FAIR principles for data and beyond

Semantic Technologies Team – Knowledge Mng Group
Dr. Leyla Jael Castro

ljgarcia@zbmed.de

C3RDM
COLOGNE COMPETENCE CENTER FOR RESEARCH DATA MANAGEMENT

ZB MED
Information Centre for Life Sciences
www.zbmed.de

The FAIR Guiding Principles

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

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

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.

Let's go FAIR

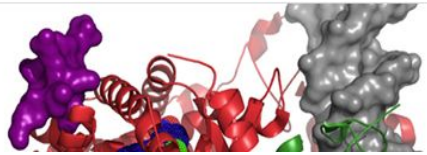
Why should I (researcher) go FAIR?

- I read about FAIR and it's cool, I want my data to be FAIR
- I got reviews on my submission and reviewers say my data is not FAIR
- My PI is telling me I have to make my data FAIR
- The Data Steward tells me that according to the RDM plan, my data is not ready because is not FAIR

Why should we (researchers) go FAIR?

- FAIR helps us reflect on our data (in combination with RDM plans)
- FAIR provides some guidelines for data sharing (so we can reuse other's data less work, more collaboration)
- FAIR provides some guidelines for data to interoperate (so we can connect our data to other data good approach to more complex problems)
- FAIR is a step forward to reproducibility (but on its own does not solve the crisis)

FAIR metadata for humans and machines alike



IUPHAR/BPS
Guide to PHARMACOLOGY

Search Database

Home About Help

Search GtoPdb...

Ligands relevant to SARS-CoV-2(COVID-19) - ligand name links to detailed information in GtoPdb, or to our pre-release blog.

Download as CSV

Ligand (Therapeutic)	Ligand ID	Comments
Abl kinase inhibitors (e.g. imatinib)	5687	Abelson kinase (Abl) inhibitors are reported to block Spike protein-induced SARS-CoV and MERS-CoV fusion in vitro PMID: 29557770, potentially by blocking Abl2 at the endosomal membrane PMID: 27466418. It will be informative to determine if this holds true for SARS-CoV-2, and whether re-purposing of imatinib and/or newer Abl kinase inhibitors (dasatinib; bosutinib) could be a viable strategy against COVID-19. Likely to be most effective during the early stage of infection.
acalabruntinib	8912	acalabrutinib is an approved BTK inhibitor that is used to treat B cell malignancies. It is being progressed to clinical trial in COVID-19 patients as part of the UK's ACCORD initiative https://bit.ly/2ZXjXOw.
ACE2 ligands	n/a	ACE2 ligands that block the site of viral spike protein interaction could offer anti-SARS-CoV-2 infection potential. In hypertensive COVID-19 patients maintained on ACE2 inhibitor therapy whilst hospitalised, ACE2 inhibitors are reported to reduce all-cause mortality compared to patients with hypertension not on ACE inhibitors PMID: 32302265. But we need to be cognisant that the protective effects may not be evident in COVID-19 patients with normal blood pressure, and that a proper controlled clinical trial is needed before drawing firm conclusions.
anakinra	6972	An approved IL-1 pathway blocking peptide, that has anti-inflammatory action. This Comment in The Lancet Rheumatology discusses factors to be considered when planning anakinra trials for COVID-19 hyperinflammatory response https://bit.ly/2yu89lg.
anti-TNF therapy (e.g. infliximab, adalimumab and others)	5004, 4860	The case for re-purposing approved anti-TNF biologic therapies as a means to combat cytokine storm and inflammation in COVID-19 patients is presented in PMID: 32278362.
apilimod	9895	A clinical stage inhibitor of the type III phosphoinostol kinase, PIKfyve. Anti-SARS-CoV-2 activity in vitro https://www.biorxiv.org/content/10.1101/2020.04.16.044016v1.
ASC09F	n/a	A combination drug containing ASC09 (a viral protease inhibitor) + ritonavir is an example of an existing HIV therapy being repurposed for COVID-19.

```
<!--- BioSchemas ---->
<script type="application/ld+json">[

{
  "@context": "https://schema.org",
  "@type": "Protein",
  "@id": "https://www.guidetopharmacology.org/GRAC/ObjectDisplayForward?objectId=1614#",
  "http://purl.org/dc/terms/conformsTo": "https://bioschemas.org/profiles/Protein/0.9-DRAFT-2019_08_20/",
  "identifier": "1614",
  "name": "ACE2",
  "associatedDisease": "COVID-19",
  "description": "Receptor on host cells that is exploited by some betacoronaviruses for viral entry. Engaged by
  SARS-CoV-2 spike protein as the first step towards infection of host cells",
  "url": "https://www.guidetopharmacology.org/GRAC/ObjectDisplayForward?objectId=1614"
},
{
  "@context": "https://schema.org",
  "@type": "MolecularEntity",
  "@id": "https://www.guidetopharmacology.org/GRAC/LigandDisplayForward?ligandId=8912#",
  "http://purl.org/dc/terms/conformsTo": "https://bioschemas.org/profiles/MolecularEntity/0.4-DRAFT-2019_11_11/",
  "identifier": "8912",
  "name": "acalabruntinib",
  "url": "https://www.guidetopharmacology.org/GRAC/LigandDisplayForward?ligandId=8912",
  "associatedDisease": "COVID-19",
  "description": "acalabrutinib is an approved BTK inhibitor that is
  progressed to clinical trial in COVID-19 patients as part
```

Images taken from a slide by Alasdair Gray from a presentation given at LEIXIR Data and Tools Workshop, November 2021

Ok, let's go FAIR but how?

Start with explicit and concrete steps coming directly from the FAIR Guiding Principles

R1.1. (meta)data are released with a clear and accessible data usage license.

R1.2. (meta)data are associated with their provenance.

F3. (meta)data are registered or indexed in a searchable resource.

F1. (meta)data are assigned a globally unique and eternally persistent identifier.

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

DRYAD

zenodo

You are not alone, link to others

Focus on the linked aspect

I3. (meta)data include qualified references to other (meta)data.

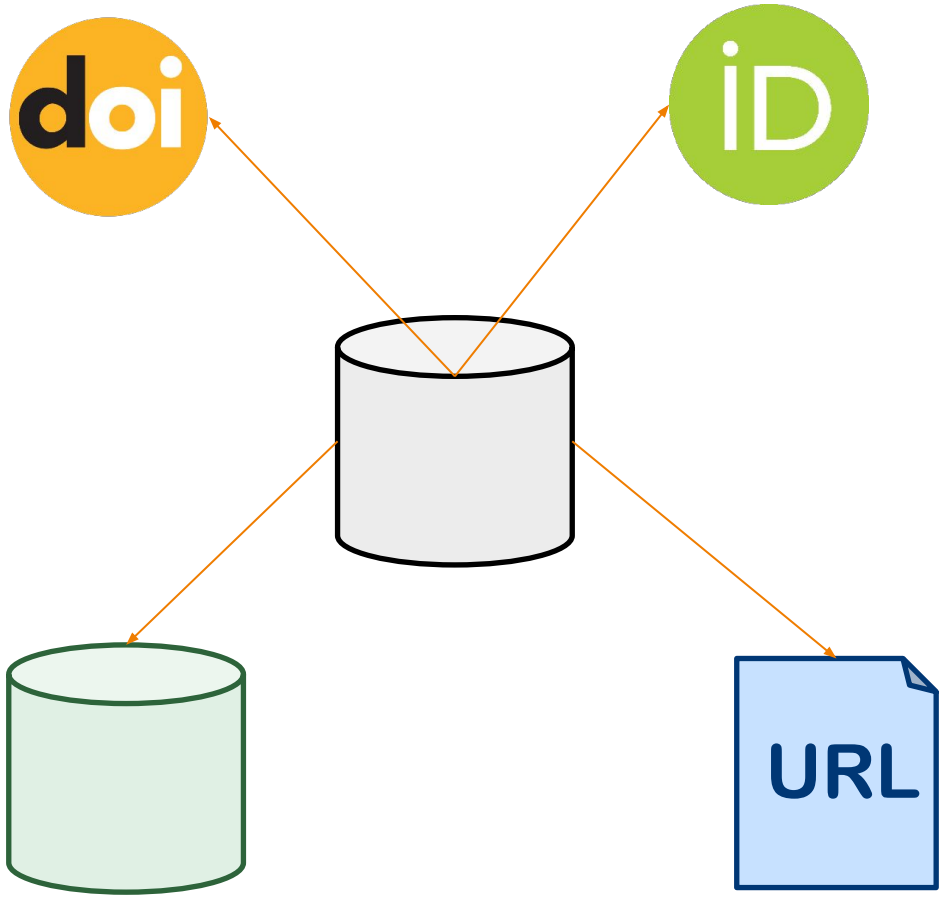


Image taken from http://sites.linkeddata.center/help/devop/training/introduction-to-linked-data

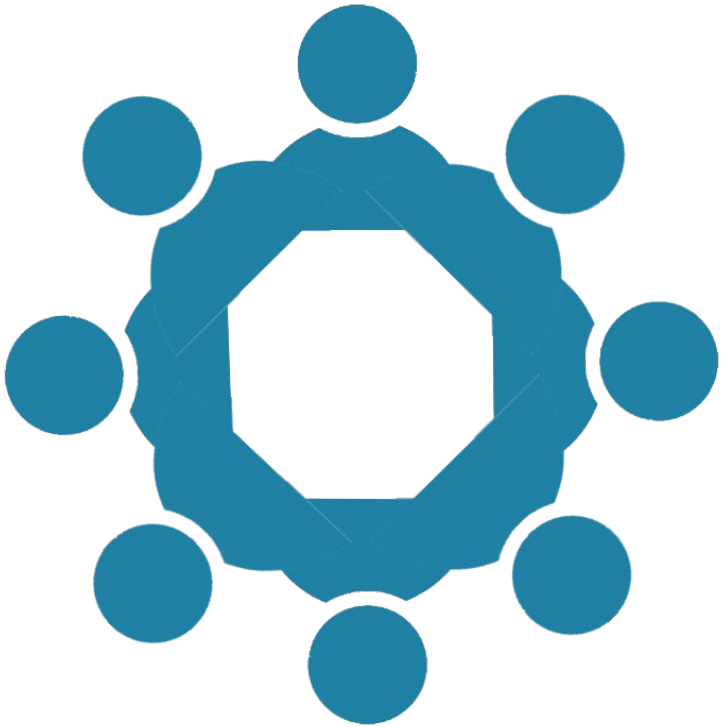
Rich, plurality, community

Start with (good enough) metadata

F2. data are described with rich **metadata**.

R1. **meta**(data) have a plurality of accurate and relevant attributes.

R1.3. (**meta**)data meet domain-relevant community standards.



What about the rest of the principles?

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.

Step by step rather than all at once

Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud

Cite

Article type: Research Article

Authors: Mons, Barend[a; b; c; *] | Neylon, Cameron[d] | Velterop, Jan[e] | Dumontier, Michel[f] | da Silva Santos, Luiz Olavo Bonino[b; g] | Wilkinson, Mark D.[h]

Keywords: FAIR Data, Open Science, interoperability, data integration, standards

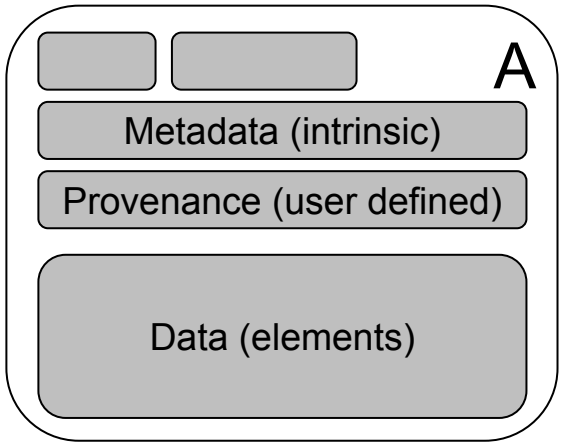
DOI: 10.3233/ISU-170824

Journal: Information Services & Use, vol. 37, no. 1, pp. 49-56, 2017

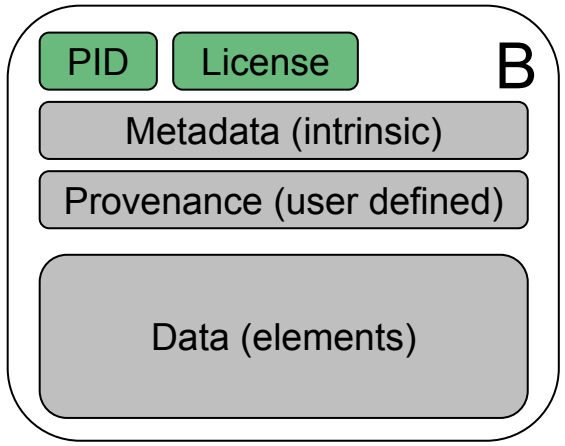
Published: 7 March 2017

Note: diagram has been modified, in particular, it adds a explicit mention to the license

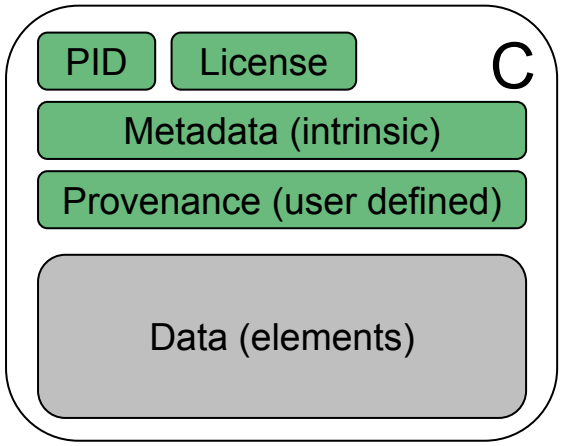
Re-useless data (common case)



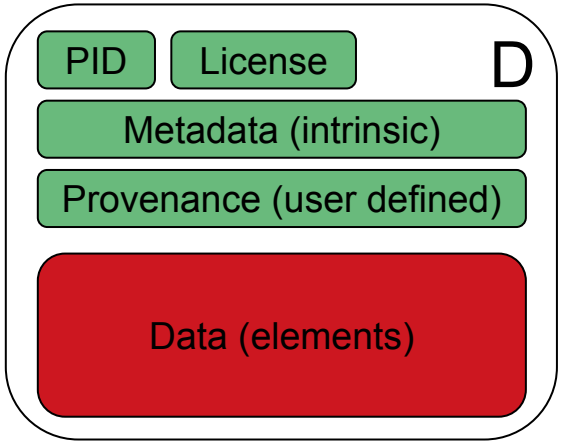
Findable and reusable



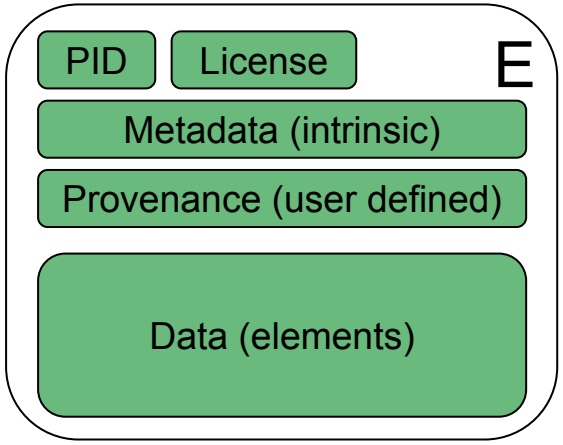
FAIR metadata



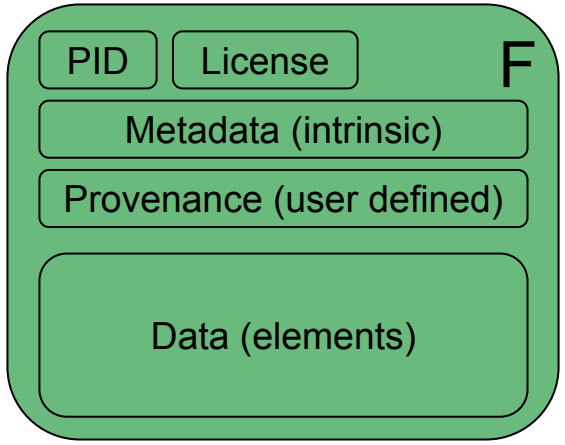
FAIR data - restricted access



FAIR data - **open access**



FAIR data - **open access** and functionally linked



I think my data is FAIR, how can I know for sure?

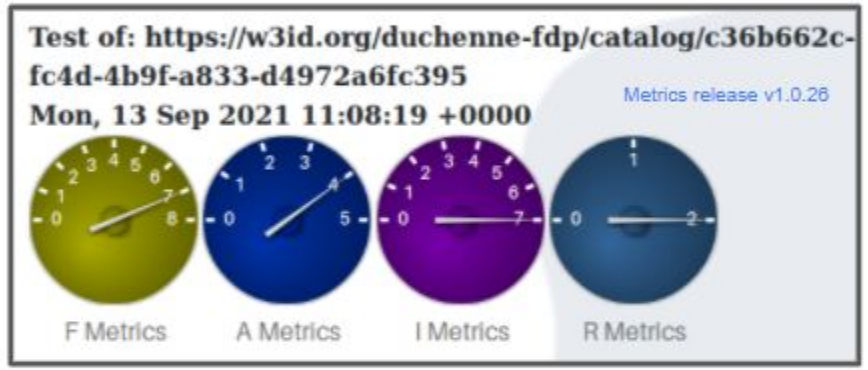
FAIRassist.org

Resource	Execution Type
5 Star Data Rating Tool	Manual - questionnaire
AutoFAIR	Semi-automated
Data Stewardship Wizard	Predictive; based on a manually filled questionnaire
F-UJI	Automated
FAIR Data Self-Assessment Tool	Manual - questionnaire
FAIR Evaluator	Automated
FAIR enough?	Manual - checklist
FAIR-Aware (BETA)	Manual - questionnaire
FAIR-Checker	Automated
FAIRdat	Manual - questionnaire
FAIRness self-assessment grids	Manual - checklist
FAIRshake	Manual - questionnaire, Semi-manual
GARDIAN FAIR Metrics	Manual - checklist
RDA Maturity Model	Manual - checklist

How different can they be?

Comparison of The Evaluator with F-UJI, on the same URI

(a Catalog record in the Duchenne Muscular Dystrophy FAIR Data Point)



20/22 Tests Pass

2/24 Tests Pass

Image corresponds to a slide by Mark D. Wilkinson from a presentation given at FAIReScience Workshop @ IEEE eScience Conference 2021

My data is FAIR, is that FAIR enough?

January 01 2020

FAIR Computational Workflows

Carole Goble, Sarah Cohen-Boulakia, Stian Soiland-Reyes, Daniel Garijo, Yolanda Gil, Michael R. Crusoe, Kristian Peters, Daniel Schober

> Author and Article Information

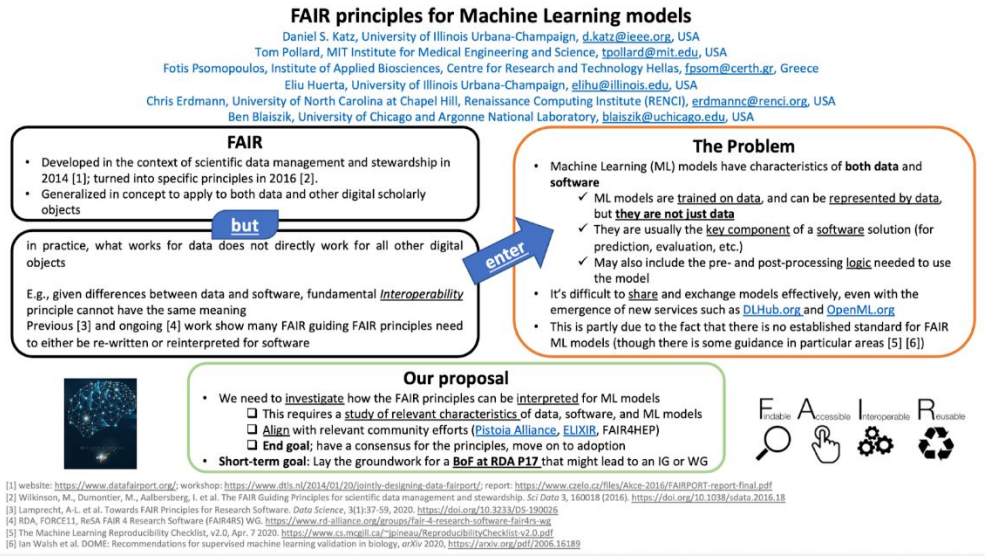
Data Intelligence (2020) 2 (1-2): 108–121.

https://doi.org/10.1162/dint_a_00033

Ten simple rules for making training materials FAIR

Leyla Garcia, Bérénice Batut, Melissa L. Burke, Mateusz Kuzak, Fotis Psomopoulos, Ricardo Arcila, Teresa K. Attwood, Niall Beard, Denise Carvalho-Silva, Alexandros C. Dimopoulos, Victoria Dominguez del Angel, Michel Dumontier, Kim T. Gurwitz, [...], Patricia M. Palagi [view all]

Published: May 21, 2020 • https://doi.org/10.1371/journal.pcbi.1007854



FAIR principles for Machine Learning models

Daniel S. Katz, University of Illinois Urbana-Champaign, d.katz@ieee.org, USA
Tom Pollard, MIT Institute for Medical Engineering and Science, tpollard@mit.edu, USA
Fotis Psomopoulos, Institute of Applied Biosciences, Centre for Research and Technology Hellas, fpsom@certh.gr, Greece
Eliu Huerta, University of Illinois Urbana-Champaign, elihu@illinois.edu, USA
Chris Erdmann, University of North Carolina at Chapel Hill, Renaissance Computing Institute (RENCI), erdmannc@renci.org, USA
Ben Blaiszik, University of Chicago and Argonne National Laboratory, blaiszik@uchicago.edu, USA

FAIR

- Developed in the context of scientific data management and stewardship in 2014 [1]; turned into specific principles in 2016 [2].
- Generalized in concept to apply to both data and other digital scholarly objects

but

in practice, what works for data does not directly work for all other digital objects

E.g., given differences between data and software, fundamental *Interoperability* principle cannot have the same meaning
Previous [3] and ongoing [4] work show many FAIR guiding FAIR principles need to either be re-written or reinterpreted for software

enter

The Problem

- Machine Learning (ML) models have characteristics of **both data** and **software**
  - ✓ ML models are trained on data, and can be represented by data, but **they are not just data**
  - ✓ They are usually the key component of a software solution (for prediction, evaluation, etc.)
  - ✓ May also include the pre- and post-processing logic needed to use the model
- It's difficult to share and exchange models effectively, even with the emergence of new services such as DLHub.org and OpenML.org
- This is partly due to the fact that there is no established standard for FAIR ML models (though there is some guidance in particular areas [5] [6])

Our proposal

- We need to investigate how the FAIR principles can be interpreted for ML models
  - ❑ This requires a study of relevant characteristics of data, software, and ML models
  - ❑ Align with relevant community efforts (Pistoia Alliance, ELIXIR, FAIR4HEP)
  - ❑ **End goal**; have a consensus for the principles, move on to adoption
- **Short-term goal**: Lay the groundwork for a **BoF at RDA P17** that might lead to an IG or WG

Findable Accessible Interoperable Reusable

[1] website: https://www.datafairport.org/; workshop: https://www.dtls.nl/2014/01/20/jointly-designing-data-fairport/; report: https://www.copdess.org/files/Akce-2016/FAIRPORT-report-final.pdf
[2] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
[3] Lamprecht, A-L. et al. Towards FAIR Principles for Research Software. *Data Science*, 3(1):37-59, 2020. https://doi.org/10.3233/DS-190026
[4] RDA, FORCE11, ReSA FAIR 4 Research Software (FAIR4RS) WG. https://www.rd-alliance.org/groups/fair-4-research-software-fair4rs-wg
[5] The Machine Learning Reproducibility Checklist, v2.0, Apr. 7 2020. https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist-v2.0.pdf
[6] Ian Walsh et al. DOME: Recommendations for supervised machine learning validation in biology, *arXiv* 2020, https://arxiv.org/pdf/2006.16189

FAIR Principles for Research Software (FAIR4RS Principles)

DOI: 10.15497/RDA00065

Citation and download: 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., et al. (2021). FAIR Principles for Research Software (FAIR4RS Principles). *Research Data Alliance*. DOI: 10.15497/RDA00065

Breakout 7 | Data Infrastructures - Organisa… | The FAIR Agenda | WGs Getting started

WG FAIR for Virtual Research Environments: FAIR for VREs - The Path Forward

7:30 AM - 9:00 AM

Room E

FAIR for research digital objects

Recommendation n°5 :

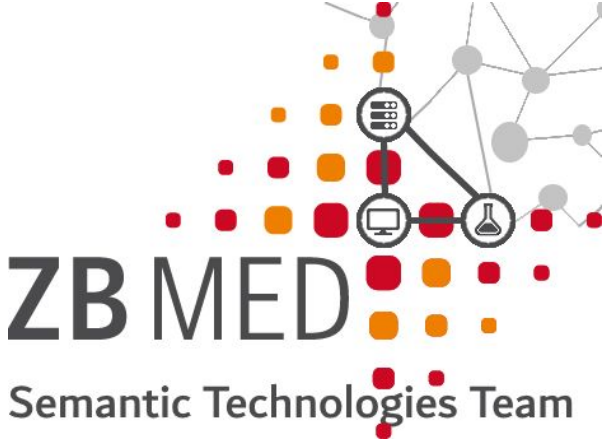
Recognise that FAIR guidelines will require ***translation for other digital objects*** *and support such efforts.*

2020: 'Six Recommendations for Implementation of FAIR Practice'

(FAIR Practice Task Force EOSC, 2020)

Thanks! Danke!

www.zbmed.de



Semantic Retrieval Team – Knowledge Management Group
Dr. Leyla Jael Castro

ljgarcia@zbmed.de