://www.handle.net/*. For those language data repositories that made use of other PID systems, such as URNs, a solution was offered to instead integrate their data into the CLARIN domain and to make* the data resolvable. *Universal Resource Name (URN) is a system for naming a given resource, which may have multiple instances. URNs cannot be used for routing to a specific instance of an object. CLARIN is suggesting a way for building PID resolving mechanism starting from URN. NBN stands for National Bibliography Number*[14]*.*

PID-13. [CLA] {best practice, citation, implementation}

CLARIN endorses the FORCE11 citation principles (https://www.force11.org/datacitation).

> *The FORCE11 principles include "A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community" and "Unique identifiers, and metadata describing the data, and its disposition, should persist - even beyond the lifespan of the data they describe".*

PID-14. [CES] {availability}

Research infrastructures shall use a global PID services that ensure 24/7 resolvability of PIDs.

> *Here the need for high availability of such a system is being stressed, since users need to rely on fast resolution of PIDs.*

3.2 Relevance of PIDs and PID Systems

PID-15. [PID WS] {strategy, policy}

Proper PID usage and support will become key for competitiveness in science and industry.

[14] See https://wiki.surfnet.nl/display/standards/URN-NBN or https://tools.ietf.org/html/rfc3188.

PID-16. [PID WS] {strategy, policy}

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.

PID-17. [PID WS] {strategy, policy}

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.

PID-18. [COR] {community-specific}

Clinical trial datasets should be considered legitimate, citable products of research, which needs persistent identifiers as a prerequisite.

> *Biobanks and their collections of biological samples and/or data should also be considered legitimate identifiable and citable product of research.*

PID-19. [COR] {best practice}

Persistent Identifiers such as the already widely used DOI should be applied to datasets to improve discoverability and to allow correct citation.

> *DOIs are Handles with the prefix 10 and thus form a specific community based on a business model and agreements within the Handle domain.*

PID-20. [COR] {implementation, sensitive data}

The requester [wishing to have access to datasets] should also provide information on his/her expertise, possibly making use of persistent digital identifier systems (e.g. ORCID).

> *Actually "expertise" in this context should be interpreted as "affiliation" and "role", so as to provide a basis for authentication and authorization in connection with accessing the DO in question. The assertions we are gathering discuss the mechanisms for building resolvable routes between identifiers and identified digital objects. Once built, these routes may be restricted to some users. Access rules and privileges required for resolving a particular PID are out of the scope of this document, but this could be very important in case of clinical/medical data.*

3.3 Assigning PIDs

PID-21. [RDA DFT 1.1] {definition}

A digital object (DO) is represented by a bitstream, is referenced and identified by a persistent identifier and has properties that are described by metadata.

> *This can be seen as a core data model that allows organising data in a FAIR compliant way.*

PID-22. [USE] {best practice}

Each meaningful digital object should be assigned a PID to facilitate re-use and recombination.

PID-23. [USE] {granularity}

In particular automatic solutions [for PID assignment] in general require a high granularity.

> *Here, "automatic solutions" refer to both the creation & subsequent registration of DOs and to their subsequent use in research, e.g. by machine-actionable workflows. Processing modules in workflows in general require a detailed typing of the data they can operate on, i.e. rich metadata is required. "Granularity" is here related to resolution and scale of the DO. In a data context, examples of high granularity include treating data collected at the same time, but at different locations and with different sensor types, as individual distinguishable datasets. Correspondingly, by gathering all of those datasets into a single dataset(a "collection"), a low granularity dataset is created.*

PID-24. [FAIR F1] {best practice}

(Meta)data are assigned a globally unique and eternally persistent identifier

> *Here "globally unique" means that a given PID is just assigned to one single specific digital object. One cannot have a PID pointing to two different objects. However a given object may be assigned several different (unique) identifiers. The term "eternal" is debated, but it is widely agreed that if the DO is deleted, the PID should remain and point to so-called tombstone information, including basic metadata about the (now defunct) DO's properties , the reasons for its deletion and, if available, pointers to any succeeding versions. See also Appendix 3.*

PID-25. [RDA DFIG] {best practice}

A PID needs to be requested as early as possible, at least at the time of registration at a trustworthy repository a PID record needs to be available.

> *Registration here is meant as an action by the repository to make the digital object part of the globally accessible data domain. This registration therefore includes a step of requesting a PID from an authorised service provider. (Actually, in general any agent may "request" an assignment of a PID to any digital resource.) Note that scientific repositories may contain data objects that have a rather limited period of existence, such as outputs from intermediate processing steps, or those resulting from failed experiments. These can be kept for internal use, but will probably never be "registered" by the producers in the context of this assertion, which implies that these DOs will not be made visible and accessible.*

PID-26. [RDA DFIG] {granularity}

PIDs are associated with collections which can consist of a number of digital entities, i.e. the level of granularity at which PIDs will be assigned is left to the communities and repositories.

> *Granularity is related to the smallest part of a DO that it is reasonable to refer to and/or cite, and is thus connected to both collections as well as subsetting. The question of granularity is complex, and best practises are by necessity quite domain-specific. There seem to be cases where a high granularity is called for, but in other situations this cannot be recommended. Use cases need to be specified by the specific community or the specific repository to provide guidance.*

PID-27. [COR] {community-specific, best practise}

The generic metadata scheme [of a community] will need to include a common identifier scheme for clinical research data objects.

> *This refers to a certain implementation for storing information about clinical research. In general, it is wise to add the PID of a DO to its metadata in a structured way so that it can be found by both humans and machine-driven workflows.*

PID-28. [USE] {policy}

Repositories need to clearly state which policies they follow in terms of granularity, versioning, time of PID assignment, binding, etc.

> *In this context, "repositories" refers to the organizational entity that curates and stores DOs (on behalf of the object's producer and/or owner). Also, "binding" refers to the practise of adding - already in the PID record of a DO - machine-actionable links to all relevant information required to understand and use the DO - not only paths to the locations where the bitstreams are stored, but also links to the original metadata record, landing pages, access rights record, etc. Preferably, all of the latter should themselves have been assigned separate PIDs.*

PID-29. [USE] {best practice, versioning}

Previous and subsequent versions can be indicated as machine readable types in the PID record.

> *Some repositories use attributes in the PID record to refer to the previous and/or subsequent version. If these attributes are typed (using a standardized format and/or attribute name, preferably defined in a type registry), also machines can use the information. Other repositories use separate metadata records to include this information which is probably not as efficient as using the PID record. Importantly, the versioning metadata should provide a linkage between all versions of a DO.*

PID-30. [USE] {PID record}

It is reasonable to use the PID record to add relational information binding to crucial digital objects such as bit sequences, metadata of different types (descriptive md, provenance, access rights etc.), landing pages, etc.

> *The PID which is assumed to be persistent can be used to associate pointers to digital objects that are closely related to the PID of the digital object concerned. There is a need to standardize the types of the corresponding attributes in a type registry in the PID record to enable machine processing.*

PID-31. [USE] {best practice, semantics}

PIDs [string] should in general not contain semantic information.

> *See Appendix 2 for more discussion of this.*

PID-32. [USE] {best practice, semantics}

The prefixes define name spaces and the service providers owning the prefix need to specify their naming policy.

> *The prefixes are globally unique requiring clear statements how these prefixes are constructed and assigned.*

PID-33. [USE] {resolution}

Some PID systems allow the use of fragment specifications (appended to a PID, usually after a hash character: <PID>#<fragment-spec>). The fragment specifer string is not part of the identifier itself, but can be used to address parts of the digital object - such as a short video clip within a lengthy recording, or an XML sub-structure contained in a larger XML document.

> *Importantly, the responsibility for resolving the fragment does not lie with the PID system and/or its maintainers, but instead it rests solely with the owner/provider of the digital object to which the PID itself resolves. The PID can be used for citations, a PID extended by a fragment specified can be used for referring to parts for practical reasons, such as for visualisations. After resolution of the PID, the fragment specifier will be appended to the path information which could be a URL.*

PID-34. [BIO 3] {best practice}

Make Local Resource Identifiers rugged to real-world use.

> *In this context, “rugged” refers to robustness and stability. This statement makes practical suggestions about the type of characters one should avoid in the (local) identifier string of the digital object, as they may have special meaning in the protocols used, or not be interpretable by browsers.*

PID-35 [BIO 4] {best practice, resolution, landing page}

Make the full URI simple and durable.

> *The statement explains why it is important to implement full URIs to enable resolution to state information or a landing page and gives practical advice which kind of information should not be included in the string. It applies to basically all types of PIDs, but is especially pertinent to non-Handle identifier systems.*

PID-36. [BIO 5] {best practice}

Carefully consider whether to embed meaning [into the PID sequence].

> *The statement explains that all meaningful information about a DO should be incorporated into its associated metadata, rather than be embedded into the identifier. Indeed, including any "meaningful", (human) interpretable information into a PID should only be considered for exceptional cases. Compare PID-31 and PID-35.*

PID-37. **[BIO 9]** {best practice}

Document the identifiers you issue and use.

> *The statement explains that the identifiers one is issuing (both the types and the system) need to be documented. Issuing PIDs can become a complicated issue, and may be touching on very domain-specific needs and practices.*

PID-38. **[BIO 2]**{best practice}

Help identifiers travel well: don't let them leave home without a Prefix and a Namespace.

> *The statement states that using a Local Resource Identifier alone can lead to collisions. For Handle/DOI the inclusion of the prefix as namespace is crucial since they make PIDs globally unique. To make them resolvable in the web, one needs to extend them by the URL of a Proxy service as long as a special internet protocol has not been accepted. see also Appendix 1.*

PID-39. **[BIO 7]** {versioning}

Implement a version-management policy.

> *Some PID schemes allow adding a version indicator at the end of a PID separated by a dot: <PID>.<version>. This statement explains that either the change history of the digital object needs to be documented, or the identifier should be versioned, or both should be done. The statement also indicates that if the resource has been removed the full URI should continue to resolve, but to a "tombstone" page. There is an ongoing debate how to best solve the versioning problematic and how PIDs could be used efficiently. Some will assign PIDs for each new version of a digital object. In this case the PID registry should contain the history of a given digital object, allowing end users (humans or machines) to traverse the DO version history, if needed.*

PID-40. **[BIO 8]** {versioning}

Manage complex life cycles without deletion (of PIDs).

> *The statement explains that generated and publicly advertised PIDs must never be reassigned to a different record/resource or deleted. The statement indicates a difference between static records such as experimental data and dynamically evolving entities such as concept descriptions. Numerical suffixes (version indicators) will not be sufficient to cover dynamics.*

PID-41. **[BIO 8a]** {provenance}

When two or more [DOs] that have different identifiers are merged, the [resulting new DO] should have a new identifier, [and] information about the merging action [provided in its metadata].

> *Note that the original assertion text refers to "entries", e.g. in the form of database records. In general, "merging" applies to any combination of pre-existing content. Specifically, database records identified by a query may be considered as persistently identified digital objects (DOs) in their own right, if the query itself is sustainably stored (preserved) and associated with a PID.*

PID-42. [BIO 8b] {provenance}

When [a DO] is being split into two or more [units], new identifiers should be assigned to the new entries, also here some history information needs to be made.

> *Note that the original assertion text referred to "an entry", e.g. in the form of a database record which can be considered as a DO (see PID-41).*

PID-43. [BIO 8c] {provenance}

If [a DO] has been removed or deprecated, the original identifier must still [be resolvable].

> *Note that the original assertion text referred to "an entry", e.g. in the form of a database record which can be considered as a DO (see PID-41). A special case is represented by physical objects that are consumed as a part of the research (e.g., in medical research, such a consumable resources are biological samples). After such samples are depleted, their digital representation (metadata record and associated PID registry records) should indicate that the sample is no longer available. Also compare PID-24.*

PID-44. [LSID 2] {LSID-specific}

For assigning LSIDs the following recommendations are made:

1. Providers should use separate authority identifications for objects where there is any reasonable possibility of future need to separate the name space.
2. Providers should not use separate authority identifications to split LSIDs by categories (departments, collections, data types, etc.) - unless the objects are likely to be transferred to new owners or served from different servers.
3. Providers should use namespace identifiers to split LSIDs across different categories (object type, discipline, departments, collections, projects, etc.).
4. Providers should use well-established locally unique and immutable object identifiers as LSID object identifiers.
5. LSID authorities should not use the primary key of relational database tables as object identifications.
6. LSID authorities should use appropriate metadata properties to represent relationships between revisions of an object.
7. LSID data must never change.
8. LSID metadata may change.
9. The default metadata must be RDF serialised as XML.
10. Non-binary encoded objects should be served as LSID metadata.
11. Objects in the biodiversity domain that are identified by an LSID should be typed using the TDWG[15] ontology or other accepted vocabularies in accordance with the TDWG common architecture.
12. Providers should not dynamically encode data in formats such as XML which may change the exact sequence of bytes.
13. Providers should tag their objects with LSIDs and encourage clients to use LSIDs to refer to those objects.

[15] TDWG refers to the Taxonomic Database Working Group, which is the biodiversity informatics communities' standardization authority.

> *Most statements are specific to the deprecated LSID concept and cannot be easily translated to non UNR-schemas except for the following cases: [1]-[2] indicate that it may make sense in special cases to think about the granularity of prefixes. [3] Some repositories use semantics in the suffixes to separate by categories. [5] Using a preserved and persistently identified query (see PID-41) can make sense. [6] It is shared by many that complex relationships should be expressed by metadata and not within a PID record. [7] In general, repositories make a difference between mutable and immutable data and indicate this in the PID record. [9] In general, PIDs or PID systems do not make statements about the formal specification language of DO's metadata.*

PID-45. **[CLA]** {community-specific}

- Centres need to associate PID records according to the CLARIN agreements with their objects and add them to the metadata record. Handle Assignment policies should be made clear and it should be clear where to find the PID in the metadata record.
- Centres need to associate Handle PIDs with their metadata.
- Non-metadata files should receive a PID or a PID in combination with a part identifier, if these files are accessible via internet, are considered to be stable by the data provider, are considered to be worth to be accessed directly.
- Centres using for example URNs are suggested two options to make their service compliant: a) a Handle is created that points to the URN:NBN resolver URL, b) URNs are transformed into Handles (example: urn:nbn:fi:lb-20140421 will become hdl:11113.1/20140421).

> *CLARIN agreements are rule sets defined for all participating centres which include for example DSA/WDS certification. CLARIN requires having a metadata record for all DOs registered and stored in repositories and important is that one can find the PID of the DO in the metadata record in a way that machines can find it. The third statement basically requires persistent accessibility of the bit sequences of a DO via Internet when they are assigned a PID, i.e. this is the case when they have been uploaded into a repository and been assigned a PID. A discussion about URN:NBN can be found under PID-12.*

PID-46. **[CES]** {community-specific}

- Each CESSDA service provider (SP) shall use globally unique and persistent identifiers to identify their data holdings.
- All data holdings of each CESSDA SP shall be findable by their global PID via the Internet.
- Persistent Identifiers need to be used to ensure referencing and citation of the data holdings of each CESSDA SP.
- Persistent Identifiers must be included in the resources-discovery metadata provided by CESSDA SP.

> *Basically the CESSDA community established criteria comparable to the CLARIN community.*

PID-47. **[CMIP]** {versioning}

[The CMIP project has defined] a hierarchically organised collection of PIDs allowing finding the latest version of a DO.

> *The CMIP6 specifications to be used by the international climate simulation community to generate their world climate reports use specific PID attributes to refer to previous and subsequent versions, i.e. just by using PID record information one can span the history graph of a DO.*

3.4 Using PIDs

PID-48. **[FAIR A1]** {best practice}

(Meta)data should be retrievable by their identifier …

> *The PID should provide a resolvable route to the data and/or metadata object, albeit not necessarily a direct one as part of the state information returned by the PID resolver. See Appendix 3 for a discussion on landing pages.*

PID-49. **[RDA PIT 1]**

PID systems should provide the attribute profile they are supporting under their prefix root.

> *This refers to the so-called attributes which are used in the PID records of a PID services provider and which are returned as state information when requesting to resolve a PID. The attribute schema should be easily available to be able to check which attributes are used and what their constraints and meaning is. Typically, this is a small subset of all the metadata associated with the DO. Yet there is no common agreement about a core set although experts in RDA are working on this.*

PID-50. **[RDA DFIG]**

The PID Record can be used to store the context of digital objects (bitstream locations, metadata, PID, rights information, landing page, etc.).

> *Some communities/repositories prefer to make use of attributes in the PID record to store state information, others prefer to use structured landing pages. See appendix 3!*

PID-51. **[RDA DFIG]** {best practice}

A metadata description contains the PID of the corresponding object. The PID record contains the metadata PID to ensure at all times that DO's context can be retrieved.

> *When having done a metadata search for useful DOs, users need to use the metadata information to see how the corresponding bit sequence can be accessed, i.e. a PID is needed that can be resolved into path information. On the other hand it is useful for people who find a reference to a DO to also find the pointer to the metadata in the PID record to be able to do interpretations.*

PID-52. **[RDA DFIG]** {best practice}

The PID record should include an expiration date for the digital object. Even for digital objects that have been deleted a PID record should still exist, to indicate the deletion and if possible to point to the metadata record.

> *For DOs whose bit sequence is no longer available there should be a reference to a "tombstone" page in the PID record, i.e. a landing page with at least a minimum set of metadata including an explanation of why the DO was deleted. See Appendix 3 for more information on landing pages.*

PID-53. [PID WS] {best practice}

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 ways of data handling and finally to a Global Digital Object Cloud as a generic, non-proprietary virtualization layer.

> *The Global Digital Object Cloud (GDOC, http://hdl.handle.net/11304/a8877a1a-9010-428f-b2ce-5863cec4aff3) refers to a concept worked out in the context of the RDA Data Fabric IG. In this model, a cloud of Digital Objects (each of which is persistently and uniquely identified) exists as a virtualization layer on top of network resources and services. An end user, human or machine-based, does not need to know any particulars of where or how the information about DOs is stored in order to operate on them. In addition, following a PID-centric view, the DOs are described and typed by their metadata, ideally using links to definitions stored in Data Type Registries (DTRs).*

PID-54. [BIO 10] {best practice}

Reference responsibly and rely on full URIs.

> *The statement describes the responsibilities of those who are using PIDs for referencing and citation where full URIs are important since they are actionable.*

PID-55. [LSID 3] {community-specific, best practice}

Usage recommendations for LSID are as follows:

- Clients in general must not try to infer relationships between objects based on the revision identification or any other part of an LSID. Instead clients must retrieve revisions related information from the returned metadata. Clients in general must consider LSIDs as opaque strings.
- In an HTML document, an LSID appearing within the description of the object it identifies should be presented in plain text and in its original form.
- In HTML web pages, LSIDs that refer to objects other than that being described should be presented as hyperlinks, with their original form as link text, and their proxy version as the link URL.
- In documents that support hyperlinks, LSIDs should be presented as hyperlinks with their original form as link text, and their proxy version as the link URL.
- In printed documents, LSIDs should be presented in their original form.
- In RDF documents, objects must be identified by an LSID in its standard from using the rdf:about attribute.
- The description of all objects identified by an LSID must contain an owl:sameAs, owl:equivalentProperty or owl:equivalentClass statement expressing the equivalence between the object identifier in its standard form and its proxy version.
- All references to objects identified by LSIDs using the rdf:resource attribute must use a proxy version of the LSID.

> *The LSID community also made statements about how to include PIDs in documents or in semantic assertions. The first statement is generic and recommends including most of a DOs relationships in metadata. Although the other statements seem to be specific to LSIDs they can be transformed to statements about the embedding of other types of PIDs in different contexts.*

3.5 Handles and DOIs

PID-56. [DOI] {best practice}

For electronic documents and published digital objects register a digital object identifier (DOI, which is a Handle with prefix 10) and associate suitable information with it (such as citation metadata).

> *There is a growing agreement that Handles issued by trustworthy service providers should be used as early as possible in data management to enable stable referencing for example in workflows. DOIs (which constitute a subcategory of Handle-based PIDs) however should be used when data (collections) are being published, i.e. in general this requires for example quality checks and additional metadata, as prescribed by e.g. DataCite.*

PID-57. [PID WS] {strategy}

We urgently need to come to a structured and integrated domain of Handle Service Providers.

> *There is a growing number of communities and repositories worldwide that are relying on the resolution of Handles. Therefore, the urgency for a better support of the domain of Handles services is increasing.*

PID-58. [PID WS] {strategy}

Service providers need to ensure that these two interoperable domains are part of one integrated landscape of rich services.

> *With respect to various services built on top of (especially Handle-based) PIDs, there is a gap between the increasingly rich choice - actively pushed ahead by publishing industry - that is available to the DOI user community, and the comparatively small set that is targeting those communities that are using Handles in general. It is time to work towards bridging this split, but the success of this effort will critically depend on the level of engagement and trustworthiness of the service providers.*

3.6 Communications/discussions/networking with communities, funders, users

PID-59. [PID WS] {strategy}

Setting up and maintaining trustworthy repositories is key for a structured data landscape guaranteeing access to data and its accompanying metadata.

PID-60. [PID WS] {strategy}

We need to design the required mechanisms (for facilitating automatic data processing) and build the needed tools now with high urgency.

PID-61. [PID WS] {strategy, best practice}

The PID centric approaches that are key to manage the data Tsunami require simple and clear messages for the users.

> *Here, “PID-centric approaches” refer to a concept described by e.g. the RDA DF IG where PIDs are put in the centre of solutions for data management and access. It stresses the role of*

> *attributes stored in the PID record called state information allowing machines to immediately act when references to other DOs of relevance for further processing are returned such as for type information enabling visualisation, transformation, etc.*

PID-62. [USE] {best practice}

Mechanisms need to be defined to bridge between the Semantic Web community preferring Cool URIs for identification and the Data Community often using PID systems such as Handles/DOIs.

> *Cool URIs are based on Uniform Resource Identifiers (URI) which, by proclamation, will not change. They make use of standard HTTP functionalities, in particular content negotiation, to enable the URI to be resolved to different representations (RDF, HTML) of the same object. Cool URIs allow webmasters to maintain the persistence of their resource identifiers, the URIs, with a minimum of effort and without a PID system. Practice shows that it is not easy to fulfil the promise of Cool URIs. For access to data broken links are dramatic, since mostly machines will operate on data and they will get stuck in case of access problems. This and speed considerations is why the data community mostly relies on Handles/DOIs. To support seamless crosswalks a better integration of the two existing approaches is requested.*

PID-63. [BIO 6] {Best practice, Community-specific}

Make the full URI and CURIE clear and easy to find.

> *The statement explains that it must be easy for users to find the location where the PID can be found and it gives some suggestions. It is specific to the Bio and Life Sciences domains as explained in PID-10, but it can be extended to other PID usages.*

4 Conclusions and outlook

This work was undertaken by the GEDE PID Focus Area group in close synchronisation with the RDA Data Fabric Interest Group (DFIG), ensuring a global focus for both the collection and validation of the assertions. The intention of this cross-disciplinary report is to overcome still existing confusions about the nature, the creation and the usage of Persistent Identifiers (PIDs), and to address the lack of detailed knowledge in many research disciplines. The document is not meant to produce yet another comprehensive document on PIDs, but identify agreements across documents — in the form of assertions — that have been suggested to be included by experts. With this document GEDE is happy to help demystifying PIDs, by helping to overcome confusions and to create bridges between the various disciplines.

Acknowledgements

The work in the GEDE PID Focus Area group in general, and on this document in particular, has benefited greatly from discussions with a large number of colleagues from all over the world. We would like to especially acknowledge valuable inputs from Larry Lannom (CNRI), Andrew Treloar (ANDS), Alexander Ntoko (ITU) and members of the RDA Data Fabric Interest Group.

Sources

The following table (Table 2) lists the sources and documents from which we collected assertions and statements related to PIDs and their usage.

Table 2. Sources used in this report, and corresponding references. (Complementary to Table 1.)

Abbrev	Source	Reference
BIO	Life science community report by McMurry et al.	McMurry et al. (2015), "10 Simple rules for design, provision, and reuse of persistent identifiers for life science data", preprint available at https://zenodo.org/record/18003#.WJHDt7YrJbU
CES	CESSDA Documentation	communication by Email
CLA	CLARIN PID policy summary	https://www.clarin.eu/node/3965
CMIP	Persistent Identifiers for CMIP6: implementation plan	Contributed by T. Weigel, based on an CMIP6 project-internal working document. See also https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6.
COR	CORBEL Consensus document on providing access to individual participant data	http://www.corbel-project.eu/
DOI	International DOI Foundation documentation	https://www.doi.org/
FAIR	FAIR Principles	https://www.force11.org/group/fairgroup/fairprinciples
ITU	ITU Documentation	ITU-T X.1255: https://www.itu.int/rec/T-REC-X.1255-201309-I/en
LSID	Taxonomic Database Working Group (TDWG) Globally Unigue IDentifiers (GUID)	https://github.com/tdwg/guid-as
PID WS	RDA EU PID Workshop	https://www.rd-alliance.org/views-about-pid-systems-workshop https://www.rd-alliance.org/views-about-pid-systems-training-course
RDA DFT	RDA Data Foundation & Terminology WG	DFT Core Terms and Model http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318
RDA DFIG	RDA Data Fabric IG Discussion	https://www.rd-alliance.org/group/data-fabric-ig.html
RDA PIT	RDA PID Information Type WG	https://www.rd-alliance.org/groups/pid-information-types-wg.html
USE	GEDE Usage Document	Aspects of PID Usage – Discussion Document

In addition, the following documents were also studied:

- Global Digital Object Cloud: http://hdl.handle.net/11304/a8877a1a-9010-428f-b2ce-5863cec4aff3
- Handle System: https://www.handle.net/index.html
- DONA: https://www.dona.net/
- Anton Güntsch, et. al.: Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. https://doi.org/10.1093/database/bax003
- Jens Klump, Robert Huber: 20 Years of Persistent Identifiers - Which Systems are here to stay? https://doi.org/10.5334/dsj-2017-009

- Stian Soiland-Reyes, Alan Williams: What exactly happened to LSID? http://dev.mygrid.org.uk/blog/2016/02/what-exactly-happened-to-lsid/
- Netherlands Coalition for Digital Preservation (NCDD): Introduction to Persistent Identifiers". http://www.ncdd.nl/en/pid/
- The International Virtual Observatory Alliance (IVOA): Table Access Protocol. http://www.ivoa.net/documents/TAP/
- Report from Dynamic Data Meeting (2014): "Data citation and digital identifiers for time series data / environmental research infrastructures". https://www.bodc.ac.uk/about/outputs/presentations_and_papers/documents/datacitation_juck.pdf
- R. Huber: How dead is the PID Zombie zoo? https://www.rd-alliance.org/sites/default/files/attachment/20160902-RDA_EU_View_on_PID_Systems_Garching-Robert_Huber-Jens_Klump-How_dead_is_dead_in_the_PID_Zombie_zoo.pdf
- PID Centric operation: https://www.rd-alliance.org/group/data-fabric-ig/wiki/df-configuration-pid-centric-data-management-and-access.html

Appendix 1. Elaboration on PID Forms

This is a short elaboration on PID forms to clarify some terminology and conceptual differences between URIs or Handle/DOIs, both being used as persistent identifiers.

Short Summary

- Cool URIs are being used as unique identifiers although they practically refer to locations on the web. Cool URIs should not change to maintain address stability. URIs are coupled with the web protocol.
- A Handle/DOI is an identifier that is independent from any protocol anticipating that on internet a variety of protocols will exist. However, to make it actionable in the web an extension to a URL by specifying a proxy URI is required.

Life Sciences Document

The document from the life science colleagues ("10 Simple rules for design, ..."; BIO in Table 1) states the following about PIDs:

An Identifier is a sequence of characters that identifies an entity. A Local Resource Identifier (LRI) is an identifier that is only guaranteed to be unique within a single database. A Uniform Resource Identifier (URI) is an identifier that is guaranteed to be both uniform and globally unique. A CURIE is a compact URI comprised of <prefix>:<LRI>. A full URI is an identifier that also resolves to a webpage containing information about the identified entity.

An example for a full URI is: http://zfin.org/ZDB-GENE-980526-166

The compact URI for this is: ZFIN:ZDB-GENE-980526-166 with ZFIN as prefix.

The basic assumption of this approach has been described by T. Berners-Lee: *Cool URIs don't change*, i.e. the maintainers of the server behind the URL "http://zfin.org" will take care of the resolving steps and pretend to stay forever. A full URI needs to be used to make the PID actionable. The string "ZDB-GENE-980526-166" must be locally unique in the ZFIN name space, but not globally.

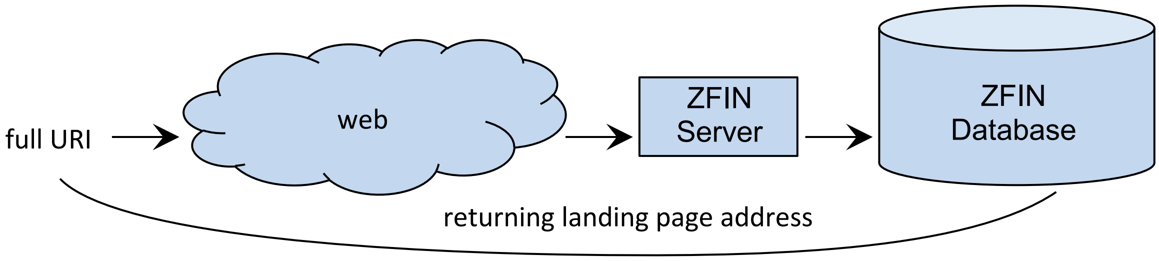

Figure Appendix 1- 1. The LSID approach. A landing page address (URL) is returned, which needs to be processed and turned into action.

A landing page address (URL) is returned, which needs to be processed and turned into action.

Handle/DOI System

For the Handle System which is also being used for DOIs (Digital Object Identifiers) the following is being specified:

Within the handle identifier space, every identifier consists of two parts: its prefix, and a unique local name under the prefix known as its suffix. The prefix and suffix are separated by the ASCII character "/". A handle may thus be defined as <Prefix> "/" <Handle Local Name>. For example, *handle "12345/hdl1" is defined under the Handle Prefix "12345", and its unique local name is "hdl1".*

Two examples for real Handles and DOIs (which are Handles) are:

Handle: 11304/a3d012ca-4e23-425e-9e2a-1e6a195b966f

DOI: 10.23728/b2share.3d2296cd14e74e74b9c960a2fafb5ff5

In the first example, the string "a3d012ca-4e23-425e-9e2a-1e6a195b966f" must be unique within the name space "11304".

In the second example "b2share.3d2296cd14e74e74b9c960a2fafb5ff5" must be unique within the name space "10.23728". Handle prefix allow some sub-structure as it is being used in the DOI example.

Since there is yet no accepted web-protocol called "HDL" proxy servers are being used to make the Handle/DOI actionable. The handle/DOI proxy servers must have the characteristics of a Cool URI, i.e. they need to be maintained at least as long as the HDL is not accepted as a standard web-protocol. The "full URIs" as introduced above for the two examples are as follows:

Handle: https://hdl.handle.net/11304/a3d012ca-4e23-425e-9e2a-1e6a195b966f

DOI: http://doi.org/10.23728/b2share.3d2296cd14e74e74b9c960a2fafb5ff5

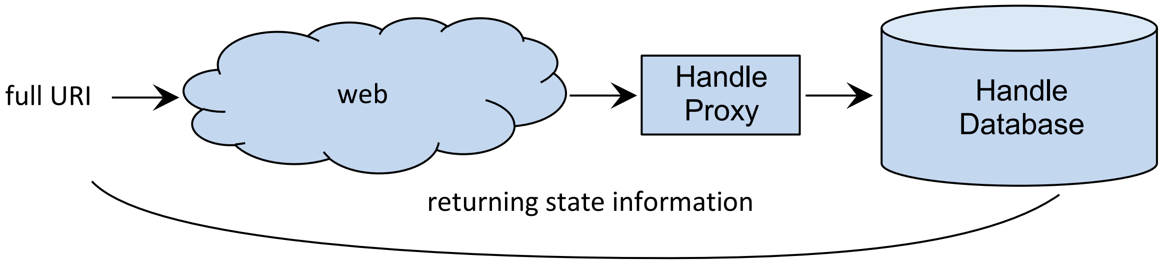

Figure Appendix 1- 2. The Handle system approach. Typed state information is returned to the caller, which will lead to actions.

Typed state information is being returned to the caller which will lead to actions. The Handle database can also be contacted via an API which means that no web protocol based actions are required.

Differences

Syntactically, there is hardly any difference between the URI and the handle system except for the delimiter separating the prefix and the suffix, and the address space is unlimited. However, there are a few differences with respect to processing:

- Without specifying the zfin.org URL the string "ZDB-GENE-980526-166" is meaningless. However, the Handle System is a networked global naming system that uses proxy servers to fit in with the current web protocols. The Handle remains meaningful without the proxy extension since it can be resolved independent of any URL.
- The current proxy does not include complex logic except for redirecting to the Handle database and sorting through the returned values. In the DOI case, the proxy executes linked data and multiple resolution. When "HDL" would be accepted as standard internet

protocol there would be no need for a proxy server any more and each Handle would be directly resolved by the Handle Services.

- Nevertheless, maintaining proxies, servers etc. is a burden and since the Handle System is already being used by a large number of communities it is easier to keep the domain names and services constant[16].
- The Handle/DOI system just has two resolution layers (global, local) and it directly returns state information which will be typed, i.e. machines know which actions are to be taken.
- Based on requests by scientific communities the Handle Syntax allows adding fragment identifiers which can be interpreted by the receiving repository. *<Prefix> "/" <Handle Local Name>#<fragment id>*

It should be stated clearly that a PID identifies a Digital Object as a whole, making it referable and citable. The fragment identifier can only be used to point to parts within a DO. Example: A PID identifies a video film of 1 hour, however in a piece of text someone wants to refer from a paper to a moment of simply 10 seconds. They could first extract this fragment from the video as a new DO associated with a new PID and then refer to it, but this is not economic. Better is to add to the PID a time specification (begin time + end time), send the whole identifier to the streaming server which will then only return the selected video frames.

Appendix 2. Use of Semantics

Different data communities discuss about the relevance of putting semantic information into PIDs, in other words to give a semantic meaning to the PID Strings.

If it is not discouraged to use semantics for building PIDs (for some reasons the PID creators may want to include some semantics in the PID string, cf. the LSID specifications for example), it is highly discouraged to build PIDs resolvers based on the semantic meaning of a PID. Indeed, the information "semantically coded" into a given PID by its creator may evolve over time and thus become meaningless or even confusing for users: in order to guarantee the persistent nature of an identifier, a PID string should be always seen only as an identifier that will be resolved to a useful set of information (often called state information) that points to all kinds of semantic messages. This implies that the resolution process should be transparent with respect to the PID string content and treat it just as a reference. PIDs consumers need to pay great attention while interpreting semantics "coded" into PID-Strings.

Appendix 3. PID records, landing pages etc.

The terms *Digital Object, PID, PID-records, landing pages, Digital Object metadata, etc.* are widely used in research data landscape (and of course into the present document).

Experience has shown that sometimes a given term may be interpreted in different ways. In this appendix, we would like to confirm the definitions we adopted through this document by clarifying them with practical examples.

[16] The example of LSIDs shows that uptake is crucial (http://dev.mygrid.org.uk/blog/2016/02/what-exactly-happened-to-lsid/) and that in the case of LSIDs uptake was NOT sufficient. The report indicates how sensitive it is to create a useful and working PID system.

For the *Digital Objects,* we adopt the definition given by the RDA Data Foundation Terminology Interest Group: a Digital Object is a triplet composed by:

- A bit sequence, stored into some repositories;
- Metadata and state information. These are specific metadata information that describe the current properties of the DO that are relevant for proper management and access;
- A resolvable Persistent Identifier (which is a String).

The Persistent Identifier is associated with a PID record. It may be seen as an "incarnation" of the Digital Object and gives access to the bit-sequence, metadata and state information, etc.

The resolvable character of PIDs consists in establishing the route to find an identified digital object (or the digital object built with information from a physical object or people). In what follows, we define the *Landing Page* as the set of information that is retrieved when a PID is resolved (note that this definition is "local" to this document and other communities may give it a different meaning). A landing page may be human oriented (typically an HTML page formatting the metadata/state information) or be machine actionable. It is important to underline that, at present, there is no broadly accepted specification about the attributes that a landing page should include: the *PID Kernel Information* RDA working group is active for specifying a set of *core attributes* (https://www.rd-alliance.org/groups/pid-kernel-information-wg). Although different user communities may define their own set of attributes (e.g. Zenodo, EUDAT, ENES). The RDA *Data Type Registry* is currently working for achieving interoperability between different sets of attributes (https://www.rd-alliance.org/group/data-type-registries-wg/outcomes/data-type-registries

The routing process between a given PID and the relevant landing page is not only agnostic about the content of the landing page (as we explained in the previous paragraph), but is also agnostic about the nature of the adopted PID system: some communities rely on URIs, other Handles (DOIs are a Handle subset).

In the two following examples the PIDs are DOIs. This particular choice does not decrease the generality of our reason for highlighting the specifics of different landing pages. The first example comes from digital scientific publication, the other from data published into repositories.

- **The case of digital publications**: let us consider the publication referenced by the PID http://doi.org/10.5334/dsj-2017-039. This is an article titled *"Persistence Statements: Describing Digital Stickiness"* published on the digital open review *Data Science Journal.* The PID landing page is displayed when one follows the link. The underlying bit-sequence is the text of the publication (this is displayed directly on the landing page, or may be downloaded as a PDF). The associated metadata are the name of the journal, the names of the authors (with their academic affiliation), the keywords, the citation snippet, the publishing date, the license protecting the paper and some statistics (number of views, number of downloads, number of twit associated with this publication). The information displayed on the DOI landing page and the presentation style are arbitrarily chosen by each journal. Some information (e.g. the number of *Tweets* or the list of citing articles) may be present for a given review and missing for other ones.
- **The case of digital published data**: A dataset is a collection (an aggregation of Digital Entities) of data of different sort. Let us consider a dataset published and referenced by

the PID 10.5281/zenodo.804836. This is a dataset published on the Zenodo repository. The PID is not a direct link to the raw bit-sequence, but a link to a human readable page containing a set of information concerning the data-set:

- The title and a detailed human-oriented description of the data,
- The names of the authors,
- The publication date and the version of the data,
- A set of keywords associated with the data-set,
- The licence protecting the data and the access rights,
- A set of links for downloading the bit-sequence of each element of the collection and the files types encapsulating each bit-sequence.

As we see with these examples, the resolution of two different PIDs points to two completely different kind of landing page. Some elements may be common but other may diverge.

We may also note that in the included examples the landing pages are human readable. Other service may deal with machine actionable PIDs resolutions systems.

Appendix 4. Considerations around Persistence

During the 9th plenary meeting of the Research Data Alliance (April 5-7, 2017 in Barcelona), Andrew Treloar started a discussion on the meaning of the term “persistence”. This particular point was not discussed before. This appendix resumes the outcome of that discussion: persistence occurs at different levels in the PID context, as we are going to explain in the following paragraphs.

Persistence of PIDs

The concrete, non-abstract form of any PID is an arbitrary character string, built adhering to some syntax and norms. It can exist in various forms (e.g. in an ASCII file, printed on paper, etc.). Beyond any resolution system that may be associated with- (or actuated by) the PID, this may be always seen just as a string. In this context, their persistence is not an issue.

Persistence of PID Resolution

The resolution system (i.e. the mechanisms establishing the route between the PID and the identified digital object) gives a PID life. Indeed it turns the PID into useful state information about a Digital Object (such as the locations where the bit sequences can be found, where the metadata is located etc.)

The persistence of the resolution system is crucial. This must follow specifications, norms and syntaxes stable over the time in order to achieve persistence. It should be mentioned here that there could be, for a given resolution system, different implementations and/or evolving adopted technologies, without persistence being affected (as long as the specifications and norm are respected).

Persistence of PID Management

In the previous paragraph we highlighted the importance of the rules for building routes between a PID and its identified digital object. Even if the rules for building these routes are stable over time (persistence of the resolution system), a given Digital Object may move over time. The route should then be adapted/rebuilt (always following the specifications and norms): this means that, in addition to the PID Resolution service, it is necessary to persist the

PID Management service. Indeed a large part of the value of any PID system is in how it engineers away some of the inherent brittleness in the location of Digital Objects.

Persistence of Access to the DO

According to their definition (cf. Appendix 3), the Digital Objects have bit sequences stored in some repositories, are assigned a PID and are associated with various types of metadata and state information.

- **Persistence of access to the bit sequences.** This requires that the bit-sequences are accessible over long periods of time. Some data needs to be stored only for 10 years (to be able to reproduce scientific results) whereas other data should be available forever, as the data from cultural heritage.

  Bit Sequences are stored in repositories. It is therefore the responsibility of repositories to execute proper procedures to take care that the bit sequences are accessible according to what is specified and that the resolution system is updated such that the location information associated with the PID is correct.

  It is now generally accepted that, when a repository deletes bit sequences, the corresponding PID is resolved to a tombstone information page: This will inform the users that the DO once existed and will describe some of its characteristics.

- **Persistence of access to the landing page, metadata, state information, etc.** Very often PIDs also refer to metadata, landing pages, state information. This set of information is crucial for interpretation, reproduction and re-use of data. The repositories managing this type of information need to have procedures in place to make it accessible via the PID as long as the bit sequences exist, or via a tombstone page (see previous item) when the DO is removed.