NFDI4DataScience registry for reproducible Data Science and Artificial Intelligence

Leyla Jael Castro [1, 3 [0000-0003-3986-0510]], Zeyd Boukhers [3 [0000-0001-9778-9164]], Olga Giraldo [1, 3 [0000-0003-2978-8922]], Adamantios Koumpis [2, 3], Oya Beyan [2, 3 [0000-0001-7611-3501]], Dietrich Rebholz-Schuhmann [1, 2, 3 [0000-0002-1018-0370]]

1 ZB MED Information Centre for Life Sciences
2 Faculty of Medicine, University of Cologne
3 NFDI4DataScience consortium

Scientific advances are built on previous work, and for it, all the involved pieces are needed, i.e., goals, methods, results (usually reported as a scholarly publication), data, software and any other Digital Object (DO) used in the research process. The Findable, Accessible, Interoperable and Reusable (FAIR) Principles [1] provide a metadata-based approach for research to improve on these four dimensions. Although aiming to cover all sorts of (research) DOs, the FAIR principles mainly focus on data. Additional efforts to adjust and extend their coverage to software [2], workflows [3] and machine learning [4] have been established in the last few years. Despite improvements brought by FAIRification efforts, e.g., DOs are more findable nowadays, there are scientific desirable aspects, such as reproducibility, which are beyond the scope of FAIR and still pose a challenge [5]. In the case of Data Science (DS) and Artificial Intelligence (AI), the discussion around FAIR, reproducibility and other *ilities is a recent ongoing effort.

The (German) National Research Data Infrastructure for Data Science consortium (NFDI4DS) brings together 16 partners aiming at creating an infrastructure to support interdisciplinary research involving DS and AI. In particular, the consortium will create a registry encompassing metadata for a variety of DOs involved in AI approaches, including data and its configuration for a particular approach, software used to process the data together with its hyper-parametrization, underlying AI model, evaluation process and additional elements supporting benchmarking, comparability and reproducibility. The registry will build on top of FAIR metadata models for research data and software and the notion of FAIR DOs (FDOs). FDOs [6] have been proposed to improve access to DOs thanks to the formalization of their metadata, types, identifiers and the explicit declaration of their computational operations, making them actionable FAIR objects. We plan to use Research Objects Crates [7] to package research outputs along with their metadata and turn them into FDOs; implementation allows for a broad range of use cases, across scientific domains. As for the AI tailored metadata, the registry will formalize and extend the Data, Optimization, Model, and Evaluation recommendations for reporting supervised machine learning on computational biology [8] to cover further cases and disciplines.

As efforts around FAIR and reproducibility for AI are still recent, and the NFDI4DS has just started this year 2022, there are open questions including what aspects of reproducibility can be tackled with metadata, what dimensions can be used to compare AI approaches (possibly)

across disciplines, and how much metadata can be obtained by automatic means. In this workshop, we want to promote an open discussion with the community around these topics.

Funding
All the authors are partially funded by a German Research Foundation DFG grant corresponding to the NFDI4DataScience project.

References

1. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3: 160018. doi:10.1038/sdata.2016.18
2. Chue Hong NP, Katz DS, Barker M, Lamprecht A-L, Martinez C, Psomopoulos FE, et al. FAIR Principles for Research Software (FAIR4RS Principles). 2022. doi:10.15497/RDA00068
3. Goble C, Cohen-Boulakia S, Soiland-Reyes S, Garijo D, Gil Y, Crusoe MR, et al. FAIR Computational Workflows. Data Intelligence. 2020;2: 108–121. doi:10.1162/dint_a_00033
4. Castro LJ, Katz DS, Psomopoulos F. Working Towards Understanding the Role of FAIR for Machine Learning. PUBLISSO; 2021. doi:10.4126/FRL01-006429415
5. Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016;533: 452–454. doi:10.1038/533452a
6. Schultes E, Wittenburg P. FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure. In: Manolopoulos Y, Stupnikov S, editors. Data Analytics and Management in Data Intensive Domains. Cham: Springer International Publishing; 2019. pp. 3–16. doi:10.1007/978-3-030-23584-0_1
7. Soiland-Reyes S, Sefton P, Crosas M, Castro LJ, Coppens F, Fernández JM, et al. Packaging research artefacts with RO-Crate. Data Science. 2022; 1–42. doi:10.3233/DS-210053
8. Walsh I, Fishman D, Garcia-Gasulla D, Titma T, Pollastri G, Capriotti E, et al. DOME: recommendations for supervised machine learning validation in biology. Nature Methods. 2021. doi:10.1038/s41592-021-01205-4