publications

MDPI

Article in Special Issue

FAIR Digital Objects for science: From data pieces to actionable knowledge units

Peter Wittenburg [1], Koenraad De Smedt [2] and Dimitris Koureas [3,*]

[1] Max Planck Computing and Data Facility; peter.wittenburg@mpcdf.mpg.de
[2] University of Bergen; desmedt@uib.no
[3] Naturalis Biodiversity Center; dimitris.koureas@naturalis.nl
* Correspondence: peter.wittenburg@mpcdf.mpg.de; Tel.: +49-151-41858784 (P.W.)

Received: date; Accepted: date; Published: date

Abstract: Data science is facing the following major challenges: (1) develop scaled cross-disciplinary capabilities, (2) improve our capabilities to deal with the increasing data volumes and their inherent complexity, (3) build tools that help creating trust, (4) implement new mechanisms to efficiently operate in the domain of scientific assertions, (5) turn data into actionable knowledge units and (6) harmonize standards to overcome their existing proliferation. The new concept of FAIR Digital Objects (DO) is presented, indicating how the challenges could be overcome. The FDO concept goes back to early work on Digital Objects by Internet pioneers which was taken up by various groups within the Research Data Alliance. Based on a variety of use cases, basic components of a DO Architecture have been developed. Recent discussions revealed that semantic explicitness needs to be added to fully meet the FAIR principles, thus allowing us to speak about FAIR Digital Objects. A survey of use cases within a collaboration of ESFRI initiatives indicated the growing interest of research communities in the FDO concept. We conclude that the FDO concept has the potential to act as the interoperable federative core of an hyperinfrastructure initiative such as EOSC.

Keywords: digital objects; data infrastructure; research infrastructure; data management; data science; grand challenges; FAIR data; open science; EOSC

1. Introduction

From about the turn of the millennium it has become apparent that the rapid acceleration in the production of research data has not been matched by an equivalent acceleration in our access to all that data. Around the same time, the realization dawned that data science is at the core of our ability to address many global challenges in research as well as in society, e.g. the construction of climate models, the monitoring of threatened species, the detection of fake news, etc. With "data" we do not just refer to what is called published data, but to the huge mass of data being created in data labs and being maintained in many repositories with various thematic foci. It may be expected that more than 90% of the data that has a value in scientific processing should be reusable and be referenced in a stable way. This is depicted in the middle layer of the diagram in Figure 1, taken from EUDAT.[1] In contrast to temporary data we call this the registered data domain which is being used in analytics tasks in various contexts. Cross-silo data science attempting to include all relevant data is known to be extremely inefficient and thus expensive, hampering progress in tackling the global challenges.

[1] https://eudat.eu/

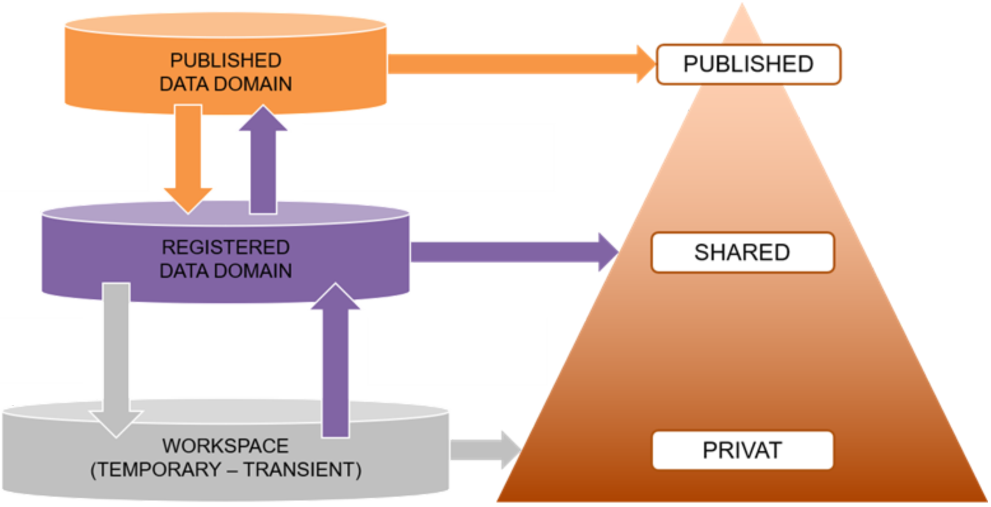

Figure 1. Layers of data (from EUDAT) with some data being published, most data being generated in the labs and reused within collaborations, stored in repositories and to be assigned PIDs.

In 2006 the first ESFRI roadmap[2] was published, which provided an impetus for a large variety of disciplines build research infrastructures providing advanced facilities for conducting research and fostering innovation in their fields. For the first time, distributed databases were accepted as research facilities comparable to large physical infrastructures. The construction of domain-based infrastructures sparked a harmonisation of standards, an exchange of tools and methods, the establishement of unified data catalogues, the development of trustworthy repositories and interfaces for making data accessible. Despite large improvements within domains, data integration and reuse *across* lab and discipline boundaries are however still highly inefficient. The European eInfrastructures[3] were not helpful in this respect, since they have been constructing specific technologies (grids, clouds, portals) while being mainly interested in offering core services such as compute cycles, storage and networking capacity.

Consequently, the research data infrastructures are not sufficiently interconnected and interoperable. It is increasingly understood that the sheer volume of data and its inherent complexity makes manual search, evaluation, access and processing of datasets by individual researchers no longer feasible. In addition, we still lack means to guarantee data accessibility and reusability over time. New strategies were required to improve practices. First, an initiative towards open science and data was started to convince researchers to make research data and other results available to their colleagues. Second, the convergence on principles for the creation, management and dissemination of data, canonicalized in terms of the FAIR principles [1], was meant to inform researchers of how to deal with data.

In this context, the European Open Science Cloud (EOSC)[4] was initiated as a next step to overcome the inefficiency hurdles and to implement the FAIR principles. In this respect, EOSC must be seen as a great opportunity for research, in particular for interdisciplinary data science. While a number of exploratory cluster projects in the EOSC context have done interesting work, the core concept of EOSC is not well defined and remains elusive. Many research infrastructure experts see the EOSC promise as a federated shell to combine the data and services created by the various disciplines. Basically, the research infrastructures expect to make their data, tools, services and repositories visible via the EOSC expecting that others might benefit from their work and knowledge, thereby promoting interdisciplinary data use. Accordingly, some strategists are focusing on building an EOSC Portal[5] which, however, may fail as long as more fundamental issues in managing the billions of data sets and thousands of tools are not tackled.

[2] https://www.esfri.eu/
[3] https://ec.europa.eu/digital-single-market/en/policies/einfrastructure
[4] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
[5] https://www.eosc-portal.eu/

Currently EOSC does not address the fundamental questions of how to improve interoperability in the federating core, how to move to automated data discovery, access and aggregation of data across repositories and other sources, as necessary to deal with the sheer volume and heterogeneity of the available data, and how to address the persistence requirement enabling data and references to survive technological changes that can be expected in the coming decades.

In this paper, we present a way forward by focusing on a core model for interoperable data, instead of on tools that will come and go with frequent technology changes. We propose to take the work on persistent identifiers a major step forward by encapsulating sufficient information about a dataset into a FAIR Digital Object (FDO). This new approach will enable automated systems to interact with data in a reliable way over long periods of time. Previously this was called Digital Object (DO) for short, as discussed at various meetings.[6]

In the remainder of this paper we will explain the FDO concept and we will show how it meets the FAIR principles. We will discuss its potential, in particular for science, and we will describe how different areas of science are moving into this direction. Thus we envisage the evolution of the EOSC into a highly interoperable Global Digital Object Domain guided by strict FAIR compliant guidelines allowing different implementations and boosting innovation. The FDO architecture has the potential to be proliferated into science, industry and public services since the challenges described hold across sectors.

2. Principal Challenges for Data-intensive Science

2.1. Addressing grand challenges is linked to scaled cross-disciplinary capabilities

Humanity continues to face ‘grand’ challenges towards its sustainable development. Problems such as climate change, biodiversity loss, fake news and social inequality require concerted efforts across the realms of science, policy and technology. The need for a globally coordinated response to those challenges has been already well documented, with important policy goals being articulated by international organisations.[7] The scope and complexity of these problems present unprecedented barriers, and as such, require novel approaches in order to produce effective results. Arguably, the solution space for these challenges would include an interconnected multi-actor and multi-level ecosystem that supports cross-disciplinary socio-technical interfaces and enable meaningful and scalable integration and interpretation of evidence across different realms of science.

Our ability to combine and process complex information across scientific disciplines and draw conclusions that robustly inform policy making is predicated on the capacity of different communities of practice to navigate, understand and use increasingly complex data from across many fields of science. Such practices go above and beyond the traditional ways through which scientific communities operate, as they require them to transcend disciplinary boundaries defined by traditional epistemic objects. In this endeavor, the capacity to find, access, understand and re-use data across scientific domains is of pivotal importance.

2.2. Drowning in data?

Today, we experience the digital transformation of most sectors of our societies. Mass-scale digitisation efforts, including the use of high-throughput monitoring and analytical devices, produce more data today than ever, and do so at ever lower cost. A prominent example can be found in the field of genomic studies. The National Center for Biotechnology Information (NCBI) runs one of the largest genomic sequences open repositories, doubling in size (in terms of the number of bases) on average every 18 months.[8] Similarly, the Global Biodiversity Information Facility publicly serves more than 1.3 billion organism occurrence records today, having doubled its size in the last five years.[9] Similar examples of ever-increasing data availability can be found in other fields of science.

[6] https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/FDO-Contributions

[7] https://en.unesco.org/sdgs

[8] https://www.ncbi.nlm.nih.gov/genbank/statistics/

In the field of material sciences, the results of millions of simulations and experiments are currently being aggregated to find hidden patterns in the data that may help to better categorize compound materials along reduced descriptors.[10] In the domain of research on languages, large collections of textual and audiovisual materials, are being collected. The CLARIN Virtual Language Observatory has more than one million datasets about languages, of various granularities.[11] These datasets, and technologies to process them, are transforming scholarship in the humanities.

The availability of large amounts of data, as promoted in part by Open Data policies, has great potential for the advancement of knowledge. One of the key questions, however, is whether the increasing volume of generated research data has further enabled scientific communities to produce meaningful knowledge from the data and, in particular, to engage in large-scale cross-disciplinary research as required for finding solutions to some of our grand challenges. Although there is an increase in interdisciplinary data research (IDR), this increase is not on par with the much larger increase of data potentially available across disciplines. The level to which the scientific data are open and FAIR significantly affects the ability of researchers to make use of diverse data sources in their scientific practice. But even the properties of openness and FAIRness, despite being prerequisites for data discoverability and re-use, are in themselves not sufficient to support the transformation in community practices towards IDR. The fact that much data currently has a short shelf life is not so much because it ceases to be relevant, but because sufficient knowledge of the context in which the data originated has gotten lost and it is unclear how the data can be used in evolving contexts.

There is yet another cultural aspect in science that prevents cross-disciplinary, even cross project data science. C. Borgmann pointed to a paradox when she applied an expression from S.T. Coleridge “Water, water, everywhere, nor any drop to drink” to data [2]. There is already much data out there, but we seem to have a skills gap to make use of this richness even in cases where context information and technical circumstances would make reuse possible. Many researchers still prefer to refer to research results as documented in written papers since their reuse is only dependent on proper language understanding. Data science as it is currently being done, requires mastering new skills many researchers do not have and requires various transformations which include the risk of reducing transparency.

Further support to the assertion that, despite the increasing cross-disciplinary data availability, IDR is still not a comfort zone for scientific communities, can be drawn from the relatively lower success rates of IDR project proposals in relation to non-IDR proposals. Despite unprecedented data generation capabilities, our collective capacity to produce actionable knowledge out of data seems not to be on par with the increase in its volume and diversity. In this paper we briefly discuss some of the current limitations that potentially impede the amplification of actionable knowledge production from the vast volume of newly produced and available data, and discuss the scientific value of digital objects, as a unified data organisation model.

2.3. Interpreting scientific evidence in a trusted context

The global scientific domain is organised in distinct communities of practice. These communities consist of scientists who not only focus on a certain field, but also operate within a common socio-scientific context. They share mechanisms and processes through which they construct knowledge and attribute credit. Scientific outputs deriving from their members are interpreted, usually also through an implicit understanding of the context. This context is pivotal to the ability of a practitioner to evaluate the fitness-for-purpose of the information consumed. Traditionally, this information is circulated within the community’s trusted communication channels and interpreted based on the community-specific criteria for quality and fitness.

As data travels across domain-agnostic repositories and aggregators and becomes available to a variety of scientific communities, it is gradually uncoupled from the original context in which the data was generated. It effectively loses the contextual information, which is essential for

[9] https://www.gbif.org/news/6s1uLwFkwGq0oSHGNnzfUE/doubling-up-two-year-ebird-refresh-adds-200-million-more-observations

[10] https://nomad-coe.eu/

[11] http://vlo.clarin.eu

communities to understand and evaluate the quality and fitness for purpose of the data. This, in turn, can reduce the re-usability of the data, especially across scientific disciplines. The greater the conceptual and methodological distances between the data-producing communities and the data-consuming communities are, the less contextual information can be automatically induced, and the lower the ability is for the consuming community to trust the available data.

Our personal experience has shown some of the origins of this problem. In a research institution, one of us observed a difference in behavior between researchers who made observations in the field and those who designed lab experiments. While it was relatively easy to capture the context in which observations were made, it was much more difficult to capture the precise intentions, restrictions and settings of experiments. Experiments are typically part of a set of very specific (control) experiments being part of the context that needs to be understood to reuse the results. As a consequence, some experimentalists doubted that their experimental data could be repurposed and thus stopped to create the necessary metadata for others to interpret, and therefore trust, their results.

A special case of reuse, relevant in the area of experimental research, is replicability and reproducibility of experimental results. Traditionally, researchers expect that the methodology is captured as precisely as possible in the scientific paper. In data science, the methodology description can become highly complex, prompting the need for ready, sealed and fingerprinted packages of data and algorithms to enable verification of results.

2.4. Domain of Reasoning

Research already has quite a long experience with digital data in specific fields. Complex processes as they happen in fusion reactors, for example, could only be understood by including a multitude of sensors, which already in the 1960s produced a lot of data; this required new data management approaches such as put forward by F. Hertweck by designing the AMOS system.[12] However, the rapid proliferation of data in many different fields and its inherent complexity due to the huge data variety and the many relationships between them needs another change in the data science landscape. Soon it will no longer be feasible for researchers to find, extract, evaluate and process digital data manually. Instead, successful data science will be dependent on highly automated methods for identifying and extracting datasets from repositories, aggregating selected data, and analyzing the combined data for given purposes. This will not only require the application of new types of algorithms summarised under the heading “AI”, but these algorithms will also require that data be stored and disseminated in more robust and informative ways, thus allowing automatic systems to make sense of the huge number of individual pieces of knowledge results with varying provenance, size, certainty, context, license, etc.

Some scientific communities are already experimenting with new forms of knowledge representations such as nano-publications which are basically assertions described in some formal semantic language such as RDF, augmented by sufficient metadata. These nano-publications can be subject to smart statistical processing. Surveys in the biomedical area indicate that there are about 10^{14} such assertions with an enormous increase every year. Eliminating duplicate findings still amounts to 10^{11} canonical assertions and further processing yields about 10^{6} so-called *knowlets* which can be seen as core concepts in this endless space of assertions related with sets of different findings [3]. Finally, this highly reduced space of knowlets can be used to draw conclusions, for example, for proper health treatments.

What we realise in a variety of scientific disciplines is that we create complex relational buildings between the basic findings, layers of derived data and knowledge sources of different kinds (ontologies, etc.). Scientific disciplines will make huge efforts in creating and curating these relational buildings in the coming decades and it is obvious that they will form an essential part of our scientific knowledge that must be preserved. Traditional form of publications will lose relevance compared to these relational buildings. Therefore, new methods are required to preserve this new digital scientific memory to not fall into a dark digital age.

[12] https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_3055671

2.5. Advancing data to actionable knowledge units

The ability of a scientist to verify the relevance, provenance, completeness and fitness is an essential process in the scientific pipeline. As such, it is essential that the practice of data sharing includes the preservation and reconstruction of the contextual information in which data was generated. Such contextual information needs to be permanently coupled to the dataset, irrespective of the mode in which the data is shared. This information goes beyond the typical domain-agnostic metadata such as introduced by Dublin-Core[13] and requires rich metadata sets as they have been developed in many scientific disciplines already to not only enable discoverability but also access in a machine-readable way. We need data sets that are fully actionable and comprehensible knowledge units, to be shared across digital environments and re-used across scientific disciplines in both human and machine actionable ways, without degradation of its rich contextual information. These actionable knowledge units need to be stable across cyberspace and time.

The persistent encapsulation of data and different types of metadata into such a knowledge unit should significantly improve the level of trust with which it can be understood across disciplines and modes of sharing. For such a transformation be achieved, it is essential to employ a set of robust conceptual and technological models, which enable sharing and re-use of data-in-context.

2.6 Tool Proliferation and Fundamental Decisions

Another challenge data scientists are confronted with is the proliferation of tools and standards they need to chose to tackle new questions. The problem is not so much the heterogeneity they are faced with, but the knowledge that decisions for tools seem to be fundamental with the risk to be locked-in and the fact that they need to make decisions about aspects they are in principle not interested in. Figure 2 shows a diagram by L. Lannom indicating the important role of abstraction in this domain of increasingly complex and heterogeneous landscape of technology [4]. The user should only be confronted with FAIR Digital Objects which are represented by metadata and PIDs where the PID details in general will be hidden to the user. Metadata, data and PIDs are serviced by a landscape of registries and repositories that are connected with the help of the unifying DO Interface Protocol and thus build the federative core of a research infrastructure.

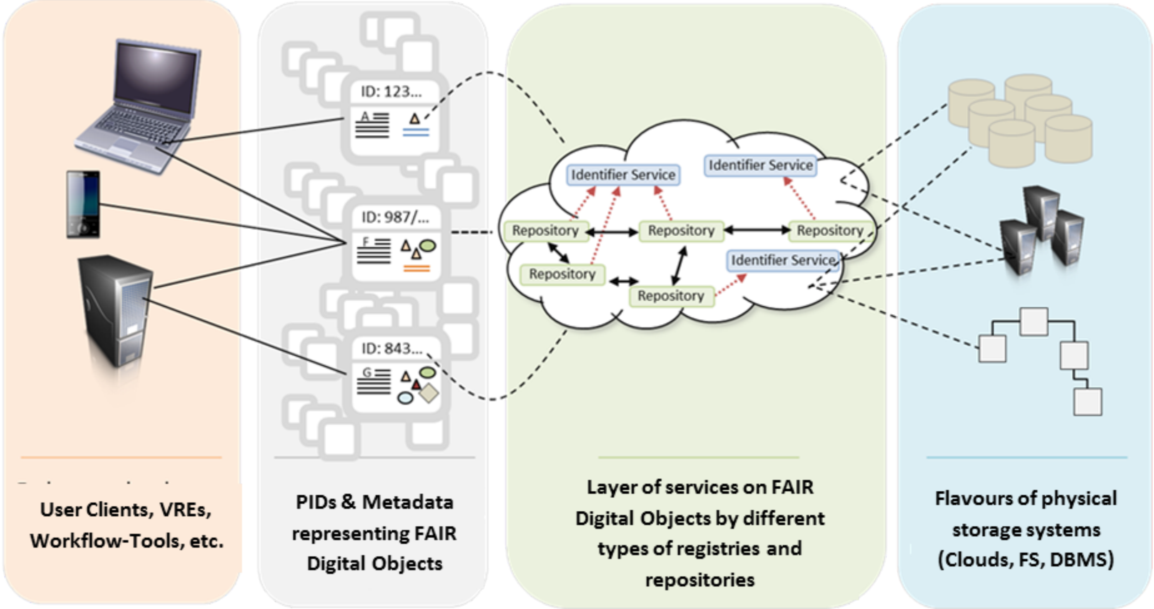
ID: 123...
ID: 987/...
ID: 843...
Identifier Service
Identifier Service
Repository
Repository
Repository
Repository
Identifier Service
Repository
User Clients, VREs, Workflow-Tools, etc.
PIDs & Metadata representing FAIR Digital Objects
Layer of services on FAIR Digital Objects by different types of registries and repositories
Flavours of physical storage systems (Clouds, FS, DBMS)

Figure 2. Layers of abstraction in the data domain (from L. Lannom).

Technology will continue to develop rapidly allowing to generate continuously new user clients, Virtual Research Environments, collection builders, workflow frameworks etc. User communities will make their choices orienting on efficiency, but when these tools apply the FDO concept and use

[13] https://dublincore.org/

standardised protocols to interact with the service layer, there will be no danger of a technology lock-in, i.e. when better client technology would become available researchers could switch without risking their huge investments in creating the digital knowledge domain.

Currently, research communities are offered funding possibilities to adapt to flavours of cloud systems and funds are spent on defining a common interface between cloud systems. However, the researcher is not interested whether data will be stored in clouds, high performance file systems or in database management systems. Instead they are interested in parameters such as capacity, access speed, etc. Also with respect to storage systems we can expect new technological developments driven for example by quantum computing. These technological changes should be transparent to the user and implementers need to take care that the digital knowledge domain remains stable. The service layer is a virtualisation step hiding the exact functioning of the underlying technology.

3. FAIR Digital Objects

3.1. The Scientific View on FAIR Digital Objects

To meet the scientific needs and expectations described in the previous chapter, we envisage a stable and actionable atomic unit of information as indicated in the diagram which we call a FAIR Digital Object (FDO) that exists in cyberspace over long times. In this section we will give an introductory description of the FDO concept without going into technical realisation details. From the perspective of a data scientist, an FDO is a stable actionable unit that bundles sufficient information to allow reliable interpretation and processing of the data contained in it.

In order to reach this goal, FDOs have several general properties. FDOs enable abstraction, i.e. it does not matter at management level whether the FDO content is about data, metadata, software, assertions, knowlets, etc. An FDO binds its encoded content and its metadata of different kinds such as descriptive, scientific, system, access rights, transactions, etc. so that all relevant information to access and process the FDOs content is available at all times dependent on their type, as schematically shown in Figure 3. Operations can also be encapsulated in FDOs, which has shown to be a powerful concept in designing complex systems. While the PIDs and metadata of FDOs are open, the access their bit sequences can be protected, for example, in case of sensitive or personal data.

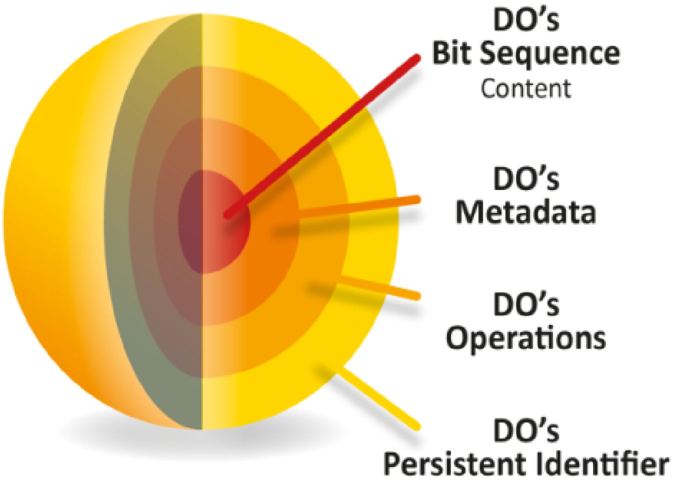

Figure 3. Layers of information associated with an FDO.

From a researcher's point of view point, we can imagine the following ideal scenario for data: A sensor produces data, associates metadata with it and places both via specific mechanisms in the care of trustworthy repositories. These repositories analyze the metadata and may decide to host and manage the new data as a FDO. Doing so implies that the data and metadata will be bundled in an FDO to which assign a type and a PID will be assigned. Also, the repository may extend the metadata based on known contexts (for example, by turning access permission assertions into Access Control Lists). Then repositories may propagate the FDO to other agents on the Internet which compare the metadata of the newly created FDO with their own profile and decide whether the new data set is of interest. If so, they will seek to get access to the content, based on licensing and legal constraints associated with the FDO. All actors in cyberspace dealing with FDOs are connected using a unifying protocol which guarantees interoperability. For this purpose we propose the DO Interface Protocol (DOIP) of which the version 2 specification is now published.[14]

The concept of FAIR Digital Objects which will enable this revolutionary scentario goes back to papers in the 1990s by Kahn & Wilensky [5,6]. Earlier, when Kahn designed the basic principles of the Internet where scientifically meaningless datagrams were being routed, it was already understood that the objects to be exchanged between senders and consumers must be assigned some meaning. This evolution is schematically shown in Figure 4. Shortly after, the design of the World Wide Web by Berners-Lee[15] represented a first step in this direction. However, despite its benefits, the Web remains an ephemeral technology, of which the high numbers of link rot are a symptom. Therefore Cerf, a colleague of Kahn in the development of TCP/IP, stated that we risk sinking into a "dark digital age".[16]

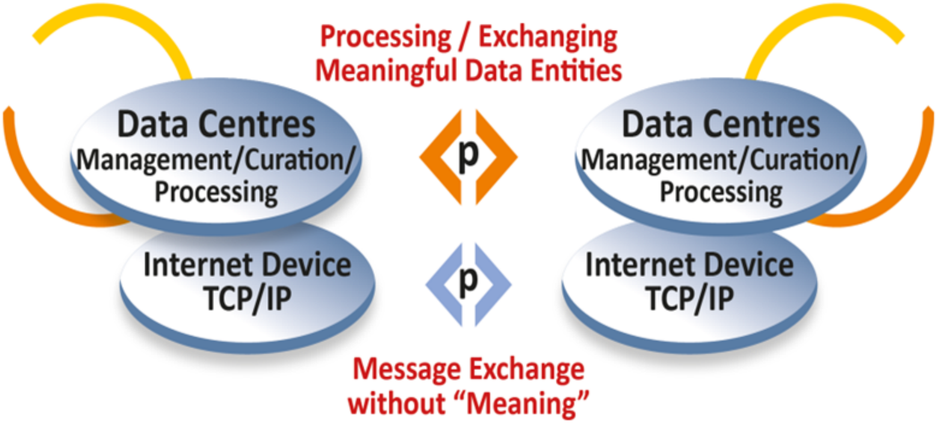

Figure 4. Evolution from message exchange without meaning to the exchange of meaningful data entities.

The introduction of stable atomic FDOs based on persistent identifiers that exist independently of all technological changes which can be expected in the coming decades will overcome the ephemeral nature of the Web. Given such long-term perspectives researchers and other stakeholders will be ready to finally rely on digital technologies as an alternative to the paper-based scientific memory and invest time, effort and huge funds to create the "digital scientific memory". This will provide the following advantages for research and development.

Scaled Cross-Disciplinary Capabilities: FDOs are a way to create the interconnected multi-actor and multi-level ecosystem as requested in 2.1, since a protocol can be envisaged that talks "FDO" to all actors in this interoperable global domain of digital objects. This will allow us to invest in a new set of tools supporting cross-disciplinary at a much more efficient level compared to the current data practices.

Data Made Accessible: In section 2.2 the gap between the amount of data being created and our capability to make use of it is being described. Different reasons such as lack of skills and little recognition for cross-disciplinary work cannot not be addressed directly by FDOs. However, one frequent reason, the lack of contextual information, will partly be addressed by FDOs, since their atomic nature will bind contextual information in a stable and persistent way. It is up to the researchers and improved workflow supporting tools to create sufficiently rich contextual information.

Interpreting Scientific Evidence in a Trusted Context: Trust in scientific results has different dimensions as described in 2.3. FDOs are addressing directly some of these dimensions, since (a) contextual and fingerprint information can be associated with digital objects at different steps of their

[14] https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf
[15] https://en.wikipedia.org/wiki/Tim_Berners-Lee
[16] https://www.bbc.com/news/science-environment-31450389

life-time improving their detailed documentation and (b) privacy information can be associated with each digital object in a way that cannot be manipulated. FDOs cannot directly influence the effects related with the increasing distance between creator and consumer communities. In this respect they only have an indirect influence by binding rich metadata persistently to the encoded content. FDOs enable the data sharing in context, support provenance, assessing fitness for purpose – thus building trust.

Domain of Reasoning: In 2.4 we outline the evolving complex domain of knowledge in and across all scientific domains which increasingly often requires automatic processing to be able to draw conclusions. FDOs with their stable atomic nature being also actionable units are excellent to structure knowledge and to capture complex relationships over long-time periods. It makes sense to implement and preserve complex knowledge structures using FDOs as building blocks for the evolving digital scientific memory.

Advancing Data to Actionable Knowledge Units: Relevant data objects need to travel through cyberspace and time in an encapsulated way, so that even after decades and in spite of changing technology and changing actors, data will remain available as fully actionable units, binding all relevant information for access and reuse. FDOs have exactly this capability.

Tool Proliferation and Fundamental Decisions: The FDO concept achieves abstraction that hides technological details to the researcher, thus preventing technological lock-in and allowing technological innovation without putting the evolving Digital Knowledge Domain at risk. Virtualised registry and repository services can be connected into a federated core using unified DO protocols and offering understandable client interaction at the service layer.

3.2. A Technical View on FDO

Although FDOs represent a quantity leap in data management, the concept can be implemented on top of the existing Internet protocols, as already suggested in earlier proposals [5,6]. After the World Wide Web established HTML resources as referentiable and exchangeable digital entities, this technology started to open possibilities for the Semantic Web[17] and the Linked Data Platform.[18] However, as was concluded at a recent workshop, the data community was not convinced [7]. Subsequently, the term DO found its revival in two developments: (a) the Data Foundation & Terminology Group of RDA[19] after having extracted a core data model from many scientific use cases and (b) the design of large cloud systems, also called "object stores". In the end, the definitions of DO made by the RDA DFT Core Model [8] did not differ so much from the definitions used by Kahn & Wilensky. For more details on the term and concept of DO see [7].

Two diagrams were essential in explaining the nature of DOs in the DFT Core Model. The diagram in Figure 5 indicates the simplicity of the DO concept. A DO has some content encoded as a structured bit-sequence and stored in some repositories. It has also assigned a globally unique, persistent and resolvable identifier (PID) and different types of metadata (descriptive, scientific, system, provenance, rights, etc.). Metadata descriptions themselves are DOs. DOs can be aggregated to collections which are also DOs with a content consisting of the references to its components. This simple definition abstracts away from the nature of the content of a DO and covers the whole domain of digital entities including digital representations of physical objects. The detailed metadata specifications imply a type specification, i.e. a DO has a type that allows to associate operations with it. DOs therefore can be compared with "books" standing for the work of an author and not the printed copy in a book shelf. It has an ISBN number which can be resolved to some information and it is associated with library cards containing some metadata.

[17] https://de.wikipedia.org/wiki/Semantic_Web
[18] https://www.w3.org/TR/ldp/
[19] https://www.rd-alliance.org/group/data-foundation-and-terminology-wg.html

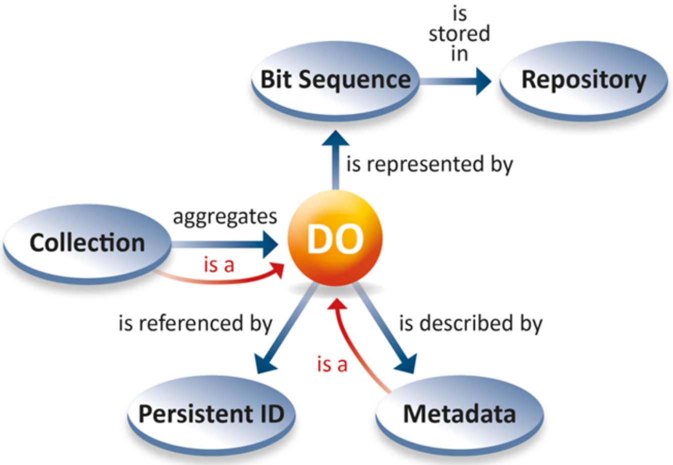

Figure 5. Relations between information associated with a FDO.

The diagram in Figure 6 indicates the crucial role of the PIDs in binding different information entities that belong to a DO, making the PID the anchor point for accessing and reusing the DO's content. Assuming that the PID indeed is persistent, which is based on a cultural agreement, it makes sense to bind essential information into the PID record which is the result of the PID's resolution. This can be the paths to access the bit sequence, the PID of the metadata, the PID to the access permission specifications, a pointer to a blockchain entry storing the transactions, a checksum to proof that the DO is indeed the one looked for, the type specification, etc. The RDA Kernel Information group[20] defined a first core set of attributes which are of relevance for scientific disciplines and registered them in a public type registry. The nature of type registries has been specified by another RDA group.[21]

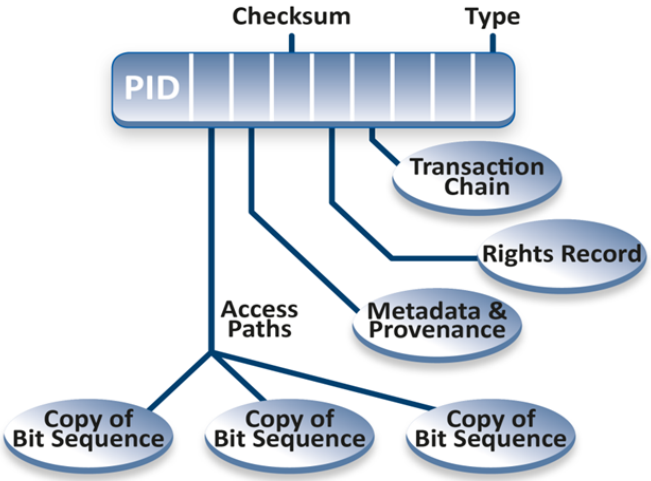

Figure 6. Components of a FDO.

This definition of a DO already implements some of the FAIR principles [9]. Intensive discussions pushed forward in the interaction between RDA and GOFAIR experts during the last year revealed that additional specifications were required to make DOs fully FAIR compliant. Early papers coined the term FAIR Digital Object [10,11], but it was L. Bonino who recently pointed clearly at the missing parts [12]. It has become obvious that the specifications of the DFT Core Model were not sufficient to guarantee machine actionability with respect to all FAIR principles. The RDA Kernel group defined kernel attributes and registered them, but the DO model did not make

[20] https://www.rd-alliance.org/group/pid-kernel-information-wg/wiki/pid-kernel-information-guiding-principles

[21] https://www.rd-alliance.org/group/data-type-registries-wg/outcomes/data-type-registries

any statements about their usage. The FAIR Digital Object (FDO) needs to be specific on three aspects:

1. The FDO model requires to define PID attributes and register them in a trustworthy type registry or a more complex type ontology and trustworthy repositories are requested to use these defined attributes to achieve interoperability and machine actionability.
2. The FDO model requires to make metadata descriptions machine actionable, i.e. to define its semantic categories and register them. A moderate requirement could be to declare at least those metadata categories that are necessary for proper management such as where to find the PIDs of relevant information entities.
3. The FDO model requires to make the construction of collections machine actionable enabling machines to parse collection descriptions and to find its component DOs.

It is still an ongoing task to specify the required semantic explicitness in the necessary detail to support the FAIR principles and make FDOs machine actionable. A recent workshop[22] resulted in the formation of a coordination group and a technical implementation group that will define formal processes around requirements for FDO, called FDO Framework (FDOF),[23] and advance the specifications. FDOF will allow different technological implementations, nevertheless guaranteeing interoperability.

From the beginning research disciplines were active in formulating organisational and technical specification details of FDOs since it was realised that they are not just a technical concept, but are scientifically relevant to structure and represent the increasingly complex digital domain of knowledge in a persistent and technologically independent way. RDA Data Fabric,[24] GEDE-DO[25] and C2CAMP[26] were used as platforms to discuss DO matters. For FDOs these platforms will merge and additional actors such as from GOFAIR[27] will join. Workflow frameworks creating and consuming FDOs will become increasingly popular as will be indicated in chapter 4. All components being used in a specific workflow (workflow script, software tools being used, data being processed, ontologies being applied, etc.) can be seen as one complex collection consisting of different object types. Reproducibility suggests to put such collections into a container that can be transferred to another computational environment for example to be executed. In this respect the Research Objects initiative[28] was very active in specifying standards, so that a close collaboration should be envisaged.

4. Scientific Use Cases

Based on early discussions about Digital Objects and FAIR Digital Objects a survey was held among 47 scientific disciplines organized mainly as ESFRI initiatives (ref). Using the RDA GEDE collaboration platform,[29] the survey focused on the questions of whether scientific communities see any potential in the FDO concept, and if they have any experiences with advances going in that direction. The results of the survey showed that there is a wide recognition that with the current approaches and solutions fragmentation will continue preventing a breakthrough towards higher efficiency, effectiveness and trust. The research infrastructure experts are therefore looking for new options and where inspired by recent papers describing the FAIR Principles and the FDOs, as a way to implement the FAIR principles, as possible major anchors driving convergence and thus as effective means to help tackling the principle challenges mentioned in section 2.

The survey resulted in 31 case descriptions driven by scientific needs and interests in the participating research communities [13]. The results reflect the three major areas of challenges:

1. knowledge extraction from increasing amounts of complex data and this in an interdisciplinary context,

[22] https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/Paris-FDO-workshop

[23] https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/FDOF

[24] https://www.rd-alliance.org/group/data-fabric-ig.html

[25] https://github.com/GEDE-RDA-Europe/GEDE/tree/master/Digital-Objects

[26] https://www.rd-alliance.org/sites/default/files/2018Jan_ChairsMtg_C2CAMP.pptx

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

[28] http://www.researchobject.org/

[29] https://www.rd-alliance.org/groups/gede-group-european-data-experts-rda

2. the combination of knowledge which is represented in suitable forms to yield evidence and enable decision taking, and
3. the provisioning of an ecosystem of infrastructures that enables efficient and effective work in the two first mentioned areas.

Therefore, the suggested descriptions can be grouped into two major classes of interest: 1) Scientific and 2) Infrastructure and Networking. The providers of scientific interests are aware that their intentions can only be realised when an appropriate infrastructure will be available and when networking is being addressed to exchange information about this new FDO based approach.

4.1. Scientific Interests

4.1.1. Automatic Processing

Many research domains see the need to increase the degree of automation in the phase we above called "knowledge extraction" which is the domain of cutting edge statistical procedures such as machine learning. In this paper we will not discuss the nature of the scientific algorithms being applied, but the framework in which they can be executed. Introducing automatic workflows has a number of advantages for researchers:

- Assuming rich metadata and clear identifiers for data, researchers could specify requirement profiles for data and for trustworthy sources; they can then leave the job of finding suitable data to cyberspace agents which crawl data sources and build useful collections. In other words, the researcher takes a declarative role and can focus on the algorithms that will use the data and on the evaluation of the results.
- Workflow management tools will use rich metadata and identifiers to determine at every step what the state of computation is and which tools could be executed next in order to reach a certain goal. If properly designed by experts, such workflow managers would allow complex calculations to be carried out based on intentionally aggregated data collections.
- The introduction of actionable digital object collections, in which every included FDO is associated with a type as the FDO concept suggests, enables a high degree of automation. The creation or uploading of a new FDO with a specific type will trigger automatic procedures for data management or analytical tasks. This is what some labs are already doing.
- Large existing software stacks, for example for distributed cloud computing, could be amended to support FDOs by recording rich metadata, thereby enabling responsible data science and increasing reproducibility thanks to documentation in terms of provenance of data and production steps towards the computational results.

Well-known communities such as from climate modeling (ENES[30]), language data and technology (CLARIN[31]) and material science (NOMAD[32]), and research groups such as from the University of Illinois[33] or DEWCom,[34] a cloud software provider, have already started brainstorming and have initiated implementations in the above-mentioned direction. The motivation for this work is obvious, but the communities also realise the difficulties on this path. Steps which may lead to clear evidences in cutting edge research are often not predictable, i.e. researchers need to be able to react in a flexible way changing the steps to be followed and the parameters to be used. Workflow frameworks must therefore be handy and flexible on the one hand, but nevertheless exactly document all steps on the other hand. New interactive frameworks such as Jupyter[35] and Galaxy[36] amended with ready-made software libraries supporting FDOs seem to be promising for the labs. The extension of broadly used orchestration frameworks such as Weblicht[37] will enable even IT-laymen to create FAIR compliant FDOs with ease [14,15].

[30] https://is.enes.org/
[31] https://is.enes.org/
[32] https://nomad-coe.eu/
[33] https://illinois.edu/
[34] http://www.dewcomputing.org/
[35] https://jupyter.org/
[36] https://en.wikipedia.org/wiki/Galaxy_%28computational_biology%29
[37] https://www.clarin-d.net/en/language-resources-and-services/weblich

4.1.2 Stable Domain of Scientific Entities and Relationships

It is obvious that by using automatic workflows for a variety of management and analytic tasks as described above we are creating large numbers of relationships which need to be part of the scientific memory and thus need to be stored safely based on stable PIDs and accessible provenance (metadata) records. But these will be only one fraction of the relationships that will be created by research work.

The DISSCO[38] (biodiversity), ELIXIR[39] (biomedical research), E-RIHS[40] (cultural heritage), and EISCAT[41] (atmospheric research) initiatives give excellent examples about the challenges and needs which many scientific disciplines are faced with. At the bottom layer of the scientific knowledge space are the digital representations of gigantic numbers of physical objects such as, for example, specimen and/or observations of phenomena such as caused by, for example, diseases, treatment of diseases, chemical processes in the atmosphere and many others. Exemplars of the corresponding digital objects are hosted in many institutions and labs worldwide. They will be annotated based on multiple information sources, taxonomies, and ontologies, and as described above will be part of workflows to generate derived data. Specifically designed collections serve as basis for theorisation. Layers of digitally represented knowledge are thus created on top of the bottom layer of digital objects and form the incrementally growing scientific knowledge space in a similar way as papers with their many references to other papers formed the scientific memory until now. The inherent capabilities of abstraction, binding and encapsulation of FDOs based on stable identifiers will establish the trust of researchers to invest their time in developing and maintaining these knowledge spaces over the next decades. There is no doubt that researchers globally and in many disciplines are waiting on signals of convergence on FDOs since this would mean that their investments will not be lost.

In the health sector (ECRIN[42]) the additional requirements posed by sensitive data need to be considered, i.e. in addition to increased security requirements, transaction relations and contractual specifications on data are of greatest importance. Also here FDOs with their characteristics and clear identification are an excellent starting point to link the required type of information such as for example event descriptions in a blockchain in a stable way. For agricultural research, such as carried out at Wageningen University, a temporal aspect needs to be considered as well. Entire food creation chains from production to consumer products need to be registered and monitored. As has been shown already by the implementation of the Chinese supply chain control system for baby milk powder, the FDO have the potential to solve such complex systems with the promise to store temporal relations in a persistent and stable way.

Virtually integrating many existing databases increasingly often in an interdisciplinary setting requires proper strategies for enabling semantic cross-walks based on ontologies, but also in these cases the stability of the linked structures that are being created is required. This extension of what was said above was brought forward by the e-RIHS (cultural heritage), MIRRI[43] (microbial databases), GESIS[44] (social sciences), ForumX[45] (experimental sciences) and Instruct[46] (structural biology) communities.

4.1.3. Various Advanced Plans

An immediate follow-up concern in all these scientific fields that are creating these complex constructions of relationships between their digital objects is characterised by two questions: 1) How should we express knowledge in this endless mass of DOs and their relationships? 2) How should we identify relevance in this complex domain? There are no clear answers yet and there may be disciplinary differences in the approaches.

[38] https://www.dissco.eu/
[39] https://elixir-europe.org/
[40] http://www.e-rihs.eu/
[41] https://en.wikipedia.org/wiki/EISCAT
[42] https://www.ecrin.org/
[43] https://www.mirri.org/defaultinfo.aspx?page=Home
[44] https://www.gesis.org/home
[45] https://www.forumx.org/
[46] https://instruct-eric.eu/

A large community in particular in the biomedical domain believes that the way to go is to document knowledge in nano-publications which are essentially augmented RDF assertions extracted from lengthy papers, for example. Written papers in this scenario have the role to verify details in case of obvious questions, but the sheer amount of scientific papers requires a more condensed form of knowledge representations that is suitable for computational analysis.

Following this approach (presented by GOFAIR) for the highly condensed form of representing knowledge formally, smart statistics such as calculating cardinal assertions representing many others could be used to identify so-called "knowlets" which are clusters with high connectivity, their central concepts and their internal and external relationships [3]. These knowlets could be used to unleash unseen patterns and stimulate further theorising and investigations. Investing in establishing such a domain of billions of assertions requires trust in the stability of the underlying mechanisms. The systematic use of the FDO concept with its binding to metadata, relationships and in particular provenance to represent such knowlets builds a stable fundament for analyses that, for instance, examine older states of knowlets and their evolution. In this domain, each assertion, each concept and each central concept will be identified by a PID, and the FDO will have pointers to all related concepts.

A Virtual Research Environment is a concept that points to another virulent challenge IT-naive users are confronted with. They need an as simple as possible desktop with simple-to-use applications to deal with the possibilities state-of-the-art infrastructures are offering. Basically it is about reducing complexity at the man-machine interface which needs to be highly discipline specific also covering the terminology people are used to. In the climate modelling domain (ENES) the development of suitable VREs is a high priority task to allow many researchers to participate in their emerging FDO based landscape. Updating the already existing Virtual Collection Builder tool (CLARIN) to meet the requirements of the FDO model and thus to make it widely discipline independent and essential part of any VRE, since in all research disciplines collection building as a preparatory step towards processing will be needed.

In experimental fields the correct maintenance of lab-books is crucial where every experiment is being documented serving different purposes. Turning these lab-books into electronic versions including all important relationships is a step suggested by material science (NOMAD), but certainly of relevance for many other disciplines. Also here the FDO model with its binding capability could link an entry (a Digital Object) to an entry in a tamper-free blockchain if this would be required. In such disciplines an electronic lab-book would also be part of a VRE enabling fast linking.

Often errors are detected in data that have already been offered for re-use. It would be important for trust building, if the re-usage of data could be traced and a warning could be sent to those who already used the erroneous data. Tracing the re-usage of data in different contexts is a hard problem especially in science where flexibility and openness are important criteria. Systematically applying the FDO concept as suggested by VAMDC[47] experts working on atomic and molecular data could offer the opportunity to make real steps in checking authenticity and tracing re-use to finally increase the quality of scientific results.

4.2. Infrastructural and Networking Interests

All participating research infrastructures stressed the need for an FDO based infrastructure with appropriate basic services, and of funding flexible and extendable testbeds as a way to come to satisfying solutions. Many components are in the core of such an infrastructure and some of them have been specified by DONA[48] and RDA.[49] An elaborate list can be found in the FDO roadmap document [16] and will not be commented in detail in this paper.

In order to make progress in fairly new areas, much networking will be needed, not only amongst the key experts, but also including potential users. More than 150 experts from many research infrastructures supported the idea to submit a proposal to closely interact on FDO matters and organised several activities under the umbrella of RDA GEDE DO[50] that emerged from the RDA

[47] http://www.vamdc.org/

[48] https://www.dona.net/

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

[50] https://github.com/GEDE-RDA-Europe/GEDE/tree/master/Digital-Objects

Data Fabric group.

In addition to the communities already mentioned, we should here refer to the following communities and/or institutions which emphasise the importance of these aspects: CNRI-US,[51] DONA, GWDG/ePIC,[52] CNIC-CAS[53] and ICOS.[54]

5. Conclusion

We have presented the background against which the FDO concept has been evolving, and we have described its rationale as well as its potential for science. Our proposal has been supported by the results of a survey of use cases in a substantial number of scientific areas. There is still a lot of work ahead, including implementation and testing in realistic research contexts. For this purpose, a substantial interdisciplinary group of researchers needs to be mobilized.

The proposed approach not only meets the FAIR principles for scientific data management but also extends its scope through more explicit mechanisms allowing machine-actionability. FDOs are findable through their persistent identifiers, and contain rich metadata. They are accessible using a standardized communications protocol, for which we propose DOIP. Furthermore, FDOs achieve a higher level of interoperability because of the standard way in which their operations are encapsulated in the objects, and for that reason they are also highly reusable. These properties are important for the advancement data science, especially in interdisciplinary, cross-domain and cross-sector contexts.

In sum, we feel that the concept has matured enough to warrant our suggestion that the EOSC consider the option of using FDO as a basic mechanism for achieving global interoperability in data management. Thus we envisage the evolution of the EOSC into a Global Digital Object Cloud. The FDO architecture has the potential to be proliferated into both science, industry and public services.

Author Contributions: Conceptualization and writing, all authors. All authors have read and agreed to the submitted version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The conceptualization and writing of this article has benefited substantially from discussions with and comments from Carsten Baldauf, Dirk Betz, Claudia Biniossek, Christophe Blanchi, Claudia Draxl, Alex Hardisty, Ingemar Häggström, Margareta Hellström, Liu Jia, Nick Juty, Tibor Kalman, Willem Jan Knibbe, Larry Lannom, Giridhar Manepalli, Peter Mutschke, Christian Ohmann, Marlon Pierce, Robert Quick, Antonio Rosato, Erik Schultes, Ulrich Schwardmann, Karolj Skala, Hermon Sorin, Hannes Thiemann, Dieter van Uytvanck, Alexander Vasilenko, Tobias Weigel, Joachim Weimann, Harald Wypior, Tian Ye, Carlo Maria Zwölf and the RDA GEDE group.

Conflicts of Interest: The authors declare no conflict of interest.

Note about the References list: It would be advantageous to the reader to include URLs for accessing reading materials that are published online only. We are however unsure as to how to format these. It would be a great help if MDPI could provide an appropriate Zotero bibliography style that outputs the desired format.

References

1. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, Ij.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018.
2. Borgman, C.L. Data, Data, Everywhere, Nor Any Drop to Drink.; Bepress, 2014.
3. Mons, B. FAIR science for social machines: Let’s share metadata Knowlets in the Internet of FAIR data and services. *Data Intell.* **2019**, *1*, 1–15.
4. Lannom, L. Digital Object Architecture Primer 2019.
5. Kahn, R.; Wilensky, R. *A Framework for Distributed Digital Object Services*; CNRI, 1995;
6. Kahn, R.; Wilensky, R. A framework for distributed digital object services. *Int. J. Digit. Libr.* **2006**, *6*, 115–123.

[51] https://www.cnri.reston.va.us/

[52] https://www.pidconsortium.eu/

[53] http://english.cnic.cas.cn/

[54] https://www.icos-ri.eu/

7. Wittenburg, P. *Moving Forward on Data Infrastructure Technology Convergence: GEDE Workshop, Paris, 28–29 October 2019*; Paris, France, 2019;
8. Berg-Cross, G.; Ritz, R.; Wittenburg, P. RDA DFT Core Terms and Model 2016.
9. Schultes, E.; Wittenburg, P. FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure 2019.
10. Hodson, S.; Collins, S.; Genova, F.; Harrower, N.; Jones, S.; Laaksonen, L.; Mietchen, D.; Petrauskaité, R.; Wittenburg, P. *Turning FAIR into reality: Final Report and Action Plan from the European Commission Expert Group on FAIR Data*; Publications Office of the European Union, 2018;
11. Strawn, G. Open Science, Business Analytics, and FAIR Digital Objects 2019.
12. Bonino, L. Internet of FAIR data and services: Center of the hourglass 2019.
13. De Smedt, K.; Koureas, D.; Wittenburg, P. An Analysis of Scientific Practice towards FAIR Digital Objects 2019.
14. Van Uytvanck, D. Digital Objects – Towards implementation: The CLARIN use cases 2019.
15. Weigel, T. DO management for climate data infrastructure support 2019.
16. Digital Object Roadmap Document (V 3.0) 2019.