FAIR Digital Object Demonstrators 2021

Peter Wittenburg (0000-0003-3538-0106), Ivonne Anders (0000-0001-7337-3009), Christophe Blanchi (0000-0003-2277-5176), Merret Buurman (0000-0001-6945-456X), Carole Goble (0000-0003-1219-2137), Jonas Grieb (0000-0002-8876-1722), Alex Hardisty (0000-0002-0767-4310), Sharif Islam (0000-0001-8050-0299), Thomas Jejkal (0000-0003-2804-688X), Tibor Kálmán (0000-0001-5194-5053), Christine Kirkpatrick (0000-0002-4451-8042), Laurence Lannom (0000-0003-1254-7604), Thomas Lauer (0000-0002-7662-1132), Giridhar Manepalli, Karsten Peters-von Gehlen (0000-0003-0158-2957), Andreas Pfeil (0000-0001-6575-1022), Robert Quick (0000-0002-0994-728X), Mark van de Sanden (0000-0002-2718-8918), Ulrich Schwardmann (0000-0001-6337-8674), Stian Soiland-Reyes (0000-0001-9842-9718), Rainer Stotzka (0000-0003-3642-1264), Zachary Trautt (0000-0001-5929-0354), Dieter Van Uytvanck (0000-0001-7807-186X), Claus Weiland (0000-0003-0351-6523), Philipp Wieder (0000-0002-6992-1866),

Abstract

This paper gives a summary of implementation activities in the realm of FAIR Digital Objects (FDO). It gives an idea which software components are robust and used for many years, which components are comparatively new and are being tested out in pilot projects and what the challenges are that need to be urgently addressed by the FDO community. After basically only one year of advancing the FDO specifications by the FDO Forum we can recognise an increasing momentum to test and integrate essential FDO components. However, many developments still occur as soloistic engagements that offer a scattered picture. It is widely agreed that it is now time to combine these different pilots to comprehensive testbeds, to identify still existing gaps and to turn some services into components of a convincing and stable infrastructure. This step is urgently needed to convince even more institutions to invest in FDO technology and therefore to increase FAIRness of the evolving global data space.

1. Foreword

Beginning 2021 the FDO Forum[1] was established as the place to discuss all aspects related to FAIR Digital Objects (FDOs) [1,2], to advance the specifications and to closely collaborate with other initiatives focusing on FDOs such as the RDA FDO Fabric IG[2]. There are indications that the concept of FDO is gaining traction, since an increasing number of experts are looking for basic agreements for establishing a FAIR data space across research communities. In the realm of EOSC[3] FDOs have been introduced as a way to turn FAIR principles into practice and in the current EOSC discussions they are one of the "nuggets" for establishing cross-disciplinary and cross-country data infrastructures. In recent NFDI Projects (National German Data Infrastructure)[4] FDOs have been deeply embedded in the infrastructure concepts (e.g. NFDI-MatWerk, NDFI4Ing, NFDI4Phys, NFDI4Earth) and quite a number of institutions and initiatives are already testing out aspects of FDOs.

[1] http://fair-do-org
[2] https://www.rd-alliance.org/group/data-fabric-ig.html
[3] https://eosc.eu/
[4] https://www.nfdi.de/?lang=en

The DOBES example referred to in the already mentioned EOSC document "Turning FAIR into Reality" [3] refers to the DOBES online archive[5] as an example that has achieved almost FAIR compliance although it was started already in 2000. The data collected by an internationally distributed team of researchers is about languages and cultures that are endangered, i.e., the generated material is part of our cultural heritage and cannot be generated again. Therefore, the technological work was driven by the need to create data structures and organisations that guarantee sustainability and allow researchers worldwide to uniquely refer to each digital object, be it data or metadata. Each object stored was assigned a PID (Handle[6,7]) and metadata was created according to a community standard[8]. At that time machine actionability was not yet an issue, but it would not be so difficult to add a PID profile describing the kernel attribute set and to register the kernel attributes used.

In May 2021 the FDO Forum was invited to an eIRG workshop[9] on interoperability and three presentations related to FDOs were given. One talk explained the FDO concept, a second talk described the results of the EOSC PID Architecture group [4] and a third described the concept of flexible semantic mapping (SEMAF) [5] as key for supporting crosswalks. As is known, PIDs are key for creating and reusing FDOs, since they not only give clear identities but also do the binding of all relevant information of a digital entity that is necessary to find, access, interpret and reuse the entity. The goal of assigning PIDs is to create a stable and manageable global data space organised by self-standing FDOs. Flexible semantic mapping enables registration of mapped relations between concepts as they are used in communities to create metadata, annotations, measurements etc. Such semantic mappings are part of our scientific memory and need to be persistently assigned with stable references, thus they should be organised as FDOs.

The FDO Forum knows that working only on specifications will not be sufficient, since proper specifications can only emerge when they are tested in practice and since experts want to see FDOs in practice to understand their potential and to estimate the required effort. Therefore, a few selected demonstrator cases were also presented at the eIRG meeting. In this paper we will therefore discuss (1) some basic software components which are already available, (2) the demonstrators presented at the eIRG meeting, and (3) other demonstrators from work in progress. Finally, we will summarise the contributions, give an overview, identify challenges and draw some concluding remarks.

Especially the last two sections are important, since until now we see that many groups are working on individually designed projects as required by the funding sources. What is lacking is to integrate all these projects based on unified infrastructure pillars.

2. Basic Tools

A few basic tools have been developed, have been tested and are ready for use:

- the Handle System to manage PIDs including an Identifier Resolution Protocol
- the Digital Object Interface Protocol (DOIP)[10] which interacts with FDOs independently of the way they are configured in repositories and other FDO servers and clients
- the Data Type Registry as defined within the RDA DTR WG[11] and installed in some sites

[5] https://dobes.mpi.nl/
[6] https://www.handle.net/
[7] Certain commercial software, equipment, instruments, or materials are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.
[8] https://www.clarin.eu/content/component-metadata
[9] https://indico.fccn.pt/event/15/overview
[10] https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf
[11] https://www.rd-alliance.org/group/data-type-registries-wg/outcomes/data-type-registries

- the set of PID Kernel Attributes as defined by the RDA Kernel Type WG[12] and as registered at GWDG DTR[13]
- the PID Information Types (PIT)[14] service to offer a simple interface to create, update and validate PIDs and their records following FAIR DO concept. It is still in development, but already usable.

Handle System

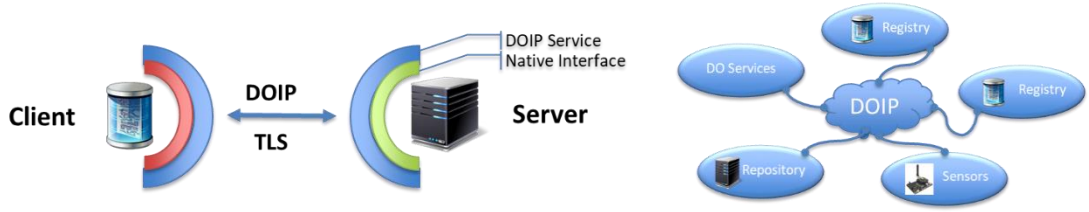
The Handle System is a globally used PID registration and resolution system based on RFC specifications. It is used by about 3000 institutions worldwide and one of its most active user communities is the International DOI Foundation[15]. DOIs are very well-known and emerged in the field of ePublications. They are using Handles with the prefix 10 and have defined a specific business model that implies a fee structure for its members.

The Handle System is governed by the independent international DONA Foundation registered in Geneva[16]. The DONA Foundation is responsible for maintaining the distributed Global Handle Registry system which exists of a set of Multi-Primary Administrators (MPA). The handle has a two-level syntax: <prefix>/<suffix>. The MPAs run redundant global resolvers and also act as authorities to issue prefixes to other institutions and initiatives. These in turn then manage Local Handle Registries, generate suffixes according to their own schemes and define their own business model. While the GHR software is controlled by the DONA Foundation to guarantee seamless operation, local Handle registries can operate their own software as long as they adhere to the Handle specifications.

The Handle software is highly optimised to offer fast resolution of Handles and has now proven to operate reliably for more than 20 years. Together with specifications made in RDA groups the Handle System can be used as an ideal basis for the creation and management of FAIR Digital Objects.

DOIP

The Digital Object Interface Protocol (DOIP) is defined as a simple and universal protocol to interact with all entities (clients and servers) that manage or process FDOs. DOIP can be tunnelled through any secure communications protocol and the DOIP itself can be used to determine the choice of such



This figure shows the purpose of DOIP as a protocol allowing to combine any client and server talking to each other if the FDO model is supported. In the right diagram it indicates how DOIP can be used to create single integrated FDO space independent on how repositories organise their data and on which technology they have chosen.

protocol. A minimum requirement is that TLS[17] be supported for network communications. DOIP[18] is independent of how data are being organised in different servers such as for example repositories. To support DOIP, however, adaptors might have to be developed if the server does not support FDOs natively. Introducing DOIP as the interaction protocol allows interaction among all kinds of devices that

[12] https://www.rd-alliance.org/groups/pid-kernel-information-profile-management-wg
[13] https://dtr-pit.pidconsortium.net/#urls/intro.html
[14] https://www.rd-alliance.org/groups/pid-information-types-wg.html
[15] https://www.doi.org/
[16] https://www.dona.net/index
[17] https://en.wikipedia.org/wiki/Transport_Layer_Security
[18] https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf

create and manage FDOs and thus reduce complexity to 1*N. Adapters need to be maintained locally and a server only needs to maintain one adapter.

The protocol elements show the concept behind DOIP. All items are uniquely specified by PIDs, i.e., it puts global unique and resolvable persistent identifiers, as for example provided by Handles, into the core. With respect to operations, it should be mentioned that DOIP comes with basic operations to create, read, manipulate and delete FDOs, however, any other operation defined by users can be linked as well.

The following set attributes need to be provided in a request package:

- **requestId**: The identifier of the request provided by the client to keep track of responses.
- **clientId**: The identifier of the requesting client.
- **targetId**: The identifier of the DO on which to perform the operations
- **operationId**: the identifier of the operation to be performed on the DO.
- **attributes(optional)**: Stipulated by the Operation ID.
- **authentication(optional)**: used by clients to authenticate.
- **input(optional)**: Payload stipulated by the operation.

The DOIP v2.0 specification has been implemented, tested and is available as software development kit.

CORDRA Refence Implementation

Cordra[19] is highly configurable open source software offered to software developers for managing digital objects, and thus FDOs, with resolvable identifiers, at scale. It integrates popular storage and indexing solutions, presents a unified interface and provides configurable hooks to validate and update information at various stages of a digital object lifecycle. Cordra can save substantial development time as it comes with built-in functions that developers need ranging from user authentication and access control to information validation, enrichment, storing, and indexing.

It can also be seen as a reference implementation of a DOIP-supporting server. The software can be downloaded and is open source. It is currently being used by several projects for a variety of goals.

Data Type Registry

Each FDO has a type and also a variety of attributes in PID and metadata records that need to be "typed", i.e., defined and registered, to make them machine actionable. Therefore, registries must be available for users. In many cases user communities have already created different kinds of semantic artifacts where agreed concepts are being defined and registered. If PIDs are supported clearly and persistently identifying concepts these can be integrated in the FDO domain.

Often, however, there is no simple facility to define and register concepts and assign types. For this purpose a Data Type Registry (DTR) has been specified in RDA and open-source software has been developed according to the specifications including an API. Such DTRs have been installed at some places like at GWDG and are being used by a variety of projects.

PID Kernel Attributes

All FDOs are assigned a PID and when such a PID is submitted to a PID system such as the Handle System the clients expect to receive useful attributes that allow to access to the different informational entities of this FDO (important attributes such as checksum, type of data, references to bit-sequence, metadata, access permissions, etc.). We call these metadata attributes Kernel Attributes.

[19] https://www.cordra.org/

A general recommendation is to not overload the PID record with all sorts of metadata, but to link to a separate metadata description that will include contextual and provenance information. On the other hand it is important to provide recommendations about the set of widely agreed Kernel Attributes. The RDA WG on Kernel Information Types defined a first set of such attributes, which user communities want to get immediately when the PID has been resolved. These suggested attributes have been defined and registered in the DTR maintained at GWDG[20].

The FDO Forum is now further discussing the requirements for Kernel Attributes (KA) for the specific needs of FDO. It is widely agreed that

- Kernel Attributes must be defined and registered in a DTR
- FDO servers need to specify the profile of KA they are providing as a selection of the registered KA
- there will be mandatory and FDO-community wide agreed KA, but that user communities could add KA as long as these are registered in an open DTR (detailed specifications are in progress)

PID Information Types (PIT) service

The PIT implementation of KIT[21] is an adaption of the RDA concept and implementation[22]. It is a generic abstraction for creating, updating and validating PIDs and their records, which involves requests to a Data Type Registry and a PID service. For example, to maintain machine actionability, it is recommended to validate PID record information before creating PIDs or maintaining (modifying) PID records. It can be configured to use arbitrary instances of DTRs or Handle prefixes.

3. eIRG Demonstrators

For the eIRG workshop three demonstrators were selected and presented:

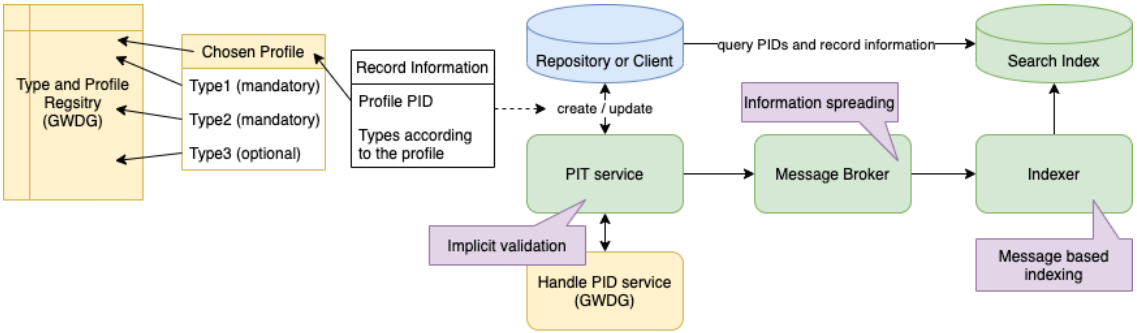
- proper PID creation and management at Karlsruhe Institute of Technology (KIT)
- workflow creation at Indiana University (IU) using systematically PIDs and applying DOIP
- FDO exchange between different repository types using DOIP

It should be noted that the creation of Handles (PIDs) is being carried out at many institutes and in many initiatives worldwide already.

[20] https://dtr-pit.pidconsortium.net/#urls/intro.html
[21] https://github.com/kit-data-manager/pit-service
[22] https://rd-alliance.org/group/pid-information-types-wg/outcomes/pid-information-types

3.1 KIT[23] FAIR Digital Object Ecosystem Testbed

The FDO concept promises a method to harmonize data interoperability. But from a developer's perspective, a lot of open questions remain to start with implementations of clients that can make use of it. What are the standard procedures for creating, updating and retrieving PIDs? When should validation take place and how does it work? How does the knowledge about newly created or updated



This figure shows the testbed at KIT implemented as a set of services that work together to demonstrate feasibility and offer simple APIs for client developers. This image shows the most important components, which are discussed here. Yellow components are external to the testbed. The client (blue) is represented by an interface for purely demonstrational purposes within the testbed. Green components are contained in the testbed and can also be reused independently.

FDOs spread across institutes, services and technologies? This testbed implements such processes and offers them as a simple-to-run container to developers. The ecosystem, consisting of a set of services, can be executed on everyday computers. The PIT service is part of the testbed and offers a simple REST-API for clients to create PIDs, update PID records and resolve PIDs. By default, it uses sandboxed PIDs, which are only resolvable from within the testbed, so developers can test and experiment without access to a PID system. It can be configured to a handle prefix if this is not desired.

Validation of the records will take place implicitly when creating PIDs or updating a PID record. The client does not have to explicitly trigger the validation process and can instead simply rely on the checked assumptions. In this context, validation means that a record is expected to contain a typed property, that has a PID as a value, pointing to a profile. A profile contains a set of typed properties and their obligations. Profiles and Types are registered in a data type registry (DTR), for example the one of GWDG. The validation process will ensure that

1. all mandatory properties defined within the profile are contained within the record.
2. only properties defined within the profile are contained (if the profile does not allow such).
3. all properties have valid values according to their definitions in the DTR.
4. every property can be resolved and is defined in a DTR.

This way, the testbed enforces the usage of registered profiles. The usage of profiles enforces verifiable machine actionability under the assumption that they only contain registered, machine actionable types as properties, which is ensured by DTRs. For each type-value pair in a record, the client knows whether it is a checksum of an object, the date of ingestion, etc. and how to interpret the value.

Finally, for each created or modified PID (record), the PIT service will spread the information that this PID contains new or updated information, by sending it to a message broker. This can be used, for example, to track changes on PIDs, link PIDs and for general information federation. Each PID is disseminated with a topic identifier string, indicating if the PID was just created or if the record has been changed. Any interested service might register for one or both topics.

This principle has been applied to build a local search index within the testbed. An indexer service registers itself at the message broker for all events, to track every new PID and changes within existing

[23] https://www.kit.edu/

PIDs. It will then resolve the PIDs, transform the record information and put it in a search index service. If the indexer would register at additional message brokers, their events could be included, resulting in the inclusion of information in external search indices.

Proper PID generation is an essential basis for all kinds of FDO applications. The components such as the PIT service and the indexer can be reused and are available as free and open-source software[24], although it should be noted that most of the services are still in development. Especially the dissemination function is of general interest, since a variety of interesting services can be associated. Possible applications would be to identify duplications based on checksum and perhaps other metadata information included in the PID record, updating PID graphs or building search indices. The testbed is in continuous development to evaluate and improve different aspects of a FAIR ecosystem. KIT is planning to adopt DOIP where applicable. The testbed has been supported by the research program 'Engineering Digital Futures' of the Helmholtz Association of German Research Centers and the Helmholtz Metadata Collaboration Platform (HMC).

3.2 Systematic Workflow Documentation at Indiana University[25]

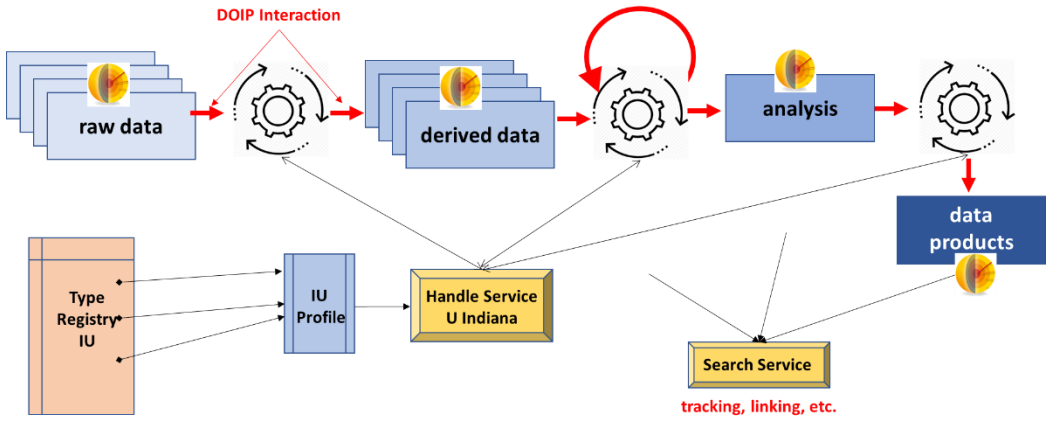
At Indiana University (IU) a variety of FDO testbeds are being developed. For the demonstration at the eIRG meeting we selected a workflow use case which generates a set of specific data products from some input collection with resources from material science to improve reproducibility. The workflow was implemented by systematically assigning proper PIDs for all digital objects, data and software, which are involved in the steps. For each separate object type different PID profiles were specified, and the used attributes were registered in a data type registry. This systematic and proper PID assignment guarantees machine actionability and excellent documentation of all workflow steps. Since the automation of data processes with the help of workflows will become increasingly important it is of great relevance to design proper mechanisms for documentation and reproducibility. Systematically using Handles with machine actionable resolution results pointing to all digital objects and creating metadata descriptions including all relevant provenance information is the basis for creating and managing FAIR Digital Objects. PID record information and metadata descriptions are exported to a search service which enables the development of a variety of useful added-value services.

This figure illustrates the set of different object types that are involved in the IU workflow demonstrator and their assignment with Handles (PIDs) using different profiles with registered attributes. Such an assignment implements FAIR-compliance and thus machine actionability.

It should be noted, that during workflow processing the DOIP protocol is used to carry out all exchanges of information, since all digital objects including the software are FDOs as is indicated in the figure

[24] https://github.com/kit-data-manager/testbed4inf
[25] https://www.indiana.edu/

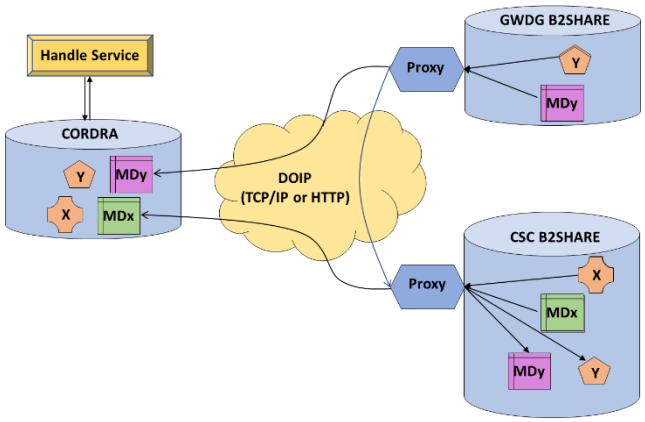
below. In addition, a Virtual Research Environment was developed to support workflow orchestration and execution.



This figure illustrates the workflow process as has been implemented at IU. By assigning systematically Handles and metadata descriptions that include provenance information, a proper and persistent documentation and reproducibility are achieved. All FAIR Digital Objects being created can be found, accessed, interpreted, and reused by others.

3.3 Repository Integration Demonstrator

The third demonstrator presented at the iRG meeting was designed and developed to show the interchange of digital objects between repositories applying different technologies and data organisations. This demonstrator shows that B2SHARE[26] repositories as developed in the EUDAT initiative can be easily integrated with a US-based CORDRA repository by developing a DOIP proxy that acts as an adapter to B2SHARE. The focus for this demonstrator was on indicating the strength of the FDO model and the DOIP protocol. B2SHARE has a specific way to organise data and metadata, but it is not fully FAIR compliant. However, Handles are assigned to all "bundles", which are described by metadata descriptions and to all individual "data streams" that are associated with this bundle. In doing so and offering an API, which enables easy PID based access to all entities, it was indeed fairly straightforward to develop a DOIP proxy[27].



This figure illustrates the integration between two differently structured repositories with the help of a proxy that adapts B2SHARE instances to DOIP and facilitates easy integration.

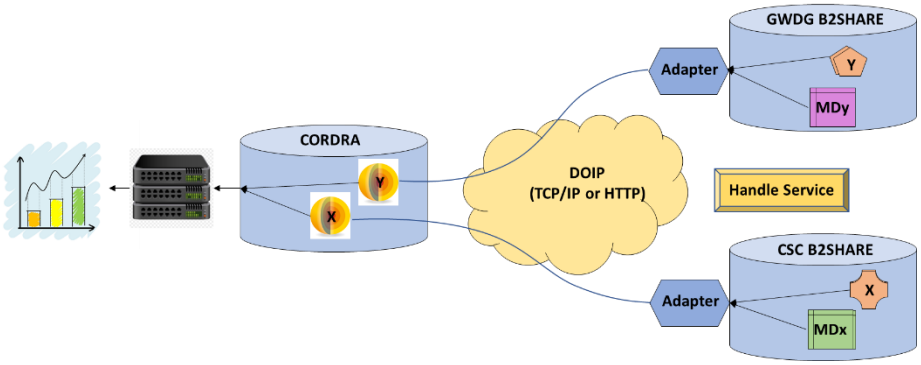
For the demonstrator simple files (spreadsheets) - described by some metadata and stored at the GWDG B2SHARE instance - were copied to the CORDRA store at CNRI. For this first demonstrator we created the same data organisation in the CORDRA repository as in the B2SHARE instance, i.e., data streams and metadata were in this bundle structure. This proxy could now be used in the same way to adapt other B2SHARE instances,

[26] https://b2share.eudat.eu/

[27] The effective development time was less than a week.

such as the one at CSC[28]. Due to time restrictions this could not be tested. For the demonstrator we used command line operation launching.

A logically next step would be not to create the same data organisation as in B2SHARE at the receiving site. This would enable to immediately create true FAIR Digital Objects, which would bring data together from different sources and would make it possible at one of the sites to execute statistical applications, for example as indicated in the following figure. A simple VRE would be necessary to replace the command line feature by easy-to-use frontends and to enable a set of operations to manage FDOs (copy, move, delete, execute). Due to time restrictions this needed to be postponed.



This figure illustrates an enhanced version of the Repository Integration Demonstrator where the results of the copying actions are true FAIR Digital Objects making it easy to execute analytics, for example, on data aggregated from different sources.

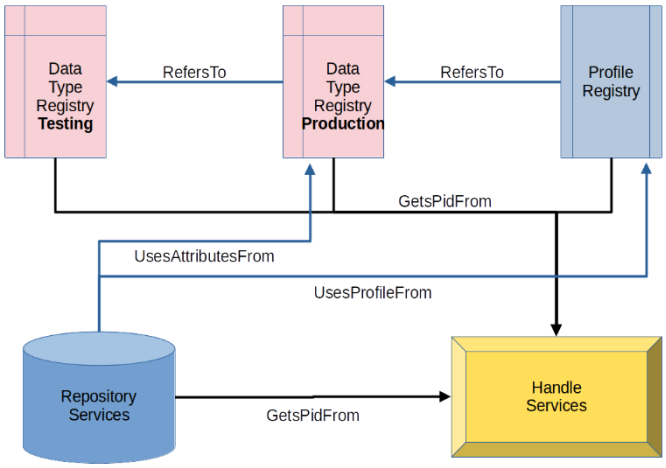
3.4 PID, Type and Profile Management at GWDG

GWDG[29] provides for its customers and on behalf of ePIC several Handle PID services, data type registries (DTRs) and a profile registry, which were partly used for the e-IRG demonstrators. The usual life chain of types starts with user or community driven type definitions in a testing environment, where all types get a Handle PID, but are still due to changes and also deletions.

After the maturity of a type definition is acknowledged by the community and certain consistency conditions are proven, the application for a candidate type leads to an instantiation of the type in the production DTR. This registration of the type implies its permanent maintenance. After approval, the candidate becomes a productive type. In case it is replaced by a newer version, the type becomes deprecated, but it still stays maintained.

[28] https://www.csc.fi/
[29] https://www.gwdg.de/

All such types can be bundled into so called profiles that describe the types to be expected with a given FDO, and all profiles are provided with a PID. Several repository services are provided by GWDG for its customers and other institutions, as well as for communities and projects. The resulting digital objects can become FAIR, whenever these repositories are used together with this typing framework.



This figure illustrates the GWDG repository services and its embedding in a framework of typing services.

All DTRs and the profile registry are based on the CORDRA software. In the specific case, where the repository contains type instances, CORDRA is also used as the underlying repository technology.

4. Further Demonstrators

In this chapter some projects will be presented that could not be selected for the eIRG meeting, but which are carried out or are planned.

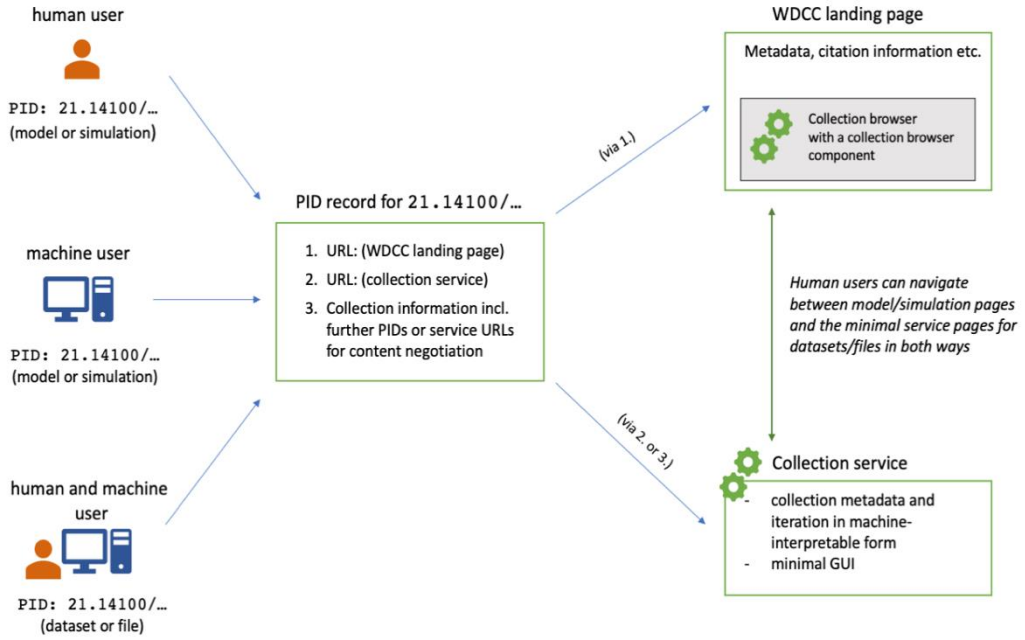
4.1 Work in Climate Modeling

The global climate science community has for several years now established efficient ways of disseminating high-volume climate model simulation output at global scale. The system of choice is the ESGF (Earth System Grid Federation[30]) infrastructure, which is a globally distributed network of nodes facilitating efficient data findability and access. DKRZ[31] is a major partner in the ESGF and takes the lead for implementation and maintenance of a PID-system [6] specifically for the most recent global climate model output inter-comparison effort CMIP6. The implementation has been operational since 2016. The PIDs for individual files are provided by a handle-server hosted at DKRZ, are contained inside the files and are equipped with Kernel Attributes (dataset ID, file size, checksum (SHA256), data access URL, related datasets, errata information, information of precedence / supersedence by older / newer file versions). The PID landing pages are also hosted and maintained at DKRZ (see Figure below). The current implementation allows the CMIP6 data collection, currently 9.82 PetaBytes in volume, to be human- as well as machine-actionable. Together with the domain-specific (meta)data standards required for CMIP6 (meta)data to be published in the ESGF, the demonstrated approach is close to an FDO infrastructure.

Plans to advance and establish the ideas in this demonstrator include efforts to build connections to the European Open Science Cloud (EOSC) and formalize the approach by incorporating the domain-specific PIDs and the corresponding attributes in DTRs to enable reuse beyond the climate science community.

[30] https://esgf.llnl.gov

[31] https://www.dkrz.de/en/dkrz-partner-for-climate-research?set_language=en



This figure indicates entry points, pathways and resulting GUIs or web services that human users or machine agents (e.g. smart clients) can access from a given PID. The pathways can lead to metadata and eventually data. (Source: Figure 4 in [3] modified.)

4.2 Work at NIST[32]

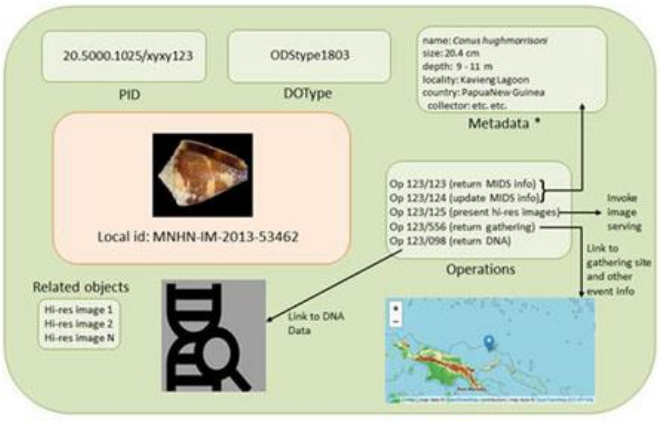
Researchers within the Material Measurement Laboratory within the National Institute of Standards and Technology are spearheading early-stage research and development activities centered around FDO use cases within materials science and engineering. Work has initially been focused on high-level metadata schema development for FDO Types, addressing concerns within the materials science and engineering community. We have emphasized the need for PIDs for samples, instruments, and data, and illuminating their relationships via PIDs in metadata. We are working with two partners on pilot projects. One pilot project is centered around fully representing process-structure-property relationships in metallic materials and the other pilot is centered around creating and selecting materials sourced from locally abundant biomass that are part of a regenerative circular economy.

4.3 DiSSCo Digital Specimen Repository

In DiSSCo, the FDO model was used to design a digital specimen repository for natural science collections data. As is shown in the diagram below, a digital specimen representing the physical specimen of a cone snail is identified by a Handle (20.5000.1025/xyxy123). This persistent identifier can be resolved to a rich information bundle: a local identifier (MNHN-IM-2013-53462 is stored in the French National Museum of Natural History), a digital object type definition (ODSType1803), some metadata according to an agreed standard, and the thumbnail image of the snail as payload. Besides the metadata, a variety of operations have been defined for this digital object which gives access to associated information such as methods to access the metadata, the high-resolution image of the snail, DNA sequence data and a map indicating the location where the snail was collected. DiSSCo is working on a specification (https://github.com/DiSSCo/openDS) according to which all digital specimens need

[32] https://www.nist.gov/

to be described. On the one hand, this open specification guarantees flexibility and the other hand, harmonises data across different types of specimens with local data elements.

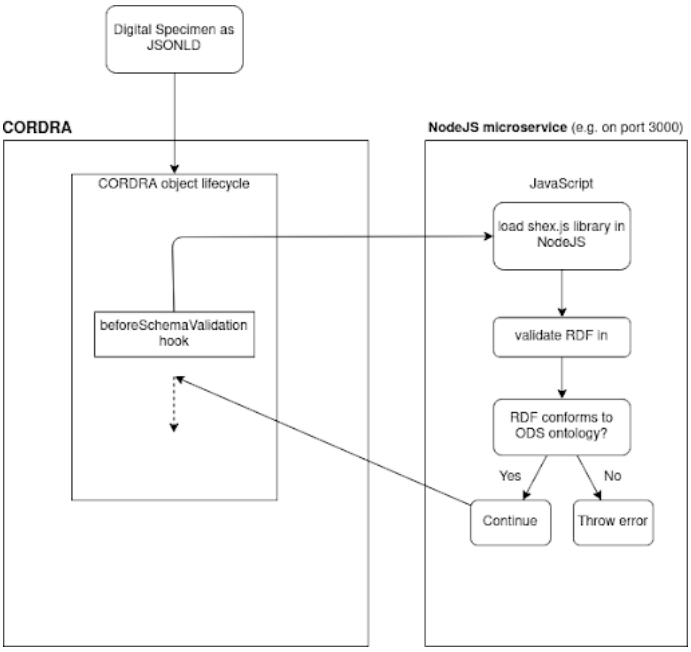


This figure shows gives an impression of the configuration of the Digital Specimen FDO as designed in DISSCo.

DiSSCo's implementation of the Digital Specimen repository (https://nsidr.org) is based on the Cordra (https://cordra.org) open-source software that provides both Digital Object Interface Protocol (DOIP) and REST interface. This repository will serve as a reference implementation for the partners involved in the DiSSCo project and different services can be built using the FDOs as building blocks.

4.4 DiSSCO[33] Work with CORDRA

DiSSCo'S Digital Specimen architecture (DS arch) follows the principle "FAIR by design", which led to a set of implementation considerations described in the paper by Islam [7]. A key factor for DS arch is the extensive integration and meshing of our Cordra-based system architecture (Section 4.3) and machine-actionability as a core objective of the FAIR principles [8,9].



This figure shows the integration of an RDF validation pipeline into the CORDRA based repository of Digital Specimen.

Of particular importance in this context are aspects of interoperability and reusability as GO-FAIR's FAIRification process outlines: (I3) Qualified cross-references and mappings to semantic artifacts (controlled vocabularies, taxonomies, ontologies) describing digital resources enable machines to detect and classify the data regarding the suitability for a specific purpose [4] and (R1.1, R1.2) the provision of rich license and provenance data facilitates machine decision making if the resource is actually usable in a given context[34].

To implement these objectives, we complemented Cordra's digital object validation mechanism based on JSON Schema by a semantic validation pipeline (Figure). We developed a JSON for Linked Data[35] representation for Digital Specimens (building on the standard openDS[36]), which is both valid JSON to sustain usability within Cordra as well as a serialization format for RDF.

Accordingly, we used the Shape Expressions Language (ShEx) [10] to define a schema of data types and properties which a digital object is expected to satisfy. A Cordra Life Cycle Hook[37]

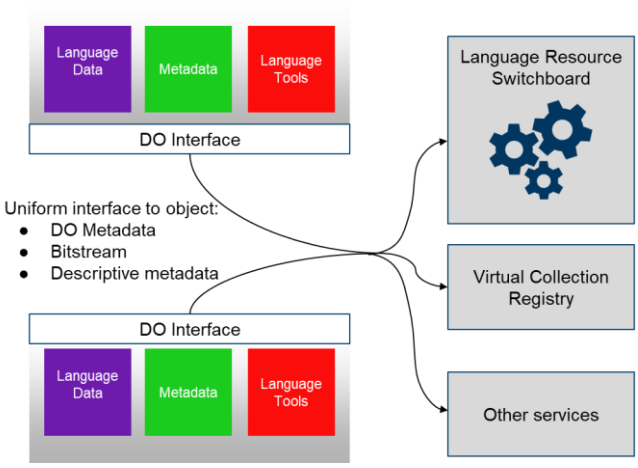
[33] https://www.dissco.eu/
[34] https://www.thehyve.nl/articles/fair-data-for-machine-learning
[35] https://www.w3.org/TR/json-ld11
[36] https://github.com/hardistyar/openDS
[37] https://www.cordra.org/documentation/design/javascript-lifecycle-hooks.html

("beforeSchemaValidation") sends the digital object to a microservice which performs the RDF validation and triggers - if the object satisfies all asserted constraints - the next step of the Create Lifecycle of Cordra. We regard this service as an initial step to foster machine actionability of data by expressing interrelations of these data in a semantic form that machines can process.

The source code is available at: https://github.com/jgrieb/ODS_rdf_validator

4.5 Switchboard[38] work in CLARIN

The CLARIN ERIC research infrastructure includes currently more than 50 regularly audited centres mainly in Europe but also in the US and South Africa which are offering language resources of various kinds and from all regions worldwide and tools/services. One big challenge CLARIN is addressing is the question how to offer analytic and visualisation capacities to any researcher interested in language resources for a virtual collection aggregated for a specific research purpose. Trustworthy centers in CLARIN need to assign PIDs (mostly Handles are used, some use DOIs[39]), associate kernel attributes according to a defined profile and describe their resources with the help of community wide agreed CMDI metadata. Metadata of all these centres and beyond are harvested, mapped and offered via the Virtual Language Observatory. A collection builder allows users to create virtual collections according to a standard complying widely with the RDA research collection specifications. Thus, the CLARIN language resource community is fairly FAIR compliant which is an excellent starting point to improve integration and interoperability issues.



This figure shows show the switchboard solution as implemented in CLARIN to facilitate the connection between data and operations.

CLARIN implemented a switchboard concept allowing users to map language resources to tools based on metadata descriptions, i.e. dependent on categories such as "language in resource", "type of resource", etc. A set of suitable tools will be presented to the user facilitating advanced use of technology even for lay-persons. As the diagram indicates CLARIN will now turn to use a unified interface to all resource and tool providers to simplify the switchboard solution. The Digital Object Interface Protocol will be tested for these purposes. Aligned with this change CLARIN will also improve its services by registering all kernel attributes at the DTR offered by the GWDG based on a registered profile.

4.6 RO-Crate FDOs in Biomedical Research Infrastructures

The practice of performing multi-step computational processes using workflows has taken hold in the biosciences as the discipline becomes increasingly computational. These workflows perform such tasks as data preparation, data analysis, simulation sweeps and AI/model predictions. ELIXIR[40] is leading 13 European Research Infrastructures in the biomedical sciences to create an open, digital and

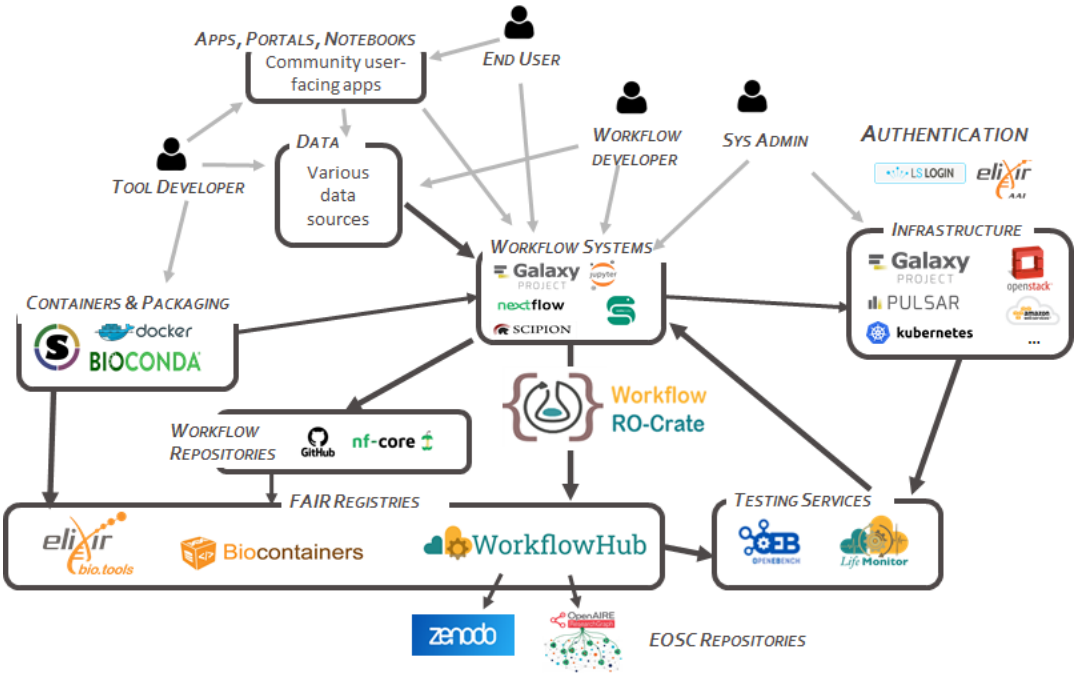
[38] https://www.clarin.eu/

[39] DOIs are Handles with prefix 10 and thus part of the global Handle resolution system and a specific business model.

[40] https://elixir-europe.org/

collaborative space for biological and medical research. This EOSC-Life[41] cluster project is building a cloud-based "collaboratory" for researchers so that they may make, use and share workflows that are FAIR. Researchers of all skill levels can use workflows as an entry point to access tools and datasets, and reuse complex processing methods. Thus workflows democratise access to EOSC resources and spread processing know-how for FAIR data processing.

A range of popular workflow management systems (Galaxy, Snakemake, Nextflow etc) are supported by a suite of FAIR services and registries and the workflows are recognised as FDOs, to be created,



This figure highlights the role of RO-Crate as a powerful approach which is (i) Web-native, using existing techniques familiar to web developers and search engines, and (ii) extensible in order to easily cope with the diversity and legacy of the many workflow systems, data types, catalogues, repositories and pre-existing platforms in the biosciences. Figure adapted from https://www.eosc-life.eu/tools-workflows/

consumed, and exchanged between these services. RO-Crate[42] implements these workflow FDOs, and the metadata framework based on schema.org Linked Data provides common descriptions. RO-Crate is a self-describing JSON-LD format that is human- and machine-readable, and that packages resources together with their descriptions and associations. For workflow FDOs [13], this includes: the description, input and intermediate data, parameter settings, final outputs, software programs and workflows, configuration information and so on. These RO-Crate objects are created, exchanged and activated by the services in the collaboratory (e.g. continuous testing), and published by the WorkflowHub registry that supports continual workflow evolution lifecycle (e.g. from GitHub) with release PIDs using Datacite DOIs (e.g. 10.48546/workflowhub.workflow.230.1). Indexing and long-term archiving is being developed with federated EOSC registries and repositories such as OpenAIRE and Zenodo.

5. Integration and Gaps

5.1 Basic Components

In this table we summarise the state of some specifications and components which are already being used in various projects beyond the selected demonstrators.

[41] https://www.eosc-life.eu/
[42] https://www.researchobject.org/ro-crate/

Basic Components	Specification	comment
Handle-System	RFC, IPR	ready service, robust since more than 20 years, broadly used by publishing & film industry and by many institutes with large data volumes, it is maintained by DONA Foundation
DO Concept	papers, RDA DFT	described early by Kahn & Wilensky [11], taken up in RDA DFT which defined the Data Core Model, then taken up in RDA Data Fabric to spin off more work
DOIP	DOIP V2.0	DOIP V2.0 was worked out within the DONA foundation as a protocol to interact with Digital Objects, since FDOs are a subset of DOs DOIP will also work for FDOs.
DOIP SDK	SDK	DOIP Software Development Kit in use by some institutions
CORDRA	CORDRA	reference software to demonstrate DOIP interaction, is used by some projects as registry and repository software
Kernel Type Concept	RDA PIT, RDA Kernel Info Types	the concept of Kernel Types associated emerged in some RDA groups and was well-defined at the end resulting in some suggestions for Kernel Types
PID Profiles	RDA Kernel Info Types	this concept was defined by the RDA KIT group
Data Type Registry	RDA DTR	the DTR was specified by RDA and is put to service in some institutions
FDO Concept	specified by FDO Framework V1.02	FDO concept was born as result of a collaboration between GEDE and GOFAIR bringing together the concept of DO and FAIR. FDO Framework V1.02 was endorsed by the Paris Meeting

5.2 Short Summaries of Demonstrators

Here we summaries the FDO related essentials of the different projects being described in this paper.

1. PID, Type and Profile Management at GWDG

The GWDG has been providing FDO related services since many years. It started by providing Handle services for a variety of European projects under the umbrella of ePIC. It extended its services by offering Data Type Registry services for interested institutions and added also a service to register PID Profiles. These two services are now being used by several projects as well.

2. KIT FAIR DO Ecosystem Testbed

KIT is developing a generic ecosystem to create, update, validate and search PIDs. PIDs are registered using well-defined PID profiles and kernel attributes, both being registered at the data type registry managed by GWDG. In doing so PID resolution not only delivers predictable results but also type-value pairs that can be interpreted by machines.
In addition, KIT offers layer services that allow to validate machine actionability of PID records, that send unified messages to brokers about new and updated PID records enabling for example searches. All code is released under an open-source license and available on github[43].

3. Systematic WF Documentation at IU

At IU one of the FDO related projects is focussing on the systematic creation of proper PIDs for all Digital Objects including data, software, workflow scripts etc. with the goal to improve reproducibility. Handles are being generated according to a registered profile that include defined kernel attributes, all being registered in a Data Type Registry making PID resolution predictable and machine actionable. The DOIP protocol is being used to interact with the created FDOs.
A VRE is being used to orchestrate and execute the workflow and a search service has been implemented to quickly find FDOs.

[43] https://github.com/kit-data-manager/testbed4inf

4. Repository Integrator Demo

One of the major challenges is the heterogeneity of the many repositories with respect to the way data and metadata are being organised and to the technologies being used. The use of DOIP implementing an interaction with FDOs independent of all these organisational and technological differences would reduce the effort of integration to a 1*N problem. In this demo case two different repositories were integrated using DOIP. For the B2SHARE repository an adapter was developed. Two aspects are important: (1) This adapter could now be used by all B2SHARE instances to integrate them in the FDO domain and (2) effectively developing the adapter cost an experienced programmer less than a week.

5. Work in Climate Modeling

The climate science community is assigning proper PIDs systematically and according to a profile to all digital objects being created and exchanged where the kernel attributes have been specified by a community wide process. Using a Data Type Registry to register the used kernel attributes and the profile(s) are in the process of being implemented. The Handle service has been embedded in a queuing setup to map between the volume and speed requirements with which digital objects are being created during the simulations on HPC systems and the capacity of the Handle service.

6. Work at NIST

NIST FDO work is in an early stage with the intention to make material science work more efficient. Handles are being created systematically for all kinds of resources being used (samples, files, instruments, etc.) and an FDO model is in development with specific emphasis on suitable metadata per object type.

7. DiSSCo Digital Specimen Repository

DiSSCo developed a model for a Digital Specimen compliant with the FDO Framework requirements indicating that a few requirements are not yet specific enough. The model was then used to build the Digital Specimen Database. One of its main characteristics is the use of operations to access different sub-parts of the FDO such as extensive metadata, large data files, etc. The PID resolves to a set of typed kernel attributes and these types are linked with operations all being identified by Handles.

8. DiSSCo Work with CORDRA

The “FAIR by Design” principle implemented in the DiSSCo architecture design requires to add a semantic validation pipeline to the CORDRA based DO store. For this purpose a JSON representation which is a valid serialisation format for RDF has been developed and a CORDRA trigger yields the execution of a microservice to perform the RDF validation and in case of success starts the subsequent steps in the objects life cycle. In this way it is ensured that digital specimens are stored in a semantic form machines can process.

9. Switchboard Work in CLARIN

Audited centers are requested to assign PIDs and metadata to all data they are hosting. PIDs are created according to a profile making use of registered kernel attributes. A Switchboard tool allows to map data collections and software tools based on community agreed metadata descriptions to facilitate usage by end users. The use of the DOIP protocol is in progress to unify the access to the different resources and the registration of kernel attributes and PID profiles will improve machines actionability.

10. RO Crate

The RO Crate contribution is a slightly different but nevertheless important contribution in so far as it offers a framework to bundle digital objects of different types as they typically appear in workflow scenarios into collections and describe them with metadata that are based on well-defined and registered categories based on an already existing framework. Such bundles help increasing re-

usability and reproducibility. RO Crate is now being extended to make these bundles FAIR Digital Objects to increase its exchangeability.

5.3 Summary

	PIDs	PID Profile	Kernel Types	DTR	FDO	DOIP	CORDRA	RDF validation
1. GWDG services								
2. KIT								
3. IU								
4. Rep Int.[44]								
5. Climate								
6. NIST								
7. DiSSCo Spec								
8. DiSSCo Cordra								
9. CLARIN								
10. RO Crate								

From the table it is obvious that for almost all projects following the DO approach there is an interest to implement an FDO domain which is achieved by systematically associating PIDs in the sense of globally unique and resolvable persistent identifiers, and using PID Profiles and Kernel Attributes defined and registered in an open Data Type Registry. The CORDRA reference system is being used by some as default system to register data types and PID profiles. DiSSCo has chosen CORDRA to store its Digital Specimen Objects and it was used in the repository integration testcase as second data repository since it did not require to develop an adapter. DOIP is being tested in a few cases as universal interface between digital objects ignoring all repository/registry differences.

All projects working with PID profiles and kernel attributes are satisfied by having registered the specifications in data type registries as defined by RDA. Machine actionability is given by the fact that the resolution of a PID delivers the reference to the profile and other attribute-value pairs with attributes being defined and registered. Yet only DiSSCo makes use of linking kernel attribute types with operations which is an interesting option. Most other projects assume typed attributes in metadata descriptions to allow interpretations and connection to operations.

Several projects include validation aspects at various levels ranging from correctness of PID assignments up to the machine actionability of FDOs as repository entries. RO Crate as an already existing description standard for collections that emerged in the context of workflows is now adapting to support the FDO requirements.

5.4 Challenges Ahead

In this chapter we highlight some challenges that need to be addressed in the coming months to come to an integrated domain of FDOs. It should be noted that the following statements are also being inspired by the work on Canonical Workflows for Research[45] where about 30 teams described their workflow solutions with several teams describing how FDOs could be used as a unified and easy to exchange documentation standard.

- With one exception we only report about projects that follow the DO approach towards FDOs. In the next phase we also should look for projects using the LD approach and address

[44] The repository integration experiment focussed on the development of a DOIP Proxy to create an integrated domain of DOs. Therefore, the use of PID profiles and Kernel Types was not an issue.
[45] https://osf.io/2cy86/

the challenges of interoperability across the two approaches. The exception is RO-Crate which emerged in an environment inspired by LD.

- Increasingly more projects are being carried out that are implementing and testing FDO related technologies. However, they work mainly on individual solutions, i.e., we clearly miss a more coordinated approach where developed software snippets can be reused.
- The document on FDO Configurations and this project overview indicate that there are various ways to implement FDOs bearing the risk that interested people are overwhelmed by the possible choices. Therefore, the FDO Forum should come up with recipes for various steps such as for example establish an FDO compliant repository. Default configurations and mechanisms need to be identified.
- Some people have the impression that it will take much effort to be part of an integrated FDO domain. We need to indicate that in cases where, for example, repositories have been setup in well-structured way the steps that need to be taken to integrate them into the global FDO domain imply moderate efforts.
- An aspect of great interest is the option of FDO to do encapsulation, i.e., to link FDO types with operations. Yet, this has not been tested widely, i.e., new projects are needed to understand the full power of DOIP.
- The other big challenge is to create an integrated domain of FDOs of a critical mass allowing to apply DOIP across repositories of different disciplines etc. This implies that more repositories need to be adapted to FDOs by using DOIP adapters or proxies to the repository software systems.
- It is obvious that decisions need to be taken to stabilise the DTR model, to specify the way PID profiles can be found by machines, to establish processes to define the FDO-wide kernel attributes and the way to incorporate well-defined community attributes.

6. Conclusions

The concept of Digital Objects, which was documented [11] in 1995 by Kahn & Wilensky, was developed to meet a need for a unified approach to data being exchanged on top of the Internet protocol stack. This work included the design and deployment of the Handle System which is now broadly used in industry (publishers, film industry, etc.) and by many research centres that need to do proper management of large volumes of data. The underlying idea of digital objects evolved slowly for many years due to the great success of the Web protocol stack. It was accelerated by the appearance of cloud systems, which are essentially “object stores”, and in 2013 when the Research Data Alliance was founded and several of its working groups started delivering concrete results for the research community. The Data Foundation and Terminology Group defined the Data Core Model which is basically an elaboration of the Digital Object model based on insights in research data management (RDM). The Data Type Registry Working Group defined the DTR Model which is at the basis of several implementations now. A few more RDA groups could be mentioned which contributed to the revival of the DO concept.

In 2016 the Nature paper about the FAIR Principles [8] was published which would soon be accepted worldwide. Since they could be interpreted in a variety of ways, soon many projects and implementations claimed to be FAIR compliant with the consequence that the term “FAIR” alone could not be used anymore to drive RDM out of its crisis which can best be described by the huge inefficiencies due to data wrangling [12]. The merger of the Digital Object Model and the FAIR principles agreed upon in the Paris workshop intensified the discussions about the FAIR Digital Objects as a way out of this crisis. Since then we see an increasing number of projects taking up the concept as sketched in this paper. Basically, it is just one year of hard work that brought the FDO community up to speed and gave directions for implementations.

Yet we need to admit that the projects described in this paper and many others are widely designed and implemented in isolation. It is the priority in the coming months to bring the different developmental groups closer together and design a joint large demonstrator and to provide a testbed with validators that allows arbitrary groups which developed components to check FDO compliance. The FDO Forum is working hard to advance the specification work so that we can assume that in 2022 a full-fledged specification will become available. A collaboration with classical standardisation organisations has already been started to turn FDO Forum specifications to international standards. The implementations described in this paper and others FDO related projects will help to guide the FDO Forum in its specification work.

References

[1] Smedt, Koenraad de, Koureas, Dimitris, Wittenburg, Peter (2020). FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. MDPI Publications. https://www.mdpi.com/2304-6775/8/2/21

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

[3] Hodson, Simon, et.al. (2018). Turning FAIR into reality. https://ec.europa.eu/info/sites/default/files/turning_fair_into_reality_1.pdf

[4] https://op.europa.eu/en/publication-detail/-/publication/3136c3e6-4f07-11eb-b59f-01aa75ed71a1

[5] Broeder, Daan, Budroni, Paolo, et al. (2021). SEMAF: A Proposal for a Flexible Semantic Mapping Framework (1.0). Zenodo. https://doi.org/10.5281/zenodo.4651421

[6] Tobias Weigel, Michael Lautenschlager, & Martin Juckes. (2018). Persistent Identifiers for CMIP6: Implementation plan. Zenodo. https://doi.org/10.5281/zenodo.4751913

[7] Islam, S., Hardisty, A., Addink, W., Weiland, C. and Glöckler, F., 2020. Incorporating RDA outputs in the design of a European Research Infrastructure for natural science collections. *Data Science Journal*, *19*(50), pp.1-14. http://doi.org/10.5334/dsj-2020-050

[8] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[9] Annika Jacobsen, Ricardo de Miranda Azevedo, et al. FAIR Principles: Interpretations and Implementation Considerations. *Data Intelligence* 2020; 2 (1-2): 10–29. doi: https://doi.org/10.1162/dint_r_00024

[10] Jose E. Labra Gayo, Eric Prud'hommeaux, Iovka Boneva, Dimitris Kontokostas (2018) Validating RDF Data, Synthesis Lectures on the Semantic Web: Theory and Technology, Vol. 7, No. 1, 1-328, DOI: 10.2200/S00786ED1V01Y201707WBE016, Morgan & Claypool

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

[12] Peter Wittenburg, George Strawn (2018). Common Patterns in Revolutionary Infrastructures and Data. http://doi.org/10.23728/b2share.4e8ac36c0dd343da81fd9e83e72805a0

[13] Stian Soiland-Reyes, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M. Fernández, Daniel Garijo, Björn Grüning, Marco La Rosa, Simone Leo, Eoghan Ó Carragáin, Marc Portier, Ana Trisovic, RO-Crate Community, Paul Groth, Carole Goble (2022): Packaging research artefacts with RO-Crate**.** *Data Science* (pre-press).https://doi.org/10.3233/DS-210053