AI-Ready EO Training Datasets : Final report

Subject:	ESA AO/3-16408/20/I/NB
Title:	AI-Ready EO Training Datasets
Document:	Final report
Prepared By:	Alastair McKinstry
Issue/Revision:	0.1
Date of Issue:	deliverableDate
Status:	Draft

Approval

Title	Final report and summary of the AI-Ready EO training datasets project
Issue Number 1	**Revision Number** 1
Author Alastair McKinstry	**Date** September 21, 2021
Approved By Alastair McKinstry	**Date of Approval** September 21, 2021

Change Log

Reason for change	Issue Number	Rev. Number	Date
Initiated	0	1	07Sep-2021

Distribution

Name/Organisation
Patrick Griffiths/ESA
Sara Arpacio/ESA

Contents

List of Figures

1 Executive Summary

This report describes the AI-Ready EO training datasets project, undertaken by the Irish Centre for High-End Computing (ICHEC) and the National Centre for Applied Data Analytics and AI (CeADAR, UCD) and supported by ESA Phi Lab.

In this project, a specification and best practice guidelines for producing AI-ready EO datasets were created, describing how to prepare both EO and reference datasets (such as ground measurments) for machine learning. A Python library was developed to enable AIREO training datasets (TDS) to be created and easily used by ML practitioners.

Four pilot datasets were adapted to comply to the specification and as examples to demonstrate the library functionality: Forest biomass (a regression and correlation task), Arctic sea-ice via Sentinel-1 and AMSR (boundary detection, segmentation and identification), the Austrian CAP set (segmentation, classification and boundary detection) and SpaceNet7, a multi-temporal urban development dataset.

A community network was created with over 100 organisations who were surveyed at the beginning of the project and the final specifications and work were presented at a workshop, with detailed engagements with standards groups and major industry players (OGC, the STAC community).

All materials were published at https://www.aireo.net, with the datasets available from the EODC and the Python library from both GitHub and PyPI.

2 Introduction

In this project, we set out to research, develop and formulate dataset specifications and best practice guidelines that define what constitutes AIREO training datasets. This included adapting a set of pilot datasets to be "AIREO-compliant", exposing this to the community, analysing and incorporating any feedback, and developing a Python library to make it easy for Machine Learning practitioners to use these datasets to test new algorithms, as well as for EO practitioners to develop new training datasets for this purpose.

3 Community consultation

A consultation process was set up with the EO and ML communities, including setting up a network of contacts (the AIREO Network) with email and a survey, including detailed one-to-one consultations. This was summarised in the Community Consultation report, Deliverable 1, submitted and accepted in November 2020.

4 State of the Art

Based on contact with the AIREO network and survey, and research into the State of the Art, a report was written and submitted as Deliverable 2 to ESA in December 2020. This covered file formats, metadata specifications, data sharing and principles, existing projects and future requirements.

5 Workshops

Once initial investigations into the state of the art were complete, a Workshop was held as a side-event to the ESA EO Phi week in 2020. As well as presentations on AIREO work to date and the feedback on the survey of the network, a number of break-out sessions were held to gather detailed feedback on the proposed specification and pilot datasets.

When the initial v0 of the AIREO resources were ready for comment, a second round of consultation and a second workshop was held on 20th July 2021. Again, breakout sessions were held and in-depth consultations were carried out with representatives from the OGC and STAC communities amongst others.

The workshop and evolution report was submitted as Deliverable 8.

6 Presentations and Publication

In addition to the dedicated workshops held with the AIREO network, the AIREO project was presented in public at the Open Geospatial Consortium GeoAI DWG meeting, 14th September 2020, and at the European Geosciences Union (EGU) 2021 Virtual meeting, 26 April 2021. It was well received at both, with follow-on inquiries. A poster was presented at ESA EO Phi Week, 2020.

A dedicated website (www.aireo.net) has been created that contains the Specification, best practices and documentation on the pilot datasets, along with links to the Jupyter notebooks hosted on EuroDataCube. The software and notebooks are also available in GitHub (aireo-project); the libraries are available in GitHub and PyPI, the Python package index.

A blog post was also published at medium.com. The communications report was published as Deliverable 5.

7 Specifications

From the initial requirements and State of the Art in Deliverable 2, a first draft (v0) of the AIREO specifications was created. It includes requirements, such as licensing and that any TDS adheres to FAIR (Findable, Accessible, Interoperable, Reusable) standards, introduces quality assurance standards that must be matched and the concept of compliance levels. It builds on the STAC data model so that TDS are cloud-native and may be dynamic (built upon dynamic collections, such as the growing Sentinel imagery datasets). The data model is extended to include the concept of "Areas of Interest" : sub-collections of geospatial data over a particular geometry that match the reference data.

The concept of "Profiles" was added: a TDS specifies which profiles it adheres to, which enables the user/library to know which metadata will be present in the dataset. Metadata were specified for each profile, mostly building on existing standards. New metadata were added for quality indicators and a "compliance level" defined: whether the TDS contains just essential metadata, recommended metadata or all optional metadata.

The standard includes tracking of provenance, but also embedded "feature recipes" : we record how to automatically regenerate the data both to inform the user but also to redo the work later, such as when more data is available.

The specification was presented at both workshops, and published as v1 after inclusion of feedback from the final workshop as Deliverable 3.

8 Best Practices Guidelines

Accompanying the formal specification, a set of best practices were developed. These cover all the issues raised in the specification (FAIR attributes, licensing, etc.) but also how metadata should be approached: what issues a TDS creator may need to pay attention to, such as best practice for splitting the dataset into training/validation/test splits, handling class imbalance, data versioning, etc. Again, recommendations were provided for both Earth Observation and Machine Learning audiences, who may not be familiar with each others' practices.

The best practices were published alongside the specifications at both workshops, and updated following the final workshop, and both are available on the website, www.aireo.net.

9 Python Library

A Python Library was developed to aid the development and use of AIREO TDS by both dataset creators and users. It has been designed around principles common across popular Machine Learning libraries, and it makes it possible to easily load data from an AIREO TDS into commonly used Machine Learning data formats (e.g. *xarray* format) and to train and test a model with only a few lines of code. This full load, train and test end-to-end ML workflow was demonstrated for one of the pilot datasets, with a demonstration model trained using TensorFlow.

The library also provides a framework for the creation of AIREO TDS, and is used in the pilot data sets. It provides functionality for the analysis of the TDS and for statistics of the data (useful for studying e.g. class imbalance) and visualisation of both the feature data (typically EO images) and reference data. The library was provided as Deliverable 7, uploaded and available at GitHub (aireo-project), PyPI.org (the Python package index), installed on the Euro Data Cube, and via the website www.aireo.net.

10 Pilot Datasets

Four pilot data sets were chosen in agreement with ESA:

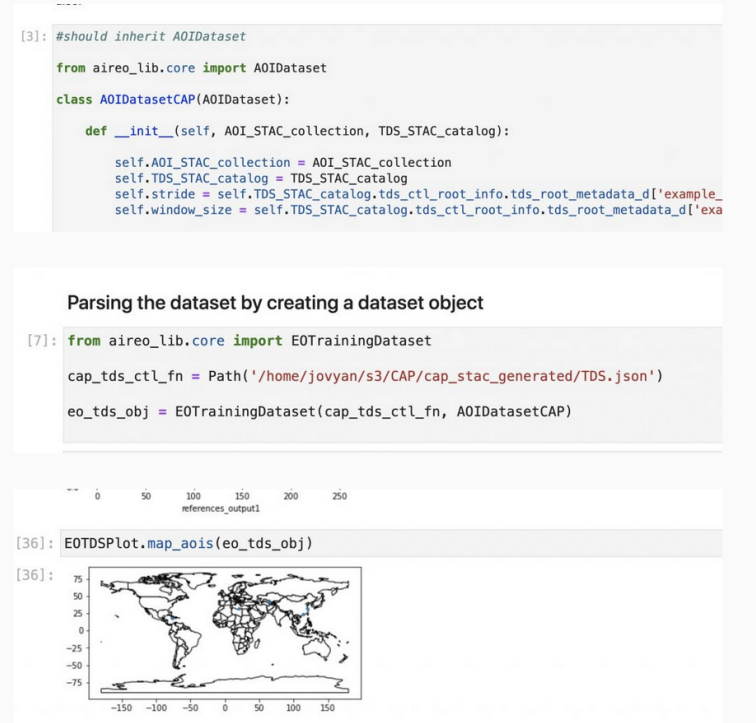

Figure 1: Example use of library, from CAP data set

1. *Biomass retrieval.* This provides canopy and biomass measurements for 260 forest plots, and corresponding Sentinel-2 imagery. This is an example of regression and correlation task.

2. *AI4Arctic Sea Ice.* Sentinel-1 (SAR) and Microwave (AMSR) imagery and corresponding Sea Ice concentrations. This provides boundary detection, segmentation and identification challenges.

3. *Common Agricultural Policy (CAP) Austria.* This includes information about crop types and field boundaries for segmentation, classification and boundary detection.

4. *SpaceNet7.* This is a multi-temporal urban development dataset, providing one image per month of 100 locations with building footprints for urban segmentation and detection challenges.

These datasets were with Jupyter notebooks available for exploration by users, demonstrating the use of the AIREO library. The datasets are annotated and broken into example sets for training, test and validation, with measured class imbalance. They are available on the Euro Data Cube, and through the website www.aireo.net. The datasets and documentation were submitted as Deliverable 6.

11 Summary

Specifications and best practice guidelines were developed to support the creation and use of EO training datasets for ML and AI applications, aimed at both Machine Learning and Earth Observation practitioners. Four pilot datasets were created adhering to the AIREO specifications and best practices, along with a Python library providing functionality for creating and working with AIREO TDS. The primary innovations of this work are Quality Assurance automation in the library, to provide consistency and completeness; Provenance tracking, to provide full data traceability and compliance with FAIR principles and embedded Feature engineering recipes, to

enable automated recreation of Training data sets. AIREO STAC extensions were created to enable cloud-native metadata and datasets which have been presented to the OGC and STAC communities. This is continuing with the merging of these extensions into STAC.

We have proposed further work for a CCN or AIREO2 project to include both continued work with the standardisation committees in merging the changes, additions to the AIREO library for ease of use in handling new reference data, extensions to new data set types, including profiles for non-imagery data such as LIDAR and 3d Point cloud data (e.g. atmospheric composition), benchmarking of defined TDS and the creation of a platform for AIREO datasets.