1st Conference on Research Data Infrastructure

Poster presentations II

https://doi.org/10.52825/CoRDI.v1i.406

Published: 07 Sept. 2023

Enhancing Reproducibility in Research through FAIR Digital Objects

Zeyd Boukhers[1,2] [https://orcid.org/0000-0001-9778-9164], and Leyla Jael Castro[3] [https://orcid.org/0000-0003-3986-0510]

[1] Fraunhofer Institute for Applied Information Technology, Sankt Augustin, Germany

[2] University Hospital of Cologne, Cologne, Germany

[3] ZB MED Information Centre for Life Sciences, Cologne, Germany

Abstract:

The FAIR principles were introduced to enhance data reuse by providing guidelines for effective data management practices. In the broader context of research, assets encompass not only data but also other artifacts such as code, software, and publications. FAIRifying these artifacts is as essential as FAIRifying data, especially in Data Science and Artificial Intelligence, where the complexity of current AI approaches makes reproducibility extremely challenging. Therefore, facilitating the easy reuse of these artifacts represents a significant stride towards mitigating this problem. The concept of FAIR Digital Objects (FDOs) presents a solution to FAIRify these artifacts, treating them as FDOs. NFDI4DataScience is embracing FDOs and proposing an architecture to efficiently manage them.

Keywords: FAIR Digital Object, Reproducibility, Research Data, Metadata, NFDI4DS

1 Background

Since introducing the FAIR principles in 2016 [1], they have gained widespread acceptance among data management professionals, projects, and initiatives, including the EOSC roadmap (European Commission 2018). These principles were later expanded to encompass other digital objects (e.g., software [2] and workflows [3]). Recognizing the importance of these principles such as making metadata available, significant efforts have been mobilized to motivate researchers to FAIRify their resources and supply metadata, with various strategies employed to extract metadata from existing resources such as publications [4]. However, it's become clear that availability alone is insufficient to achieve the desired level of reproducibility and reusability. Metadata must not only be accessible but also complete, accurate, and readily findable. The FAIR Digital Objects (FDOs) emerged as a more detailed version, where objects, their metadata and their materialisations are clearly separated. An FDO is a conceptual approach [5] that offers a technical solution for implementing FAIR principles across various digital

object types [6], [7] while also adding an operations layer boosting interoperability beyond the initial FAIR possibilities as machines can understand and directly use those operations to (semi)automatically use the objects and connect them with each other via operational workflows.

Numerous initiatives have emerged in Europe and beyond to establish a comprehensive infrastructure ecosystem for research data, such as the EOSC at the European level and NFDI at the German national level. It is crucial that these infrastructures are founded on robust fundamentals, like those offered by FDOs. The implementation of FDOs helps ensure that research digital objects are findable, accessible, interoperable, and reusable, thus paving the way towards improved transparency, and reproducibility in data-driven research disciplines.

2 FDO Concept

The adoption of FAIR Digital Objects (FDOs) enables researchers and institutions to effectively manage digital objects while ensuring compliance with evolving research management standards. This, in turn, enables more effective sharing and collaboration among researchers, fostering innovation and accelerating scientific discovery. Additionally, as more funding agencies and publishers begin to require adherence to FAIR principles, the adoption of FDOs will become increasingly important to secure funding and disseminate research findings. Consequently, the implementation of FDOs can create a foundation for more transparent, efficient, and collaborative research processes that ultimately benefit the entire scientific community.

Recognizing the importance of FDOs, the NFDI4DataScience[1] consortium is adopting them to manage research artefacts within the realm of data science, including essential components such as datasets, publications, software, and models. The objective is to assist researchers in easily locating and utilising the resources they need, whether they correspond to the metadata or the actual object in a particular format together with some basic operations on top of them. This is particularly useful when the objects are private as their metadata will still be open and available to others.

3 FDO Architecture

In accordance with the FDO specifications[2] released by the FDO Forum[3], we propose an architecture[4] that adheres to both FAIR principles and FDO specifications while encompassing all artifacts in a data science workflow. As illustrated in Figure 1, the architecture begins with a Registry of FDOs that stores all the PID records along with minimal metadata records such as the FDO type.

Each PID in the registry resolves to a more extensive metadata piece containing the FDO Profile, the PID of the metadata, and the PID of the digital object itself (e.g., publication, dataset). The rationale for having separate PIDs for the digital object and metadata is to comply with the FDO specification, which emphasizes separating the metadata from the digital object and ensuring that metadata persists even if the digital object does not.

[1] https://www.nfdi4datascience.de/
[2] https://fairdo.org/specifications/
[3] https://fairdo.org/
[4] https://fdda1.gitlab.io/fdom

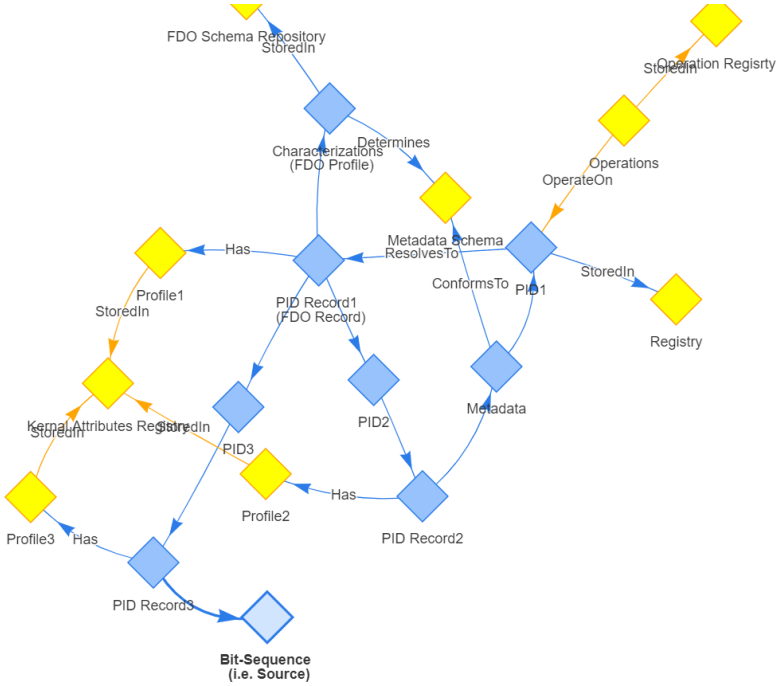

Figure 1. FDO Architecture. See footnote 4

While the digital object PID resolves to the bit-sequence, the metadata PID resolves to the corresponding metadata of the FDO, which follows predefined metadata schemas. These schemas are determined by the FDO profile, such as software or dataset, and obtained from Schema.org[5]. Additionally, a set of operations from a predefined registry is assigned to each FDO based on its profile.

This architecture fosters compliance with FAIR principles and FDO specifications, ensuring that all artifacts in a data science workflow are managed effectively. An example of this architecture can be found here[6].

Author contributions

Z.B: conceptualization, writing – original draft, writing – review & editing. LJC: conceptualization, project administration, writing – review & editing.

Competing interests

The authors declare that they have no competing interests.

Funding

This work received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number: NFDI4DataSciene (460234259).

[5] https://schema.org/
[6] https://ai-research.net/FDO_Example.html

References

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[2] M. Barker, N. P. Chue Hong, D. S. Katz, *et al.*, "Introducing the fair principles for research software," *Scientific Data*, vol. 9, no. 1, p. 622, 2022.

[3] C. Goble, S. Cohen-Boulakia, S. Soiland-Reyes, *et al.*, "Fair computational workflows," *Data Intelligence*, vol. 2, no. 1-2, pp. 108–121, 2020.

[4] Z. Boukhers, N. Beili, T. Hartmann, P. Goswami, and M. A. Zafar, "Mexpub: Deep transfer learning for metadata extraction from german publications," in *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, IEEE, 2021, pp. 250–253.

[5] K. De Smedt, D. Koureas, and P. Wittenburg, "Fair digital objects for science: From data pieces to actionable knowledge units," *Publications*, vol. 8, no. 2, p. 21, 2020.

[6] E. Schultes and P. Wittenburg, "Fair principles and digital objects: Accelerating convergence on a data infrastructure," in *Data Analytics and Management in Data Intensive Domains: 20th International Conference, DAMDID/RCDL 2018, Moscow, Russia, October 9–12, 2018, Revised Selected Papers 20*, Springer, 2019, pp. 3–16.

[7] U. Schwardmann, "Digital objects–fair digital objects: Which services are required?" *Data Science Journal*, vol. 19, no. 1, 2020.