

Working Towards Understanding the Role of FAIR for Machine Learning

Slides: <https://doi.org/10.5281/zenodo.5594990>
Paper: <https://doi.org/10.4126/FRL01-006429415>

Daniel S. Katz
Fotis E. Psomopoulos
Leyla Jael Castro



CERTH
CENTRE
FOR RESEARCH
& TECHNOLOGY
HELLAS



FAIR principles

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, [...] Barend Mons 

Scientific Data **3**, Article number: 160018 (2016) | [Cite this article](#)

194k Accesses | **2450** Citations | **1852** Altmetric | [Metrics](#)

A set of principles, to ensure that **data** are shared in a way that enables and enhances reuse by humans and machines

Findable

- F1.** (meta)data are assigned a globally unique and eternally persistent identifier.
- F2.** data are described with rich metadata.
- F3.** (meta)data are registered or indexed in a searchable resource.
- F4.** metadata specify the data identifier.

Accessible

- A1** (meta)data are retrievable by their identifier using a standardized communications protocol.
 - A1.1** the protocol is open, free, and universally implementable.
 - A1.2** the protocol allows for an authentication and authorization procedure, where necessary.
- A2** metadata are accessible, even when the data are no longer available.

Interoperable

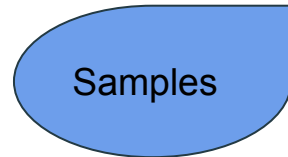
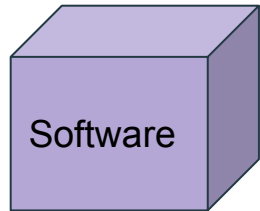
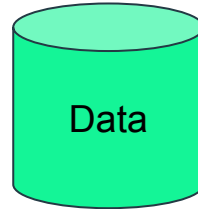
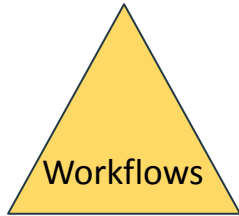
- I1.** (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2.** (meta)data use vocabularies that follow FAIR principles.
- I3.** (meta)data include qualified references to other (meta)data.

Reusable

- R1.** meta(data) have a plurality of accurate and relevant attributes.
 - R1.1.** (meta)data are released with a clear and accessible data usage license.
 - R1.2.** (meta)data are associated with their provenance.
 - R1.3.** (meta)data meet domain-relevant community standards.

<https://doi.org/10.5281/zenodo.5594990>

Not all research objects are the same



FAIR for other (non-data) research objects

- 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
- Text in principles often includes "(Meta)data ..."
 - Shorthand for "metadata and data ..."
- Principles applied via dataset creators and repositories, collectively responsible for creating, annotating, indexing, preserving, sharing the datasets and their metadata
- What about non-data objects?
 - While they can often be stored as data, they are not **just** data
- While high level goals (F, A, I, R) are mostly the same, the details and how they are implemented depend on
 - How objects are created and used
 - How/where the objects are stored and shared
 - How/where metadata is stored and indexed
- Work needed to define, then implement, then adopt principles

FAIR for Machine Learning

- FAIR Principles, are intended to apply to **all digital objects** ([Wilkinson et al. 2016](#))
- We focus on the **adaptation and adoption** of the FAIR principles to machine learning

Recommendation n°5 :

*Recognise that FAIR guidelines will require **translation for other digital objects** and support such efforts.*

2020: 'Six Recommendations for Implementation of FAIR Practice'

([FAIR Practice Task Force EOSC, 2020](#))

FAIR for ML: Data + Software

- As previously stated*, original FAIR principles
 - Claim to apply to "scholarly digital research objects"
 - But actually focus on metadata and data
- FAIR for Research Software work and FAIR Workflows focusing on how to translate/interpret the principles for research software & workflows
- What about machine learning (ML) models?
 - Are they data?
 - E.g., a set of parameters and options for a particular framework
 - Are they software?
 - E.g., an executable object that takes input and provides output
 - Are they something else?

* see [Wilkinson et al. 2016](#) and [Katz & Barker 2021](#)

FAIR for ML: Stakeholders

- Information and data scientist
- Researchers
 - Create, train, share and use ML models
- Platforms
 - Publishing, sharing, running, comparing ML approaches
 - [DLHub](#) - Find, share, publish, and run machine learning models and discover training data for science
 - [Kipoi](#) - API & repository of ready-to-use trained models for genomics
 - [OpenML](#) - 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
 - [Pistoia Alliance](#) - a global, not-for-profit members' organization working to lower barriers to innovation in life science and healthcare R&D through pre-competitive collaboration
 - [ELIXIR](#) - An intergovernmental organisation that brings together life science resources (including databases, software tools, training materials, cloud storage and supercomputers) from across Europe
 - [CLAIRE](#) - Confederation of Laboratories for Artificial Intelligence Research in Europe
- Projects
 - [FAIR4HEP](#) - Using high-energy physics (HEP) as the science driver, developing a FAIR framework to advance understanding of AI, applying AI techniques, and exploring approaches to AI

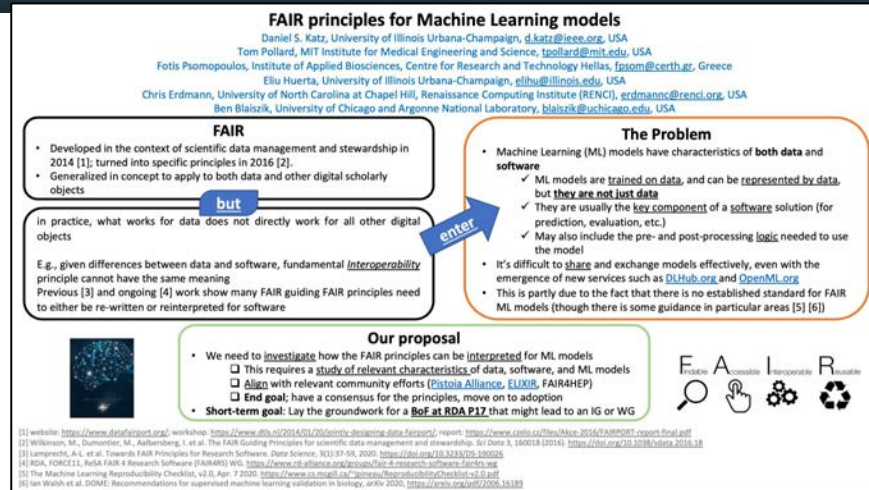
Current FAIR (data) practices

- Dataset creators, repositories, registries
 - creating, annotating, indexing, preserving, and sharing datasets and their metadata.
- Large elements of FAIR for data are dependent on archival repositories (e.g. Zenodo, re3data.org)
 - Hold data and/or metadata, provide search and access capabilities
- Software is different, since it typically isn't shared via archival repositories but instead via social coding platform (e.g., GitHub) and package management systems (e.g. PyPI, CRAN)
- What about ML models?
 - Searched and shared via repositories?
 - Searched and shared via executable platforms?
 - Searched and shared via something else? (e.g., DLHub, OpenML, ...)
 - Models and training data are linked - should they be shared together?

Ongoing effort

- Poster at RDA VP16 (Nov 2020):
<https://doi.org/10.5281/zenodo.4271995>
- BoF at RDA VP17 (April 2021):
<https://www.rd-alliance.org/defining-fair-machine-learning-ml>
- FAIR for Machine Learning Models (June 2021), FAIR Festival
- 1st Community call July 21 2021
- BoF at RDA VP18: "[Steps towards defining FAIR principles for Machine Learning \(ML\)](#)"
 - Discuss status and form new working group
 - Find co-chairs & general contributors

<https://doi.org/10.5281/zer>



Defining FAIR for Machine Learning (ML)

Home

25
JAN
2021

Defining FAIR for Machine Learning (ML)

Submitted by Daniel S. Katz

Meeting objectives:

Discuss:

- Current projects (both research and infrastructure) in machine learning (ML) that are considering FAIR,
- If there's value in and a need for defining FAIR for ML, and if so,
- How to move forward to do so, ideally under the RDA umbrella based on the current role of RDA in FAIR activities

Beyond FAIR: Reproducibility & *ilities

- Other goals (not part of FAIR): reproducibility, comparability, explainability
- But 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
 - Minimal metadata describing them exists
 - Addresses part of the reproducibility puzzle
- Metadata can also support comparison and benchmarking of ML approaches
 - Provides a common underlying tissue
 - 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.

Beyond FAIR: Management plans

- FAIR approach to ML also helps research objects management plans
- FAIR metadata added to data, software, workflows, ML and other research objects makes it easier to package all these metadata together and connect them to machine-actionable plans
- Machine-actionable Research Data Management (ma-RDM) plan provides researchers with a way to systematically manage data along its research lifecycle
- 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
- Similar, (Research) Software Management Plans (SMPs) should also evolve towards ma-SMPs
- ma-DMPs can be easily connected to Research Object packages, e.g., RO-Crates
- 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

Status

- Now starting to understand landscape relevant to FAIR for machine learning
 - Including researchers, communities, and infrastructure such as execution platforms and repositories
- FAIR for data is mature; FAIR for other research objects (software, workflows, ...) being developed
- Traditional ML process is research software & training data connected via workflows
 - So FAIR principles could apply to each independently
- But machine learning goes beyond individual components
 - Especially when considering platforms and processes necessary to successfully create a model

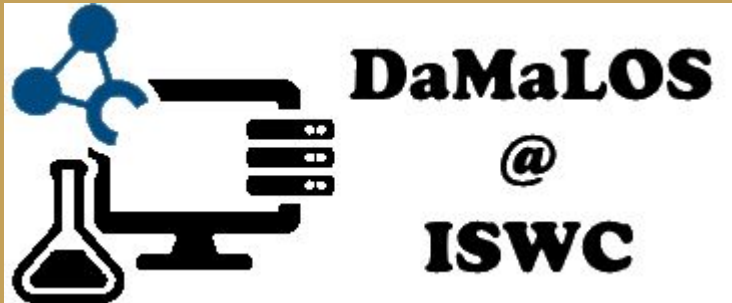
Open questions

- Does this effort merit further discussion?
 - If so, under which context (e.g. RDA WG?)
- How do you think FAIR should be applied to ML?
 - What changes/definitions are most needed?
 - What are the questions that need discussion?
 - Should FAIR address only ML models, and/or also processes, and/or also platforms, etc.?
 - Are ML models 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, ...)
 - Should the fact that models are trained on specific data be reflected in how the models and data are shared?

Conclusions

- Still a lot of work to do in defining how FAIR should be applied to machine learning
- Discussion & potential answers to open questions requires careful analysis of FAIR principles
 - Similar to current work for software & workflows, and past work for data
- Want to continue working on this using a community-focused approach
 - With different kinds of stakeholders participating & shaping FAIR ML principles
- After principles are defined, next challenges include
 - Documentation
 - Identifying relevant metrics and indicators for ML
 - Adoption examples

Thanks for your attention!



Slides: <https://doi.org/10.5281/zenodo.5594990>
Paper: <https://doi.org/10.4126/FRL01-006429415>