Commenting on “Digital Object” Aspects

Peter Wittenburg (Max Planck Computing & Data Facility, Garching/Munich)
George Strawn (US National Academy of Sciences, Washington)
February 2019
Corresponding author: Peter Wittenburg (peter.wittenburg@mpcdf.mpg.de)

1. Introduction

The concept of Digital Objects (DO) goes back to discussions raised by Kahn & Wilensky in 1995 [1] which shortly afterwards was over-shadowed by the discussions about and implementations of the very successful World Wide Web. The concept of Digital Object was put back on the agenda by the discussions within the Data Foundation & Terminology group, which published a Core Data Model in 2014 [2] using the input from Kahn and many use cases from different scientific communities. Since that time much has happened within different initiatives and organizations such as RDA [3], CNRI [4], DONA [5], CODATA [6], BRDI [7], GOFAIR [8], expert groups initiated by funders such as the European Commission [9], and many others. Although other initiatives also started to use the term Digital Object, in this paper we will limit ourselves to explain the small differences between the definitions being used in the above mentioned initiatives and to put forth suggestions for future consideration.

The recent publication of the DO Interface Protocol v2.0 [10] represents a significant milestone in the increasingly active discussions and events referenced above and it reflects the ongoing intensive discussions. These resulted in sharpening our minds about the concepts and components essential for building DO-based infrastructures. Necessarily, some of the concept and term definitions have been subject of changes that may confuse those who have not been involved so deeply in the core discussions. This note is intended to clarify some confusions and open questions.

In addition, we will address a few topics that were raised in recent discussions about DOs.

2. Development of the DO Concept

2.1 Early Kahn & Wilensky Definitions

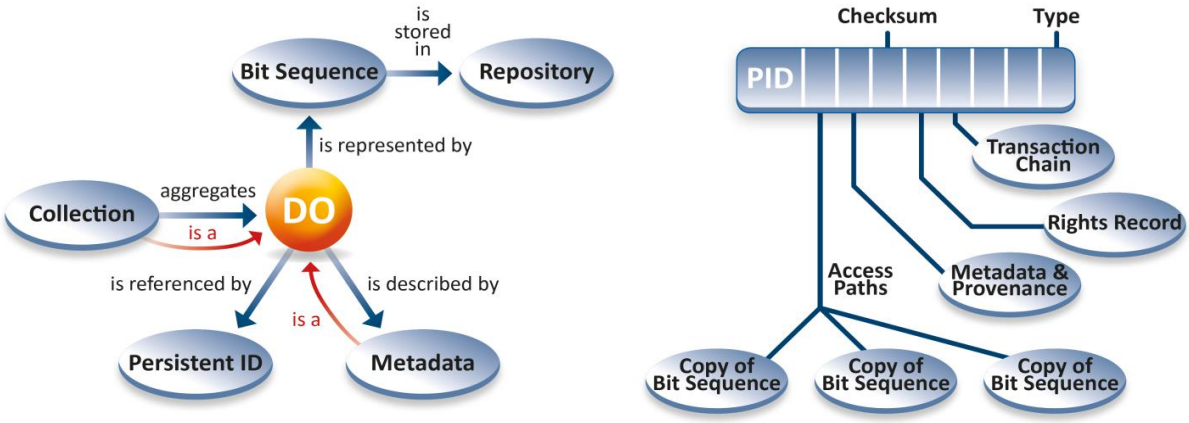
In their very first paper in 1995 Kahn & Wilensky defined a DO as having a structured bit sequence, a unique PID, and at least key metadata which includes the PID. In their revised paper in 2006, Kahn & Wilensky [11] introduced the Digital Object Architecture where they specify the DO as "an instance of an abstract data type that has two components, data and key-metadata. The key-metadata includes a handle i.e., an identifier globally unique to the digital object. It may also include other metadata, to be specified. In fact, the 2006 paper confirmed the Kahn & Wilensky notion of what a Digital Object is.

A DO has a structured bit sequence and key-metadata that includes a Handle (PID) and it is typed.

2.2 RDA DFT Definition

In 2014, the RDA Data Foundation and Terminology group published its core data model which was based on various contributions, including those from Kahn and many research community use cases. It has two major aspects: 1) It describes from a user point of view what a DO is and 2) it specifies the role of the PID as one that needs to include "passport-type of metadata" and persistent references to relevant information about the DO. These can be summarized by the two diagrams in figure 1.

A Digital Object is represented by a bit sequence (the content) stored in some repositories. It is referenced by a PID and is described by metadata. The *PID record* associated with the PID is the result of the PID resolution process and has attributes that describe essential properties of the DO (checksum, type, creation time, time of deletion, etc.) and/or point to essential information that is



required to work with the DO's bit sequence (locations of the bit sequence, descriptive metadata, deep metadata, provenance metadata, access rights, transactions stored in blockchains for example, etc.). Metadata descriptions of different types are themselves DOs and thus have a PID. DOs can be aggregated to collections which also are DOs. This definition of the DO includes the strong concepts of abstraction, binding, and encapsulation. Metadata will include a "type" specifications allowing researchers to define a set of operations associated with the type.

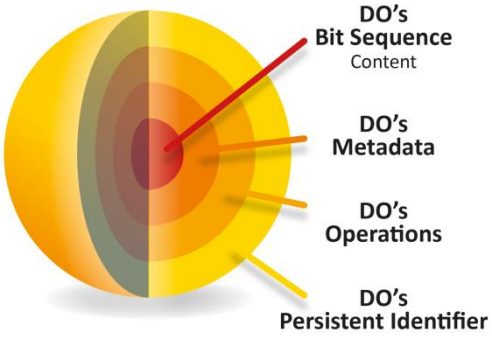
The DO has a bit sequence, a persistent PID and (different kinds of) metadata.

2.3 DO in the DOIP V2.0 Specification

In 2018, the DO Interface Protocol (DOIP) V2.0 specifications were published by the DONA Foundation. The DO definition is determined by the specification of the DOIP which focuses on the essentials and separates the DO level interaction from the step of interacting with the PID resolution system, which is specified by a separate protocol. From this perspective, it makes sense to define the DO as having a bit sequence, a PID and a type. The "type" of a DO is a special metadata assertion about the DO, which allows operations to associate with it. The DOIP V2.0 specifications include a set of standard operations such as create, delete, move, etc. to achieve uniformity.

The DO has a bit sequence, a persistent PID and a type.

2.4 DOs in the recent Wittenburg, Strawn, Bonino, Schultes paper



In 2019, Wittenburg, Strawn, Mons, Bonino and Schultes published their paper "Digital Objects as Drivers towards Convergence in Data Infrastructures" [12] and used the "atomic" metaphor that was introduced by the expert group report. After intensive discussions they designed the diagram in such a way that it expresses the abstraction, the binding and the encapsulation capabilities of the DOs. The diagram indicates that once you have the PID of the DO and it has been resolved to "state" information which basically are special metadata assertions about the DO then you have machine

actionable access to rich information which you need to make the DO FAIR such as the "type" information that allows you to link operations with the DOs and thus provides encapsulation[1], the pointers to different kinds of metadata (descriptive, deep scientific, provenance, rights, transactions, etc.) and thus provides binding, and pointers to bit sequences of any sorts and thus providing abstraction.

The limitation of this diagram is that it does not make explicit that the different kinds of metadata descriptions are also DOs and thus also have PIDs. And it also does not make explicit that DOs can be aggregated to virtual collections, also being DOs with metadata descriptions and PIDs.

The resolution of a DO's Persistent Identifier gives access to a broad range of machine actionable information that includes different kinds of metadata and the locations of the bit sequences

2.5 DO in the Expert Group Report

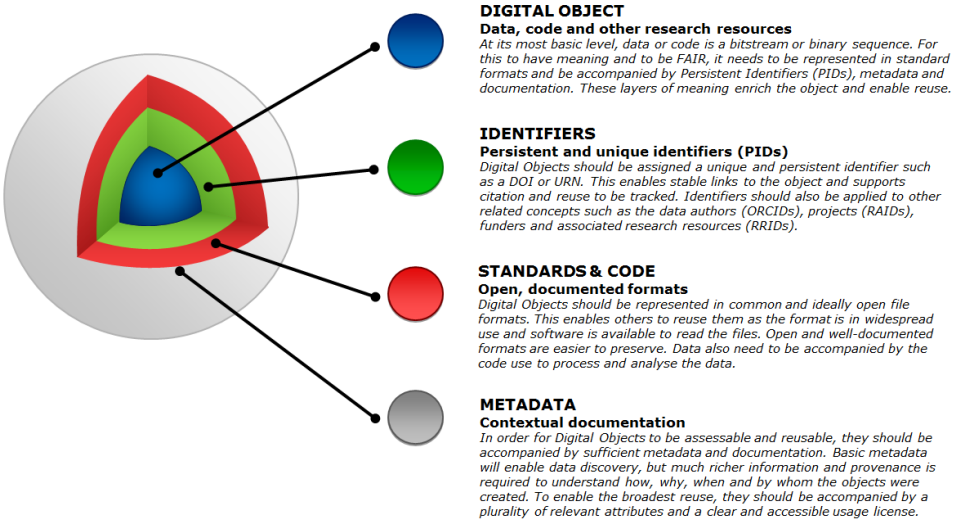
In 2018, the EC's expert group on FAIR Implementation published its report[2] introducing the term "FAIR DO" to express the relevance of the DO concept for the implementation of the FAIR principles. The document states the following about FAIR DOs: "*Central to the realisation of FAIR are **FAIR Digital Objects**. These objects could represent data, software, protocols or other research resources. They need to be accompanied by Persistent Identifiers (PIDs) and metadata rich enough to enable them to be reliably found, used and cited. Data should, in addition, be represented in* common *– and ideally open – formats, and be richly documented using metadata standards and vocabularies adopted by the related research community to enable interoperability and reuse. Software and algorithms, when shared, should include not just the source itself but also appropriate documentation including machine-actionable statements about dependencies and licensing".*

For data to be FAIR it needs to be Findable, Accessible, Interoperable and Reusable.

The document continues with *"FAIR Digital Objects, which sit in a wider **FAIR ecosystem** comprising services and infrastructures for FAIR. The realisation of FAIR relies on, at a minimum, the following essential components: policies, DMPs, identifiers, standards and repositories. In this ecosystem, data policies are issued by several stakeholders and help to define and regulate requirements for the running of data services. Data Management Plans provide a dynamic index that articulates the relevant information relating to a project and linkages with its various FAIR components. Persistent Identifiers are assigned to many aspects of the ecosystem including data, software, institutions, researchers, funders, projects and instruments. Specifications and standards are relevant in many ways, from metadata, vocabularies, and ontologies for data description to transfer and exchange protocols for data access, and standards governing the certification of repositories or composition of DMPs. Repositories offer databases and data services and should be certified to ensure trust."*

This more associative view on Digital Objects is also indicated by the diagram that the document is using. It wants to express that quite a number of activities such as data management planning, the definition of standards, the use of contextual information etc. need to be accompanied by the creation of digital objects. In doing so, it takes a broader view on DOs and does not focus on technical implementation objectives.

[1] A DO's "Type" is a metadata assertion about the DO. It can be used to access a data type registry to find out which kind of operations are defined for the DO dependent on the context.
[2] FAIR Implementation Report: https://doi.org/10.2777/1524



FAIR DOs identified by PIDs emerge and exist in a rich context of metadata and other relevant activities and information making DOs FAIR.

2.6 Conclusions about the DO Concept

We do not know whether the definition of the term "Digital Object" is now sufficiently stable, but for implementation purposes we see the need to distinguish between two major flavours of the term:

1. The **first definition** emerges from the requirement of the DO Interface Protocol for a minimal specification and a separation between the two interaction levels - the DO and PID levels. We suggest that in adherence to the DOIP specifications we call this the "**Digital Object (DO)**".
2. The **second definition** results from the scientific requirements as described in 2.2, 2.4 and 2.5 and the need to support the FAIR principles. To make the difference to the first definition clear we suggest making this the "**FAIR Digital Object (FAIR DO)**". Since the description of FAIR DOs in the expert group report is referring to the context in which they emerge and exist, we do not see a source of confusion.

When using the term FAIR-DO we need to make the relation to the FAIR principles. The DO as specified by the DOIP guarantees that once you have the PID that you can access an instance of its bit-sequence FAIR A2 and that open protocols are being used (FAIR A3). The definition of FAIR-DOs implies that when the PID is resolved it gives access to all kinds of metadata assertions that are required to meet the FAIR principles. The FAIR-DO definition implies that there is a basic mechanism to access all relevant information, however, "FAIR-DO" does not imply that a specific DO is 100% FAIR compliant. Having a FAIR-DO could still mean that for example the license conditions are not specified in a machine actionable form. The relationships between the FAIR principles and (FAIR-)DOs are explained in more detail in a recent Schulthes & Wittenburg paper [13].

The definition of the term "*Digital Objects*" results from the *DO Interface Protocol*.

The definition of the term "*FAIR Digital Object*" covers the scientific and FAIR views and is compliant to what the *RDA DFT group* specified and is further explained in the expert group report.

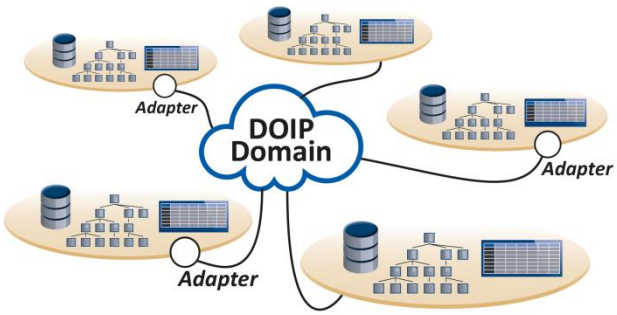
3. Collections of DOs

Collections include other collections and individual DOs, have a PID, metadata and a bit sequence. The bit sequence mainly consists of references, in general PIDs, to the collection's parts. Since collections can be complex hierarchical structures the processing of a collection requires a variety of repeated steps. Nevertheless, the DOIP will allow clients to connect to collections in the same way as individual DOs, i.e., its PID is used to get the pointers to its bit sequence and its metadata. It is up to the client to continue processing by resolving all part PIDs.

The RDA Collection group discussed the nature of collections and designed a "Collection Interface" [14]. It is an urgent task now to find out whether this Collection Interface is compliant with the DOIP V2.0.

4. DO Interface Protocol Adapters

It is obvious that DOIP creates an interoperability layer between sub-domains based on the same data organization and data model, i.e., it reduces an N*N to a 1*N problem as TCP/IP did in the Internet and had the capacity to create a huge momentum. In general, these sub-domains will be made up of repositories and/or registries storing DOs' bit sequences and metadata. This implies that repositories and registries that want to connect to the DOIP data domain need to develop and maintain adapters that translate DOIP interaction requests to internal processes. If, for example, a repository uses a relational database system to store data and metadata, the resolution of a PID needs to lead to an execution of queries that yield the required content. If, for example, the resolution results in accessing an object in a cloud store, the reference is a local identifier allowing the cloud-store to access and deliver the appropriate bit-sequence. Depending on the local configurations - for example, repositories could use a Hierarchical Storage Management System - the adapters could be more or less complex. The RDA "Practical Policy" [15] group analyzed the complexity of adapters and produced a cookbook of microcode templates and code snippets which illustrates the complexity of the task.



It is an urgent task to understand the amount of work required to develop and maintain such adapters for the DOIP domain and to provide reusable templates for different repository setups. A library of reusable code should become available. We expect, however, that repositories will stepwise adapt their local data organization to reduce the efforts for maintaining adaptors.

The simplicity of the DOIP should not mislead us to assume that adapting to the Digital Object domain will be trivial.

5. Role of DONA Foundation

In several discussions, the question was raised whether the use of DOIP will be restricted by specific licenses and fees. The intention of the authors of the DOIP and the involved community is very clear: DOIP has to facilitate a global Digital Object Space (DOS) which is not determined by commercial interests, but must be as open as the Internet is: if you have a TCP/IP compliant device, you are able to connect it to the Internet. The Internet is not owned by a company or any organization, it is basically owned by the global Internet community.

To protect the domain of Digital Objects and DOIP against against capture by any special interest group, commercial or otherwise, the authors of DOIP associated the rights on DOIP with the International DONA Foundation which is a non-profit and independent organization located in Switzerland. Its mission is clearly defined by its statutes in so far as it will maintain the core components of the DOS and take care that the global DOS community will "own it": if you have a DOIP compliant object server you should be able to connect it to DOS.

The core components maintained by the DONA Foundation, summarized as Digital Object Architecture, are the Global Handle Root System, represented by a number of globally distributed Multiple Primary Agencies, the Identifier Resolution Interface Protocol and the DO Interface Protocol. The DONA Foundation will take care that innovative solutions can emerge around the DOA components which define an island of stability in a highly dynamic domain and that urgent community needs will be taken up.

The international and independent DONA Foundation, located in Switzerland, will protect the core elements of the Digital Object Architecture against capture by any special interest group, commercial or otherwise and thus enable the next phase of innovation and exploitation.

Acknowledgements

We would like to thank Larry Lannom for his valuable comments.

References

[1] R. Kahn, R. Wilensky: A Framework for Distributed Digital Object Services, 1995; http://www.cnri.reston.va.us/home/cstr/arch/k-w.html
[2] RDA DFT Core Terms and Model; http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318
[3] https://www.rd-alliance.org/
[4] https://www.cnri.reston.va.us/
[5] https://www.dona.net/
[6] http://www.codata.org/
[7] http://sites.nationalacademies.org/PGA/brdi/index.htm
[8] https://www.go-fair.org/
[9] S. Hodson, et al: FAIR Implementation Report; https://doi.org/10.2777/1524
[10] https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf
[11] R. Kahn, R. Wilensky: A framework for distributed digital object services, 2006; https://www.doi.org/topics/2006_05_02_Kahn_Framework.pdf
[12] 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
[13] Schultes&Wittenburg: GO-DO-FAIR; to come
[14] https://www.rd-alliance.org/groups/research-data-collections-wg.html
[15] https://rd-alliance.org/groups/practical-policy-wg.html