Challenges for Putting FAIR into Practice

Peter Wittenburg
Max Planck Computing and Data Facility
Max Planck Society
Munich, Germany
peter.wittenburg@mpcdf.mpg.de

***Abstract*—*The understanding is growing that the emerging integrated and interoperable data domain (I2D2) will have far reaching consequences for society, research and economy comparable to the changes caused by the Internet, for example. Such large infrastructures have a global dimension and require global agreements which in general are simple pre-competitive standards. Despite the FAIR Principles and the concept of FAIR Digital Objects (FDO), it seems that we are still far away from agreements on convergent specifications if we for example look at practices in the data labs. In this paper we describe the evolution of the FDO concept and point to the crucial role of global, unique, persistent and resolvable identifiers as basis for FDO.*

Keywords*—*Data Management, FAIR data, Digital Objects

I. Introduction

In 2018 Wittenburg and Strawn wrote a paper with the title "Common Patterns in Revolutionary Infrastructures and Data" [Wittenburg 2018] in which they found common patterns in the evolution of a few large infrastructures such as electrification, Internet and the Web. As figure 1 indicates, an early vision is taken up by an increasing number of people that explore the landscape of possible solutions. This creolisation phase leads to many different suggestions, testbeds and implementations increasingly lacking coherence and creating interoperability challenges. As consequence, the wishes to improve convergence get into the focus of developments leading to some attractors which are then evaluated and discussed at various levels such as by technologists, economists and politicians. Finally, after some time of debates there are decisions about convergence such as 50 Hz AC for electricity transmission, TCP/IP for Internet message exchange and HTTP/HTML for the Web information exchange. More examples can be mentioned from history such as for railroads and telephone systems. These mostly simple core standards lead to convergence which people could then rely on and build applications which led to enormous exploration waves. These simple standards reduce complexity and are pre-competitive, i.e. they have the chance to be globally accepted.

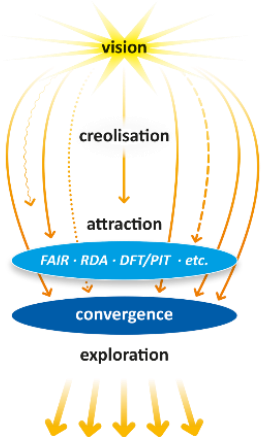


Figure 1 describes the typical pattern found in the evolution of large infrastructures – from an early vision to an utterly dynamic exploration phase

Wittenburg and Strawn applied this pattern now to the area of data since there is no doubt that we will need a globally Integrated and Interoperable Data Domain (ISD2) to make use of the value of the increasing volumes of data that will become available for the benefits of society, research and economy. They concluded that with the publication of the FAIR principles [Wilkinson 2016] and the specification of the FAIR Digital Objects [DFT 2016, Paris 2019] implementing the FAIR principles enormous steps have been made towards a convergence for building a global I2D2. Of course, from their study it was obvious that it often takes decades between launching a vision and coming to agreements on convergence. One reason for the delays is to find in the possible impact and political relevance of such infrastructures, i.e. different stakeholders want and need to have a saying in this process.

The assumption was that with the FAIR Principles and the FAIR Digital Objects basic elements of a future I2D2 are available and would accelerate the discussion about convergence finding. However, this agreement cannot yet be seen for FDOs, although there is broad agreement on the FAIR Principles. However, principles are not blueprints for building infrastructures and different interpretations of the FAIR principles already emerged.

In chapter II I will relate the hopes on fast agreement finding with the reality in many data labs. In chapter III I will elaborate on the development of the concept of Digital Objects which was recently extended to FAIR Digital Objects (FDO). In chapter IV I will briefly explain the crucial role of identification and in chapter V draw some conclusions.

II. Data Practices

It is well-known that about 80% of the time in data projects is spent on efforts related to what is called data wrangling and these inefficiencies seem to be in the same order in all sectors [Wittenburg 2018]. Data wrangling includes all steps that are necessary to be carried out before one can start the analytics. From industry it is known that about 60% of data projects fail which to a large extent is devoted to underestimating the costs of data wrangling. Given my own experience in a Max Planck Institute which from its beginning was data-driven and also dependent on data from other institutes I would claim that in the research domain the failure rate is comparable. The high costs associated with data intensive research have as consequence that many smaller departments and individual researchers are widely excluded from data science, since a technical support staff and student assistents that carry out all the data wrangling are needed. In many institutes PhDs need to carry out this basic work as part of their thesis work.

Recently, we had the chance to analyse about 70 research reports in great detail and we could identify a number of paradoxes which illustrate some additional challenges [Jeffery 2021]: (1) Researchers have already heard about the FAIR principles and support their basic messages. However, they like to shift making digital objects FAIR to the end of the project, since this allows them to continue in the labs with what they are used to and to not suffer from disruptions. We know however, that this does not work well and that the costs for delayed curation processes are by factors higher than for immediate actions [Beagrie 2008]. (2) Many researchers support the open science principle, however, only make commitments about the data that is being associated with publications. Since more than 90% of the data being created is being reused in the processes in the data labs, this has the consequence that much data will not be part of open sharing. We should note here, that of course data is already being exchanged, but often without being FAIR, i.e. people exchange files without submitting identifiers and metadata. (3) The statement of G. Strawn that "standards are good for science, but not for scientists" was confirmed [Strawn n.d.]. In general, researchers are not really interested in standards, but in tools that allow them to address their challenges at that very moment. Therefore, they like tools and of course are happy when tools support some useful standard. Research organisations have a different attitude towards standards, since they will at the end reduce inefficiencies and thus lead to more results for the investments made. (4) The number of excellent tools which is being used across all the labs is increasing, but this does currently not increase interoperability between the data silos. There are so many smart young researchers and developers who just take any new technology available and apply them. Some of these tools incorporate some workflow steps and increase local efficiency. (5) Discipline experts believe that their methods and processes are unique. Comparing processes, however, across disciplines indicates that there are many recurring patterns that could be subject of automation by prefabricated workflows. (6) In modern data science different researchers and institutions are involved in the processes due to different types of expertises. In such situtations responsibility for making data FAIR is not clarified and thus shifted between the actors. (7) DMPs are now seen by funders as a tool to improve data practices. But in general researcher see it as bureaucratic act which may help at the very beginning of a project to clarify basic needs. Experience shows that to a large extent DMPs are not useful. This may change when they would be pro-actively supported by data stewards.

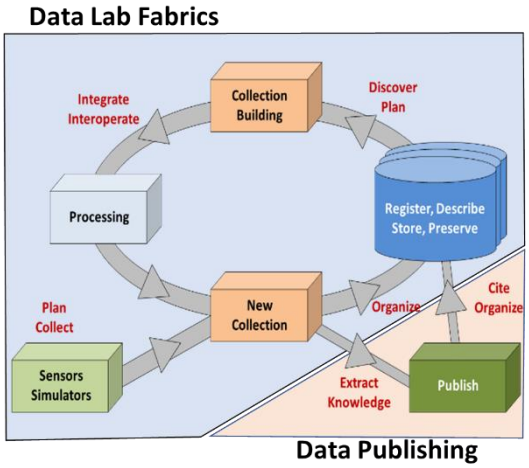


Figure 2 points to the typical data processing cycles in data labs where researchers organize new and existing data stored in repositories to collections to carry out some processing. Often these cycles are repeated until results can be produced that show evidence that will lead to a final publications often with large delavs.

Summarising, I can state that the recent study widely confirms the results which we gained in 2014 in the RDA EU project [Stehouwer 2016] which led to building the RDA Data Fabric Interest Group (DFIG)[49]. The DFIG had as goal to not analyse and optimise the final step of data publishing, but to look at the processes in the data labs for two reasons: (1) The publishers and librarians are already highly active to optimise the data publishing processes as indicated in the low-right corner of the diagram. (2) If we ever want to realise Open Science (or FAIRness) by Design [BRDI 2018], i.e. from the beginning of projects, we need to optimise data practices in the labs as indicated in the centre of the diagram which is of course a much more complex task. But it will be the only way to make data practices more FAIR and efficient.

After long discussions the DFIG came to the conclusion that the concept of Digital Objects (DO) which later became FAIR Digital Objects (FDO) will be the most promissing way in the long-term to change practices based on smoothly integrated data

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

infrastructures. Therefore in the next chapter I will focus on the evolution of the concepts of DO and FDO.

III. (FAIR) Digital Objects

A. *Digital Objects*

When R. Kahn started the Internet and invented TCP/IP it was obvious that the messages being exchanged at the TCP/IP level are meaningless. Meaningful messages that will be exchanged between two centres are chopped into small pieces in the sending centre, transmitted and aggregated on the receiving side again. Already when the Internet

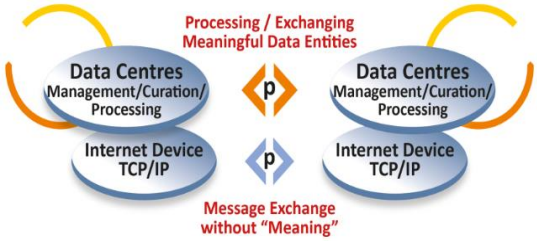


Figure 3 indicates the two layers of information exchange using the Internet. At the TCP/IP layer in general chunked and thus meaningless messages are exchanged. At a higher-level, meaningful messages must be exchanged such as files, emails. messages, etc.

was born it was obvious to the designers that there need to be unifying mechanisms at the application level. We have seen that in the early days for example FTP was designed to transmit files and SMTP as protocol to exchange emails. Roughly a decade later T. Berners-Lee invented the pair HTML and HTTP as mechanism to exchange "web-pages" on top of TCP/IP. Web-pages are written in HTML and HTTP understands HTML encoded information streams. The Web is using URNs as identifiers which come in two forms: URNs as used in some national libraries and URLs which are generally used for all kinds of information.

Roughly at the same time R. Kahn and his team developed the Handle System[50] to have a means for global, unique, persistent and resolvable identifiers (PID). Handles are independent of any technology like an ISBN is, while URLs always encode semantics such as ownership and location and are dependent on the HTTP protocol. The publishers realised the fundamental difference between URNs and Handles at an early phase and soon defined DOIs[51] which are basically Handles with a prefix 10 combined with a specific business model. Instead of web-pages the Handle resolver returns structured data which can be interpreted by machines.

At the same time (around 2000), first labs dealing with large amounts of data started using Handles. My own team decided to follow that path, i.e. in our repository with about 80 TB of organised data in 2010, all stored data and metadata items have assigned a unique Handle. Other repositories with even more data took the same step and in 2014 the Max Planck Society, for example, decided to run a persistent Handle services for all its institutes and researchers. Motivated by the uptake of the Handle System, in 2005 R. Kahn and R. Wilensky revised their early paper from 1995 on digital objects [Kahn 1995, Kahn 2006]. It was the first time that the term Digital Object was coined to indicate the items that were being exchanged via the Internet protocols. A Digital Object can contain bit-sequences of any type: data, metadata, software, assertions, etc. DOs therefore are the most abstract definition of the content that can be transferred. Due to the great success of the Web which was soon be used for all kinds of applications, the notion of Digital Objects was widely forgotton although it exists per definition in the term Digital Object Identifier (DOI).

[50] https://www.handle.net/
[51] https://www.doi.org/

In 2013 when the research Data Alliance[52] was set up one of the first groups that was established was the Data Foundation and Terminology working group co-chaired by the author[53]. Based on many use cases from various disciplines we ended up in defining the Core Data Model [Berg-Cross 2016] which is an extended version of the DO model as introduced by Kahn & Wilensky. The RDA DFT Digital Object model states the following: (1) Each DO has a structured bit-sequence encoding its content. (2) Each DO is assigned a PID and associated with metadata (can be of different types from type, descriptive, provenance to rights and transactions). (3) Each metadata description is a DO in itself. (4) DOs can be aggregated to Digital Collections which are also DOs.

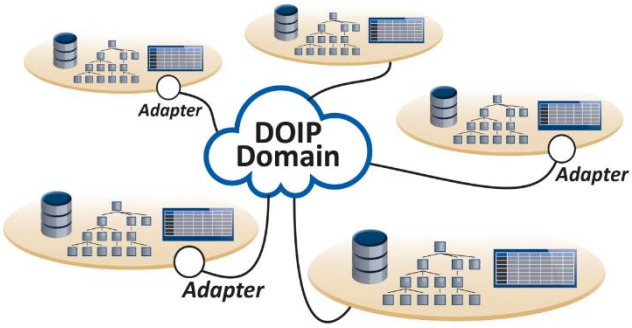


Figure 4 describes the emerging landscape of data repositories and other data providers/consumers all using different technologies and data organizations. A domain integrated by the Digital Object Interface Protocol would reduce complexity substantially, opening the path to an integrated data domain.

Two other RDA groups are of relevance it defining the core of DOs. The PID record can contain a number of Kernel attributes[54] which are returned during the resolution step, and which need to be typed and registered to make the resolution result fully machine actionable. The RDA Kernel group defined a set of attributes which are often being used and they were registered in an open Data Type Registry instance. The RDA Data Type Registry group defined a model for defining and registering Kernel types[55]. Note that the DO definition does not make any specifications about the nature of metadata.

A. Extension to FAIR DO

In 2019 experts from RDA Data Fabric, RDA GEDE[56] and GOFAIR[57] started interacting about the FAIRness of DOs [Schultes 2018]. FAIR implies machine actionability of data and metadata. As indicated above, the definition of DOs makes recommendations about typing PID attributes, but does not strictly require their definition and registration which needs to be changed. In addition, it was found that a linear registry of types might not be sufficient at the end to cover the complexity of the domain of digital object types. Therefore, an ontological approach was suggested. One aspect is clearly underspecified: metadata standards are in the hand of the research communities and most of them spent much effort during the last decades to specify their metadata schemas and concepts. Most of these specifications do not meet the criterium of machine actionability, but communities will hesitate to adapt their standards quickly. At the Paris workshop the FDO Framework was specified which should be the basis of all FDO discussions[58].

Summarising, we can state that the difference between DOs and FDOs are as follows: (1) The DO concept does not strictly request typing of all PID Kernel attributes, while the FDO concept does. It is recommended to use RDF compliant type specifications. (2) Over time the current linear type registry should be extended by a more complex ontology. (3) The DO model does not make strong requirements about the metadata provided by communities, while the FDO concept requires machine actionability. However, it will take time to meet this criterium since communities need to be convinced to do adaptations. Therefore, the DOs/FDOs including their specifications of PID Kernel Types and about Data Type Registries are widely FAIR compliant. Yet the DO/FDO domain is lacking a systemic implementation appraoch.

[52] https://rd-alliance.org/
[53] https://www.rd-alliance.org/dft-work-group
[54] https://www.rd-alliance.org/groups/pid-kernel-information-profile-management-wg
[55] https://www.rd-alliance.org/groups/data-type-registries-wg.html
[56] https://www.rd-alliance.org/groups/gede-group-european-data-experts-rda
[57] https://www.go-fair.org/
[58] https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/FDOF

B. Integrative (F)DO Domain

One of the big sources of inefficiencies in the data domain is the variety of technologies and organisations schemes of data and metadata in the thousands of existing and emerging repository systems. In figure 4 repositories with cloud, file and database systems are indicated and there are different variants of them. Much more problematic are the differences in data organisation, i.e. while data is often stored in files for example, different types of metadata (descriptive, scientific, rights, transactions, etc.) are stored in a variety of different containers and mostly without machine actionable relations between the different entities. This is the reason why in daily practices still mostly only files are being exchanged and the metadata is lost or forgotten on the way, which implies that reuse is often associated with time consuming wrangling.

Figure 5 indicates the referencing structures used for the domain of books. The ISBN numbers point to some metadata representing the "book as a work". Separated from this are local catalogues that refer to shelf locations where the "printed books" can be found.

Digital Objects as well as FAIR Digital Objects offer now a completely different opportunity since they can be used as common glue to achieve interoperability. In the world of FDOs the Digital Object Interface Protocol (DOIP)[59] acts as an interoperable gateway like TCP/IP acted as a gateway between the different network types that had been developed beforehand. What does DOIP solve? It reduces complexity from N*N to N*1 as TCP/IP did for networking since one protocol is now sufficient to interconnect the thousands of data repositories and clients. It puts responsibility for the integration to each repository (or other DO service) and is relying on persistent identifiers which may turn out as one of the few salient corner stones in an utterly dynamic scenario. In addition, it opens the path for service providers to stepwise make their data organisation FDO compliant implying that the connector will then be trivial. For any large data infrastructure such as EOSC and NFDI, FDOs therefore offer big chances to create the Integrated and Intereoperable Data Domain (I2D2) which we are dreaming of. Of course, DOIP is not a protocol addressing, for example, all the challenges related with semantic cross-walking, it is just the stable basis for the complexity of the future global data domain.

IV. Note on Persistence

V. Cerf is warning for the Dark Digital Age[60] we can enter when we will not take appropriate measures. There is much talk about making data more persistent, however, it is mostly not spelled out whether we speak about strategies for 10 years, 100 years, or even longer time periods. In this paper I do not want to elaborate on this aspect implying long-term data curation which will not be trivial to achieve. Here I want to focus on an aspect which is often overlooked: the stability and long-term persistence of the relations between digital objects which we are creating manually or increasingly often automatically.

Research infrastructure projects such as DISSCO [Koureas 2019] in the area of biodiversity indicate the challenges we will have: they deal with 1.5 billion specimens – now as digital twins – from about 500 different natural history museums. Each digital specimen is part of different contexts such as classifications, relationships to other specimens, is associated with different kinds of observations such as photos, gene sequence data etc. Therefore, the number of relationships for each of these specimens is estimated to be around 30 times as big, i.e. we speak about roughly 50 billion relationships which are important to understand the details of specimens. It would be a disaster when these relationships would be lost, since they incorporate the accumulated research knowledge about nature in the digital age where we cannot go back to paper anymore.

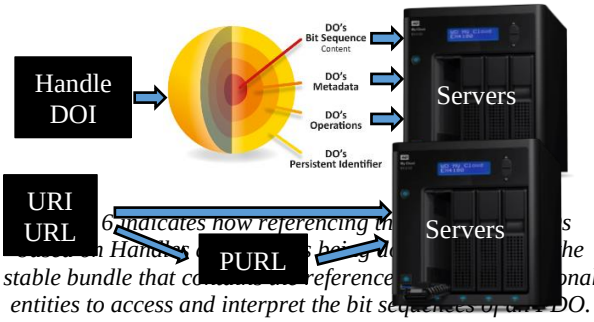
The aspect of long-term stability of relations is widely ignored in our discussions. Librarians and publishers have addressed this issue when preparing themselves for the digital age. They first created the ISBN numbering system which makes a difference between the "book as a work" and the "printed book on a book shelf" (Figure 5). The domain of the "books as works" is nicely separated from the book on shelves, since the first has to be persistent while the latter is ephemeral.

[59] https://www.dona.net/specsandsoftware
[60] https://www.theguardian.com/media-network/2015/may/29/googles-vint-cerf-prevent-digital-dark-age

With Fair Digital Objects we apply an equivalent 2-step strategy. Global, unique and persistent identifiers that resolve associated Kernel Attributes to machine actionable metadata point to all essential information components of an FDO (Figure 6). The attributes will then contain path and other crucial information that is needed to access and interpret FDOs content. This principle is also applied by the publishers which request to associate a DOI issued by the service providers organised in the International DOI Foundation to any electronic publication. The rational behind this is that the publishers will take care that the DOIs based on the Handle System will be persistent. It should be noted here that "persistence" is not only a characteristic achieved by technological choices, but mainly a community responsibility and effort.

To support data labs where the usage of DOIs is not the prefered choice currently more than 3000 Handle services[61] have been established worldwide. It is up to the community behind those services to guarantee persistence. In several countries national data providers and in some large research organisations data centres have taken responsibility to run such services[62]. Handles/DOIs have the capability to include references to an unlimited number of copies of the various components, however, requesting an effort from the remote repositories to change update information in case of changes. This will only work out if the processes in repositories will be automated.

Updating values of attributes associated with PIDs is a highly sensitive operation, since wrong code or operations could destroy all crucial information. Therefore, in the case of the Handle System such operations are highly protected. Each record has a clear owner and for management operations a public key infrastructure (PKI) system is being used to protect records against unallowed access.



6 indicates how referencing th… s … he stable bundle that co… reference… onal entities to access and interpret the bit se… DO.

Figure 7 indicates how referencing in the Web is being done. In general, references occur as URLs encoding ephemeral semantics such as location. PURLs compensate for this deficit.

In the case of URIs, which in general are URLs, the mechanism is different. URLs point directly to locations at specific servers, i.e. they include location information that will change over time and is therefore hardly persistent. To cope with this deficite the Web community invented the PURL system, for example, which allows an indirection step (Figure 7). As in the case of Handles/DOIs repositories are responsible to submit updated information.

However, a PURL cannot cope with multiple copies which is increasingly important in our digital domain for several known reasons and with the association of types and other attributes. The Web community has addressed this gap by suggesting to use Signpost[63] (Figure 8). Ideally a URL points to a PURL which then redirects to a structured HTML landing page which contains standardised and thus machine actionable attributes, and therefore can direct machines to other references. With this mechanism basically the same goals are being achieved as by using standardised kernel attributes in the Handle/DOI case. The major differences can be described as follows: (1) Handle resolution offers immediately structured information while in the Signpost case another indirection step is required

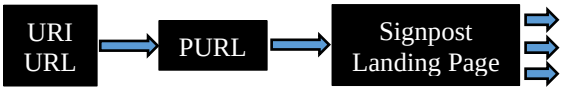


Figure 8 indicates another deficit in the Web-protocol stack which is the lack of a standardized and machine actionable information where different information about a digital object can be found. Signposted landing pages are meant to overcome this gap.

[61] It should be noted that the differences between DOIs and other types of Handles can be found mainly in the differences between the fee structure, the degree of flexibility, for example, in assigning kernel attributes and performance considerations.
[62] https://www.pidconsortium.net/
[63] https://signposting.org/

and HTML needs to be parsed. (2) The directly implemented security mechanisms for Handles/DOIs are stronger compared to those for web-pages.

V. Conclusion

In this paper we started describing our hopes that the FAIR Principles and the FAIR Digital Objects are the needed attractors to come to an Integrated and Interoperable Data Domain (I2D2) which is highly needed to come to a new stable situation for digital objects and not take the risk to enter a Digital Dark Age. I still believe that this is the way to go, however, we still seem to be far away from fast agreements on a core for the global I2D2.

I described then the situation in the data labs where the mass of data is being managed of which only a small amount will be associated with publications (<10%). Not so much has been changed yet with respect to data management efficiency and interoperability despite that researchers are provided by increasingly powerful tools to carry out their research. In general, researchers are not interested in standards that may have a long-term relevance. Due to the pressure on them to show relevant research results they will focus on solutions that are available on short term. Impulses for innovation, if they include the risk of disruptive phases, need to come from other stakeholders.

I then explained why we believe that FDOs can be the solution we are looking for to implement FAIR data and that major components such as, for example, the Digital Object Interface Protocol and the Data Type Registry are ready to be used. Some communities have started to make use of the FDO concept. However, we need to admit that (a) stricter specifications and (b) reference implementations with a reasonable size are yet missing. Only these latter will convince policy makers that the suggested approach will work and scale. Many other voices especially from an IT side can be heard that are not convinced about the FDO approach. Some colleagues believe that a stepwise improvement of the existing service landscape will remove the critical roadblocks on the way towards an I2D2, overlooking that services populate an infrastructure but do not define an integrative core. Others argue that we should leave leadership to big cloud companies observing the huge amounts of investments they currently do, overlooking that cloud systems do not solve the FAIR challenges and that big companies are not interested in achieving an open data domain. Again, others argue that the Web does all we need without knowing even the details as suggested by, for example, Signpost and overlooking speed and security aspects.

Also, we see that the FDO user community is currently fragmented. Various groups and initiatives are trying out the FDO concepts relying on Handles/DOIs for stable referencing but yet do not have an effective global forum to exchange experience, to increase power to achieve necessary changes in the service landscape and to work on additional specifications for standardisation. The RDA Data Fabric IG is focusing on FDOs and is an excellent platform for exchange, however, it lacks the power to set standards and push developments. In this respect the FDO community needs to take urgent action.

Acknowledgment

I need to thank all those who contributed until now to the specification of Digital Objects and here in particular R. Kahn for his early papers and inspirations, a variety of RDA groups, the GEDE group with many excellent experts from research infrastructure projects and those who contributed to the Paris and other workshops[64].

References

Beagrie, N. (2008). Keeping Research Data Safe - JISC Research Data Digital Preservation Costs Study. APA-Conference Budapest.

Berg-Cross, G., Ritz, R., Wittenburg, P. (2016). DFT Core DFT Core Terms and Model. http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318

Committee on Toward an Open Science Enterprise - Board on Research Data and Information (2018). OPEN SCIENCE BY DESIGN - Realizing a Vision for 21st Century Research. https://www.nap.edu/read/25116/chapter/1

Jeffery, K., Wittenburg, P., Lannom, L., et.al. (2021). Not Ready for Convergence in Data Infrastructures. Data Intelligence. Vol.3:1. https://doi.org/10.1162/dint_a_00084

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

Kahn, R., Wilensky, R. (1995). A Framework for Distributed Digital Object Services. http://www.cnri.reston.va.us/k-w.html

Kahn, R., Wilensky, R. (2006). A Framework for Distributed Digital Object Services. https://www.doi.org/topics/2006_05_02_Kahn_Framework.pdf

Koureas, D., (2019). DISSCO – Distributed System of Scientific Collections. https://github.com/GEDE-RDA-Europe/GEDE/blob/master/FAIR%20Digital%20Objects/Paris-FDO-workshop/GEDE_Paris_Session%202_%20Koureas.pptx

Herzcog, E., Mons, B., Lannom, L., et.al. (2019). Report Paris Meeting on Moving Forward to Data Infrastructure Convergence. https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/Paris-FDO-workshop

Schultes, E., Wittenburg, P., (2018). FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure. In DAMDID Conference on Data Analytics and Management in Data Intensive Science, 2018.

Strawn. G., personal communication (see also Jeffery 2021)

Stehouwer, H., Wittenburg, P., (2016). RDA Europe : Data Practices Analysis. http://hdl.handle.net/11304/6e1424cc-8927-11e4-ac7e-860aa0063d1f

Wilkinson, M.D., Dumontier, M., et. al. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. Scientific Data, Vol. 3, Article Number 160018

Wittenburg, P., Strawn, G. (2018). Common Patterns in Revolutionary Infrastructures and Data. http://doi.org/10.23728/b2share.4e8ac36c0dd343da81fd9e83e72805a0