GEDE Workshop on Digital Objects Report

At 26.9.2018 the first workshop on Digital Objects was held by the DO Topic Group within GEDE. It was organised by the Co-Chairs of the GEDE Digital Object Topic Group (Peter Wittenburg, Dimitris Koureas, Koenraad de Smedt) in collaboration with GEDE (Margareta Hellström, Carlo Maria Zwölf and Zsuzsanna Szeredi) and brought together 25 face-to-face and 35 remote participants.

The workshop was held in a week where Dr. Robert Kahn, the inventor of TCP/IP and one of the Internet pioneers, was visiting Brussels to have many high level interactions about the early evolution of Internet, the main characteristics of the Internet success, his role in DARPA and the Digital Object Architecture described by him and others in early papers from 1995 and refined in 2006. Therefore, we were happy to have Dr. Kahn as the keynote speaker.

Executive Summary

This workshop of the RDA GEDE topic Group on Digital Objects was right in time. It allows us to draw a number of essential conclusions:

- The concept of Digital Objects is not just a useful IT concept, but it can play a crucial role in organising the increasingly complex domain of digital entities within and across the scientific communities if a close interaction can be continued as planned in GEDE DO and if DOs will be used to re-establish trust in distributed and anonymous scenarios.
- This DO concept shows a way forward that would allow us to build the stable basis for a FAIR compliant digital domain for centuries.
- R. Kahn's explanations show clearly that
    - one of the most crucial ingredients for realising the DO domain is the existence of a globally resolution system and that the neutral and international DONA foundation is in charge to maintain the Handle System and make it available to everyone interested
    - the DO Interface Protocol as the other crucial ingredient has been implemented as version 2 and will soon be released and
    - with respect to the DOIP he will follow the same principles of availability to everyone interested free of charge as he decided to do it for TCP/IP.
- EOSC will need to be implemented as a federation of amongst others repositories and registries already being established and maintained by the ESFRI projects and many other projects. DO based architectures can exactly be the interlinking and interoperability mechanisms at the data organisation layer that could make this federation happen as tCP/IP interlinked the various existing networks decades ago. In addition, it would facilitate the interoperability work at other layers.
- A number of European experts working on data intensive science projects are global drivers in implementing the DO concept to the benefit of science. ESFRI ERICs and similar initiatives are active in including the DO concept in their plans. In addition, European eInfrastructures such as ePIC are key players for the global Handle System and are testing new DO related concepts such as the Data Type Registry allowing to relate operations with DO of specific types.
- The DO concept and the FAIR principles are complementary which will be further explained in a paper currently being written.
- The discussion also showed that blockchain technology and open science are widely contradictory despite some usage scenarios where highly sensitive data is being shared.

To remain at the cutting edge of the global developments, it will not be sufficient to work with funds already assigned. It will be crucial for the European actors to apply for funds where the concept of DO will be central which will also be a way to start the implementation of EOSC.

Report

Peter Wittenburg opened the workshop by explaining first that GEDE is a collaboration of about 47 large European research infrastructure projects and ERICs (RI) organising their work around discussing special topics of concern such as now Digital Objects, best practices for Citations, the role of Repositories and the use of Blockchain technology in science. The goal of the topic discussions is always to summarise and evaluate practices in the different RIs, to relate them with state-of-the-art discussions and to find broad agreements where possible. These groups are open to all experts from the GEDE members. In addition he described the motivation of starting the DO topic group. Three large surveys from science and industry indicate clearly that about 80 % of experts' time in data-intensive science (DIS) projects is wasted by data wrangling. This does not even consider the effort needed for semantic interoperability. Due to these inefficiencies the costs are huge, many DIS projects simply fail and many researchers are excluded from DIS. Different approaches can be seen in science and industry to tackle the inefficiencies, but despite some good first steps and results in initiatives such as RDA, CODATA, WDS, DONA and GO FAIR no breakthrough has yet been achieved. New comprehensive initiatives such as EOSC offer a chance to change this. Until now, however, EOSC lacks a convincing architectural approach. The concept of Digital Objects, in combination with making data and services 'FAIR', offers a great potential, since it implies principles of abstraction and encapsulation which have already shown their potential for dealing with complexity. Many questions remain and the workshop is one step towards finding answers.

Robert Kahn first spoke about the early history of the Internet and the key factors that were essential to its huge success. The unique role of the DARPA agency at that time allowed experts to follow their visions about new inventions and realizing them thanks to excellent funding support. Technically the Internet was such a success because a) packet switching as opposed to circuit switching was a much more efficient means of communication, b) the basic protocols (IP and TCP) were defined on purpose to be uncomplicated, c) the protocols were general purpose and all specifications were open and free to be used by anyone and d) there were no intellectual property or ownership issues to deal with as the Internet was developed with government support in the public interest. Visionary people might have seen the need to connect the different emerging computer networks and to create a global "information system", but no alternate technology took hold at the time. Further, although the scalability of the architecture was predicted, no one anticipated the widespread adoption globally or that it would continue to work without significant architectural change as the underlying technology scaled up by factors of up to 10 million (computation, communication and storage).

He described the concept of a Digital Object Architecture (DOA) that evolved from early ideas about mobile program environments. The same overriding principles that applied to the development of the Internet protocols were adopted, but by evolving the Internet to simplify the overall management of information in digital form. He defines a Digital Object as a sequence of bits or a set or sequences of bits, incorporating a work or a portion of a work or other information in which a party has rights or interests, or in which there is value. Each DO has a type and has, as an essential element, an associated unique persistent identifier. The DOA is an open architecture, independent from any underlying technology, minimizes complexity for users, is non-proprietary and publicly available without charge. Components of the DOA are an identifier/resolution system that resolves identifiers to "state information" about the DOs, repositories that store DOs and enable access to them, and registries that provide information (generally known as metadata) about DOs.

The concept of a DO allows a uniform conceptual approach to managing and interacting with information represented in digital form. A DO Interface Protocol (DOIP) provides a means of interoperability in accessing any kind of DO (including entire information systems as DOs along with

the information entities incorporated within). The DOIP version 2.0 has been specified and will soon be released by the DONA Foundation. An early version of DOIP was known as the Repository Access Protocol. After further development, it was later described in ITU-T Recommendation X.1255 and defines the interaction between a client and any digital object, including, for example, a repository hosting a collection of DOs. All entities and actions are identified by persistent identifiers in the DOIP.
In the ensuing discussion, Robert Kahn expressed his conviction that the concept of DO can enable a new wave of interoperability between the many different information systems in the world. He also made clear that blockchain technology is simply a special way of connecting blocks which are, in essence, DOs and that first ideas about connecting blocks in safe and useful ways were first specified decades ago. Transforming the huge volume of legacy data to be part of the world of known DOs is a huge task and should be addressed by the various research communities. Part of the discussion was dedicated to issues of monopolisation and monetisation of crucial technology, where Robert Kahn and Patrice Lyons took a very clear position: DOIP (as with TCP/IP before it) are crucial pieces of technology and should be free of any monopolisation and monetisation strategies - it would hamper broad uptake. However, implementations of the protocols or components of the DOA are the province of individuals and/or organizations that create them. This is basically what happens in the Internet today, where equipment and services are generally made available commercially, but the overall architecture is open, non-proprietary and free of charge.

Tobias Weigel introduced the C2CAMP collaboration which emerged from some RDA groups and now covers 20 institutions from 5 continents. They all share the interest to improve the management of data by using the concept of Digital Objects and wish to synchronize on DO related developments. Most of the members have already discussed their intentions to build in the DO concept into their infrastructure building and started looking at opportunities within their current and future work. A few others that were not so deeply involved in the RDA Data Fabric discussions currently take a more observational role. C2CAMP is a loose collaboration sharing the abstraction view outlined by L. Lannom where users working with some clients on notebooks, smartphones or other devices operate on logical representations of Digital Objects such as PIDs and Metadata, but in fact ignore all details where and how the metadata and data are exactly stored. This PID and metadata information will be provided by registries and repositories which are meant to be virtualized services decoupling concerns of storage and access. Everyone interested and committed to implementing DO related infrastructure elements can participate in C2CAMP. The US colleagues have already got two grants for implementing DO related aspects.

Ulrich Schwardmann introduced the ePIC collaboration and the DONA foundation briefly and the role of his organisation (GWDG) in these activities, before he elaborated on DOs, Types and DOIP. DONA is a Swiss foundation guiding the future of the Handle System[1] which is broadly used globally. It maintains a highly reliable system of Multi-Primary Agencies (MPAs) that build the distributed and redundant root system to resolve Handles and to act as registration authorities to give prefixes to interested parties. ePIC is a collaboration of some major computing and data centres in Europe that issue prefixes and suffixes of Handles to their many customers. They have also established a scheme of redundancy to guarantee high availability. GWDG is one of the currently 10 MPAs and also leading the ePIC consortium which gives services to the scientific community.
He confirmed the important roles of abstraction to hide heterogeneity, virtualisation to provide a layer of abstraction between data and applications and encapsulation to provide a layer of abstraction between inner heterogeneity and complexity and outer simplification, all three aspects being addressed by the DO concept. He also stressed the idea of "binding" all entities of a DO by using the attributes in the PID record; however, in order to make them machine actionable their types (being key-value pairs) need to be registered in a Data Type Registry, which is also a service from ePIC. In addition, he pointed out that data driven research very much relies on methods using

[1] It should be mentioned that DOIs are Handles associated with a specific business model defined by the DOI Foundation.

data. In this respect the suggestion for a DOIP, to provide a framework for the usage of registered services, comes very close to the ideal representation of mathematical functions, which he sees as very promising.

Erik Schultes explained the emergence of the GO FAIR initiative he is part of and the major topic of his talk which is to draw relations between the FAIR principles and the DO concept. Both have been discussed in different communities. The FAIR principles evolved from a workshop in Leiden and they see their current actions in GO FAIR very much in the line of the convergence forming described in the Wittenburg & Strawn paper. GO FAIR evolves in three pillars all related to the the FAIR principles: awareness raising, stimulating capacity building and stimulating converging implementations.
He then gave a first impression on the relationship between FAIR principles and DOs by looking into the 4 different dimensions of F.A.I.R and explained that currently a more elaborated paper is being written on this matter. There seems to be a large overlap. One of the major metaphors driving the GO FAIR work is the one of the propeller with three major blades: data, tools and compute. Presently, data, tools and compute are largely non-FAIR. He sees the FAIRification of data, tools and compute along axes that take us closer to the hub of the propeller. DO based infrastructure is located at the hub of the propeller, and is connecting all three dimensions as well. Seeing it this way one can imagine the architecture for a data infrastructure as an extension of the well-known hourglass representation of the Internet to a 3 lobe “hourglass”, one lobe each for data, tools and compute.

Dimitris Koureas elaborated on the scientific relevance of the DO concept and on the expectations from the scientific point of view. He started with referring to the continuously increasing volumes of data and a possible paradox mentioned by C. Borgman paraphrasing S. T. Coleridge "data, data, everywhere, nor any drop to drink". We are creating huge amounts of data, but exploiting them scientifically to a small extent. Data driven solutions are urgently required, but we need to understand to deliver data at scale, form and precision to make that happen. The big question is whether the availability of data will help us to significantly contribute to the grand dimensions of sustainable development as identified by the UN assembly. These identify global challenges which are carried out locally by researchers who are acting in general in their small communities of practice. This bears the risk that data is being disconnected from the context in which they were produced which can lead to lack of trust in data and the services built on them. The big question for him is therefore whether DO-based architectures can help to rebuild trust in the various dimensions. Therefore, further developing DO-based architectures cannot just be a realisation of a technical concept, but needs close interactions with domain scientists to consider the domain specific needs, to operate in trusted frameworks and to deliver a clear added value. In short, the implementation needs to be driven by strong scientific cases which puts research infrastructure in a front role. Science requires light and when the source of light changes, science will change. There is a huge chance for DO-based architectures to change light conditions promising new things.

Dieter van Uytvanck explained the nature of the CLARIN ERIC, a pan-European infrastructure in the domain of language resources, before he described the intended application of DOs in their infrastructure. The CLARIN infrastructure maintains a large amount of data in different languages and tools ready to operate on the data all being described by component metadata and referenced by PIDs. CLARIN is using Handles to refer to all data[1] and beyond. It recently developed a switchboard concept that allows matching of language resources and tools based on profile descriptions such as language, format and task to be carried out. The switchboard is a lightweight construction that can easily be installed in different contexts and that can easily be invoked from these contexts. If for example a certain text is being uploaded to a cloud service, the intention is to be able to use a mouse click to invoke syntactical processing of that text. To improve the switchboard mechanisms, CLARIN

[1] Some national CLARIN consortia are now forced by policy requirements to also register DOIs pointing to landing pages, which leads to the strange situation that objects now have two identifiers for different purposes.

has the following minimal requirements for a DO representation: access not only to the PID of the bit sequence of the text, but also to the detailed metadata, to its MIME type, to a language identifier according to ISO 639-3 and a usage license specification, all in machine readable form. Another application where CLARIN intends to apply the concept of DOs has to do with their way of allowing users to build virtual collections. When making use of DOs the binding concept is of great importance, since for example data and metadata are closely coupled through the PID record.
CLARIN is given a great opportunity to make use of the DO Interface Protocol in the future. All repositories should provide a DO compliant interface which would give machine readable information on, for example, where to find the DO's bit sequence, the metadata and how to access applicable tools. This could open the way to improved automation since all repositories would react in the same way and provide the relevant information about the DO.

Maggie Hellström explained the ICOS (Integrated Carbon Observation System) research infrastructure she is working for and which has as major tasks to offer open access to harmonised and integrated measurements on carbon cycle, greenhouse gas emissions and atmospheric concentrations of greenhouse gases from Europe and beyond and to carry out data curation so that the data is ready to be used for research work. In addition to the assignment of PIDs and rich metadata as required by the FAIR principles, it is important for ICOS to offer scalable and documented services with high availability. PIDs which have the checksum as suffix point to landing pages that support content negotiation. Metadata is based on a documented ontology and all metadata assertions are stored as RDF triples with a SPARQL endpoint to support human and machine access.
An increasing number of data services are offered through the portal to support discovery, data carting, collection building and previewing interactive maps indicating the location of measurements. A REST interface supports machine access to services where applicable. The ICOS repository is set up in a way very much in line with the concept of DO which would facilitate its integration into a DO-based data infrastructure.

Alex Hardisty described the huge challenges to create a consistent European domain of digital information about the objects and observations in natural sciences (plants, animals, fossils, rocks, etc.) to be covered by the DiSSCo initiative (www.dissco.eu). Central are the specimens (about 1.5 billion) curated in the more than 115 facilities in 21 European countries where each specimen is representing one physical object with an increasing number of digital representations. "Digital specimen" becomes the centrepiece of the DiSSCo knowledge base - they can be viewed as "dynamic knowledge boxes" covering all links to all core information about a "thing" stored in one place and how they relate to other things and classes. The great challenge for DiSSCo is to create a single data-driven European virtual collection of digital specimen (DS) from all the many contributions and to provide a domain of stable references supported by Handle services and to demonstrate its added value to science and the public.
It can be easily recognised that the number of links that need to be maintained in a stable way is in the order of 50 billion. A major challenge for the implementation will be to find a balance between a fast presentation of informative registry records and the need to fetch and unpack comprehensive object content from a repository. DS are characterised by a set of mandatory and optional elements, but extension mechanisms based on clear rules must be in place to cope with new requirements that will evolve over time. The Handle System has been selected as basis where different alternatives for prefixes are being studied and where flexibility with respect to the kernel attributes will be required. The project aims at sustainability of at least 30 years, but during the discussion it became clear that due to the huge development costs for all these digital infrastructures we need to aim for mechanisms that will hold for longer time periods.

Carlo Maria Zwölf first described the basic tasks of and mechanisms in VAMDC - the "Virtual Atomic and Molecular Data Centre" - and then elaborated on a few questions that cannot be easily solved. VAMDC is aggregating information about many different distributed nodes which host detailed

information about atomic and molecular structures to enable user searches via a unified portal. A user query is thus dispatched to the corresponding nodes, the responses are being collected and returned to the user. In case of necessity the user can directly contact a corresponding node. This information exchange requires a harmonised interface which is specified by the widely supported XSAMS standard. XSAMS is a rigorous and unambiguous object model for atomic and molecular physics serialised into an XML schema and it contains descriptors for attributes, relations, rules, restrictions etc. One could view the returned information as Digital Objects since bit-sequences containing the information, together with related metadata and state information, are identified with PIDs. Q1: However, if these results are interpreted as DOs there is a problem linked with the data cross matching. For that, CMZ gave the example of the quantum-state description: the same physical state may have different digital representation (based on quantum numbers, energy from a fundamental state or ionisation energy). How to automatically switch from one representation to the other, for enabling machine driven comparison?
Q2: The data extracted from VAMDC are then stored as items in Zenodo, associated with a DOI. These different items may contain data for specific elements/molecules at given frequency ranges. But currently there is no easy way to perform complex queries in Zenodo (e.g. “extract the data for Helium in the range [100A, 2000]”). How to solve such challenges in a framework such as EOSC?
Q3: There is a huge challenge with respect to data curation and error tracking in DO architectures. In case an error has been detected in a given DO as sketched above, there is no way to correct all the results produced by reusing the “error”-DO. Also, how can we handle provenance and plagiarism in such scenarios?
The discussion revealed that the three presentation versions need to be identified as being derived from one object having its own identifier with references to the derived versions, that blockchain technology does not help in open scientific frameworks and that there is hardly any chance to roll back citations to erroneous DOs.

Tobias Weigel presented the DO related work within the climate modelling community which is organized within the Earth System Grid Federation (ESGF) of which IS-ENES is the European branch. This community agrees to run a number of clearly specified experiments which allow a comparison of the simulation results to explain the differences between models and derive new insight into climate change. The basis of these experiments and the scientific driver is the CMIP (International Climate Model Intercomparison Projects) specification which evolved over the last decades also with respect to the amount of data involved. While the results amounted to 1 GB in 1995, the CMIP6 experiments currently conducted are expected to produce between 300 PB to 3 EB of data. First runs have been conducted, but the whole exercise will span multiple years from 2018 on. The data volume and complexity pose new technical and organizational challenges and demand solutions such as automated digital object management, increasing workflow support and provenance aggregation and supporting to work at higher levels of abstraction. Key for the community is the systematic use of PIDs (Handles) in their data distribution workflows. To address scalability and availability concerns, a queueing system is used to support peak performance in registration activities.
Future work such as replicating data and supporting HPC workflow will require clearly specified interfaces such as DO Interface Protocol supporting mechanisms to record what modules actually did (provenance). A workflow system based on type-triggered automated processing making use of the DO model as being discussed in RDA Data Fabrics and C2CAMP is therefore highly attractive to the community. Therefore, DOs now have the best chances to become true "primary citizen" within ENES and revolutionize the data management for future CMIP and other community projects. Part of this is also increasing interest by a wider variety of users, including “downstream” users such as from public services and climate adaptation, to employ server-side virtual research environments such as Jupyter to create workflows for example to carry out machine learning calculations.

Finally, **Koenraad de Smedt** and **Peter Wittenburg** wrapped up the meeting by summarising the next steps for the GEDE DO topic group work. Currently a few additional papers about DO related aspects are being in progress:

- One paper is currently being written by some C2CAMP and GO FAIR experts to look at the DOs from a computer science point of view.
- Another paper is being prepared currently by Dimitris Koureas and Peter Wittenburg which will look at DO from a scientific point of view and we will invite all GEDE DO members to provide use cases to be included in the paper.
- A fourth paper is currently being written by Erik Schultes and Peter Wittenburg to explain in more detail what the relation is between DOs and the FAIR principles.
- Finally we should discuss within the GEDE DO group about assertions about DOs as it is a good habit in the GEDE topic groups to push agreement finding.

Appendix A

GEDE Workshop on Digital Objects

Date: 26.9.2018, **Time**: 8.30 to 16.00
Place: APCO, Rue Montoyer 47, 1000 Bruxelles, Belgium

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

Organisers:
Co-Chairs of the GEDE Digital Object Topic Group[2] (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.

[2] All supporters of the EUDOn proposal are invited to this event. The pressure is high to make progress in the area of data management.

Embedding

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

Program

Remote Login

Please join my meeting from your computer, tablet or smartphone.

https://global.gotomeeting.com/join/177773245

more information see below

25.9 Joint Dinner		
20.00	joint dinner for interested GEDE people Brasserie 1898, Avenue d'Auderghem 4, 1040 Bruxelles, Belgium	
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	Objects, types, collections and operations in DOIP	Ulrich Schwardmann (ePIC)
	GOing FAIR & DOing FAIR	Erik Schultess (GO FAIR)
12.00	Lunch	
13.00	DOs- The Scientific Case	Dimitris Koureas (chair)
13.20	Digitals Objects as direct input into the CLARIN Language Resource Switchboard	Twan Gosen, Dieter van Uytvanck (CLARIN)
	How a Digital Object Architecture could help ICOS streamline data service provisioning	Margareta Hellström (ICOS)
	DiSSCo Digital Specimens- Widening access to natural science collections	Alex Hardisty (DISSCO)
	The DO Case in Virtual Atomic and Molecular Data Centre	Carlo Maria Zwölf (VAMDC)
	Digital Object Management for ENES: Challenges and opportunities	Tobias Weigel (ENES)
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.

Abstracts

Ulrich Schwardmann (ePIC): Objects, types, collections and operations in DOIP

The DOIP describes a world of digital objects and operations provided by repositories, referenced by identifiers and enhanced by type metadata, which is made interoperable again by type definitions in data type registries. In the world of digital objects one always can see valuable applications like collection repositories to structure these objects in a generic way and to build user specific views and work benches on such structures by a digital object browser. The next step is to include also operations into this picture. Since with REST services operations are already ubiquous in the HTTP world, the question is, how this huge amount of existing technology can be adapted into the DOIP world. The talk will also try to show a possible bridge here.

Erik Schultess (GO FAIR): GOing FAIR & DOing FAIRGOing FAIR & DOing FAIR

The Digital Object framework is an abstraction layer striving towards technology-indepdent, future-proof, and increasingly automated operations between data, software, and compute resources. The 15 FAIR Principles are high-level specifications for the automated Findability, Accession, semantic Interoperation, and Re-use of data, software and services. Although the DO Framework and the FAIR Principles have very different origins, they nonetheless share overlapping and complementary features. In their current states of development, the DOIP specifies elementary “informatics" operations on DOs, while the FAIR Principles provide some specification for the technical, domain

specific, and provenance metadata that are necessary to inform DOIP operations. I will describe GO FAIR, a bottom initiative coordinating a very large and diverse stakeholder community actively building implementations (including “Metadata for Machines”) that demonstrate the FAIR Principles in practice. I will describe how ongoing GO FAIR activities, including the C2camp Implementation Network, could play a role in accelerating the adoption and application of DOs in an emerging data infrastructure.

Twan Gosen, Dieter van Uytvanck (CLARIN): Digitals Objects as direct input into the CLARIN Language Resource Switchboard

Numerous repositories offer data and associated metadata, following the CMDI specification, within the CLARIN infrastructure. The CMDI metadata is harvested and collected in a central repository: the Virtual Language Observatory (VLO). After finding a relevant piece of information, the data object and metadata object can be provided, either as bit streams or by identifiers, to the Language Resource Switchboard (LRS), to get a suggestion of tools available to operate on the data object. In this context, digital objects are a natural fit, provided the available metadata matches our needs, to enhance the LRS. By providing a digital object identifier to the switchboard, the switchboard can utilize the DO protocol to obtain all relevant information about the digital about, such as mime type and language. Any of the suggested tools can in turn utilize the DO protocol to obtain a bitstream to the actual object data itself in order to run the tools processing pipeline.

Margareta Hellström (ICOS): How a Digital Object Architecture could help ICOS streamline data service provisioning

The ICOS Carbon Portal (CP) manages, curates and disseminates both greenhouse gas observational data (measured by its own station networks), as well as outputs of advanced atmospheric and ecosystem models (provided by external parties). The CP assigns Handle-based PIDs (from ePIC or DataCite) to all data objects it manages. All relevant metadata are kept in the CP catalogue , which is based on semantic web & open linked data (LOD) concepts. We are now actively looking into the best way to support (automated) workflows and processing of ICOS data in cloud environments (like EGI federated cloud). Important aspects include optimizing access to, and linking of, all relevant metadata from various sources (the ICOS catalog, databases at PID registries, and others), including descriptive information (context of acquisition), data contents (variable types), and data processing basics (model version). It remains to be seen to what degree ICOS can adapt the DO Network approach, but we are willing to be one of the test cases – perhaps with a special focus on combining data type definitions stored across both LOD-based and PID-based registries.

Alex Hardisty (DISSCO): DiSSCo Digital Specimens- Widening access to natural science collections

For more than 300 years, scientists have collected and studied plants, animals, rocks, minerals, and fossils from our planet. Constituting 55% of the global asset base, and representing 80% of the world's species, more than 1.5 billion specimens are housed, organised and catalogued as collections in many hundreds of institutes and museums across Europe. Together, these represent an unparalleled resource, a scientific infrastructure for knowledge and discovery about the world's biodiversity; it's past, present and future and its influence on global challenges in environment and society. In June 2018 the European Strategy Forum on Research Infrastructures (ESFRI) accepted the importance of this resource and included the Distributed System of Scientific Collections (DiSSCo) into the ESFRI Roadmap 2018 as a priority research infrastructure to commence operations in 2025.
With an expected 30-year lifespan, DiSSCo aims at digital transformation of today's slow, expensive, inefficient and limited system where the need to physically visit collections and the absence of linkages to relevant information represent significant impediments. Our architecture inserts a 'Digital Specimen Object Layer' unifying natural science collections into a single data-driven European virtual Collection offering wider, more flexible, 'FAIR' access for a range of biodiversity science and policy

applications. This is expected to lead to faster insights for lower cost. Acting as surrogates for the physical specimens in collections, DiSSCo places Digital Specimens at the heart of an interconnected graph of diverse and dispersed data classes that can include imagery, taxonomy, relevant scientific literature, genetic sequence and trait data, agricultural, toxicology and ecosystem data and much more.
Immediate concerns for consideration in the present workshop include: i) the social aspects of 'selling' the benefits of the digital object approach; ii) achieving balance between fast presentation of informative registry records and the need to fetch and unpack comprehensive object content from a repository; and iii) extensibility of dynamic Digital Specimens to support new information types whilst maintaining backwards compatibility for older systems.

Carlo Maria Zwölf (VAMDC): The DO Case in Virtual Atomic and Molecular Data Centre

The VAMDC e-infrastructure federates into an interoperable way ~30 heterogeneous and independent atomic and molecular databases. All the outputs produced by this infrastructure are identified by a resolvable PID (Persistent Identifier) and are formatted using a rigorous XSD schema (XML Schema for Atoms Molecules and Solids; This schema is a computer model for all the processes and physics contained in VAMDC).
The data extracted from VAMDC are indeed a good approximation of what is designed by "Digital Object".
In our presentation we will describe the main issues we experienced:

- In using for scientific purposes generic repositories (e.g. Zenodo, Eudat) for storing VAMDC-DO.
- In automatic handling (e.g. comparison, cross-matching) of VAMDC-DO.

Tobias Weigel (ENES): Digital Object Management for ENES: Challenges and opportunities
The ENES[3] community develops and operates e-infrastructures and end-user tools and services which are used by the European but also the global climate data community. An integral part of this is the global Earth System Grid Federation (ESGF[4]), in which ENES participates with operative nodes and contributes key developments.

The ENES community faces manifold challenges for data management, processing, and overall aspects of making data FAIR, based on technologies from ESGF and other initiatives such as EUDAT and EOSC:

- Data volumes, but also the number of data objects, are increasing exponentially. The CMIP6 data aggregation is expected to hold more than 100 PB of data, distributed among multiple 100's of millions of files. The key operational processes dealing with data objects must scale up, yet they rely on technologies that are difficult to automate individually. Resources will not increase as much as data object number and volume, and therefore, approaches for **automated digital object management** are highly desired.
- Data workflows are becoming more complex, covering the entire range of individual stages from data production, dissemination (e.g., via the ESGF), quality control and long-term archival, reuse and repurposing by a wide area of actors. However, there are not yet conceptually mature and operationally feasible approaches for **enabling automated workflows** and ensuring that **provenance is gathered automatically and exposed in a user-friendly way to data consumers, producers and funders.**

[3] European Network for Earth System Modelling, https://www.enes.org
[4] https://esgf.llnl.gov

- An increasing drive to keep middleware services and direct end-user facing services alive beyond individual project lifetime calls for **sustainable funding and business models**, which has given ENES an impetus to participate in initiatives such as EUDAT and EOSC. While the integration with generic infrastructures and services used in other disciplines opens up innovation potential, it also generates additional friction as it adds complexity to technical and organizational architectures. Common agreements and models for Digital Objects may help to pave the way here.
- Data consumers will work at **higher levels of abstraction** in the areas of data processing and analytics. ENES is reacting by offering more sophisticated Virtual Research Environments (VREs) to an increasingly diverse user base with less expertise in core modelling areas. **Tailored services** with shallower learning curves, strong user integration and good usability are needed. The way in which ENES establishes Digital Objects as primary citizen is a necessary and logical part of the abstraction effort, and also supports and facilitates the understanding of workflow processes.
- Also beyond data processing, the diversity of consumers for ENES data products and data services is increasing. This includes users largely unfamiliar with the data generation and refinement process, which presuppose knowledge of limitations and assumptions not obvious to those unfamiliar with Earth system modelling. ENES therefore has a strong motivation to enter discussions with **new user communities**, particularly in the area of **social sciences, public administration, planning and policy making.**

The ENES community looks towards initiatives such as C2CAMP to address these points, starting with individual prototypes and small-scale solutions. There is however now a good opportunity to reach out to other communities and work towards a larger, overarching architecture and operationally capable solutions, based on the concepts and models the C2CAMP participants have discussed and matured in the past. Central design principles are adherence to agreed interfaces, for example based on specifications from RDA, IETF or W3C; finding the balance between complexity and feasibility, particularly regarding metadata and semantics; and building long-term sustainable services for data processing and provenance tracking.

Case Descriptions

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 inter-connected 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.