

MLentory, an FDO registry for machine learning models

Dhwani Solanki¹, Nelson Quiñones¹, Dietrich Rebholz-Schuhmann^{1,2} and Leyla Jael Castro^{1,*}

¹ ZB MED Information Centre for Life Sciences, Cologne, Germany

² University of Cologne, Cologne, Germany

Abstract

Here we introduce MLentory, an FDO registry for Machine Learning models and their corresponding workflows, from creation to deployment. MLentory relies on FAIR Digital Objects (FDOs) to improve Findability, Accessibility, Interoperability, and Reusability while also improving reproducibility and transparency. MLentory aggregates, harmonizes and FAIRifies data from various ML model and model-related repositories and platforms. Here we present the initial architecture for data extraction, transformation, and loading.

Keywords

Machine learning models, FAIR, FDOs, registry

1. Background

Due to the proliferation of Machine Learning (ML) models, it is necessary a systematic approach to report them. To this end, the ML model cards were proposed in 2019 [1], complemented by Dataset cards [2] by providing additional information on the training datasets. A parallel and complementary effort are platforms facilitating the storing, sharing and reporting of ML models and other artifacts needed for training and deployment, e.g., HuggingFace, neptune.ai, SpaceML, Kipoi, BioImagine Model Zoo, etc. Interoperability across platforms and connection to other ML experiments and related artifacts are still challenging as the corresponding metadata is stored in different ways by different platforms. Harmonization across different efforts is a gap being addressed by different communities, e.g., Research Data Alliance (RDA) FAIR4ML Interest Group, ELIXIR ML Focus Group (ELIXIR MLFG), and National Research Data Infrastructure for Data Science (NFDI4DS) in Germany.


SW4THCLS 2024: Semantic Web Applications and Tools for Health Care and Life Sciences, February 26–29, 2024, Leiden, The Netherlands

* Corresponding author.

✉ ljarcia@zbmed.de (LJ Castro)

ORCID: [0009-0004-1529-0095](https://orcid.org/0009-0004-1529-0095) (D. Solanki); [0000-0002-5037-0443](https://orcid.org/0000-0002-5037-0443) (N. Quiñones); [0000-0002-1018-0370](https://orcid.org/0000-0002-1018-0370) (D. Rebholz-Schuhmann); [0000-0003-3986-0510](https://orcid.org/0000-0003-3986-0510) (LJ. Castro)

© 2024 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

2. MLentory

MLentory aims at providing a registry (aka directory, inventory) for ML models and corresponding workflows, from creation to deployment. MLentory relies on FAIR Digital Objects [3] to improve Findability, Accessibility, Interoperability, and Reusability (FAIR layer in FDOs) while also improving reproducibility and transparency (operations layer in FDOs). It will rely on metadata agreements reached by the aforementioned communities, mapped (whenever needed) to schema.org [4], a lightweight approach to semantics already considered by the scientific community for datasets and software. Here we introduce MLentory architecture together with an initial proposal for ML models metadata based on schema.org. MLentory relies on data harvesting from third-party platforms, with aggregation and harmonization modules for the final shape of the ML model FDOs. A scheduler is available to keep the inventory continuously updated. The data storage corresponds to an RDF graph with an ElasticSearch module for indexing and communication with the front-end interface and corresponding RESTful services.

3. Conclusions and future work

Here we have outlined the MLentory architecture to collect, aggregate and harmonize reporting of ML models together with the initial consideration for a possible metadata schema. We aim to share our framework to improve ML model metadata, paving the way for more robust and transparent ML practices.

Acknowledgements

This work has been partially supported by NFDI4DataScience, a consortium funded by the German Research Foundation (DFG), project number 460234259.

References

- [1] Mitchell M, et al. Model Cards for Model Reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019. doi:10.1145/3287560.3287596
- [2] Bender EM, Friedman B. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Transactions of the Association for Computational Linguistics. 2018. doi:10.1162/tacl_a_00041
- [3] De Smedt K, Koureas D, Wittenburg P. FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. Publications. 2020;8: 21. doi:10.3390/publications8020021
- [4] Guha RV, Brickley D, Macbeth S. Schema.org: evolution of structured data on the web. Communications of the ACM. 2016;59: 44–51. doi:10.1145/2844544