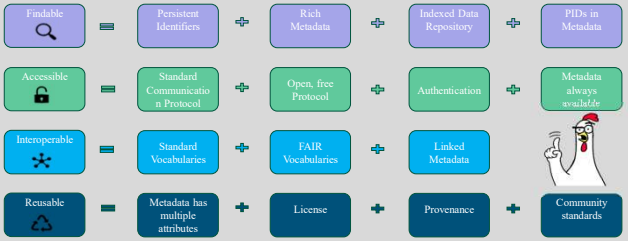
FAIRSCAPE: A FRAMEWORK FOR AI READINESS

Sadnan Al Manir[1], Maxwell Adam Levinson[1] and Timothy W. Clark[1,2,3]
[1]Public Health Sciences [2]Department of Neurology [3]School of Data Science
University of Virginia

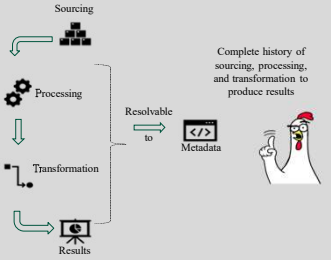
What is AI-Ready Data?

AI-Ready data are fully characterized FAIR data of known provenance, which can be ethically and reliably processed by AI applications; whose models and software are available and well-described for validation and re-use; and whose predictions may be fully explained and interpreted to the user as needed. Regardless of the specific definitions adopted, in all cases significant requirements are placed on the datasets used to train AI/ML models and on datasets analyzed using these AI/ML models. Datasets that meet these requirements are called "AI-Ready". We define *FAIR*, *Provenanced*, *Characterized*, *Explainable*, and *Ethical* as the criteria.

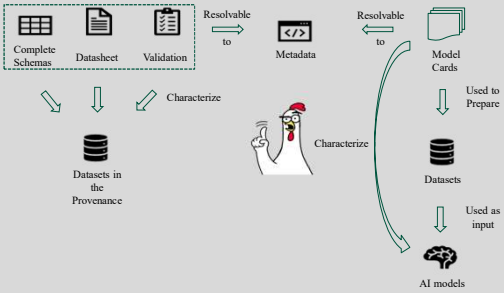
FAIR



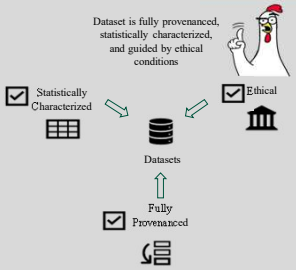
Provenanced



Characterized



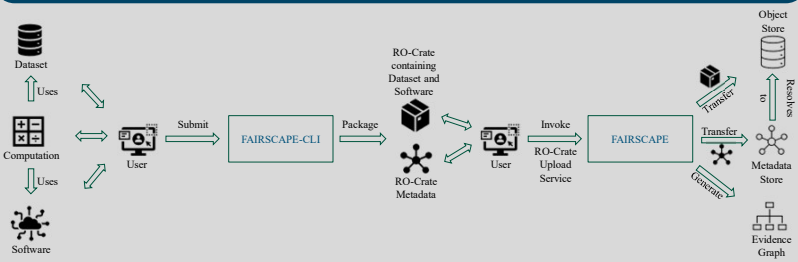
Explainable



Ethical

Ethical conditions include but not limited to
Ethical treatment of human and animal subject data.
Proper non-biasing data analysis and scientific integrity of methods, conclusions, and publications.
Full documentation of scientific methods and reagents.
Licensing and distribution of data, software, and models for reuse by the scientific community.
Anticipate potential applications of the data and derived AI/ML systems to be integrated into data governance frameworks.
Optimize benefits, mitigate potential risks, and promoting the use of the system for the collective good.

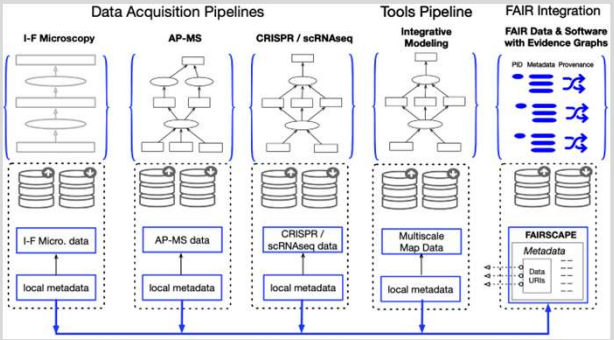
FAIRSCAPE-CLI + FAIRSCAPE



Summary

FAIRSCAPE consists of a client-side Python3 application, called either from the command line or as a set of Python functions by the Tools Module's data integration pipeline, and a server application, also in Python3, which completes the packaging.
The client side computations may constitute a pipeline of arbitrary complexity, or they may be a single step. The client-side package, FAIRSCAPE-CLI, is called when any computation or coherent set of computations in the pipeline is completed, and it is passed metadata which defines schemas in JSON-Schema for the datasets in the computational unit, as well as the inputs, computations, software, models, and outputs. FAIRSCAPE-CLI creates an RO-Crate package with the datasets, metadata, and software – or resolvable references to these components – and unique stubs for identifier creation on each of these components.
The RO-Crates are then sent to the FAIRSCAPE server where they are registered and assigned persistent, resolvable, globally unique IDs (PIDs). The RO-Crates are then decomposed into their individual components – datasets, models, software – which are also registered and assigned PIDs. The PID system currently in use is the ARK scheme – with DOIs a future feature as supplementary PIDs for final-state publishable work.
Lastly, the server computes end-to-end entailments on each RO-Crate's provenance as expressed in the EVI Evidence Graph Ontology and links them together where possible. PIDs generated by the server will resolve to machine-readable and/or human-readable landing pages containing the metadata, expressed in the JSON-LD graph language using vocabularies from Schema.org, EVI, and other well-defined public ontologies.
Both packages are PIP-installable and licensed under the MIT open-source license.

FAIRSCAPE in Cell Maps for AI



References

[1] Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018 (2016).
[2] Katz, D. S. et al. Recognizing the value of software: a software citation guide. F1000Res. 9, 1257 (2020).
[3] Gebru, T. et al. Datasheets for Datasets. arXiv (2018) doi:10.48550/arxiv.1803.09010.
[4] Mitchell, M. et al. Model cards for model reporting. in Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19 220–229 (ACM Press, 2019). doi:10.1145/3287560.3287596.
[5] Sendak, M. P., Gao, M., Brajer, N. & Balu, S. Presenting machine learning model information to clinical end users with model facts labels. npj Digital Med. 3, 41 (2020).