earthpulse!

EOX

space·tec PARTNERS

esa

European Space Agency

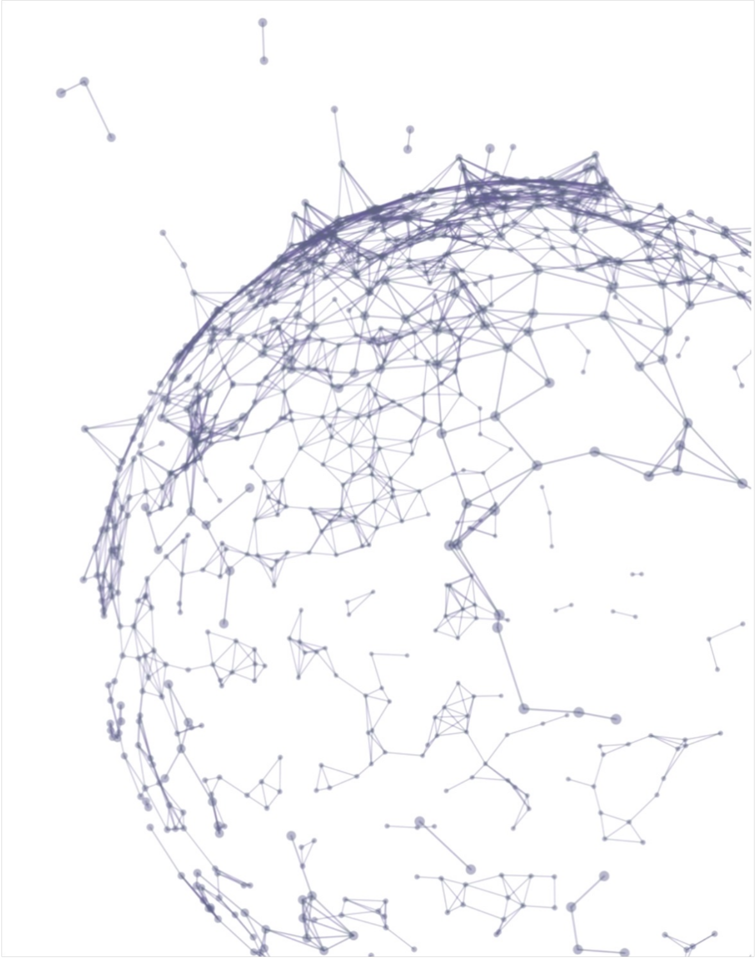
SINERGISE

Brockmann Consult

EARTH OBSERVATION TRAINING DATA LAB

Draft Brochure

27/10/2022



The Earth Observation Training Data Lab

- The European Space Agency (ESA) **Earth Observation Training Data Lab** (EO-TDL) will revolutionise the field of Earth Observation (EO) by providing an accessible repository of training data for machine learning and artificial intelligence as well as open-source tools to create, share, and improve datasets. Machine learning algorithms can be explored and trained with ease thanks to integrated cloud computing and a well-documented toolbox.

- The EO-TDL is expected to launch in September 2023.

AI4EO

- One of the most limiting factors of AI is the scarcity of suitable and accessible **Training Datasets** (TDS). As the name suggests, TDS are used to train an AI model to perform a specific task or identify patterns. The main barrier currently is that gathering and labelling EO data is a convoluted process. Some techniques exist that can help alleviate this issue, for example transfer learning or unsupervised learning, but annotated data is always required for fine-tuning and final validation of AI models.

- Gathering and labelling EO data is especially time consuming and expensive. Data access is usually limited and costly, especially for Very High Resolution (VHR) images that allow objects like trees to be clearly identified. Even after it is obtained, domain experts or even in-person (in-situ) trips are required to manually confirm the objects in a satellite image are correctly annotated with a high degree of quality. This results in the field of AI4EO lagging when compared to other fields, impeding the development of new applications and limiting the full potential of AI in EO .

Objectives

- To contrast the lack of suitable, open data, the EO-TDL has the **objective of providing a set of open-source tools to generate, curate, analyse, and use AI-ready EO datasets.** This platform will act as a cloud repository, where TDS can be created, imported, maintained, and improved by everyone. Additionally, cloud computing infrastructure will be available for training models with the datasets directly on the cloud with multi-GPU machines.
- The EO-TDL will thus create a hub for ML training data, with a focus on EO applications. There will be a particular emphasis on datasets to overcome AI4EO issues: Unsupervised learning, Data Fusion, multitask Learning and the development of custom architectures for EO.

Who is this for?

- The EO-TDL will be available to researchers and engineers, but also to AI4EO enthusiasts!
- Many areas will benefit from this platform. Having a repository of AI-ready EO datasets will strengthen industry capabilities for exploiting EO data as a whole and help accelerate EO market penetration. Furthermore, to enable Digital Twin Earth simulations , access to these quality datasets is necessary for researchers and engineers as they build and apply quality models.
- Data with large geographical or temporal scopes will be prioritised using the platform, meaning researchers that require more computational power will not be obstructed by sporadic users. However, smaller-scale datasets will also be provided, allowing novice users to familiarise themselves with the platform and encourage faster prototyping.

Community

- The flagship characteristic of the EO-TDL is its community and open-source nature. Given the users' capacity to create their own TDS and access others', alongside the possibility of using pre-trained ML models, the EO-TDL's growth will be directly linked to the number of its users.

- Not only will users be able to train their AI models on the cloud with the available datasets, but they will also be encouraged and incentivised to contribute to the enrichment of the platform. Users making significant contributions, such as the addition of new datasets or the enhancing of existing ones, will be rewarded in the form of credits. A larger community will thus correspond to a better and more versatile platform for everyone.

- The EO-TDL will also allow integration by popular Machine Learning libraries, like Scikit-Learn, Pytorch and Tensorflow. Furthermore, the platform will provide tools to ingest datasets (including legacy data) with automated quality assurance procedures: such tools include python libraries, web APIs and UIs.

- Additionally, the EO-TDL's source code will be hosted on Github as a public repository, encouraging the open-source community to contribute however possible. Users will also be allowed to link and publish curated datasets with a Digital Object Identifier (DOI) that follows a known standard. The platform will also support versioning of datasets and models.

EARTH OBSERVATION TRAINING DATA LAB

Draft Brochure

27/10/2022