23rd Plenary Meeting Collaborative Notes

Group(s) name(s) organising the session: *FAIR for Machine Learning (FAIR4ML) IG*
Session link: *FAIR4ML IG Activity Updates and Feedback*
Session scheduled date/time/breakout session: *Tuesday, 12 November, Breakout 1, 10:00 – 11:30 CST*
Venue: *Edificio de Ingeniería, Auditorium 2*

Session summary (for Group co-chairs)

We will use the content in the table below to highlight your work to the RDA community as a report organised by the Technical Advisory Board and to a wider audience through English & Spanish social media mentions.
Please complete ALL fields below by **Friday, 29th November, close of business** to be included in the report & social media activities.

Summarise the session in three sentences:
Updates on progress in TF1 and TF2; invitation to members to contribute; invitation for new members to join the IG
Key outcomes/actions/takeaways
1. *Progress continues slowly towards a first draft of a white paper on FAIR and ML* 2. *Progress is occurring more quickly on defining metadata for ML, attendees interested in joining* 3. *Collaboration happening with Skills4EOSC*
Synergies and/or possible collaborations identified with RDA groups and other groups:
Collaboration between FAIR4ML TF1 (White Paper) and Skills4EOSC (Top 10 FAIR Data Things in AI)
Highlight text that will be used in social media mentions (please make sure the text is clear and appropriate for public consumption & comprehension)
The interest group updated the community on its work, and invited attendees to join. Attendees brought up issues related to future adoption of group work.
Direct link to Group home page
FAIR for Machine Learning (FAIR4ML) IG

*Get involved in **RDA Community***
*Check out **P23 programme sessions***

This meeting will take place according to the **RDA Code of Conduct**

Attendee Check-in

Please complete this table to indicate your attendance (add rows as needed):

Full Name	Affiliation	Location	Email
David Elbert	Johns Hopkins University	USA	elbert@jhu.edu
Graham Parton	Centre for Environmental Data Analysis (CEDA)	UK	graham.parton@stfc.ac.uk
Gerhard Goldbeck	Goldbeck Consulting	UK	gerhard@goldbeck-consulting.com
Raphael Cobe	Sao Paulo State University	Brasil	raphael.cobe@unesp.br
Hugh Shanahan	Royal Holloway, University of London	UK	hugh.shanahan@rhul.ac.uk
Christian Pagé	CERFACS	France	christian.page@cerfacs.fr
Florina Piroi	TU Wien	AT	florina.piroi@tuwien.ac.at
Gwenaelle Moncoiffe	NOC-BODC	UK	gmon@noc.ac.uk
M-Christine Jacquemot	CNRS	FR	marie-christine.jacquemot@inist.fr
Daniel S. Katz	U Illinois Urbana-Champaign	USA	d.katz@ieee.org
Tomasz Miksa	TU Wien	AT	tomasz.miksa@tuwien.ac.at
Sveinung Gundersen	University of Oslo / ELIXIR Norway	Norway	sveinugu@uio.no
Line Pouchard	Sandia National Labs	US	lcpouch@sandia.gov
Lisa Mayer	NIH	US	Lisa.mayer@nih.gov
Yueyang Liu	Intersect	AU	yueyang@intersect.

			org.au
Glen Charlton	Intersect	AU	glen@intersect.org.au
Claudia b medeiros	unicamp	BRBR	cmbm@unicamp.br
Anne-Sophie Bage	INRAE	Fr	Anne-sophie.bage@inrae.fr

Collaborative Session Notes *(To be used by participants and chairs during the session)*

Relevant projects & stakeholders: please include EMMC ASBL: https://emmc.eu/ Community of stakeholders in Europe with interest in materials modelling (including data-based/ML) and interoperability/metadata/ontologies.

Useful Links

- Collaborative notes (this document): https://tinyurl.com/RDA-P23-FAIR4ML-notes
- RDA P23 Slides: https://tinyurl.com/RDA-P23-FAIR4ML-slides
- RDA FAIR4ML IG: https://www.rd-alliance.org/groups/fair-for-machine-learning-fair4ml-ig
- GitHub Space: https://github.com/RDA-FAIR4ML
- Zotero Group: https://www.zotero.org/groups/5157406/fair4ml
- ML Lifecycle and FAIRness: ML Lifecycle and FAIRness (*work in progress*)
- Preprint: https://zenodo.org/records/10407265
- Vocabulary: https://w3id.org/fair4ml

Agenda

Welcome and introduction

Intro to FAIR4ML IG, the objectives, structure, how to join

Part 1 - Updates

Task Force 1 - Overview

Task Force 2 - Overview

David E.: apologies for not being active, but I have not seen an articulation of what queries are of interest, which I think of as the first step in focusing a schema. ***Is there a record of queries compiled to guide the schema work?*** I can quickly think of lots of queries but don't want to be presumptuous that this hasn't been done! I guess I can list mine just in case some aren't on the table. I think of:

1. A description of the specific field it might work in like "Find a machine learning model in molecular biology or materials property prediction"
2. Compatibility with the input data I have or can get "ML model for image classification using microscopy data or a regression model for numerical data in ecology"
3. Filtering by the output I need like "classification models for cell types or anomaly detection for manufacturing sensor data"
4. ML technique or algorithm type (or specific algorithm) like "neural network for predicting yield strength of alloys or reinforcement learning models for control in automated laboratories"
5. Finding pre-trained or transfer learning models: "What are the pre-trained models for medical image analysis"
6. Data compatibility or source matching like "ML models trained on crystal structure data" or "ML models for datasets similar to mine"
7. Might want to query on performance metrics like "high-accuracy ML models for..." or "interpretable ML models for climate impact studies"
8. Query about accessibility "open-source or publicly available"
9. Community use recommendation or benchmarks "ML models with standard benchmarks in xyz" or something
10. Should link to domain schemas like being able to query for "ML for alloy strength prediction" which might be somewhat domain specific. (Could use boolean queries here or have the info available as RDF and train an LLM for the query (or make it a content store to augment an LLM)

David E: I'm interested in the use of model cards. My understanding is that model cards are primarily human readable and give an overview but are not specific like a schema description. This brings me to a question: ***Is the goal in Task Force 2 to provide both a schema and model cards in the traditional sense (perhaps created automatically from the schema) or is it to present a true schema in some new type of model card that really has all the specific data***

fields and types, but presented in a somewhat more human readable way than just a JSON or XML?

Part 2 - Invited talks

Flash Talk #1 - Line Pouchard - Challenges for FAIR4 Machine Learning (slides to be added)

Flash Talk #2 - audience volunteers?

Flash Talk #3

Flash Talk #4

Wrap-up and next actions

https://fair-universe.lbl.gov/

(Not sure where this goes) This effort is putting together an ecosystem, inc. data, experiments, infrastructure, analysis, all permeating with ML about cosmology