Quality Considerations for TrainingDML-AI

ESIP IQC Monthly Telecon
Peng Yue
Wuhan University
24 August 2021

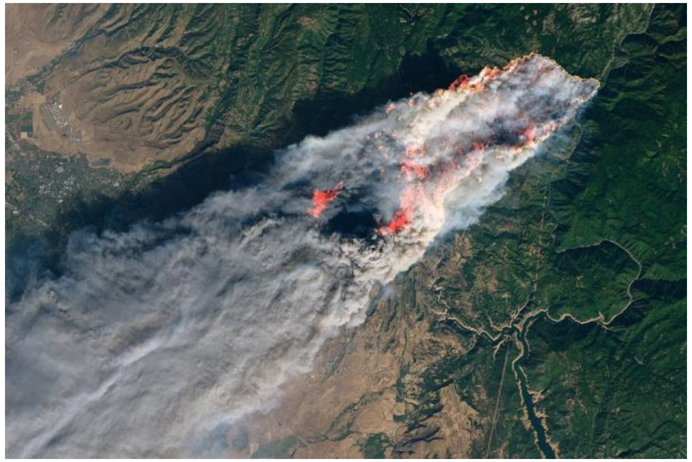
Training Data for AI/ML

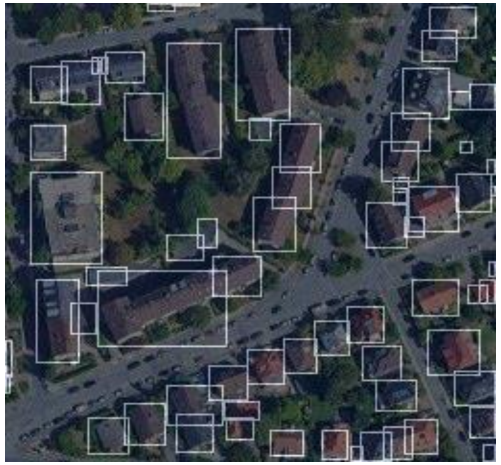
❖ Remote Sensing Machine Learning Scenarios:

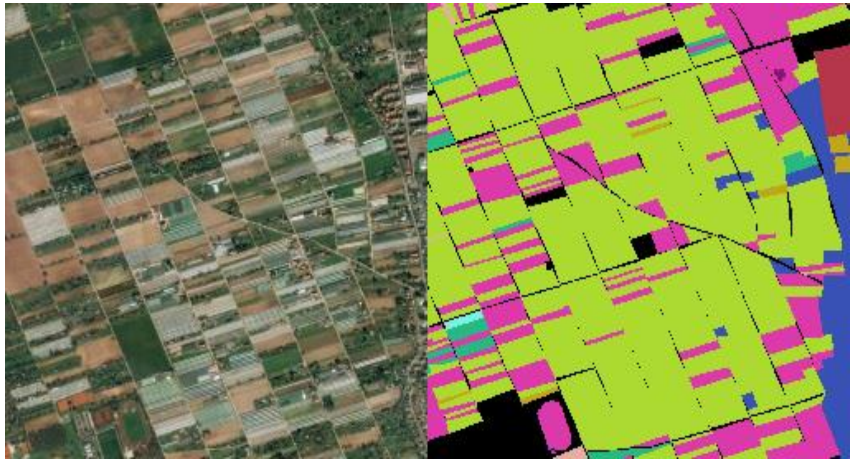
- In the scene level, e.g., the wildfire scene classification, the training data content includes an image and its corresponding binary label

- In the object level, e.g., the building detection, the training data content includes an image with several polygons indicating the position of buildings

- In the pixel level, e.g., the landcover classification, the training data content includes the Earth Observation (EO) imagery and the landcover class of each pixel







Background

❖ Purpose of the Standard Working Group:

- The TrainingDML-AI SWG is chartered to develop the UML model and encodings for Artificial Intelligence/Machine Learning Training Data. The usage of TrainingDML is to train AI/ML models, and to validate model results.
- The SWG will investigate the feasibility and interoperability of OGC standards to use and share geospatial Training Data in AI/ML applications and describe gaps and issues that can lead to a new geospatial standard.
- The geospatial Training Data categories will include, but are not restricted to, remote sensing imagery, moving features (e.g., vehicle trajectories), and related spatial content.
- The UML model and encodings will consistent with the OGC standards baseline to exchange and retrieve the geospatial training data in the Web environment.

OGC®

Background

❖ Scope of Work for TrainingDML-AI SWG :

- Discuss the cutting-edge issues of training data for AI in the geospatial community;
- Design the UML model and encoding of TrainingDML for AI ;
- Define the description of spatial and temporal representativeness;
- Define the description of whether it’s a classification or object detection task, type of applicable AI/ML model/algorithm, preferred accuracy level, techniques used to generate the training data, original data used to generate labels;
- Define the description of the permanent identifier, version, license, training data size, dates of measurement or imagery used for annotation, uncertainty of the measurement, data privacy, etc;
- Define the description of quality evaluation (e.g., training data errors, training data sparsity) and the provenance (all intermediate data and training process);
- …

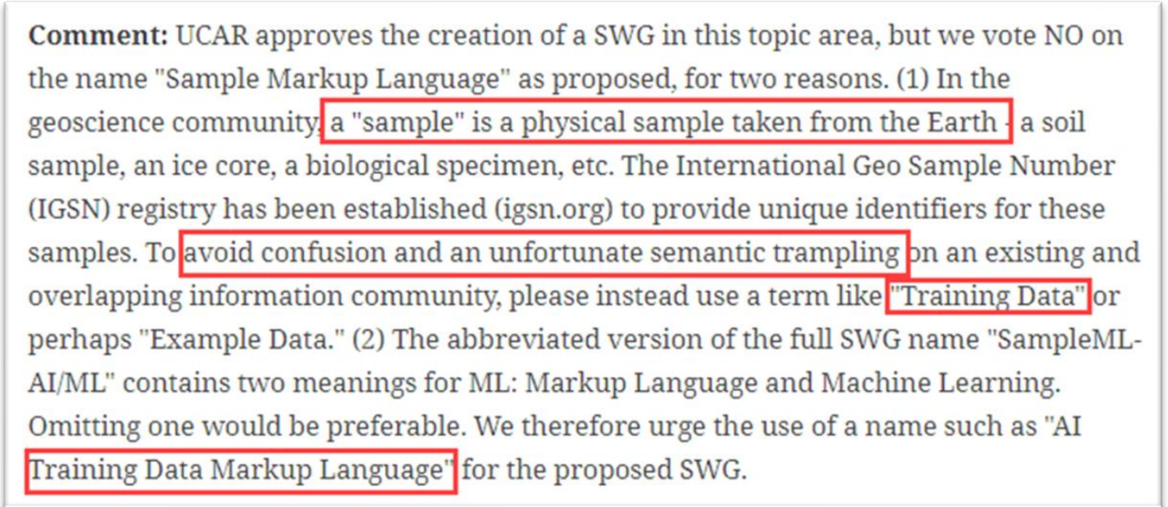
OGC®

History

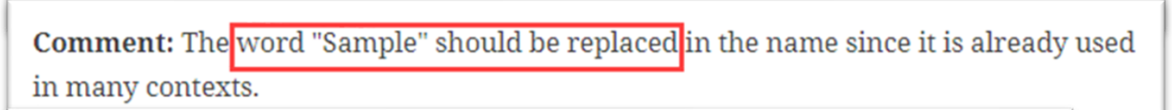
- Stage 1 (Proposal):
  - 01/30/2021 OGC 21-003 submitted to portal
  - Hydrology and Geociences DWG mailing lists discussion on names
- Stage 2 (Ad hoc session):
  - 02/27/2021 OGC 21-003r1 submitted to portal
  - 03/22 SampleML ad hoc session at 118TH OGC MEMBER MEETING, discussion on Name Abbreviation and O&M harmonization
- Stage 3 (Seeking Public comment):
  - 04/09/2021 OGC 21-003r2 submitted to portal
  - Seeking public comment for SampleML Charter: 04/15/2021-05/05/2021
- Stage 4（TC Vote）
  - 05/26/2021 OGC 21-003r3 submitted to portal after addressing public comments
  - Reporting to 119TH OGC MEMBER MEETING and starting TC vote to approve charter

History

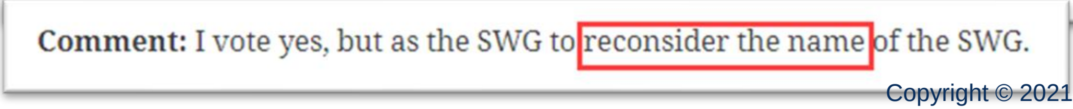
- Stage 5:
  - 08/09/2021 The SWG vote in the TC has passed
  - Based on the comments on the votes, the full name of SWG/standard has been changed to "Training Data Markup Language for Artificial Intelligence". The abbreviation has been updated to "TrainingDML-AI". The data quality in TrainingDML has been aligned with ISO 19157



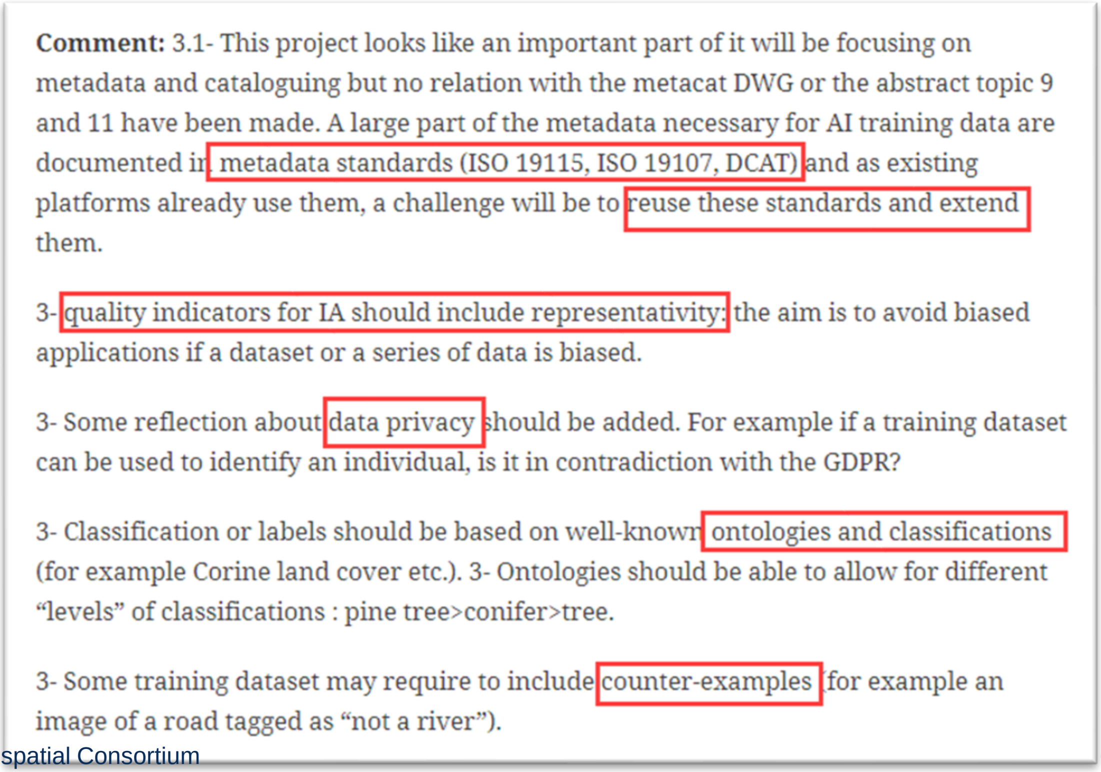
Comment: UCAR approves the creation of a SWG in this topic area, but we vote NO on the name "Sample Markup Language" as proposed, for two reasons. (1) In the geoscience community, a "sample" is a physical sample taken from the Earth - a soil sample, an ice core, a biological specimen, etc. The International Geo Sample Number (IGSN) registry has been established (igsn.org) to provide unique identifiers for these samples. To avoid confusion and an unfortunate semantic trampling on an existing and overlapping information community, please instead use a term like "Training Data" or perhaps "Example Data." (2) The abbreviated version of the full SWG name "SampleML-AI/ML" contains two meanings for ML: Markup Language and Machine Learning. Omitting one would be preferable. We therefore urge the use of a name such as "AI Training Data Markup Language" for the proposed SWG.



Comment: The word "Sample" should be replaced in the name since it is already used in many contexts.



Comment: I vote yes, but as the SWG to reconsider the name of the SWG.



Comment: 3.1- This project looks like an important part of it will be focusing on metadata and cataloguing but no relation with the metacat DWG or the abstract topic 9 and 11 have been made. A large part of the metadata necessary for AI training data are documented in metadata standards (ISO 19115, ISO 19107, DCAT) and as existing platforms already use them, a challenge will be to reuse these standards and extend them.

3- quality indicators for IA should include representativity: the aim is to avoid biased applications if a dataset or a series of data is biased.

3- Some reflection about data privacy should be added. For example if a training dataset can be used to identify an individual, is it in contradiction with the GDPR?

3- Classification or labels should be based on well-known ontologies and classifications (for example Corine land cover etc.). 3- Ontologies should be able to allow for different "levels" of classifications : pine tree>conifer>tree.

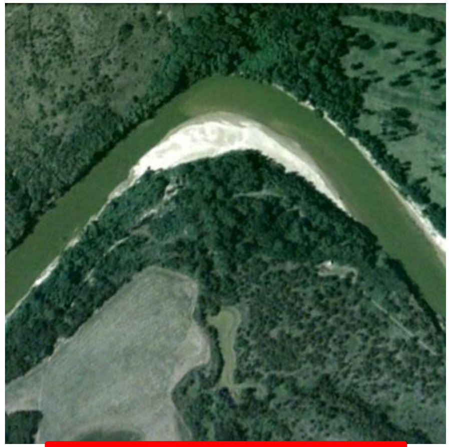
3- Some training dataset may require to include counter-examples for example an image of a road tagged as "not a river").

Quality considerations for TrainingDML-AI

❖ Remote Sensing Machine Learning Scenarios

- **Scene level**: It is necessary to consider the scene label correctness, image size, degree of class balance, etc.

Scene Label Correctness



Forest.jpg

Label :Forest

Ground truth: River

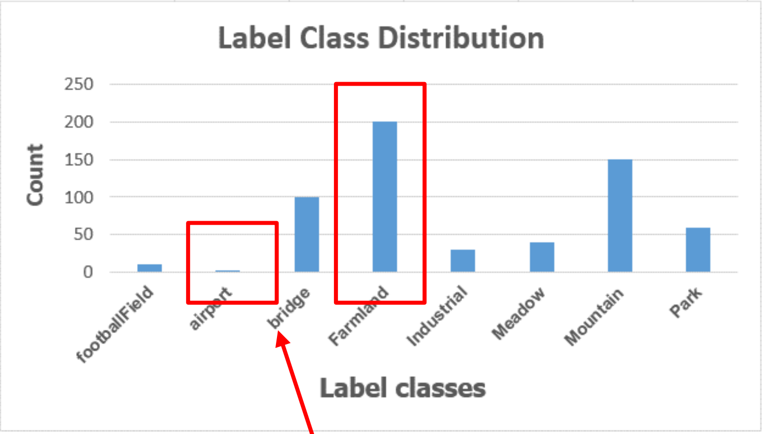
Image Size

image size 256×256

image size 16×16

The image size is too small for the general AI model

Class Balance Degree



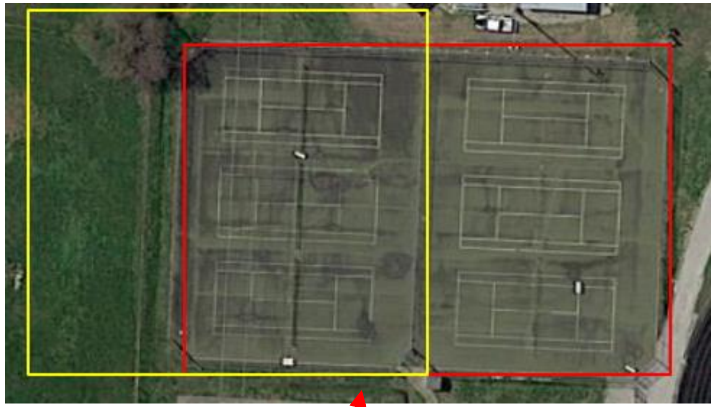
The distribution of label classes is skew

Quality considerations for TrainingDML-AI

❖ Remote Sensing Machine Learning Scenarios

- **Object level**：The annotation of the object detection records the coordinate of the object, it is necessary to evaluate the positional accuracy of the object label, omission of object labels, label overlap rate, etc.

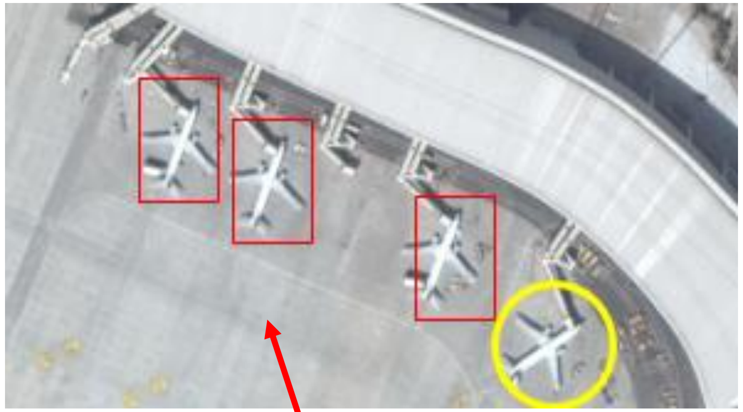
Label Positional Accuracy



Ground truth: tennis court

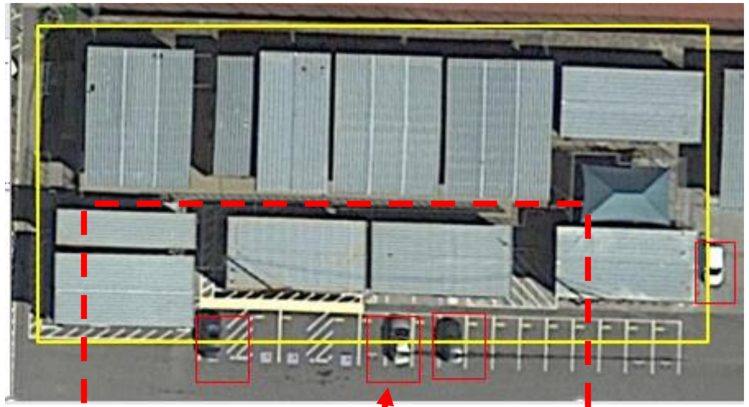
Red: correct label
Yellow : offset label

Object Label Omission



Red : correct labels
Yellow : missing label

Label Overlap Rate



Red Object : Car
Yellow Object: Building

Quality considerations for TrainingDML-AI

❖ Remote Sensing Machine Learning Scenarios

- **Pixel level**：The label of segmentation record the land cover/land use type in every pixel and the semantic relationship between each category, considering pixel label correctness, image size, degree of category balance.

Pixel Label Correctness

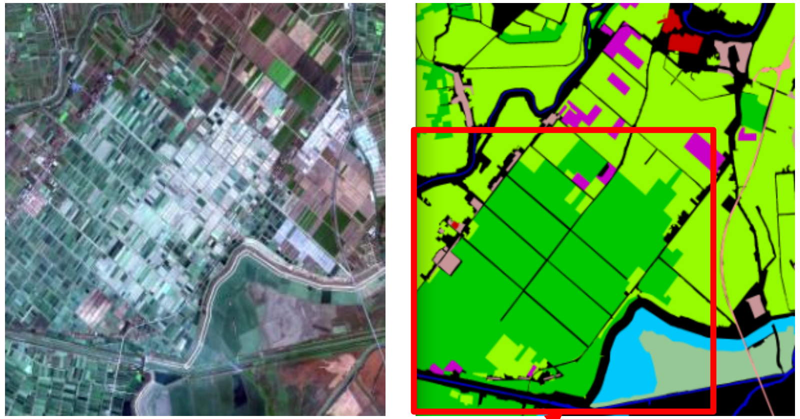


Image source 🟩 paddy field 🟩 irrigated land

Label :paddy field

Ground truth: wheat

Image Size

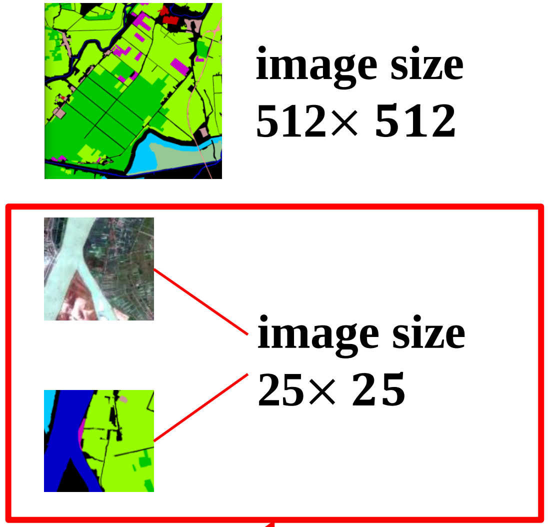
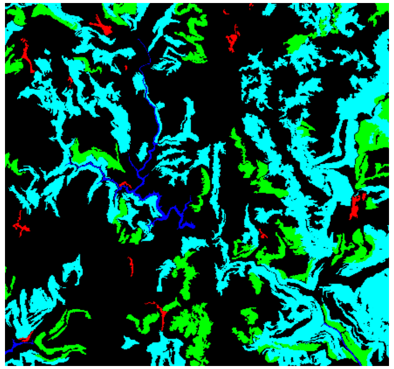


image size 512× 512

image size 25× 25

The image size is too small for the general AI model

Class Balance Degree



className	ratio
🟥 built-up	3%
🟩 farmland	10%
🟦 forest	20%
🟨 meadow	0%
🟦 water	5%

The quantity distribution of land cover species is unbalanced.

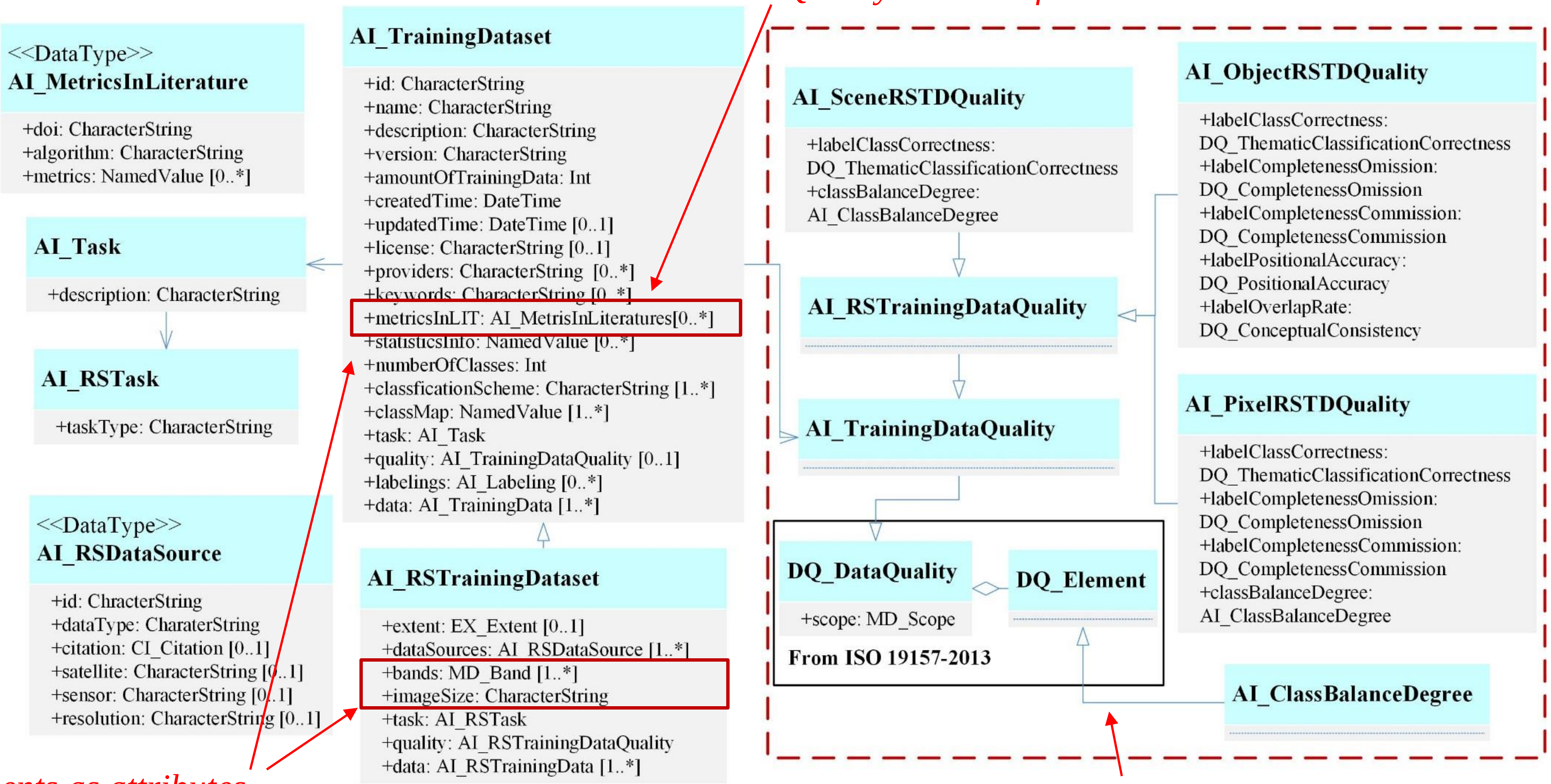
Quality considerations for TrainingDML-AI

❖ The AI Training Data Quality Model

We defined the AI Training Data Quality using two strategies: add quality elements as attributes, and map quality elements to ISO 19157.

- AI_TrainingDataset
- AI_RSTrainingDataset
- AI_TrainingDataQuality
- AI_RSTrainingDataQuality
- AI_SceneRSTDQuality
- AI_ObjectRSTDQuality
- AI_PixelRSTDQuality

Quality elements from user-centric view



<<DataType>>
AI_MetricsInLiterature
+doi: CharacterString
+algorithm: CharacterString
+metrics: NamedValue [0..*]

AI_Task
+description: CharacterString

AI_RSTask
+taskType: CharacterString

<<DataType>>
AI_RSDataSource
+id: ChracterString
+dataType: CharaterString
+citation: CI_Citation [0..1]
+satellite: CharacterString [0..1]
+sensor: CharacterString [0..1]
+resolution: CharacterString [0..1]

AI_TrainingDataset
+id: CharacterString
+name: CharacterString
+description: CharacterString
+version: CharacterString
+amountOfTrainingData: Int
+createdTime: DateTime
+updatedTime: DateTime [0..1]
+license: CharacterString [0..1]
+providers: CharacterString [0..*]
+keywords: CharacterString [0..*]
+metricsInLIT: AI_MetrisInLiteratures[0..*]
+statisticsInfo: NamedValue [0..*]
+numberOfClasses: Int
+classficationScheme: CharacterString [1..*]
+classMap: NamedValue [1..*]
+task: AI_Task
+quality: AI_TrainingDataQuality [0..1]
+labelings: AI_Labeling [0..*]
+data: AI_TrainingData [1..*]

AI_RSTrainingDataset
+extent: EX_Extent [0..1]
+dataSources: AI_RSDataSource [1..*]
+bands: MD_Band [1..*]
+imageSize: CharacterString
+task: AI_RSTask
+quality: AI_RSTrainingDataQuality
+data: AI_RSTrainingData [1..*]

AI_SceneRSTDQuality
+labelClassCorrectness:
DQ_ThematicClassificationCorrectness
+classBalanceDegree:
AI_ClassBalanceDegree

AI_RSTrainingDataQuality

AI_TrainingDataQuality

DQ_DataQuality
+scope: MD_Scope

DQ_Element

From ISO 19157-2013

AI_ObjectRSTDQuality
+labelClassCorrectness:
DQ_ThematicClassificationCorrectness
+labelCompletenessOmission:
DQ_CompletenessOmission
+labelCompletenessCommission:
DQ_CompletenessCommission
+labelPositionalAccuracy:
DQ_PositionalAccuracy
+labelOverlapRate:
DQ_ConceptualConsistency

AI_PixelRSTDQuality
+labelClassCorrectness:
DQ_ThematicClassificationCorrectness
+labelCompletenessOmission:
DQ_CompletenessOmission
+labelCompletenessCommission:
DQ_CompletenessCommission
+classBalanceDegree:
AI_ClassBalanceDegree

AI_ClassBalanceDegree

Add quality elements as attributes

Map quality elements to ISO 19157

Future Work

- Refining the data quality characteristics for TrainingDML-AI
- Seeking inputs from wider communities

OGC®