GEDE

Group of European Data Experts

Digital Object Topic Group Report

Some Terminology Issues

V3.0, Status of August, 2019

Editor:
Peter Wittenburg

Contributors:
We thank for the many contributions from members of the large the GEDE- DO Topic Group.

It should be noted that this document version does not yet include the concept of FAIR Digital Objects which was coined in 2019 but still needs further work. An extension to include FDOs will be left to further work in 2020.

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

2018/10/05	V 1.0 Initiating the document
2019/02/13	V 1.0 Intensive commenting on Drive by GEDE - DO members
2019/March	V 2.0 Summarising the discussion by editors
2019/03/14	V.2.0 Intensive commenting by GEDE- DO members
2019/06	V2.0 Summarising the discussion by co -chairs
2019/08	V3.0 Current state of the document

Abstract

At the last DO Meeting we identified terminology issues and I promised to give it a start to clarify some terms (from my point of view). This document is meant as a start and I will use definitions that were made in RDA DFT based on many use cases as a pivot point and add others for example from DONA, ITU X.1255, etc. There are many other definitions out there and GEDE should feel free to add notes on definitions or add additional terms.

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.

Contents

At the last DO Meeting we identified terminology issues and I promised to give it a start to clarify some terms (from my point of view). This document is meant as a start and I will use definitions that were made in RDA DFT based on many use cases as a pivot point and add others for example from DONA, ITU X.1255, etc. There are many other definitions out there and GEDE should feel free to add notes on definitions or add additional terms.

1. Digital Object (DO)

RDA DFT Definition

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

Note*: There is now wide-agreement to call this the FAIR-DO since it includes the mechanisms to make DOs FAIR.*

Note*: In the meantime we speak about different types of metadata such as descriptive, provenance, context, system, access rights, license, relationships of various sorts etc. metadata, all describing different properties of a DO.*

Note: *Characteristically for a DO is that it has "constitutional internal" relations. But of course a DO can also have many different external relations to other DOs which is apparent for example in the case of collections which can consist of many other DOs.*

Definition DONA DOIP (compliant with ITU X.1255)

A digital object is a sequence of bits, or a set of sequences of bits, incorporating a work or portion of work of other information in which a party has rights or interests, or in which there is value, each of the sequences being structured in a way that is interpretable by one or more of the computational facilities. Each DO has, as an essential element, an associated unique persistent identifier, known as digital object identifier[1] (referred to informally as a Handle).

Note*: The two definitions are widely overlapping. The DONA definition is more elaborate and does not mention "metadata" while the DFT definition stresses the relevance of metadata for interpretations etc. Both will rely on the availability of crucial information such as metadata types via the PID record.*

Note*: Digital Objects are a generic term that could include different types of information such as data, software, configurations, semantic assertions, etc. Some speak about* ***Digital Data Objects*** *as a specification of the generic DO term when the content contains data[2].*

Note*: The interim FAIR Implementation Report[3] speaks about FAIR Data Objects which indicates the close relationship between the FAIR principles and the DO concept which will be explained in a separate paper to come soon.*

2. DO Interface Protocol[4]

DONA DOIP Definition

The DOIP specifies a standard way for clients to interact with DOs.

Note*: The DOIP is a unifying protocol that allows to access, create, delete, etc. DOs independent of the way repositories and registries organise and model their data and other types of information[5]. In doing so, it will have a similar role for the domain of DOs as TCP/IP had for interconnecting the many island networks decades ago. This comparison does not want to give the impression that the challenges of networking computers is of the same complexity than the networking of data.*

[1] This may not be mixed up with DOIs which are special types of Handles issued by a Handle user community characterised by a defined business model.

[2] It is obvious that what even software could be seen as data dependent on the type of operations intended.

[3] https://doi.org/10.2777/1524

[4] Sometimes this is also called "DO access protocol".

[5] Some are using file systems, cloud objects, database tables, etc. and also the way they relate different entities such as bit-sequences, different types of metadata, PIDs, etc. is very different.

3. Persistent Identifier (PID)

RDA DFT Definition

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

***Note**: We see the terms "Persistent Resolvable Identifier" and "Unique Persistent Identifiers" as synonyms, i.e. PIDs are meant to be globally unique of course.*

4. PID Record

RDA DFT Definition

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

***Note:** The PID record can be seen as a DO (however, it does not have metadata)*

***Note**: Pointers (PIDs) to different types of metadata should be included in the PID record.*

***Note**: The RDA Kernel group started specifying such types of attributes; it will remain a continuous effort.*

5. PID Resolver (aka Resolution System)

RDA DFT Definition

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.

***Note**: With the Handle System as managed now by the international DONA Foundation located in Geneva is such a global resolution system being used by the DOI Foundation, the ePIC consortium, and more than 3000 Handle service providers world-wide.*

6. Metadata

RDA DFT Definition

Metadata contains descriptive, contextual and provenance assertions about the properties of a DO.

***Note**: Such metadata will make the DO for example discoverable, accessible and usable/interpretable[6].*

***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 (it describes).*

***GOFAIR Assertions on Metadata** (a paper will be made available soon)*

- *Metadata statements are assertions about Digital Objects (which can be anything digitally represented) and are part of the DO to make it findable, accessible, interpretable and re-usable.*
- *Metadata assertions are made by different actors with different roles for different purposes at different times. Some metadata will be created already by the sensor equipment, other metadata will be added by the responsible researcher, again others will be added by the repository manager etc.*

7. Digital Collection

RDA DFT Definition

A digital collection is an aggregation which contains Digital Objects and Digital Entities[7]. The collection is identified by a PID and described by metadata.

Note: A digital collection is a (complex) DO and has the structure of a hierarchy. In general, several hierarchies can be built on top of the same set of basic DOs.

8. Repository

RDA DFT Definition

[6] Currently, one would refer to FAIR which was not a label when the DFT definitions were finsihed.

[7] "Digital Entities (DE)" in this context are referring to elements that do not exist as self-standing DOs. Researchers, for example, may bundle thousands of photos to one DO without taking the effort to associate PIDs and metadata with all individual photos.

A digital repository is an infrastructure component that is able to store, manage and curate DOs and return their bit-streams when a request is being issued.

Note*: The Curation of a DO can imply several different aspects. The bit sequence encoding DO's content could be changed which is often the case for mutable DOs. Or the metadata description of a DO could be changed. The repository need to define their policies whether such changes automatically lead to new DOs with a new PID being associated or whether the PID will remain the same. Users need to know what the policies are.*

Note: In general the task will be to return the DO's bitstream since some operations will be carried out. But there will also be cases where all essential entities should be delivered which would require some selection and container mechnisms.

DONA DOIP Definition

The repository system is a DO service that provides the necessary functionality to manage DOs including the provision of access to such DOs based on the use of identifiers.

Note*: PID records need to be protected against unauthorised usage, i.e. the PID resolution system has security mechanisms built in. Additional security mechanisms ensuring control of data access and transactions need to be implemented at the application layer to not overload the protocol.*

Note*: In the current discussion we separate three views on repositories:*

- *In the Computer Science view a repository is a complex collection and therefore a Digital Object, i.e. it has a PID and metadata and supports the DOIP - a functional interface.*
- *In the above mentioned view a repository is also an institution with "people" taking care that all relevant functions which range from storing DO bit-sequences up to doing curation are being taken care of. This definition does not exclude that the different functions are mapped on different organisational structures.*
- *Some see a repository simply as a big store taking just care of maintaining the bit sequences. This is just one basic function a "full repository" takes care of. It refers to one specific type of possible organisational structures.*

9. Registry

DONA DOIP Definition

The registry system is a specialised repository system intended to store metadata about DOs rather than the digital information itself, and when used as a standalone component, typically stores metadata of DOs that are managed by one or more repository systems.

Note: This definition states that there is only a functional difference between a repository and a registry. Typically, registries aggregate different types of metadata about DOs (which can be anything digital) and offer specialised services on the aggregated data. Metadata portals supporting searches are typical registries, for example. They store metadata descriptions, however, these metadata descriptions could be enhanced or specialised versions and in general such registries do not have to offer persistence for their stored information, i.e. it is obvious that such registries register their metadata versions as self-standing DOs. However, there are cases of registries such as a PID registry that need to have persistent storage functionality.

10. State Information

RDA DFT Definition

State information is “metadata” information that describes those properties of the DO that are relevant for proper management and access.

Note*: There is yet no "exact" definition of what the scope of "state information" is. It was introduced by the RDA Practical Policy group*[8] *and in lack of another term used by others as well. It will depend on what kind of information about DOs communities want to associate directly with PIDs in the PID*

[8] https://rd-alliance.org/groups/practical-policy-wg.html

record. It is generally agreed that the PID record has the binding role that is so important for working with DOs, i.e. link to other information such as different metadata types, to different versions etc.
Note*: A well-known analogy is the identity of persons specified by a passport for example. Passports have some metadata associated with the ID, but there may be differences between countries what kind of metadata is being associated with the ID in the passport.*

11. Type of DO

DONA DOIP "Definition"
The "Type" informs DOIP services what operations can be invoked against the DO.
Note*: In the recent discussions of the RDA Data Fabric*[9] *group, which led amongst others to the paper from Wittenburg, Strawn, Mons et al.*[10]*, it was obvious that DOs are the gate towards increased automation in the data domain. Crucial metadata assertions can be summarised to a "Type" that can be linked to a set of "operations" one can carry out on DOs, i.e. we can see them as an extended MIME Type concept for the more complex scientific DO domain. This "typing" supports the encapsulation of the "object" concept as we ca find it in "Abstract Data Types" and "Object Oriented Programming".*

12. Data Type Registry

RDA DTR Goal Specification[11]
The overall goal of the group was to define a data model for data types, to prototype its use in a functioning registry, to experiment with some domain-specific data type sets, and to define a federation strategy across multiple registries. The assumed benefit of this, on which there was fairly general agreement in Plenary Breakout sessions, which is where most of the discussions took place, was that precise typing of data sets and collections, combined with one or more registries that define those types in a standard fashion, would benefit every sector of data management, especially interoperability and reuse. The WG further agreed that it would not attempt to define the methods of association of data and type but would work towards a standard approach for registering and discovering types as well adding value to their use by linking types to services.
Note*: DTRs are registries that allow users to link "types" with "operations" in a standard format, i.e. the researcher specifies operations for classes of DO (types) dependent on the aim and context of processing.*

13. Research Objects

"*www.researchobject.org*" Statements
Research objects are not just data, not just collections, but any digital resource that aims to go beyond the PDF for scholarly publishing!
The reuse and reproduction of scientific experiments as they are described in publications can be hard. Often it requires additional information, data, tooling or support beyond that provided in the text of a traditional publication.

Note: "Research Objects" are being defined by the above mentioned initiative and address the issue of how to package complex information in a way that allows the reproduction of scientific results. ROs could also be seen as a specification of how to pass over complex information that is necessary to orchestrate a workflow.

[9] https://www.rd-alliance.org/group/data-fabric-ig.html
[10] Digital Objects as Drivers towards Convergence in Data Infrastructures; http://doi.org/10.23728/b2share.b605d85809ca45679b110719b6c6cb11
[11] https://www.rd-alliance.org/group/data-type-registries-wg/outcomes/data-type-registries