Recommendations for Implementing a Virtual Layer for Management of the Complete Life Cycle of Scientific Data

January 2017

Edited by: Tobias Weigel, Peter Wittenburg,
Supported by: Bridget Almas, Reinhard Budich, Sandra Collins, Michael Diepenbrook, Ingrid Dillo, Francoise Genova, Frank Oliver Glöckner, Rebecca Grant, Wilco Hazeleger, Margareta Hellström, Keith Jefferey, Franciska de Jong, Tibor Kalman, Rebecca Koskela, Dimitris Koureas, Wolfgang Kuchinke, Leif Laaksonen, Larry Lannom, Michael Lautenschlager, Damien Lecarpentier, Jianhui Li, Jay Pearlman, Luca Pezzati, Ralph Müller-Pfefferkorn, Beth Plale, Stefano Nativi, Raphael Ritz, Ulrich Schwardmann, Rainer Stotzka, Achim Streit, Dieter van Uytvanck, Anwar Vahed, Doris Wedlich, Colin Wright, Ramin Yahyapour, Thomas Zastrow, Carlo Maria Zwölf

This note does not deal with research infrastructures in the general sense, but only with those aspects that are related with data. Appendices refer to details.

1. Recommendations Summary

We are at a tipping point in the development of a common conceptual framework and set of tools and components which will revolutionize the management of scientific data. It is widely acknowledged, as detailed below, that the current volumes and complexity of data now being collected and, even more so, the inevitable and enormous increase in that volume and complexity, have reached the point where action is required. At the same time, and largely in response to this perceived crisis, a number of principles for the management of scientific data have arisen and been widely endorsed. The danger now is that agreement will stop at the level of principles and that multiple non-interoperable domain and technology specific silos will continue to arise, all based on the abstract principles, and we will lose the opportunity of leveraging the current crisis to create a common set of tools and components based on an agreed conceptual approach.

There is wide agreement on a set of principles. Action is now required to put in place operational components of a common infrastructure.

What follows is our summary of the current agreed-upon principles, a more detailed analysis of the requirements implied by those principles, and the current state of work on those requirements, as reflected in the work of RDA, which we believe has the broadest base and most neutral view of the situation. This includes brief summaries of the requirements and the current state of work on repositories, registries, identifiers, metadata, types, licenses, and, in general,

The real risk at the moment is in not building a common core infrastructure according to our best current information.

the whole ecosystem of interlinked digital objects needed for managing the life cycle of scientific data. We end with a more detailed view of the requirements for selected components extending partly the FAIR principles (Findable-Accessible-Interoperable-Reusable, Appendix A).

Action is now required to put in place operational infrastructural components based on this and similar analyses. Some of these components already exist at an operational level with wide experience across communities, while others are yet at a prototype or concept stage, i.e. no design from scratch is intended and we can build on extensive knowledge built up in various regions. We should now install a systematic approach where these components can mature and ultimately enable communities to build new services and proof that added value can be achieved by means of combination of components. There will, of course, be a risk in doing this and some of these components will surely fail or otherwise prove inadequate. In some cases waiting another five or ten years would perhaps result in better designs and implementations based on technology advances between now and then but by that time the interoperable silo problem will have gained ground and be difficult to displace. The real risk at the moment is in not building a common core infrastructure according to our best current information.

Many core infrastructure components are already in use. We need to validate these components, encourage their use, connect the components, and begin building a common core infrastructure.

In addition to the current widely adopted recommendations by funders the essence of which recommendations will come next can be summarized as:

1. *Digital objects should be stored in trustworthy repositories that are assessed regularly using DSA/WDS[1] guidelines and those repositories should be registered in open registries such as re3data[2].*
2. *Trustworthy repositories need to assign PIDs to all digital objects and register them with trustworthy PID service providers, such as the International DOI Federation[3] and the European Persistent ID Consortium for eResearch[4], that guarantee their resolution to meaningful state information.*
3. *The digital objects referenced in points 1 and 2 above are not restricted to the data itself but also include schemas, queries, concepts and concept vocabularies, all of which need to be registered in open registries and assigned PIDs if they are cited or referenced.*

There are still many issues to be explored and questions to be answered, but we believe that science would be well-served if future scientific data infrastructure projects accepted and followed these high level recommendations.

[1] https://www.rd-alliance.org/groups/repository-audit-and-certification-dsa%E2%80%93wds-partnership-wg.html
[2] http://www.re3data.org/
[3] http://www.doi.org/doi_handbook/7_IDF.html
[4] http://www.pidconsortium.eu/

2. State of Recommendations

There is wide agreement that on the one hand data volumes and complexity of scientific data are increasing exponentially and that on the other hand cross-border (disciplines, countries) re-usage is an emerging paradigm in almost all research fields. As a result of these developments a number of deficits have been identified:

- enormous inefficiencies (80% of data inaccessible after short periods, 80% of specialists time wasted for typical data management tasks)
- lack of conceptual frameworks to unify infrastructures and best practices resulting in huge fragmentation and lack of interoperability
- lack of experts who could help in a fast transition to common and efficient practices

Faced with these developments a number of recommendations have evolved and consensus is now being reached:

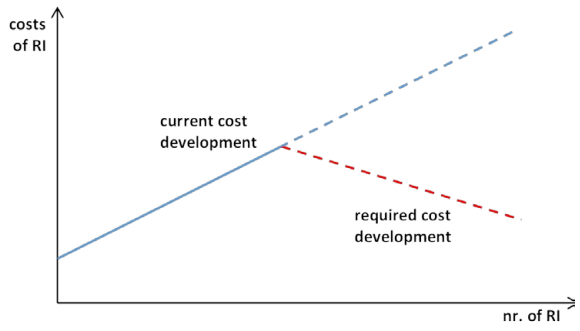
- projects need to submit and then follow **Data Management Plans** that cause researchers to take data management seriously
- **Open Data** is an agreed default principle
- the **FAIR principles** (appendix A) have been agreed widely as basis for proper data management
- it is agreed that a certain percentage of the funds of grants must be devoted to data **curation and stewardship**, even following project completion

In particular the FAIR principles, the result of broad discussions in various forums[5] have the potential to lead to improvements. At its core are statements about Digital Objects stored in trustworthy repositories, assigned PIDs, and associated with metadata:

- All Digital Objects (data, collections of data, metadata, software, configurations, etc.) need to be registered in trustworthy repositories[6] to make them findable, accessible and persistent.
- All such registered DOs must have assigned Persistent Identifiers (PIDs) and Metadata (MD) to improve interoperability and re-use.

3. Challenges Ahead

The highest priority for actions is **now** the need to put these recommendations in practice to make data management, access and curation much more efficient. Yet we are still missing a compelling conceptual framework to create global momentum. Currently the number of research infrastructures is growing worldwide with many components being re-invented in different flavours. The costs for building data infrastructure are thus increasing, as are the costs for data re-use, due to the huge fragmentation. Therefore, this trend needs to be counteracted by identifying common components and means of interoperability and then implementing them in a cost-effective manner, as illustrated by the diagram. Increasing numbers of experts agree that we are at a tipping point (see also appendix B) whereby new approaches are required to create a global momentum



[5] We refer here to the discussions at ICRI 2012, within the G8 group and within the RDA. Similar statements can even be found in the Report "Managing Access to Digital Information" formulated by a cross-industry team of delegates from 47 of the most relevant IT companies in 1997!

[6] We understand a "repository" as 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.

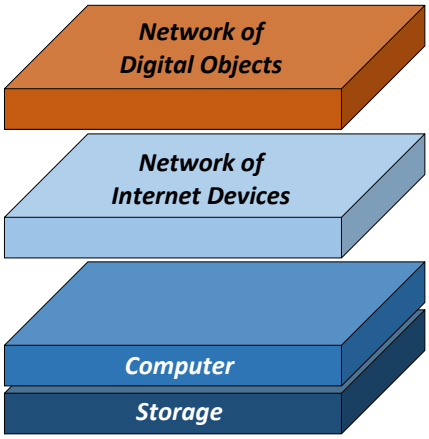
comparable in many ways to the creation of the Internet. The European Open Science Cloud is one such important step to create this momentum, yet it needs to be empowered by an implementation concept.

There is no doubt that compute, storage and network capacity are essential resources for data-driven science and each society that wants to remain competitive needs to invest continuously in maintaining and upgrading these resources. Yet to support data-driven science this is not sufficient. We need to address the layer of data organisation, typing and re-use facilitation to be able to make use of the value of data. We are lacking a "virtual layer of stable and robust components" which is accessible for common usage, enables the increase of automated data processing by machines, and guarantees reproducible, efficient and more interoperable data science in future. This layer is currently dominated by discipline specific views and solutions in science and by the huge numbers of cloud solutions in industry.

When designing this virtual layer we need to not just look at the amounts and complexity of data we see now, but we need to anticipate the coming challenges that are defined by the estimated 50 billion smart devices being used for scientific, societal and industrial applications. Developing a smoothly functioning virtualisation layer to connect Digital Objects and enable type-based processing will allow us to fight the emerging data monster. The challenges in science and industry are now so dramatic that we need to start working on the virtualisation layer now and thus get beyond the agreed-upon principles.

4. Virtual Layer Requirements

A few decades ago a network layer was designed on top of the computers and other edge devices so that such devices could exchange messages. This network was based on IP numbers, a number of protocols such as TCP, and, after its wide acceptance, a number of global registries. Thus a network of Internet devices was established that created enormous momentum for science, industry and society. All kinds of experts could suddenly turn all their energy into other aspects layered on top of the Internet infrastructure, such as the World Wide Web, and leave the details of computer networking to the experts. We are so dependent on a functioning Internet that all stakeholders participate in maintaining a stable and robust infrastructure despite knowing that the chosen protocols have their weaknesses.



We are now in the phase where we need to interlink data professionally with the help of PIDs and a number of protocols and registries. The landscape of registries necessarily is more complex than the one for the Internet.

A few decades ago a first virtualisation layer was added by stating that the computers we use do not have to be local, but can be accessed via network protocols. The IP system identifies all devices in the network so that they can exchange messages using some protocols and registries.

CPU and storage layers have existed from the start of modern computing. They have undergone technological and architectural changes, but the concepts have basically remained stable. Cloud stores in their basic form are new types of stores where fast logic translates internal IDs into internal paths.

We are now far enough along that we recognise the significant need of a layer interlinking Digital Objects. We feel that we need momentum to overcome all the fragmentation and in doing so create another innovation wave – as positively disruptive as the previous network innovation. What will be the basic components of such a virtualisation layer to put the principles mentioned above into practice? We can describe some of the components that will be required to implement the FAIR principles and evolving frameworks such as being developed in the RDA DFT[7] and Data Fabric[8] groups knowing that some of the specifications need further clarification, that in some cases specifications may change and that further components will be identified as essential for the functioning of the virtualisation layer:

1. a network of **trustworthy repositories** (*T-REP*) that are available for every researcher to register, store and manage data, that have a clear interface to access DOs and that are certified to guarantee a certain quality of service,
2. a **trustworthy registry of such T-REPs** that is human and machine readable to enable efficient and goal-driven access,
3. a system to **register and resolve PIDs** available for every researcher that we can rely on and that offers adequate security mechanisms,
4. a system to **register types** of DOs allowing machines to relate actions with types as the basis of automating data processing,
5. a definition of a set of **core types** that are being used to describe the state of DOs enabling machine action,
6. a system to **register metadata schemas** and **metadata descriptions** to enable re-use and machine processing,
7. a system to **register concepts and concept vocabularies** to enable re-use and machine processing,
8. a system of **authorisation record registries** to enable efficient access control in large federations of repositories,
9. a system of **license registries** to efficiently deal with licenses and their acceptance,
10. an ecosystem of **tools and operating procedures** that enable data service providers to efficiently manage digital objects as part of their daily business and collectively populate the digital object network layer and to put data into context (software, publications, etc.).

It is obvious that a number of characteristics must be associated with such systems. They need to be global, stable, robust and secure so that we can rely on their long term availability as we rely on the availability of the Internet infrastructure.

Virtual Layer: Current State and Needed Actions

Most of these components have been tested in different varieties in recent years so that we have collected a deep knowledge about their pros and cons. It is now time to critically look at those varieties and make choices so that efficiently functioning systems[9] can be set up and maintained and that software builders can systematically include the components in their software. The challenge is not to design yet other components in most cases, but to enable systematic implementation and use. In principle all of these components need to be certified regularly to make sure that we can rely on them - yet assessment rules have been developed only for repositories.

Items	Comments
Trustworthy Repositories	it is known how to set up DFT certified repositories supporting a high level of FAIR-ness, it is a matter of consequent realisation

[7] https://www.rd-alliance.org/groups/data-foundations-and-terminology-ig.html
[8] https://www.rd-alliance.org/group/data-fabric-ig.html
[9] The term „systems“ is used here to indicate that we talk about distributed solutions that can for example include meta-registries to make it functionable for machines.

	and funding support to make the network available for everyone
Assessment of Repositories	DSA/WDS[10] is in use and available, but probably will evolve over time and applied by all repositories which want to appear as serious repositories
Repository Registry System	with re3data[11] a registry is available for human processing, a system needs to be added for machine processing and an RDA WG has been started
Repository API	a new RDA group and others have defined an API or are in process of defining one based on earlier experiences
PID registry and resolving system	the Handle System[12] is tested and available, service providers need better funding support
Type registry system	the specifications have been worked out by an RDA WG, but still much work will be necessary to make it professionally usable
set of core types	this is currently being investigated based on earlier work within an RDA WG
MD schema registry system	there are many registries - one has been setup by an RDA WG, we need to turn the many flavours to a usable and maintained system
MD registry system	same
Concept Registry	same
Vocabulary Registry	same
Authorisation Record Registry	this needs to be developed urgently
License Registry	this needs to be developed urgently, there is a first test solution in Finland
Ecosystem of Tools	this does not yet exist, but is urgently needed, it will include a wide range of services from simple operations such as "replication" to "complex brokering"

5. Additional Requirements for Selected Virtual Layer Components

In addition to what is being described by the FAIR principles we can derive a number of further principles that should be taken care of. These are partly more detailed specifications of what the FAIR principles describe and are centred on topic areas, driven by specific technological perspectives on the common components stated above. Those from the recommendations that are currently being discussed in RDA (appendix C) that are widely agreed are summarised here:

Persistent Identifier View

- A persistent identifier (PID) needs to be supported by a sustainable and trustworthy resolution system that will resolve PIDs to meaningful state information for machines and humans which are metadata attributes describing essential properties of a Digital Object (DO).

[10] DSA and WDS joined to create one set of rules to assess quality of T-REPs. A working group in RDA is working on an interface for repositories.

[11] http://www.re3data.org/

[12] With the Handle System a global, independent, robust and secure system is ready to be used and for years experts already use it in form of DOIs and for example EPIC Handles. The stakeholders need to make sure that the system will be stable over time. Some communities will continue to work with legacy ID systems, but may map their IDs with Handles to make use of one global reference system.

- A trustworthy PID resolution system needs to fulfil quality criteria still to be defined and needs to undergo regular quality assessment.
- The persistent PID record should be used to persistently bind the context of digital objects.
- A PID should be assigned to a Digital Object when it is registered at a trustworthy repository and thus becomes part of the domain of visible and findable data.
- A DOI should be registered when Digital Objects (data) are being published and citation metadata should be associated with it.

Repository View

- The bit sequences of DOs need to be stored in trustworthy digital repositories.
- A trustworthy repository needs to be assessed regularly to guarantee the correctness of their processes and procedures. The joint requirements defined by Data Seal of Approval and World Data Systems are the basis for a successful assessment.
- A trustworthy repository specifies APIs that allow programmed access to the digital content stored.

Interoperability[13] and re-usability View

- Data needs to be structured according to open and registered standards
- Schemas structuring data need to be registered in open registries
- Data and metadata should be exposed in form of broadly used knowledge representation languages where applicable
- Concepts and vocabularies should be registered in open registries
- Data Types should be registered in data type registries and linked with interpretation procedures to enable automatic processing by machines

[13] The FAIR principles speak about a "knowledge representation" language for data representation. The term is being defined by AI and used by the Semantic Web, but as such is not applicable to large domains of data.

Appendix A: FAIR Data Principles

To be Findable:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

To be Accessible:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
A2 metadata are accessible, even when the data are no longer available.

To be Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.

To be Re-usable:

R1. meta(data) have a plurality of accurate and relevant attributes.
R1.1. (meta)data are released with a clear and accessible data usage license.
R1.2. (meta)data are associated with their provenance.
R1.3. (meta)data meet domain-relevant community standards.

Apendix B: Global Digital Object Cloud (GDOC) – A guiding vision

This appendix is an updated and slightly edited version of an original discussion document from September 2016 that worked out the details of the GDOC idea, concept and possible implementation pathways. The original document was published here: http://hdl.handle.net/11304/13e717c7-22c2-45f0-90c6-938c579d891a

Authors: Larry Lannom, Peter Wittenburg, Tobias Weigel

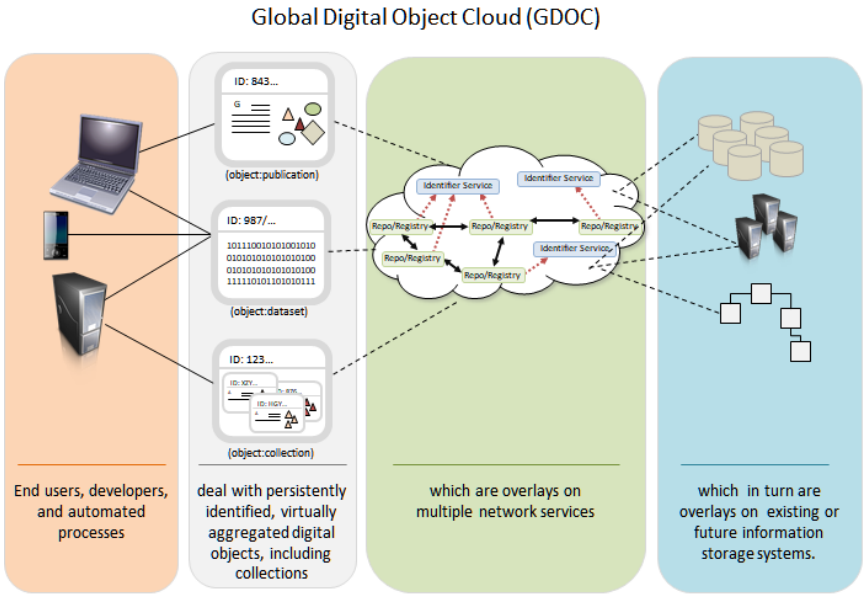
Background

It is now widely agreed in the Research Data Alliance (RDA) and beyond that concept of the "Digital Object (DO)" is central to proper data management, access and use. A DO has a bit sequence that can be stored in multiple repositories and is associated with a Persistent Identifier (PID) and quality metadata [1]. The persistent identifier, augmented with usable attributes about the DO, can provide identification, location, and other functions. The metadata description has many different functions amongst which is offering information allowing proper interpretation and reuse of the bit sequences. Due to the claim of persistence PID records are increasingly seen in a binding role, i.e., storing persistently all the necessary actionable references to the locations of the bit sequences, to the metadata and other useful information. Given such a domain of registered DOs it is compelling for users to just deal with PIDs and metadata as widely as possible. We call this layer of virtualization the Global Digital Object Cloud (GDOC), which is based on the ideas of a Digital Object Architecture [2] and fully compliant with the FAIR principles [3].

Global Digital Object Cloud

The concept of the Global Digital Object Cloud is illustrated in the figure below. Key to this model are Digital Objects, which comprise a virtualization layer on top of various network resources and services, much as files and databases currently are virtualization layers on top of raw computer storage and sets of standard processes, but in this case extended to the network level. Each Digital Object is persistently identified, with the persistence guarantee dependent on use case, such that every object on the network can be referenced and, given adequate permissions, can be the subject of stated operations, including but not limited to raw access. Further, the objects are described and typed by metadata such that their structure can be well understood through the mechanisms of type registries. These objects are shown in the second panel from the left in the figure.

The ability for clients – both humans and machines – to call, access, and act upon these objects is provided by the network services shown in panel three of Figure 1. Here the set of repositories, registries, and identifier resolution services provide the structure and processing that enable the array of storage and specific data management services, shown as the rightmost and lowest level panel of Figure 1, to be consistently addressed as coherent objects. The repositories essentially serve as unifying portals into the lower level storage and heterogeneous information management technologies and they have the responsibility of presenting the consistent structured object view to clients, regardless of the details of the underlying data stores and management systems. The registries provide discovery services for objects, by providing searchable metadata, as well as providing the information needed to access and act upon objects, through registration of types and related services that are relevant to specific types. The identifier resolution services enable the objects to be directly addressed, regardless of their current state and location, again depending on permissions as well as the capabilities of the individual objects and the repositories providing the portal service.

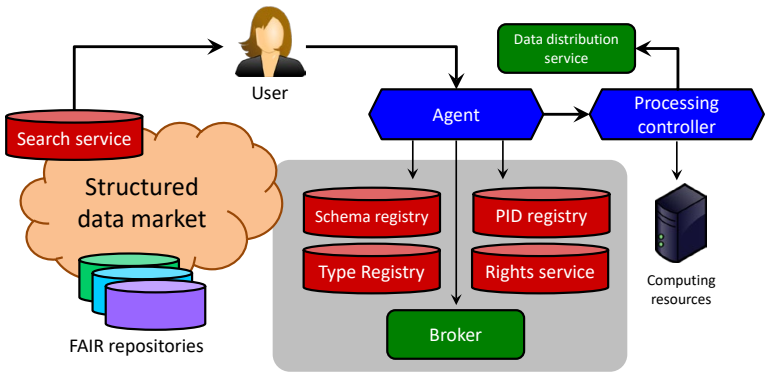


Note that registries, repositories, and resolution services are functional categories and not necessarily software categories. Metadata registries, for example, are best regarded as metadata object repositories and the same software-based service can provide both object discovery and access services.

Necessary Steps

While we are far away from a complete implementation of this GDOC model, we find that it is necessary to define such a goal, and to keep it in mind as we work towards revolutionizing our up to now inefficient approaches. One of the Data Fabric configuration options currently being pursued, for example, is a PID Centric approach to data management [4] and it embodies several principles from the model, i.e., persistent identification, the centrality of well-managed repositories, type registries and collections with respective harmonized interfaces.

A particular application area that combines multiple of the envisioned virtual layer components is type-triggered automated processing (T-TAP). The figure below illustrates the general idea, where agents acting on human users' behalf will coordinate a data processing and analysis workflow. This should happen autonomously to the extent possible, and will involve interaction between the agents and the necessary components such as PID and type registries, metadata catalogs, brokers and so on, accompanied by a structured market that provides seamless discovery and access to FAIR repositories and other resources. Towards the end of the workflow, actual computing tasks will be executed on possibly distributed HPC resources, and resulting data of particularly high volume will be delivered back to the user through adequate data distribution services.



Summary

This GDOC model does not, of course, solve any problems on its own but serves instead to guide our work in the direction of a non-proprietary and highly efficient data management and access infrastructure where globally available and stable PIDs are anchors for all activities. GDOC serves as a framework into which multiple lower-level data storage and management solutions can be placed as well as accommodating a wide variety of applications. Users deal with a domain of registered Digital Objects where they primarily just deal with entities such as PIDs and metadata as long as they do not start calculations on the bit sequences themselves. But it is the task of the GDOC middleware to transparently access and operate on the bit sequences and for example sort out which copy should best be taken or whether the data should go to the algorithms or vice versa. GDOC based on these principles opens up completely new perspectives that also have the power to involve industry in providing compliant software.

Further applications have recently been discussed within the Internet of Things (IoT) domain, where there are continuing discussions about the creation of a backend system that is powerful enough to allow identification of all the billions of data objects being created by IoT devices. As yet industry does not have a solution, but is aware that a solution is urgently needed. In consequence, a first joint workshop of the IoT Forum and RDA was held at the IoT Week 2017 in Geneva to bring RDA and industry experts together.

It is obvious though that in such PID centric models we are highly dependent on a functioning, highly available and powerful PID infrastructure which allows everyone to register and resolve PIDs to meaningful information about the DOs. Similarly to how we now rely on a worldwide IP-based infrastructure to connect computers, we need to be able to rely on a worldwide infrastructure for PIDs. The Handle System, now owned by the Swiss DONA Foundation and governed by an international board is with its multinational root nodes and numerous service providers, including DOI providers, is a strong candidate for such an infrastructure.

Given the existence and broad support for such an infrastructure trust and validation, for example, could be applied at the object level instead of depending on a variety of underlying heterogeneous systems. As has been shown by Crossref, for example, a multitude of valuable services can be implemented on top of the PID infrastructure.

Appendix C: Recommendations under RDA Discussion

This appendix summarises the recommendations that are currently in discussion within RDA groups: https://www.rd-alliance.org/group/data-fabric-ig/wiki/recommendations.html.
These recommendations are collected from various important initiatives including FAIR. Many of them are still under discussion.

Discovery/Finding Topic Area

MD1. RDA DFT-1.1: A digital object (DO) ... has properties that are described by metadata.

MD2. RDA DFT-1.2: A persistent identifier is ... persistently resolved to meaningful state information (systems metadata) about the identified DO.

MD3. RDA DFT-1.5: Metadata contains descriptive, contextual and provenance assertions about the properties of a DO. Note: To make metadata referable it needs to be associated with a PID and thus is a DO. Note: Metadata minimally needs to contain the PID of the DO.

MD4. RDA DFT-1.7: A Digital Collection is identified by a PID and described by metadata.

MD5. RDA DFT-1.13: A digital metadata repository is a digital repository that is able to store, manage and curate metadata.

MD6. FAIR-F1: metadata are assigned a globally unique and eternally persistent identifier.

MD7. FAIR-F2: data are described with rich metadata.

MD8. FAIR-F3: metadata are registered or indexed in a searchable resource.

MD9. FAIR-F4: metadata specify the data identifier

MD10. FAIR-A1: metadata are retrievable by their identifier ...

MD11. FAIR-A2: metadata are accessible

MD12. FAIR-I1: metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

MD13. FAIR-I2: metadata use vocabularies that follow FAIR principles.

MD14. FAIR-I3: metadata include qualified references to other (meta) data.

MD15. FAIR-R1: metadata have a plurality of accurate and relevant attributes.

MD16. FAIR-R1.1: metadata are released with a clear and accessible data usage language.

MD17. FAIR-R1.2: metadata are associated with their provenance.

MD18. FAIR-R1.3: metadata meet domain relevant community standards.

MD19. RDA: Metadata needs to be openly accessible and harvestable via common protocols such as OAI-PMH or ResourceSynch.

MD20. RDA: Metadata schema must be registered in an open schema registry.

MD21. RDA: Concepts used in metadata need to be registered and define in open concept registries using an accepted formal semantic language.

MD22. RDA: Metadata needs to be created as early as possible, at least at the time of registration at a repository a metada description needs to be available.

MD23. RDA: Metadata will be harvested, used and changed/extended by users dependent on the needs requiring an information in the PID where the original metadata can be found.

MD24. RDA: Metadata structure and semantics need to be in a form that service providers can turn the decsriptions into semantic assertions in RDF format.

MD25. RDA-DFT: Metadata is associated with collections which can exist of a number of digital entities, i.e. the level of granularity at which metadata will be assigned is left to the communities and repositories. A high granularity is recommended to anticipate future applications.

Repository Topic Area

REP1. RDA DFT1.9: A digital repository is an infrastructure component that is able to store, manage and curate Digital Objects and return their bistreams when a request is being issued.

REP2. RDA DFT1.13: A digital metadata repository is a digital repository that is able to store, manage and curate metadata.

REP3. FAIR-A1: (meta) data are retrievable by their identifier using a standardized communications protocol. *(this statement indirectly refers to the existence of digital repositories by stressing accessibiity via a protocol, i.e. a digital repository must offer a standardized protocol to access data)*

REP4. RDA: Trustworthy repositories are digital repositories that untertake regularly quality assessments successfully such as Data Seal of Approval / World Data Systems. *(these two initiatives harmonized their requirements)*

REP5. RDA: Digital objects need to be stored in trustworthy digital repositories.

REP6. RDA DFT: Digital repositories should have a repository software system that supports the data organisation as defined in DFT.

REP7. RDA: digital repositories should expose their characteristics and services in widely recognized schemas to enable service providers to create useful services for human and machine processing.

REP8. RDA: One of the services of a digital repository to be indicated in the schema is the metadata harvesting port supporting a standard protocol such as OAI-PMH.

REP9. RDA-PP: A trustworthy repository must specify auditable practical policies for its various tasks, turn them into executable procedures and workflows, and systematically apply them in all cases to document provenance of all its digital objects.

REP10. RDA-DFT: Data copies will reside in several trustworthy digital repositories. It is recommended to indicate in the PID record which repository is the original one and thus has authority about setting access permissions and original metadata descriptions.

REP11. RDA: The global Internet of Data is domain of registered digital objects, at registration with a digital repository a PID is associated and metadata are created.

REP12. DSA4: The data repository has an explicit mission in the area of digital archiving and promulgates it.

REP13. DSA5: The data repository uses due diligence to ensure compliance with legal regulations and contracts including, when applicable, regulations governing the protection of human subjects.

REP14. DSA6: The data repository applies documented processes and procedures for managing data storage.

REP15. DSA7: The data repository has a plan for long-term preservation of its digital assets.

REP16. DSA8: Archiving takes place according to explicit work flows across the data life cycle.

REP17. DSA9: The data repository assumes responsibility from the data producers for access and availability of the digital objects.

REP18. DSA10: The data repository enables the users to discover and use the data and refer to them in a persistent way.

REP19. DSA11: The data repository ensures the integrity of the digital objects and the metadata.

REP20. DSA12: The data repository ensures the authenticity of the digital objects and the metadata.

REP21. DSA13: The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS.

Persistent Identifier Topic Area

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. (*state information can be interepreted as systems metadata*)

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 (*properties of DOs)* 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 (*DOI, which is a Handle with prefix 10*) 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 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.

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)

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 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.

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

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.

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.

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 automtic data processing) and build the needed tools now with high urgency.

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.

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

Data Topic Area

DAT1. FAIR-F1: data are assigned a globally unique and eternally persistent identifier.

DAT2. FAIR-F2: data are described with rich metadata.

DAT3. FAIR-F3: data are registered or indexed in a searchable resource.

DAT4. FAIR-A1: data are retrievable by their identifier ...

DAT5. FAIR-I1: data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

DAT6. FAIR-I2: data use vocabularies that follow FAIR principles.

DAT7. FAIR-I3: data include qualified references to other (meta)data.

DAT8. FAIR-R1: data have a plurality of accurate and relevant attributes.

DAT9. FAIR-R1.1: data are released with a clear and accessible data usage license.

DAT10. FAIR-R1.2: data are associated with their provenance.

DAT11. FAIR-R1.3: data meet domain-relevant community standards.

DAT12. RDA: The schemas of digital objects need to be registered in open schema registries.

DAT13. RDA: Concepts used in digital objects need to be registered and define in open concept registries using an accepted formal semantic language.

DAT14. RDA: The provenance of digital objects need to be stored in provenance records making use of the W3C PROV standard and referenced by metadata.

DAT15. RDA-DFT1.1: A digital object is represented by a bitstream, is referenced and identified by a persistent identifier and has properties that are described by metadata.

DAT16. RDA-DTR: For machines accessing a digital object via a PID, either as a direct reference or as the result of a search, it is crucial to find the type information to understand how to interprete it. Type information can be found in the metadata and/or in a type registry. While the metadata refers to a registered schema and the schema to registered concepts, the type registry specifies the kind of action to be taken for a specific concept in its type context.

DAT17. RDA DTR: Every type included in a digital object and registered in a data type registry must be identified with a resolvable persistent identifier.

DAT18. DOI: A DOI needs to be registered for a published DO and it should be associated with citation metadata.

DAT19. DSA1: The data producer deposits the data in a data repository with sufficient information for others to assess the quality of the data and compliance with disciplinary and ethical norms.

DAT20. DSA2: The data producer provides the data in formats recommended by the data repository.

DAT21. DSA3: The data producer provides the data together with the metadata requested by the data repository.

Data Access Topic Area

ACC1. RDA: Trustworthy repositories need to support a standard API that includes path information, PID, user credentials, checksum and other important information)

ACC2. RDA: Secure and standardized mechanisms (SAML) are supported to interact with an authentication system checking user's identity.

ACC3. RDA: Secure and standardized mechanisms (??) are supported to interact with an authorisation system checking access permissions.

ACC4. 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)

Data Processing Topic Area

PRC1. RDA-PP: A trustworthy repository must specify auditable practical policies for its various tasks, turn them into executable procedures and workflows, and systematically apply them in all cases to document provenance of all its digital objects.

PRC2. RDA: Workflows and procedures that create new digital objects need to include software components that read the existing PID record and metadata and that associate a new PID, create new metadata incl. provenance both to be associated with the new DO and upload it into a trustworthy repository.

PRC3. RDA: Annotations need to be created in stand-off manner and where suitable Open Annotation Format should be applied.