Open
Geospatial
Consortium

TESTBED-18: MACHINE LEARNING TRAINING DATA ER

ENGINEERING REPORT

PUBLISHED

CONTENTS

LIST OF TABLES

LIST OF FIGURES

I   ABSTRACT

This OGC Testbed 18 Engineering Report (ER) documents work to develop a foundation for future standardization of Training Datasets (TDS) for Earth Observation (EO) applications. The work performed in the Testbed 18 activity is based on previous OGC Machine Learning (ML) activities. TDS are essential to ML models, supporting accurate predictions in performing the desired task. However, a historical absence of standards has resulted in inconsistent and heterogeneous TDSs with limited discoverability and interoperability. Therefore, there is a need for best practices and guidelines for generating, structuring, describing, and curating TDSs that would include developing example software/packages to support these activities. Community and parallel OGC activities are working on these topics. This ER reviews those activities in parallel with making recommendations.

II   EXECUTIVE SUMMARY

This OGC Testbed-18 ER begins by providing an introduction to Artificial Intelligence/ML in the context of EO. The introduction is followed by a review of existing approaches to creating and storing TDSs. Then, TDS formats are reviewed in terms of metadata, creating a catalog, expressing quality, and adherence to Findability, Accessibility, Interoperability, and Reuse (FAIR) principles. Finally, the summary reviews the next steps, best practice ideas, and the geoethics of generating and distributing training data.

III   KEYWORDS

The following are keywords to be used by search engines and document catalogues.

Artificial Intelligence, Earth Observation, Machine Learning, Training Dataset

IV PREFACE

IV.A. Foreword

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. The Open Geospatial Consortium shall not be held responsible for identifying any or all such patent rights.

Recipients of this document are requested to submit, with their comments, notification of any relevant patent claims or other intellectual property rights of which they may be aware that might be infringed by any implementation of the standard set forth in this document, and to provide supporting documentation.

V SECURITY CONSIDERATIONS

No security considerations have been made for this document.

VI SUBMITTERS

All questions regarding this submission should be directed to the editor or the submitters:

NAME	AFFILIATION	ROLE
Sam Lavender	Pixalytics Ltd	Editor
Kate Williams	FrontierSI	Editor
Caitlin Adams	FrontierSI	Editor
Ivana Ivánová	Curtin University	Editor
Jim Antonisse	NGA	Contributor
Sara Saeedi	OGC	Contributor
Sina Taghavikish	OGC	Contributor

1

SCOPE

1 SCOPE

The Open Geospatial Consortium (OGC) Testbed-18 initiative aimed to explore six tasks, including advanced Interoperability for: Building Energy; Secure; Asynchronous Catalogs; Identifiers for Reproducible Science; Moving Features and Sensor Integration; 3D+ Data Standards and Streaming; and Machine Learning (ML) Training Data (TD).

The goal of this Testbed-18 task is to develop the foundation for future standardization of Training Datasets (TDS) for Earth Observation (EO) applications. The task has included evaluating the status quo of TD formats, metadata models, and general questions of sharing and re-use. It has taken into account several initiatives, such as the European Space Agency's Artificial Intelligence-Ready EO Training Datasets (AIREO), the Radiant MLHub, and the SpatioTemporal Asset Catalog (STAC) family of specifications.

For the purposes of this Engineering Report (ER), the authors define EO data as data that has been collected through remote sensing, including passive and active sensors carried on drones, airplanes, helicopters, or satellites.

In terms of ML applications, the most appropriate scope is supervised learning, as this type of ML directly leverages labeled training datasets. However, unsupervised learning will also be considered. These types of learning are also appropriate for the context of this work within the field of EO, as the application of ML in EO is often focused on the goal of identifying meaningful features from input EO data using a set of known mappings between inputs and desired outputs (the training dataset).

In laying out a path for future standardization of training datasets for EO applications, the ER has also taken into account and collaborated with the Training Data Markup Language (DML) for AI Standards Working Group (SWG). The SWG is chartered to develop the Unified Modelling Language (UML) model and encodings for geospatial ML training data. While these Testbed-18 activities have progressed, the SWG have released draft versions of their Conceptual Model Standard (part 1) and JSON Encoding (part 2).

2

NORMATIVE REFERENCES

---

2 NORMATIVE REFERENCES

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

AIREO: **Best Practice Guidelines** https://www.aireo.net/aireo-training-dataset-best-practice-guidelines/

AIREO: **Specification** AIREO specification

W3C: **Data on the Web Best Practices**, W3C Best Practice, 2017 https://www.w3.org/TR/dwbp/

Gebru, T. , J. Morgenstern , B. Vecchione, J.W. Vaughan, H. Wallach, H. Daume III, and K. Crawford. Datasheets for datasets. Communications of the ACM. 2021, 64(12):86–92. https://doi.org/10.1145/3458723

Lavender, S. Detection of Waste Plastics in the Environment: Application of Copernicus Earth Observation Data. Remote Sens. 2022, 14, 4772. https://doi.org/10.3390/rs14194772

McKee, L., C. Reed, and S. Ramage. 2011. "OGC Standards and Cloud Computing." OGC White Paper. Accessed 29 March. http://www.opengeospatial.org/docs/whitepapers

OGC: **API-Records** https://github.com/opengeospatial/ogcapi-records/

Oxford Reference: **Artificial intelligence** https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095426960

Yue, P., Shangguan, B., Hu, L., Jiang, L., Zhang, C., Cao, Z., Pan, Y., 2022. Towards a training data model for artificial intelligence in earth observation. International Journal of Geographical Information Science, 36(11), pp. 2113-2137, https://doi.org/10.1080/13658816.2022.2087223

3

TERMS, DEFINITIONS AND ABBREVIATED TERMS

———

3 TERMS, DEFINITIONS AND ABBREVIATED TERMS

This document uses the terms defined in OGC Policy Directive 49, which is based on the ISO/IEC Directives, Part 2, Rules for the structure and drafting of International Standards. In particular, the word "shall" (not "must") is the verb form used to indicate a requirement to be strictly followed to conform to this document and OGC documents do not use the equivalent phrases in the ISO/IEC Directives, Part 2.

This document also uses terms defined in the OGC Standard for Modular specifications (OGC 08-131r3), also known as the 'ModSpec'. The definitions of terms such as standard, specification, requirement, and conformance test are provided in the ModSpec.

For the purposes of this document, the following additional terms and definitions apply.

3.1. Application Programming Interface

An Application Programming Interface (API) is a standard set of documented and supported functions and procedures that expose the capabilities or data of an operating system, application, or service to other applications (adapted from ISO/IEC TR 13066-2:2016).

3.2. OGC APIs

The family of OGC standards developed to make it easy for anyone to provide geospatial data to the web.

3.3. Abbreviated terms

ADES	Application Deployment and Execution Service
AI	Artificial Intelligence
AP	Application Package
ARD	Analysis Ready Data

AWS	Amazon Web Services
CEOS	Committee on Earth Observation Satellites
DML	Data Markup Language
EMS	Exploitation Platform Management Service
EO	Earth Observation
ER	Engineering Report
ESA	European Space Agency
FAIR	Findability, Accessibility, Interoperability, and Reuse
ML	Machine Learning
OGC	Open Geospatial Consortium
STAC	SpatioTemporal Asset Catalog
SWG	Standards Working Group
TD	Training Data
TDS	Training Dataset
TrainingDML-AI	Training Data Markup Language for Artificial Intelligence
UML	Unified Modelling Language

4
ENGINEERING REPORT OVERVIEW

———

4 ENGINEERING REPORT OVERVIEW

Artificial Intelligence (AI) and Machine Learning (ML) algorithms have great potential to advance processing and analysis of Earth Observation (EO) data. Among the top priorities for efficient machine learning algorithms is the availability of high-quality Training Datasets (TDSs). Training data (TD) is the initial dataset used to train ML algorithms. Models create and refine their rules using this data. Training data are also known as a training dataset (TDS), learning set, or training set.

TDSs are crucial for ML and AI applications, but they can also become a significant bottleneck in EO's more widespread and the systematic application of AI/ML due to:

- the absence of standards resulting in inconsistent and heterogeneous TDS (data structures, file formats, quality control, meta data, repositories, licenses, etc.);

- limited discoverability and interoperability of TDS; and

- lack of best-practices and guidelines for generating, structuring, describing, and curating TDS.

The Engineering Report (ER) starts by providing an introduction to AI/ML in the context of EO (Clause 5) followed by a review of existing approaches to creating and storing TDSs (Clause 6). The relevant standards applicable in terms of metadata (Clause 7), creating a catalog (Clause 8), and expressing quality (Clause 9) are reviewed in subsequent sections. Unlocking the power of geospatial resources requires that those resources are stored following the Findability, Accessibility, Interoperability, and Reuse (FAIR) principles. The FAIR principles are reviewed concerning TDS (Clause 10). Finally, the summary (Clause 11) reviews the next steps, best practice ideas, and geoethics of generating and distributing TD.

5

INTRODUCTION TO AI/ML WITHIN THE CONTEXT OF EARTH OBSERVATION

---

5 INTRODUCTION TO AI/ML WITHIN THE CONTEXT OF EARTH OBSERVATION

This section outlines the overall scope for this ER, beginning with a summary of Artificial Intelligence (AI) and Machine Learning (ML). Next is a discussion of how AI/ML are used in the field of Earth Observation (EO) followed by a series of case studies that demonstrate the application of ML techniques to EO data in a variety of contexts. Finally, this section concludes with a discussion of foreseen issues and opportunities in relation to the creation of an OGC standard for Training Datasets (TDSs).

5.1. Defining AI and ML

As a field, AI covers "the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages" Oxford Reference. ML is then a subset of the AI field, specifically focusing on the creation of algorithms that learn from data without explicit programming. The output of these algorithms is then a trained ML model, which can process new inputs.

Within the domain of ML, there are three categories of application, each distinguished by their learning method.

- Supervised learning: The algorithm is provided with a labeled TDS which pairs input data with a training label. The algorithm creates an output from the input data and compares this with the training label then iteratively updates itself to maximize its accuracy in comparison to the desired output.

- Unsupervised learning: The algorithm is provided with input data and attempts to identify commonalities and differences in the data that allow it to be grouped. Once groups are established, the algorithm is able to identify which group new input data should belong to.

- Reinforcement learning: The algorithm is exposed to an environment to which it must respond and is then rewarded if it responded appropriately.

This Engineering Report (ER) primarily focuses on supervised learning applications. Supervised learning algorithms include linear regression, decision trees, support vector machines, and neural networks. In particular convolutional neural networks are commonly used in EO applications due to their ability to identify features in images. Convolutional neural networks can be used for semantic segmentation (pixel-level classification) and object detection (bounding box identification).

5.2. Typical formats for TDSs in EO Applications

A TDS is made up of two components: input data and training labels. For EO applications, the input data are remote sensing observations: Multispectral and hyperspectral satellite imagery, red, green, and blue (RGB) aerial photography, drone-mounted LiDAR point clouds, and so on. The training labels capture the classification and location of known features in the input data, such as the location and bounds of a building, or the type of crop at a given location. Training labels can be provided in vector or raster format, depending on the ML application. Supervised learning applications require both the input data and the training labels whereas unsupervised learning applications only require the input data.

5.3. Example use cases

This ER section describes some ML for EO use cases to provide context. The use cases are based on Testbed participants' experience particularly where ML is used to create spatial data (vector and raster) from EO data for a broad range of use cases.

The use cases are from efforts from state and national governments to create spatial data from ML on jurisdictional areas as a method of automating the creation of these products. In these use cases, reusability is a key consideration. These agencies (for example) may wish to later extend the existing training set to keep the dataset current, as well as to explore new questions and produce new information products.

As part of each use case, the authors provide a summary of implications for a TDS standard. For a specific review of the suitability of the proposed TraningDML-AI standard for use cases, see Table A.1.

5.3.1. Use case: Mapping vegetation from aerial imagery in Australia

Project Participants: The Department of Environment, Land, Water and Planning (Victoria, Australia), FrontierSI (Australia), Orbica (New Zealand).

Project Goal: Mapping tree cover consistently over time can help governments understand land use, urban heat, and fire risks. The Victorian Department of Environment, Land, Water and Planning sought to develop an automated machine learning approach that could map tree cover from collected high resolution aerial ortho-photography.

Challenge: Due to the labor-intensive process that was traditionally used to create and update the data, Vicmap Vegetation statewide data had not been updated for 20 years. The State of Victoria wanted to trial the use of ML for statewide data maintenance, using vegetation as a test case. Project participants wanted to create a repeatable method to update data as new imagery was acquired.

Input data: The input data was 10-20cm red, green, and blue aerial ortho-photography. All imagery was resampled to 20cm before being labeled and provided to the ML algorithm.

Training labels: Training labels were created by hand digitizing vector polygons delineating areas of a range of tree coverage and density. Initial labels were then used to train a semantic segmentation algorithm to produce further Training Data (TD), which was then refined into polygons by human examination.

Training data selection: TD were stratified using a vector dataset of ecological bioregions to capture training data covering the range of tree species and ecosystems in Victoria.

Method: The project used semantic segmentation (U-Net) with transfer learning as the ML model architecture. The project successfully created a statewide dataset and delivered the resulting vector data, training data, and scripts to re-run the ML process in the future.

Key metadata

- Geographical extent and coverage

- Extent and data summarizing ecological bioregions

- Resolution, spectral wavelength range, date, seasonality, and quality of input data

- Method of generation (human only or ML with human revision)

- ID for association with ML model metadata to understand which data were used to train the model

Implications for a TDS standard

Input data may be modified from the original source for the purpose of creating a TDS. In this case study, all input data were resampled to a uniform resolution and labels were delineated from the resampled data. As such, the labels are appropriate for the resampled imagery and should be used with caution on imagery at a higher resolution. For input data supplied alongside training labels, a TDS standard should capture any modifications that have been made relating to the creation of the labels.

ML projects may use multiple methods to create TDSs. In this case study, humans provided an initial set of labels and used these to train an ML process to produce additional labels. For quality control, a TDS standard should capture how a given training data label was produced.

5.3.2. Use case: Capturing footprints of building roofs (roofprints) from aerial imagery in Australia

Project Participants: The Department of Environment, Land, Water and Planning (Victoria, Australia), DSM Geodata.

Project Goal: Detailed and accurate building outlines can assist decision making for planning, infrastructure, and risk modelling. The project goal was to derive high-accuracy building roofprint models from existing aerial ortho-photography.

Challenge: The derived roofprints needed to be highly accurate to meet the needs of various sectors. At the time, no commercially available products met these needs. The project approach involved training an ML model and then manually revising the predicted rooflines to match the underlying imagery.

Input data: The input data was 10cm red, green, and blue aerial ortho-photography.

Training labels: Training labels were created by hand digitizing vector polygons delineating rooflines.

Training data selection: Validation data were collected from either the residential or non-residential zones of a core urban area.

Method: Computer vision (exact model architecture unknown) with outputs reviewed and cleaned by humans to achieve high accuracy.

Key metadata

- Geographical extent and coverage

- Definition of residential and non-residential zones

- Designation of whether the data belongs to the training or validation set

- Resolution, spectral wavelength range, date, seasonality, and quality of input data

- Method of generation (human only or ML with human revision)

- ID for association with ML model metadata to understand which data were used to train the model

Implications for a TDS standard

The TDS in this use case contained a specific validation set with labels from both residential and non-residential zones of a specific area. Validation sets may be specifically designed to represent the expected variability and presence of features in the domain. A TDS standard should provide an optional way for a creator to distinguish between elements of the TDS that belong to the training, validation, and test sets. A future user could then review validation samples to understand the domain the training data were designed for or use the same validation set with a new ML process and fairly compare the performance of the new method with an existing one.

5.3.3. Use case: Capturing flood extent from aerial imagery in Australia

Project Participants: The Department of Customer Service, Spatial Services Division (New South Wales, Australia), Charles Sturt University (Australia), Deloitte (Australia), Intellify (Australia)

Project Goal: Emergency response efforts require timely access to flood boundaries to aid planning, rescue, recovery, and rebuilding. The project goal was to automatically delineate flood extent from post-flood aerial ortho-photography.

Challenge: Imagery alone is challenging for humans to interpret, particularly those in emergency response that have not been trained to interpret four-band imagery. ML for automated boundary detection provides an opportunity to deliver an easily interpretable data product soon after imagery capture, aiding emergency response.

Input Data: The input data was 15cm red, green, blue, and near infra-red aerial ortho-photography. For the final ML process, three-channel imagery was used containing the near infra-red, red, and green values.

Training labels: Areas identified as flood or non-flood for captured imagery.

Method: The project used an unsupervised Gaussian mixture model which identified clusters in the data. The identified clusters were then compared to labeled imagery to assign either flood or non-flood labels. When run on imagery the Gaussian mixture model returned each pixel's probability of being drawn from each of the identified clusters. Once clusters were labeled as flood or non-flood, pixels that had a high probability of having been drawn from a flood cluster could be labeled as flood.

Key metadata

- Geographical extent and coverage

- Definition of flooded areas

- Resolution, spectral wavelength range, date, seasonality, and quality of input data

Implications for a TDS standard

While the ML process used in the case study was an unsupervised learning method, the project team still used training labels to identify clusters and then classify new input data as flood/not flood. A TDS standard should allow flexibility in how training labels are specified relative to the input imagery because the same TDS can be used as the input to many different ML approaches. Unnecessary rigidity in a TDS standard may prevent it from being applicable to newly developed TDS formats and ML applications.

5.3.4. Use case: Classifying crops by type from satellite imagery in Zambia

Project Participants: FrontierSI (Australia), Tetra Tech (United States of America), Digital Earth Africa (South Africa).

Project Goal: Food security is a key issue in Africa. Knowledge of crop extents and types can help governments ensure access to food and plan for future. The project goal was to use ML analysis of satellite imagery and other EO products to estimate the extent of major crop types and thus availability of produce to assist with food security management in Zambia.

Challenge: The use of on-ground surveys to understand distribution of crop extent for food security is a time consuming and expensive process. Zambia needed to develop a repeatable and scaled country-wide process to provide estimates of crop type to inform food availability in a timely manner.

Input data: The input data was analysis-ready Sentinel-2 (multispectral) and Sentinel-1 (radar) satellite imagery between 10-60m resolution as well as ancillary data sets such as rainfall, digital elevation models, and analytic products derived from Landsat (multispectral) satellite imagery.

Training labels: Training labels were created using on-ground field collection with GPS-enabled mobile device to associate human identified crop type with point location. If the collector could not enter the field, the point was captured on the road and later moved into the area of the relevant crop. The vector points were labeled with the crop type with each point associated with a specific field.

Training data selection: Unsupervised learning was applied to satellite data over known cropping areas to identify clusters of spectral variability. These were then sampled to suggest locations for collecting training data randomly stratified by the area covered by each class from the unsupervised learning process.

Method: The project used supervised random forest as the ML model architecture. The project successfully created a country wide dataset and delivered the resulting raster data, training data, and scripts to re-run the process in future.

Key metadata

- Geographical extent and coverage

- Sampling strategy

- Date and time of field collected label

- Crop status at time of label (e.g., sown, ready for harvest, harvested, fallow)

- List of classes and number of observations of each

- Location of point relative to target field (e.g., center, roadside)

- GPS accuracy of the point location

- Whether the point has been updated by a human reviewer (e.g., moved from road to field while completing quality assurance)

Implications for a TDS standard

In this use case, the training data labels were collected through field sampling with a GPS-enabled device, meaning that they are not specifically tied to an input data source. If the date of capture for the training data labels is supplied then the data should still be considered a valid TDS. This is because such information would be sufficient for a user to select appropriately matched input data. A TDS standard should be aware that TDSs do not necessarily require associated input data, even though most TDSs will have this.

5.3.5. Use case: Detection of plastics and waste across the world in terrestrial and marine environments.

Project Participants: Pixalytics (UK), CLS (France & Indonesia), RisikoTek Pte Ltd(Singapore), rasdaman GmbH (Germany).

Project Goal: Focuses on the use of a ML TDS for the detection of waste plastics. It was developed in conjunction with two projects.

- Marlisat: European Space Agency (ESA) funded study with the overarching objective of developing a unique combination of three innovative components to constitute a plastic anthropogenic marine debris monitoring system. The components were EO for detecting the source and impact of plastics, a low-cost satellite tracker deployed at sea, and a modeling tool to understand the at-sea plastic debris transport.

- Space Detective: Singapore-funded project with the goal of detecting waste plastics, including tires on land, so they can be recycled.

Challenge: The detection of plastics, whether it be plastic, tires, or mixed waste in waste sites across the globe in multiple land cover environments.

Input data: The primary input data was Sentinel-2 and Sentinel-1 satellite data which was supplemented by a digital elevation model and vector layers for roads and coastlines for background mapping with high-resolution commercial data to support focused activities.

Training labels: Training pixels were manually identified using a combination of the high spatial resolution satellite imagery within Google Earth and the Sentinel-2 RGB color composites for the different land cover types. Where the locations of the plastics could not be reliably identified, these land cover classes were not digitized, and only the background land cover classes were digitized so as not to reduce the accuracy of the overall dataset.

Training data selection: The test sites were accumulated over several years by reviewing peer-reviewed papers, reports, and news articles on plastic waste and its detection using remote sensing. This work continues as new sites are discovered and new versions of the model are generated. The test sites are separated into training/validation/testing datasets so that the data used to validate the model is not the same as the training data. Also, test data were chosen carefully as ML models often exhibit unexpectedly poor behavior when they are deployed in real-world domains which has identified as being caused by underspecification — where observed effects can have many possible causes. Also, as the plastics classes have low numbers of pixels compared to the broader land cover classes, such as clouds, there was class imbalance during the training. Therefore, in training the model, a re-weighting is applied to reduce the number of pixels for the classes with high numbers and increased the number of pixels for classes with low numbers through duplication.

Method: The project used a sequential artificial neural network (ANN) and post-ANN decision tree. The ANN on its own experienced confusion due to the broad range of environments it was applied within. A post-ANN decision tree allowed the user to decide whether the results were conservative or relaxed. For example, a conservative approach was adopted when time-series

datasets were automatically processed to prevent a build-up of false positives which became distracting to users when observing composite outputs. See Lavender 2022 for further details.

Key metadata:

- Geographical extent and coverage

- Resolution, spectral wavelength range, date, and quality of input data

- List of classes and number of observations of each

- Links to literature sources identifying test sites

- Designation of whether the data belongs to the training or test set

Implications for a TDS standard

In this use case, the TDS was compiled over multiple years due to the discovery of new test sites. Metadata identifying when a given TDS element was added, the test site it relates to, and whether it should be used within the training, validation, or test set, is valuable to a new user who may only want to use data related to a particular test site. A TDS standard must support a TDS to be updated over time including the addition of new entries which would also include being able to "version" the TDS, so that analyses can be compared over time as new entries are added.

While the entries in this use case came from the same input data, TDSs could be compiled from multiple input sources if they are maintained for a long period. This implies that individual elements of a TDS need to clearly capture the metadata of their associated input data.

5.3.6. Use case: mapping coastal bathymetry using ML by combining multiple data sources.

Project Participants: Satellite-Derived Bathymetry (SDB) activities to increase the global coverage of accurate bathymetry maps.

Project Goal: This use case focuses on the use of a ML TDS to support the extraction of information from multiple types of EO data in support of extending the limited bathymetric data collection possible from vessels and airplanes.

Challenge: The techniques used for the detection of bathymetry varies according to both water depth and turbidity, and the TDS could therefore contain data from multiple source types that have different operating characteristics and uncertainties.

Input data: The input data can be any of the following.

- **Lidar optical data** such as airborne LiDAR measurements, and satellite ICESat-2 data.

- **Multispectral optical data** sources such as high resolution Landsat and Sentinel-2 data alongside very-high resolution satellite missions such as WorldView.

- **Sonar data** from underwater instruments such as single and multi-beam echo sounders.

Training labels: The input data will be labeled with the bathymetric depth.

Training data selection: Depending on the location of interest, e.g., whether it is small area such as a port or global coverage, the source of the training data will vary.

Method: The approaches use ML methods such as random forest, e.g., TCarta who used ICESat-2 and Sagawa et al. 2022 who used multi-temporal satellite EO data to create a generalized model, and Zhong et al. 2022 who used a deep learning framework containing a 2D convolutional neural network.

Key metadata:

- Geographical extent and coverage

- Input source, including spatial resolution

- So far, SDB data are not considered as hydrographic data (i.e., can be used for charting for navigational purposes) because of their lower accuracy and the difficulty of estimating uncertainties compared to data from conventional sensors (such as echo sounders and LiDAR). Therefore, uncertainties of the input TDS are vital for progress to be made.

Implications for a TDS standard

A TDS standard must consider how to describe data from multiple source types and their associated uncertainties. Also, the SDB TDS will need to store the location in terms of both horizontal (latitude, longitude) and vertical (depth below defined water surface or height above a reference surface) coordinates.

5.4. Opportunities

As highlighted in the above case studies, ML has become widely used in the automated creation of insightful data products from EO data. As TDSs form the basis of ML approaches, a TDS standard has the potential to improve the quality and consistency of the application of ML to EO. This section covers specific opportunities that a TDS standard could enable.

The generation of a TDS is context-specific. The process is directly linked to the geospatial and temporal domain over which it is created, as well as the features it includes. A TDS standard would encourage TDS creators to provide this context along with the TDS. This allows future users to understand the TDS's applicability to a new domains, or to refresh or augment the TDS to capture different features of interest. As TDSs are often time- and resource-intensive to create, improved reusability of TDSs would be valuable for the EO community.

Our world is constantly changing, and features captured by a TDSs may become outdated over time. As such, the ability to describe and trace changes to features, along with versioning of TDSs, is important for ensuring ML applications are using valid data. A TDSs standard can aid the cataloging and versioning of TDSs.

ML processes need high quality and consistent TDSs to perform well. This may relate to either consistency in labeling across the TDS, or measures of its similarity to associated ground truth data. Having provenance and automated quality metrics captured by the standard would serve to help creators serve reliable and consistent TDSs and provide users with confidence in the TDS.

As agencies begin to rely on ML to produce automated products from EO data, it is critical that they are well-informed when creating or procuring TDSs. A TDS standard would support these agencies to request metadata that enables use and reuse as described above, without needing deep ML expertise. Clear descriptions of TDS metadata would also allow ML projects to be worked on by multiple providers, helping set clear expectations between the TDS creator and the TDS user, and allowing for transfer of a TDS across multiple parties.

5.5. Challenges

The case studies presented in this ER also highlight several challenges that must be considered in the development of a TDS standard. This section describes specific challenges that arise when working with TDSs and how these relate to the creation of a standard.

The use cases demonstrate that TDSs are created for highly specific domain problems. The challenge for a TDS standard will be to support creators in providing sufficient information about the domain. Without this, a new user cannot easily assess whether the TDS can be leveraged in their domain. Relevant domain information includes the following.

- Total geographic extent

- Spatial distribution of individual TDS elements

- Date and time of labeling

- Date and time of input data capture

- Properties of the input data and labels, including (but not limited to):

  - the source of the input data (e.g., a specific satellite or LiDAR instrument);

  - any corrections applied to the source data (e.g., terrain correction, top of atmosphere correction);

  - the features of the input data (e.g., spectral bands, derived features);

  - any properties of those features (e.g., spectral range, definition of any derived features, spatial resolution); and

  - uncertainties associated with the input data or labels (e.g., positional uncertainty from GPS, depth uncertainty from SDB).

- Designation to training, validation or test set, for individual TDS elements

- Description of sampling strategy

- Description of methods used to stratify the data

- Description of class imbalances present in the TDS

There are many methods that can be used to create training labels or input data, and multiple of these may be used within a single TDS. This may affect the overall quality of a TDS (discussed further in Clause 9), and a new user may wish to include, exclude, or revise particular elements based on their creation method. A TDS standard will need to ensure each element in a TDS can be labeled with the following information.

- Who created the label (with each individual assigned a unique ID), including (but not limited to)

  - a domain expert

  - a non-expert

  - a machine learning process

- The process for creating the label, including (but not limited to)

  - labeled from imagery by a human

  - generated by a machine learning process

  - collected in the field by a human

- Version history of the label

- Any accuracy measures related to the label (e.g., GPS accuracy for field-collected labels)

- The path to corresponding input data

- The process for creating the corresponding input data, including (but not limited to)

  - direct from source

  - augmented from source (e.g., rotated, shifted, mirrored)

  - synthesized (e.g., generated by a Generative Adversarial Network (GAN) or from simulations)

The development of a standard for TDSs should anticipate that TDSs will evolve over time, as new algorithms are developed and popularized. The challenge for a TDS standard will be in capturing the critical domain information described above while remaining flexible enough to accommodate future changes in the way TDSs are generated.

By having a TDS standard there is potential for TDSs to become more interoperable This is due to users having information on the limits of the domain of application for a given TDS. As such,

the standard needs to address the idea that a new TDS could comprise selected elements of existing TDSs and that the lineage is appropriately recorded.

6

CURRENT STATE OF ART

This section reviews previous and on-going activities relevant to the definition of an Artificial Intelligence (AI)/Machine Learning (ML) Training Dataset (TDS) Standard.

6.1. Training Data Markup Language for Artificial Intelligence Draft Standard

[peng] recognized that existing TDS, including open source benchmarks, can lack discoverability and accessibility plus there is often no unified method to describe the Training Data (TD).

The Training Data Markup Language for Artificial Intelligence (TrainingDML-AI) standard, released for internal review on August 2, 2022 is a conceptual model defined using Unified Modelling Language (UML) as a series of modules.
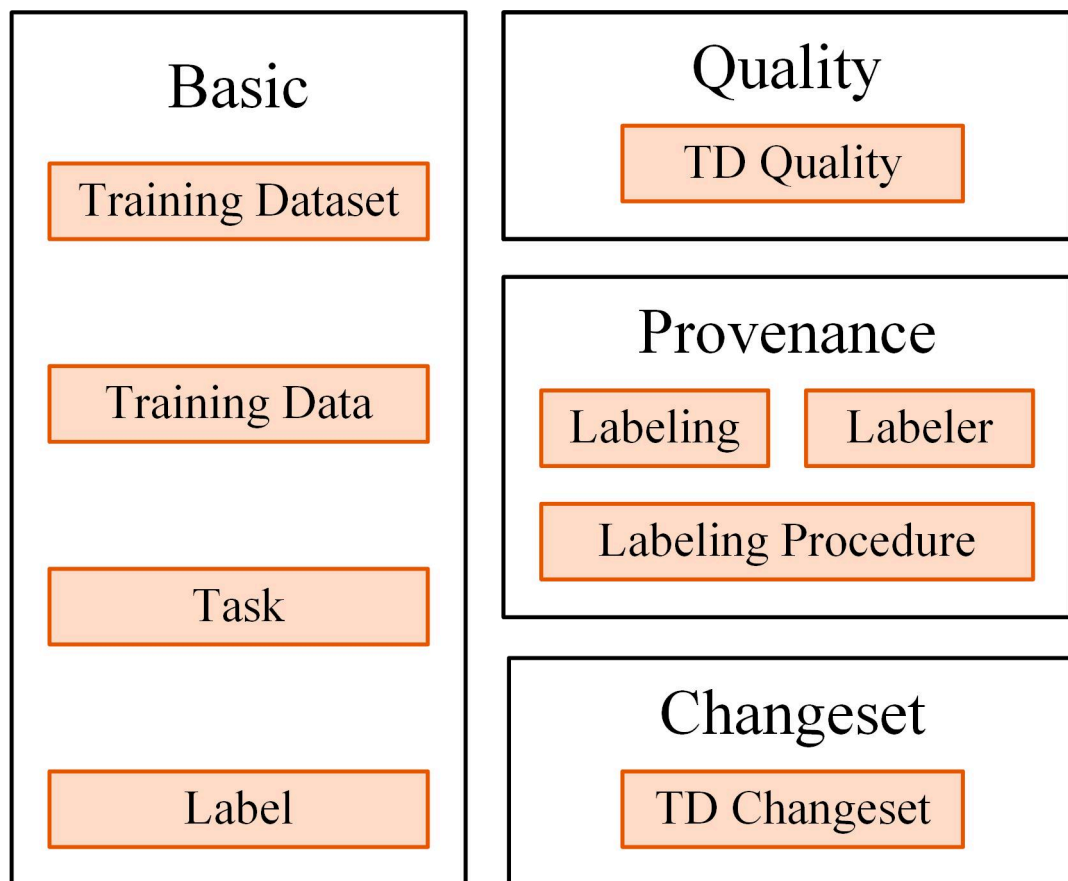
Figure 1 — TrainingDML-AI module overview.

TrainingDML-AI is designed as a universal information model that defines elements and attributes which are useful for a broad range of AI/ML applications. Any TD element may be augmented by additional attributes and relations whose names, data types, and values can be provided by a running application without requiring extensions to the TrainingDML-AI conceptual schema and respective encodings.

TrainingDML-AI builds on the ISO 19100 family of standards.
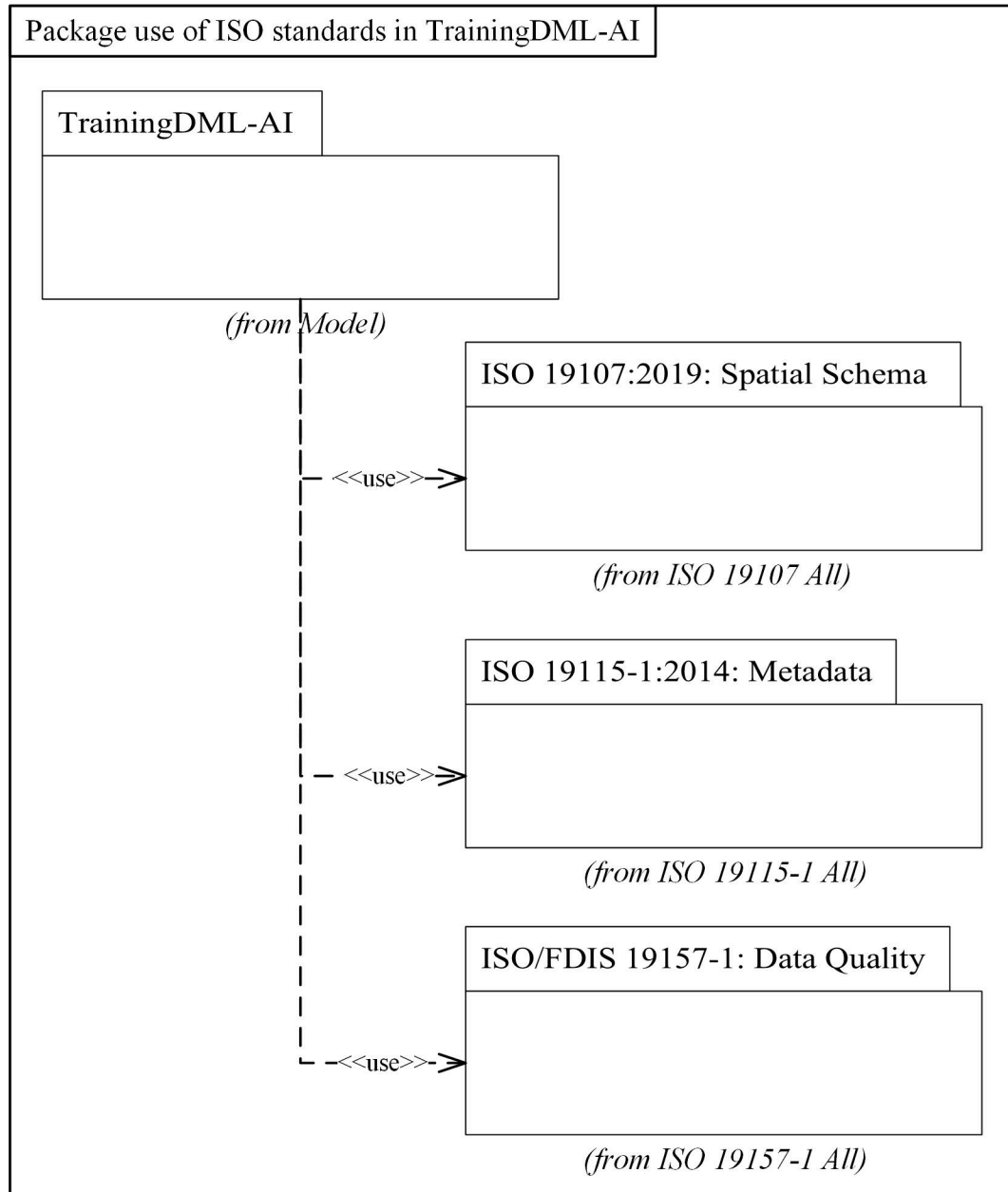
Figure 2 — Use of ISO Standards in TrainingDML-AI.

Annex B of the specification showcases the JSON encoding of example TDSs.

6.2. SpatioTemporal Asset Catalog (STAC)

The goal of the SpatioTemporal Asset Catalog (STAC) family of specifications is to standardize the way geospatial asset metadata is structured and queried. Of relevance to this Testbed-18 activity are the following extensions, which are currently (as of 03 August 2022) classed as Work In Progress.

- ML AOI: An Item and Collection extension to provide labeled training data for ML models.

- ML Model: An Item and Collection extension to describe ML models that operate on Earth observation data.

The ML AOI (Area of Interest) extension relies on, but is distinct from, the existing label extension. STAC items using the label extension link label assets with the source imagery for which they are valid. This is often as a result of human labeling effort. By contrast STAC items using the 'ml-aoi' extension link label assets with raster items for each specific ML model that is being trained.

6.3. ESA funded initiatives and projects such as AIREO

The European Space Agency (ESA) funded Artificial Intelligence (AI) Ready Earth Observation activity AIREO provides resources and tools to data creators and users to ensure their TDS are FAIR and to standardize aspects of TDS such as quality assurance and metadata completeness indicators.

The aim is for the AIREO TDS Specification to be applicable to all levels of predictive feature data and to other target variable types. The purpose of the first version of the specification was to generate feedback on the content and requirements from the community to ensure a more useful and relevant V1 specification. As such, the focus has been on a limited number of datasets with examples in the AIREO pilot datasets.
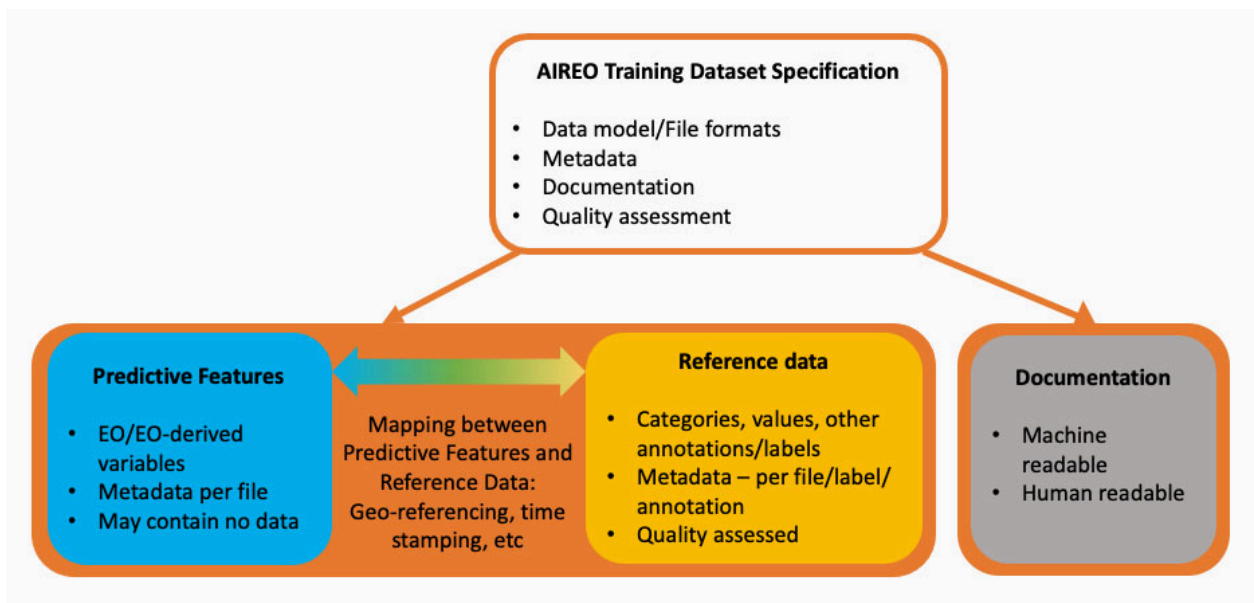
Figure 3 — Overview of the AIREO TDS specification.

Figure 4 shows the STAC implementation of the data model.
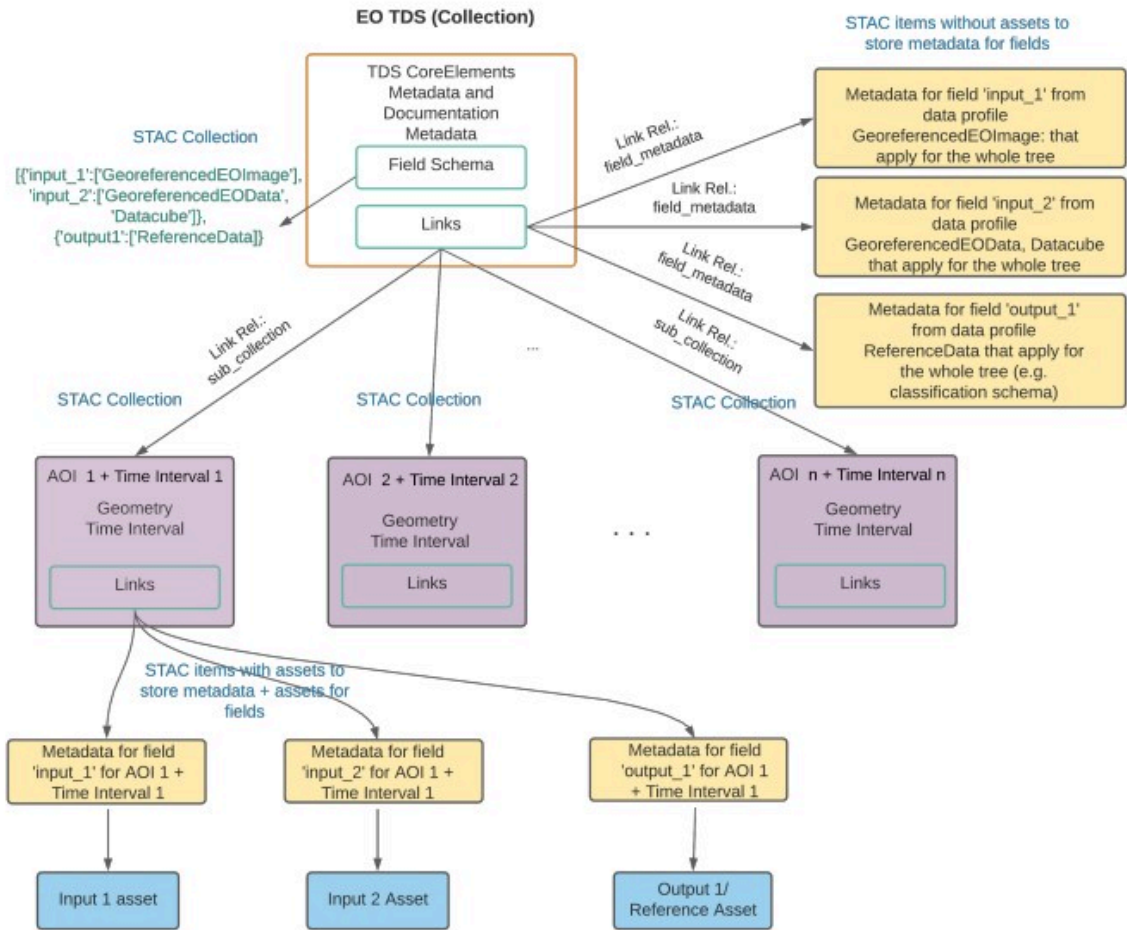
Figure 4 — STAC implementation of the AIREO data model.

There are also the AIREO Best Practice Guidelines that outline how to generate and document AIREO-compliant datasets following the AIREO specifications.

Separately, the AI4EO initiative supports the aims of ESA's Φ-lab that are to accelerate the future of EO by means of transformational innovations. AI4EO hosts challenges that bring AI and EO together.

6.4. ANZLIC considerations of TDSs as foundational data

In the geospatial domain, the United Nations Committee of Experts on Global Geospatial Information Management (the UNGGIM) has identified 14 Global Fundamental Geospatial Data Themes. These data themes are considered fundamental to support global initiatives, such as reporting on the Sustainable Development Goals. These themes are being adopted by nations across the world and are driving the need to create and maintain these datasets in an efficient and effective manner. In Australia, the Australian and New Zealand Land Information Council (ANZLIC) has adopted these themes and challenged jurisdictions to develop and maintain these

data. The ANZLIC community is considering the role of ML in this activity (as described in Clause 5.3.1) and also how the TDSs themselves may be considered as a critical building block, and perhaps may in future be recognized as "foundational," leading to the consideration of whether TDSs might in future become a recognized and authoritative data product.

6.5. Public TDS repositories

6.5.1. Kaggle

The TDS stored on Kaggle are in CSV, JSON, SQLite, and BigQuery as well as other formats.

When the word "satellite" is used to filter the datasets to find those that are most relevant, then CSV, JSON, and "other" are the predominant formats. NASA provides data in CSV, JSON, and NetCDF. "Other" includes a variety of image file formats, such as GeoTIFF, PNG, and JPG as well as Shapefiles. The commercial operator, Satellite Vu, has provided a wildfire dataset in the Tensor tfRecords binary format with TD stored in features.

Kaggle has a usability rating index that is a single number, maximum of 10. This number is used to rate how easy-to-use a dataset is based on a number of factors, including level of documentation, availability of related public content like kernels as references, file types, and coverage of key metadata.

6.5.2. LuoJiaSET, Wuhan University

LuoJiaSET is an Open AI/ML Training Data Hub, with datasets collated from several sources including competitions, review articles and papers with code, Kaggle, blogs, and GitHub. LuoJiaSET is a draft TrainingDML-AI API implementation.

6.5.3. Radiant Earth Foundation & Radiant MLHub

Radiant Earth Foundation is focused on applying ML for EO to meet the UN Sustainable Development Goals and is developing the ML Model Extension to STAC.

The Radiant MLHub hosts open ML TDSs and models generated by Radiant Earth Foundation, partners, and community. A Python client allows users to search and download TDSs. Users may also use other scripting languages and the REST Application Programming Interface (API).

There are several online TDSs focused on applications such as building detection, crop classification, flooding, land cover, and marine debris.

6.5.4. SpaceNet

SpaceNet is a nonprofit organization founded in 2016 by IQT Labs' CosmiQ Works and Maxar to accelerate open source geospatial ML. They run data challenges and release the TDSs, baseline algorithms, winning algorithms, and detailed evaluations under an open-source license.

As of 2020, the Radiant Earth Foundation announced the registration of a STAC-compliant version of SpaceNet's high-quality geospatial labeled datasets for roads and buildings on Radiant MLHub. The broader SpaceNet Dataset is hosted as an Amazon Web Services (AWS) Public Dataset.

6.5.5. Zenodo

Zenodo was originally developed by the European Organization for Nuclear Research (CERN) as part of an EC project to support Open Data. The goal was to be a catch-all repository for EC funded research. Through various sources of funding, CERN makes Zenodo publicly available. Advantages of using Zenodo are that DOIs are created and Zenodo automatically maintains a list of uses and citations.

Zenodo contains many ML TDS and users uploading data may choose the format of what is being uploaded. One example is the The WorldStrat Dataset that includes open high-resolution satellite imagery from Airbus supplied SPOT 6/7(1.5 m spatial resolution) paired with multi-temporal low-resolution satellite imagery from Sentinel-2 (10 m spatial resolution). The metadata are stored in a CSV file within the datasets, which are held in TAR gzipped files. As the WorldStrat creators wanted to lower the barrier to entry, the dataset and PyTorch DataLoader are provided in a format most accessible to the ML community. The code is also open-source and available on GitHub.

6.6. Previous OGC activities

6.6.1. Testbed-16

The OGC Testbed-16 Machine Learning (ML) task focused on understanding the potential of existing and emerging OGC standards for supporting ML applications in the context of wildland fire safety and response. Relevant recommendations for this broader activity included the following.

- Make sure that datetimes are properly and accurately set in datasets.

- Provide accuracy information in the metadata of each training dataset.

- Establish a standard way to store and reuse a model.

As future work, the following was suggested.

- There is a real need to work out a best practice for a generalizable metadata model (framework) for ML TDSs. The Key Elements for Metadata Content section contains several items that could form a basis of this best practice in the future.

- Furthermore, Earth Observation (EO) datasets should be rendered AI-ready as described, for example, by the aireo.net reference to this topic. Also in this context, efforts towards Analysis Ready Data (ARD) such as those proposed by the Committee on Earth Observation Satellites (CEOS) will likely become vital for future ML applications.

- Solid and reliable ground truth datasets should be developed, including accuracy levels of the ML training data.

6.6.1.1. Summary of the Testbed-16 Metadata Content Section

The main points in the Testbed-16 ER metadata content section are as follows.

- Metadata should at least contain statistical information about the data set, such as source, size, dimension, license, update status, and other elements, as well as of course features.

- Creating and generating metadata for ML or research data and datasets in the ML training data "lifecycle" preserves the data in the long run and will also facilitate the use of ML training data for non-experts.

- The reader is also referred to the CDB SWG research on metadata standards and common mandatory elements across standards.

A set of rules and recommendations from OGC Testbed-16 are as follows. (Source: DMPTool. Digital Curation: A How-To-Do-It Manual; Digital Curation Centre).

- Consider what information is needed for the data to be read and interpreted in the future.

- Understand requirements for data documentation and metadata. Several instructive examples can be found under the Funder Requirements section of the Data Management Plan Tool (DMPTool).

- Consult available metadata standards for the domain of interest. Refer to Common Metadata Standards and Domain Specific Metadata Standards for details.

- Describe data and datasets created in the research lifecycle, and use software programs and tools to assist in data documentation. Assign or capture administrative, descriptive,

technical, structural, and preservation metadata for the data. Some potential information to document includes the following.

- Descriptive metadata

  - Name of creator of data set

  - Name of author of document

  - Title of document

  - File name

  - Location of file

  - Size of file

- Structural metadata

  - File relationships (e.g., child, parent)

- Technical metadata

  - Format (e.g., text, SPSS, Stata, Excel, tiff, mpeg, 3D, Java, FITS, CIF)

  - Compression or encoding algorithms

  - Encryption and decryption keys

  - Software (including release number) used to create or update the data

  - Hardware on which the data were created

  - Operating systems in which the data were created

  - Application software in which the data were created

- Administrative metadata

  - Information about data creation (e.g., date)

  - Information about subsequent updates, transformation, versioning, summarization

  - Descriptions of migration and replication

  - Information about other events that have affected the files

- Preservation metadata

  - File format (e.g., .txt, .pdf, .doc, .rtf, .xls, .xml, .spv, .jpg, .fits)

  - Significant properties

- Technical environment

- Fixity information

- Adopt a thesaurus in the relevant field [i.e., common terminology] or compile a data dictionary for the dataset.

- Obtain persistent identifiers (e.g., DOI) for datasets, if possible, to ensure data can be found in the future.

6.6.2. Testbed-15

The Testbed-15 activity explored the ability of ML to interact with and use OGC Web Service Standards (OWS) in the context of natural resources applications; including WPS, WFS, and CSW.

The work exercised OGC standards using five different scenarios incorporating use cases that included traditional ML techniques for image recognition; understanding the linkages between different terms to identify a dataset; and vectorization of identified water bodies using satellite imagery. The Testbed-15 Engineering Report noted that the web service-based standards would soon be complemented by OGC API Standards based on OpenAPI descriptions and RESTful principals.

The Testbed recommendations were primarily linked to the OGC standards. However, the ER noted that even if the source code used to implement predictive models is kept static, the behavior of the models can change due to the varying availability and constant evolution of their training data. This affects the reliability of models and reproducibility of the experiments, which is a cornerstone of scientific research. Keeping track of changes in data is not an easy task, as Version Control Systems (VCS) are not typically made to track large binary files and solutions for small projects are often limited to locally hosted datasets that are not frequently updated.

Adding rigorous metadata fields related to data sources and modification times to standardized web service requests were seen as greatly improving the robustness of ML training and evaluation services.

6.6.3. Testbed-14

The Testbed-14 ML activity also focused on how to support and integrate emerging AI and ML tools using OWS, as well as publishing their input and outputs. A proof-of-concept client application executed processes offered by the ML system and displayed its results found in an Image and Feature Repository.

7

METADATA REQUIREMENTS AND RECOMMENDATIONS

7 METADATA REQUIREMENTS AND RECOMMENDATIONS

Metadata are crucial for ensuring lossless data interchange and their appropriate use. Metadata can be created automatically during data capture (e.g., timestamps of a data record, or an automatic label of data production software), or added before advertising the data object to provide context for understanding the creation of a dataset (e.g., through detailed description of dataset's provenance information).

7.1. Current structure and usage of metadata in ML TDS

As outlined in Clause 6, most current ML TDS models use the STAC family of specifications as the basis to structure the TDS and related metadata. The STAC specification defines only a limited set of 'STAC Core Metadata' elements used for STAC Catalog 'Collection' and STAC Catalog 'Item'. The following core metadata elements are required.

- Basic metadata to provide an overview of a STAC Item

    - title

    - description

- Date and Time definition

    - datetime, created and updated — to allow recording of temporal capture via information about

    - start_datetime and end_datetime — to allow specification of ranges of capture datetimes

- License information for data and metadata

- Provider information — to allow defining information about provider (e.g., name, description, and url) and their roles (e.g., processor, producer, licensor, host)

- Instrument information — to allow specifying the information about platform, instrument, mission, constellation and ground sampling distance used for data acquisition

Given its modular nature, the STAC specification allows enhancing the metadata definition of STAC objects through extensions. One of the stable STAC extensions recommended for definition of ML items and collections (see Clause 6) is the 'Scientific Citation Extension'. This extends STAC core metadata elements with reference information about which publication a STAC object originates and how it should be cited or referenced. Additional scientific citation

metadata, such as the Digital Object Identifier (DOI), citation, and indication of relevant publications help to increase reproducibility and findability of a STAC object, and thus improving its FAIRness (more detail in Clause 10).

7.2. Review and application of ISO metadata standards for ML TDS

This section discusses four key ISO metadata standards: ISO 19115-1, ISO 19115-2, ISO 19157-1, and ISO 19157-3.

7.2.1. ISO 19115-1 and ISO 19115-2 for geographic information

ISO 19115-1:2014 Geographic information — Metadata — Part 1: Fundamentals

ISO 19115-1:2014 defines the schema required for metadata about geographic datasets and services. The standard defines the structure for information about data and metadata identification, spatial and temporal extent, quality, distribution, and licenses. This standard is applicable to the definition of metadata catalogs (typically used in a Spatial Data Infrastructure — SDI) as well as for describing geographic resources of various kinds (i.e., datasets or services, maps, charts, or textual documents about geographic resources) and at various levels of detail (e.g., dataset, feature, or attribute). Figure 5 illustrates the metadata schema defined in ISO 19115-1.

Figure 5 — ISO 19115-1 Metadata schema.

ISO 19115-1:2014 also identifies the minimum metadata set required to serve most metadata applications, including data discovery, access, transfer, and use, and a decision on dataset's fitness for use (see Table 1).

Table 1 — Metadata for the discovery of geographic datasets

METADATA ELEMENT	OBLIGATION	COMMENT
Metadata reference information	Optional	Unique identifier for the metadata
Resource title	Mandatory	Title by which the resource is known
Resource reference date	Optional	A date which is used to help identify the resource
Resource identifier	Optional	Unique identifier for the resource

METADATA ELEMENT	OBLIGATION	COMMENT
Resource point of contact	Optional	Name, affiliation, and role of the person responsible for the resource
Geographic location	Conditional	Geographic coordinates or description of metadata location — Mandatory if the described resource is not a 'dataset'.
Resource language	Conditional	Language used to describe the resource — Mandatory if other than default (English).
Resource topic category	Conditional	A selection from the list of topics defined in ISO 19115-1 — Mandatory if the described resource is not a 'dataset' or a 'dataset series'.
Spatial resolution	Optional	The nominal scale and/or spatial resolution of the resource
Resource type	Conditional	ISO 19115-1 standard code (e.g., dataset, feature, attribute, product) the metadata describe — Mandatory if the described resource is not a 'dataset'.
Resource abstract	Mandatory	A brief description of the content of the resource
Extent information for the dataset	Optional	Temporal or vertical extent of the resource
Resource lineage/provenance	Optional	Source and production steps used in producing the resource.
Resource on-line link	Optional	URL for the resource.
Keywords	Optional	Words and phrases describing the resource to be indexed and searched.
Constraint on the resource access and use	Optional	Restrictions on the access and use of the resource.
Metadata date stamp	Mandatory	Reference date for the creation (and update) of metadata
Metadata point of contact	Mandatory	The party responsible for the metadata.

Note in Table 1 that if a described resource is a 'dataset' or a 'dataset series' ISO 19115-1 mandates only four metadata elements to describe a such resource.

1. Resource title

2. Resource abstract

3. Metadata date stamp

4. Metadata point of contact

Arguably, this is insufficient to ensure findability, accessibility, interoperability, and reuse of a resource, especially by machines, which is often the case in ML.

ISO 19115-2:2019 Geographic information — Metadata — Part 2: Extensions for acquisition and processing

ISO 19115-2:2019 extends ISO 19115-1:2014 by defining the schema required for describing the acquisition and processing of geographic information, including imagery. This standard defines the structure for describing properties of measuring systems and the numerical methods and computational procedures used to derive geographic information from the acquired data. Figure 6 illustrates the metadata schema defined in ISO 19115-2.
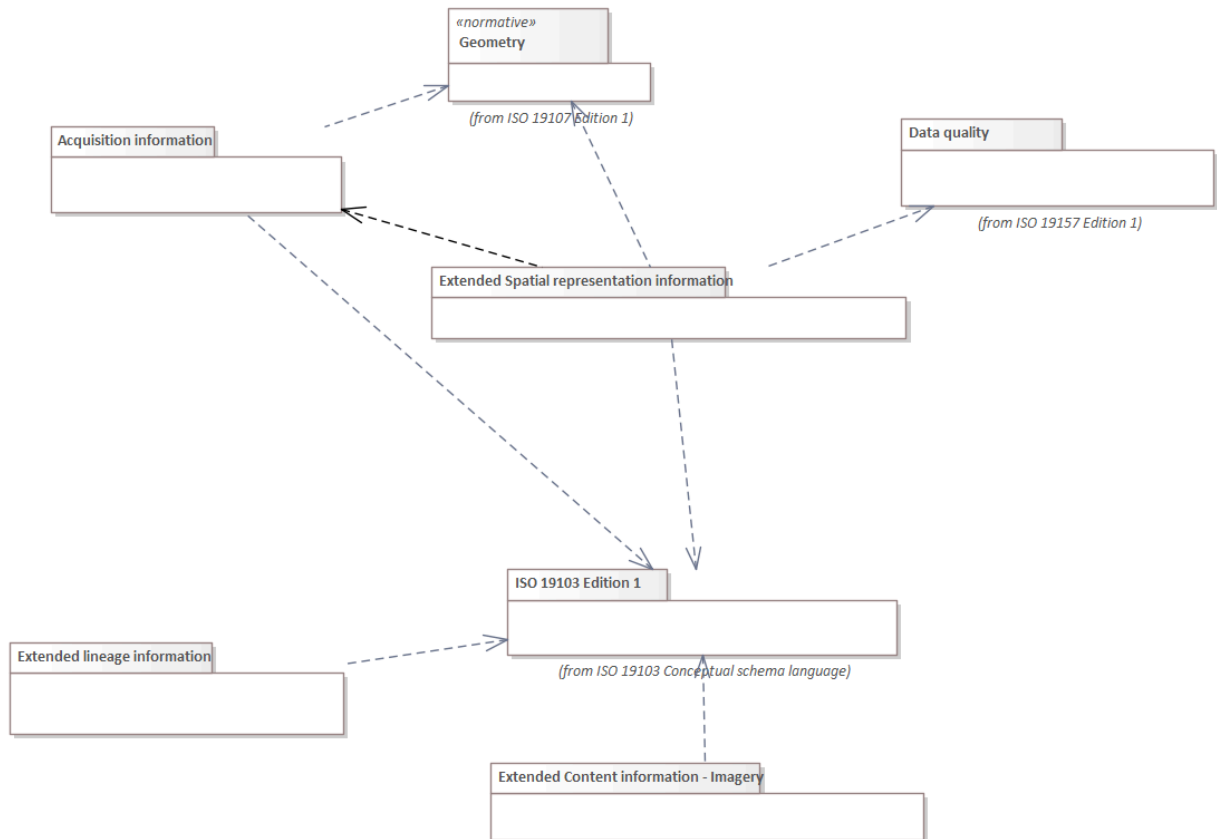
Figure 6 — ISO 19115-2 Metadata schema.

ISO 19115-1 and ISO 19115-2 are intentionally generic. Although they are applicable in most geospatial domains, some specific disciplines may need extensions or modifications to the model. ISO 19115-1:2014 allows the creation of metadata extensions. The following are the permitted types of extensions in the current metadata standard.

1. Adding a new metadata package

2. Creating new metadata codelist elements (expanding a codelist)

3. Adding new metadata elements

4. Adding new metadata classes

5.   Imposing a more stringent obligation on an existing metadata element

6.   Imposing a more restrictive domain on an existing metadata element

An extension mechanism defined in item five in the list above would ensure more comprehensive discovery metadata, and thus help increase the FAIRness of a resource.

ISO 19157-1 Geographic information — Data quality — Part 1: Fundamentals

In the 19100 series of standards, data quality information is considered to be a specialized type of metadata about the quality of a geographic information resource. ISO 19157-1 provides a framework for defining the quality of geographic data. This includes principles for evaluating quality, a conceptual model for handling quality information, a structure and content of data quality measures, and guidelines for reporting a quality evaluation. The framework is extensible, with rules for how to add additional data quality measures including a provision for complex dimensions of data quality. Figure 7 illustrates the data quality model defined in ISO 19157-1.
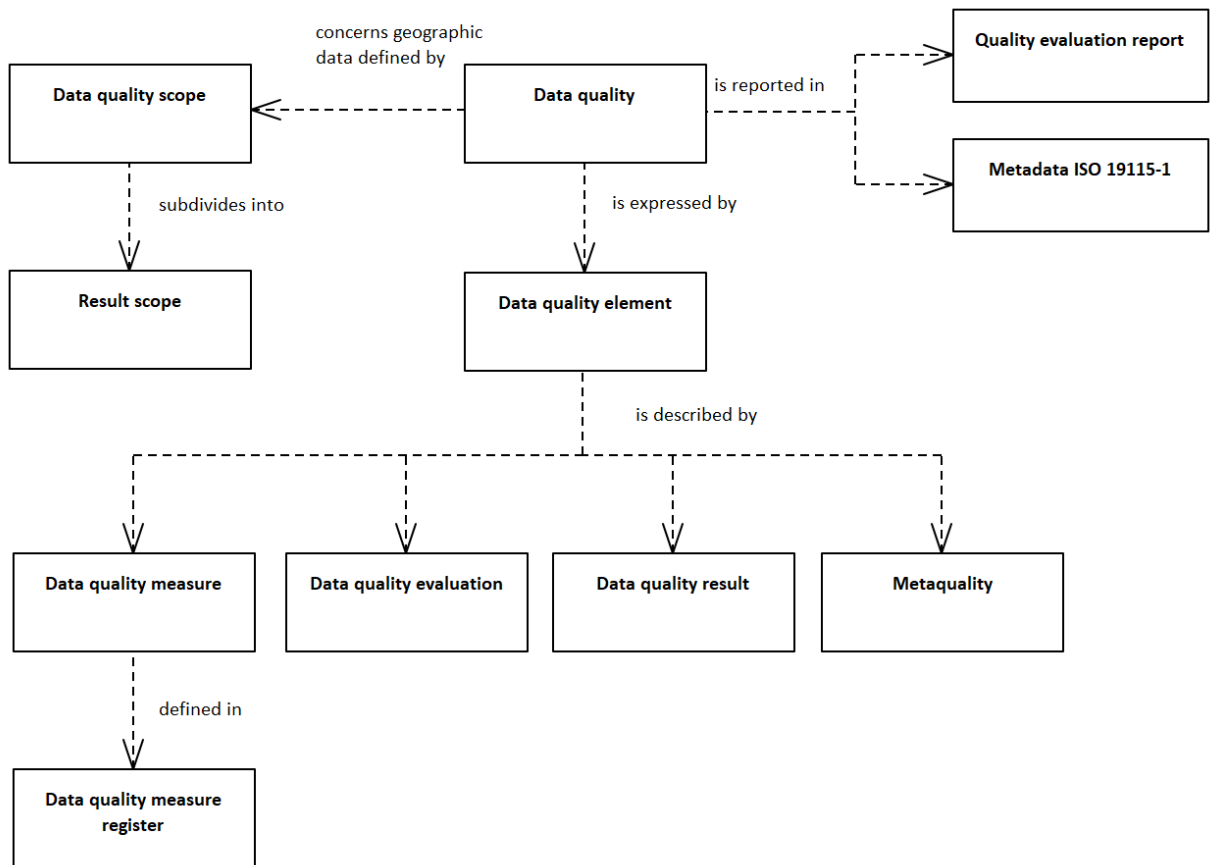
Figure 7 — ISO 19157-1 Conceptual model of quality for geographic data.

According to the standard, working with data quality includes:

- understanding the concepts of data quality related to geographic data;

- defining data quality conformance levels in data product specifications or based on user requirements;

- evaluating data quality and metaquality; and

- reporting data quality and metaquality.

Where metaquality is a way of describing the quality of the data quality evaluation.

A data quality evaluation can be applied at various scopes (see Figure 11) to dataset series, a dataset, or a subset of data within a dataset, sharing common characteristics so that its quality can be evaluated. Data quality elements and their descriptors are used to describe how well a dataset meets the criteria set forth in its data product specification or user requirements and provide quantitative quality information.

Data quality elements and their descriptors are used to describe how well a dataset meets the criteria set forth in its data product specification or user requirements and provide quantitative quality information. ISO 19157-1 defines the following data quality elements.

- **positional accuracy**: Closeness of agreement between a measured position of features and a position accepted as true within a spatial reference system. The following are the types of positional accuracy.

  - absolute positional accuracy

  - relative positional accuracy

  - gridded data positional accuracy

- **thematic accuracy**: The accuracy of quantitative attributes and the correctness of non-quantitative attributes and of the classifications of features and their relationships. Depending on the type of attribute, this can be expressed with the following.

  - classification correctness

  - non-quantitative attribute correctness

  - quantitative attribute correctness

- **completeness**: The presence or absence of features, attributes, or relationships in a resource and can be expressed by the following.

  - omission

  - commission

- **logical consistency**: The degree of adherence to logical rules of data structure, attribution, and relationships as follows.

    - conceptual logical consistency

    - domain consistency

    - format consistency

    - **topological consistency**

- **temporal quality**: The quality of the temporal attributes and temporal relationships of features. The following aspects can be used to express temporal accuracy.

    - accuracy of time measurement

    - temporal consistency

    - temporal validity

In addition to the elements describing the quality of a geographic resource, the standard defines a way of describing the quality of the quality evaluation, i.e., **metaquality** of a resource. Confidence, representativity, and homogeneity are example elements suggested in the standard.

ISO 19157-1 recognizes that for many domain-specific purposes it is necessary or convenient to extend the standard data quality information model as defined in the standard. An extension includes adding data quality elements (e.g., 'class balance degree') and data quality measures (e.g., 'number of imbalanced classes in the training dataset').

ISO 19157-3 Geographic information — Data quality — Part 1: Data quality measures register

To facilitate comparisons of datasets expressing the quality in a comparable way and having a common understanding of the data quality measures that have been used is essential. These data quality measures provide descriptors of the quality of geographic data through comparison with the universe of discourse. ISO 19157-1 standardizes the components and structures of data quality measures, and ISO 19157-3 (currently under development) establishes a machine-actionable data quality measures register. An ISO 19157-3 compliant register will contain an extensible curated set of data quality measures. The structure of a data quality measure (as defined in ISO 19157-1) is illustrated in Figure 8 and an example is in [igure-DQmeasureExample].
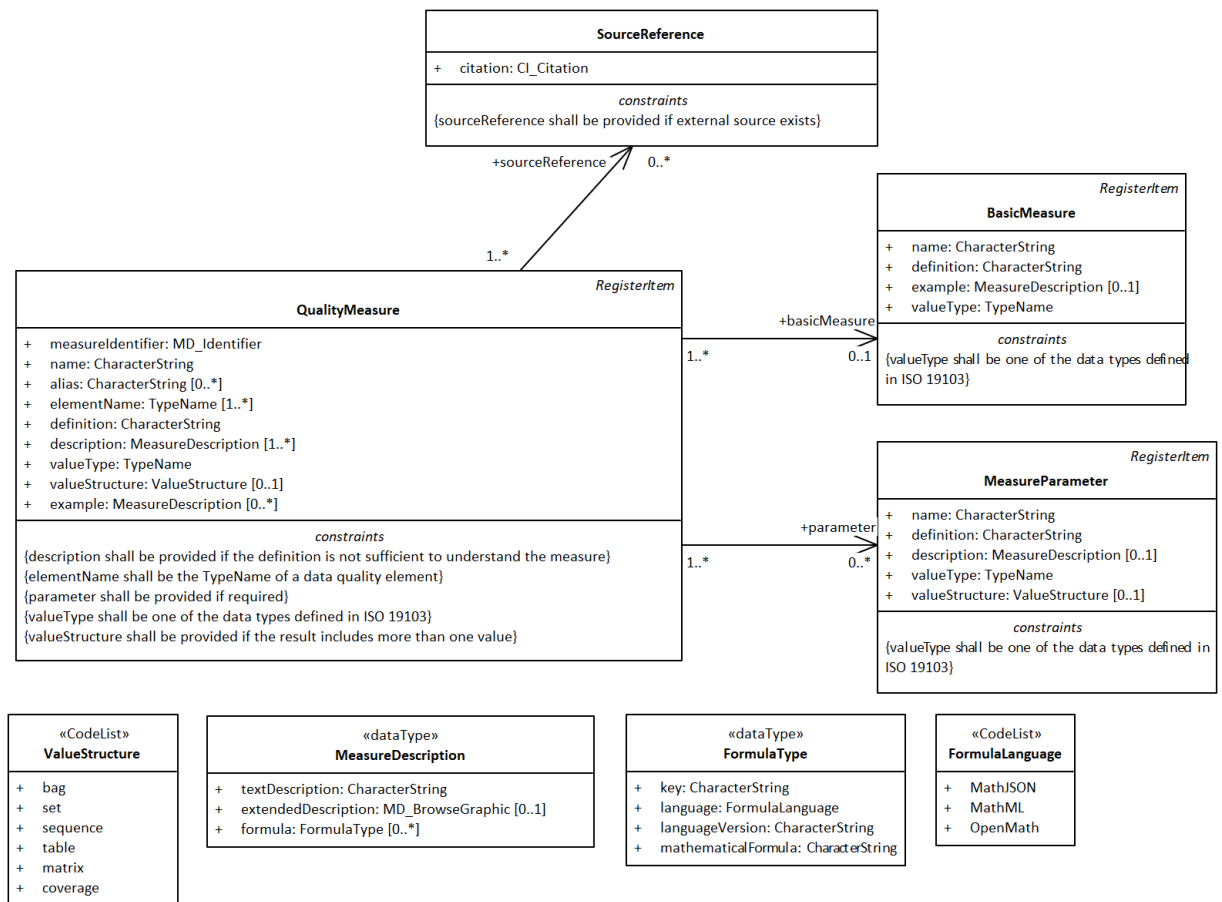
Figure 8 — Data quality measure structure as defined in ISO 19157-1.

Line	Component	Description
1	Name	misclassification matrix
2	Alias	confusion matrix
3	Element name	classification correctness
4	Basic measure	-
5	Definition	matrix that indicates the number of items of class (i) classified as class (j)
6	Description	The misclassification matrix (MCM) is a quadratic matrix with n columns and n rows. n denotes the number of classes under consideration. MCM (i,j) = [# items of class (i) classified as class (j)] The diagonal elements of the misclassification matrix contain the correctly classified items, and the off diagonal elements contain the number of misclassification errors.
7	Parameter	Name: n Definition: number of classes under consideration Value Type: Integer
8	Value type	Integer
9	Value structure	Matrix ($n \times n$)
10	Source reference	-
11	Example	
12	Identifier	62

Figure 9 — Example of a data quality measure.

7.2.2. Metadata reporting at various granularity levels

ISO 19115-1:2014, ISO 19115-2:2019, and ISO 19157-1 are applicable at various levels of detail. These are defined by the MD_Scope property in the metadata information (MD_Metadata) and of the resource (see Figure 10).

Figure 10 — ISO 19115-1 Metatadata on Metadata.

The default scope for metadata is the 'dataset', but the domain of the scope is not restricted to the dataset. The MD_ScopeCode (see Figure 11) defines an extensible list of possible types of resources. The choice of suitable type is left to the data provider.
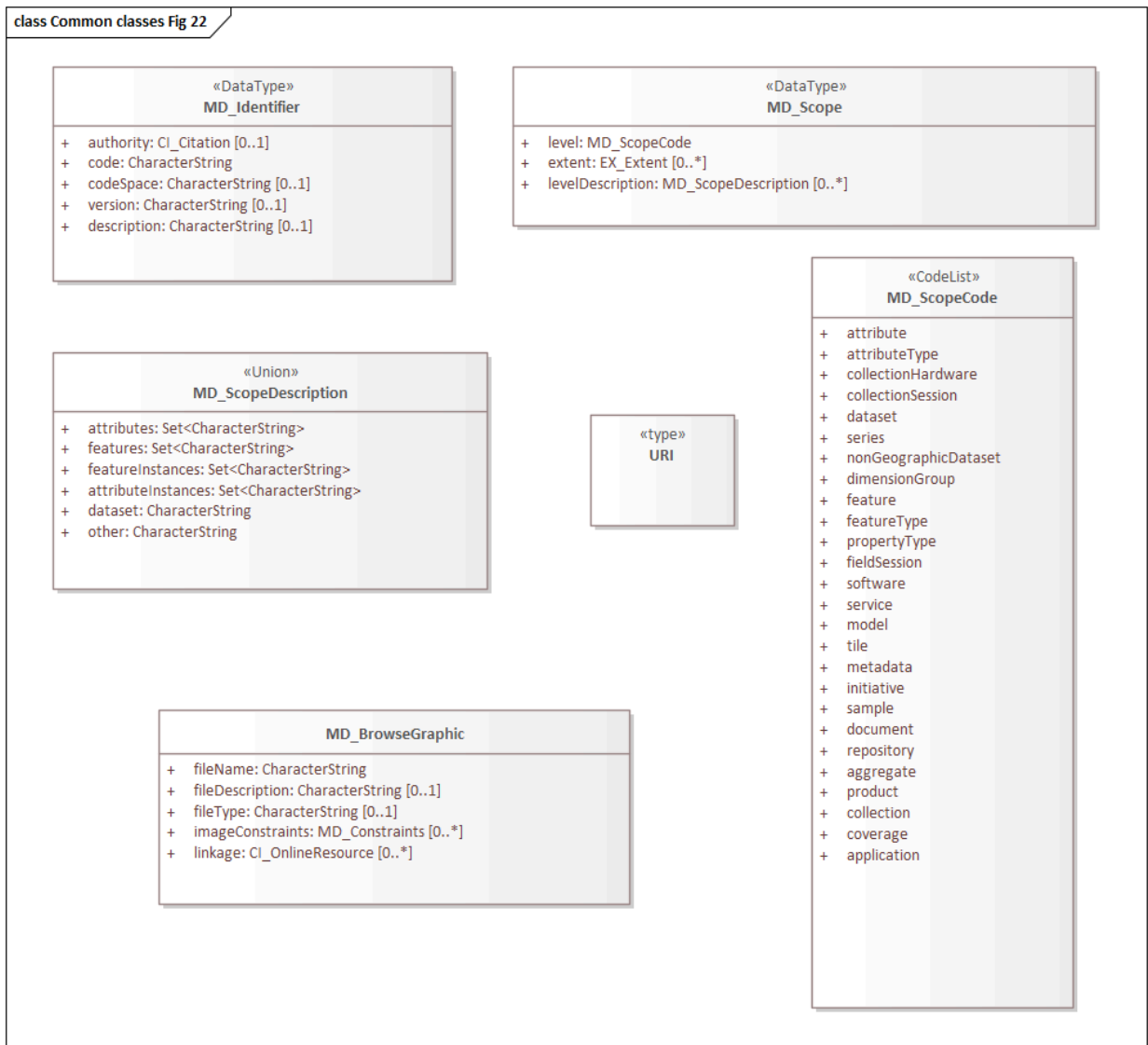
Figure 11 — ISO 19115-1 Metatadata scope types

7.2.3. Applicability of ISO 19115-1 and ISO 19115-2 for ML TDS

From the ISO 19115-1 and ISO 19115-2 standards' perspective, a TDS for an ML application is 'just a dataset' and thus both standards are applicable for the definition of TDS metadata. When looking at the current proposal of OGC specification for TrainingDML-AI, there is a need for more precise mapping between ISO 19115 scope terminology and the AI_TrainingDataset conceptual model — see Table 2.

Table 2 — Mapping ISO 19115-1 scope codes into Draft TrainingDML-AI concepts

ISO 19115-1 SCOPE CODE	TRAININGDML-AI CONCEPT
metadata	AI_AbstractTDQuality (and all child elements)
	AI_TDSChangeset
	AI_Task
	AI_Labelling
dataset/dataset series	AI_TrainingDataset
feature	AI_TrainingData
attribute	AI_Label (and all child elements)

7.3. Examples of human and machine-readable metadata for a TDS

Producers use metadata to advertise their resources. This is typically done through metadata catalogs or via embedded metadata in the resource that can be indexed by search engines. Metadata also serve as the essential vehicle for users' (humans or machines) decision on resources' fitness for use. Each type of user requires their own way of metadata expression, i.e., human readable (as in the example in Figure 12) or machine readable (as in the example in Figure 13).

Figure 12 — Human readable metadata documentation

```xml
                  ▼<gmd:topicCategory>
                       <gmd:MD_TopicCategoryCode>geoscientificInformation</gmd:MD_TopicCategoryCode>
                  </gmd:topicCategory>
              </gmd:MD_DataIdentification>
          </gmd:identificationInfo>
      ▼<gmd:contentInfo>
        ▼<gmd:MD_CoverageDescription id="col0.ds14152625">
          ▼<gmd:attributeDescription>
                <gco:RecordType>File content</gco:RecordType>
            </gmd:attributeDescription>
          ▼<gmd:contentType>
                <gmd:MD_CoverageContentTypeCode
                codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#MD_CoverageContentTypeCode"
                codeListValue="thematicClassification">thematicClassification</gmd:MD_CoverageContentTypeCode>
            </gmd:contentType>
          </gmd:MD_CoverageDescription>
      </gmd:contentInfo>
      ▼<gmd:contentInfo>
        ▼<gmd:MD_CoverageDescription id="col1.ds14152626">
          ▼<gmd:attributeDescription>
                <gco:RecordType>Binary Object (File Size)</gco:RecordType>
            </gmd:attributeDescription>
          ▼<gmd:contentType>
                <gmd:MD_CoverageContentTypeCode
                codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#MD_CoverageContentTypeCode"
                codeListValue="physicalMeasurement">physicalMeasurement</gmd:MD_CoverageContentTypeCode>
            </gmd:contentType>
          ▼<gmd:dimension>
            ▼<gmd:MD_Band id="col1.ds14152626.band">
              ▼<gmd:units>
                ▼<gml:UnitDefinition gml:id="col1.ds14152626.band.unit">
                     <gml:identifier codeSpace="PANGAEA">Bytes</gml:identifier>
                     <gml:name>Bytes</gml:name>
                  </gml:UnitDefinition>
                </gmd:units>
              </gmd:MD_Band>
            </gmd:dimension>
          </gmd:MD_CoverageDescription>
      </gmd:contentInfo>
      ▼<gmd:contentInfo>
        ▼<gmd:MD_CoverageDescription id="col2.ds14152626">
          ▼<gmd:attributeDescription>
                <gco:RecordType>Binary Object</gco:RecordType>
            </gmd:attributeDescription>
          ▼<gmd:contentType>
                <gmd:MD_CoverageContentTypeCode
                codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#MD_CoverageContentTypeCode"
                codeListValue="thematicClassification">thematicClassification</gmd:MD_CoverageContentTypeCode>
            </gmd:contentType>
          </gmd:MD_CoverageDescription>
      </gmd:contentInfo>
```
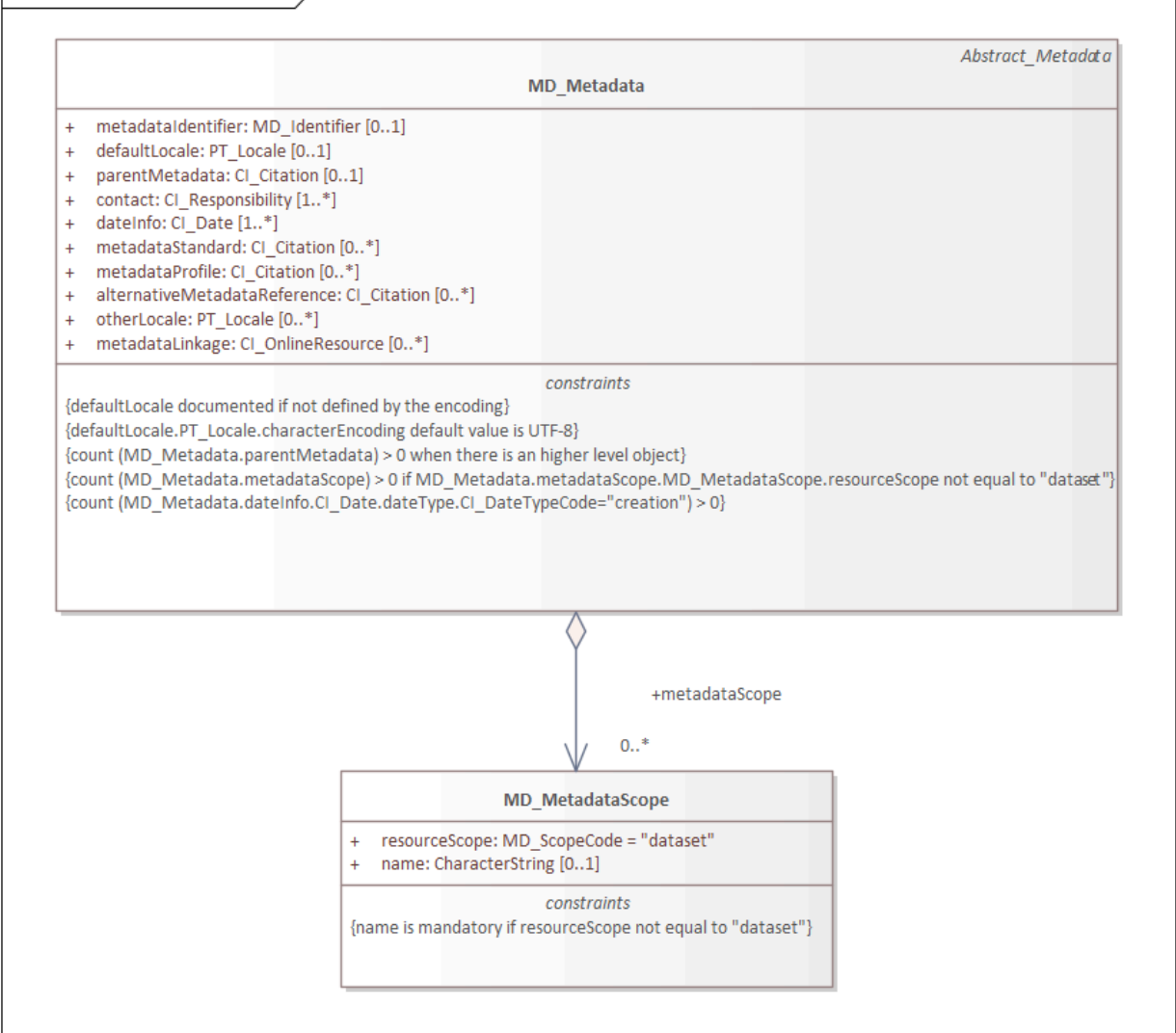
Figure 13 — Machine readable metadata documentation

8

TDS CATALOGS

---

8 TDS CATALOGS

This section discusses key features that must be considered for how Machine Learning (ML) Training Datasets (TDS) are cataloged.

8.1. What is a catalog?

In an OGC White Paper on standards and cloud computing McKee et al. 2011, the cloud is described as based on a standards framework for service-oriented architectures that provides a "publish, find, bind" model.

- Publish: Resources can be hosted and their description, network location, and interfaces can be published in standards-based registries or catalogs.

- Find: Client applications can search the registries or catalogs to find a resource.

- Bind: The client application can invoke the server through standard interfaces.

As a result, catalogs are important for both publishing and finding information.

ISO 19115-1 and ISO 19115-2 are applicable to the metadata for cataloging geographic services/datasets as previously detailed in Clause 7.

OGC-API Records provide discovery and access to metadata about geospatial resources (e.g., data, services, ML models, etc.), and has three main building blocks: record, collection, and records Application Programming Interface (API) as a web interface. A record provides a description (i.e., metadata) about a resource that the provider of the resource wishes to make discoverable. A collection is used to describe a collection of resources. There are several ways that records can be deployed as a "collection of records" or a catalog. Three deployment patterns are envisioned as follows.

1. A catalog deployed as a crawlable collection of records

2. A catalog deployed as searchable endpoint(s)

3. A local resources catalog

For a SpatioTemporal Asset Catalog (STAC), the terms static and dynamic are used to describe these deployment patterns. However, a static catalog is not really static since additional records can be added at any time. Therefore, the terms crawlable and searchable are proposed for OGC-API Records.

8.2. Version control for TDS

The Data on the Web Best Practices (DWBP) states that datasets published on the Web may change over time and describes example scenarios. Even for small changes, it is important to keep track of the different dataset versions to make the dataset trustworthy. Publishers should remember that a given dataset may be in use by one or more data consumers. Therefore they should take reasonable steps to inform those consumers when a new version is released. Also, the publisher should take a consistent, informative approach to versioning, so data consumers can understand and work with the changing data.

For TDSs, questions include the following.

- In the context of maintaining fundamental data, new imagery sources may be added, and a trained model would need to be exposed to this imagery. Training data collection will expand, or may be updated. How is this tracked?

- Reprocessing of the used imagery: Many satellite missions have their data reprocessed, which creates a consistent dataset. When a TDS is made available, users may be downloading older versions of extracted data, which have anomalies that have since been corrected. Alternatively, if new versions of such data are download the ML model may not work as expected, e.g., if there is a change to the radiometric calibration of how the data is stored as has occurred with the Copernicus Sentinel-2.

- Temporal element: Training data is tied to imagery from a given date. What happens when the underlying landscape changes? Training data cannot necessarily be used on new imagery if the landscape has changed.

- Research papers would need to state which version of a training data set were used for future reproducibility of results.

For the Training Data Markup Language for Artificial Intelligence (TrainingDML-AI) standard, and as shown in Figure 1, there is a Changeset module that defines changes in the Training Data (TD) between different versions. Three kinds changes to training samples are included: add, modify, and delete.

Public repositories, such as Zenodo (see Clause 6) provide a means to archive dataset versions with version-specific Digital Object Identifiers (DOIs). However, EO data is often processed, including sub-setting, so will not be in its original form and hence will probably have lost its original metadata. Therefore, keeping a record of the input dataset versions can be complicated as datasets may not be consistently processed. For example, Copernicus Sentinel-2 is a frequently used dataset where the processor has been updated during the mission lifetime and the archive has not currently (as of August 2022) been reprocessed consistently. Therefore, imagery from different dates will have different corrections applied that may influence the ML inputs, and hence, model.

8.3. Splitting source data and annotated training data

Source data may change, either intentionally or unintentionally. Therefore, in storing a TDS, there is a question about what should be contained versus referenced by the catalog.

For example, if the underlying imagery is not provided, only a query for it or a description of how it was generated is? Is this the best approach, or should the extracted source data be included as well as any derived products that act as inputs to the ML model?

8.4. Making TDS catalogs self-explanatory

"Self-explanatory" should equate to sufficient metadata within the catalog, including onward web links, that enables a user to answer questions they might have. Difficulties can occur as time passes because web links become out-of-date/broken and the information becomes lost — a particular problem with online storage where storage is reliant on individuals/organizations paying for storage. Another example would be a disaster response scenario where users do not have access to the internet, or setups where the TDS is distributed to a disconnected environment.

The AIREO Best Practice Guidelines encourages data creators to describe the metadata and documentation in the catalog richly to enable attribute-based searches and facilitate data findability. The AIREO Best Practice Guidelines also discusses the need for eternally persistent identifiers, with the metadata containing machine-processable and verified elements for citation and version control.

Gebru et al. 2021 proposed datasheets for datasets to support a perceived gap in the standardized process for documenting ML TDS, which was applied by the WorldStrat TDS. The datasheet is as an appendix to the accompanying paper, and part of the accompanying documentation within the WorldStrat Zenodo archive. It contains more detailed information than is potentially feasible to store in the metadata alone.

9

TDS QUALITY

9 TDS QUALITY

The ability of a machine learning process to correctly classify data is highly dependent on the quality of examples it has been exposed to during training. Labelled training data is created by having an annotator (which could be a person or a process) assign a label to data. This could be through a field survey, noting the crop type present at a specific location as in Clause 5.3.4, or by the delineation of features directly from imagery as in Clause 5.3.2. For machine learning training data, there are two key factors that influence the quality: accuracy and consistency.

- Accuracy describes how well the training data labels reflect reality. This captures both the accuracy of individual training data labels as well as whether the validation and testing data sets have distributions that reflect reality. In Clause 5.3.1, the goal was to capture tree canopy, where a tree was defined as vegetation taller than two meters. Human annotators relied on the presence of shadow and vegetation texture to distinguish between trees and shrubs but, without explicit LiDAR height data, some shrubs were labelled as tree cover. This confusion on the part of the annotator lowers the accuracy of the training dataset, and may impact the performance of a model trained on this dataset.

- Consistency describes how frequently different annotators agree on the assigned label. Low consistency may indicate that the classes that need to be labelled do not have sufficiently clear definitions for annotators to consistently agree on the best classification. In the case of delineation of vector features from imagery, such as the identification of roofprints from aerial imagery in Clause 5.3.2, consistency could also describe how well-matched different annotators' delineations of roofprints are.

For each factor, there are several elements to be considered.

- Accuracy

  - Inclusion of a benchmark set, with examples that have been delineated or created by a subject matter expert

  - Statistical distribution of classes measured or expected in reality

  - Spatial distribution of classes measured or expected in reality

  - Description of methods used to select the training data instances, including any stratification criteria and resulting proportion of labels for each class

  - Identification of outliers and duplicates

  - Relative number of items in each class, indicating class balance

- Consistency

    - Inclusion of a labelling guide that demonstrates how to identify each class and any useful information for disambiguating classes

    - Number of times each training data instance was labelled by independent annotators

    - Ratio of training data instances with consistent labels to all training data instances

9.1. Biases and domains in TDS

The domain in which training data is created has significant influence on whether it can be reused. For example, the training dataset of roofprints created for Clause 5.3.2 was created for urban areas in the state of Victoria, Australia. Whether it can be reused in a new domain depends on the similarities between the original domain and the new domain. For example, the roofprint dataset might be reusable in other urban settings of Australian cities but might not be suitable to rural areas of Australia, due to differences in size, density, ground cover, or roof material.

When delineating and predicting features from Earth observation data, the year and season may also impact whether the training data can be applied to a new domain. This is particularly true in the case of training datasets that label the presence of vegetation, such as in Clause 5.3.1 and Clause 5.3.4, as vegetation will have different spectral qualities depending on the time of year the associated satellite data was captured. By capturing domain information in training dataset metadata, future users can make informed decisions about whether existing training data can be considered accurate enough for a new application.

Domains for consideration include the following.

- Geographical extent

- Capture date of reference or ground truth data

- Seasonality of reference or ground truth data

Understanding any biases that may be present in the data is important. Bias can be thought of as any process that occurred during training data creation that led to the data being unrepresentative of the real world. For example, when conducting ground surveys such as in Clause 5.3.4, the most efficient collection method may be to work along main roads, but this may mean the distribution of classes in the training data is not representative of the distribution across the whole area of interest. Lack of domain knowledge when producing labels may also lead to biases in the training data, such as through the systematic mislabeling of certain examples due to annotator confusion.

Along with the domain description, training dataset metadata should also describe the method used to generate the training dataset, especially sampling strategies that were used to ensure the training dataset was representative of the real world.

9.2. Auto-generation of quality indicators

Consistency, measured as the ratio of consistently labelled instances to all instances, can be measured automatically, but will require each training data instance to be tagged with each annotator's label.

Accuracy is more challenging and may require a subset of the training data to be identified as the benchmark for assessing correctness. With such a benchmark set, it would be possible to compare the agreement between labels from annotators with those from the benchmark. This would only give a representative idea of the accuracy but may be helpful in assessing the appropriateness of a training dataset for a new application.

ENABLING FAIR IN THE FUTURE TDS STANDARD

10 ENABLING FAIR IN THE FUTURE TDS STANDARD

The FAIR (findable, accessible, interoperable and re-usable) data principles are a valuable framework in assessing the accessibility and usability of resources. This section introduces the principles and discusses how they apply to ML TDS.

10.1. The FAIR guiding principles

According to Wilkinson2016 resources are FAIR when they are findable, accessible, interoperable and reusable. A summary of the FAIR principles are found in Table 3:

Table 3 — The FAIR principles

ASPECT	PRINCIPLES
Findable	F1. (meta)data are assigned a globally unique and persistent identifier
	F2. data are described with rich metadata (defined by the R1 below)
	F3. metadata clearly and explicitly include the identifier of the data it describes
	F4. (meta)data are registered or indexed in searchable resource
Accessible	A1. (meta)data are retrievable by their identifier using a standardized communications protocol
	A1.1. the protocol is free, open and universally implementable
	A1.2. the protocol allows for an authentication and authorisation procedure, where necessary
	A2. metadata are accessible, even when the data are no longer available
Interoperable	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
	I2. (meta)data use vocabularies that follow FAIR principles
	I3. (meta)data include qualified references to other (meta)data

ASPECT	PRINCIPLES
Reusable	R1 (meta)data are richly described with a plurality of accurate and relevant attributes
	R1.1. (meta)data are released with a clear and accessible data usage license
	R1.2. (meta)data are associated with data provenance
	R1.3. (meta)data meet domain relevant community standards

In more detail, data are findable when they are sufficiently described by their metadata and when they are registered and indexed in a searchable resource that is known and accessible to potential users (EC2018; Wilkinson2016).

Digital resources are accessible, when anyone (human or machine) with access to the Internet understands via provided metadata exactly how to access the digital resource and what the conditions on its reuse are (EC2018; Wilkinson2016). A common misinterpretation of this concept is the expectation that accessible (and hence FAIR) digital objects should be 'open' and/or 'free'. This is not what FAIR guiding principles define. The only condition for FAIR digital objects is the clarity and transparency on the conditions of access and reuse of these objects (Mons2017).

Resources are interoperable when they use "normative and community recognized specifications, vocabularies, and standards that determine the precise meaning of concepts and qualities that the data represent" (EC2018, p.19). Although presence of these vocabularies and standards in a format compatible with semantic web would undoubtedly increase their interoperability (vanDenBrink2019;Mons2017), this requirement does not mean that vocabularies and standards used to describe the resource have to 'be on the web.' Use of a well-defined community profile (e.g., an ISO 19115-1:2014 Metadata Profile) and providing metadata in a machine-readable format (e.g., XML) ensure sufficient interoperability of a resource.

License information and the description of the provenance are the two crucial factors determining the reuse of a digital resource by both humans and machines (Mons2017; EC2018). This requires that the description of the license and provenance information need to be provided in a suitable format (e.g., XML or RDF).

Implementation of FAIR varies by community regarding which data should be FAIR and to what degree. In this sense, FAIR needs to be understood as scale and various degrees of FAIRness for different types of digital objects, such as datasets, are hence possible.

10.2. Metadata – crucial element for ensuring FAIRness

Metadata are crucial for ensuring FAIRness of digital resources (Wilkinson2016; Musen2022). Metadata can be intrinsic or user-defined (Mons2017). Intrinsic metadata is created automatically during data capture (e.g., timestamp of a data record, or an automatic label

of a data production software) and user-defined metadata is added to provide context for understanding digital object's creation through provenance information. Both types of metadata should be added to a digital resource to ensure its FAIRness.

10.3. Defining a TDS standard that enables FAIR Principles

In addition to general metadata requirements to meet the FAIR principles, a TDS standard should the following.

- Have all labels and input data (and any other descriptive metadata) link to the assigned globally linked and persistent identifier (F3)

- Provide clear links between labels and input data (R1)

- Capture provenance for both labels and input data (R1.2)

A review of compliance of the proposed TrainingDML-AI standard with the FAIR principles is in Table A.2.

11
SUMMARY

SUMMARY

This section summarizes lessons learned from the Testbed-18 Machine Learning (ML) Training Dataset (TDS) activity and proposes future activities to build on this work and broader OGC activities. The use of Artificial Intelligence (AI), and in particular ML, is an area that is seeing exponential growth both in the Earth Observation (EO) and the broader geospatial community.

11.1. Standards

- Underlying geospatial standards have been developed by ISO that support the definition of a TDS, including the the data quality.

- These ISO standards are being built upon by the Training Data Markup Language (DML) for AI (TrainingDML-AI) Standards Working Group alongside community activities such as the AI Ready EO (AIREO) activity and SpatioTemporal Asset Catalog (STAC).

11.2. Next steps

- Feedback from this Tested-18 activity has been provided to the community drafting the TrainingDML-AI standard and this should continue alongside providing inputs to/collaboration with the broader community activities. A summary of open recommendations are within Annex A.

- TrainingDML-AI introduces proposals to standardize both Provenance and Versions for TDS. These two elements of a TDS standard could end up being exceptionally important in the future because of the centrality of TDS to AI/ML applications, and also because there is an increasing focus on reproducability as part of open science.

- Establish how the metadata of a ML model can link to the metadata of a TDS used to train the ML model.

11.3. Best practice ideas

- A TDS defines and delimits the operational competence of an ML model inferred from it. Therefore, to enable reusability, the domain of the TDS, such as geographical extent, date, and seasonality, must be sufficiently described so a new user can judge the suitability of a TDS for their application. Also, for applications such as object detection, it could also

be helpful to describe the range of sensor orientations represented or the kinds of sensor distortions.

- There would be a benefit in having controlled vocabulary and semantics to specialize, generalize, or combine, etc., existing ML TDS within application domains. Current labeling regimes are almost always taxonomically "flat," i.e., they seldom (if ever) reflect hierarchical concepts such as cars/volkswagons or animals/dogs. Yet such labeling will be required to sustain a healthy market for, and long-term maintenance and extension of, TDSs. One example of a hierachal system is the CORINE land cover nomenclature.

- It may be useful to distinguish the quality of particular labels in the TDS, such as those delineated by experts, as these could be used as a benchmark to assess the accuracy of labels from other annotators.

- To enable the automated generation of metrics such as TDS consistency, Training Data (TD) labels must be attributed with an ID that is unique for each annotator. This is so the number of annotators per label can be counted, as well as whether all unique annotators agreed on the label.

11.4. GeoEthics

- Encouraging the open availability of TDSs so training models can reuse an existing TDS rather than developing a new one. This is costly and time consuming. Still, application-specific TD may be required, with more general TDs being a helpful starting point.

- Alternatively, it may be helpful to store and make the trained ML models publicly available and then "transfer learning" is applied for the specific application. This is a significant approach to reuse in AI/ML with reuse carried out at the model rather than TD level. In this case, knowledge of the used TDS would remain vital as additional TD would need to follow the same input specification. Also, it would be important to have developed an open standard for interoperability between ML architectures, e.g., as developed by ONYX.

- Supporting ML scientists in storing data within standard formats by developing open-source tools/code that allows formatting data in standard formats and ingesting said data into typical ML tools/packages.

A

ANNEX A (NORMATIVE) FEEDBACK ON THE DRAFT TRAININGDML-AI STANDARD

———

A ANNEX A
(NORMATIVE)
FEEDBACK ON THE DRAFT TRAININGDML-AI
STANDARD

There are references within the document to the draft TrainingDML-AI Standard. The draft standard is described in the Current state of art section (Clause 6).

As part of a series of discussions between the Testbed-18 Machine Learning and TrainingDML-AI groups, the following points have been raised and remain open as further discussion is needed.

A.1. How is the geometry specified in TDML?

Currently, the information is expressed as objects or pixels, and the object labels can be expressed by pixel labels. The group agreed that what is needed further is a clarification on what's expected from a compliant training dataset.

A.2. Should there be an option to qualify Training Data with a probability or other confidence score?

The current class "AI_TDQuality" is inherited from the class "DQ_DataQuality" in ISO19157. "DQ_DataQuality" already includes a confidence element, which can show how confident the precision of the training data (TD) is. Also, finding a practical way to calculate quality metrics has not been possible. So, this issue needs more investigation so appropriate metadata can be added in a future version of the standard.

There is another possible meaning for a confidence score, i.e., the confidence of the training data set (TDS) annotator. Consider, for Clause 5.3.5. The resolution of the satellite Earth Observation (EO) data used for labeling means it is not always possible to see what is being labeled. Instead, a third data source is used to intimate that there are plastics, and it is assumed that EO data can positively detect it.

A.3. Use of "Revision" in Update module

The Update module seems to be carrying the notion of "Revision," which is how the software world would characterize changes in its code base. Since TDS are now essential parts of the definition of a codebase functionality, reliability, etc., it might be useful to align with the Software Engineering world.

The resulting discussion was unable to determine whether changing the module's name to "Changeset" to match the class name "AI_TDChangeset" would be helpful for users.

A.4. Requirements identified by use cases

The use cases provided in Clause 5.3 demonstrate a variety of metadata requirements. Table A.1 captures these requirements, the related sections of the draft standard, and comments and recommendations from the ER authors.

Table A.1 — Use case metadata requirements and comments for the draft standard

USE CASE	REQUIRED METADATA	TRANINGDML-AI IMPLEMENTATION	RECOMMENDATION
DELWP Vegetation	Provenance/ manipulation of input data (e.g., resampling, atmospheric corrections, terrain corrections).	AI_EODataSouce covers the whole dataset and does not describe individual EO inputs. AI_EOTraining Data does not include an attribute for whether the data has had additional processing, but this may be covered by the genericAttributes attribute of AI_TrainingData. AI_Labeling supports the modelling of provenance information of the training dataset, but its classes and their attributes specifically capture provenance of the label, rather than the input datasets.	AI_Labeling class should include attributes that describe whether the input data has been manipulated (e.g, resampled, color corrected, atmospherically corrected, terrain corrected). This would address the use case from DELWP, which had imagery from multiple sources, with either 10cm, 15cm, or 20cm resolution. Imagery with 10cm and 15cm resolution was resampled to 20cm prior to labeling. 20cm imagery was not resampled.
	Method of creation	Covered by the AI_Labeler and AI_LabelingProcedure classes.	None.
DELWP Roofprints	Designation of a label to training, validation, test.	Covered by the trainingType attribute of AI_AbstractTrainingData class.	None.
Spatial Services Floodmappir	The task should not restrict presence or type of label	AI_TrainingDataset has AI_Task and AI_Labeling as dependencies. AI_Labeling does not have AI_Task as a dependency.	None.

USE CASE	REQUIRED METADATA	TRANINGDML-AI IMPLEMENTATION	RECOMMENDATION
DE Africa Crop Type	If collected in the field, object labels may exist without associated EO data. They can later be tied to EO data as long as sufficient spatio-temporal information is captured per object label.	AI_ObjectLabel does not have attributes that would cover the date and time of capture. The object attribute of AI_ObjectLabel is sufficient to cover spatial information associated with the object label.	A dateTime attribute should be added to AI_ObjectLabel as an optional obligation. The use-cases/ examples folder on the TrainingDML-AI should include an example where labels were collected in the field, rather than annotated from EO.
	Positional uncertainty of each label, for example, as measured by a GPS device in the field.	This is captured for the TDS as a whole as part of AI_TDQuality, but is not provided for individual labeled objects.	A label-level quality class should be included, which would capture label-specific quality information, such as positional accuracy. The label-level quality class could also include information about whether the labeler was a domain expert, which may help calculate automated quality metrics as discussed in Clause 9.
	The sampling strategy for selection of training data (such as stratification) should be documented.	Overall statistical distribution of training data is captured by statisticsInfo attribute of AI_Training Dataset class, but this does not cover designed distribution (such as stratification conditions) if any.	Two optional attributes should be added to AI_TrainingDataset, one to describe the sampling strategy and one to provide any additional geospatial data that was used as part of the sampling strategy. For example, the Digital Earth Africa crop type use case conducted unsupervised classification to identify variability within the expected cropping regions. The sampling strategy was then developed to sample fields in proportion to their unsupervised class. The standard should include a description of the sampling method, and any additional data (in this case, a raster with the unsupervised classes) that was used to select training data samples. This will help future users understand potential biases in the data, as well as extend the dataset using the same sampling strategy, if required.
World Plastics	New entries must be able to be added.	Captured by the add, change, and delete attributes of AI_TDChangeset class.	None.

USE CASE	REQUIRED METADATA	TRANINGDML-AI IMPLEMENTATION	RECOMMENDATION
	Individual entries must link to sufficient metadata for their associated EO data	Captured by the extent, dateTime, and dataSourceId attributes of the AI_EOTrainingData class. Additional information not specified may be covered by the genericAttributes attribute of AI_TrainingData.	None.
	A whole TDS should be able to be versioned.	Covered by the version, createdTime and updatedTime attributes of AI_AbstractTrainingDataset class, as well as AI_TDChangeset	None.

A.5. Compliance with FAIR principles

In the current TrainingDML-AI draft, attention is given to ensure metadata are defined for each training dataset. A review of compliance of proposed training dataset model with the FAIR principles is in Table A.2:

Table A.2 — TrainingDML-AI compliance with the FAIR principles

FAIR PRINCIPLE	TRANINGDML-AI COMPLIANCE	RECOMMENDATION
F1	The 'id' for the AI_AbstractTrainingDataset or its metadata (e.g., in AI_AbstractTDQuality) is defined as 'CharacterString and there is no explicit element for recording persistent identifier (e.g., DOI or PID). There is an optional 'doi' property through the 'metrics InLIT' property, but the full meaning of this property is unclear. Is this a DOI of related papers or is this a DOI of the training dataset?	DOI or PID should be added to the current model as a property of the AI_Abstract TrainingDataset class
F2	TrainingDML-AI defines ISO 19115-1 compliant metadata where applicable as well as W3C PROV compliant provenance model	None.
F3	Some part of metadata such as AI_TDChangeset clearly and explicitly include the identifier of the data it describes, some, such as AI_AbstractTDQuality, do not.	Although the model preserves these links, it is important to ensure there is a clear link between metadata and the data.
F4	This is not applicable as TrainingDML-AI defines a conceptual schema	None.

FAIR PRINCIPLE	TRANINGDML-AI COMPLIANCE	RECOMMENDATION
A1	This is not applicable as TrainingDML-AI defines a conceptual schema	None
A1.1	This is not applicable as TrainingDML-AI defines a conceptual schema	None
A1.2.	This is not applicable as TrainingDML-AI defines a conceptual schema	None
A2	This is not applicable as TrainingDML-AI defines conceptual schema	None
I1.	Both ISO 19115-1 and W3C PROV are a formal, accessible, shared, and broadly applicable language for knowledge representation	None
I2.	Not specified	ISO 19100 series is FAIR (see in <Ivanova2020>). Vocabularies not yet available in a machine-actionable way, but as this is currently in progress in ISO/TC211, compliance with ISO 19100 series ensures I2.
I3.	Not specified	This can be achieved through ISO 19115-1 compliant 'Citation and responsible party' information.
R1	Scope and usage limitations are not explicitly defined for a TDS, but there is an option, 'genericAttributes,' that allows any additional attributes for a TDS.	More detailed model of provenance and license for TDS is required.
R1.1.	An optional 'license' property is defined in AI_AbstractTrainingDataset class	The license information should be mandatory
R1.2.	Data are associated with data provenance through 'AI_Labeling' class	None.
R1.3.	Both metadata and data meet domain relevant community standards	None.

B

ANNEX B (INFORMATIVE) REVISION HISTORY

B

ANNEX B (INFORMATIVE) REVISION HISTORY

DATE	RELEASE	AUTHOR	PRIMARY CLAUSES MODIFIED	DESCRIPTION
2022-10-31	0.1	S. Lavender	all	Initial version
2022-11-25	0.2	S. Lavender	all	Version for review
2022-12-21	0.3	S. Lavender	Sections 3, 5, 6 and 7	Feedback from TrainingDML-AI approval process
2023-01-05	0.4	C. Adams	Section 5	Additional feedback from TrainingDML-AI approval process

BIBLIOGRAPHY

BIBLIOGRAPHY

[1]   OGC: OGC xx-xxx: Training Data Markup Language for Artificial Intelligence (TrainingDML-AI) Part 1: Conceptual Model Standard, Draft

[2]   Wilkinson M.D. et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3:160018, http://doi.org/10.1039/sdata.2016.18

[3]   Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B. and Wilkinson, M. D., 2017. Cloudy, increasigly FAIR; revisiting the FAIR Data guiding princples for the European Open Science Cloud, Information Services & Use, 37(1), http://doi.org/10.3233/ISU-170824.

[4]   European Commission (EC), 2018. Turning FAIR into reality — Final Report and Action Plan from the European Commission Expert Group on FAIR data, EC: Brussels, https://doi.org/10.2777/1524

[5]   van den Brink, L. Barnaghi, P., Tandy, J., Atemezing, G., Atkinson, R., Cochrane, B., … Troncy, R. (2019) Best practices for publishing, retrieving and using spatial data on the web, Semantic Web 10(1), 95-114, http://www.semantic-web-journal.net/system/files/swj1785.pdf

[6]   Ivánová I. et al., 2020. Ensuring FAIR access to precise positioning data by improving geodetic interchange standards. https://frontiersi.com.au/wp-content/uploads/2020/11/P1003-Geodetic-Standards-Final-Report.pdf

[7]   Musen, M.A. (2022). Without appropriate metadata, data-sharing mandates are pointless, Nature 609, 222 (2022), doi: https://doi.org/10.1038/d41586-022-02820-7