GEDE

Group of European Data Experts

Digital Object Topic Group Report

Digital Object Roadmap Document

V3.0, Status of August, 2019

Editor:

Peter Wittenburg

Contributors:

We like to thank the many contributor from GEDE, in particular Dieter van Uytvanck, Koenraad de Smedt, Dimitris Koureas, Tobias Weigel, Per Öster, Ari Asmi, Werner Kutsch, Colin Mc Murtrie, Thomas Schulthess, Daniel Mallmann and Thomas Lippert, for their many contributions and comments.

It should be noted that this document does not yet include roadmap aspects that include FAIR Digital Objects, since discussions about the future devlopments are still ongoing. We leave it to the FDO initiative to extend this document.

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

Document Revision History

2018/10/13	V 1.0 Initiating the document
2019/01/14	V 1.0 Intensive commenting on Drive by GEDE - DO members
2019/February	V 2.0 Summarising the discussion by editors
2019/March	V.2.0 3 months intensive commenting by GEDE – DO members
2019/ July	V3.0 Summarising the discussion by co -chairs
2019/08	V3.0 Current state of the document

Contents

Abstract

This document indicates the areas of participation in and contribution to establishing a stable domain of digital entities in the sciences and a DO based eco system of infrastructures. This document includes three sections:

1. A summary of actions with prioritisation
2. A list of actions that are required to build the basic DO-based interoperable infrastructure
3. A list of actions that are of relevance from a scientific point of view, knowing that this list can be infinite.

This paper is a working document and will be extended stepwise. It will be discussed within the GEDE DO collaboration.

About GEDE

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

This document indicates the areas of participation in and contribution to establishing a stable domain of digital entities in the sciences and a DO based eco system of infrastructures. This document includes three sections:

4. A summary of actions with prioritisation
5. A list of actions that are required to build the basic DO-based interoperable infrastructure
6. A list of actions that are of relevance from a scientific point of view, knowing that this list can be infinite.

This paper is a working document and will be extended stepwise. It will be discussed within the GEDE DO collaboration.

1. Action Summary and Prioritisation

The added value of implementing Digital Objects

Out of the past efforts of various RDA working groups it has become clear that several stakeholders from several communities see general value in the concept of Digital Objects, where an object is a conceptual entity consisting of a persistent identifier, associated metadata and a bitstream. This is the most general model that can be applied and understood across many application scenarios and can be supported by infrastructure services.

The value of the Digital Object concept, however, resides not in establishing Digital Object services, tools or even whole infrastructures for the sake of establishing just the concept. We understand that the reason for pursuing efforts concerning Digital Object applications and community uptake in the general domain of research data and research services must ultimately be driven by the needs of three main stakeholder groups: Researchers, their supporting infrastructures, and funders.

For researchers, implementing a coherent, reliable Digital Object concept can build trust in the objects that researchers deal with in their daily work: If the relevance and fitness-for-purpose of objects is clearly exposed, if object provenance and completeness are clearly visible, then the trust of one researcher in another researcher’s generated objects can be improved. Realizing this could remove a large barrier for data sharing and re-use, and building trust in objects is the key to this.

For research infrastructures, the concept of Digital Objects offers the potential for more efficient, scalable data management processes that span the boundaries of institutions, communities of practice and formal or legal barriers. Data management at the technical infrastructure level today still relies on creating, moving and copying files within and across a great diversity of technical storage systems, which offers a high potential for faults, errors and poorly managed processes.

For funders, the concept of Digital Objects opens up the chance to trace the impact of research in a new way. If objects are reliably identified, associated with metadata and put into a coherent object graph, the impact of data generating research projects can be assessed. This can also stretch to the tools and services that are used on Digital Objects, and thus increase transparency and facilitate insight into the research process and return on investment.

It is clear that an implementation across disciplines and technical infrastructures is the next immediate thing to do to realize the potential that the Digital Object concept promises.

However, an implementation should first and foremost be measured by the added value it provides for the aforementioned stakeholder categories. The following concrete actions can facilitate this:

1. **Improve FAIR metadata servicing** across the data lifecycle and across disciplines. Research infrastructures already have metadata repositories, and therefore, the existing services, schemas and interfaces need to be adapted. Concerning provenance in particular, many communities already use W3C PROV to expose metadata but are struggling with the conceptual overhead of ontology engineering. A combined approach with the RDA PID Kernel Information[1] concept can help.

2. **Establish a common approach to DO management** at the research infrastructures level and within their data services. DO-based systems will co-exist with current systems and services, but a DO-based approach will be encouraged leading to stepwise replacements and transitions. trust in the flexibility of DO-based solutions must be secured first.

3. A common Digital Object Interface Protocol needs to be evaluated and co-developed with multiple communities, based on the existing DOIP effort. The key success factors include how well such a protocol can be combined with a modular, component-based software architecture that many infrastructures already use or that they would benefit from in terms of long-term maintenance.

4. **Improve and automate data object management processes** within the research infrastructures, facilitating better scalability and automating significant parts of the still manual processes. A working software component ecosystem to manage objects needs to be implemented and evaluated in multiple research infrastructures. This includes components to manage Research Data Collections, facilitating common create, read, update and delete actions on multiple objects, and components to type Digital Objects, register types and associated actions and thus facilitate automated service discovery and execution.

2. Contributions to a DO based Eco System of Infrastructures

This list includes all actions that are required to establish the basic interoperable infrastructure eco-system as it is for example indicated by the EOSC plans in Europe.
However, we assume that the neutral Swiss DONA foundation will

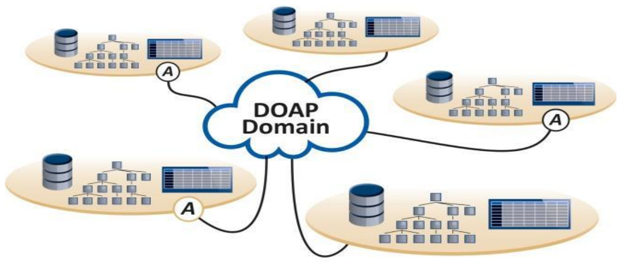
- continue to provide the first class and global Handle Resolution System making it possible for everyone to register and resolve Handles in a stable way for decades
- provide and maintain the specifications of the DO Interface protocol and also provide prototypical implementations that will also include canonical basic operations on DOs
- provide software that describes the usage of DOIP in a canonical way.

2.1 Interoperability by Adapters

Action: Make sure your repository is ready to participate in the DOIP domain by adaptation.
Sub-actions:

- **Make sure that your digital items are referenced by X.1255 compliant PID such as the Handle Service.**
- **Make sure that you can provide all needed information to satisfy seamless machine processing.**
- **Reuse for your PID record attributes registered types and where necessary start a process to define and register new types.**

[1] https://www.rd-alliance.org/groups/pid-kernel-information-wg

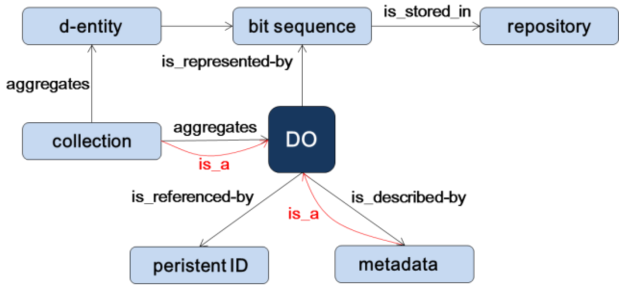


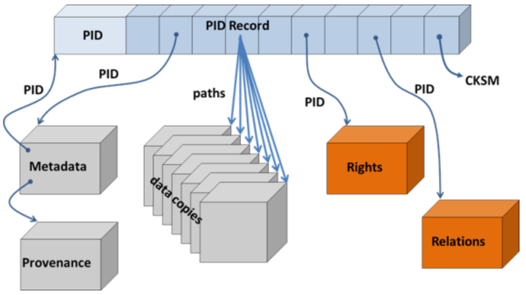
The DO Interface Protocol[2] defined an interoperable domain based on the definition of a DO, i.e. those that do not have a DO-based setup of their repository or data service already in place need to develop an adapter that can be complex dependent on the way the digital entities are organised[3]. Data, Metadata, relations, provenance etc. can all be stored and linked in different ways[4]. DOIP, however, will clearly specify how the different information entities need to be provided. Basis will be that your repository is working with ITU X.1255 compliant PID systems such as the Handle System, since otherwise DOIP cannot work. To give an example of a typical adapter problem that may occur when metadata is being stored in database tables: You would have to execute a query to extract the relevant data and transform it in the required way.

DOIP has a very simple structure based on the ITU X.1255 standard:

- *EntityID: the identifier of the digital entity requesting invocation of the operation;*
- *TargetEntityID: the identifier of the digital entity to be operated upon;*
- *OperationID: the identifier that specifies the operation to be performed;*
- *Input: a sequence of bits containing the input to the operation, including any parameters, content or other information; and*
- *Output: a sequence of bits containing the output of the operation, including any content or other information.*

All elements in the DOIP are referenced by a global resolution system such as the Handle System. In addition, from a scientific perspective the RDA DFT definitions[5] are important, since they partly define which information needs to be made available by using the protocol. Two diagrams are of relevance. The first clearly states that the DO has a bit sequence stored in some locations, has a PID and has metadata which can be of different types. The second image specified the binding role of the PID record which allows even stupid machines to find the (descriptive) metadata, the locations, other types of metadata such as rights, attributes such as checksums etc.





The *TargetEntityID* in the DOIP packet is exactly the PID of the DO we are talking about. The repository needs to make clear that it uses typed PID attributes[6] so that the machines can find these

[2] Sometimes DOIP is also called the DO Access Protocol. We see them as synonyms here.

[3] This is exactly what happened in early days of Internet when for example IBM which had its own protocol stack for networking adapted to the TCP/IP domain to participate and later stepwise adapted their internal structure to natively support TCP/IP. When a repository stores all metadata and data in one relations database, DOIP based requests need to be translated to more or less complex queries for example where the queries need to be assigned a PID and the content must be time-stamped to get the intended bit sequences.

[4] Repositories use different containers such as file systems, cloud stores, SQL/NoSQL databases, etc., use different representation models and often lack exolicitness..

[5] http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318

[6] It should be noted here that data intensive science will often require flexibility with respect to PID attributes which is not given by data publishing communities such as issuing DOIs (which are also Handles).

elements that are needed for interpretation, lending access etc. In general, those attribute types need to be used that already have been defined by the RDA Kernel group and openly registered in a Data Type Registry (such as the one from GWDG - ePIC[7]). In case that new attribute types are required this can be notified to the GEDE secretariat which would take care that it will be commonly defined and registered in collaboration with the requester and the community in charge.

2.2 Interoperability by Repository Organisation

Action: Make sure your repository is organised according to the DFT model and uses relevant attributes in the PID record. In this case it is straightforward to write an adapter that can interact with the protocol.

In this case the adaptors are much more straightforward since all information elements can directly be provided.

With respect to the PID record attributes the same remarks as made in 1.1 hold.

2.3 Specialisation of Standard DOIP Operators

Action: Develop microcode that implements the standard DO operators for the specific repositories.

Sub-actions:

- **It should be checked in how far microcode types can be generated for repository types to simplify the task.**

DOIP will incorporate a number of standard operators for Digital Objects such as create, delete, change, move, etc. These will be defined in a canonical way. Since all repositories will need to transform these operations in different ways dependent on their setup, microcode needs to be developed that will transform canonical operators into real ones performing the task given the special architecture of the repository.

GEDE should check in how far quasi generic microcode can be developed for specific types of solutions such as relational databases etc. so that not all parties need to start developing adapters. It seems to be very important to provide

2.4 Forming DOIP Service Provider

Action: Not all repositories need to be made ready to implement the DOIP server software. Repositories could form associations to share the work.

Sub-actions:

- **A survey needs to be carried out to understand how clustering of repositories could be done.**

Models need to be worked out and tested how repositories could share DOIP service providing nodes (proxies) that would act on their behalf in the DOIP landscape. Still microcode would have to be developed to expand the interaction to the different repositories. In Europe the ESFRI projects and ERICs would have an important role to organise such models.

2.5 Forming a PID Kernel Information Type Community

Action: A formal community needs to be defined that can take care of the standardisation, definition and registration of Kernel Types. Procedures need to be defined that guide this process.

The standardisation, definition and registration of PID record attributes (Kernel Info Types) requires the setup of a community that takes decisions according to processes that need to be defined. There has been such a discussion in the RDA PID Kernel Group, but it is not yet sufficient. International collaboration needs to be built into the process.

Also here the ESFRI projects and ERICs would have an important role to start organise this interaction. But it seems to be obvious that the RDA interaction process cannot replace a decision structure.

[7] https://www.pidconsortium.eu/

2.6 Integration into Metadata Registries

Action: Many communities have setup their metadata aggregation, indexing and servicing systems such as the CLARIN Virtual Language Observatory. They need perhaps to do updates so that they can play an active role in the DO based domain.

Most scientific communities have organised their metadata domain insofar as schemas and vocabularies have been defined and indexing and portal structures have been worked out[8]. It may be necessary that some specifications and mechanisms need to be adapted, although we assume the adaptations to be small.

2.7 Integration into Repository Registries

Action: It seems to be useful to have a registry of repositories and perhaps registries that indicate DO compliance. It needs to be discussed with existing registries (re3data) in how far they can be of use for the DO-based domain.

It is important that a registry exists that covers all repositories that participate in the DO domain and that includes all relevant information (metadata) about the repository. Most probable is the need to define a new type of registry, but interactions with initiatives such as re3data should happen to look for ways of collaboration.

2.8 Certification of Repositories and Registries

Action: There are already certification schemes for repositories such as CoreTrustSeal, but there is nothing yet for registries. We need to find out whether current schemes can be extended if necessary or whether new schemes are required.

We have a few suggestions for certifying repositories where CoreTrustSeal seems to be the most widely used[9] and the most pragmatic one. Nevertheless, CTS needs to be further developed. In particular when we apply the DO concept systematically we can add useful rules to the existing CTS set. This needs to be worked out in collaboration with the CTS experts.

In addition, there are no widely accepted certification schemes for registries yet although the proper functioning of registries is a key requirement. Work needs to be invested to identify crucial types of registries and to develop rule sets.

2.9 DOIP Testbed

Action: The ultimate goal of course is to develop an evolving testbed for a DO-based infrastructure at an early point in time to test feasibility and to require improvements. This needs to be done in close international collaboration making use of C2CAMP etc. eRPID (US) can be seen as such a tesbtbed.

The testbed exists of a number of participating repositories and the required registries that have been described above. As usual the integration always needs additional work that needs to be planned.

2.10 DOIP Integration into EOSC

Action: We need, over time, to confirm which components of the testbed are solid and robust enough to be integrated into the EOSC plans in order to form a common infrastructure.

Given that the testbed results are excellent we need to make it official as part of EOSC. Technically that requires integrating even more candidates, extending the required registries and giving a lot of support and training.

2.11 Formalising the GEDE DO Community

Action: Many of these actions require the existence of a formalised European GEDE DO community. Currently, we have the GEDE interaction, which exists on a voluntary basis, but this

[8] This does not mean that still much has to be done at all aspects.

[9] In some disciplines Core Trust Seal is not that well-known as in many others.

will not be sufficient when implementation work based on firm funding has to be carried out. We need to understand how this can go together with structures in EOSC.
We need a formal structure to take decisions on various matters (registries, certification, kernel attributes, etc.), to undertake reviews (for example of the testbed state), to have a continuously acting body to push RDA results, to receive funds and assign them to people, to check delivery etc.

3. Scientific Case Driven Activities

This is a wide field of activities driven by the needs of the scientific communities and making use of the basic infrastructure. At the end, it will be the scientific cases that will drive the basic infrastructure to work robustly, to add urgently needed features and to understand the potential of DO based infrastructures. As a start we take those cases described already by domain scientists.

3.1 Methods to Organise a Stable Domain of Digital Entities

Action: Establish and maintain an interaction platform bringing together data scientists and IT experts to push the social and technical dimensions of the DO-related work.
We need to maintain an interaction platform at expert level that includes domain scientists and IT experts driving the infrastructure work to discuss the scientific needs, the state of the infrastructure work, priorities etc. Such an interaction platform will address issues of relevance, new ideas for scientific applications and build task forces where necessary.

Some topics which are not per se of technical but sociological nature have been mentioned that should be taken up:

- There is a paradox: On the one hand we are creating huge amounts of data, but on the other hand we are only exploiting them scientifically to a very small extent. Data driven solutions are urgently required, but we need to understand to deliver data at scale, form and precision to make that happen. What is missing to overcome the current barriers? What needs to be done? Can DO-based approaches help to overcome the barriers?
- With a DO based eco-system of infrastructures in place there should be new opportunities to integrate data from various sources. We need to identify how this eco system can give new impulses to significantly contribute to the grand dimensions of sustainable development as identified by the UN assembly? These identify global challenges which are carried out locally by researchers who are acting in general in their small communities of practice. How can we make use of the new eco-system to enlarge the community of practices, to better document the context in which data were produced and in doing so to improve trust in them and the services built on them?

3.2 Design and Implementation of a Very Large DO Domain

Action: Very large systems come with extreme requirements which need to be understood and tackled at an early stage. Careful designs and stepwise testing is crucial.
Some scientific domains such as biodiversity are characterised by a huge number of digital representations of physical objects. The only possible way to build a stable domain of digital data and information in such complex domains seems to be to define a bottom layer of these digital representations of physical observations as a layer of basic DOs which are clearly identifiable at global level and to allow building complex virtual structures (different classifications, geographical and other relations, etc.) on top of them. Core of such infrastructures is that the huge amount of references is independent of any technological change and stable for centuries. It is important to define a testcase with such extreme requirements, to design and specify the major components and to a stepwise implementation.

3.3 Methods for a DO supported Switchboard

Action: All participating instances need to implement the DOIP protocol, give access to various DO-related information via the Handle record and provide sufficiently rich metadata. The switchboard orchestration facility needs to be developed on top of DOIP.

Matching metadata and type profiles between tools and data is an important step towards automatic workflow orchestration helping non-professionals to carry out complex work. A switchboard solution where all DOs are identified by Handles and where Handles are being resolved to crucial metadata, type and location information is a promising way to test such workflow solutions, i.e. the binding function of the DO concept is crucial to let procedures seamlessly find all relevant information. In the area of language resources for example minimal requirements are access not only to the PID of the bit sequence of the text, but also to the detailed metadata, to its type, to a language identifier according to ISO 639-3 and a usage license specification, all in machine-readable form. The switchboard is a lightweight construction that should be easily installable in different contexts and be easily to be invoked from these contexts. If for example a certain text is being uploaded to a cloud service, the intention is to be able to use a mouse click to invoke syntactical processing of that text.

3.4 Virtual Collection Builder

Action: Existing virtual collection builders need to be analysed. A selected number of them need to be extended to make them DO and DOIP compliant. The RDA collection interface need to be looked at and perhaps needs to be adapted as well.

Building virtual collections combining different types of resources from different repositories is one of the major applications data intensive scientists will carry out daily across all disciplines. Virtual collections are aggregations of metadata of different DOs which are identified by a Handle where the Handle record binds important information about the DO. The resulting collections are themselves DOs, i.e. have a Handle and the record referring to metadata about the collection, the place where it is stored, etc. The type is most probably “collection” since it can include a wide variety of different types bundled together. Some scientific communities already developed collection builders[10]. Few of them with excellent functionality need to be extended to make them DO and DOIP compliant and be provided to the interesting cross-disciplinary community. The RDA collection interface[11] is of great relevance and should be considered while developing the VCB tools. If necessary, the collections interface needs to be expanded.

3.5 Type-Triggered Automatic Processing Domain

Action: The adoption of workflow-based methods will be crucial for success in many different communities where large amounts of data need to be processed. Projects need to be designed to stepwise implement workflow scenarios based on the DO approach.

In many scientific disciplines there is an urgent need to move towards automatic data processing with the help of smart workflows such as in climate modelling where the new phase of comparison studies will include up to 3 EB of data. The data volumes and complexities in such sciences pose new technical and organizational challenges and demand solutions such as automated digital object management, increasing workflow support and provenance aggregation and supporting to work at higher levels of abstraction. Key for the success in such scientific disciplines with advanced computational needs is the establishment of a domain of digital objects with clear and stable identities and with DOs binding all relevant information in a way that machines can find it. DOs have the best potential to become true "primary citizen" within such communities and revolutionize the data management/stewardship. A DO-based infrastructure opens the way to even allow laymen and collaborators to employ server-side virtual research environments such as Jupyter to create workflows for example to carry out machine learning calculations.

[10] Examples can be found here: http://www.researchobject.org/; https://www.clarin.eu/content/virtual-collections
[11] https://www.rd-alliance.org/groups/research-data-collections-wg.html

We urgently need to explore the potential of DO-based infrastructures for such demanding scenarios, do design studies and carry out implementations.

3.6 Specific Challenges Emerging from Science

Action: Specific cases from scientific practice need to be addressed at an early stage to explore the potential of the DO approach.

For example, search results emerging from distributed databases can be valuable Digital Objects in their own right. However, they can be represented in substantially different ways such as a quantum-state description in atomic and molecular physics where the same physical state may be represented as based on quantum numbers, as energy from a fundamental state or as ionisation energy. It would require a system design change to indicate that they are different appearances of the same physical entity, i.e. a PID needs to identify the common source allowing users to quickly switch between appearances which would otherwise be impossible.

There are many more such specific challenges emerging from concrete scientific projects. A number of such cases should be defined and tackled at an early stage to demonstrate the usefulness of the DO approach, but to also discuss its limitations.