FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units

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

[1] Department of Linguistic, Literary and Aesthetic Studies, University of Bergen, P.O Box 7800,
5020 Bergen, Norway

[2] International Biodiversity Infrastructures, Naturalis Biodiversity Center, P.O. Box 9517, 2300 RA Leiden,
The Netherlands; dimitris.koureas@naturalis.nl

[3] Max Planck Computing and Data Facility, Max-Planck-Gesellschaft, Gießenbachstraße 2, 85748 Garching,
Germany; peter.wittenburg@mpcdf.mpg.de

[*] Correspondence: desmedt@uib.no; Tel.: +47-95836558

Abstract: Data science is facing the following major challenges: (1) developing scalable cross-disciplinary capabilities, (2) dealing with the increasing data volumes and their inherent complexity, (3) building tools that help to build trust, (4) creating mechanisms to efficiently operate in the domain of scientific assertions, (5) turning data into actionable knowledge units and (6) promoting data interoperability. As a way to overcome these challenges, we further develop the proposals by early Internet pioneers for Digital Objects as encapsulations of data and metadata made accessible by persistent identifiers. In the past decade, this concept was revisited by various groups within the Research Data Alliance and put in the context of the FAIR Guiding Principles for findable, accessible, interoperable and reusable data. The basic components of a FAIR Digital Object (FDO) as a self-contained, typed, machine-actionable data package are explained. A survey of use cases has indicated the growing interest of research communities in FDO solutions. We conclude that the FDO concept has the potential to act as the interoperable federative core of a hyperinfrastructure initiative such as the European Open Science Cloud (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 [1]. 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., through the construction of climate models, the monitoring of threatened species, or the detection of fake news. With "data" we do not just refer to published data, but to any data that has been created in research labs and lifted from its original workspace to a domain where it can be managed and shared.

The arrows in the diagram in Figure 1 indicate possible movements of data. Newly created data and collected data reside in temporary workspaces. Most of the data to be shared with others in internal or external workflows will move to the registered data domain in which rapidly increasing amounts of data are being amassed and managed for reuse. A small fraction of those data will be formally documented and published so that it can be properly cited based on metadata and according to publishers' requirements.
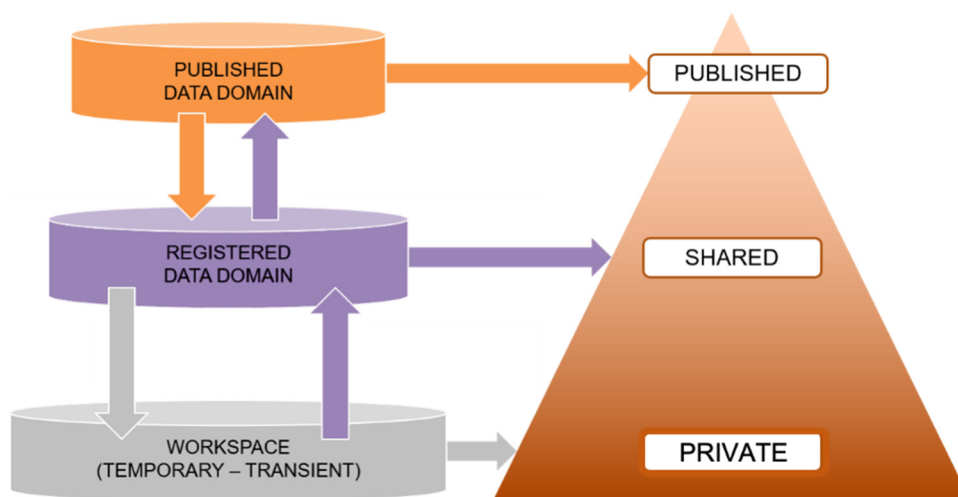
Figure 1. Layers of data with some data being published, more being shared for reuse in labs and collaborations, and a large amount residing in transient storage.[1]

Most of the data with a potential value in scientific processing should be reusable and should be referenced in a stable way in the registered data domain. From this level, which is spread over many repositories with various thematic foci, some data are being collected and copied down again into transient storage to be used in analytics tasks. Cross-silo data science, attempting to track down and collect all relevant data for certain tasks, is however known to be extremely inefficient and thus expensive, which hampers progress in tackling the global challenges.

In 2006, the first European Strategy Forum on Research Infrastructures (ESFRI) roadmap[2] was published, which provided an impetus for a large variety of disciplines to build research infrastructures which provide advanced facilities for supporting 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 establishment of unified data catalogues, the development of trustworthy repositories, and interfaces for making data accessible. These important advances were however gained only within each domain or discipline. Data integration and reuse across discipline boundaries are however still highly inefficient. The European eInfrastructures[3] have so far not been 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, cloud storage and networking capacity. Despite the provision of those core services, scientific communities still have not fully benefited from this public offering [2].

Consequently, today's research data infrastructures are neither sufficiently interconnected nor 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 the means to guarantee data accessibility and reusability over time. New strategies are required to improve practices. National and international funding bodies started an initiative towards open science and open data. Second, the wide adoption of principles for making data *findable*, *accessible*, *interoperable* and *reusable*, canonicalized in terms of the FAIR Guiding Principles, was meant to inform researchers on good data management but also to "provide 'steps along a path' toward machine-actionability" [3].

---

[1] Diagram adapted from EUDAT, a European Collaborative Data Infrastructure, https://eudat.eu/.
[2] European Strategy Forum on Research Infrastructures, https://www.esfri.eu/.
[3] https://ec.europa.eu/digital-single-market/en/policies/einfrastructure.

In this context, the European Open Science Cloud (EOSC)[4] was initiated as a next step to overcome the hurdles for cross-disciplinary data science and to progress towards implementing the FAIR principles. While a number of exploratory 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 only 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 potentially 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 datasets and thousands of tools, including the need for easily integrating them, are not being tackled.

Currently, EOSC does not address the fundamental questions of how to improve interoperability in the federating core, how to move toward 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 requirement of persistence, 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), as discussed at various meetings, among others, in Research Data Alliance[6] (RDA) groups.[7]

In the remainder of this paper, we will explain the FDO concept and we will show how it meets the FAIR principles. We will not go into details of implementations, on which progress is being made, but will explain the concept. 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 FAIR Digital Object Domain while 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. Global issues such as climate change, food security and lifelong health and wellbeing require concerted efforts across the realms of science, policy and technology [4,5]; so do the challenges of biodiversity loss, fake news and pandemic tracking and mitigation. The need for a globally coordinated response to global challenges has already been well documented, with important policy goals being articulated by international organisations.[8] The scope and complexity of these problems create 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 thereby enables meaningful and scalable integration and interpretation of evidence across different realms of science.

4　https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud.
5　https://www.eosc-portal.eu/.
6　https://rd-alliance.org/.
7　https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/FDO-Contributions.
8　https://en.unesco.org/sdgs.

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 reuse data across scientific domains is of pivotal importance.

2.2. Drowning in Data?

Today, we have experienced 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 costd, while published knowledge continues to accelerate [6]. 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.[9] Similarly, the Global Biodiversity Information Facility (GBIF) publicly serves more than 1.3 billion organism occurrence records today, having doubled its size in the last five years.[10] Similar examples of ever-increasing data availability can be found in other fields of science. 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 with reduced descriptors.[11] In the domain of research on languages, large collections of textual and audiovisual materials are being collected. The CLARIN[12] Virtual Language Observatory (VLO)[13] has brought together metadata on more than one million datasets related to language, of various granularities. These datasets, and the 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. The current interest in interdisciplinary data research is however not on par with the much larger increase in data that is potentially, but not in practice, available across disciplines. The properties of openness and FAIRness, despite being prerequisites for data discoverability and reuse, are in themselves not sufficient to support the transformation in community practices towards interdisciplinary data research. 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 reused in evolving contexts. More generally, the assessment of data quality is crucial for analyzing and using big data [7].

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 [8]. There is already much data out there, but we seem to have a skills gap in terms of making use of this richness, even when reuse is possible. Many researchers still prefer to refer to research results as documented in written papers since their reference is only dependent on proper language understanding. Data science, as is currently being conducted, requires mastering new skills which many researchers do not have, including approaches to maintain transparency in their reuse of data. Further support to the assertion that data science across discipline boundaries, while becoming important [9], is still not a comfort zone for

---

9　https://www.ncbi.nlm.nih.gov/genbank/statistics/.
10　https://www.gbif.org/news/6s1uLwFkwGq0oSHGNnzfUE/doubling-up-two-year-ebird-refresh-adds-200-million-more-observations.
11　https://nomad-coe.eu/.
12　The European Research Infrastructure for Language Resources and Technology, http://www.clarin.eu.
13　http://vlo.clarin.eu.

scientific communities, can be drawn from the relatively lower success rates of project proposals on interdisciplinary data research as compared to proposals that stay within disciplinary boundaries [10].

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 FAIR Digital Objects as a unified data organisation model, in particular for data science across the boundaries of domains and disciplines.

2.3. Interpreting Scientific Evidence in a Trusted Context

The global scientific domain is traditionally organized 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 through an implicit understanding of the context that 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 community-specific criteria for quality and fitness.

As data pieces "travel" across domain-agnostic repositories and aggregators, thus becoming available to a variety of scientific communities, they tend to become gradually uncoupled from the original context in which the data were generated. They effectively lose their contextual information, which is essential for communities to understand and evaluate the quality and fitness for purpose of the data, and to verify research results through replication studies. This carries the risk of reducing the reusability 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.

2.4. Advancing Data to Actionable Knowledge Units

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 data 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" (Artificial Intelligence), but these algorithms will also require that data and information about data be stored and disseminated in more robust and informative ways, thus allowing automatic systems to make sense of the vast number of individual pieces of knowledge.

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,[14] 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 [11]. Finally, this highly reduced space of knowlets can be used to draw conclusions, for example, for proper health treatments.

Thus a variety of scientific disciplines are creating complex relationships between basic findings, layers of derived data and knowledge sources of different kinds (ontologies, etc.). In the coming decades, scientific communities will spend huge efforts in creating and curating these relationships, which will form an essential part of our scientific knowledge that must be preserved. Traditional forms

---

14 Resource Description Framework, https://www.w3.org/2001/sw/wiki/RDF.

of publications will lose relevance compared to the knowledge captured in these relational frameworks. Therefore, new methods are required to preserve this new digital scientific memory.

The ability of a scientist to verify the relevance, provenance, completeness and fitness of data is an essential process in the scientific pipeline. As such, the practice of data sharing must include 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 are shared. This information goes beyond the typical domain-agnostic metadata as introduced, for instance, by Dublin Core[15] and instead requires standards for rich metadata that are not only discoverable and accessible (as by OAI-PMH[16]) but also interpretable in a machine-readable way. We need datasets that are fully actionable and comprehensible knowledge units, to be shared across digital environments and reused 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 that enable sharing and reuse of data-in-context.

2.5. 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 in itself, but the fact that any choice of tools brings with it the risk of a lock-in to particular standards, formats and other details of storage and processing that researchers are in principle not interested in. Researchers would, in general, prefer to deal with data objects at a more abstract level which are stable and not dependent on particular technologies.

The important role of abstraction in this respect is schematically presented in Figure 2 adapted from L. Lannom [12]. The user only needs to interact with FAIR Digital Objects which are identified by their persistent identifiers (PID)[17] and gets information from them through operations on metadata contained in the digital objects themselves. Metadata, data and PIDs are serviced by a federative core of registries and repositories, connected with the help of a unifying Digital Object Interface Protocol, as explained below.

Currently, research communities are offered funding to adapt to flavors of cloud systems and substantial funds are being spent on defining interfaces between such systems. However, the researcher is not interested in 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 or access speed. With respect to storage systems, we can expect new technological developments, driven by, for example, 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 that avoids the necessity for the researcher to interact with the underlying technology.

---

15   https://dublincore.org/.
16   Open Archives Initiative Protocol for Metadata Harvesting, https://www.openarchives.org/pmh/.
17   On the Web, a PID has the form of a Persistent Uniform Resource Locator (PURL), which curates redirection by means of a resolver. This scheme attempts to solve the problem of transitory locators in location-based schemes like HTTP. Example types of persistent identifiers are the Handle, the Digital Object Identifier (DOI), and the Archival Resource Key (ARK). There are also different types of globally unique identifiers which do not involve automatic curation by a resolver, such as the International Standard Language Resource Number (ISLRN).
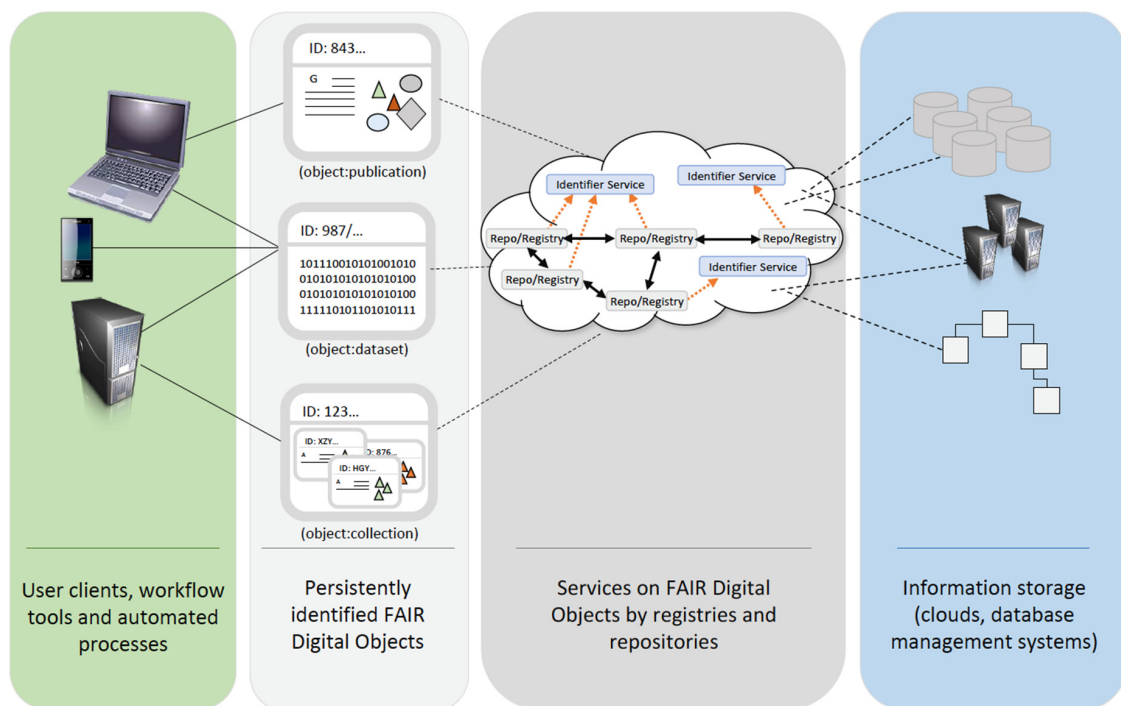
Figure 2. Layers of abstraction in the data domain.[18]

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 unit of data that is able to interact with automated data processing systems. We call this a FAIR Digital Object (FDO). From the perspective of a data scientist, an FDO is a stable actionable unit that bundles sufficient information to allow the reliable interpretation and processing of the data contained in it. In this section, we will give an introductory description of the FDO concept without going into technical details.

The encapsulation of information in an FDO is illustrated in Figure 3. FDOs are accessed through their PID. They may receive requests for operations, which they may inherit from their type, as known from object-oriented programming. Through operations, their metadata can be accessed, which in turn describes the enclosed data content (a bit sequence). FDOs enable abstraction, i.e., at the object management level it does not matter whether the FDO content is data, metadata, software, assertions, knowlets, etc. In an FDO, a data bit sequence is bound to all necessary metadata components (descriptive, scientific, system, access rights, transactions, etc.) at all times. While the PIDs and metadata of FDOs are normally open, access to their bit sequences may be subject to authentication and authorization, for example, to secure personal or otherwise protected data.

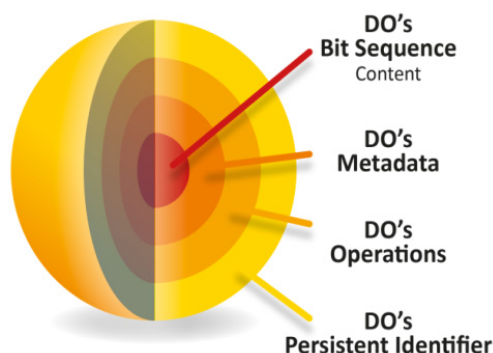18　Diagram adapted from L. Lannom [12].

Figure 3. Layers of encapsulation in an FDO. Objects are accessed through a PID and may receive requests for operations. Their metadata describes the enclosed data (bit sequence) [13].

From a researcher's point of view, we can imagine the following ideal data flow scenario: A sensor or other source produces a bit sequence of data, associates metadata with it and places both in the care of trustworthy repositories. These repositories analyze the metadata and may decide to host and manage the new data as an FDO. Doing so implies that the data and metadata will be bundled in an FDO to which a type (with associated operations) and a PID will be assigned. Furthermore, the repository may extend the metadata based on known contexts; for instance, policies for permission to use the data may be turned into licenses and access control lists. Then repositories may propagate the FDO to other agents on the Internet, which examine the metadata of the new FDO and decide whether the new data are of interest. If so, they will seek to get access to the content, contingent upon licensing, security measures and legal constraints encapsulated in the FDO. Compared to older protocols such as OAI-PMH,[19] new types of protocols will be used by repositories to offer their holdings, which are not restricted to metadata, but to any type of Digital Object.

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), the role of which can be compared to that of HTTP on the Web. For technical details on DOIP, we refer to the version 2 specification which has been published.[20]

The concept of FAIR Digital Objects which will enable this innovative scenario goes back to papers in the 1990s by Kahn and Wilensky [14,15]. 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. Shortly after, the design of the World Wide Web by Berners-Lee[21] represented the 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".[22]

The introduction of stable FDOs based on persistent identifiers will not change the current bulk of the web, which is unstructured information, but it will offer a more stable and lasting solution for datasets in the registered domain, which need to be preserved for a long time. This stability will increase the level of trust by researchers and other stakeholders who are investing big efforts towards the preservation of scientific knowledge. This will, in turn, provide the following advantages for research and development to address the challenges that were discussed in Section 2.

---

[19] https://www.openarchives.org/pmh/.
[20] https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf.
[21] https://en.wikipedia.org/wiki/Tim_Berners-Lee.
[22] https://www.bbc.com/news/science-environment-31450389.

Scaled Cross-Disciplinary Capabilities: FDOs are a way to create an interconnected multi-actor and multi-level ecosystem, since there is a protocol (DOIP) that speaks the "FDO language" 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 research more efficiently compared to the current data practices.

Data Made Accessible: The gap between the amount of data being created and our capability to make use has different causes. Among these, the lack of specialized skills and the low level of recognition for cross-disciplinary work cannot be addressed directly by FDOs. However, one reason, the lack of contextual information, will partly be addressed by the FDO concept, which enables binding of contextual information to data in a stable and persistent way.

Interpreting Scientific Evidence in a Trusted Context: In FDOs, contextual and fingerprint information can be associated with digital objects at different steps during their lifetime. Privacy information can and must be associated with each digital object in a tamper-free way. In fact, all metadata are always bound to the data, so that researchers always have access to provenance and other information necessary for assessing fitness for purpose—thus building trust.

Domain of Reasoning: The evolving complex domain of knowledge in and across all scientific domains drives us towards automatic processing in our quest for data-driven conclusions. As actionable units, FDOs capture and build complex relationships over long time periods. Thus, they form building blocks that build knowledge structures for our evolving digital scientific memory.

Advancing Data to Actionable Knowledge Units: As FDOs travel through cyberspace and time, their encapsulation ensures that even after decades and in spite of changing technology and changing actors, data and their context will remain available as complete units. They will not lose any information but may accumulate contexts of reuse over time.

Tool Proliferation and Fundamental Decisions: The FDO concept achieves abstraction that hides technological details from 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. Digital objects in the DFT Core Model

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 [14,15]. After the World Wide Web established HTML resources as referenceable and shareable digital entities, this technology started to open possibilities for the Semantic Web[23] and the Linked Data Platform.[24] However, the linked data concept is not sufficient [16], as also concluded at a recent workshop [17]. Subsequently, the term Digital Object (DO) found its revival in two developments: (a) the Data Foundation and Terminology Group of RDA,[25] 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 [18] did not differ so much from the definitions used by Kahn and Wilensky. For more details on the term and concept of DO see [17].

Two diagrams explain the pervasive nature of DOs in the DFT Core Model. The diagram in Figure 4 indicates the simple structure of this model. The content of a DO is encoded as a structured bit-sequence and stored in repositories. It is assigned a globally unique, persistent and resolvable identifier (PID), as well as rich metadata (descriptive, scientific, system, provenance, rights, etc.). Metadata descriptions themselves are DOs. Moreover, DOs can be aggregated to collections which are also DOs with a content consisting of the references to its components. This simple definition makes

---

DO a generic concept, abstracting away from the many possible types of content of a DO, and covering the whole domain of digital data entities.
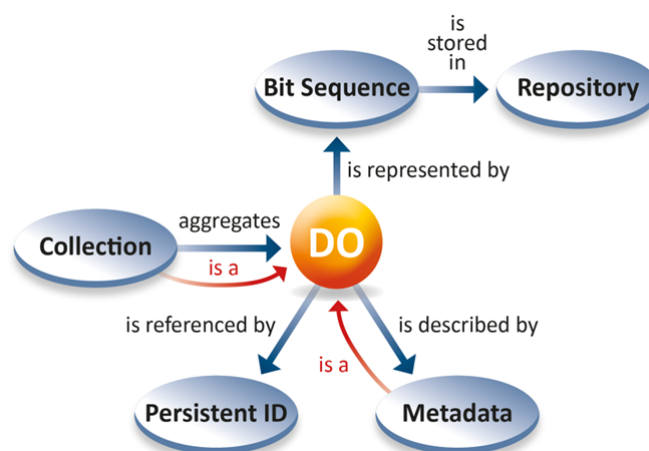

Figure 4. Relations between information associated with a Digital Object [13].

The diagram in Figure 5 indicates the role of the PID as the anchor point for accessing and reusing the DO. Assuming that the PID is indeed persistent, which is based on a cultural agreement, it makes sense to bind essential information into the PID record which will be returned to the user when a PID is being resolved. This essential information may contain paths to access the bit sequence, the metadata (also a DO), the rights record containing permission specifications, a pointer to a blockchain entry storing the transactions, a checksum for verification, etc. Furthermore, DOs are typed; operations are associated with a DO based on its type, which is a familiar, powerful concept from object-oriented programming. The RDA Kernel Information group[26] 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.[27]
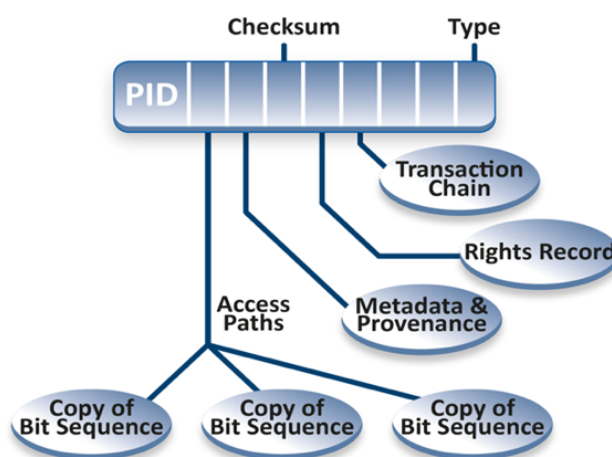

Figure 5. Components and pointers in a Digital Object [13].

3.3. From Digital Objects towards FAIR Digital Objects

The above definition of a DO already satisfies some of the FAIR principles and has been used as a blueprint for their implementation [19]. Intensive discussions between RDA and GOFAIR experts over

---

[26] https://www.rd-alliance.org/group/pid-kernel-information-wg/wiki/pid-kernel-information-guiding-principles.
[27] https://www.rd-alliance.org/group/data-type-registries-wg/outcomes/data-type-registries.

the past year revealed that additional specifications were required to make DOs fully FAIR compliant. Early papers, including one by a European Commission Expert Group on FAIR Data, coined the term FAIR Digital Object [20,21], but it was L. Bonino who recently identified the missing parts [22]. It has thus 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 any statements about their usage. The FAIR Digital Object (FDO) model needs to be specific on three aspects:

1. The FDO model requires the definition of PID attributes and their registration in a trustworthy type registry or a more complex type ontology, while trustworthy repositories are requested to use these attributes in order to achieve interoperability and machine actionability.
2. The FDO model requires metadata descriptions to be interpretable by machines. This implies that their semantic categories must be declared and registered. A moderate requirement could be to declare at least metadata categories strictly necessary for basic management, such as where to find the PIDs of relevant information components.
3. The FDO model requires the construction of collections to be machine actionable, thus 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 necessary detail to support the FAIR principles and make FDOs fully machine actionable. A recent workshop[28] 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),[29] and will elaborate the specifications. FDOF will allow for different technical implementations, nevertheless guaranteeing interoperability.

Researchers from many disciplines have been active in formulating organizational and technical specification details of FDOs since their scientific relevance became clear against the present background, revealing a pressing need to structure and represent increasingly complex scientific knowledge in a way that will be not only persistent but also independent of evolving underlying technology. International discussions have been held in two RDA groups, the Data Fabric Interest Group[30] and the Digital Objects subgroup of the RDA Group of European Data Experts (GEDE-DO[31]), as well as in the C2CAMP[32] cooperation. We are currently seeing a tendency for these discussions to converge, while additional actors, e.g., from GOFAIR,[33] are joining [19].

Workflow frameworks that create and consume FDOs are expected to become increasingly popular, as further discussed in the next section. All components in a specific workflow (workflow script, software tools being used, data being processed, ontologies being applied, etc.) can be seen as a complex collection consisting of different object types. The goal of reproducibility suggests putting such collections into a container that can be transferred to another computational environment where it retains its full functionality. In this respect, the Research Objects (RO) initiative[34] has been very active in specifying container and transmission standards that would ideally allow the execution of a workflow, including all its components on different virtual machines. Thus, the RO initiative is working towards complementary goals, so that a close collaboration can be envisaged.

28　https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/Paris-FDO-workshop.
29　https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/FDOF.
30　https://www.rd-alliance.org/group/data-fabric-ig.html.
31　https://github.com/GEDE-RDA-Europe/GEDE/tree/master/Digital-Objects.
32　https://www.rd-alliance.org/sites/default/files/2018Jan_ChairsMtg_C2CAMP.pptx.
33　https://www.go-fair.org/.
34　http://www.researchobject.org/.

4. Scientific Use Cases

Based on ongoing discussions about Digital Objects and FAIR Digital Objects in the RDA GEDE collaboration,[35] involving about 150 experts from more than the 47 scientific disciplines, also representing many infrastructures on the ESFRI[36] roadmaps, a survey was held in 2018. This survey was formulated as an open request for use cases, i.e., descriptions of ongoing work related to the DO concept and of the potential of the DO concept for this work. Details on the survey and summaries of the different use cases have been described elsewhere [23]. Therefore, in this section, we will only summarize the results and discuss their relevance in the present context.

The results of the survey showed that there is a wide recognition that the current status of fragmentation continues to hamper breakthroughs towards higher efficiency, effectiveness and trust. The research infrastructure experts are therefore looking for new options and were inspired by recent papers that describe FDOs, implementing the FAIR principles, as possible major anchors driving convergence and thus as effective means to help to tackle the major above-mentioned challenges.

The survey resulted in 31 case descriptions driven by scientific needs and interests in the participating research communities [23]. The results reflect the three major areas of challenges:

1. knowledge extraction from increasing amounts of complex data, especially in an interdisciplinary context;
2. the aggregation of knowledge which requires a broad domain of trust and compatibility to yield evidence and enable decision-taking; and
3. the provisioning of an ecosystem of research infrastructures that enables efficient and effective work on the two previous points.

These challenges link scientific interests to issues of broad infrastructure and networking. Providers of scientific data and methods are aware that their intentions can only be realized when an appropriate infrastructure becomes available and when networking is supported at appropriate levels, including that of hyperinfrastructure initiatives such as EOSC.

4.1. Scientific Interests

4.1.1. Automatic Processing and Workflows

The need to increase the degree of automation in knowledge extraction is apparent in many domains, especially in data science, where large and heterogeneous amounts of data must be aggregated and fed into machine learning algorithms. 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, leaving the job of finding suitable data to cyberspace agents that crawl FDO repositories and build useful collections; thus, researchers can focus more on the algorithms that will use the data and on the evaluation of the results.
*   Workflow management tools could use rich metadata and identifiers in FDOs to track the status at every step of computation and determine which processes 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, 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. Some research labs report that they are already doing something similar.

---

[35] https://www.rd-alliance.org/groups/gede-group-european-data-experts-rda.
[36] European Strategy Forum on Research Infrastructures, https://www.esfri.eu/.

- 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 data provenance and tracking production steps taken towards computational results.

Well-known communities such as those engaged in climate modelling (ENES[37]), language data and technologies (CLARIN[38]) and material science (NOMAD[39]), research groups at the University of Illinois[40], the cloud provider DEWCom,[41] and others have started designing and implementing solutions in the above-mentioned direction. The motivation for such innovations is obvious, but the communities also face difficulties along this path. Experimentation needs the flexibility to change parameters while still keeping stable data objects. The amendment of interactive frameworks such as Jupyter[42] and Galaxy[43] with software libraries that will support FDOs, therefore seems promising. The extension of broadly used orchestration frameworks such as Weblicht[44] will enable even IT-laymen to create FAIR compliant FDOs with ease [24,25].

4.1.2. Stable Domain of Scientific Entities and Relationships

Automatic workflows for a variety of data management and analytic tasks, as described above, create large numbers of relationships that need to be part of the scientific memory and therefore must be stored with PIDs and metadata records that crucially record data provenance. However, the relationships that follow from explicit data management and processing will be only a fraction of the relationships that will be automatically created.

The DiSSCo[45] (biodiversity), ELIXIR[46] (biomedical research), E-RIHS[47] (cultural heritage), and EISCAT[48] (atmospheric research) initiatives provide excellent examples of the challenges and needs which many scientific disciplines are faced with [26,27]. At the bottom layer of the scientific knowledge space are the digital representations of large numbers of physical objects such as, for example, biological specimens [28], or observations of phenomena such as caused by, for example, diseases, treatments 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 are annotated, based on multiple information sources, taxonomies, and ontologies, and, as described above, the resulting digital objects will be part of workflows to generate derived data. Specifically designed collections serve as a basis for new theories. 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 way similar to the knowledge network created by scientific papers with their references to other papers 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.

---

37 https://is.enes.org/.
38 http://www.clarin.eu/.
39 https://nomad-coe.eu/.
40 https://illinois.edu/.
41 http://www.dewcomputing.org/.
42 https://jupyter.org/.
43 https://en.wikipedia.org/wiki/Galaxy_%28computational_biology%29.
44 https://www.clarin-d.net/en/language-resources-and-services/weblicht.
45 https://www.dissco.eu/.
46 https://elixir-europe.org/.
47 http://www.e-rihs.eu/.
48 https://en.wikipedia.org/wiki/EISCAT.

In the health sector (e.g., ECRIN[49]), the additional requirements posed by sensitive data need to be considered. In addition to increased security restrictions on transfer of data, transaction relations and contractual specifications on data are of the greatest importance. Furthermore, in this respect, properly typed FDOs can provide inherent security measures, e.g., by recording transactions in a blockchain. 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 shown already by the implementation of a Chinese supply chain control system for baby milk powder, the FDO approach has the potential to solve such complex systems with the promise to store temporal relations in a persistent and stable way.

The virtual integration of many existing databases, increasingly common in an interdisciplinary setting, requires proper strategies for enabling semantic crosswalks based on ontologies; but also in these cases, the created linked structures must be stable. Suitable extensions of the concept are brought forward by the e-RIHS (cultural heritage), MIRRI[50] (microbial databases), GESIS[51] (social sciences), ForumX[52] (experimental sciences) and Instruct[53] (structural biology) communities.

4.1.3. Advanced Plans for Management and Security

A follow-up in scientific communities that are experimenting with complex constructions of relationships between digital objects concerns two questions: How should we express what constitutes "knowledge" in this endless mass of digital objects and their relationships? And how should we identify "relevance" in this complex domain? There are no clear answers yet and there may be disciplinary differences in approaching the questions.

A large community, in particular in the biomedical domain, believes that the way forward is to document knowledge in nano-publications which are essentially augmented RDF assertions, which are, for example, extracted from an extensive literature. In this scenario, written papers have a role in verifying details in cases, but the sheer amount of scientific papers requires a more condensed form of knowledge representation that is suitable for computational analysis.

Following an approach suggested by GOFAIR,[54] smart statistics such as calculating cardinal assertions that represent many others could be used to identify "knowlets" as clusters with high connectivity, their central concepts and their internal and external relationships [11]. These knowlets could be used to unleash unseen patterns and stimulate further theory development and investigations. Investing in establishing 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 (VRE) is a concept that points to the virulent challenge IT-naive users are confronted with. Such users cannot grasp all the possibilities that state-of-the-art infrastructures are offering, so they need a very simple desktop with easy-to-use applications that brings all those possibilities at their fingertips. In this context, the challenge is one of reducing complexity at the man-machine interface, which needs to be tailored to the disciplinary terminology and jargon that researchers are used to. In the climate modelling domain, ENES reports that the development of suitable VREs is a high priority task to allow many researchers to participate in their

---

49 https://www.ecrin.org/.
50 https://www.mirri.org/defaultinfo.aspx?page=Home.
51 https://www.gesis.org/home.
52 https://www.forumx.org/.
53 https://instruct-eric.eu/.
54 https://www.go-fair.org/.

emerging FDO-based landscape. Likewise, it seems possible for CLARIN to update their existing Virtual Collection Builder tool to meet the requirements of the FDO model and make it more available to related humanities disciplines [24].

In experimental fields, the maintenance of lab books, in which every experiment is being documented in relation to its purpose, is crucial. Turning these lab books into electronic versions which contain all important relationships is a step suggested by material science (NOMAD), but certainly of relevance for many other disciplines as well. The FDO model with its binding capability could achieve traceability by binding entries to a tamper-free blockchain if required, in addition to making electronic lab-books part of VREs.

Often errors are detected in data that have already been offered for reuse. An important step towards trust building would consist of tracing the reuse of data and issuing a warning to those who already used the erroneous data. Tracing the reuse of data is currently not easy, especially in contexts where open access is important. Systematically applying the FDO concept, as suggested by VAMDC[55] experts working on atomic and molecular data, would offer the opportunity to make real steps toward checking authenticity and tracing reuse in order to validate and monitor 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 satisfactory solutions. Many components are at the core of such an infrastructure and some of them have been specified by DONA[56] and RDA.[57] An elaborate list can be found in the FDO roadmap document [29] 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 among the key experts, but also including potential users. More than 150 experts from many research infrastructures have supported the idea to submit a proposal to closely interact on FDO matters and organised several activities under the umbrella of RDA GEDE DO[58] that emerged from the RDA Data Fabric group.

In addition to the communities already mentioned, we should here refer to the following communities and/or institutions which have participated in the discussions and emphasise the importance of these developments: CNRI-US,[59] DONA, GWDG/ePIC,[60] CNIC-CAS[61] and ICOS.[62]

5. Conclusions

We have presented the background against which the FAIR Digital Object (FDO) concept has been evolving, and we have described its rationale as well as its potential for science. This concept is fundamentally changing our view of what a dataset should be. In current practice, the world of digital data lacks explicit semantics, and despite best practices,[63] the information obtained by resolving a PID is not yet fully interpretable by machines. Instead, we propose that data should be a stable, typed bundle of information, directly accessed by a PID and able to return requests for information in a semantically defined way.

The proposed approach not only meets the FAIR principles for scientific data management, but also extends their scope through more explicit mechanisms allowing machine-actionability. FDOs

---

55 http://www.vamdc.org/.
56 https://www.dona.net/.
57 https://www.rd-alliance.org/.
58 https://github.com/GEDE-RDA-Europe/GEDE/tree/master/Digital-Objects.
59 https://www.cnri.reston.va.us/.
60 https://www.pidconsortium.eu/.
61 http://english.cnic.cas.cn/.
62 https://www.icos-ri.eu/.
63 https://support.datacite.org/docs/landing-pages.

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 typed objects, and for that reason, they are also highly reusable. These properties are important for the advancement of data science, especially in interdisciplinary, cross-domain and cross-sector contexts.

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. 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. Furthermore, the FDO architecture has the potential to be widely proliferated in science, industry and public services.

References

1. Borgman, C.L. The conundrum of sharing research data. *J. Am. Soc. Inf. Sci. Technol.* **2012**, *63*, 1059–1078. [CrossRef]
2. Koureas, D.; Arvanitidis, C.; Belbin, L.; Berendsohn, W.; Damgaard, C.; Groom, Q.; Güntsch, A.; Hagedorn, G.; Hardisty, A.; Hobern, D.; et al. Community engagement: The 'last mile' challenge for European research e-infrastructures. *Res. Ideas Outcomes* **2016**, *2*. [CrossRef]
3. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.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. [CrossRef] [PubMed]
4. O'Rourke, M.; Crowley, S.; Gonnerman, C. On the nature of cross-disciplinary integration: A philosophical framework. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* **2016**, *56*, 62–70. [CrossRef] [PubMed]
5. Efstathiou, S. Is it possible to give scientific solutions to Grand Challenges? On the idea of grand challenges for life science research. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* **2016**, *56*, 48–61. [CrossRef] [PubMed]
6. Bornmann, L.; Mutz, R. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *J. Assoc. Inf. Sci. Technol.* **2015**, *66*, 2215–2222. [CrossRef]
7. Cai, L.; Zhu, Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Sci. J.* **2015**, *14*, 2. [CrossRef]
8. Borgman, C.L. *Data, Data, Everywhere, Nor Any Drop to Drink*; Springer: Berlin/Heidelberg, Germany, 2014. Available online: https://works.bepress.com/borgman/322/ (accessed on 21 April 2020).
9. Chen, S.; Gingras, Y.; Arsenault, C.; Larivière, V. Interdisciplinarity patterns of highly-cited papers: A cross-disciplinary analysis. *Proc. Am. Soc. Inf. Sci. Technol.* **2014**, *51*, 1–4. [CrossRef]
10. Bromham, L.; Dinnage, R.; Hua, X. Interdisciplinary research has consistently lower funding success. *Nature* **2016**, *534*, 684–687. [CrossRef]
11. 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. [CrossRef]

12.	Lannom, L. Digital Object Architecture Primer. Digital Objects- from RDA Results towards Implementation. RDA Side Meeting, Philadelphia. Available online: https://github.com/GEDE-RDA-Europe/GEDE/blob/master/Digital-Objects/DO-Workshops/Workshop-Philadelphia-2019/lannom-do-p13.pdf (accessed on 1 March 2020).

13.	Wittenburg, P.; Strawn, G.; Mons, B.; Bonino, L.; Schultes, E. *Digital Objects as Drivers towards Convergence in Data Infrastructures*; EUDAT: Helsinki, Finland, 2019. [CrossRef]

14.	Kahn, R.; Wilensky, R. *A Framework for Distributed Digital Object Services*; CNRI: Reston, VA, USA, 1995. Available online: https://www.cnri.reston.va.us/k-w.html (accessed on 21 April 2020).

15.	Kahn, R.; Wilensky, R. A framework for distributed digital object services. *Int. J. Digit. Libr.* **2006**, *6*, 115–123. [CrossRef]

16.	Bechhofer, S.; Buchan, I.; Roure, D.D.; Missier, P.; Ainsworth, J.; Bhagat, J.; Couch, P.; Cruickshank, D.; Delderfield, M.; Dunlop, I.; et al. Why linked data is not enough for scientists. *Future Gener. Comput. Syst.* **2013**, *29*, 599–611. [CrossRef]

17.	Wittenburg, P. Moving Forward on Data Infrastructure Technology Convergence: GEDE Workshop, Paris, 28–29 October 2019; Paris, France, 2019. Available online: https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/Paris-FDO-workshop (accessed on 1 March 2020).

18.	Berg-Cross, G.; Ritz, R.; Wittenburg, P. *RDA DFT Core Terms and Model*; EUDAT: Helsinki, Finland, 2016. Available online: http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318 (accessed on 21 April 2020).

19.	Schultes, E.; Wittenburg, P. *FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure*; EUDAT: Helsinki, Finland, 2019. [CrossRef]

20.	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*; European Union: Brussels, Belgium, 2018. [CrossRef]

21.	Strawn, G. *Open Science, Business Analytics, and FAIR Digital Objects*; EUDAT: Helsinki, Finland, 2019. [CrossRef]

22.	Bonino, L. Internet of FAIR Data and Services: Center of the Hourglass. Moving Forward on Data Infrastructure Technology Convergence: GEDE Workshop, Paris, 28–29 October 2019; Paris, France, 2019. Available online: https://github.com/GEDE-RDA-Europe/GEDE/blob/master/FAIR%20Digital%20Objects/Paris-FDO-workshop/GEDE_Paris_Session%201_Bonino.pptx (accessed on 1 March 2020).

23.	De Smedt, K.; Koureas, D.; Wittenburg, P. *An Analysis of Scientific Practice towards FAIR Digital Objects*; EUDAT: Helsinki, Finland, 2019. [CrossRef]

24.	Van Uytvanck, D. Digital Objects–Towards Implementation: The CLARIN Use Cases. Digital Objects– from RDA Results towards Implementation: RDA Side Meeting, Philadelphia, 2019. Available online: https://github.com/GEDE-RDA-Europe/GEDE/blob/master/Digital-Objects/DO-Workshops/Workshop-Philadelphia-2019/uytvanck-do-p13.pdf (accessed on 1 March 2020).

25.	Weigel, T. DO Management for Climate Data Infrastructure Support. Moving Forward on Data Infrastructure Technology Convergence: GEDE Workshop, Paris, French, 28–29 October 2019. Available online: https://github.com/GEDE-RDA-Europe/GEDE/blob/master/FAIR%20Digital%20Objects/Paris-FDO-workshop/GEDE_Paris_Session%202_Weigel.pptx (accessed on 1 March 2020).

26.	Addink, W.; Koureas, D.; Rubio, A.C. DiSSCo as a New Regional Model for Scientific Collections in Europe. *Biodivers. Inf. Sci. Stand.* **2019**, *3*, e37502. [CrossRef]

27.	Casino, A.; Raes, N.; Addink, W.; Woodburn, M. Collections Digitization and Assessment Dashboard, a Tool for Supporting Informed Decisions. *Biodivers. Inf. Sci. Stand.* **2019**, *3*, e37505. [CrossRef]

28.	Lannom, L.; Koureas, D.; Hardisty, A.R. FAIR Data and Services in Biodiversity Science and Geoscience. *Data Intell.* **2020**, *2*, 122–130. [CrossRef]

29.	Digital Object Roadmap Document (V 3.0); 2019. Available online: https://github.com/GEDE-RDA-Europe/GEDE/blob/master/Digital-Objects/Foundation-Documents/DORoadmap.pdf (accessed on 1 March 2020).