GEDE Workshop on Digital Objects

25.09 GEDE Dinner at 20.00 in Brasserie 1898, Avenue d'Auderghem 4, 1040 Bruxelles, Belgium

**Date**: 26.9.2018, **Time**: 8.30 to 16.00

**Place**: APCO, Rue Montoyer 47, 1000 Bruxelles, Belgium

**Organisers**:

Co-Chairs of the GEDE Digital Object Topic Group[[1]](#footnote-1) (Peter Wittenburg, Dimitris Koureas, Koenraad de Smedt) in collaboration with GEDE (Margareta Hellström, Carlo Maria Zwölf and Zsuzsanna Szeredi)

**Keynote Speaker**:

Dr. Robert Kahn (CNRI) (explanation see below)

The presence of Dr. Kahn in Brussels is an excellent opportunity to discuss in depth the concept of DOs, the suggestions for DO based infrastructures, and the DOs' potential value in structuring the domain of scientific digital entities. We suggest a workshop that starts along the key elements of the recently submitted proposal (Efficiency Case, DO Case, Scientific Case) and ends in an open discussion.

**Workshop Focus:**

For the workshop we will focus on a number of questions and every participant (in person or remote) should feel motivated to come up with comments on them. Here are some typical questions:

* What is the motivation behind DOs from a computer science perspective and what is their expected impact to build infrastructures and architectures based on them?
* What is the motivation behind DOs from a scientific perspective and what is their expected impact to structure the world of digital entities in the sciences?
* What types of interoperability problems can DOs solve, where can they facilitate finding solutions and where will they not help? Will they help to achieve FAIR compliance?
* Will DOs help to move towards automatic workflows and what is required to make them work?
* Which are the essential components to realise DO based architectures?
* Do we have already clear cases (designed or implemented) that show the potential of DOs?

The workshop is not meant to come up with final answers to these questions, but we should get first answers, an idea of the dimensions that are relevant and guidelines for the follow up discussions such as a confirmation of the need to discuss and agree upon on a DO Interface Protocol.

**Embedding**

At the evening of 25.9 we will organise an informal dinner for interested GEDE members (own costs).

**Program**

|  |  |  |
| --- | --- | --- |
| 25.9 Joint Dinner | | |
| 20.00 | joint dinner for interested GEDE people |  |
|  |  |  |
| 26.9 GEDE Workshop | | |
| 8.30 | Reception and Coffee |  |
| 9.00 | Welcome and Introduction | Peter Wittenburg (chair) |
| 9.20 | Keynote on Digital Object Architecture | Robert Kahn |
| 10.30 | Coffee |  |
| 11.00 | Implementation work in C2CAMP | Tobias Weigel (chair) |
| 11.20 | - Putting DO components in action  - GO FAIR and Digital Objects | Ulrich Schwardtmann  Erik Schulthess |
| 12.00 | Lunch |  |
| 13.00 | DOs- The Scientific Case | Dimitris Koureas (chair) |
| 13.20 | - The DO concept in Language Research  - The DO concept in Biodiversity Research  - The DO concept in Climate Modeling  - The DO concept in Athmospheric Research  - The DO concept in Atomic & Molecular Res. | Dieter v Uytvanck (CLARIN ERIC)  Alex Hardisty (DISSCO)  Tobias Weigel (ENES)  Margareta Helström (ICOS)  Carlo Maria Zwölf (VAMDC) |
| 15.00 | Final Discussion | Koenraad de Smedt (chair) |
| 16.00 | end |  |

In addition to the keynote from Dr. Kahn who will address the computer science view of the global data management challenges, the computer science aspects of DOs and the scientific views should be in the focus to guide our future work. Therefore, the workshop is split mainly into two sessions. While the morning is devoted to discussing computer science aspects of Digital Objects, the afternoon should be devoted to views from scientific domains, i.e. we expect contributions from scientific domains how they see the use of the DO concept in their discipline and how it would help to structure the digital domain.

**Registration and Restrictions**

The workshop is planned as a face-to-face meeting with the option to include other GEDE DO experts via virtual conference techniques. Please, register yourself when you plan to come to Brussels to participate in person. We will have a room with limited capacity, i.e. when the capacity limit has been reached we will use the first registered, first reserved principle. The costs for lunch and coffee will be taken from the support funds from the RDA Europe 4 project. Registration will be closed at 15.9 or when the limit is reached to make contracts.

**Please, register asap but at least until 15.9. for participation at this survey monkey site:** <https://www.surveymonkey.com/r/NP3YDK5>

**We will let you know immediately after whether the capacity limit has been reached.**

**Reporting**

A report will be created from this meeting and the presentations will be made available. We also intend to write and discuss a position statement which will be part of the report.

## Introduction of Dr. Robert Kahn

Dr. Kahn was not only one of the two designers of TCP/IP and thus at the source of our current days Internet, he also wrote in 1995 the first paper on Digital Objects together with Robert Wilensky. This early paper was revised in 2006. Since TCP/IP only describes the exchange of in general meaningless messages between Internet devices with an IP address, there was a need to exchange meaningful entities. FTP was an early protocol to exchange files and HTTP was another very successful protocol to exchange Web information. Kahn & Wilensky realised the need to define a protocol that is generic and thus includes files, web pages and other possible entities which they called Digital Objects. Kahn and his team at CNRI then started to design and develop components that could realise such a world of DOs. The most well-known component is the Handle System which allows anyone now to assign Handles to Digital Objects and to resolve identities into useful "state" information. With the setup of the DONA Foundation the Handle System can now be seen as a common good not owned by one person or company anymore. Other components have been design and partly developed such as the Digital Object Interface Protocol which may indeed change our practices.

It should be mentioned that Dr. Kahn got many awards for his work amongst which is the Turing award.

## Case Descriptions (copied from EUDOn Proposal)

For easy reference we include here the "case" descriptions which were used in the recent proposal.

**Scientific Case**

Data-intensive science, the systematic analysis of large scale data, is rapidly permeating all areas of research in academia and industry, as it offers promise to revolutionize our understanding of the world. One of the keys for the success of data intensive science is FAIR, a globally agreed acronym for Findability, Accessibility, Interoperability and Re-Usability. In turn, one of the keys to FAIR is an inter­connected ecosystem of infrastructures. These infrastructures currently serve their scientific communities of practice, curating and annotating data in a domain-specific manner. Arguably, interoperation of research e-infrastructures is a prerequisite for streamlining cross-disciplinary organis­ational, syntactic and semantic interoperability and largely depends on the adoption of common data systems. Establishing such a complex system will require global cooperation on the introduction of new concepts. We explore in particular one concept called "Digital Object (DO)" and its different flavours to help implementing FAIR compliant infrastructures.

A few examples may indicate the kind of research questions that data-driven scientists want to solve, but they still find many obstacles to carry out such work.

For discovering the causes of for example brain diseases it is widely agreed that machine learning algorithms could be used to find hidden patterns in a variety of data such as genetic data, brain-imaging data and others when being correlated with typical phenomenological patterns. However, large amounts of training data are required which come from many different labs all using different methods and tools. The hurdles for integrating useful data, which in these cases are sensitive data, from different sources are currently so high that such research is hardly doable even for big research centres. Knowing that increasingly more humans suffer from dementia and Alzheimer disease, for example, better solutions for overcoming these hurdles are urgently needed. DOs offer the ground for clear identification and thus findability of data, for making proper usage agreements, for effectively tracing proper re-usage, for associating verified tools with data, and for facilitating interoperability.

Digital language data, as studied in the humanities, cognitive science and related fields, are extremely diverse in their nature, formatting and annotation and also the domain of tools is rather diverse. Integrating data and tools alone on a platform with selection menus would not help since the "normal" researcher working with language data would be completely lost in this heterogeneity and only a few specialists would be capable to manoeuvre in subspaces of the field of acceptable combinations. A large European infrastructure is currently developing a switchboard solution linking specific data types to tools that have proven their usefulness for these types. DOs are exactly the type of solution allowing realising such a switchboard elegantly since they are typed and have other relevant metadata that can automatically be retrieved and these types can be associated with specific tools using the same basic mechanisms. Using these methods would enable a "computer-naive" researcher to create his/her workflows only applying domain knowledge that then can automatically process data of specific types in the intended way.

In material science a trend to make better use of the experimental and simulation data generated in thousands of labs worldwide is clearly visible combined with the expectation that access to massive amounts of results will enable the researchers to come to new categorisations of compound materials applying smart machine learning algorithms. Such multidimensional categorisations would allow researchers and industry to much more quickly find suitable material combinations given a specific new application. Large initiatives in the US and Europe are working in this direction indicating the huge relevance of this data driven approach. Also here the identification of the results, of the included materials and their attributes and the used creation processes, specific relationships and more are of crucial importance to make progress. Due to their binding capacity DOs have the potential to facilitate such research and make the high expectations reality.

In biodiversity, our extensive natural science collections (natural history specimens) have been, for hundreds of years, the focal point of research for new species discovery. Genomic information, morphological and ecological traits, and occurrence records are among the data classes individually extracted by those physical objects. Despite originating from the same physical object, data is currently fragmented and isolated, across small and larger repositories, with no or minimum capacity to bring this information together. The introduction of DOs in the field of biodiversity studies, could provide a technologically and socio-culturally acceptable way through which currently dispersed information is brought back together as a meaningful and machine actionable digital object, which effectively acts as the digital surrogate of the physical object.

Many other cases from different research disciplines could be mentioned and will be studied in the proposed Action. It is important to note that DOs are not just a technical concept, but a way to optimally structure domain data in a way that facilitates Data-Intensive Science.

**Efficiency Case**

Recent surveys indicate how important increased efficiency in the future handling of data will be in order to tackle unsolved scientific challenges, to include a much broader group of researchers in data-intensive science, to be able to monitor usage, increase trust and foster the path towards automatic processing which will be a must in order to keep a competitive edge.

Different studies show that most of the time of data scientists is wasted with “data wrangling” which includes the steps before the real analysis can be done. The study from RDA Europe mentions 75% of wasted time in 2014, a study from MIT cited by M. Brodie mentions 80% in 2015, a study from CrowdFlower in industry mentions 79% in 2017. From the research domain we know that many projects cannot be started and that many contributors cannot participate in data science. Huge fragmentation is hampering fast progress. The survey from CrowdFlower does not even include semantic interoperability, which relates to the analysis of knowledge after being extracted from a series of measurements. The main inefficiencies are caused by bad data organisation and quality. For accessible data it is often hard or impossible to find or interpret metadata that enables valid processing of the data streams.

The degree of automation in data-driven science globally, but in particular in Europe, is not adequate given the increasing quantities of data and their inherent complexity. Manual and ad-hoc operations do not scale and in general lead to undocumented and non-reproducible results which have been identified in several publications as a huge problem for scholarly communication. Automation of complex and data-demanding methods such as machine learning requires systematic and systemic approaches to the organisation of data. The development of easy to use workflow systems which are flexible enough to cope with various conditions requires harmonisation of the basic organisation of data. In summary, we can state that one of the major factors preventing the broad take-up of data-intensive science is the lack of broadly agreed basic operating mechanisms such as potentially offered by DOs.

**Digital Object Case**

Broad interactions at different global platforms such as RDA, FORCE11, C2CAMP, GO FAIR, at PIDapalooza and at workshops in Europe, the US and China have resulted in a broad agreement about the crucial role of persistent and globally resolvable identifiers (PIDs) for all data entities as a basic requirement for a fundamental change. The GEDE collaboration which brought together delegates of 47 large European research infrastructures (most ESFRI projects) agreed after a year of intensive discussions on a paper about the need to use PIDs and their patterns of usage. One of the key messages in this paper is that the granularity with which PIDs should be associated to data is dependent on what is meaningful in a given scientific context.

However, it has been shown recently that simply assigning PIDs is not in itself sufficient to achieve convergence on essential data management principles and thereby overcome the huge fragmentation that obstructs progress. Within global initiatives there is the growing conviction that we must properly exploit the increasing global use of PIDs and the availability of global PID resolving systems.The concept of DO architecture, as suggested already 2006 by Kahn and Wilensky indicated a path towards fundamental changes of data practices. Furthermore, the RDA Data Foundation and Terminology working group describes a DO as a structured bit sequence stored in some repositories, associated with a persistent, unique and resolvable Identifier (PID) and described by metadata. Some kernel metadata are being associated with the PIDs to achieve the high degree of binding different types of information necessary for efficient and especially automatic processing. RDA working groups are working on standardising these kernel attributes. DOs can be simple or complex, i.e. the latter exist of aggregated collections of DOs, and their content can include data of different types, metadata, software code, machine configurations, etc. DOs have a type enabling their association with functions by use of Data Type Registries as has been defined by another RDA working group.

The term "object" is widely used in modern software technology since it implies an encapsulation of its internal structure by offering a set of tested functions that can be executed. The term became also very popular through the introduction of cloud stores which are also been called "object stores". Representing each "object" by a locally valid hash value is a step of virtualisation since the user does not need to know anymore how and where exactly the object is stored. Associated with this hash value is also the metadata information needed for finding and processing.

Much work is being done to define the term "research objects" with the intention to capture the complex context of digital object to improve scholarly publishing. Closely related to this concept is the concept of packaging which discusses ways to pack such rich contexts into containers that can be exchanged easily to different environments to be used for further processing. These concepts are complemented by approaches such as Linked Open Data that offer ways to expose and exploit complex semantic relationships.

1. All supporters of the EUDOn proposal are invited to this event. The pressure is high to make progress in the area of data management. [↑](#footnote-ref-1)