International FAIR Digital Objects Implementation Summit 2024

Extended Abstracts

https://doi.org/10.52825/ocp.v5i.1194

Published: 18 Mar. 2025

Layered Dataspaces in GIDS

Peter Wittenburg[1,*], George Strawn[2], Christophe Blanchi[3], Daan Broeder[4], and Ulrich Schwardmann[1]

[1]GWDG, Germany

[2]National Academy of Sciences, Board Director, US

[3]DONA, Switzerland

[4]CLARIN, Netherlands

*Correspondence: Peter Wittenburg, peter.wittenburg@mpi.nl

Abstract. The Global Integrated Dataspace (GIDS) will come to finally reduce the costs for data-driven work substantially. It will be based on a minimal standard which will be transparent to rights on data, but will nevertheless transport usage information to additional federation layers which are now called dataspaces. Partners in dataspaces may agree on terms such as roles, their rights to use data, etc. and they will apply some dataspace technology that will control the usage of the data and metadata. For the basic minimal standard the FAIR Digital Objects are suitable candidates. For advanced usage control technologies such as the IDSA Eclipse Dataspace Connector might be used.

Keywords: Data Management, Global Dataspace, FAIR Digital Objects

1. Introduction

The Global Integrated DataSpace (GIDS) will come, we just do not know how it will exactly look like and how the different interests will be mapped to regulations and technologies. Both recent success stories, the Internet and the Web, started with the simplest assumptions, i.e. they did not include rights issues in their basic protocols, but shifted it to the application layer. This decision was wise, since including rights is always associated with “federation agreements” which need to respect different legislations. This paper discusses whether the GIDS based on FDOs [1] should chose the same approach.

In the research domain there is now a global agreement on the Open Science principles which includes Open Data. In the same way as open printing stimulated research decades ago, open data will stimulate data driven research which is already now the default case in many research institutes. However, we cannot ignore those sectors and domains where data will be protected, as for example with personal data, commercial data, clinical data, etc. Companies which are creating data may want to trade it, as any other good[1]. Therefore, it is not surprising that large initiatives such as the International Data Space Association (IDSA) [2] and the Industry 4.0 [3] worked out comprehensive reference architectures which both were started in industrial environments and therefore include data protection and data sovereignty as core requirements of their design.

[1] Here we do not want to address the unsolved issues on rights of data where citizens create data and companies claim rights.

2. GIDS Layers

We believe that there should be only one basic integration layer as in the cases of the Internet and the Web, since every other solution would create silos of information. This integration layer needs to be transparent to all specific legislations and agreements on rights, roles and also technologies. In figure1 (left), we indicate the usual situation as it is given for any large repository managing valuable data: (1) It needs to follow the regulations of its organisation embedded in a national legislation (green), (2) it may also be embedded in regional collaborations where additional regulations may apply (red) and (3) increasingly often it will be participating in global collaborations where again different agreements may apply (blue). To satisfy all requirements we need to define a framework which is transparent to all these differences in federation agreements which are embedded in different legislative systems. In addition, all these collaborations often make statements on technologies, so that a generic solution needs to be found that is also transparent to technological variants.

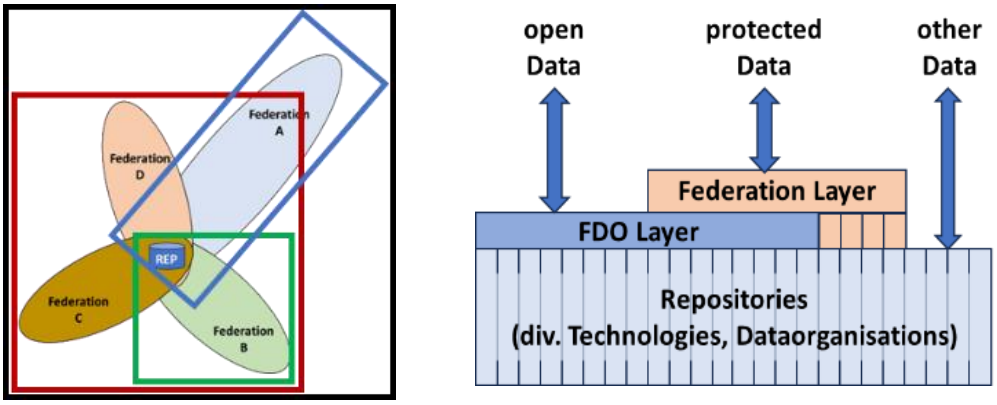

Figure 1. *The left figure indicates current practices indicating that repositories with relevant data are in many cases already part in a number of federations (dataspaces) requesting specific regulations and technologies. The right figure indicates the layer structure of the emerging global integrated dataspace where many repositories will be connected via the minimal FDO standard and may apply specific federation technology to control usage rights where this is required.*

This allows us to propose an architectural solution as indicated in figure 1 (right): The FDOs have the potential to form a basic integration layer for all open data, since they are independent of any agreement on any specific regulations and technologies. However, FDOs are units of information that bundle all kinds of metadata including access rights information. This means that on top of FDOs federation layers could be implemented that agree on specific sets of regulations and technologies. Of course, the diagram also indicates areas which do not make use of the FDO layer.

It should be noted that repositories have internal data protection regulations to guarantee data integrity and trustworthiness relevant for the FDO layer, but these regulations are in accordance with the external interface standards of the repository or the FDO layer.

3. Federation Agreements

For the generic FDO solution, a repository will need to act as an FDO server (i.e., contain an adapter that is capable of offering all metadata information associated with a bit-sequence encoding some useful content to the DOIP interface protocol) [4]. That's all—there is no need to adapt or to map crucial information. In many collaborations metadata categories need to be mapped to different semantic spaces etc. This is something that needs to be solved by the client software if required. One can say that the FDO presents the internal information being managed by the repository in a neutral way.

As an example for a solution where data protection plays a key role we can refer to the Eclipse Data Connector (EDC) [5] solution which was developed within the IDSA and which is currently being tested in industrial pilot applications which define roles, rights associated with roles, app stores etc. for their federations. On each side a specific connector needs to be installed that takes care that all data and metadata transactions are occurring as it is agreed between the two partners (see for a schematic indication figure 2). These can include complex license agreements such as: use the data for a limited time. Only applications that have been checked and added to the app store can be used to carry out operations on the data which, for example, would prevent unwanted data copying during processing. The EDC solution requires agreements between two or more parties and to establish a joint governance.

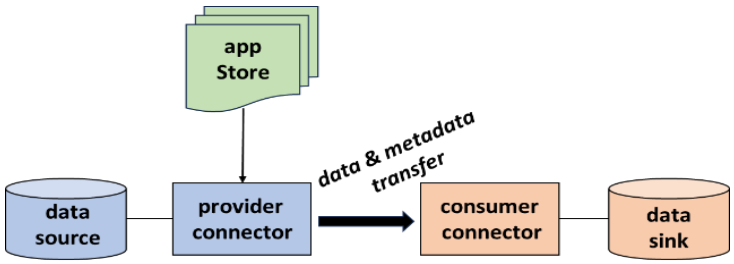

Figure 2. *This figure shows schematically the configuration of the Eclipse Dataspace Connector technology as it was developed by ISDA. It uses two specialized connectors that negotiate and control the usage of data according to some specifications as agreed in a specific data space.*

It can be seen that there would be no harm to use FDOs as the data source, i.e. the EDC could be used on top of the common FDO layer. In addition, one could argue that using the FDO layer unifies across all different data sources, i.e. applying the EDC would be easier since only one interface between FDOs and EDC would have to be provided.

4. Application Relation

We should point to a difference between the two approaches when it comes to linking operations which are crucial to establish trust. In the EDC example, the applications need to be checked by the federation boards and added to the app store and they need to be executed by the provider connector, so that only specific data will be transferred.

In the case of FDOs, the data provider can specify a registry that contains relations between types and operations, i.e. in this way the data provider can specify which software can be executed on the data. There are no specific checks involved in a dynamic research environment which would be difficult to realize and it would also be difficult to carry out checks for case of large software packages. Both solutions may have their relevance for specific cases, but the FDO solution would be simpler to realize

Data availability statement

There is no data directly involved in this paper.

Author contributions

All authors contributed at the same level to this paper.

Competing interests

The authors declare that they have no competing interests.

References

[1] FDO Overview, https://zenodo.org/records/7824714
[2] IDSA Dataspaces, https://internationaldataspaces.org/
[3] Industry 4.0 Platform, https://www.plattform-i40.de/IP/Navigation/EN/Home/home.html
[4] DOIP Protocol, https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf
[5] Eclipse Dataspace Connector, https://international-data-spaces-association.github.io/DataspaceConnector/