EGU21-12384, updated on 16 Sep 2021
https://doi.org/10.5194/egusphere-egu21-12384
EGU General Assembly 2021

AI-Ready Training Datasets for Earth Observation: Enabling FAIR data principles for EO training data.

Alastair McKinstry[1,2], Oisin Boydell[3,4], Quan Le[3,4], Inder Preet[3,4], Jennifer Hanafin[1,2], Manuel Fernandez[1,2], Adam Warde[1,2], Venkatesh Kannan[1,2], and Patrick Griffiths[5]

[1]NUI Galway, Irish Centre for High-End Computing, Galway, Ireland
[2]Irish Centre for High End Computing, NUI Galway, Galway, Ireland
[3]University College Dublin, Belfield, Dublin 4, Ireland
[4]CeADAR, UCD, Ireland
[5]ESA ESRIN, Via Galileo Galilei, 1, 00044 Frascati RM, Italy

The ESA-funded AIREO project [1] sets out to produce AI-ready training dataset specifications and best practices to support the training and development of machine learning models on Earth Observation (EO) data. While the quality and quantity of EO data has increased drastically over the past decades, availability of training data for machine learning applications is considered a major bottleneck. The goal is to move towards implementing FAIR data principles for training data in EO, enhancing especially the finability, interoperability and reusability aspects. To achieve this goal, AIREO sets out to provide a training data specification and to develop best practices for the use of training datasets in EO. An additional goal is to make training data sets self-explanatory ("AI-ready) in order to expose challenging problems to a wider audience that does not have expert geospatial knowledge.

Key elements that are addressed in the AIREO specification are granular and interoperable metadata (based on STAC), innovative Quality Assurance metrics, data provenance and processing history as well as integrated feature engineering recipes that optimize platform independence. Several initial pilot datasets are being developed following the AIREO data specifications. These pilot applications include for example forest biomass, sea ice detection and the estimation of atmospheric parameters.An API for the easy exploitation of these datasets will be provided.to allow the Training Datasets (TDS) to work against EO catalogs (based on OGC STAC catalogs and best practises from ML community) to allow updating and updated model training over time.

This presentation will present the first version of the AIREO training dataset specification and will showcase some elements of the best-practices that were developed. The AIREO compliant pilot datasets will be presented which are openly accessible and community feedback is explicitly encouraged.

[1] https://aireo.net/