Digital Objects as a Concept to Help Implementing FAIR and EOSC

© Peter Wittenburg
Max Planck Computing and Data Facility,
Garching/Munich, Germany
peter.wittenburg@mpcdf.mpg.de

Abstract. Current data practices are very inefficient and many global initiatives (RDA, CODATA, GO FAIR) are working hard on finding agreements to reduce the ongoing proliferation of standards and tools. With the FAIR principles a global language has been widely agreed upon to guide data practices, however, the principles cannot be seen as guidelines of how to build data infrastructures. The definition of the Internet, where in general meaningless messages are being exchanged, needs to be amended by a new paradigm for global data management that helps to organize data for the next 100 years. The concept of Digital Objects that has already shown its usefulness in programming complex systems is suggested by a few initiatives, since it opens the path to proper and stable data organizations and to encapsulation with all its proven advantages.

Keywords: data analytics, data management, data interoperability, digital objects.

1 Introduction

In their recent paper Wittenburg & Strawn [1] summarized a few facts indicating that the data domain in science and industry suffers from huge fragmentation at all levels, from a proliferation of tools to meet the detailed needs of users and as a consequence of this from an extreme inefficiency in data driven projects where about 80% of the work is simply wasted with data wrangling, where many projects fail due to the required efforts and where many researchers are excluded from big data work. They compare the situation in the data domain with other domains which had similar phases of proliferation of standards, practices, tools and services which they call "creolization", point to a few global initiatives that aim at finding convergence and thus conclude that the time is ripe to come to revolutionizing agreements as they happened in the other investigated domains. Their conclusion is that the concept of Digital Objects has the potential to change our data practices fundamentally.

For some years some initiatives are active to work out solutions pushing agreements in the data domain and thus decrease the size of the solutions space. The Research Data Alliance (RDA) [2] is relying on the bottom-up organization of data professionals to work out recommendations of policies, procedures, components and their interfaces, CODATA [3] is a global organization focusing on policies, and the recently started GO FAIR initiative [4] wants to foster implementation work by forming so-called Implementation Networks.

In parallel, the FAIR principles have been published [5] summarizing years of discussions about Findability, Accessibility, Interoperability and Re-Usability of data in a way which can now be called a global language most data scientists agree upon. An expert group of the European Commission has recently published a preliminary version of its recommendation paper on how to put FAIR principles into practice to open it for broad discussions [6]. Also funders are acting globally to promote the idea of convergence. Here, I only want to refer to the European Open Science Cloud [7] as an initiative of the European Commission to bundle forces towards building an interoperable eco-system of data infrastructures based on the work that has already been done. Yet, however, it lacks a unifying concept.

The question remains whether the suggested concept of Digital Objects indeed does help to achieve these goals and to make FAIR common practice.

2 Internet Gap

The Internet specifies devices that exchange basically meaningless messages. At the sender side meaningful entities are chopped into fragments and routed through a network of nodes to the receiver which then puts the fragments together again using standard protocols such as TCP. Finally, however, users/clients want to exchange meaningful entities of different types which is why early applications such as FTP and later HTTP were introduced. While FTP helps exchanging files, HTTP was defined to exchange HTML encoded information. As is known the latter, defining the World Wide Web, was very successful and extended in many different ways to exchange different types of meaningful information. The Web now is used for "managing" and in particular exchanging the increasing amount of data and information although it has not been designed for such a comprehensive task. The web gave us a way to point at all manner of remote files and databases, send commands and requests to those data sources, and build pockets of networked information management. The web, however, suffers from a few basic deficits:

Proceedings of the XX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2018), Moscow, Russia, October 9-12, 2018

- Identification of digital entities is not stable and not independent of protocols which will change over the coming decades.
- The Web is by design ephemeral, i.e. it is changing continuously which can be seen as one of its strengths, but which is not adequate for data management.
- The Web does not provide consistent and consistently applied security and quality measures.
- The Web can basically be seen as a delivery system.

It becomes increasingly obvious that we need a new consolidated paradigm for global data management given the dramatically increasing amounts and complexity of data and the necessary shifts towards automatic procedures. It needs to be a paradigm that helps to organize data for the next 100 years to make sure that we do not lose our digital memory, i.e. we may not only think about short term solutions that may help to bridge between the many solutions which are out there.

3 Digital Objects

Digital Objects were already introduced in an early paper by Kahn & Wilensky in 1995 and then in an updated version in 2006 [8,9]. The concept is very much related with computer science concepts such as "object-oriented programming" [10], "abstract data types" [11] and "object stores" [12] which are at the basis of state-of-the-art cloud systems such as Amazon's S3. We can therefore claim that the concept of "objects", is closely related with ideas such as "encapsulation", "virtualization", and "interfacing by defined methods", has shown its great importance to help designing complex systems.

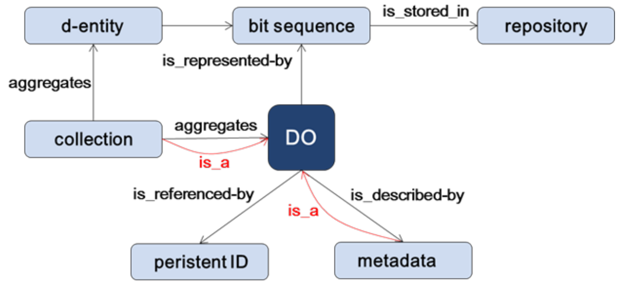

Figure 1 This figure indicates the core data model as it was worked out within RDA DFT.

In 2014, the RDA group "Data Foundation and Terminology" (DFT) published its results on a core data model and the corresponding basic terminology [13]. It summarized the discussions about Digital Objects (DO) in so far as it stated that

- DOs are at the core of our data organizations in so far as it has the capacity to bind crucial entities which are necessary for a stable and reusable domain of data;
- DOs have a bit sequence (content) which can be stored in various repositories, are referenced by a unique and persistent identifier (PID) issued by a trustworthy globally available resolution system and is described by various types of metadata that can include descriptive, system, access rights, license, contractual, transactional and other kinds of meta information about the DO;
- Metadata itself are DOs;
- DOs can be combined to collections which also are DOs, i.e. have a PID and are described by metadata;
- DOs can include all kinds of digital information such as data, software, configurations, representations of persons, institutions, semantic concepts, etc.

We can also look schematically at DOs from a different point of view, if we extend the above definition by encapsulation principles as being introduced by the RDA group "Data Type Registry" [14]. One of the metadata types describing a DO is its "type" which is summarizing several technical metadata attributes. A Data Type Registry allows users to relate data types with operations which are also DOs of a specific type. These defined operations allow users to realize the encapsulation principle as requested by Abstract Data Types. Figure 2 indicates this encapsulation which can be implemented when strong and stable binding is being realized. The usage of PID systems such as the Handle System [15] allows to create such a strong and stable binding, since the PID records allow to include pointers (PIDs) to all relevant entities and metadata types associated with a DO.

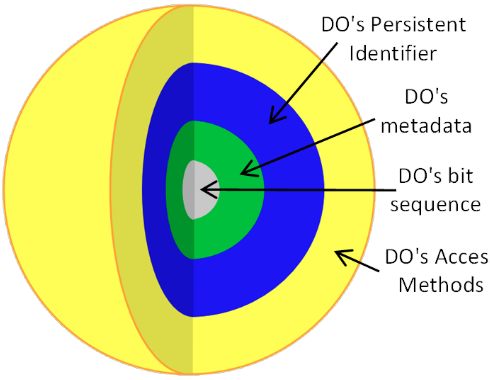

Figure 2 This figure schematically indicates the types of encapsulations that can be implemented with DOs.

Recently, a first version of a protocol to interact with DOs, the DO Interface Protocol (DOIP), has been opened for broad discussion by DONA [16]. It basically describes how clients interact with DOs where all involved actors are represented by PIDs. DOIP is meant to have a relevance that is comparable to TCP/IP for the Internet, i.e. it should become a fundamental protocol to manage and exchange digital objects.

The C2CAMP initiative [17] is devoted to implement a DO based infrastructure including to understand DOs as active entities that have methods associated with them.

4 Conclusions

The concept of "objects" is not new in computer science and it has shown its strengths in designing complex systems. Also in relationship with data the concept has been successful in so far as cloud storage systems are making us of its characteristics.

DOs as introduced in this paper are by definition FAIR compliant, i.e. they can be used to build the basis of a FAIR compliant eco system of data infrastructures. In addition, when systematically assigning PIDs it comes along with a strong binding system that allows us to build data organizations which can be maintained over many decades, to implement proper encapsulation, to increase data security, to document and control transactions of data where necessary. DOs open the way towards automatic procedures where specific data that are useful for specific analytics can be found by software agents and where procedures can be found to start automatic workflows on collections of data of a specific type. DOs thus help directly to make data Findable, Accessible and Re-Usable, and it will facilitate interoperability.

Acknowledgments. I would like to thank all colleagues in CNRI, DONA and RDA for all the various contributions which led to the development of the concept of Digital Objects.

References

[1] Wittenburg, P., Strawn, G.: Common Patterns in Revolutionary Infrastructures and Data. http://doi.org/10.23728/b2share.4e8ac36c0dd343da81fd9e83e72805a0

[2] https://www.rd-alliance.org/

[3] http://www.codata.org/

[4] https://www.go-fair.org/

[5] https://www.force11.org/group/fairgroup/fairprinciples

[6] https://zenodo.org/record/1285272

[7] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

[8] Kahn, R., Wilensky, R.: Key Concepts in the Architecture of the Digital Library. http://www.dlib.org/dlib/July95/07arms.html

[9] Kahn, R., Wilensky, R.: A framework for distributed digital object services. https://www.doi.org/topics/2006_05_02_Kahn_Framework.pdf

[10] https://en.wikipedia.org/wiki/Object-oriented_programming

[11] https://en.wikipedia.org/wiki/Abstract_data_type

[12] https://en.wikipedia.org/wiki/Object_storage

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

[14] RDA DTR: https://rd-alliance.org/group/data-type-registries-wg/outcomes/data-type-registries

[15] https://en.wikipedia.org/wiki/Handle_System

[16] DOIP is not yet published

[17] https://github.com/c2camp/core