OGC Testbed 18 work relating TrainingDML-AI to the ISO 19000 series

Ivana Ivánová (SA & OGC DQ DWG & Curtin University/FrontierSI)
Testbed-18 Machine Learning Training Datasets

• **Goal**: To develop the foundation for future standardization of Training Datasets for Earth Observation applications.

• **Participants**:

<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sam Lavender</td>
<td>Pixalytics Ltd</td>
<td>Editor</td>
</tr>
<tr>
<td>Kate Williams</td>
<td>FrontierSI</td>
<td>Editor</td>
</tr>
<tr>
<td>Caitlin Adams</td>
<td>FrontierSI</td>
<td>Editor</td>
</tr>
<tr>
<td>Ivana Ivánová</td>
<td>Curtin University</td>
<td>Editor</td>
</tr>
<tr>
<td>Jim Antonisse</td>
<td>NGA</td>
<td>Contributor</td>
</tr>
<tr>
<td>Sara Saeedi</td>
<td>OGC</td>
<td>Contributor</td>
</tr>
<tr>
<td>Sina Taghavikish</td>
<td>OGC</td>
<td>Contributor</td>
</tr>
</tbody>
</table>
Future standardisation of TD @OGC

Training Data Markup Language for AI SWG

- **Mission**: to develop the UML model and encodings for geospatial Machine Learning training data.

- **Draft standard was prepared for TC approval at the time of T18-D27 start**
Future standardisation of TD @OGC

• TrainingDML-AI SWG and T18-D27 aligned efforts:
  • T18-D27 joined the SWG and reviewed the Draft TrainingDML-AI standard
  • The SWG joined T18-D27 and is one of the approvers of T18-D27 Engineering Report
• Draft TrainingDML-AI standard revised on-the-fly
Draft TrainingDML-AI standard & ISO

- Modular standard

- **Quality**
  - TD Quality

- **Provenance**
  - Labeling
  - Labeler
  - Labeling Procedure

- **Changeset**
  - TD Changeset

- **Basic**
  - Training Dataset
  - Training Data
  - Task
  - Label

- **DataQuality**

- **AI_Task**

- **AI_TrainingDataset**

- **AI_Labeling**

- **AI_Label**

- **AI_TrainingData**

- **AI_TDChangeset**
Draft TrainingDML-AI standard & ISO

- Dependent on latest ISO 19100 series

10. TrainingDML-AI Data Dictionary
   10.1. ISO Classes
       10.1.1. Feature (from ISO 19107:2019)
       10.1.2. MD_Band (from ISO 19115-1:2014)
       10.1.3. EX_Extent (from ISO 19115-1:2014)
       10.1.4. CI_Citation (from ISO 19115-1:2014)
       10.1.5. DataQuality (from ISO 19157-1)
       10.1.6. QualityElement (from ISO 19157-1)
Draft TrainingDML-AI standard & ISO

• Taking advantage of the latest developments – *e.g. ISO 19100 dependency*

---

**Before T18-D27**

- **ISO 19107:2003: Spatial Schema**
  - (from ISO 19107 All)

- **ISO 19115-1:2014: Metadata**
  - (from ISO 19115-1 All)

- **ISO 19157:2013: Data Quality**
  - (from ISO 19157 All)

---

**After T18-D27**

- **ISO 19107:2019: Spatial Schema**
  - (from ISO 19107 All)

- **ISO 19115-1:2014: Metadata**
  - (from ISO 19115-1 All)

- **ISO/FDIS 19157-1: Data Quality**
  - (from ISO 19157-1 All)
• Taking advantage of the latest developments – e.g. AI_TDQuality
D27-ER Table of Contents

1. SCOPE .............................................................................................................. 2
2. NORMATIVE REFERENCES ............................................................................. 4
3. TERMS, DEFINITIONS AND ABBREVIATED TERMS ................................. 6
   3.3. Abbreviated terms .................................................................................... 6
4. ENGINEERING REPORT OVERVIEW ................................................................. 9
5. INTRODUCTION TO AI/ML WITHIN THE CONTEXT OF EARTH OBSERVATION
   .............................................................. 11
   5.1. Defining AI and ML .................................................................................. 11
   5.2. Typical formats for TDSs in EO Applications ............................................ 12
   5.3. Example use cases ................................................................................... 12
   5.4. Opportunities ........................................................................................ 19
   5.5. Challenges ............................................................................................ 20
6. CURRENT STATE OF ART ............................................................................... 24
   6.1. Training Data Markup Language for Artificial Intelligence draft standard ... 24
   6.2. Spatial/Temporal Asset Catalog (STAC) ..................................................... 25
   6.3. ESA funded initiatives and projects such as AI4EO .................................. 25
   6.4. ANZLIC considerations of TDS as foundational data ............................ 26
   6.5. Public TDS repositories ........................................................................... 27
   6.6. Previous OGC activities ......................................................................... 29
7. METADATA REQUIREMENTS AND RECOMMENDATIONS ..................... 34
   7.1. Current structure and usage of metadata in ML TDS ................................. 34
7.2. Review and application of ISO metadata standards for ML TDS ................. 35
7.3. Examples of human and machine-readable metadata for a TDS .................. 46
8. TDS CATALOGS .............................................................................................. 50
   8.1. What is a catalog? ................................................................................... 50
   8.2. Version control for TDS ........................................................................... 51
   8.3. Splitting source data and annotated training data ................................. 52
   8.4. Making TDS catalogs self-explanatory .................................................... 52
9. TDS QUALITY ................................................................................................. 54
   9.1. Biases and domains in TDS .................................................................... 55
   9.2. Auto-generation of quality indicators ...................................................... 56
10. ENABLING FAIR IN THE FUTURE TDS STANDARD ................................. 58
   10.1. The FAIR guiding principles ................................................................. 58
   10.2. Metadata - crucial element for ensuring FAIRness ............................... 59
   10.3. Defining a TDS standard that enables FAIR Principles .......................... 60
11. SUMMARY ..................................................................................................... 62
   11.1. Standards ............................................................................................ 62
   11.2. Next steps ........................................................................................... 62
   11.3. Best practice ideas ............................................................................... 62
   11.4. GeoEthics ............................................................................................ 63
ANNEX A (NORMATIVE) FEEDBACK ON THE DRAFT TRAININGDML.AI STANDARD .......................................................................................... 65
   A.1. How is the geometry specified in TDML? ........................................... 65
   A.2. Should there be an option to quality Training Data with a probability or other confidence score? ................................................................. 65
   A.3. Use of “Revision” in Update module ...................................................... 66
   A.4. Requirements identified by use cases ................................................... 66
   A.5. Compliance with FAIR principles ......................................................... 68
ANNEX B (INFORMATIVE) REVISION HISTORY ............................................. 71
BIBLIOGRAPHY ................................................................................................. 73
Relevant insights for the standard's development

• Review of how this standard compares to State of the Art?

• Use cases from around the world commonly use multiple data sources as input, the standard must be able to treat these as a single Training Dataset

• Label geometries can be collected in the field and through delineation from input data. Both use cases must be supported

• Data quality is important at the level of labels, such as positional accuracy from field data collection, or annotator confidence
T18-D27 ER and the ISO 19100: Section 7

- ISO 19115-1
  - Discovery metadata
  - Applicability of MD_Scope

- ISO 19115-2
  - Extension for imagery applications
  - Extended lineage

- ISO 19157-1
  - Data quality fundamentals
  - Treatment of domain specific data quality

- ISO 19157-3
  - Data quality measure structure
  - Data quality measure register

- Extension options of ISO 19115-1 & ISO 19115-2 & ISO 19157-1 & ISO 19157-3
Recommendations for use-case support

• Describe input data manipulation in AI_Labeling
  • e.g., resampling, terrain correction, atmospheric correction

• Include dateTime attribute for field-collected data in AI_ObjectLabel

• Create a quality class at the label level
  • e.g., with positional accuracy, capture whether labeller was expert/non-expert

• Add optional attributes for sampling strategy (description and supporting geospatial data) to AI_TrainingDataset
Recommendations for FAIR compliance

• Add DOI or PID to AI_AbstractTrainingDataset
• Improve consistency and clarity of links between metadata and data
• Develop a more detailed model of provenance and license
• Make license information mandatory

• All of these recommendations apply to ISO 19100 series too! ...but that is a presentation for another session...
Thank you!

ivana.ivanova@curtin.edu.au