Working Towards Understanding the Role of FAIR for Machine Learning

Daniel S. Katz[1 [0000−0001−5934−7525]], Fotis E. Psomopoulos[2 [0000−0002−0222−4273]], and Leyla J. Castro[3 [0000−0003−3986−0510]]

[1] University of Illinois, Urbana IL 61801, USA
d.katz@ieee.org
[2] Centre for Research and Technology Hellas, Thessaloniki, Greece
fpsom@certh.gr
[3] ZB MED Information Centre for Life Sciences, Cologne, Germany
ljgarcia@zbmed.de

Abstract. The FAIR Guiding Principles aim to improve findability, accessibility, interoperability and reusability for both humans and machines, initially aimed at scientific data, but also intended to apply to all sorts of research digital objects, with recent developments about their modification and application to software and computational workflows. In this position paper we argue that the FAIR principles also can apply to machine learning tools and models, though a direct application is not always possible as machine learning combines aspects of data and software. Here we discuss some of the elements of machine learning that lead to the need for some adaptation of the original FAIR principles, along with stakeholders that would benefit from this adaptation. We introduce the initial steps towards this adaptation, i.e., creating a community around it, some possible benefits beyond FAIR, and some of the open questions that such a community could tackle.

Keywords: FAIR · Machine Learning · Metadata

1 Introduction

The idea of FAIR (findable, accessible, interoperable,and reusable) in the context of scientific data management and stewardship was developed in 2014 and turned into specific principles in 2016 [23]. Along the way, the idea was generalized in concept to apply to both data and other digital scholarly objects, but it has become clear in practice that what works for data may not directly work for all other digital objects. For example, both previous and ongoing work show that many of the guiding FAIR principles need to either be re-written or reinterpreted for software, and this is being done [6].

This position paper discusses the beginning of a process for extending the FAIR principles to machine learning (ML) models, which have characteristics of both data and software. Although an argument can be made that a combination of the existing FAIR principles for data and the FAIR principles for software could be used to address elements in machine learning, the authors suspect that this may not be sufficient. As such, additional work is needed to understand how to apply FAIR concepts to machine learning.

We have begun a community-building process to investigate this, specifically to:

1. identify relevant projects,
2. identify interested people, and
3. determine the next steps needed to understand how FAIR can be applied to ML.

The community-building process began with a virtual poster at the Research Data Alliance (RDA) virtual plenary 16 (VP16) meeting in late 2020 [15]. Next, we held a birds of a feather session at the RDA VP17 meeting in April 2021 [16], with the intent of discussing and gathering relevant projects and people. We also held a community call in June 2021 where we invited participants from the previous events and others who had been recommended in these events, as well as those who responded to a set of social media and Slack posts.

2 Context

While the work of defining how FAIR applies to machine learning can be done in part by information scientists, there are other stakeholders who need to be involved, starting with the researchers who train, share, and use machine learning models with a particular and well-defined goal defined by their own research topic. In some communities, these researchers can be found in conferences such as NeurIPS [17] or the International Conference in Machine Learning (ICML) [13]; special interest groups in professional societies nowadays commonly grouped under Data Science or Artificial Intelligence umbrellas; and via research projects that work in this space, e.g., FAIR4HEP [10], which, using high-energy physics (HEP) as the science driver, is developing a FAIR framework to advance understanding of AI, applying AI techniques, and exploring approaches to AI.

Another category of stakeholders is platforms that are relevant in machine learning. One example is DLHub [5], which lets a user find, share, publish, and run machine learning models and discover training data for science. Another is Kipoi [2], which is an API and repository of ready-to-use trained models for genomics. And a third is OpenML [21], which lets users build open source tools to discover (and share) open data, draw them into machine learning environments, build models, analyse results, get advice on better models.

Communities are another class of stakeholder. These include, for example, the Pistoia Alliance [18], a global, not-for-profit members' organization working to lower barriers to innovation in life science and healthcare R&D through pre-competitive collaboration; the National Research Infrastructure for Data Science (NFDI4DS) [11], a German national consortium aiming at delivering standards and tools to support all steps of the complex and interdisciplinary research data lifecycle, including collecting, creating, processing, analyzing, publishing, archiving, and reusing resources in Data Science and Artificial Intelligence; ELIXIR [8], an intergovernmental organisation that brings together life science resources (including databases, software tools, training materials, cloud storage and supercomputers) from across Europe, which has a dedicated focus group working on ML [9] that produced the DOME recommendations around it [22] with a clear connection to the FAIR principles; and CLAIRE [7], the Confederation of Laboratories for Artificial Intelligence Research in Europe.

3 FAIR

The FAIR Principles, at a high level, are intended to apply to all research objects; both those used in research and those that are research outputs. They claim to be applicable to "scholarly digital research objects." [23, 14]

But they actually focus on both metadata and data: the text of the principles often includes "(Meta)data ...", which is shorthand for "metadata and data ...". This is reasonable when discussing data, in part because metadata is a kind of data. However, when applied to other objects, this terse formulation breaks down, as some of the discussion of metadata for data translates to metadata for other objects, depending on the type of object, and some of the discussion of data translates to other objects, again depending on the object, but this is not uniform across all the FAIR principles.

There has been work in FAIR for Research Software and FAIR workflows focusing on specifically how to translate or interpret the principles for research software and workflows [6, 12]. But as of yet, there has been little work about how FAIR applies to ML, and specifically ML models. Are they data, for example, a set of parameters and options for a particular framework? Or are they software, for example, an executable object that takes input and provides output? Or, are they both or neither?

One important specific aspect is that the FAIR principles are applied to data via dataset creators and repositories. These two types of entities are collectively responsible for creating, annotating, indexing, preserving, and sharing the datasets and their metadata. However, while non-data objects can often be stored as data, they are not just data. While the high level FAIR goals are mostly the same for these other objects as they are for data, the details and how they are implemented depend on

- how the objects are created and used;
- how and where the objects are stored and shared; and
- how and where metadata is stored and indexed.

Thus, the FAIR guiding (detailed) principles developed for data are not directly applicable to machine learning models; work is needed to define, then implement, then adopt FAIR principles in this case.

Specifically, we can start by considering that large elements of the FAIR principles for data are dependent on archival repositories (e.g., Zenodo and others registered in re3data.org). These repositories hold data and/or metadata, and they provide search and access capabilities. Software is different, since it typically is not shared via archival repositories but instead via social coding platform (e.g., GitHub) and package management systems (e.g., PyPI, CRAN) so the assumptions in the principles about the role of repositories do not work for software; this is one of the main reasons different FAIR principles are needed for software. Although the practices for software are changing, both driven by the wider scientific community as well as by funding agencies and policy makers, the use of archival repositories (such as Zenodo) and registries (such as bio.tools) for software is still limited, and likely will remain so. The relationship between machine learning and archival repositories and other publishing and/or sharing platforms is in flux and has not yet reached a point of standardization, which makes the definition of FAIR principles for machine learning equally uncertain and in need of development and consensus.

4 Beyond FAIR

Although reproducibility is not part of the FAIR principles nor is comparability or explainability in the case of ML models, making ML and ML models FAIR will have an impact on these topics, at least at a basic level. A first step towards reproducibility is sharing and linking together data and software. If both data and software, e.g., ML tools, are FAIR, there will be at least minimal metadata describing them, making it possible to at least have access to a set of initial pieces of the puzzle that reproducibility poses. Metadata can also improve the terrain for comparison and benchmarking of ML approaches as it provide a common underlying tissue that can be used to, for instance, group together approaches working with similar machine requirements, data types, and underlying algorithms. Initiatives such as OpenML, DLHub and NFDI4DS are working towards this direction, offering a not only a combination of data and software repositories and registries but sandbox platforms where different ML approaches can be found, compared, tried, and ideally understood within the context provided by the platform.

Research objects management plans will also benefit from a FAIR approach to ML. As FAIR metadata is added to data, software, workflows, ML and other research objects, it becomes easier to package all these metadata together and connect them to machine-actionable plans. A machine-actionable Research Data Management (ma-RDM) plan provides researchers with a way to systematically manage data along its research lifecycle. While DMPs help describe techniques, methods, and policies in relation to data as well as activities and their relations across the lifecycle, ma-DMPs structure and standardize the way such descriptions are provided [3]. Similar to DMPs, there is also the concept of (Research) Software Management Plans (SMPs) [1, 20] that should also evolve towards ma-SMPs. ma-DMPs can be easily connected to Research Object packages, e.g., RO-Crates [4]. An RO-crate provides a way to package together one or more research objects and their corresponding metadata using schema.org [19] as the supporting structure data model. DMPs, SMPs, and RO-crates all benefit from metadata and will become more powerful (e.g., standardized, compatible, extensible) as metadata becomes FAIR itself. This holds not only for data but also for other research objects, including ML.

5 Conclusions and open questions

We have begun to understand the landscape that is relevant to FAIR for machine learning, including researchers, communities, and elements of infrastructure such as execution platforms and repositories, but there is still a lot of work to do in defining how FAIR should be applied to machine learning. While FAIR has been defined for data, and is increasingly being adapted to other research outputs (e.g., software, workflows), it is currently unclear if FAIR should only apply to machine learning models or if there is a way that FAIR could apply to other parts of machine learning processes. An argument can be made that a traditional machine learning process comprises research software and training data connected via workflows, and therefore the FAIR principles could be applied to each independently. However, machine learning goes beyond the individual components, especially when taking into considerations the respective platforms and processes necessary to successfully create a model. And ultimately, looking at the ML model as the direct digital output of the process itself, the definition of FAIR is far from a given.

For ML models specifically, questions around FAIR are not (yet) as clear as for software and computational workflows. Are they searched and shared via repositories? Or perhaps searched and shared via executable platforms? Or maybe searched and shared via something else? (e.g., DLHub, OpenML, ...) Additionally, given that models are trained on specific data, they are really linked to that data. Should this be reflected in how they are shared?

We believe that discussion and potential answers to these questions requires careful analysis of the FAIR principles, similar to the one that is taking place for software and workflows, combined with work already done on FAIR for data. We will continue working on this using a community-focused approach so different stakeholders can participate and shape the FAIR principles for ML. Once these principles have been defined, we can move towards the next relevant challenges, which include identifying relevant metrics and indicators that can be directly applicable for ML.

References

1. Alves, R., Bampalikis, D., Castro, L.J., Fernández, J.M., Harrow, J., Martin Del Pico, E., Psomopoulos, F., Via, A.: ELIXIR Software Management Plan, https://github.com/elixir-europe/smp
2. Avsec, Ž., Kreuzhuber, R., Israeli, J., Xu, N., Cheng, J., Shrikumar, A., Banerjee, A., Kim, D.S., Urban, L., Kundaje, A., Stegle, O., Gagneur, J.: Kipoi: accelerating the community exchange and reuse of predictive models for genomics. bioRxiv (2018). https://doi.org/10.1101/375345

3. Cardoso, J., Garcia Castro, L.J., Ekaputra, F., Jacquemot-Perbal, M.C., Miksa, T., Borbinha, J.: Towards semantic representation of machine-actionable Data Management Plans. In: DaMaLOS 2020. Online (2020). https://doi.org/10.4126/FRL01-006423289
4. Carragáin, E.O., Goble, C., Sefton, P., Soiland-Reyes, S.: A lightweight approach to research object data packaging. In: Bioinformatics Open Source Conference (BOSC). ISMB/ECCB 2019, Basel, Switzerland (Jun 2019). https://doi.org/10.5281/zenodo.3250687
5. Chard, R., Li, Z., Chard, K., Ward, L., Babuji, Y., Woodard, A., Tuecke, S., Blaiszik, B., Franklin, M.J., Foster, I.: DLHub: Model and data serving for science. In: 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS). pp. 283–292 (May 2019). https://doi.org/10.1109/IPDPS.2019.00038
6. Chue Hong, N.P., Katz, D.S., Barker, M., Lamprecht, A.L., Martinez, C., Psomopoulos, F.E., Harrow, J., Castro, L.J., Gruenpeter, M., Martinez, P.A., Honeyman, T.: FAIR principles for research software (FAIR4RS principles). Tech. rep., Research Data Alliance (2021). https://doi.org/10.15497/RDA00065
7. Confederation of Laboratories for Artificial Intelligence Research in Europe (CLAIRE), https://claire-ai.org
8. ELIXIR, https://elixir-europe.org
9. ELIXIR machine learning focus group, https://elixir-europe.org/focus-groups/machine-learning
10. Findable, accessible, interoperable, and reusable frameworks for physics-inspired artificial intelligence in high energy physics (FAIR4HEP), https://fair4hep.github.io
11. German national research infrastructure for data science (NFDI), https://www.nfdi4datascience.de/
12. Goble, C., Cohen-Boulakia, S., Soiland-Reyes, S., Garijo, D., Gil, Y., Crusoe, M.R., Peters, K., Schober, D.: FAIR computational workflows. Data Intelligence **2**(1-2), 108–121 (01 2020). https://doi.org/10.1162/dint_a_00033
13. International Conference on Machine Learning (ICML), https://icml.cc/
14. Katz, D.S., Barker, M.: Generalizing fair (Mar 2021), https://danielskatzblog.wordpress.com/2021/03/29/generalizing-fair/
15. Katz, D.S., Pollard, T., Psomopoulos, F., Huerta, E., Erdmann, C., Blaiszik, B.: FAIR principles for machine learning models (Nov 2020). https://doi.org/10.5281/zenodo.4271996
16. Katz, D.S., Psomopoulos, F.E.: Defining FAIR for machine learning (ML) (Apr 2021), https://www.rd-alliance.org/defining-fair-machine-learning-ml
17. NeurIPS conference, https://nips.cc/
18. The Pistoia Alliance, https://www.pistoiaalliance.org
19. Schema.org, https://schema.org
20. The Software Sustainability Institute: Checklist for a Software Management Plan. Tech. rep., Zenodo (Oct 2018). https://doi.org/10.5281/zenodo.1460504
21. Vanschoren, J., van Rijn, J.N., Bischl, B., Torgo, L.: OpenML: Networked science in machine learning. SIGKDD Explorations **15**(2), 49–60 (2013). https://doi.org/10.1145/2641190.2641198
22. Walsh, I., Fishman, D., Garcia-Gasulla, D., Titma, T., Pollastri, G., Capriotti, E., Casadio, R., Capella-Gutierrez, S., Cirillo, D., Del Conte, A., Dimopoulos, A.C., Del Angel, V.D., Dopazo, J., Fariselli, P., Fernández, J.M., Huber, F., Kreshuk, A., Lenaerts, T., Martelli, P.L., Navarro, A., Broin, P.O., Piñero, J., Piovesan, D., Reczko, M., Ronzano, F., Satagopam, V., Savojardo, C., Spiwok, V., Tangaro,

M.A., Tartari, G., Salgado, D., Valencia, A., Zambelli, F., Harrow, J., Psomopoulos, F.E., Tosatto, S.C.E., ELIXIR Machine Learning Focus Group: DOME: recommendations for supervised machine learning validation in biology. Nature Methods (Jul 2021). https://doi.org/10.1038/s41592-021-01205-4
23. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., et al.: The FAIR guiding principles for scientific data management and stewardship. Scientific Data **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18