SLOVENSKI STANDARD
oSIST prEN ISO 19178-1:2024

01-september-2024

Geografske informacije - Jezik za označevanje podatkov za usposabljanje za umetno inteligenco - 1. del: Standard konceptualnega modela (ISO/DIS 19178-1:2024)

Geographic information - Training data markup language for artificial intelligence - Part 1: Conceptual model standard (ISO/DIS 19178-1:2024)

Ta slovenski standard je istoveten z: **prEN ISO 19178-1**

ICS:

07.040	Astronomija. Geodezija. Geografija	Astronomy. Geodesy. Geography
35.060	Jeziki, ki se uporabljajo v informacijski tehniki in tehnologiji	Languages used in information technology
35.240.70	Uporabniške rešitve IT v znanosti	IT applications in science

oSIST prEN ISO 19178-1:2024 **en,fr,de**

DRAFT International Standard

ISO/DIS 19178-1

Geographic information — Training data markup language for artificial intelligence —

Part 1: Conceptual model standard

ICS: ISO ics

ISO/TC **211**

Secretariat: **SIS**

Voting begins on:
2024-07-10

Voting terminates on:
2024-10-02

This document is circulated as received from the committee secretariat.

IMPORTANT — Please use this updated version dated 2024-05-30, and discard any previous version of this DIS. The VA relation has been removed.

Reference number
ISO/DIS 19178-1:2024(en)

COPYRIGHT PROTECTED DOCUMENT

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

ISO draws attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO takes no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO *[had/had not]* received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents. ISO shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 211, *Geographic information/Geomatics*, in collaboration with the European Committee for Standardization (CEN) Technical Committee CEN/TC 287, *Geographic Information*, in accordance with the Agreement on technical cooperation between ISO and CEN (Vienna Agreement) and in collaboration with the Open Geospatial Consortium Inc. (OGC).

A list of all parts in the ISO 19178 series can be found on the ISO website.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

The Training Data Markup Language for Artificial Intelligence (TrainingDML-AI) Standard aims to develop the UML model and encodings for geospatial machine learning training data. Training data plays a fundamental role in Earth Observation (EO) Artificial Intelligence Machine Learning (AI/ML), especially Deep Learning (DL). It is used to train, validate, and test AI/ML models. This Standard defines a UML model and encodings consistent with the OGC Standards baseline to exchange and retrieve the training data in the Web environment.

The TrainingDML-AI Standard provides detailed metadata for formalizing the information model of training data. This includes but is not limited to the following aspects:

— How the training data is prepared, such as provenance or quality;

— How to specify different metadata used for different ML tasks such as scene/object/pixel levels;

— How to differentiate the high-level training data information model and extended information models specific to various ML applications;

— How to introduce external classification schemes and flexible means for representing ground truth labeling.

Geographic information — Training data markup language for artificial intelligence —

Part 1: Conceptual model standard

1 Scope

Training data is the building block of machine learning models. These models now constitute the majority of machine learning applications in Earth science. Training data is used to train AI/ML models, and to then validate model results. Formalizing and documenting the training data by characterizing the training data content, metadata, data quality, and provenance, and so forth is essential.

This document describes work actions around training data:

— Documents the UML model with a target of maximizing the interoperability and usability of EO imagery training data;

— Defines different AI/ML tasks and labels in earth observation in terms of supervised learning, including scene level, object level and pixel level tasks;

— Describes the description of the permanent identifier, version, license, training data size, measurement or imagery used for annotation, and so on;

— Defines the description of quality (e.g. training data errors, training data representativeness) and the provenance (e.g. agents who perform the labeling, labeling procedure).

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.

ISO 19101-1:2014, *Geographic information — Reference model — Part 1: Fundamentals*

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

ISO 19157-1:2023, *Geographic information — Data quality — Part 1: General requirements*

3 Terms and definitions

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

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1 Terms and definitions

3.1.1
artificial intelligence
AI
branch of computer science devoted to developing data processing systems that perform functions normally associated with human intelligence, such as reasoning, learning, and self-improvement

[SOURCE: ISO/IEC/IEEE 24765:2017, 3.234]

3.1.2
machine learning
ML
process of optimizing model parameters through computational techniques, such that the model's behaviour reflects the data or experience

Note 1 to entry: ML processes create models from training data by using a set of learning algorithms, and then can use these models to make predictions. Depending on whether the training data include labels, the learning algorithms can be divided into supervised and unsupervised learning.

[SOURCE: ISO/IEC 22989:2022, 3.3.5, modified — Note 1 has been added]

3.1.3
deep learning
DL
approach to creating rich hierarchical representations through the training of neural networks with one or more hidden layers

Note 1 to entry: Deep learning uses multi-layered networks of simple computing units (or “neurons”). In these neural networks each unit combines a set of input values to produce an output value, which in turn is passed on to other neurons downstream.

[SOURCE: ISO/IEC TR 29119-11:2020, 3.1.26]

3.1.4
dataset
identifiable collection of data

Note 1 to entry: A dataset can be a smaller grouping of data which, though limited by some constraint such as spatial extent or feature type, is located physically within a larger dataset. Theoretically, a dataset can be as small as a single feature or feature attribute contained within a larger dataset. A hardcopy map or chart can be considered a dataset.

[SOURCE: ISO 19115-1:2014, 4.3]

3.1.5
training dataset
collection of samples, often labelled in terms of supervised learning

Note 1 to entry: A training dataset can be divided into training, validation, and test sets. Training samples are different from samples in ISO 19156:2023. They are often collected in purposive ways that deviate from purely probability sampling, with known or expected results labelled as values of a dependent variable for generating a trained predictive model.

3.1.6
label
<earth observation> refers to known or expected results annotated as values of a dependent variable in training samples

Note 1 to entry: A training sample label is different from those on a geographical map, which are known as map labels or annotations.

3.1.7
class
<classification> result of a classification process as part of a classification system which subdivides concepts within a given topic area

[SOURCE: ISO 19144-2:2023, 3.1.6]

3.1.8
provenance
documents the chronology of events regarding the creation, modification, ownership and custody of a resource, such as who produced it and who has had custody since its origination; it provides information on the history of the multimedia content (including processing history)

Note 1 to entry: In this document provenance is a record of how training data were prepared.

[SOURCE: ISO/IEC 23000-15:2016, 3.4.1, modified — Note 1 has been added]

3.1.9
quality
degree to which a set of inherent characteristics of an object fulfils requirements

Note 1 to entry: Quality of training data (such as data imbalance and mislabeling) can impact the performance of AI/ML models.

[SOURCE: ISO 9000:2015, 3.6.2, modified — Notes 1 and 2 to entry have been deleted, and a new Note 1 has been added]

3.1.10
scene classification
<earth observation> task to identify scene categories of images, on the basis of a training set of images whose scene categories are known

3.1.11
object detection
<earth observation> recognition of objects from images

Note 1 to entry: The objects are often localized using bounding boxes.

3.1.12
semantic segmentation
<earth observation> task to assign class labels to pixels of images or points of point clouds

3.1.13
change detection
<earth observation> recognition of changes in an area between images taken at different times

3.1.14
3D model reconstruction
<earth observation> task that builds 3D objects and scenes from multi-view images

3.1.15
generative model
method of large model training, which improve model performance through unsupervised pre-training

Note 1 to entry: In the fine-tuning phase, labelled data plays a critical role in optimizing the model for specific vertical domains or tasks. By incorporating labelled data, the model can learn to accurately identify and extract relevant features, leading to better performance on specific downstream tasks. Overall, the combination of generative models and fine-tuning with labelled data can significantly improve the performance of large models in specialized domains or tasks.

3.2 Abbreviated terms

In this document, the following abbreviations and acronyms are used or introduced:

AI	artificial intelligence
DL	deep learning
EO	earth observation
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
LC	land cover
LU	land use
ML	machine learning
OGC	Open Geospatial Consortium
RS	remote sensing
TD	training data
UML	Unified Modelling Language
URL	Uniform Resource Locator
XML	Extensible Markup Language

4 Conventions

This section provides details and examples for any conventions used in the document. Examples of conventions are symbols, abbreviations, use of XML schema, or special notes regarding how to read the document.

4.1 Identifiers

The normative provisions in this specification are denoted by the URI

`http://www.opengis.net/spec/TrainingDML-AI-1/1.0`

All requirements and conformance tests that appear in this document are denoted by partial URIs which are relative to this base.

4.2 UML Notation

The Standard is presented in this document through diagrams using the Unified Modeling Language (UML) static structure diagram. The UML notations used in this document are described in the diagram in Figure 1.

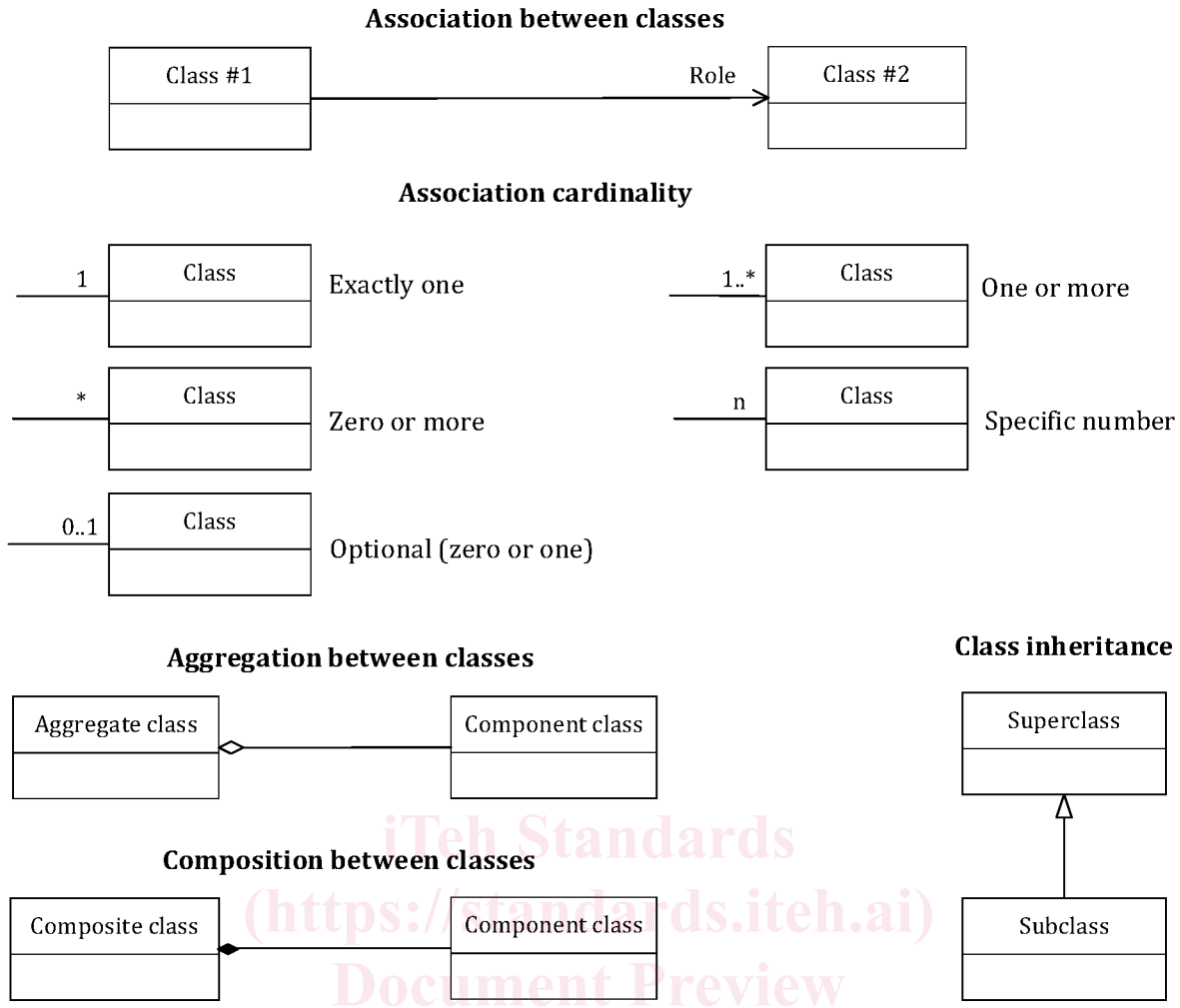

Figure 1 — UML notation (see ISO 19103, Geographic information — Conceptual schema language)

All associations between model elements in the TrainingDML-AI Conceptual Model are uni-directional. Thus, associations in the model are navigable in only one direction. The direction of navigation is depicted by an arrowhead. In general, the context an element takes within the association is indicated by its role. The role is displayed near the target of the association. If the graphical representation is ambiguous though, the position of the role has to be drawn to the element the association points to.

The following stereotypes are used in this model.

— «DataType» defines a set of properties that lack identity. A data type is a classifier with no operations, whose primary purpose is to hold information.

— «CodeList» enumerates the valid attribute values. In contrast to Enumeration, the list of values is open and, thus, not given inline in the TrainingDML-AI UML Model. The allowed values can be provided within an external code list.

5 Conformance

This TrainingDML-AI Standard defines a conceptual model that is independent of any encoding or formatting technologies. The standardization targets for this document is:

— TrainingDML-AI Conceptual Model

Conformance with this document shall be checked using all the relevant tests specified in Annex A of this document. The framework, concepts, and methodology for testing, and the criteria to be achieved

to claim conformance are specified in the OGC Compliance Testing Policies and Procedures and the OGC Compliance Testing web site.

All requirements-classes and conformance-classes described in this document are owned by the standard identified.

6 Overview

The TrainingDML-AI Conceptual Model Standard defines how to represent and exchange ML training data. The conceptual model includes the most relevant training data entities from datasets, to instances (i.e. individual training samples), to labels. The conceptual schema specifies how and into which parts of the training data should be decomposed and classified.

TrainingDML-AI Conceptual Model Standard strategically addresses geospatial requirements by providing a modular and extensible framework tailored to EO applications. The content and format of training datasets differ depending on the EO ML scenarios they were collected for (e.g. scene/object/pixel levels). A training data model defines a UML model and encodings consistent with the OGC/ISO baseline standards to exchange and retrieve geospatial training data. On the one hand, existing geospatial standards can be reused when defining geospatial requirements on source RS images, label geometry, metadata, and quality. On the other hand, while generic information entities are defined for training data at the high level, other EO-specific information, such as the size of each sample image, spatial extent, and bands, can be extended in a subclass at the low level. With a hierarchical and extensible structure, it accommodates diverse geospatial data characteristics, ensuring flexibility and interoperability.

The TrainingDML-AI conceptual model (Clause 7) is formally specified using UML class diagrams, complemented by a data dictionary (Clause 8) providing the definitions and explanations of the object classes and attributes. This conceptual model provides the basis for specifying encoding implemented in languages such as JSON, or XML.

6.1 AI Tasks for EO

In recent years AI/ML is increasingly used in the EO domain. The new AI/ML algorithms frequently require large training datasets as benchmarks. AI/ML TD have been used in many EO applications to calibrate the performance of AI/ML models. Many efforts have been made to produce training datasets to make accurate predictions. As a result, a number of training datasets are publicly available, with new datasets being constantly released. In the EO domain, examples of AI/ML training datasets have been developed in various tasks including the following typical scenarios:

— Scene classification. These algorithms determine image categories from numerous pictures (e.g. agricultural, forest, and beach scenes). The training samples are a series of labelled pictures. The data can be either from satellite, drones, or aircrafts. The metadata of the datasets often includes the number of training samples, the number of classes, and the image size.

— Object detection. These algorithms detect and localize different objects (e.g. airplanes, cars and building) in a single image. The image can be optical or non-optical, such as Synthetic Aperture Radar (SAR). Recent work also suggests an increasing focus on object detection from street view imagery. Objects can be labelled with two forms of bounding boxes, i.e., oriented and horizontal bounding boxes. The geometry of a bounding box can be expressed using top-left/bottom-right coordinates, coordinates of four corners, or center coordinates along with the length and width of the box.

— Semantic segmentation. In terms of Land cover (LC) and land use (LU) classification, this process assigns a LC/LU class label to a pixel (or groups of pixels) of RS imagery. Considering semantic segmentation of 3D point clouds, it is to classify points of a 3D point cloud into categories. TDs are usually composed of RS images/point clouds, and the corresponding labelled value of each pixel/point recording its class.

— Change detection. These algorithms identify the difference between images acquired over the same geographical area but taken at different times. The TD comprise a set of pre-change and post-change RS images, with the corresponding ground truth map labelled changed and unchanged pixels. The image can be optical or SAR images.