Group of European Data Experts

Digital Object Topic Group Report

Digital Object Assertions

V3.0, Status of August, 2019

Editors:

Peter Wittenburg, Dimitris Koureas, Koenraad de Smedt

It should be noted that this document does not yet include assertions about FAIR Digital Objects, since discussions about the specifications in detail are still ongoing. We leave it to the FDO initiative to extend this document.

RESEARCH DATA ALLIANCE

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

Document Revision History

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 3 months intensive commenting by GEDE members
2019/06	V2.0 Summarising the discussion by co -chairs
2019/08	V3.0 Current state of the document

Abstract

It is good practice in GEDE to start at an early moment with first assertions to understand where we have agreements and where further discussions are needed. There might also be more assertions over time. This version is a consolidated second version based on the commented online version[1]. It should have responded to all comments that have been made. In addition to definitional, operational and scientific assertions, it includes also a side discussion raised by a comment. This version 3.0 is now open to comments again.

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.

[1] https://docs.google.com/document/d/1Yb291ITgrqfEabmwGa4UYWi_9t47-xVRSGbYLqeY0WA/edit

Contents

It is good practice in GEDE to start at an early moment with first assertions to understand where we have agreements and where further discussions are needed. There might also be more assertions over time. This version is a consolidated second version based on the commented online version[2]. It should have responded to all comments that have been made. In addition to definitional, operational and scientific assertions, it includes also a side discussion raised by a comment. This version 3.0 is now open to comments again.

Definitional Assertions

D1: Definition of DO

The discussions about Digital Objects since the start of RDA in 2013 resulted basically in two major definitions from RDA DFT and DOIP and several others that are based on different discussion roots such as from ITU, SAA or the one suggested by Hermon Sorin.

RDA DFT[3]: A digital object (DO) is represented by a bitstream, is referenced and identified by a persistent identifier and has properties that are described by metadata.

DOIP[4]: A digital object (DO) is a sequence of bits, or a set of sequences of bits, incorporating a work or portion of a work or 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, and having as an essential element an associated unique persistent identifier.

ITU[5]**:** A digital entity is an entity represented as, or converted to, a machine-independent data structure consisting of one or more elements in digital form that can be parsed by different information systems; the structure facilitates interoperability among diverse information systems in the Internet.

Society of American Archivists (SAA)[6]: A digital object (DO) is a unit of information that includes properties (attributes or characteristics of the object) and may also include methods (means of performing operations on the object).

Hermon Sorin: DO is an abstract element that gathers information about a physical or a virtual entity, sufficient for the discovery and inquiry on this entity.

***Note**: In a recent paper Wittenburg & Strawn*[7] *[1] capture the development of the term "DO". They see the necessity of having a minimal definition for the purposes of defining a simple DO Interface Protocol (DOIP) that is identifying the resolution of a PID as a separate step. On the other hand they suggest to adopt the term "FAIR-DO" for the definition created by RDA DFT which is driven by scientific considerations, i.e. it opens the door for DOs implementing FAIR by abstracting away from the concrete protocol steps and by stressing the abstraction, binding and encapsulation aspects. The ITU insisted on using the term "entity" since due to their history the term "object" is already being used for different aspects. ITU's definition is fairly abstract, but not in contradiction with the FAIR-DO one, if you replace their term "entity" with the term "object". The SAA definition is overlapping with the FAIR-DO definition, including also the capacity of encapsulation. The RDA DFT definition in addition indicates that by making use of PIDs one can implement these properties. Hermon's definition does not cover all relevant aspects of what we mean by a "DO".*

***Note**: All definitions state in some form that a DO is an abstract concept which represents knowledge about a virtual entity (the work) be it a born-digital entity or a physical entity being represented by some digital entity. Attributes associated with the DO will include DO properties and references needed to make DOs FAIR as spelled out in the FAIR-DO definition.*

[2] https://docs.google.com/document/d/1Yb291ITgrqfEabmwGa4UYWi_9t47-xVRSGbYLqeY0WA/edit

[3] DFT Core Terms and Model; http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318

[4] DONA: DOIP specification; https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf

[5] ITU: Recommendation X.1255 (09/13); https://www.itu.int/rec/T-REC-X.1255-201309-I

[6] Society of American Archivists: Digital Object; https://www2.archivists.org/glossary/terms/d/digital-object

[7] It is quite natural that terms that have their origin in 1995 [2] undergo changes dependent on the increasing insights in the needs.

***Note**: DOs can be compositions at two levels: 1) the "work" itself can be complex (can include various bit sequences structured by different schemas), 2) a DO can be a complex collection of DOs.*

***Note**: A DO is different from a File or a HTTP-Resource. A file is just one way to organise your digital entities, you can organise your entities for example also in a relational database. The term "DO" is abstracting away from any concrete way of organisation. The same is true with respect to resources to be exchanged based on HTTP. HTTP is a specific protocol coupled also with HTML resources. A DO exists independent of the protocol and any format specifications and is therefore defined at a more abstract level.*

D2: Scope of DOs

A DO is an abstract representation of a work or entity in digital form. DOs can contain data, metadata, software, configurations, concepts, digital representations of physical entities, etc.

D3: PID Record

The PID record is a DO that contains attributes which bind all entities together that are needed to make the DOs FAIR, i.e. locations of the bit-sequences, the type, references to metadata types (descriptive, system, access rights, licences, transactions, etc.), checksum and other important properties.

***Note**: To enable machine processing all these information types need to be registered and defined in an open data type registry. This is the concept of PID Kernel Information.*

***Note**: The abstract entity that includes the binding of information through the PID record is now called FAIR-DO.*

D4: PID Kernel Information

PID Kernel Information[8] is information in the form of attributes stored within the PID record, i.e., information stored at a global or local PID registry and accessible by a resolver. PID Kernel Information supports smart programmatic decisions that can be accomplished through inspection of the PID record alone. PID Kernel Information profiles are registered schemas for PID records. PID records may be created according to specific profiles and checked for conformance against them. In other words, PID records are concrete instantiations of profiles, comparable to how we consider objects as instantiations of classes in object-oriented programming.

D5: Digital Object Collections

Digital objects can be bundled together to form collections. A collection is a digital object that consists of multiple other digital objects, referenced by their PIDs, and affords a set of individual capabilities. Capabilities would include operations such as adding, removing or iterating over collection objects, but may also cover properties of a collection such as whether its elements have an implicit ordering or whether there are roles associated with elements in the context of a specific collection. The RDA WG on Research Data Collections[9] has defined a conceptual model for collections and an API to describe the interaction with collection management services.

D6: Relation between FAIR principles and DOs

A FAIR-DO is implementing some FAIR principles directly in machine actionable ways, and facilitating others indirectly.

***Note**: Two recent papers from Schultes & Wittenburg [2] and Strawn [3] give a detailed impression about the relationship.*

***Note**: A FAIR-DO provides basic mechanisms, but the degree of FAIRness being achieved by the actors making use of these mechanisms (repositories etc.) is dependent which kind of information they provide to which degree.*

[8] RDA Kernel Group; https://www.rd-alliance.org/groups/pid-kernel-information-wg

[9] RDA Research Data Collections; https://www.rd-alliance.org/groups/research-data-collections-wg.html

Operational Assertions

O1: DO Typing

The DO has a type summarising detailed metadata allowing us

- to form classes of DOs
- to associate operations with DO classes and thus to allow encapsulation
- to let DOs “talk to” other DOs ("data to data") and thus enable automatic workflows

O2: DO for Abstraction

The DO allows users to work at a level of abstraction by only dealing with the PID, type and metadata as far as possible and thus abstracting away from how and where the bit-sequences are exactly stored.

O3: Protocol for Interfacing with the DO

A unified DO interface protocol will allow us to access DOs independently of how repositories store their bit-sequences and other associated information. It will thus guarantee that all relevant parts are being preserved for example when moving or copying DOs.
***Note**: DOIP could use TLS[10] directly, but also use REST/HTTP[11] or SOAP[12], see our elaborations in the paper from Wittenburg, Strawn, Mons et al [4].*

O4: DO relevance for data management/stewardship

The management of scientific data is currently too complex and dependent on detailed knowledge of local computing environments. Abstraction to the DO level, and standardization of types and operations, all independent of the underlying heterogeneous systems, can simplify data management and decrease the difficulties of interoperation across domains.

Scientific Assertions

S1: DOs have a Scientific Value

The concept of DOs is not just another computer science idea, but has an intrinsic value for the organisation of the domain of digital entities in science.
***Note**: A first analysis of submitted scientific cases may give a n impression [5]. A more elaborate paper is being written currently.*

S2: DO and Trust

Trust is crucial for data sharing and reuse. DOs need to be used to enable the reconstruction of trust.
***Note**: See here the contribution of Koureas [6].*

S3: DOs are Universal

DOs and their organization are not restricted to a domain, application or community, but due to their generic and extensible concept, they are intended to be universal.

S4: DOs are discoverable

DOs support collecting, searching and development knowledge by the large communities.

S5: DOs can represent physical objects

In the realm of cyberspace, Digital Objects can act as digital surrogates (representatives) for physical objects in the real World.
***Note**: This can be a rather complicated relationship. We have a classification of real world objects (tree), the concept of such objects (in our mind), and their digital representation (FAIR-DOs). We need to be clear about the nature of this representation. It is an amalgam of object and concept*

[10] TLS; https://de.wikipedia.org/wiki/Transport_Layer_Security
[11] REST; https://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm
[12] SOAP; https://de.wikipedia.org/wiki/SOAP

transformed into a digital representation which certainly will be reduced compared to the complex networks in which real-world objects exist and are represented in our mind.

S6: Social Perspective
We should include the social perspective. This is already done partly by mentioning the DO/trust relation. But there are other aspects to address, for example, the ideas of the Solid (Social Linked Data) Project from Tim Berners-Lee[13], and his "Magna Carter for the web" activity[14], or activities that support a social contract for the Internet (Open Data Institute[15]). A global DO architecture will have dire social consequences, only some have to do with data ownership and privacy.

Side Discussion

Rob: Since the beginning of the GEDE discussions on Digital Objects[16], there has been a strong parallel between a digital object and a TCP/IP packet. I would like to point out some differences: from my point of view a DO may be seen as an atomic unit of digital sense (a DO is a self-consistent digital structure providing information about itself and how to consume its content).
For TCP/IP: Once a file is copied on my computer, I can handle and use it without any needs to know where the packets came from and who produced them. This seems a major difference with the sketched behaviour for DO where one has to resolve constantly (using DOIP) elements for getting metadata+type information.
Pushing things further, my questions are:

- may a DO exist without a network connection? If not, the DO definition should point this out.
- is it possible to handle and use a DO without network connection?

These questions were inspired by a discussion with colleagues from Airbus and CNES: they are seduced by the horizons opened by DOs and see the utility, but they usually operate without internet connection (ex. numerical simulation performed on very confidential items are cutted out from the internet for security reasons. Another example: space missions producing data have no access to internet).

In this phase I think that the set (metadata + format information) a user may obtain by resolving the DO items from registries while consuming a given DO should be embedded into the DO itself.

Peter:
I think that we need to distinguish between the DO and/or FAIR-DO concept and how we store and/or transport things. The DO/FAIR-DO definition just makes assumptions about the availability of certain attributes such as a PID, different kinds of metadata, etc. and requires the capability to resolve a number of links. One could argue now that this requires an access to the Internet. But the definition is agnostic about this, what you need to have is a resolution machine that turns a PID into useful state information.

For the storage/transport aspects we could package the context of a DO/FAIR-DO in a way that is most useful for a specific application. Indeed, one could pack all metadata which is needed into a container and transmit it. The Research Object project is working on these aspects with the idea in mind to put the whole context into a container so that some calculations can be repeated in a different environment. For these special missions one could also turn the PIDs into local PIDs or addresses or one could set up a local Handle resolver and take care that the required PIDs will be

[13] https://de.wikipedia.org/wiki/Solid_(Software)
[14] https://www.theguardian.com/technology/2018/nov/05/tim-berners-lee-launches-campaign-to-save-the-web-from-abuse
[15] https://en.wikipedia.org/wiki/Open_Data_Institute
[16] https://rd-alliance.org/group/gede-group-european-data-experts-rda/wiki/gede-digital-object-topic-group

transferred to the local system. Such redundancy schemes are already in use for example by ePIC. For a large data driven project in material science this solution is being thought of since it seems to be more practical and maintains basic principles.

Rob suggests that the required "*information (metadata etc.) should be embedded into the DO itself*". I assume that Rob means that it should be embedded in the bit-sequence of the DO which some communities do when they specify data formats that include all kinds of metadata in the file header. This is another way of packaging the context of the DO.

A larger problem will be when one relies on a number of different information entities all "belonging" to the context of a DO. Take for example the case that an institution wants to associate a blockchain entry with its DO to control licensing and transactions and industry is very much interested in this combination. Blockchain technology IS per definition a solution in a distributed scenario. One will most probably not include all transactions for a specific DO in its header or in other local form. Thus, there may be limitations to the scope of context one wants to put into a container.

An instance of a DO can thus be maintained locally without Internet access; however, there may be limitations.

References

[1] Wbg & Strawn Flavours paper to come

[2] Schultes & Wbg GO-DO-FAIR to come

[3] G. Strawn: Open Science, Business Analytics, and FAIR Digital Objects; http://doi.org/10.23728/b2share.6ceeed13eb6340fcb132bcb5b5e3d69a

[4] P. Wittenburg, G. Strawn, B. Mons, L. Bonino, E. Schultes: Digital Objects as Drivers towards Convergence in Data Infrastructures; http://doi.org/10.23728/b2share.b605d85809ca45679b110719b6c6cb11

[5] analysis paper to come

[6] D. Koureas: Digital Objects: The Science Case; https://rd-alliance.org/sites/default/files/GEDE-koureas-science-case-v2.pdf