,

Assessment of metadata descriptors of AI-ready datasets

Jerven Bolleman[1], Leyla Jael Castro[2,*], Alban Gaignard[3], Agoritsa Kalampaliki[4], Edwin Jun Kiat Ong[5], Núria Queralt-Rosinach[6,*], Nelson David Quiñones[2], Rohitha Ravinder[2], Dhwani Solanki[2], David Steinberg[7], Claus Weiland[8] and Daphne Wijnbergen[6]

[1]*SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland*
[2]*ZB MED Information Centre for Life Sciences, Cologne, Germany*
[3]*Nantes Université, CNRS, INSERM, l'institut du thorax, Nantes, France*
[4]*Biomedical Sciences Research Center "Alexander Fleming", Vari, Attiki, Greec*
[5]*Queen's University of Belfast, Belfast, Northern Ireland, UK*
[6]*Leiden University Medical Center, Leiden, The Netherlands*
[7]*University of California Santa Cruz Genomics Institute, Santa Cruz CA, USA*
[8]*Senckenberg Nature Research Society, Frankfurt, Germany*

Abstract

To advance the use of Artificial Intelligence, including notably Machine Learning, for the understanding of diseases and conservation of biodiversity, it is important to promote FAIR AI-ready datasets. However, it is not clear how much AI-ready metadata is covered in well-known dataset repositories such as OpenML, Hugging Face or Kaggle. During the BioHackathon Europe 2024, we tackled this problem following a programmatic approach and applying Semantic Web technologies. Here, we show our preliminary results on the coverage of the implemented Croissant metadata format and discuss its implications in ML data management and future steps.

Keywords

Life Sciences, Machine Learning, AI-ready datasets, FAIR

1. Background

To advance the use of Machine Learning (ML) for the understanding of diseases and conservation of biodiversity, it is important to promote Findable Accessible Interoperable Reusable (FAIR) AI-ready datasets since data scientists and bioinformaticians spend 80% of their time in data finding and preparation. The aim to provide AI-ready datasets is to support ML analysis over complex and integrable data, yet the criteria for AI-readiness of biomedical data is currently under debate [1, 2]. The tagging of datasets from the Life Sciences is non-uniform across the platforms involved in this study, e.g., on Kaggle, we selected data using a combination of tags such as "Biology", "Earth and Nature", "Genetics" or "Disease" to select data with a thematic focus on Life Sciences. Furthermore, AI-ready datasets, whether by design or after pre-processing, can be enriched with metadata so they become FAIRer, with all benefits that come from FAIR. Metadata descriptors for datasets are pivotal for the creation of ML models as they facilitate the definition of strategies for data discovery, feature selection, data cleaning and data pre-processing. The Croissant metadata format, developed within the scope of the MLCommons initiative [3], is an extension of schema.org to better describe AI-ready datasets, released early 2024 and already adopted by several popular ML dataset repositories such as Hugging Face [4], Kaggle [5] and OpenML [6]. The Tensorflow Datasets utility library (TDFS) provides a CroissantBuilder class that allows instantiating

SWAT4HCLS 2025: 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences 2025
ljgarcia@zbmed.de (L. J. Castro); n.queralt_rosinach@lumc.nl (N. Queralt-Rosinach)
0000-0002-7449-1266 (J. Bolleman); 0000-0003-3986-0510 (L. J. Castro); 0000-0002-3597-8557 (A. Gaignard); 0009-0005-9948-2242 (A. Kalampaliki); 0009-0009-8910-444X (E. J. K. Ong); 0000-0003-0169-8159 (N. Queralt-Rosinach); 0000-0002-5037-0443 (N. D. Quiñones); 0009-0004-4484-6283 (R. Ravinder); 0009-0004-1529-0095 (D. Solanki); 0000-0001-6683-2270 (D. Steinberg); 0000-0003-0351-6523 (C. Weiland); 0000-0002-7449-6657 (D. Wijnbergen)

of Croissant-annotated data automatically within ML frameworks such as PyTorch or TensorFlow [7]. However, as it commonly happens with metadata, there are some limitations to the amount of metadata that can be automatically extracted and it is not clear how much metadata is covered in these well-known repositories of ML datasets ready to be loaded into the most popular ML frameworks.

Recently, Steinberg provided a subset of Hugging Face datasets marked with Croissant in RDF as a use case to study its semantic status for data usability in ML tools and to bridge ML with Semantic Web, identifying semantic issues that challenge cross-dataset research [8]. Moreover, the Life Sciences community is actively utilizing Semantic Web technologies for FAIR data management and analysis. Therefore, to address the aforementioned problems we formulated the following questions: How much Croissant metadata can be programmatically extracted from AI-ready datasets? How could this automation be improved? How much metadata is covered in data repositories? Is this metadata aligned to be leveraged in ML frameworks? Our hypothesis was that an SPARQL-based assessment of Croissant metadata coverage would enable us to explore answers to these questions. During the BioHackathon Europe 2024, we tackled this assessment following a programmatic approach and applying Semantic Web technologies. Our aim was to assess, understand and compare the metadata description status of ML datasets from major data providers. Here, we show our preliminary results on the coverage of the implemented Croissant metadata format and discuss its implications in ML data management and AI-readiness and future steps.

2. Results and Discussion

This project was developed at BioHackathon Europe, 4-8 November 2024 in Barcelona, which provided a unique opportunity to integrate different expertise, vision and tools. During an intense week of hybrid work, we delivered two main outcomes. First, a Croissant RDF knowledge graph built from Hugging Face, OpenML and Kaggle. Besides, we developed croissant-rdf Python tool, a Java API, stored the RDF files into the BioHackCloud for further analysis, and a FAIR assessment. Second, a coverage status assessed by SPARQL queries. The results identified inconsistent use of Croissant, with some mandatory properties missing. Although Croissant ML vr.1.0 was released less than one year ago (around March 2024), it is getting quick attention due to its potential to improve reuse of ML datasets. Despite its adoption by well-known AI-related platforms, there is still a need to improve the use of types and properties as recommendations given in Croissant, e.g., wrt minimum properties, are not yet fully followed. As future steps we aim to finish our analysis. Our findings can help to elucidate best practices and define criteria for the AI-readiness of biomedical datasets.

Acknowledgments

Thanks to the BioHackathon Europe 2024 organizers and to ELIXIR for accepting the project and supporting participation. LJC, NQ, RR and DS are partially funded by the German Research Foundation (Deutsche Forschungsgemeinschaft - DFG) as part of the NFDI4DataScience consortium under grant 460234259. NQR is partially funded as contributing partners of SYNTHIA and ERDERA projects. SYNTHIA is supported by the Innovative Health Initiative Joint Undertaking (IHI JU) under grant agreement No 101172872. The JU receives support from the EU Horizon Europe programme and COCIR, EFPIA, Europa Bio, MedTech Europe, and Vaccines Europe and DNV. ERDERA has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement N°101156595.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] M. Y. Ng, A. Youssef, A. S. Miner, D. Sarellano, J. Long, D. B. Larson, T. Hernandez-Boussard, C. P. Langlotz, Perceptions of data set experts on important characteristics of health data sets ready for machine learning: A qualitative study, JAMA Network Open 6 (2023) e2345892–e2345892. doi:10.1001/jamanetworkopen.2023.45892.

[2] T. Clark, H. Caufield, J. A. Parker, S. Al Manir, E. Amorim, J. Eddy, N. Gim, B. Gow, W. Goar, M. Haendel, J. N. Hansen, N. Harris, H. Hermjakob, M. Joachimiak, G. Jordan, I.-H. Lee, S. K. McWeeney, C. Nebeker, M. Nikolov, J. Shaffer, N. Sheffield, G. Sheynkman, J. Stevenson, J. Y. Chen, C. Mungall, A. Wagner, S. W. Kong, S. S. Ghosh, B. Patel, A. Williams, M. C. Munoz-Torres, AI-readiness for biomedical data: Bridge2AI recommendations, 2024.

[3] A. M., O. Benjelloun, C. Conforti, P. Gijsbers, J. Giner-Miguelez, N. Jain, M. Kuchnik, Q. Lhoest, P. Marcenac, M. Maskey, P. Mattson, L. Oala, P. Ruyssen, R. Shinde, E. Simperl, G. Thomas, S. Tykhonov, J. Vanschoren, J. van der Velde, S. Vogler, C.-J. Wu, Croissant: A metadata format for ml-ready datasets, in: Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning, ACM, New York, NY, USA, 2024.

[4] Q. Lhoest, A. Villanova del Moral, Y. Jernite, A. Thakur, P. von Platen, S. Patil, J. Chaumond, M. Drame, J. Plu, L. Tunstall, J. Davison, M. Šaško, G. Chhablani, B. Malik, S. Brandeis, T. Le Scao, V. Sanh, C. Xu, N. Patry, A. McMillan-Major, P. Schmid, S. Gugger, C. Delangue, T. Matussière, L. Debut, S. Bekman, P. Cistac, T. Goehringer, V. Mustar, F. Lagunas, A. Rush, T. Wolf, Datasets: A community library for natural language processing, in: H. Adel, S. Shi (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 175–184. URL: https://aclanthology.org/2021.emnlp-demo.21/. doi:10.18653/v1/2021.emnlp-demo.21.

[5] Kaggle, kaggle, 2024. URL: https://www.kaggle.com/, online, Accessed: 2024-11-04.

[6] J. Vanschoren, J. N. van Rijn, B. Bischl, L. Torgo, Openml: networked science in machine learning, SIGKDD Explor. Newsl. 15 (2014) 49–60. URL: https://doi.org/10.1145/2641190.2641198. doi:10.1145/2641190.2641198.

[7] Tensorflow, tensorflow datasets, a collection of ready-to-use datasets, 2024. URL: https://www.tensorflow.org/datasets, online; Accessed: 2024-11-07.

[8] D. Steinberg, J. Bolleman, Bridging machine learning and semantic web: A case study on converting hugging face metadata to RDF, 2024.