TECHNICAL ANNEX — EUDOn

# 1. S&T EXCELLENCE

## 1.1. CHALLENGE

### 1.1.1. DESCRIPTION OF THE CHALLENGE (MAIN AIM)

**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 [1], 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 [2], a study from MIT cited by M. Brodie mentions 80% in 2015 [3], a study from CrowdFlower in industry mentions 79% in 2017 [4]. 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 [5], 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 [6] 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 [7]. 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 [8] 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 [9] indicated a path towards fundamental changes of data practices. Furthermore, the RDA Data Foundation and Terminology working group [10] 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 [11]. 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 [12].

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" [13] 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 [14]. These concepts are complemented by approaches such as Linked Open Data [15] that offer ways to expose and exploit complex semantic relationships.

**EUDOn Challenges**

The initiatives for Open Science and Open Data request a common understanding which is essential to achieve improvements. The proposed Action will organise networks of experts from European research infrastructures, e-Infrastructures, the RDA Europe community and other international research communities to address the eminent roadblocks by applying the DO concept. We can identify the following specific challenges to be addressed by the Action:

**C1. Bottlenecks in cutting edge data-intensive science use cases**

We need a common understanding of the bottlenecks in cutting-edge use cases for data intensive science (DIS) across various disciplines. We need a deep analysis of the great inefficiencies that hamper DIS and prohibits the execution of new projects that could lead to new scientific insights. We need a deep analysis of the best way in which the introduction of DOs could help to overcome current limitations. Only a broad network approach respecting the various positions and contributions from the different disciplines will bring us ahead.

**C2. Missing DO specifications and missing insights about their potential impact on research practices and on research infrastructures.**

The concept of DOs has been introduced by computer science in different flavours to solve specific problems in smart ways such as by introducing cloud stores which are offering the huge stores to data-intensive science that are required to tackle complex problems and a step of virtualisation allowing laymen to use such stores. However, clouds are silos with their own administrative layers. Global data-intensive science needs an open infrastructure that turns this idea of virtualisation into a global one without barriers, i.e. globally registered and resolved DOs will build the bridge between the many silos and repositories all with different data organisations. We need to draw detailed specifications of DOs in a context of trust-worthy practices and technologies, and we need to promote steps for their verification.

**C3. Missing platform to exchange knowhow about DO architecture specifications, their usefulness in various disciplines and useful software code that could be shared.**

Changes in existing infrastructures are not only painful due to the investments needed for adaptations, but in particular since they require learning new workflows and tools from the researchers. These costs result in an utterly conservative behaviour of the researchers which is contrary to the needed changes in data practices. Platforms are needed where researchers can openly exchange ideas about DO architectures, implementations, where compliance criteria are established to ensure the investments made and where useful code can be exchanged. Since data-driven science is crossing disciplinary and geographical boundaries, such a platform needs to be organised globally and cross-disciplinary. Open, technology neutral interactions will establish the trust in taking steps towards adoption of new methods and thus overcome the existing hesitations.

**C4. Lack of testbed projects that can drive the transformation, accelerate the specification work in RDA and interact with comparable initiatives working on different foci**

A stimulation of testbeds for selected scientific cases implementing the DO concept would be helpful to accelerate the specification work. Progress should be monitored and gaps and bottlenecks should be identified in concrete actions. Embedding such a DO-oriented platform in related initiatives such as GO FAIR will help to position, shape and test the DO concept in the evolving landscape.

**C5. Need to mobilise and relevant groups**

A big challenge in modern research is getting the attention of stakeholders in a time where people are overloaded with information of all sorts, with solutions full of promises and claims, with a plethora of software programs created by smart people etc. making decision taking so problematic. Only clear and trusted forums where factual and unbiased information is being exchanged will help to overcome hesitation thresholds. Therefore, suitable dissemination strategies of results and progress need to be formulated.

**C6. Demand for capacity for the specification and implementation work**

Another big challenge is to find ways to engage a young generation of data scientists to take up the concepts and participate actively in the specification, implementation and management work. Ways need to be found to quickly build up knowledge and reduce psychological barriers.

**C7. Need for training of young data scientists in order to transfer the evolving knowledge and skills from the early adopters.**

When leading data scientists with deep insights from long experience will be retiring, their knowledge and skills need to be transfered to a new generation of experts. There is a need to quickly build up a community of young experts by organising hackathons/datathons which include hands-on session and participants' active contribution towards solution.

### 1.1.2. RELEVANCE AND TIMELINESS

In the USA there is a drive to fund programs that focus on DOs. In China the establishment of a national PID infrastructure available for science and industry will open the door for innovative developments related with DOs. Therefore, in Europe there is an urgent need for a network which will bring together stakeholders in science and technology in order to drive innovation to gain a competitive edge in data driven science. The proposed Action will bring together these stakeholders across disciplines, while linking with global players. At European level we can refer to activities to improve data practices in the following arenas:

* EOSC which tries to create a momentum towards a harmonised European data landscape fostering a single data market;
* the GO FAIR initiative with its three FAIR oriented pillars to change culture, to stimulate training and to bring together developers;
* RDA and the GEDE collaboration working on an agreement about specifications of common components of a future data infrastructure;
* the Industrial Data Space, which tries to come up with an abstract reference architecture guiding the implementation work towards an efficient data infrastructure in industry.

On the brink of a major step towards convergence [8], it is time for a common understanding of how this can be achieved. The proposed Action is timely for bringing about an alignment movement and critical mass with a broad European dimension, to push the conceptualisation and implementation work in close collaboration with comparable initiatives at the global level. Crucial for the Action is that it will give networking a clear direction by putting DO in focus. Adding the concept is not risky, since: a) it is general enough to act as a platform for substantial growth (as TCP/IP did for the Internet); b) it is a strong concept that has been subject of discussions for almost 20 years; and c) it has already shown its potential strength in the cloud concept having been implemented by several companies. In contrast to many other networks, the Action is defining the conceptual nucleus which per definition is FAIR compliant and is adding an implementation direction for open science.

## 1.2. OBJECTIVES

### 1.2.1. RESEARCH COORDINATION OBJECTIVES

The proposed action will bring together, for the first time at this scale and breadth, a network of experts from different research fields to jointly look into the innovation and consolidation potential of the DO concept for data science and its direct applications to the data challenges of different scientific disciplines. This network includes European research and data infrastructure initiatives and their experts to interact about the concept, its required components, implications and impact. In particular early career data scientists need to be engaged since they will determine data-intensive science in the coming decades. The network bridges research infrastructures and e-Infrastructures, since the notion of DOs offers a unifying level of abstraction for data management, access, interoperability (at the level of data organisation) and reuse, and thus a way to implement FAIR. The Action will support work already undertaken by a few global initiatives in testing the DO concept, its flavours and its current implementations against jointly developed criteria and requirements. Of great importance will be to disseminate the acquired knowledge in suitable ways, organise a variety of capacity building events to attract early career data scientists and transfer the evolving knowledge and technology from the early adopters, and to closely interact with and contribute to European initiatives such as EOSC, PRACE, OpenAIRE, EUDAT, GO FAIR, FREYA, the ESFRI projects etc. The overall objective is to achieve a momentum towards implementing a more coherent infrastructure landscape and to stimulate and accelerate the specification work in RDA by providing clear directions. Participation in international forums which will bring together relevant initiatives from for example the US, China, South Africa, Australia and Russia need to be guaranteed, to exchange the results and commonly define future directions.

The following concrete objectives will be pursued:

**O1. Collectively carry out a SWOT/GAP analysis for the need of DOs and DO based architectures, drawing on a collection of relevant scientific use-cases from various disciplines.**

Collect relevant scientific use cases of data-intensive projects from various disciplines and at different scales. Analyse them with respect to their potential of increasing efficiency and effectiveness for the researchers in form of a SWOT/GAP analysis. Extract salient components from these use cases and identify gaps. Identify their specifications and compare them with components already specified to formulate generic requirements where applicable. The results of this work will be regular updates of online documents/wiki with snapshots every six months to indicate progress and remaining bottlenecks. The Action can build on results from different initiatives such as RDA and will include industry.

**O2. Propose specific approaches to DO specifications and chart their potential impact on research practices and on research infrastructures, including e-Infrastructures.**

Propose specific approaches to DO implementations and chart their potential impact on research practices and on research infrastructures, including e-Infrastructures. Discuss possible solutions to the fragmentation at the level of data organisation and in facilitating semantic interoperability. Based on the needs of data-driven projects it must be understood what the problems are which DOs can solve, which efficiency effect can be achieved and what the limitations are. This objective will also lead to an understanding of how the introduction of DOs would change data culture, for example by combining DOs with smart contracts implemented in blockchain technology to increase trust. It must also be understood what the transformational implications of introducing generic components are and which amount of work would be required to adapt existing systems. For interested communities, implementation plans will have to be worked out that can give directions and set priorities. Also this objective will result in regular updates of online documents/wiki with snapshots every six months and where applicable in a stimulation of the RDA specification work.

**O3. Propose specific approaches to DO architecture specifications and chart their potential impact on research practice and on research infrastructure, including e-Infrastructures.**

Propose, agree upon, and disseminate specifications for DO architectures. Integrate ideas, define compliance criteria, work out interface, linkage and harmonisation specifications, and exchange code of components from existing implementation initiatives. EUDOn will not implement software, but will establish an interaction platform to keep a momentum and to achieve increasing convergence with respect to DO related implementations. DO-related concepts are currently being discussed in Europe in initiatives such as GO FAIR, various research infrastructures, in implementation initiatives such as EOSC, EUDAT and PRACE, but also in the US, China, South Africa and other countries. EUDOn will maintain a close interaction with these networks to converge ideas. In these interactions EUDOn will also identify roadblocks that need to be tackled with a certain priority and stimulate specification and development work. The results of this work are regular updates of online documents/wiki with snapshots every 6 months and where applicable in a stimulation of the RDA specification work.

**O4. Stimulate and contribute capacity for the specification work in RDA or the implementation work in networks such as GO FAIR.**

Develop implementation concepts for selected cases, focusing on versatile architecture models as an output to demonstrate the potential of DOs to make data-driven science more efficient and to put the FAIR principles into practice. Continuously monitor progress in filling gaps and bottlenecks and stimulate concrete action, be it at the specification level by RDA groups or at implementation level within initiatives that participate in the shared programming effort such as for example GO FAIR. The effort needed for oversight and steering new specification and implementation initiatives should not be underestimated, since the endeavour is organised across disciplines and across regions. The results of this work are regular updates of online documents/wiki with snapshots every 6 months to indicate agreements, shared tasks and successful delivery and the initiation of specification and implementation groups.

### 1.2.2. CAPACITY-BUILDING OBJECTIVES

**O5. Dissemination of results and progress.**

Disseminate the results of EUDOn to stakeholders. Exchange successful implementations that have shown the usefulness in concrete data-intensive science projects. Promote the capacity building actions for early career people (see below) to the many research infrastructures that have already indicated support to EUDOn and beyond. Relevant audiences at conferences and other events will be targeted by presentations, written materials, lightning talks, exhibits etc. in order to build an active and engaged community, to openly evaluate solutions and to achieve broad consensus.

**O6. Stimulate and contribute capacity for the specification and implementation work**

Encourage learning and engagement of Early Career Investigators (ECIs) through mobility in Short-Term Scientific Missions (STSMs). Use STSMs to build bridges between research infrastructures and e-Infrastructures and between academic and industrial research. These STSMs should offer opportunities to develop different types of skills such as participating in the definition of data-intensive science projects, in the specification of DO related concepts, in corresponding implementation work or in managerial type of actions. All these skills will be needed for an efficient and competitive data science ecosystem.

**O7. Organise training courses and hackathons in particular to attract young data scientists and transfer the evolving knowledge and technology from the early adopters**

Plan and implement a training program for ECIs and data managers from a wide range of scientific communities together with IEEE BDGMM [16] and other powerful organisations, in order to secure successful changes of research practice and promote convergence. EUDOn will use various channels such as webinars, tutorials, training courses with hands-on sessions and hackathons. Where possible EUDOn will organise these events together with specific scientific communities or domains, since this has proven to be most effective. Participants will more easily understand the terminology and more easily see the effect on the science they are familiar with.

## 1.3. PROGRESS BEYOND THE STATE-OF-THE-ART AND INNOVATION POTENTIAL

### 1.3.1. DESCRIPTION OF THE STATE-OF-THE-ART

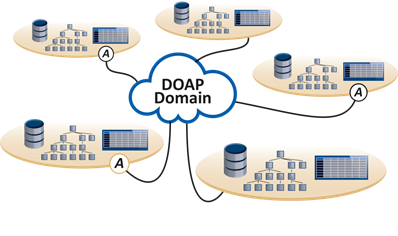
Current disadvantages and costs in the data domain are caused by the huge fragmentation creation of legacy data which will be difficult to integrate into an easily sharable and reusable data domain. The phase we are in can be described as creolisation with mixed suggestions and solutions at all levels.

Various initiatives listed below have been started from different stakeholders - some more top-down, others more bottom-up – which help to better understand the challenges and possible approaches and the FAIR principles, while broad discussions at OECD, G8+5, and RDA are milestones towards convergence. Yet we lack the core concept to turn such principles into practice, and none of the following initiatives include a broad network of practitioners to exchange views on DO as intended by the proposed Action.

* Stimulated by the availability of huge amounts of information in the web, the Semantic Web initiative has created many excellent solutions and standards to exploit this rich world. However, raw data first needs to be managed and transformed in order to enable semantic processing.
* The European Open Science Cloud is another huge initiative that wants to create a framework that facilitates the transformation towards more efficient and FAIR compliant data intensive science, but it does not yet discuss any implementation concept.
* The RDA has brought together communities that focus on scholarly communication aspects related to data and those that focus on the data fabric aspects, so as to define best practices and standards to overcome the inefficiencies. Like FORCE11, which also focuses on aspects of scholarly communication, RDA does specification work and not implementation.
* GO FAIR is a European initiative sponsored by some member states to foster implementation work towards FAIR compliance including initiatives with different foci. However, GO FAIR offers facilities to promote the work of the implementers, but not funds for a broad interaction platform.
* C2CAMP is a global initiative of some strong centers and research communities that want to start implementing and testing the DO concept and since it is an implementation network it is acting as part of the GO FAIR initiative, however, not having the financial capacity to establish a broad network

### 1.3.2. PROGRESS BEYOND THE STATE-OF-THE-ART

The Action is meant to fill this gap and offer a platform to facilitate and fund broad interactions. Due to its complementary role, members need to interact in appropriate ways with the above mentioned initiatives.

The main benefit for data science of the suggested DO-based approach can be seen in its universal way of binding essential information about data entities, either atomic or complex in form of collections. In this way the DO concept addresses aspects that are causing major inefficiencies described by the term "data wrangling". The DO model could be seen as defining a new commodity layer allowing growth again as J. Hendler, a well-known Semantic Web expert, put it recently at a US workshop [17]. G. Strawn, one of the Internet pioneers, expressed his convictions at a follow-up workshop [18] by indicating that a standardised and simple digital object access/interface protocol combined with persistent identifiers may have a potential and impact that can be compared to the one of TCP/IP some decades ago. This DO access protocol can be implemented by repositories either directly or indirectly by providing adaptors as is indicated by (A) in the diagram depending on the state of their data organisation.

The Action will not reinvent the notion of DO but discuss its implications for infrastructure building with all interested data scientists which are often collaborating with research infrastructures and e-Infrastructures and beyond. The Action wants to create the momentum that is needed to make big steps towards convergent specifications. Since this can only be achieved globally The Action is committed to collaborate closely with experts in other regions such as the US, China, South Africa, etc. It is now time to act, discuss, implement and test approaches in global synchronisation, in synchronisation with experts working on other strands such as semantic interoperability for example and by including knowledge about the already existing solutions in the different scientific disciplines.

### 1.3.3. INNOVATION IN TACKLING THE CHALLENGE

In computer science the concept of "objects", as in object-oriented programming, has led to an enormous improvement in software engineering since its encapsulation principles facilitate building complex systems. The concept of "objects" has also been realised in modern cloud systems which is why some experts call cloud stores "object stores" where all relevant information is encapsulated. This can be seen as a virtualization step since the user does not have to deal with different types of data organisations (file structures, clouds, sql-db, nosql-db, etc.), with the question where the metadata and data can be found, what kind of operations the special object type is supporting etc. However, cloud solutions are proprietary and pointers can only be resolved within the limitations of the cloud.

EUDOn wants to extend this concept to an open global space to facilitate data driven research and to make data reuse independent on the specific nature of local data organisations in the various repositories worldwide. Scientific users should be able to focus on the questions they are interested in: they want to select a certain digital object (in general a collection) identified by a PID and described by metadata (including data provenance) and carry out some operations on it or even engage agents to select a digital object based on profile descriptions and include it in workflow executions. The move towards DOs would allow us to build an interoperable domain where an agreed *Digital Object Access Protocol* links all kinds of different repositories some being connected via adaptors. This is very similar to the way the early Internet itself was built. Data scientists from various disciplines expressed their interest in DOs as follows:

* Faster selection of useful tools and services based on types of Digital Objects and their rich metadata will facilitate data re-use and their automated processing effectively.
* Scientists working in an increasingly complex data landscape want to work at higher levels of abstraction enabling scalability.
* Scientists and managers need wide scale data tracing and reproducibility to facilitate trust and verification, which are not achievable in a scalable way with today’s means.
* DOs abstract from concerns of location, storage method and community-dependent services.
* DOs will achieve a higher degree of interoperability and overcome fragmentation.
* Improved ways to automatically create scientific annotations and assertions to capture and exploit knowledge will enable machines to understand the object context and take appropriate action.
* DOs allow creators to associate mechanisms for controlled transactions and tracing of reuse which is crucial for sharing sensitive or commercial data and which finally is crucial for trust building.

## 1.4. ADDED VALUE OF NETWORKING

### 1.4.1. IN RELATION TO THE CHALLENGE

The concepts that will bring us to a change of data practices require large efforts in networking. The GEDE discussions about PIDs involving delegates and experts from 47 different European research infrastructure projects took about a year to result in a broad agreement about their relevance and their usage. This may give an indication of how much effort is required in consensus building. In iterations the commonalities of data science in different fields need to be extracted so that convergence can be achieved.

A European data infrastructure needs to be based on broad support from policy makers and data scientists being active in the research infrastructures and e-Infrastructures. A momentum towards a higher degree of interoperability and thus efficiency can only be achieved if this community will become actively engaged. At the specification level, RDA has been established and can be seen as successful. At the implementation level GO FAIR has been initiated and C2CAMP, putting the digital objects in its centre, is participating in this initiative, but neither GO FAIR nor C2CAMP offer funds for active participation from a broad group of people. The voluntary GEDE work showed that a voluntary basis alone is not sufficient to organise meetings, to support travels, to organise training courses etc.

EUDOn can build on the GEDE experience and the required trust in fair interactions has been established. But it will add funds to support networking activities and to facilitate participation of a broad group of engaged European experts. In doing so, it will also foster the work of initiatives such as GO FAIR and C2CAMP.

### 1.4.2. IN RELATION TO EXISTING EFFORTS AT EUROPEAN AND/OR INTERNATIONAL LEVEL

EUDOn fits well into the current European and global landscape in so far as it

* adds an innovative and operative dimension to the FAIR principles because the DO model is not only FAIR compliant but can also directly facilitate data intensive science;
* could help to realise the EOSC which has been focusing on important aspects of governance, business models, etc., but currently needs a common concept that facilitates data intensive science;
* adds an implementation dimension to the global RDA specification work by transforming RDA endorsed outputs into a larger testbed, so that some work in RDA is given more momentum and the speed and density of specification work will increase;
* adds a global network focussing on a specific approach to the GO FAIR implementation networks working on related challenges such as semantic interoperability;
* integrates experts from various data science communities in Europe, linking also to experts from other countries such as the US, China, Australia, South Africa and Russia.

EUDOn will enable European experts to take a leading position and attract young experts to participate in their work. It will allow EU data scientists to not only invite experts from abroad, but also to steer the global discussion, the specification and implementation work. It will allow EU data scientists to participate in corresponding discussions going on in other countries. It will help selected experts to contribute to the ongoing specification work in RDA and the implementation work in GO FAIR.

In summary, European data scientists looking for innovation in data practices will get an enormous impulse and take over a leading position. Without EUDOn, Europe risks falling behind countries such as the US and China which are already planning large investments in work related with DOs.

# 2. IMPACT

## 2.1. EXPECTED IMPACT

### 2.1.1. SHORT-TERM AND LONG-TERM SCIENTIFIC, TECHNOLOGICAL, AND/OR SOCIOECONOMIC IMPACTS

The impact of a DO based global data infrastructure ecosystem on scientific progress will be immense:

* FAIR data will be put into practice and data sharing/trading will become the accepted practice;
* data integration from different silos will be much more efficient;
* data intensive research in all disciplines will be facilitated so that more scientific questions can be addressed;
* many more researchers will be able to participate in data intensive research since efforts and costs will be reduced;
* sensitive data can be integrated into the shared data domain since it will be possible to restrict operations, to make specific "smart contracts" regulating re-usage, to improve tracing of data usage;
* automated data processing will be enabled – the only practicable way to go, given the huge amounts of data being generated in the coming decade from smart sensors, the digital marketplace, etc.

Comparable to the invention of the Internet, we cannot foresee which types of brilliant applications will be invented in a few years. No one could foresee for example the emergence of the Web application which changed the world when TCP/IP was born. However, we can imagine a situation where new data becomes available from a trustworthy source, where crawlers tuned by scientists detect the relevance of this new data for a given task, where smart contracts are available indicating the type of operations allowed on the data and where the data will immediately be included in useful processing tasks. It would free the researcher to spend time on finding hidden patterns, on checking the trustworthiness and reliability of the results, etc.

In summary, there will be considerable short-term benefits for the researchers by adopting the DO concept since it will reduce the wasted time on data wrangling. In addition, there will be even larger long-term impacts for the data volumes produced by the billions of future smart devices.

## 2.2. MEASURES TO MAXIMISE IMPACT

### 2.2.1. PLAN FOR INVOLVING THE MOST RELEVANT STAKEHOLDERS

**Data Scientists:** The most important stakeholders are data scientists, including those who are experts in applying state of the art methods such as machine learning and those who are experts in formats, metadata, etc. Many are involved in research data infrastructures; they know the scientific challenges and have a deep understanding of the roadblocks. With GEDE, a collaboration between such data scientists has been established which can serve as a basis, but this collaboration needs to be extended.

**E-Infrastructure-Providers**: Computing and data centres are currently revising their services to data scientists. They need to facilitate data workflows and to offer a variety of services on data. Computing and data cannot be separated anymore in most data intensive applications. Members from PRACE and EUDAT will be included in the Action.

**Computer Scientists**: Large computer science departments work on specific methods and also create cross-disciplinary tools. It is important for success to bind them into the discussions as well to integrate their deep methodical knowledge, to adapt their software where required and help educating a new generation of experts.

**Global Service Providers:** Another relevant group are those who offer global services such as the DONA Foundation, the DOI Foundation, the ePIC Foundation, companies offering cloud services, etc. since these are providing crucial services such as for the functioning of a global DO domain that is widely based on stable and persistent services for PIDs or by providing advanced data services.

**Global Initiatives**: A close collaboration with initiatives such as RDA, W3C, IEEE, CODATA and others is important to consider different aspects related with data.

**Policy Makers:** The Action will have a close collaboration with policy makers organised in scientific organisations and with funders at European and member state level. Here the commitment to contribute to EOSC and to the GO FAIR initiative will help EUDOn to reach out to policy makers.

**Industry**: An outreach to industry will be important because a change of practices will only be successful if industry will take up the new methods and build appropriate tools. We will reach out to start-ups and SMEs which agree with the DO model and which already started to work out innovative technology models. We will participate in industry forums and invite industry experts to specific meetings.

### 2.2.2. DISSEMINATION AND/OR EXPLOITATION PLAN

The WG on Dissemination will be responsible for producing a Dissemination and Exploitation Plan (to be ratified by the Action MC) that will guide the EUDOn outreach to various stakeholder groups described above. Data intensive ESFRI projects and other mature e-infrastructures will be included as first level targets, because they will assist in reaching out to data scientists. This has to be complemented by two directions: 1) more disciplines need to be approached and 2) more data scientists within the scientific domain of each participating ESFRI project need to be addressed. They are the drivers of change and need to be actively involved in the interactions which is not trivial given the usual workload. The ESFRI and national representatives have to take an important role in EUDOn as ambassadors and as facilitators to get DO related issues on the agendas of discipline and national meetings. Training courses and webinars driven by the scientific use cases will also help conveying the main messages and showing the implications of applying the concept.

This will be complemented by strategies to involve the young generation. Here in particular hackathons and web-based competitions turned out to be excellent mechanisms to reach out and attract young people. For this work partners have to be looked for such as IEEE BDGMM, RDA and the ESFRI projects since the effort to prepare such actions is high. We can build on some experiences made by RDA Europe and IEEE BDGMM for example. Industry will be contacted using the started collaborations by RDA Europe, using also the collaborations with the IoT Forum and various national meetings. Industry can appear in different flavours (IT service providers, sector specific service providers). In particular, innovative start-ups and SMEs working on challenging new ideas will be the target. Policy makers will be addressed e.g. by regular flyers with essential messages. Through an annual event, we will try to attract many of the actors mentioned above and bring them together using attractive formats such as interactive panels.

## 2.3. POTENTIAL FOR INNOVATION VERSUS RISK LEVEL

### 2.3.1. POTENTIAL FOR SCIENTIFIC, TECHNOLOGICAL AND/OR SOCIOECONOMIC INNOVATION BREAKTHROUGHS

Every approach that reduces the huge fragmentation and thus the inefficiencies in the data domain globally will lead to “revolutionary changes” that will not only influence science, but also industry and society as a whole. Some argue that the urgent changes needed will probably involve disruptive steps and they compare the expected transformations with those after the invention for example of the Internet.

A recent paper [8] points at global agreements on important but comparatively simple specifications that ultimately generated a boost of new services and businesses impacting all aspects of society. The question is which simple specifications will have the potential to be accepted by many as a new platform in the data domain to invest and adapt current strategies. We believe that DOs are such a fundamental enabling technology.

In order to maximize the chance of success of this promising but potentially disruptive concept, we need to agree on a roadmap and build testbeds with more investments by global actors as C2CAMP is starting it in collaboration with initiatives such as GO FAIR. Such initiatives will only succeed if we are creating networks of stakeholders around it that drive discussions based on scientific use cases and if we succeed in attracting the young generation.