Best Practices For Using Machine Learning Keywords in Collection and Service Records in the Common Metadata Repository (CMR)

Overview

This best practices document describes how to use Machine Learning keywords in curating collection and service records in the Common Metadata Repository (CMR) to improve metadata quality and discoverability of machine learning data and models.

Machine Learning Model keywords are a representation of predictive models that, when trained on a set of data containing certain features, enables a computer to identify similar features in other data. Machine Learning Training Data keywords are a representation of input data necessary for running a machine learning model.

Best Practices

Science keywords from the GCMD Keyword Management System (KMS) are important for the precise search and retrieval of data and should accurately represent the data being described. At a minimum, one science keyword must be provided, and the level of the keywords must go down to the 'Term' level. Additional keywords can be requested through the GCMD Keywords Community Forum.

- The Earth Science keywords must be picked from the GCMD KMS.
    - Examples
        - OCEANS > MARINE ENVIRONMENT MONITORING > MARINE SURFACE ELEMENTS > MARINE SURFACE DEBRIS
        - OCEANS > OCEAN CIRCULATION > DIFFUSION
        - BIOSPHERE > ECOLOGICAL DYNAMICS > COMMUNITY DYNAMICS > BIODIVERSITY FUNCTIONS
    - The full list of valid keywords can be found here.
- The Machine Learning Training Data keywords must be picked from the GCMD KMS.
    - Examples
        - MACHINE LEARNING TRAINING DATA > LABELS > RASTER LABEL
        - MACHINE LEARNING TRAINING DATA > SOURCE > VECTOR SOURCE
        - MACHINE LEARNING TRAINING DATA > SOURCE > RASTER SOURCE
    - The full list of valid keywords can be found here.
- The Machine Learning Model keywords must be picked from the GCMD KMS.
    - Examples
        - MODELS > MACHINE LEARNING MODELS > SUPERVISED
        - MODELS > MACHINE LEARNING MODELS > DECISION TREE > ISOLATION FOREST
        - MODELS > MACHINE LEARNING MODELS > DEEP LEARNING > GENERATIVE ADVERSARIAL NETWORKS
    - The full list of keywords can be found here.

Machine Learning Training Data (Model Training Data)

1. Describe your Machine Learning training data as a UMM Collection (UMM-C) compliant record in CMR.
    a. See the full list of UMM-C fields at UMM-C Schema Documentation.
2. Add relevant Machine Learning Training Data keywords describing the input data necessary for running a machine learning model.
3. Add relevant Earth Science keywords describing what is being measured as part of the training data set.
4. Ingest your UMM-C record into the CMR. *(Note: A NASA Agency AUID and Launchpad authentication is required for metadata ingest operations.)*
5. Add the 'machine.learning' CMR tag to your record. *(Note: Tag permission is granted through the provider via the INGEST_MANAGMENT_ACL. Users can add tags if they are granted the appropriate permission.)*
    a. To add a tag to a collection record, see the Search API Documentation.

Machine Learning Models (Model)

1. Describe your machine learning model as a UMM Services (UMM-S) compliant record in CMR.
2. Add relevant Machine Learning Model keywords describing the type of model that was used to train the data.
3. Ingest your UMM-S record into the CMR. *(Note: A NASA Agency AUID and Launchpad authentication is required for metadata ingest operations.)*
4. Add a collection association to relevant Machine Learning Training Data collections (from above) in the CMR. *(Note: Collection association is granted through the provider via the INGEST_MANAGMENT_ACL. Users can add tags if they are granted the appropriate permission.)*

a. To associate a service with one or more collections using the Metadata Management Tool (MMT), see the MMT User Guide.
b. To associate a service with one or more collections using the CMR API, see the CMR Search API Documentation.

Example Collection Record

- Marine Debris Dataset for Object Detection in Planetscope Imagery

References

- CMR Ingest API Documentation
- UMM Documentation