information

MDPI

Article

Revolutions Take Time

Peter Wittenburg [1,*] and George Strawn [2]

[1] FDO Forum, 2333 CR Leiden, The Netherlands
[2] National Academy of Sciences, Washington, DC 20001, USA; gostrawn@gmail.com
* Correspondence: peter.wittenburg@mpcdf.mpg.de; Tel.: +49-282149180

Abstract: The 2018 paper titled "Common Patterns in Revolutionary Infrastructures and Data" has been cited frequently, since we compared the current discussions about research data management with the developments of large infrastructures in the past believing, similar to philosophers such as Luciano Floridi, that the creation of an interoperable data domain will also be a revolutionary step. We identified the FAIR principles and the FAIR Digital Objects as nuclei for achieving the necessary convergence without which such new infrastructures will not take up. In this follow-up paper, we are elaborating on some factors that indicate that it will still take much time until breakthroughs will be achieved which is mainly devoted to sociological and political reasons. Therefore, it is important to describe visions such as FDO as self-standing entities, the easy plug-in concept, and the built-in security more explicitly to give a long-range perspective and convince policymakers and decision-makers. We also looked at major funding programs which all follow different approaches and do not define a converging core yet. This can be seen as an indication that these funding programs have huge potentials and increase awareness about data management aspects, but that we are far from converging agreements which we finally will need to create a globally integrated data space in the future. Finally, we discuss the roles of some major stakeholders who are all relevant in the process of agreement finding. Most of them are bound by short-term project cycles and funding constraints, not giving them sufficient space to work on long-term convergence concepts and take risks. The great opportunity to get funds for projects improving approaches and technology with the inherent danger of promising too much and the need for continuous reporting and producing visible results after comparably short periods is like a vicious cycle without a possibility to break out. We can recall that coming to the Internet with TCP/IP as a convergence standard was dependent on years of DARPA funding. Building large revolutionary infrastructures seems to be dependent on decision-makers that dare to think strategically and test out promising concepts at a larger scale.

Keywords: data management; data infrastructures; FAIR principles; FAIR Digital Objects

check for updates

Citation: Wittenburg, P.; Strawn, G. Revolutions Take Time. *Information* **2021**, *12*, 472. https://doi.org/10.3390/info12110472

Academic Editor: Riccardo Albertoni

Received: 10 October 2021
Accepted: 10 November 2021
Published: 16 November 2021

1. Introduction

In our 2018 paper "Common Patterns in Revolutionary Infrastructures and Data" [1], we argued that revolutionary infrastructures can be characterized by simple core standards which, on the one hand, promise some stability associated with a step of disruption but, on the other hand, act as a new common platform for dynamic evolution above that platform. We also argued that the emerging distributed data infrastructure will be such a revolutionary infrastructure and thus requires a convergence towards such a simple standard as its key pillar. We identified the FAIR principles [2] and the FAIR Digital Objects (In this paper we will not describe the FDO concept in detail but refer to some publications and the Base Definition which is published at the fairdo.org website (http://fairdo.org; accessed on 15 November 2021). In short one can state that an FDO has a structured bit-sequence, is identified by a PID and is associated with metadata. The bit-sequences of FDOs can include all kinds of types (data, metadata, software, etc.).) [3,4] as candidates for achieving convergence but also admitted that it will take time to agree on these concepts and turn them into practice. Three more years of discussions, substantial investments in infrastructure projects, and new initiatives confirm our view that indeed FAIR and FDOs are good candidates, but that more years will be needed to achieve a state where major stakeholders will have come to agreements and where industry will finally join to develop technology.

Y.N. Harari describes in his book "Homo Deus" [5] that information (*and data*) want to be free and that this is a new revolutionary step in the development of culture. Knowledge will increasingly be built on empirical data and analyzed with the help of smart mathematics. L. Floridi argues in a similar direction when he speaks about 3rd order technology in his book "4th revolution" [6] where humans are out of the decision loops in many scenarios and where machines consuming masses of data are programmed to identify small patterns in big data to help us in tackling the grand challenges. We believe that indeed "dataism" as Harari calls the new phase will become reality, but that major steps in global infrastructure building will be required to make it happen in a way that societies as a whole will profit from this revolution.

In this paper, we will touch on some aspects that hamper agreement and therefore require more attention and further discussions. We will start, however, by describing the vision which we referred to in abstract terms in our early paper. After three more years of discussion, we seem to be able to sketch the goal more clearly. Then we will relate the vision with the current initiatives realizing that already a great amount money is spent on building research/data infrastructures. Finally, we will draw some conclusions.

2. Vision

2.1. Integrated Virtual Data Collection

G. Strawn recently introduced a useful categorization to indicate where we are and where we will go [7] which is summarized in Table 1. In the 1950s we saw an increasing number of "*individual computers and each of them had separated data sets*". In the 1990s we saw a dramatic change in the concept of computing when workstations entered the market and when the phrase "the network is the computer" was born. In short, we started to speak about "*one virtual computer and still separated data sets*". Ahead of us is another dramatic change since we see the possibility to speak about "*one virtual computer and one virtual data collection*" (It should be noted here that a comparable term "international data space" has been coined recently in industry [8]). However, we need to better understand separating these two dimensions as different logical layers, i.e., cloud systems for example are not implementing this integrated virtual data collection, they can only be seen as a technology facilitating this.

Table 1. This table indicates a rough categorization of phases of virtual integrations in IT development.

1950s	many individual computers	separated data sets
1990s	one virtual computer	separated data sets
2030s	one virtual computer	one virtual data collection

The basis of the 1990 vision was the appearance of the Internet facilitated through the global adoption of TCP/IP and a change in computer technology that moved away from self-supporting mainframes towards an exponential increase of smaller computers, a trend that is still ongoing with ever-greater miniaturization of electronic circuits.

After the first applications that populated the Internet such as email exchange based on SMTP, some years later the Web application was added, which created the globally integrated landscape for human-readable information all based on the comparatively simple HTTP and HTML standards. In the meantime, the Web is being used for many different tasks, but it basically cannot overcome its design for human-centered information exchange, being ephemeral by nature and mixing identification and location.

Organizations and institutes working with huge data volumes and inherent complexity through the manyfold relationships between the data entities of different types that require persistency and stability of their digital data space opted for an approach which we describe as the domain of Digital Objects [9] referenced by Handles/DOIs (DOIs are Handles with the prefix 10 and associated with a specific business model inspired by the area of electronic publications.). This is true for communities such as the publishers, the film industry, and large institutions collaborating globally, for example, the climate modeling community. These Handles are globally unique and resolvable persistent identifiers (PID) (Persistence is based on social commitments. The DONA Foundation operating the global Handle resolution system under the umbrella of the ITU is based on strong commitments, however, local Handle resolvers can differ substantially with respect to their persistence commitments.) and are resolved to information about the digital object according to detailed specifications. These specifications introduce a predictable resolution behavior and machine actionability, thus FAIR compliance. Hence, we currently speak about FAIR Digital Objects.

Currently, there are at least two serious approaches that are working to create the "virtual data collection" which we introduced above: (1) the Digital Object (DO) approach based on Handles/DOIs and DOIP [10] and (2) the Linked Data (LD) approach [11] based on URIs and the Web protocol stack. Perhaps this situation can be compared when AC and DC current approaches in early electrification phases were investigated [12]. Finally, AC won since energy provisioning was the first priority, and it became obvious that energy transmission losses were considerably lower for AC. Today we know that both AC and DC are needed depending on the application. With respect to the creation of an integrated virtual data collection, it is too early to make final statements about the question of which of the two approaches to implement FDOs will offer advantages for the many different kinds of applications in the long run. Therefore, both approaches will need to co-evolve. In this article we will focus on the Digital Object approach, the LD approach is widely known.

2.2. *Easy Plug-In*

A second important concept is that of an easy *plug-in*. The Internet was created by defining TCP/IP as the basic unifying protocol. It was agreed that, whoever can plug-in his/her end device and show that it is fully supporting the TCP/IP protocol requirements, should be able to become part of the Internet, i.e., their device could immediately be addressed and exchange messages. This dream of a simple plug-in has become reality despite all doubts and critiques at the beginning.

As indicated in Figure 1, we should now start dreaming of a simple plug-in for data. Due to security and sustainability requirements, we see here an intermediation role for trustworthy (The term "trust" includes many different aspects. Here we focus on the capability of a repository to curate and manage user data, to provide access according to the agreements, etc., for a long time period.) repositories. Data creators will upload their data and metadata to a trustworthy repository of their choice and this repository then needs to "plug-in" that data into the global virtual collection. Therefore, we need to distinguish two slightly different "plug-in" scenarios: (1) A new repository has been established and wants to "plug-in" its FDO collection into the global data space. (2) A user uploads a new set of data and metadata to an existing collection of a repository assuming that the repository has proper procedures in place to create FDOs.

This plug-in will work when the repository has clearly identifiable, self-standing, and traceable (Metadata information closely associated with data allows to trace tracing the usage of the data.) digital entities in the collection it is offering which give access to all relevant information in a FAIR and persistent manner, i.e., enabling carrying out validations of FDO framework (The FDO Framework contains currently the specification of what an FDO is, enabling a variety of technical implementations. It can be found here: https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/FDOF; accessed on 7 Novermber 2021) conformity and machine actionability. The repository would have to simply offer all PIDs of its collections via a mechanism similar to ResourceSync [13] allowing software (and human) agents to check what is being offered and compare profiles with selected metadata categories to understand whether something is made available that is relevant for the own research. This can include agents that update indexes of search portals as well as agents that are looking for example for brain image data that are needed to calculate correlations between brain disease phenomena and patterns in measurement data with the help of deep learning. Due to the binding capacity of FDOs, i.e., its PID needs to include all references to all information to make DOs FAIR, every agent in charge of some purpose will find all information about the data that is needed including access rights and licenses.

Of course, the trustworthiness of the procedures applied by the repositories is of crucial importance, since researchers need to rely on the authenticity of the data and the quality of the metadata, for example. These procedures come into play when a user uploads a new set of data and metadata into such a trustworthy repository. A PID needs to be registered, PID record information (A PID is associated with a record containing a set of values instantiating kernel attributes.) needs to be generated according to a profile, privacy regulations need to be respected and turned into access permissions and license conditions, etc. Finally, the repository needs to indicate to the world that it has new FDOs in its collection. With this approach to FDOs, it is also possible that the submitted metadata refers to an extant data structure stored elsewhere. In this case, the repository does not have complete control over the reliability of the data.

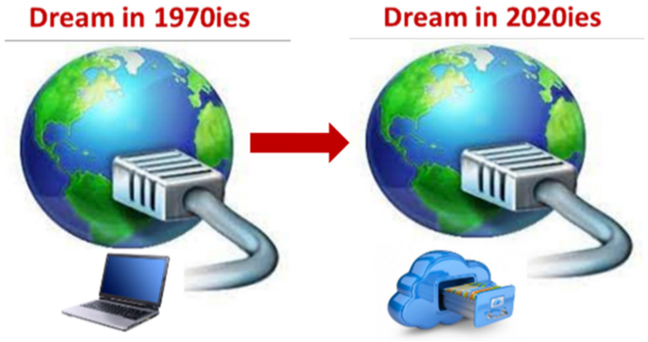

Figure 1. This figure indicates the dreams related to new paradigms. While the dreams in the 1970s of an easy plug-in of any compliant device into the Internet became reality after many years, the dream of a unified data space supporting an easy plug-in of data entities needs to be realized.

2.3. Self-Standing Entities

A third concept to make these complex networked spaces operational is the notion of “self-standing entities”. The revolutionary aspect of the Internet is based on the introduction of “self-standing complete” messages that can travel through the Internet. These messages know where they are from, where they need to go, when they have been submitted, and what their content is. Based on this information, they can now float as datagrams through the Internet and any network device knows what to do with them. This principle replaced the concept of circuit switching, which was used in telephony.

Entities that are going to populate the global data space need to be self-standing in so far as at any moment in time it must be possible to access all relevant information that is required for management or processing purposes assuming that the identifier of a certain entity is known. Ideally, with rich metadata included as the FAIR principles require, metadata should be rich enough to fully decouple creators of data from the users since users persistently find all required information (It should be noted that metadata needs to be provided by the scientific communities and the FDO concept can hardly influence the richness of the provided metadata.). Of course, these identifiers must be persistent as well as the mechanism that can resolve them into crucial information which we will call the set of kernel attributes included in a PID record as defined by the Research Data Alliance (RDA). Clients need to be able to process these kernel attributes which can be referenced, for example, to access all kinds of metadata (descriptive, scientific, access permissions, license specs, etc.) and assuming the correct permissions the bit-sequence of that entity that encodes its content. As with the famous post-box as illustrated in Figure 2 (We are using this analogy to point to the long history of self-standing units although the postbox is static while digital entities can be dynamic.), some clients will be allowed to look at the outside information which in general includes some metadata and an identifier, others will be allowed to look into the content (In the digital domain it can also happen that access to some metadata is restricted which indicates the limitations of the postbox metaphor.).

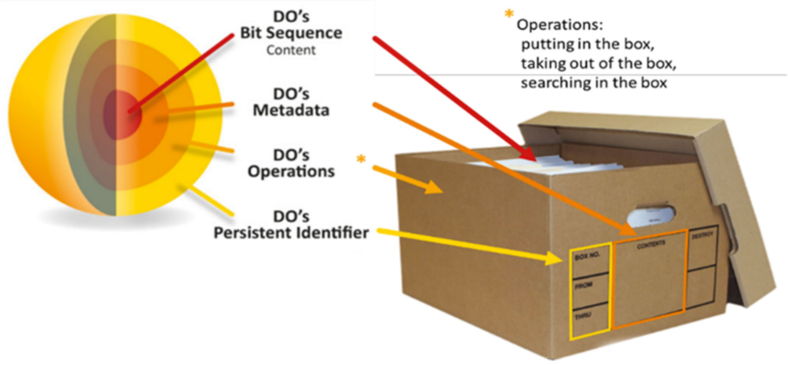

Figure 2. Based on a suggestion from A. Hardisty, this figure shows the similarity between the old post-box sent by postal offices spread across the world and the Fair Digital Object. The post-box has external information (IDs, metadata, etc.) that is needed to tell officers where the box should go to and it has internal information only accessible for the addressee. In similar ways, the FDO contains all necessary information and makes the distinction between external and internal information.

The revolutionary core concept behind FDOs is to have self-standing entities that populate the global data space and inherently bind all relevant information enabling clients to do proper management or processing assuming the required permissions.

2.4. Built-In Security

It is well known that the Internet and the Web were designed without built-in security mechanisms. What may seem today as a gap made the protocols simple and straightforward and guaranteed massive uptake. However, all kinds of mechanisms had to be invented to introduce security, but they are hardly applied systematically. Especially for data but also for other information to be shared this lack of built-in mechanisms is not acceptable anymore given the serious privacy requirements, the interest of researchers to protect their newest findings, the need for industry to protect their values, and other understandable wishes to not participate in full openness.

The FDO approach, and here we need to restrict our explanations to the DO approach, has built-in security at different levels:

- PID records that are so crucial for managing and accessing FDOs are protected using a PKI infrastructure ensuring that only accepted authorities, in general, the owners, are allowed to make changes. It is even possible to prevent access to attributes in the PID record in case that industry for example wants to protect crucial business information.
- A PID should persistently include a hash code in the PID record characterizing the bit-sequence and indicating to every user whether it is indeed the bit-sequence one is expecting and also allowing data providers to look for (unauthorized) copies in the data space (It should be noted that repositories need to state explicitly what their policies with respect to assigning PIDs are and whether they for example allow for mutable DOs where adding checksums does not make much sense.). Some repositories use such a hash code as a suffix of the PID which has the same effect. Tracking of data copying is made possible.
- Most important is that the PID record is always the anchor for dealing with digital artifacts encoded as bit-sequences of an FDO. The PID record includes all references to

relevant information including license terms, access permissions, etc. This means that the rights specifications are persistently bound with the bit-sequence independent where the copies of these may be stored. Therefore, FDOs offer a solution for proper authorization in distributed landscapes where copies are being exchanged and traded (In general, when solutions for Authentication and Authorization Infrastructures (AAI) are being offered this only includes mechanisms for distributed authentication.).

- This could be extended to persistently include a reference to a blockchain that stores transaction events if this is requested by the data provider. This would enable data providers to point to a transaction of a specific object that is characterized among others by a hash value included probably as an attribute in the PID record.
- Of course, interested data providers could extend these measures by adding hidden information into the data stream, for example, using certificates to sign data or using cryptography to increase security. These are all mechanisms changing the bit-sequence which are beyond the FDO mechanisms.

2.5. Summary

We explained that the future data space can be seen as a globally integrated virtual data collection that is populated by self-standing entities called FAIR Digital Objects (FDO) that can easily be plugged in and can be accessed by machines and humans, if access is permitted.

Therefore, the concept of FDOs can be interpreted as a universal mechanism to organize the global data space according to detailed specification standards. If this concept is used systematically, FDOs will build the basis of an interoperable data space. FDOs themselves are not forming an infrastructure, but to make an FDO-based data space operational an infrastructure with a set of basic services such as registries for trustworthy repositories, PIDs, schemas, vocabularies, etc., will be required. The global community will need to ensure that the FDO specifications and the set of basic and distributed registries is free of economic and political interests to make it serve science.

3. Current Initiatives

3.1. Research Infrastructures

In slightly more than a decade, we have observed an increase in funds for domain-driven research infrastructures addressing various aspects of data management, integration, and processing. These investments had many positive effects such as

- Almost all researchers are aware of the FAIR guiding principles.
- A variety of communities can access well-organized repositories, have developed comprehensive metadata schemas and tools useful for their research work, and built a set of supporting semantic artifacts.
- Experts are starting to experiment with workflow frameworks to automate recurring processes.
- More early career experts are being employed who are ready to make use of new technology.

The impact is huge, especially in areas such as biomedical science and neuroscience where funding is much higher than the average. In the European ESFRI (The first ESFRI projects started their work in 2009.) [14] and some nationally supported initiatives, we can observe a trend to more comprehensive frameworks that emerge from new tools. Collaborating researchers and technologists dare to think bigger to cover a higher degree of functionality in one software package, which implies that small functional cells are growing to larger islands which are not only based on some community best practices but also to include a variety of adapters to connect to other "islands of relevance" for that community. Figure 3 illustrates this extension in two exemplary ways. Functional extensions often include repository services, aggregators for metadata and search portals, transformations of data, smart analytic functions, smart visualization, workflow functionality, and support for using high-performance/throughput computers (Hadoop parallelization, Docker containers, etc.).

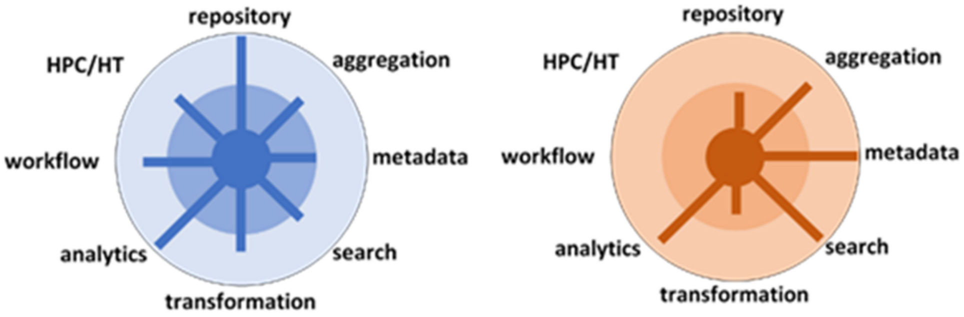

Figure 3. Illustration of two possible extension scenarios emerging from small functional nuclei. In the left diagram, functionality is added in the direction of workflow support and analytics on powerful machines, while the right is directed towards metadata search and analytics on aggregated data.

Such comprehensive packages necessarily include some infrastructural services to make them operational, but these are based on ad hoc decisions or follow industry suggestions such as using "schema.org" for the registration of schemas. Excellent examples for this trend are the Galaxy Workflow framework [15], the Icebear tool [16], the Ebrains tool [17], and the Caos database [18]. Many more examples for such extended packages could be mentioned all evolving from the direct interests of researchers in performing a specific set of tasks. For researchers in all kinds of disciplines, they have a high degree of attraction; however, they will again lead to a plethora of best practices and a growing amount of software code with recurring components and adapters that need to be maintained (A deeper analysis of such software packages would be required to make detailed statements. Due to the many development activities in the different disciplines across various countries this is hardly possible).

After a phase of many small tools which have been developed even in small departments, we can observe a phase where capable teams are extending functional nuclei to larger packages making use of best practices to improve interfacing with other packages. Often these packages are designed as all-in-one solutions to offer a specific set of functionalities. In the case of workflow frameworks often wrappers are provided to integrate other software components.

This phase can perhaps be compared with the phase-in of railway or electrification evolution when individual solutions were combined to regional islands. While the first individual solutions were purely driven by bottom-up motivations and accidental knowledge of some technology, this second phase offers already a mixture between bottom-up and top-down driven motivations. In many cases, the need for functional extensions is combined with comparisons of different technologies and best practices. Nevertheless, the many teams developing such extended packages are working separately and miss an underlying interoperability standard. This can only be introduced by top-down considerations based on abstractions from the many island frameworks. Such top-down approaches could be accelerated by increased support for interdisciplinary research, which naturally requires interoperability among the different infrastructures used by the participating disciplines.

3.2. *eInfrastructures*

In Europe, a distinction is made between research infrastructures which are mainly driven by research communities and eInfrastructures [19] which are meant to offer basic services (CPU, storage, network). At the European level large national compute and data centers or comparable centers of large national research organizations are involved. However, these centers have as the highest priority to serve the interests of their countries or organizations. This imbalance between national interests and cross-country organized research seems to hamper breakthroughs in interoperability and introduce much overhead.

For many years the concept of computing grids dominated European eInfrastructure funding and institutions, not driven by cutting-edge research, were established to ensure funding streams and are still in place. While in the US, IT experts in academia and especially in industry were looking for new concepts to meet the challenges of accessing and processing very large numbers of files and as a result, turned the grid into the cloud concept, the European eInfrastructures were continuing to exploit the grid metaphor. At a relatively late-stage cloud systems were recognized as a reformulation of the grid metaphor. Funding streams were shifted to setting up cloud services, understanding the new technology, and helping to develop knowledge at the side of research infrastructures so that they could make use of these cloud facilities. Little attention was paid to the fact that the cloud concept emerged as a side product of a wider data strategy, i.e., cloud services are still widely misunderstood as tackling the FAIR principles (The use of names such as European Open Science Cloud created much confusion since it was not clear to everyone that the term "cloud" here was meant in metaphorical sense.), for example. However, creating an interoperable data space was not the intention of cloud systems.

At a recent eIRG workshop [20], it became obvious that the eInfrastructures in Europe see the urgency to focus on data which their centers are hosting and which implies taking the FAIR principles seriously.

The innovative impact of European eInfrastructures was limited since the focus was on offering basic services, i.e., addressing the question "where to store" but not the question, "how to store" (It should be noted that in an area of increased high-performance computing needs, steps were made to offer compute cycles at European level for engaged researcher teams.).

3.3. EOSC

The European Open Science Cloud (EOSC) [21] was motivated by the insight that the researcher-driven siloed approaches being carried out in the ESFRI research infrastructure initiatives, did not foster interoperability across disciplines. In contrast, the inherent strengthening of the siloed approaches did not demonstrate a trend towards convergence despite the so-called cluster projects which forced experts from different ESFRI projects within a larger research domain (life sciences, humanities, etc.) to collaborate. Cluster projects are policy-driven and temporal and therefore did in general not lead to sustainable results and changes in organizing the silos. The effort of integrating new concepts and software into the normal default procedures and taking care of maintenance is mostly underestimated.

The starting point for EOSC was the agreement that it should be based on the FAIR principles and its architecture needs to be distributed. This led to an "epidemic" increase of projects that claimed to be FAIR compliant without proving correctness, i.e., the FAIR message was excellent to increase awareness about the proper treatment of data, but until now it had little impact on practices [22]. Due to the European construction, EC and member states had first to construct a governance system for EOSC and to not lose time, some working groups were initiated, which brought together national delegates making it impossible to separate technical and political motivations.

EOSC, therefore, was started with a weakly defined infrastructural core. As indicated, two pillars were identified at an early stage: (1) EOSC should be based on the FAIR principles, and (2) EOSC needs to be distributed. This weak definition, however, can also be seen as a unique chance if leadership manages to focus on abstractions from the many existing disciplines and nationally driven initiatives and separate technical and political arguments. EOSC is not bound by particular interests which would result in silo-based solutions. It has the potential to find a balance between research-driven evolution (often called bottom-up) and abstraction-driven disruption (often called top-down), to lead the path from growing islands to a unified common data space and thus can drive global harmonization (It has been stated clearly by the leadership that some current services such as the "EOSC Portal" are not related to the EOSC process.).

3.4. NFDI

The German National Infrastructure for Research Data (NFDI) initiative is one of the initiatives of European member states that chose an approach that is different from EOSC. An extensive discussion process resulted in a decision to foster research-community-driven infrastructure approaches and to not spend time on abstractions [23]. A first set of 9 projects started in 2020 and a second round of grants will be accepted in 2021 resulting in about 19 projects being active in 2021. In fact, NFDI is widely compatible with the ESFRI initiative, except that the projects are purely national. It was to be expected that some research domains that were already active in ESFRI as national nodes submitted proposals that were motivated by adding functionality and achieving a higher degree of maturity. A great achievement is that even more researchers are busy with thoughts about proper research management and additional research disciplines are at the stage to better organize their digital domain. NFDI is therefore a great invention to support bottom-up driven consolidation of RDM in a variety of research domains. Since researchers are primarily interested in tools facilitating their work and not in standards, there is a risk that siloed mentalities will be solidified at cost of convergence.

It is surprising that no serious discussion of the results of roughly 12 years of ESFRI funding took place, which could have resulted in recommendations to narrow the solution space for RDM in the NFDI initiative. Yet, the national discussions seem to be widely decoupled from international discussions on RDM commons. The consequence is that some projects continue with what they have done before and others seem to start from scratch, i.e., the same discussions about PIDs, metadata, repositories, etc., which have been addressed for years are being started again without guidance. The term FAIR is often used and indicates growing awareness. However, its core of "machine actionability" is hardly understood.

After 2 years of discussions, the need to discuss cross-cutting themes and basic services became more obvious and leadership is now willing to invest. A discussion between interested research communities resulted in a set of themes (https://zenodo.org/record/4593770#.YOqlB0xCTb0; accessed on 15 November 2021) such as research data commons, metadata/findability, terminologies, ethical/legal aspects, training, etc., that should be discussed across consortiums [24]. The themes are broadly described and there is the risk that instead of having a serious evaluation of solutions, those teams that have been active for some years will propagate their solutions. In addition, a discussion about basic services was started. Yet it is not clear how this discussion will be led and whether the potential for evaluations of years of experience and results of discussions in RDA, CODATA, etc., will be used or whether specific interests will dominate.

3.5. NIH Commons

Phil Bourne, now at the University of Virginia, was Associate Director for Data Science (ADDS) at NIH from 2014 until 2017. During his NIH tenure, he proposed to establish a "research data commons" in collaboration with the private sector cloud providers [25]. This proposal focused on issues of "how" to store digital entities including the provision that NIH grants would include "green stamps" that could only be spent with a certified cloud provider. This approach assured a pay-as-you-go model and that monies would not be redirected away from the data infrastructure by awardees. Unfortunately (in the authors' opinion), this effort, at least in the form described, was terminated when Bourne left NIH. The ramifications of these ideas, along with the FAIR principles, have remained at NIH in the reconstituted ADDS office and programs such as NIH Strides. The unified "research data commons" has been replaced with a number of efforts more in alignment with individual institutes and centers than the NIH writ large.

4. Stakeholder Interests

We like to speak about an "eco-system" when the system we are looking at and trying to optimize is utterly complex. Building data infrastructures is such a complex and opaque system. In our early paper, we referred to Hughes who deeply analyzed the development of electrification who indicated that for large infrastructures politics, economy and technology need to find agreements. In the current phase of data infrastructure "piloting" in the research domain, we need to extend this to politics, research stakeholders, and technology. Therefore, it makes sense to look at different stakeholders involved in the research domain.

4.1. Researchers

In general, researchers are interested in how technology can support their research work in such a way that they can get deep insights and produce recognized publications to document their advances. This defines their time horizon, which usually focuses on the next five years. In the case of setting up large research facilities which can serve as the basis for advanced and competitive research longer time horizons are, of course, considered. This implies that researchers are primarily concerned about functionality which is presented by measurement installations, simulation, and analytical capacities and tools. Researchers' primary interest is not in the standards being applied in equipment and software but in functions. In contrast, researchers often are skeptical (for understandable reasons) in discussions about standards since they often imply delays that slow down their daily work.

The analysis of many research infrastructure results and plans, as presented in Jeffery et al., support this impression. Research plans see FAIRness as a goal but shift it to the publication step which implies no change in the daily practices. This is combined with the hope that in a few years data stewards will be able to do the required collection building and screening, create the required metadata descriptions, and do the upload into repositories. At the same moment, as FAIRness is being shifted to the end stage, practices of exchanging files between networks of collaborators are being intensified with well-known inefficiencies.

In this context it is important to note the difference between FAIRness and Open Science: all data should be FAIR and as much as possible should be open. In other words, FAIRness does not require openness. However, if openness is not in the intentions of a researcher (and it is known that many researchers have objections to openness), there is less motivation to make data FAIR.

Summarizing, we can conclude that non-FAIR practices are still continued and Open Science and FAIRness by Design are not seen as broadly accepted goals yet. Any expectation that this will change quickly seems to be illusionary.

4.2. Tool Developers

Several recent discussions with academic toolmakers have in resulted in a few observations:

- Tool developers are, of course, driven by the wishes of their customers and so functional extensions have the highest priority (Of course debugging and software maintenance have highest priority and the costs for this are in general underestimated.).
- Tool developers are partial to their tool and want to keep it alive in an area of dynamic competition. Since software maintenance, in general, is not funded, they are under constant pressure to find funding sources.
- Tool developer plans have short time horizons and therefore it is not surprising that FAIRness and other emerging principles and standards do not have high relevance.
- Relevant tools need to broaden their functionality to be attractive, which implies that infrastructural elements are being integrated leading to an "all-in-one" concept. This is the opposite of how infrastructures should be designed. Often the strict separation between data and operations is not maintained due to short-term design and development considerations, which leads to unwanted dependencies.

To achieve breakthroughs towards common standards, it will be of great importance to convince toolmakers to support new concepts. However, additional funding and strict guidelines will be needed to achieve steps towards convergence. Tool developers in general will not play a driving role towards convergence and standards.

4.3. Industry

With respect to industry, we restrict ourselves to a few statements.

- Engagement of industry will be necessary to make real steps towards convergence. As in the case of the Internet standards, some form of public-private partnerships will be needed.
- Existing big data industry is not interested in new open standards, just as in the Internet case, since it would influence their market position where proprietary solutions dominate.
- New companies would have to emerge that see potential in new standards and are ready to take risks and/or the open standards would have to emerge from the science community (as in the Internet case) and then be taken up by such new companies.

In Europe, the industrial GAIA-X initiative has been started with a focus on setting up a competitive cloud system strategy, and it will require highly personal and funding investments to achieve the level of sophistication that the big market leaders already have. This means that GAIA-X will have to focus on the "where to store" question and will have to restrict their investments in the "how to store" question. In the meantime, interactions pursuing joint programs between GAIA-X and EOSC and NFDI have been started. There is potential in such collaborations, but also risk in that pure engineering solutions may dominate the discussions, since the pressure for competitive services will be enormous. This could leave little space to contribute to convergence in data sharing and management.

4.4. Policy-Makers

Policy-makers at different organizational levels are natural allies in harmonization and standardization since they have the potential to more effectively use the available funds and convince, for example, parliaments about the need for expenses. Twelve years of intensive work in ESFRI initiatives, and this was confirmed by a recent deep study of about 75 research infrastructure reports and plans, have shown that many challenges for the infrastructure work in the different research communities are identical if one dares to abstract from some flavors which can be addressed in different ways. Yet, the solutions chosen are heterogeneous which leads to some of the costly inefficiencies when integrating digital objects from different silos.

5. Conclusions

There should be no doubt that the data infrastructure revolution we speak of will happen, but that the transformation to a state where convergence standards are broadly accepted and applied will occur stepwise dependent on the state of infrastructure building in the different communities and take more time than we may have hoped. However, when we speak about the FAIR principles and the FAIR Digital Objects as candidates for this convergence, we are confronted with three challenges.

First, we need to spell out the vision more clearly. This vision is of a globally integrated data space, which is populated by clearly identifiable and self-standing digital entities that can be accessed and interpreted by anyone who has the right to do so. This virtual space is structured by trustworthy repositories that have the task to care about persistence and proper protection mechanisms. From a computer science perspective, however, it does not matter where these entities which we call FAIR Digital Objects exactly are being maintained. Some colleagues still see the FDOs as a competitive technology, but this view is wrong. FDO is a set of specifications that need to be owned by the FDO Community to prevent piracy and to ensure that the application of the specifications are as free and open as the specifications of the Internet are. The FDOs are simply a set of specifications that translate the FAIR principles into practice, and which can be implemented by different technological approaches. Finally, we need to overcome the current inefficiencies and costs of data-driven science and reduce the barriers and that is what FDOs are promising to help achieve.

Second, we believe that a level of abstraction needs to be added to globally integrate the data space. Such work, however, will not be motivated by the researchers. When a colleague taking some responsibility in guiding a large funding program recently argued that "these discussions about abstract interoperability levels" are not useful at this moment, we understand the intentions behind it, but we also see its limitations. Both approaches, the short-term evolution driven by the concrete needs of research communities and the strategic choices driving convergence, will be necessary to implement the vision. Initiatives such as EOSC, due to their "weakness" in specifying goals until now, which could at first glance be seen as a disadvantage, could be turned into an advantage if the management comes to strategic decisions involving the described visions.

Third, we were looking at the basic motivations of different stakeholders involved in current harmonization processes to understand from whom we can expect impulses. It is obvious that we will need the acceptance of researchers and the participation of software developers to come to effective changes. However, with some exceptions, both stakeholder groups will respectively need to focus on short-term goals with a limited time horizon. Building revolutionary infrastructures, however, requires a longer horizon. Industry is experiencing huge inefficiencies which are expressed in costs. While some global players are pushing their proprietary methods, creating dependencies, and hampering innovation, others such as the manufacturing industry (after some years of discussions) are now ready to develop open standards such as OPC UA. Such efforts are huge and will pay off in the long run by reducing the costs to integrate machines and software from different producers. Therefore, we argue that we can learn from industry and should not hesitate to look for collaboration.

In conclusion, we state again that the major roadblocks ahead of us to build this globally integrated data space are more social and political than technical. Wise decision-taking at this moment for us means to follow both tracks in the research world: (1) Invest in creating awareness and acceptance by programs close to the research needs and (2) by funding implementations of testbeds, reference architectures, and demonstrators of FAIR Digital Objects. With the latter, keeping in mind that multiple implementations will be the norm, as it was with the Internet.

Author Contributions: Conceptualization, P.W. and G.S.; methodology: P.W. and G.S.; investigation, P.W. and G.S.; writing—original draft preparation, P.W. and G.S.; writing—review and editing, P.W. and G.S.; visualization, P.W. and G.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

References

1. Wittenburg, P.; Strawn, G. Common Patterns in Revolutionary Infrastructures and Data. *B2SHARE Arch.* **2018**. [CrossRef]
2. 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]
3. De Smedt, K.; Koureas, D.; Wittenburg, P. Analysis of Scientific Practice towards FAIR Digital Objects. *B2SHARE Arch.* **2019**. [CrossRef]
4. FDO Forum. Available online: https://fairdo.org/ (accessed on 1 November 2021).
5. Harari, Y.N. *Homo Deus*; Harvill Secker: London, UK, 2016.
6. Floridi, L. *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*; Oxford University Press: Oxford, UK, 2014.
7. Strawn, G. Open Science, Business Analytics, and FAIR Digital Objects. In Proceedings of the 43rd Annual Computer Software and Applications Conference (COMPSAC), Milwaukee, WI, USA, 15–19 July 2019. Available online: https://ieeexplore.ieee.org/document/8754334 (accessed on 15 November 2021).
8. Int. Data Space. Available online: https://internationaldataspaces.org/ (accessed on 1 November 2021).

9. Kahn, R.; Wilensky, R. A framework for distributed digital object services. *Int. J. Digit. Libr.* **2006**, *6*, 115–123. [CrossRef]
10. DOIP. Available online: https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf (accessed on 1 November 2021).
11. Linked Data. Available online: https://en.wikipedia.org/wiki/Linked_data (accessed on 1 November 2021).
12. Hughes, T. *Networks of Power*; Johns Hopkins University Press: Baltimore, MD, USA, 1983.
13. ResourceSync. Available online: http://www.openarchives.org/rs/toc (accessed on 1 November 2021).
14. ESFRI. Available online: https://www.esfri.eu/ (accessed on 1 November 2021).
15. Galaxy. Available online: https://galaxyproject.org/ (accessed on 1 November 2021).
16. Icebear. Available online: https://www.icebear.fi/ (accessed on 1 November 2021).
17. EBRAINS. Available online: https://ebrains.eu/ (accessed on 1 November 2021).
18. CaosDB. Available online: https://caosdb.org/ (accessed on 1 November 2021).
19. eInfrastructures. Available online: https://ec.europa.eu/programmes/horizon2020/en/h2020-section/e-infrastructures (accessed on 1 November 2021).
20. eIRG Workshop. Available online: https://indico.fccn.pt/event/15/overview (accessed on 1 November 2021).
21. EOSC. Available online: https://www.eoscsecretariat.eu/ (accessed on 1 November 2021).
22. Jeffery, K.; Wittenburg, P.; Lannom, L.; Strawn, G.; Biniossek, C.; Betz, D.; Blanchi, C. Not Ready for Convergence in Data Infrastructures. *Data Intell.* **2021**, *3*, 116–135. [CrossRef]
23. NFDI. Available online: https://www.nfdi.de/ (accessed on 1 November 2021).
24. NFDI Cross-Cutting Topics. Available online: https://zenodo.org/record/4593770#.YT8VOX1CTb0 (accessed on 1 November 2021).
25. NIH Commons. Available online: https://era.nih.gov/ (accessed on 1 November 2021).