Grouped List of Assertions on PIDs

Peter

23. March 2017; Version 2

This note is summarising (section A) and clustering (section B) the various assertions about PIDs and PID systems extracted from a number of papers listed in appendix 3. In the summary section the comments made at the GEDE meetings are mentioned. All assertions marked red are "new" in so far as they were extracted from recently added documents.

A few appendices are added to the note:

* GEDE Note on Persistent Identifier (PID) Bundle
* Appendix 2: Aspects of PID Usage - Summary of GEDE Discussions
* Appendix 3: List of contributing documents/groups
* Appendix 4: Elaboration on PID Forms
* Appendix 5: Glossary

# A. Convergence Section

In this section we try to summarise the essentials ignoring some nuances in the wording.

1. Nature of PIDs and PID Systems

* Persistent Identifiers to be:
  + long-lasting,
  + uniquely identifying Digital Objects (digital entities),
  + persistently resolving to state information and/or landing pages,
  + in general existing of a name space indicator (prefix) and a local identifier (suffix),
  + issued and managed by clearly specified authorities,
  + extended by web-address to make it actionable[[1]](#footnote-1),
* A trustworthy PID system requires to
  + be maintained by a dedicated and reliable team and clear authorisation structure,
  + be based on a transparent sustainable business model[[2]](#footnote-2),
  + be provided by an organisation that has a long-term perspective,
  + be subject of regular quality assessments by external parties,
  + be governed by international boards,
  + be based on open standards such as ITU X.1255 interoperability guidelines
  + be based on a redundant and secure architecture,
  + support a huge address space
  + have 24/7 availability
  + support an openly documented API optimally supporting accepted data models.
  + have exit strategies at all levels such as in case of organisation change[[3]](#footnote-3)
* Communities such as CLARIN state that the Handle System including the DOIs[[4]](#footnote-4) fulfil most requirements.

2. Relevance of PIDs and PID Systems

* PIDs are increasingly important, are being applied almost everywhere across sectors[[5]](#footnote-5) and disciplines and for all kinds of types of digital objects and we are making ourselves increasingly dependent of a functioning PID system fulfilling the requirements, that efforts are needed to offer sustainability.
* For the identification of persons the ORCID system has been evolved and is widely being used.

3. Assigning PIDs

Many assertions go about questions such as when and how to assign PIDs and there are some differences in the way answers are given. We only present summarising assertions which show a wide agreement. Some of the LSID assertions (PID54) are specific to their URN choice (in particular the specs for "namespace", "object" and "revision") and will be ignored here although they partly contain useful practical information.

* Several messages indicate
  + the duty to assign PIDs to digital objects[[6]](#footnote-6)
  + to better not use semantics in the PID string and make it rugged
  + to include the PID and their actionable extensions (full URI)
  + to use the PID to include binding and state information via PID record attributes or landing pages[[7]](#footnote-7). However, for machine readability it is important to define the attribute types.
  + to add the PID at least to the metadata and to make it findable
  + the need to document the PID use
  + associate a flag about mutability of the related DO with the PID
* With respect to the time of PID assignment and granularity issues nuances can be indicated. However, a few rules of thumb can be given:
  + Repositories (and other data providers) need to make clear which policies they apply in particular with respect to the time of assignment and the granularity. Also the attribute types being supported need to be defined to make them machine readable.
  + In general PIDs should be assigned as early as possible to the DOs, at least at their upload into a trustworthy repository.
  + In general a high granularity is recommended due to unknown later re-use. However, there may be exceptions which need to be explained.
* Some repositories use specific options which need to be well-described:
  + Some use version indicators in the attribute list or on the landing page. Adding version information to the PID (LSID option) is not recommended. Repositories need to state clearly what their versioning policy is - also to allow machines to find versions.
  + Some use part indicator extensions to refer to a part embedded in the DO referred to. These may not be seen as part of the PID since they are being interpreted by the repositories in their ways.
  + Some recommend embedding not globally resolvable strings (such as specific URNs) in the Handles as suffix and thus making them resolvable.
  + Policies have to describe typical life-cycle actions: deleting content, splitting content, etc.

4. Using PIDs

Despite the overlap between "assigning" and "using" PIDs we maintain this distinction in this document. Here we find a number of assertions about various aspects related to PIDs. Again LSID assertions are widely ignored although they make very useful detail suggestions.

* To make DOs "findable" there must be a close relation between the PID and the metadata description, since metadata is being used for searching and the PID found in the metadata can be used for retrieval. Some use the PID record to include the relation, others refer to a landing page.
* To allow users incl. machines the interpretation of the attributes in the PID record or the specifications in the landing page a clear typing needs to be supported and be made accessible to the user. One of the standard attributes is the location of the DO's bit sequences.
* It must be clear from the metadata schema where users can find the PID to the DO and here the full URI is important to make it actionable.
* The PID record or the landing page should include an attribute informing about the expiration date of the DO and its mutability.
* Agreed citation principles should be used (can we refer to some??)
* Systematically using interoperable PIDs opens new application opportunities such as presented by the Global Digital Object Cloud.

5. Handles and DOIs

* It is widely unknown that DOIs are based on the Handle Service and thus form a community of Handle System users specifying assignment and usage practices and a specific business model. The Handle System issued about 3700 other prefixes all being authorities building their own community. One is the redundant EPIC service.
* While DOIs are mainly used for published collections, Handles from other service providers are in general used to assign PIDs to the many (partly millions) digital objects created in the labs to make them referable in software, workflows, etc.
* More efforts need to be done to better integrate the DOI and other Handle domains to enable layered services.

6. Others

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

B. Statement Section

1. Nature of PIDs and PID Systems

**PID2. RDA DFT1.2:** 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.

**PID3. RDA DFT1.3:** A PID record contains a set of attributes stored with a PID describing DO properties.

**PID4. RDA DFT1.4:** A PID resolution system is a globally available infrastructure system that has the capability to resolve a PID into useful, current state information[[8]](#footnote-8) describing the properties of a DO.

**PID8. RDA**: 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[[9]](#footnote-9),
* 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[[10]](#footnote-10) and
* support an openly documented API optimally supporting accepted data models.

**PID25. ITU**: The platform for interoperability of heterogeneous identity management systems is an open architecture based on ITU-T X.1255 and the Digital Object Architecture ... and is capable of to offer interoperability at global level.

**PID26. ITU:** The top 5 benefits from this 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.

**PID31. USE:** The sustainability of a PID system mainly depends on its relevance and its investments, ensuring sufficient financial support.

**PID32. USE**: To prevent PID Zombies it seems to be wise to define an exit strategy at all levels.

**PID39. BIO1**: If you create identifiers, do not do identifiers by yourself. *(There may be exceptions, then the statement explains the aspects to be considered)*.

**PID40. BIO**: 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[[11]](#footnote-11).

**PID53. LSID1**: The authority identification within LSID[[12]](#footnote-12) 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 (TDWG).

**PID56. CLA:** the CLARIN research infrastructure formulated one statement about systems:

* One system which is performing, scalable and robust enough and that offers enough flexibility: Handle System.
* 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).
* CLARIN endorsed the FORCE11 citation principles (<https://www.force11.org/datacitation>

**PID59. CES**: the CESSDA research infrastructure formulated one statement about systems:

* CESSDA shall use a global PID services that ensure 24/7 resolvability of PIDs.

2. Relevance of PIDs and PID Systems

**PID16. PID WS**: Proper PID usage and support will become key for competitiveness in science and industry.

**PID 18. PID WS**: 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 19. PID WS**: 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.

**PID27. COR**: Clinical trial datasets should be considered legitimate, citable products of research, which needs persistent identifiers as a prerequisite. Persistent Identifiers such as the already widely used DOI should be applied to datasets to improve discoverability and to allow correct citation.

**PID28. COR**: The requester should also provide information on his/her expertise, possibly making use of persistent digital identifier system (e.g. ORCID).

3. Assigning PIDs

**PID1. RDA DFT1.1:** A digital object is ... referenced and identified by a persistent identifier ...

**PID5. FAIR-F1**: (meta)data are assigned a globally unique and eternally[[13]](#footnote-13) persistent identifier

**PID11. RDA**: 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.

**PID12. RDA:** PIDs are associated with collections which can exist 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. A high granularity is recommended to anticipate future applications[[14]](#footnote-14).

**PID29. COR**: The generic metadata scheme will need to include a common identifier scheme for clinical research data objects.

**PID30. USE**: Repositories need to clearly state which policies they follow in terms of granularity, versioning, time of PID assignment, binding, etc.

**PID33. USE**: Previous and sub-sequent versions can be indicated as machine readable types in the PID record.

**PID34. USE**: It makes sense 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.

**PID35.USE**: PIDs should in general not contain semantic information. The prefixes define name spaces and the service providers owning the prefix need to specify their naming policy.

**PID36. USE**: Some PID systems allow using fragment specifications (<PID>#<fragment-spec>). They may not be mixed with the identifier, but can be used to address parts of a digital object such as a short video clip within a lengthy recording or a XML sub-structure in a large XML document.

**PID38. USE**: Each meaningful digital object should be assigned a PID to facilitate re-use and re-combination. In particular automatic solutions require in general a high granularity.

**PID41. BIO3**: Make Local Resource Identifiers rugged to real-world use. *(This statement makes practical suggestions about the type of characters one should not use.)*

**PID42 BIO4**: Make the full URI simple and durable. *(The statement explains why it is important to implement full URIs to enable resolution to a landing page and gives practical advice which kind of information should not be included in the string.)*

**PID43. BIO5**: Carefully consider whether to embed meaning. *(The statement explains that meaning should be better placed in metadata, that meaning is never required to be embedded in an identifier and it describes a few rules for embedding meaning in exceptional cases)*

**PID44. BIO9**: Document the identifiers you issue and use. *(The statement explains that the identifiers one is issuing need to be documented and gives an elaborate list of recommendations of how to documeent)*

**PID45. BIO2**: Help identifiers travel well: don't let them leave home without a Prefix and a Namespace. *(The statement explains that the Local Resource Identifier can lead to collisions[[15]](#footnote-15))*

**PID46. BIO6**: Make the full URI and CURIE clear and easy to find. *(The statement explains that it must be easy to users to find the location where the PID can be found and it gives some suggestions)*

**PID47. BIO7**: Implement a version-management policy. *(The statement explains that either the change history of the resource needs to be documented or the identifier should be versioned or both should be done[[16]](#footnote-16). The statement also explains that if the resource has been removed the full URI should continue to resolve, but to a "tombstone" page.)*

**PID48. BIO8**: Manage complex lifecycles without deletion. *(The statement explains that generated and publicly advertised PIDs must never be reassigned to a different record/resource or deleted. The statement indicated 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).*

**PID49. BIO8a**: When two or more entries having different identifiers will be merged for example the new recipient entry should have a new identifier, but there should be information about the merging action.

**PID50. BIO8b**: When an entry is being split into two or more, new identifiers should be assigned to the new entries, also here some history information needs to be made.

**PID51. BIO8c**: If an entry has been removed or deprecated, the original identifier must still resolve.

**PID54. LSID2**: For assigning LSIDs the following recommendations are made:

* Providers should use separate authority identifications for objects where there is any reasonable possibility of future need to separate the name space.
* 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.
* Providers should use namespace identifiers to split LSIDs across different categories (object type, discipline, departments, collections, projects, etc.).
* Providers should use well-established locally unique and immutable object identifiers as LSID object identifiers.
* LSID authorities should not use the primary key of relational database tables as object identifications.
* LSID authorities should use appropriate metadata properties to represent relationships between revisions of an object.
* LSID data must never change.
* LSID metadata may change.
* The default metadata must be RDF serialised as XML.
* Non-binary encoded objects should be serves as LSID metadata.
* Objects in the biodiversity domain that are identified by an LSID should be typed using the TWDG ontology or other accepted vocabularies in accordance with the TWDG common architecture.
* Providers should not dynamically encode data in formats such as XML which may change the exact sequence of bytes.
* Providers should tag their objects with LSIDs and encourage clients to use LSIDs to refer to those objects

**PID57. CLA:** the CLARIN research infrastructure formulated a number of assignment principles:

* Centers 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.
* Centers 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 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).

**PID60. CES**: the CESSDA research infrastructure formulated statements about assigning PIDs:

* Each CESSDA 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.

**PID61. CMIP**: Defined is a hierarchically organised collection of PIDs allowing finding the latest version of a DO.

4. Using PIDs

**PID6. FAIR-A1**: (meta)data are retrievable[[17]](#footnote-17) by their identifier ...

**PID7. RDA-PIT1**: PID systems should provide the attribute profile they are supporting under their prefix root.

**PID9. RDA:** The PID Record can be used to store the context of digital objects (bitstream locations, metadata, PID, rights information, landing page, etc.)

**PID13. RDA:** 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. (this can be compared with the reverse DNS mechanism)

**PID15. RDA**: The PID record should include an expiration date for the digital object. Even for digital objects that have been deleted the PID record should exist, indicate deletion and if possible point to the metadata record.

**PID 17. PID WS**: 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] as a generic, non-proprietary virtualisation layer.

**PID52. BIO10**: Reference responsibility 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).*

**PID55. LSID3**: 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:equivalendProperty* 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.

**PID58.CLA:** the CLARIN research infrastructure formulated one usage principle:

* CLARIN endorsed the FORCE11 citation principles (<https://www.force11.org/datacitation>

5. Handles and DOIs

**PID10. DOI:** 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)[[18]](#footnote-18).*

**PID14. DOI:** A DOI needs to be registered for a published DO and it should be associated with citation metadata.

**PID 22. PID WS**: We urgently need to come to a structured and integrated domain of Handle Service Providers.

**PID 23. PID WS**: Service providers need to ensure that these two interoperable domains are part of one integrated landscape of rich services.

6. Others

**PID 20. PID WS**: Setting up and maintaining trustworthy repositories is key for a structured data landscape guaranteeing access to data and its accompanying metadata.

**PID 21. PID WS**: We need to design the required mechanisms (for facilitating automatic data processing) and build the needed tools now with high urgency.

**PID 24. PID WS**: The PID centric approaches that are key to manage the data Tsunami require simple and clear messages for the users.

**PID37. Use**: 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.

**Appendix 1**



Persistent Identifier (PID) Bundle

1st Draft Version, July 2016

**This document will summarise the GEDE discussions on PIDs, i.e. it will be commented, extended, changed etc. dependent on the group’s progress. Mostly the Wiki space will be used for the interactions. This document will be updated at regular moments.**

**Authors: Peter (RDA), Michael (ENES), Maggie (ICOS), Rob (ELIXIR), Dieter (CLARIN)**

**Specific acknowledgements: Alexander Ntoko (ITU), Larry Lannom (CNRI)**

## **State of Discussion**

During the last decade there is a growing conviction that all digital objects (DO) (data, metadata (of objects), software, queries, ) including abstract objects (concepts, relations, etc.) should be associated with a persistent and globally unique identifier (PID) that can be resolved to useful information such as where to find the associated bit sequences and how to prove identity. In addition to the ongoing scientific debate about the use of PIDs we now can also see an increased interest in industry. It is the Internet of Things with its billions of cyberphysical devices that will create huge amounts of data entities that need to be managed. In this context we can refer to the discussion within the ITU (International Telecommunication Union) on ITU-T X.1255 (which is based on the Digital Object Architecture) and the decision to support the Digital Object Architecture which has as one of its core components a PID system. ITU is focusing on an interoperable platform for the different name spaces that are already in broad use such as Barcodes, RFIDs, OIDs and others.

In the academic and library world PIDs have been discussed and applied to digital data including electronic publications, for over 15 years and several approaches have been made (ARK, Handle, PURL, URN, domain specific solutions, etc.) all having found their specific area of application. Recently published principles, such as those from G8 and FAIR, indicate an acceleration of the discussions and deep recommendations as for example worked out by RDA based on many use cases.

Summarising we can state that

* it is time to work on management and access solutions for digital objects involving industry that are based on standardised PID systems
* it seems that we are moving towards a **Global Digital Object Cloud[[19]](#footnote-19)** which characterizes a virtualisation layer where we primarily deal with stable PIDs and useful metadata descriptions
* there is much experience now in academia and industry for a number of PID systems and their functions in digital object architectures
* there is a general and urgent need for a stable, robust, performant and globally available system that can be used easily by everyone to register and resolve PIDs for different purposes and that can incorporate widely used practices
* there will be specific solutions that have shown their usefulness and it is up to the communities to decide how these practices will evolve and how mappings will be done if necessary

## **Statements about PIDs**

Within the RDA Data Fabric Interest Group (DFIG) we started to collect recommendations related to PIDs that have emerged from the scientific community during the last few years. These can be found under <https://rd-alliance.org/group/data-fabric-ig/wiki/recommendations.html>

Up to now we distinguished a few sources of statements:

1. Suggestions which come from RDA WG outputs (RDA WG)
2. Suggestions which emerge from RDA discussions (RDA)
3. Suggestions from other initiatives such as FAIR, ITU, DOI, etc.
4. Suggestions for RDA Recommendations (RDA REC) which will be the result of RDA interactions

If there are statements that relate to this bundle that come from other initiatives we should of course add them to be comprehensive. The first number is a way to refer to the statement in GEDE discussions, the second specifies the source[[20]](#footnote-20).

**PID1. RDA DFT1.1:** A digital object is ... referenced and identified by a persistent identifier ...

**PID2. RDA DFT1.2:** 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.

**PID3. RDA DFT1.3:** A PID record contains a set of attributes stored with a PID describing DO properties.

**PID4. RDA DFT1.4:** A 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 (location, fingerprint, etc.).

**PID5. FAIR-F1**: (meta) data are assigned a globally unique and eternally persistent identifier

**PID6. FAIR-A1**: (meta)data are retrievable by their identifier ...

**PID7. RDA-PIT1**: PID systems should support the generic PIT API where Information Types[[21]](#footnote-21) are openly registered and defined.

**PID8. RDA**: A trustworthy PID system must

* be maintained by a dedicated and reliable team,
* be based on a transparent sustainable business model,
* be provided by a non-profit organisation,
* 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 (comparable or even larger than IPv6) and
* support an openly documented API optimally supporting accepted data models.

**PID9. RDA:** The PID Record can be used to store the context of digital objects (bitstream locations, metadata, PID, rights information, landing page, etc.)

**PID10. DOI:** For electronic documents and published digital objects register a Digital Object Identifier[[22]](#footnote-22) and associate suitable information with it (such as citation metadata).

**PID11. RDA**: A PID needs to be requested as early as possible, at least at the time of registration at a trustworthy repository, at which point a PID record needs to be available.

**PID12. RDA:** 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. A fine level of granularity is recommended to anticipate future applications.

**PID13. RDA:** A metadata description contains the PID of the corresponding object. A metadata description contains the PID of the corresponding object. The PID record can contain one or more corresponding metadata PIDs pointing to different metadata versions, or different components such as for provenance descriptions.

**PID14. DOI:** A DOI needs to be registered for a published DO and it should be associated with citation metadata.

**PID15. RDA**: The PID record should include an expiration date for any digital object which are scheduled to expire. Even for digital objects that have been deleted the PID record should continue to exist, indicate deletion, and if possible point to a metadata record.

**PID16. ITU**: The platform for interoperability of heterogeneous identity management systems is an open architecture based on ITU-T X.1255 and the Digital Object Architecture ... and is capable of to offer interoperability at global level.

**PID17. ITU:** The top 5 benefits from this 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.

## ITU Context

ITU is a global organization, with 193 Member States, 800+ major industry players, academia and other institutions and organizations with the mandate to harmonise and standardise telecommunications and ICTs infrastructure, applications and services worldwide, i.e. scientific usage is just a small fraction of its coverage. ITU acknowleges the advantages of Digital Object Architectures and and its relevance to meet future challenges. In particular the Handle System is seen as a platform that is ready to serve global needs. We can refer to two documents being discussed within ITU in particular:

* Information on the relationships between Recommendations ITU-T X.1255[[23]](#footnote-23) (9/2013), ITU-T X.660 (7/2011)[[24]](#footnote-24) and ITU-T X.672[[25]](#footnote-25) (8/2010)
* Platform for Interoperability of heterogeneous identity management systems (in progress) which analyses whether the Handle system can be used as platform to integrate various identity systems such as 1D and 2D Barcodes and RFIDs

In this note we summarise the ITU findings as follows:

* The Handle System which is a core component of the Digital Object Architecture is compliant to ITU-T X.1255 because this ITU-T Recommendation is based on the Digital Object Architecture and it is governed by the completely independent DONA Foundation based in Switzerland with an international board taking decisions.
* The Handle System has as major characteristics: distributed administration/ownership, enabling interoperability of heterogeneous identity management systems, administration defined per named digital object, secured data binding over public network.
* The Handle architecture allows inclusion of both flat and hierarchical OID structures, can incorporate different OID notations, allows different OID registration authorities to manage the OIDs being created by them, allows owners to associate information with Handles and offers a robust and secure resolution system.
* The Handle architecture can be seen as a platform which enables interoperability of heterogeneous identity management systems such as 1D and 2D Barcodes, RFIDs, GPS etc. on a global scale based on its compliance with ITU-T X.1255 and its embedding in a Digital Object Architecture.

## Application Notes

The range of applications of PIDs in general and in particular Handles and DOIs are commonly known. Some infrastructures such as for example ENES[[26]](#footnote-26), CLARIN[[27]](#footnote-27) and EUDAT[[28]](#footnote-28) based their designs widely on PIDs to cope with the increasing amounts of data and enable stable referencing. Very well-known is the use of DOIs in the area of ePublishing and also publishing data for quite a while. Initiatives such as CrossRef and DataCite offer added value services on top of DOIs. First communities started applying PIDs since about 15 years for their purposes and also research organisations such as the Max Planck Society decided in 2005 already to offer a Handle Service to all its scientists. The publishing industry and the research domain were testing the Handle System, but also other suggestions such as the ARK system for a long period already understanding essential requirements.

In the context of ITU the range of applications of an interoperable identifier platform such as the Handle System is huge and in particular for the IoT scenario a robust and independent system with global dimension will be crucial. Most advanced with industrial applications is probably China where Handles are used to identify digital representations of physical objects in the supply chain of child food to be able to assess state and provenance of each individual can. Millions of PIDs are associated already now giving an impression about the future use of PIDs.

## Conclusions and Recommendations

(this needs to be written after having done the discussion phases)

**Appendix 2**

Aspects of PID Usage – Discussion Document

Carlo, Maggie, Peter, Tobias, Tibor, Ulrich

30.12.2016

Background

At the GEDE meeting in Bratislava it became obvious that everyone agreed with the crucial role for PIDs in proper data management and access. Yet a number of topics were raised that need to be discussed. It was agreed that these topics should be part of the discussions of the GEDE focus group and that they should be addressed in a session at the next RDA penlary meeting in Barcelona. To move ahead communities want fast answers. Six points were mentioned

1. the related issues of granularity and collection building
   * digital objects (DOs) will be re-used and re-combined by others and we cannot predict how these objects will be used in a few years - this requires to give each scientifically meaningful object an identifier
   * DOs are not just referenced within publications, but increasingly often we will need stable references for our data processing (workflows, etc.) to guarantee reproducibility
   * there will be different strategies dependent on the discipline, the repositories storing data need to make their strategy clear
   * there seems to be a trend that people start assigning Handles at high granularity and DOIs for citable collections (climate modelling, linguistics, etc.)
   * in some labs it is already common practice to create virtual collections which are just some metadata and a whole set of PIDs pointing to DOs; collections themselves get assigned a PID
2. how to do versioning
   * some repositories use an attribute in the PID record to refer to the previous and/or subsequent version; if these attributes are typed also machines can use the information
   * other repositories use metadata records to include this information which is probably not as efficient as using the PID record
3. which kind of metadata to use in PID records
   * there is an urgent need to discuss this - a session should be organised at the Barcelona plenary
   * it is about defining a set of types, but there is no obligation to use them all
   * it is generally agreed that one should not overload the PID record
4. binding role of PIDs
   * it is obvious that we are increasingly dependent on PIDs - thus we need to work towards a stable system that is well maintained, redundant etc.
   * if we have such a system we can use the PIDs to bind various types of information (bit sequences, metadata of different types, landing pages, etc.)
5. semantic categories.
   * there is a need for using Persistent Identifiers for referring to concepts and/or categories used in specific disciplines.
   * it is not obvious which kind of references should be used to refer to semantic categories
   * the semantic web community suggests to use cool URIs
   * there are existing practices in the communities which need to be respected; in biodiversity quite a number of schemes are being used, but yet not in a systematic fashion - they are looking for an overarching schema to overcome fragmentation
6. when to assign PIDs
   * for some digital content it is obvious that they are subject to changes, therefore the question is raised when (small versus major changes) one should assign a new PID to a changed object
   * in some communities people work on such DOs and carry out many changes without “registering” a new version so that it can be accessed etc.
   * possibly the use of versionable databases in conjunction with assigning PIDs to queries - as already suggested by an RDA working group - can address this issue, but not all communities feel this is practical or implementable
   * also in this case the repositories and/or communities need to indicate which policies they follow
   * in some cases it may even be useful to assign PIDs before uploading content into a repository[[29]](#footnote-29) - however then problems may occur (what about relevance and accessibility of data on notebooks etc.)

Discussion

In this section we want to address these six issues and describe answers to provoke further discussions. For some topics several options can be given, however for future machine usage and complexity reduction it is important to narrow down the solution space.

1. the related issues of granularity and collection building

* The large data volumes we are faced with and the type of operations require to work with digital collections. Digital collections are complex Digital Objects, i.e. their content refers to other Digital Objects and they are associated with a persistent identifier and metadata enabling searchability, accessibility and reusability.
* Digital collections can thus be used to combine numbers of other digital objects whose bitsequences can be stored somewhere and to give it a referenceable and citable[[30]](#footnote-30) representation. In this very basic form, collections can help with digital object management even at the beginning of the scientific data life cycle, where the preservation status of objects is still unclear, yet large-scale automated operations must be executed by the data infrastructure machinery. In principle a PhD Student could hide the whole data set that is being used for his thesis behind one Digital Collection and just add one citation to the reference list. However, managing such large and heterogeneous collections may not be a simple task, using stable and coherent sub sets as collection elements in a recursive way seems to be more realistic.
* Using the collection mechanism makes it easy to define granularity at a low level. Since data re-use cannot be predicted it makes sense to define granularity by scientific criteria: a Digital Object should be the atomic set of data values that is scientifically meaningful and that should be referenceable in different contexts such as “a brain scan”, “a video recording of an interview”, “the set of values measured by a sensor in a specific experiment under specific conditions”, etc.
* Fragment identifiers added to PIDs can refer to a subset of data values (such as a number of frames in a video), however they require private resolution operations and cannot be mixed with generally resolvable references and citations. In addition, later operations would allow to create more granular digital objects, if this would be required by science.
* Of course there are problematic use cases such as dynamic databases where cells are filled in and changed at random or sensors generating endless streams of data where segmentation is being introduced arbitrarily. In such cases granularity needs to be defined by applying pragmatic criteria.
* Granularity should be defined by what scientifically meaningful sets of values are that will be used together. Granularity can be adapted at later stages where it turns out to be necessary.
* Important is the availability of a flexible collection builder that allows every researcher to easily combine digital objects (virtually) to sets of data that will be subject of operations or references.
* Communities and or repositories need to make statements about how granularity is being chosen to facilitate re-use.

2. how to do versioning

Versions of Digital Objects are always different Digital Objects, i.e. they have a different checksum, thus are associated with different PIDs and metadata. However different models need to be distinguished.

Model 1 (versioning)

* In this model Versions come into being due to content changes over time.
* Creators may want to express the inherent sequential logic behind versioning by attributes in the PID record. Specific attributes with a well-defined type could refer to the PID of a previous version and another one to the PID of a later version. Humans and machines could make use of this typed information.
* Some communities prefer to store versioning information as part of regular metadata that is also used for search and discovery task. Such an approach may reduce accessibility and persistence of versioning information. A better approach is to store basic versioning information (typed links between objects) close to the objects and PIDs, and keep more sophisticated information (such as associated agents or descriptions of content changes) as part of the more loosely coupled metadata.
* There are many Digital Objects such as a lexicon for a specific language which are inherently dynamic, i.e. the lexicons are continuously being updated. It is a matter of policy when a new version will be registered as an "official release", i.e. many changes will be carried out before a new version will be made public.

Model 2 (representation)

* In this model we assume the existence of a digital work which is represented in different expression formats such as an image encoded as JPG, TIF, PNG, etc. They are different Digital Objects with different checksums, which seems at a first glance similar to versions, but should not be confused.
* The question of representations is orthogonal to versioning aspects and should be modelled as such. For every distinct object version, there can be multiple representation formats, and the choice of formats can also differ between versions (for example, an obsolete format may vanish and be replaced by a new one).
* There are different ways to indicate the relationships with the help of PIDs. One way would be to create a collection with PIDs pointing to the collection members. Here again a type such as “presentation\_type” could allow humans and machines to find the different elements. Type information in the PID records would encode which encoding system is being used to implement for example content negotiation strategies.

3. which kind of metadata to use in PID records

A few initiatives within RDA are now active to define useful metadata categories that could be included in the PID record. GEDE should participate in the coming session at the RDA Barcelona plenary and prepare in a discussion the kind of information GEDE members would like to find in the PID records. A few general recommendations can be made.

* PID records should include so-called “state information” which includes metadata elements that describe the state of a digital object (some people also use the term “system metadata” in this context.
* PID records should include so-called “binding information”. The assumption is that we have persistent PID resolution systems in place. If we rely on robustness and persistence, we can add essential information to the PID such as where to find the bit sequences, where to find the metadata, where to find provenance information, where to find landing pages, where to find rights information etc.
* PID records should not be overloaded to not slow down the resolution machinery[[31]](#footnote-31). However, adding pointers and state/binding information just means to add a few bytes which will not harm functioning. Furthermore the definition and use of more complex types may shift the workload from the PID server to the client side.
* The process should define a number of attributes that can be included in the PID records.
  + Not every service provider has to use all of the defined types, but if they want to integrate a certain information type they should use the well-defined types to create interoperability.
  + In addition to the defined types service providers could add additional ones, however, these are not agreed upon and machines will have problems to interpret them.
  + Obviously we need a body within RDA to decide about which information types should belong to the types registered in a Data Type Registry.
* It seems that there is a trend that service providers are interested in their special profiles (their sets of attributes supported, recommended, excluded and/or required), that could be defined on a prefix level. Such a profile could be made public and transparent as an accordingly defined information type of the prefix record. This should be standardised and the Barcelona plenary could be used to make steps here.

4. binding role of PIDs

As already introduced under 3 an increasing number of repositories is relying on the robustness and persistence of the PID resolution system and it is obvious that we need to do even more to guarantee sustainability. Given that we can rely on robustness and persistence, we can add essential information to the PID such as where to find the bit sequences, where to find the metadata, where to find provenance information, where to find landing pages, where to find rights information etc. Such information is crucial for finding, accessing, re-use etc. Consequently it becomes also crucial to guarantee the stability and interoperability of the Type Registries, where the types that deliver this information are defined.

5. semantic categories

The Semantic Web community was and is highly active and defined a number of mechanisms to efficiently deal with complex semantics. A whole range of web-based knowledge representation standards have been designed and are in use such as RDF, SKOS etc. Technologies to aggregate semantic assertions such as triple stores and to do complex queries on such triple stores such as SPARQL have been developed and are being optimised. Finally, the Linked Open Data needs to be supported.

One of the core principles in these semantic web technologies is that URIs are being used to uniquely and persistently refer to the entities amongst which are semantic categories defined in category registries such as for example Dubin Core (example: <http://purl.org/dc/elements/1.1/contributor>). The definitions of the category "contributor" defined in the DC name space can be found using the specified URI which is using the PURL uniform resource loator to redirect to thebundle location of the requested web resource. In principle PURL is an indirection method as the use of the Handle System is. It shifts persistence from URLs which are known to be not persistent to the PURL service, i.e. if the URL of a web-location is changing it requires just a change in the central PURL database. We need to rely on the persistence of the PURL redirection service. Other communities such as the language community relied on the persistence of the URL for their categories since it is granted by ISO (<http://www.isocat.org/>).

bundle

While the PURL service for example uses the standard redirection method to generate the URL the Handle System requires to specify the location attribute it needs to be resolved to.

As indicated, there are many existing practices for referencing to semantic categories in the different scientific disciplines as well.

This approach is closely related to the metadata categories discussed in section 3. In particular versioning as in chapter 2 can give an example of this relation: a derived version of an object could have a typed entry of 'isNextVersionOf' together with its value. This way would result in a triple of PIDs referring to the new DO, the type definition of 'isNextVersionOf' and the old DO as S-P-O triple.

A discussion at the Barcelona plenary session could be used to discuss this issue in detail. Whatever is being done it needs to be compliant with the Semantic Web suggestions and needs to accept the existing solutions.

6. when to assign PIDs

A number of recommendations can be given:

* Digital Objects should be registered (assigned a PID, store pointers to metadata etc.) once it is uploaded to a trustworthy repository, i.e. it is the repository that requests a PID and updates all related attributes.
* In general it does not make sense that individuals request PIDs, since they cannot guarantee accessibility and stability of data.
* Repositories need to define their policies when and how they assign PIDs and update attributes stored together with PIDs.
* Researchers should be motivated to register their data (and other digital types) as early as possible to enable proper referencing and citing. However in case of versioning of dynamic data an explicit release strategy makes sense. Dependent on the types research communities should specify recommendations.

**Appendix 3**

List of Contribution Documents

|  |  |  |
| --- | --- | --- |
| Abbrev | Source | Reference |
| RDA DFT | RDA Data Foundation & Terminology WG | DFT Core Terms and Model  <http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318> |
| RDA PIT | RDA PID Information Type WG | <https://www.rd-alliance.org/groups/pid-information-types-wg.html> |
| RDA | RDA Data Fabric IG Discussion | <https://www.rd-alliance.org/group/data-fabric-ig.html> |
| FAIR | FAIR Principles | <https://www.force11.org/group/fairgroup/fairprinciples> |
| 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> |
| USE | GEDE Usage Document | Appendix 2:Aspects of PID Usage – Discussion Document |
| DOI | DOI Documentation | <https://www.doi.org/> |
| ITU | ITU Documentation | ITU-T X.1255: <https://www.itu.int/rec/T-REC-X.1255-201309-I/en> |
| BIO | "10 Simple rules for design, provision, and reuse of persistent identifiers for life science data" | <https://zenodo.org/record/18003#.WJHDt7YrJbU> |
| COR | CORBEL Consensus document on providing access to individual participant data | <http://www.corbel-project.eu/> |
| CES | CESSDA Documentation | communication by Email |
| CLA | CLARIN PID policy summary | <https://www.clarin.eu/node/3965> |
| CMIP | Persistent Identifiers for CMIP6: implementation plan | communication by Email |
| LSID | TDWG Globally Unigue IDentifiers (GUID) | <https://github.com/tdwg/guid-as> |
|  |  |  |

In addition the following documents were 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 4**

Elaboration on PID Forms

Peter, Larry

25. March 2017

This is a short elaboration on PID forms to prevent confusion. It is partly terminology and also a difference in concepts.

# Life Sciences Document

The document from the life science colleagues ("10 Simple rules for design, ...") 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*](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 URL "<http://zfin.org>" which will take care of resolving steps will 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.

ZFIN

Server

web

ZFIN

Database

full URI

returning landing page address

A landing page address is being 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".*

*An example for a Handle is:* [*https://hdl.handle.net/11304/a3d012ca-4e23-425e-9e2a-1e6a195b966f*](https://hdl.handle.net/11304/a3d012ca-4e23-425e-9e2a-1e6a195b966f)

*An example for a DOI is:* [*http://doi.org/10.23728/b2share.3d2296cd14e74e74b9c960a2fafb5ff5*](http://doi.org/10.23728/b2share.3d2296cd14e74e74b9c960a2fafb5ff5)

The string ["a3d012ca-4e23-425e-9e2a-1e6a195b966f](https://hdl.handle.net/11304/a3d012ca-4e23-425e-9e2a-1e6a195b966f)" must be unique within the name space "11304" in the first 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. exist as long at least as the HDL is not accepted as a standard web-protocol.

full URI

Handle

Proxy

web

Handle

System

returning state information

Typed state information is being returned to the caller which will lead into action.

# Differences

Syntactically there is hardly any difference except for the delimiter separating prefix and suffix and the address space is unlimited. With respect to processing there are a few differences:

* 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 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 anymore and each Handle would be directly be 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 amount of communities makes it easier to keep the domain names and services constant[[32]](#footnote-32).
* 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>#<part id>*

It should be stated clearly that a PID identifies a Digital Object as a whole, making it referable and citable. The part 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. He could extract this fragment from the video and give it 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 5**

Glossary

yet to be finished

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Digital Object | RDA DFT |  |
| Research Data Object |  |  |
| Digital Data Object | RDA DFT |  |
| Data Object |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

1. as long as the HDL protocol has not been accepted by IANA [↑](#footnote-ref-1)
2. Experience shows that only cross-sector and cross-disciplinary systems will have a chance to survive due to the enormous investments required including the support of layered services. [↑](#footnote-ref-2)
3. The exit strategy can be embedded in the system architecture by having a redundant setup. [↑](#footnote-ref-3)
4. DOIs are defining a community of use on top of the Handle System with additional specifications of policies and business models. [↑](#footnote-ref-4)
5. The term "sectors" covers science, industry, governments, health care, etc. [↑](#footnote-ref-5)
6. DOs can be of different types: data, metadata, collections, queries, software, workflows, configurations, etc. [↑](#footnote-ref-6)
7. Landing pages are XML structures that have machine readable sections. [↑](#footnote-ref-7)
8. State information can be interpreted as systems metadata. [↑](#footnote-ref-8)
9. The exact formulation is still subject of discussion. [↑](#footnote-ref-9)
10. Comparable or even larger than IPv6. [↑](#footnote-ref-10)
11. In the elaboration in appendix 4 we briefly indicate that there are only minor principle differences between this statement and how the Handle/DOI system is organised. [↑](#footnote-ref-11)
12. LSID seems to be not supported anymore due to not sufficient uptake. They were/are using the URN scheme: urn:lsid:authority:namespace:object:revision. For more comments see: <http://dev.mygrid.org.uk/blog/2016/02/what-exactly-happened-to-lsid/> [↑](#footnote-ref-12)
13. The term "eternal" is not convincing for some communities. [↑](#footnote-ref-13)
14. Use cases need to be given to make this assertion understandable and agreeable. There seem to be cases where a high granularity cannot be recommended. Later subsetting of data is possible, but practically questionable. [↑](#footnote-ref-14)
15. For the case of Handles/DOIs it would be extremely helpful to have it accepted as an internet protocol, since then no URL would have to be specified. The Handle/DOI includes the prefix ass namespace. [↑](#footnote-ref-15)
16. Some PID schemes allow to add a version indicator at the end of a PID separated by a dot: <PID>.<version> [↑](#footnote-ref-16)
17. FAIR combines "searchable" and "finding" under one the term "finding", i.e. metadata used for searching must include a PID allowing to find the bit sequences. [↑](#footnote-ref-17)
18. There is a growing agreement that Handles issues by trustworthy service providers should be used as early as possible in data management to enable stable referencing for example in workflows. DOIs however should be used when data collections are being published, i.e. in general this requires for example quality checks and additional metadata. [↑](#footnote-ref-18)
19. This term resulted from the discussions on concrete configuration building in RDA’s Data Fabric Group. [↑](#footnote-ref-19)
20. DFT = RDA Data Foundation and Terminology Group (<https://rd-alliance.org/groups/data-foundation-and-terminology-wg.html>) ; PIT = RDA PID Information Types group (<https://rd-alliance.org/groups/pid-information-types-wg.html>); FAIR = <https://www.force11.org/group/fairgroup/fairprinciples> [↑](#footnote-ref-20)
21. Information Types are attributes associated with a PID describing DO properties. [↑](#footnote-ref-21)
22. Technically spoken DOIs are Handles with prefix 10 [↑](#footnote-ref-22)
23. X.1255 is a framework for the discovery of identity management information and it describes an open architecture framework based on the Digital Object Architecture where a PID system is central. [↑](#footnote-ref-23)
24. X.660 defines a tree structure that supports international object identifiers (OIDs) [↑](#footnote-ref-24)
25. X.672 specifies an OID resolution system. [↑](#footnote-ref-25)
26. <https://is.enes.org/> [↑](#footnote-ref-26)
27. <https://www.clarin.eu/> [↑](#footnote-ref-27)
28. <https://www.eudat.eu/> [↑](#footnote-ref-28)
29. It may help to define the term "repository" as something "simple": a "repository" is an entity whose primary tasks are to provide services to access digital object content and essential state information, given an object’s PID, and to enable reliable and trusted data management. [↑](#footnote-ref-29)
30. To ultimately enable citation of collections, additional metadata must be gathered. [↑](#footnote-ref-30)
31. Actually the typing has much less impact on the resolution than on minting process, especially if policies about allowed types are involved [↑](#footnote-ref-31)
32. 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. [↑](#footnote-ref-32)