# Connecting Repositories to one Integrated Domain

Peter Wittenburg[‡], Christophe Blanchi[§], Daan Broeder[|]

‡ Unaffiliated, Berlin, Germany
§ DONA, Geneva, Switzerland
| CLARIN, Utrecht, Netherlands

Corresponding author: Peter Wittenburg (peter.wittenburg@mpcdf.mpg.de)

## Abstract

Information is the new commodity in the global economy and trustworthy digital repositories will be the key pillars within this new ecosystem. The value of this digital information will only be realised if these repositories can be interacted with in a consistent manner and their data accessible and understandable globally. Establishing a data interoperability layer is the goal of the emerging domain of Digital Objects. When considering how to proceed with designing this interoperability layer, it is important to state that repositories need to be considered from two different perspectives:

1. Repositories are a reflection of the institutions that make them operational (quality of service, skilled experts, accessible over many years, appropriate data management procedures).
2. Repositories are computational services that provide a specific set of functions.

Complicating the effort to make repositories accessible and interoperable across the global is that many existing repositories have been developed in the past decades using a wide range of heterogeneous technologies, organisation of data and functionality. Many of these repositories are their own data silos and not interoperable. It is important to realise that much money has been invested to build these repositories and therefore we cannot expect that they will make large changes without great incentives and funding. This heterogeneity is the core of the challenge in making digital information the new commodity in the emerging global domain of digital objects.

This paper will focus on the functional aspects of repositories and proposes the FAIR Digital Object model as a core data model for describing digital information and the use of the Digital Object Interface Protocol (DOIP) to establish interoperable communication with all repositories independently of the respective technical choices. It is the conviction of this paper's authors that this integration of the FDO model and DOIP with existing repositories can be performed with minimal effort and we will present examples that document this claim.

We will present three examples of existing integration in this paper:

- An integration of B2SHARE
- A CORDRA repository
- Integration of the DOBES archive

B2SHARE is a repository that has assigned Persistent Identifiers (PIDs) (Handles) to all of its digital files. It allows users to add metadata according to a unified schema, but also has the possibility for user communities to extend this schema. The API allows one to specify a Handle which then gives access to the metadata and/or the bit sequences of the DO. It should be noted that B2SHARE allows one to include a set of bit-sequences being linked with the Handle. The integration consists of building a proxy that would provide a DOIP interface to B2SHARE to streamline the integration of the data and metadata into a single DO. The development of the proxy was relatively simple and did not require any changes on behalf of the B2SHARE repository. CORDRA is a CNRI repository/registry/registration system that manages DO, assigns Handles to all its DOs and is accessible through DOIP. For all intents and purposes, it implements many of the features from the Digital Object Architecture.

The integration of the two repositories enables copying files or movíng digital objects. In the case of copying files (metadata and bit sequences) from B2SHARE to CORDRA, for example, all functionality of the CORDRA service such as searching would become possible. Important is that in this case the PID record identifying the digital object in the B2SHARE repository would have to be extended to point to the alternative path, and the API of B2SHARE would have to offer the alternative access paths to a client. This latter aspect has not been implemented. Moving a DO from B2SHARE to CORDRA would result in changing the ownership of the PID and adding the updated information about the DO.

This adaptation was not done yet, but since this archive has some special functionalities, it is interesting to discuss the way of adaptation which could be chosen. In the DOBES archive each bundle of closely related digital objects is assigned a Handle and also metadata is treated as a digital object, i.e., it has a separate Handle. For management reasons and especially for enabling different contributors to maintain control of access rights, a tree structure was developed to allow contributors to organise their data according to specific criteria and users to browse the archive in addition to execute searches on the metadata.

While accessing archival objects is comparatively simple, the ingest/upload feature is more complex. It should be noted that the archive supports establishing a canonical tree of resources to define scopes for authorisation (define who has the right to grant access permissions, etc.), and facilitating lookup by supporting browsing according to understandable criteria. Therefore, depositors need to specify where in the tree the new resources should be integrated, and which initial rights are associated with them. After uploading the gathered information into a workspace, the archive carries out many checks in a micro-workflow: metadata is checked against vocabularies and partly curated, types of bit-sequences are checked and aligned with the information in the metadata, etc. An

operation has been developed which is called gatekeeper to ensure a highly consistent archive despite the many (remote) people contributing to its content. Thus, the archive requires a set of 4 information units being specified:

1. the set of bit-sequences to be uploaded,
2. the metadata describing the bundle,
3. the node to be used to organise the resources and
4. the initial rights where the default would be “open”.

Adapting this archive to DOIP would imply that the proxy provides a set of operations such as “ingest a complex object”, “update metadata”, “add another bit-sequence to a specific object”, “get me the list of operations”, “give me the metadata”, etc. A client must be developed to do the front-end interaction with a user allowing them to specify the required information and to choose a suitable operation. Then the client would have to interact with the repository via DOIP by starting, for example, the gatekeeper as an external operation.

## Keywords

FAIR, FAIR Digital Objects, Data Management, Data Infrastructure

## Presenting author

Christophe Blanchi

## Presented at

First International Conference on FAIR Digital Objects, presentation

## Conflicts of interest