1st Conference on Research Data Infrastructure

Poster presentations II

https://doi.org/10.52825/CoRDI.v1i.374

Published: 07 Sept. 2023

FDO to Structure the Domain of Knowledge

Peter Wittenburg[1][https://orcid.org/0000-0003-3538-0106] and Dimitris Koureas[2]

[1] FDO Forum, International

[2] Naturalis, Netherlands

Abstract. The Globally Integrated Dataspace will evolve and result in a domain full of digital artefacts and relations which in its size and complexity is unprecedented. Traditional methods of structuring the domain such as file structures are not sufficient any longer. On top of a layer of data centres and federations which need to implement mechanisms establishing trust, there will be layer of Virtual Collections serving different purposes. These VCs to a large extent need to be FAIR and persistent as well, since they include important information. We claim that the concept of FAIR Digital Objects is an excellent framework to build and manage these VC. Currently, there are a few implementations of the FDO concept.

Keywords: FAIR Principles, FAIR Digital Objects, Data Management

1. Introduction

The increasing volumes of data and their exponentially increasing interdependencies of different types create a complexity which is challenging for management and reuse in the evolving Globally Integrated Dataspace (GIDS) [1]. The widely agreed goal is that all different types of digital artefacts (data, metadata, software, semantic assertions, configurations, etc.) in this GIDS should be FAIR [2], and we also assume that we have mechanisms in place that guarantee improving responsibility, accountability, and persistence (RAP) for establishing trust. But this will not be sufficient for users to easily navigate and operate in this almost endless space. We will need to include structure at various levels to master the complexity.

Creators of digital artefacts create “canonical structures'', which are mostly hierarchical collections created with specific views in mind to help navigation and management. A typical tree structure in experimental science looks like “*experimenter/experiment-name/experiment-day/ measurements*”. In observational research the structures that are being created by the creators are much more varied such as “*project/language/fieldtrip-name/day/observation-type/observations*”. In all cases, the individual observations or measurements are the leaves in such a tree. Another structuring element is typically the repository that hosts and manages this collection. In an IT sense, this just adds another layer to the chosen trees “*repository/project/language/...*”.

Assuming FAIR data, we can expect that increasingly “rich metadata” will be available describing the different leaves in the tree, i.e., in principle no-tree structure would have to be provided, but queries with a certain profile would help to get a result list which is an unstructured bag of digital artefacts. For some operations, such a list might be sufficient, but it will not help for many others. What we see often is that researchers want to create virtual collections following their own views on top of the leaves and perhaps reusing certain structural elements from canonical structures. With reusing digital artefacts from different disciplinary contexts for

different purposes, it is crucial to provide mechanisms to create, manage, exchange, and preserve such virtual collections. It should be noted though, that the leaves and even sub-structures in these virtual collections often will change over time – collections are living bodies.

2. Examples

In this section, we want to discuss a few examples of such recursively defined complex collections.

In the DOBES project a repository was created containing all material from about 250 researchers worldwide coming from highly different disciplines [3]. In addition to an agreed and well-defined metadata set, the creator teams wanted to define their own canonical hierarchical collection structures to simplify navigation and management, for example, to easily define rights. A simple search on specific types would not be sufficient. Also, the repository managers used the canonical tree for some operations, such as transforming all data items of type X in a certain sub-collection to a new type Y. But these canonical collection structures were not informative for other researchers who, for example, wanted to carry out an intonation analysis and comparison between languages. They wanted to create their own tree structure to facilitate the comparison and to document this structure to make it easily citable.

Another example of implying structure on a huge set of distributed data is the biodiversity digital specimen [4]. At many labs worldwide, different digital information about a specific physical object is being created, and it is related with many other objects across institutions due to a variety of classification schemes. In this example we have two canonical structures: (1) all information which is about a specific physical object, (2) different classification schemes to group the digital twins according to some criteria. In one case the leaves are the information sources about an object, in the other case the leaves are the digital twins. Also, in this case we can foresee that researchers want to carry out specific operations on virtual collections constructed according to their own criteria.

A last example shall briefly be indicated. Increasingly more people see the need to extract the major assertions in papers to create nano-publications, which are basically augmented RDF triples [5]. From insights about Medline, for example, it can be easily estimated that the number of such nano-publications will increase exponentially in the coming decade either by manual or automatic extractions. Again, we will need to create structures to be able to navigate or operate in such a huge space of semantic assertions. Different sets of criteria will be used to determine key concepts in such spaces as a start to form structures resulting in many different views on semantic spaces fit for specific purposes.

All these different virtual collections representing different views will have a high scientific value and many of them need to be preserved despite changes over time. They will be the basis of proper management and in addition, researchers and specialised brokers will put efforts in the creation of meaningful virtual collections that will have a value in themselves and will be reused, extended, changed etc.

3. Relevance of FDOs

FAIR Digital Objects (FDO) are atomic, self-containing units of information that persistently bundle all information needed for FAIRness [6]. They can be leaves but also collections due to their recursive definition which would mean that the body of each collection is the set of included elements, that the collection is assigned a PID and is associated with collection metadata. Therefore, FDOs are excellent mechanisms to organise this almost endless space of virtual structures, make them persistent, track their changes over time, and share them with others independent of the particular dataspace people are working in. FDOs are neutral with respect to the structuring of the body, i.e., it does not care how the different elements are described and referenced if the specifications will be machine actionable. Description standards such as RO Crate [7] could be used here. The type of the FDOs containing collections is

"collection" and a subtype could be "encoded_by_RO-Create" to enable machines to parse the structures.

Author contributions

Peter Wittenburg and Dimitris Koureas both contributed to the whole paper.

Competing interests

The authors declare that they have no competing interests.

Acknowledgement

We would like to thank all the colleagues who contributed to the DOBES and DiSSCO work and the growing community pushing the FDO specifications and implementations.

References

1. P. Wittenburg, G. Strawn, "Shaping and standardising the Global Integrated Data Space", https://sites.grenadine.co/sites/iot/en/iotweek-2022/schedule/8581/Shaping%20and%20standardising%20the%20Global%20Integrated%20Data%20Space The FAIR Guiding Principles for scientific data management and stewardship
2. M. Wilkinson, et al., "The FAIR Guiding Principles for scientific data management and stewardship", https://www.nature.com/articles/sdata201618
3. DOBES Archive, https://dobes.mpi.nl/
4. DiSSCO, "What is a digital specimen?", https://dissco.tech/2020/03/31/what-is-a-digital-specimen/
5. B. Mons, J. Veltrop, "Nano-Publication in the e-science era", https://ceur-ws.org/Vol-523/Mons.pdf
6. G. Strawn, P. Wittenburg, "FDO Requirement Specifications", https://zenodo.org/record/7781926#.ZEjWP87P3b0
7. RO-Crate, "Research Object Crate", https://www.researchobject.org/ro-crate/