
European Commission
21
March
2024
Introductory webinar on
MLDCAT-AP
interoperable
europe
innovation ∞ govtech ∞ community

Agenda

01 **Introduction**

02 **MLDCAT-AP & the DCAT-AP ecosystem**

03 **Guest speaker: OpenML**

04 **MLDCAT-AP: a closer look**

05 **Next steps**

Workshop practicalities

Audio

Click on ‘connect audio’ but please mute your microphones

Chat

You can also share your questions for the Q&A session via the chat

Recording

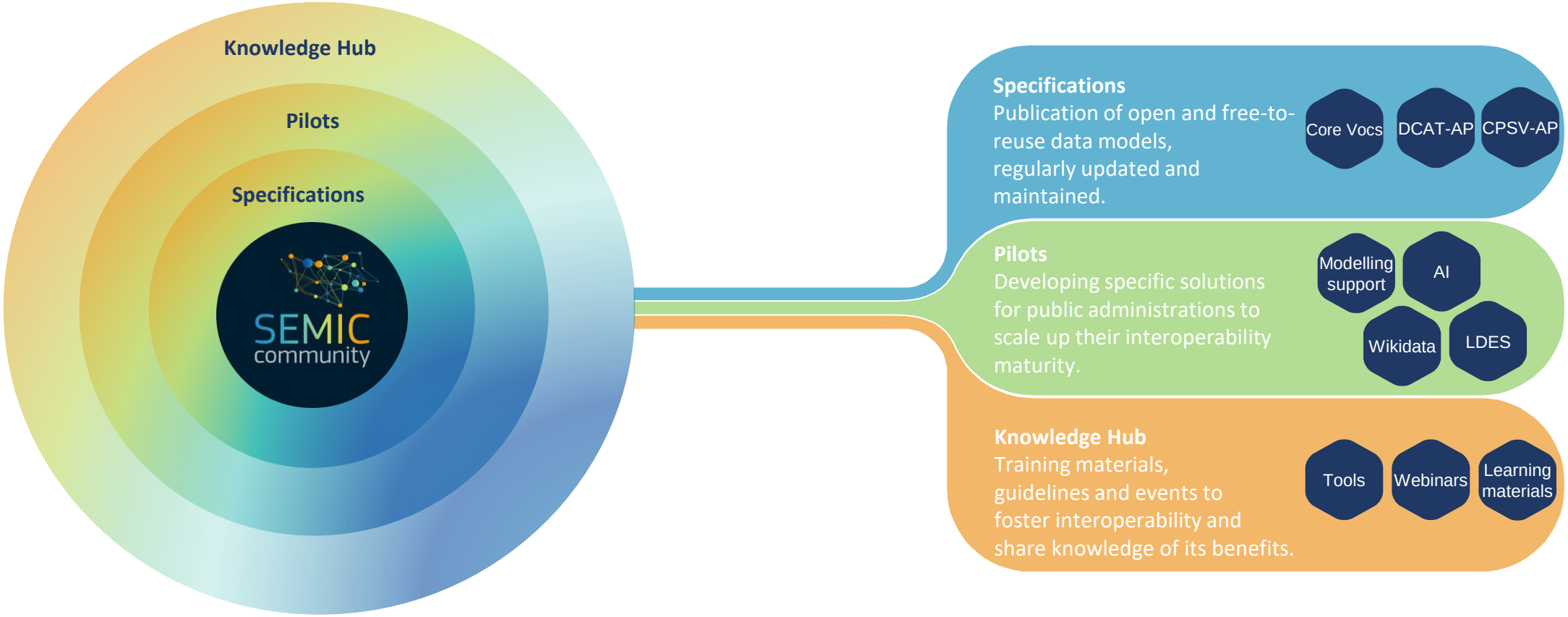
The workshop will be recorded

Context of the SEMIC assets

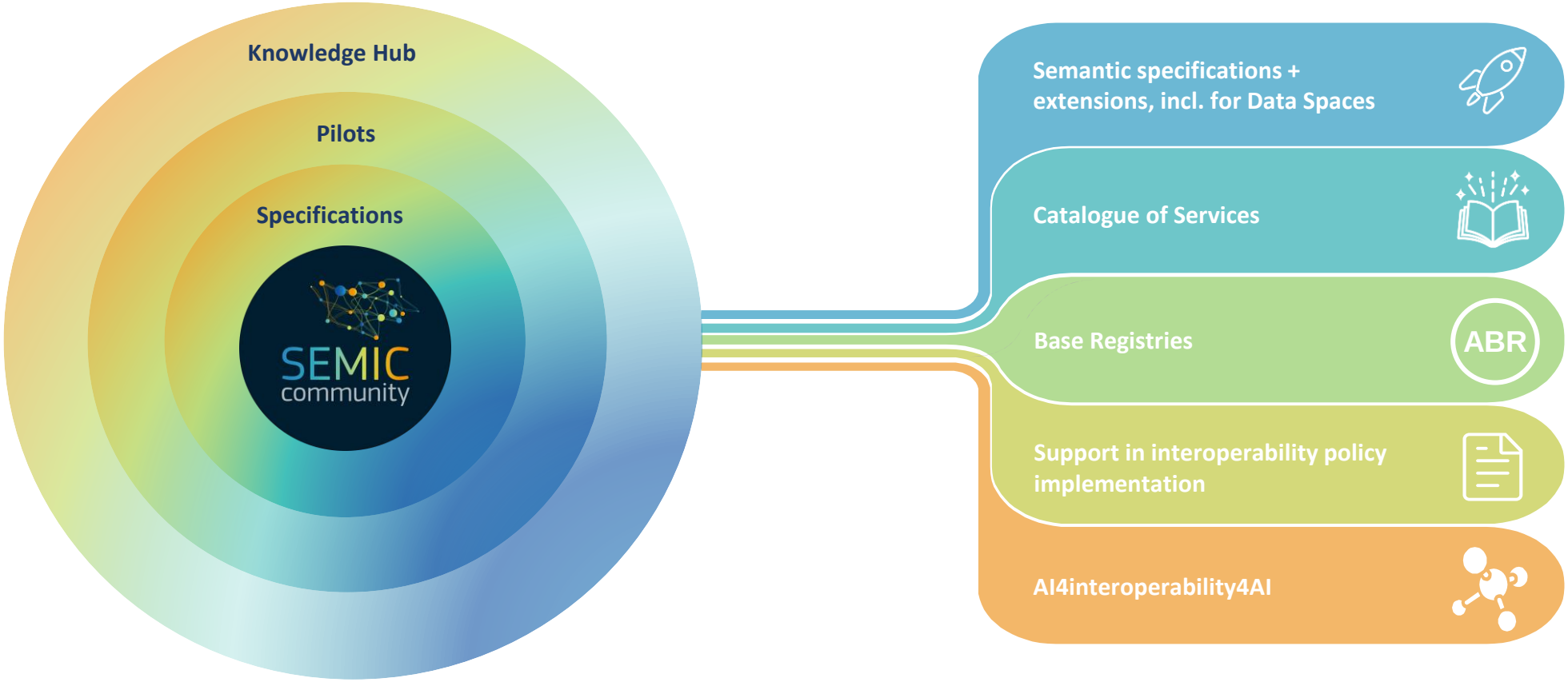
European Commission

SEMIC

SEMIC's mission is to promote Semantic Interoperability amongst the EU Member States and deliver pragmatic support to help build an Interoperable Europe.



SEMIC Focus Areas



Specifications

SEMIC specifications enable interoperability:

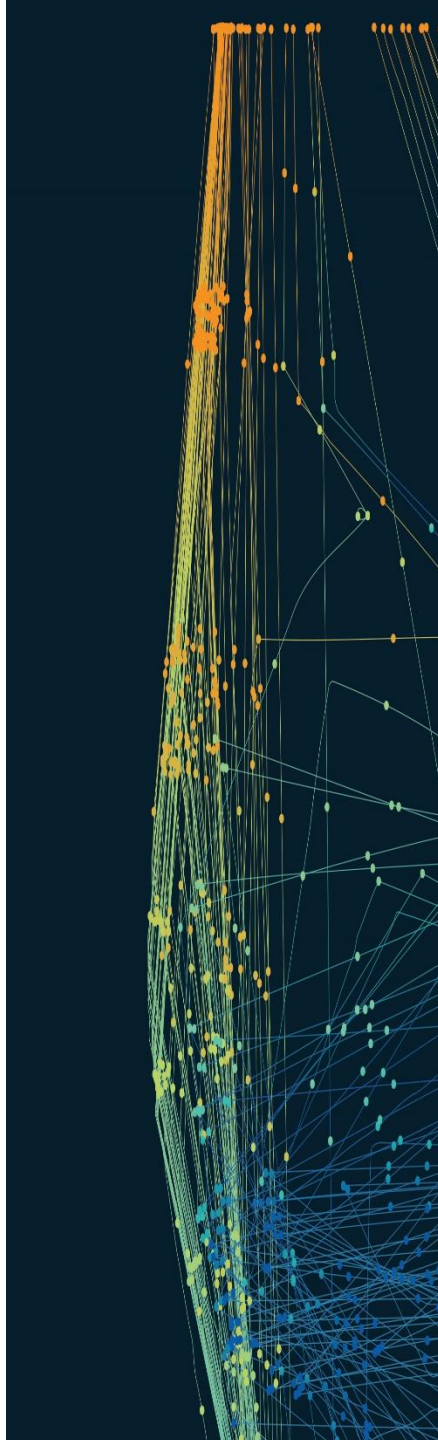
- They make data **transparent** and **available**
- They support the **coherent** implementation of laws and policies
- They help implement **cost efficiencies**
- They help **digitalisation** and **harmonising** processes

Core Vocabularies

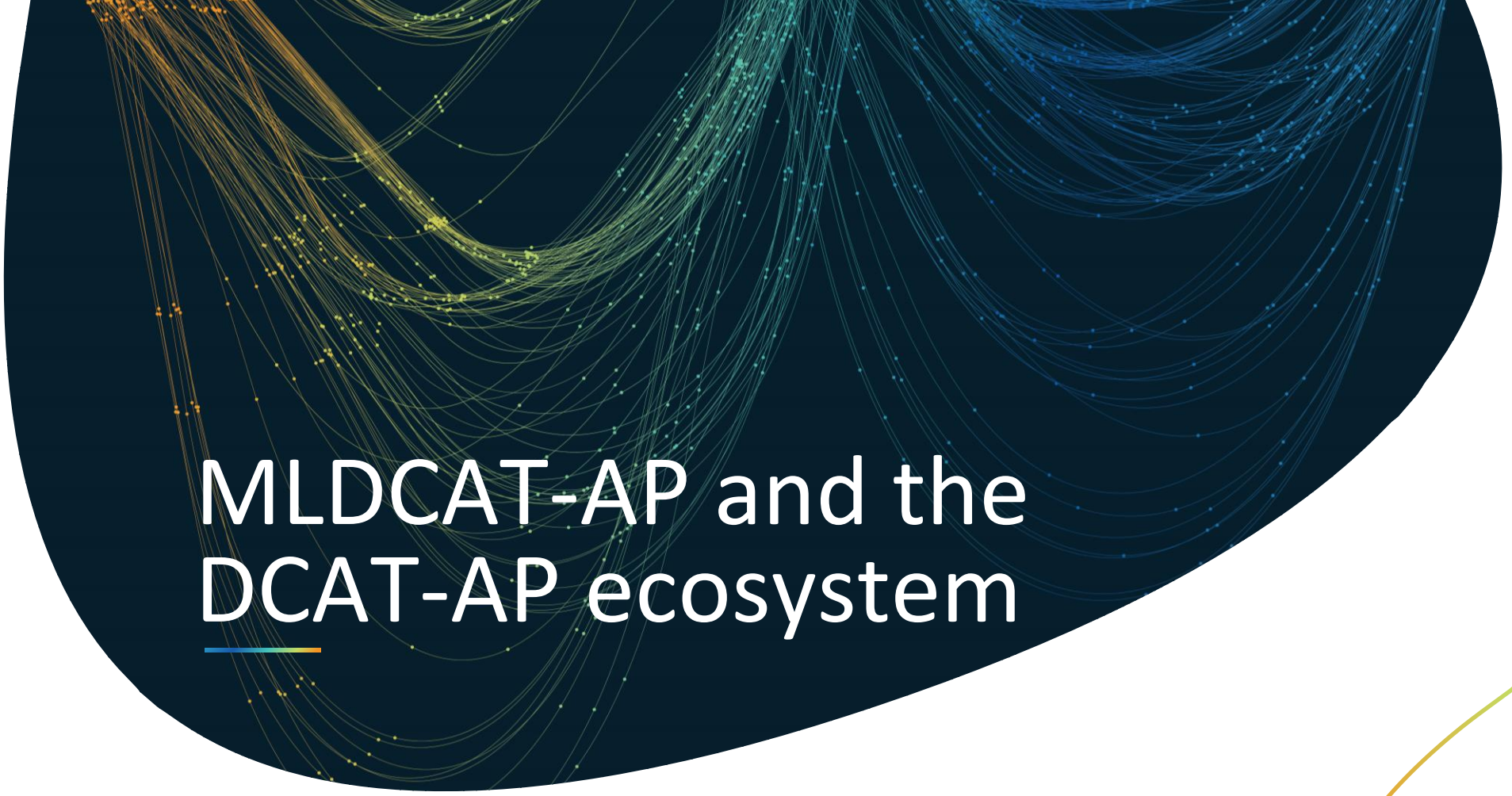
Core Vocabularies are a cornerstone element of semantic interoperability. They provide a standardised approach for describing key concepts such as locations, businesses, organisations and natural persons.

Application Profiles

Application Profiles make use of vocabularies for a detailed set of use cases to define mandatory relations, constraints and relationships.



European Commission



MLDCAT-AP and the DCAT-AP ecosystem

European Commission

Objectives of DCAT-AP

Supporting the discovery of/access to (open) data in a cross-border and cross-domain environment, by describing metadata to be harvested across a distributed network of portals.

In the form of an application profile of W3C DCAT, by

- expressing constraints and usages on DCAT properties and classes, and
- including additional properties and usages of controlled vocabularies

Domains of applications

Open data portals with an extension for statistics and geospatial data.

Base registries metadata descriptions

Data spaces

- NAPCORE-Mobility
- GeoDCAT-AP
- HealthDCAT-AP
- …

Machine Learning with MLDCAT-AP

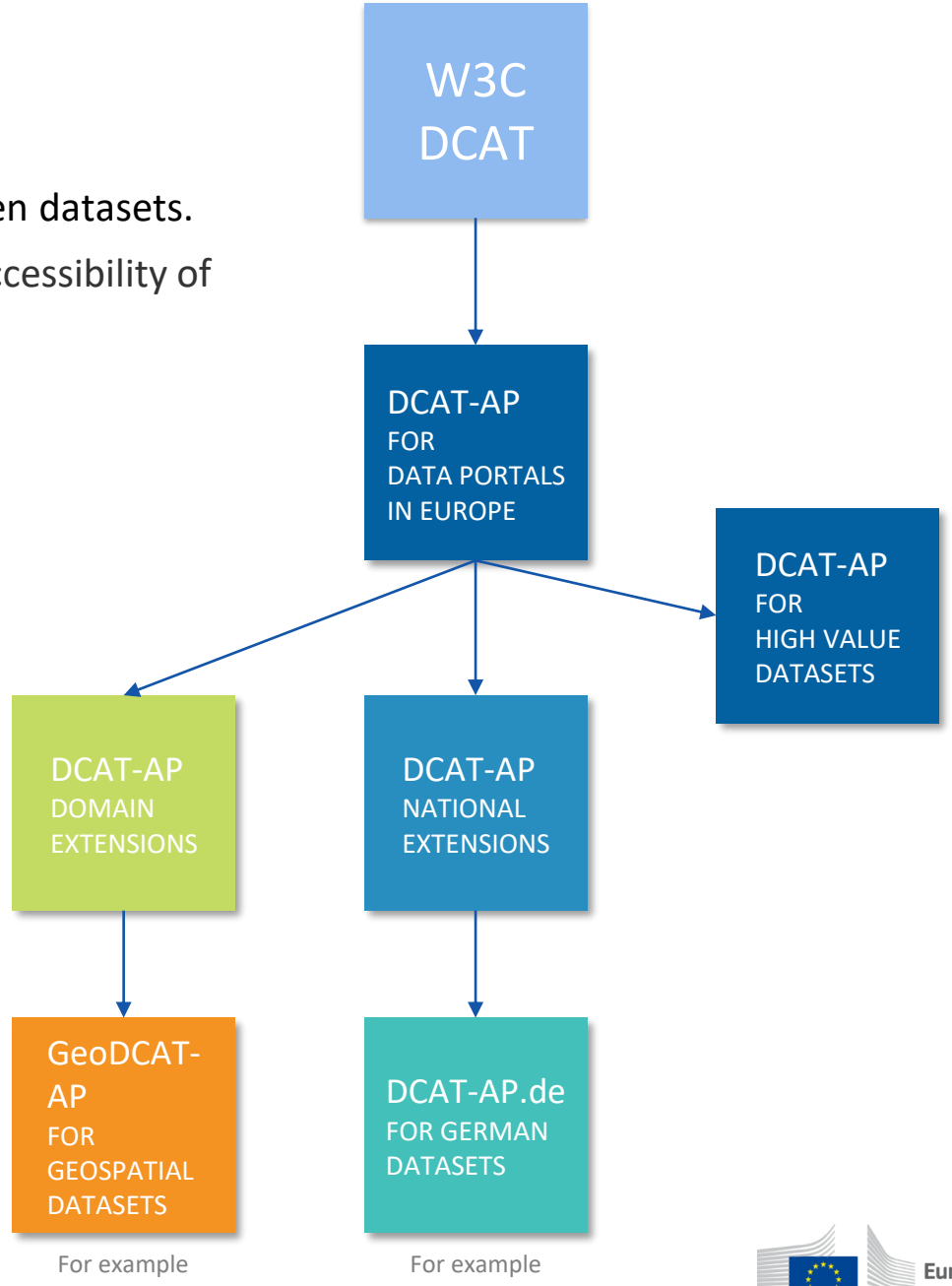
DCAT-AP ecosystem

W3C Data Catalogue Vocabulary for facilitating interoperability between datasets.

- **DCAT 3:** extends DCAT 2 and introduces classes to support offline accessibility of datasets and datasets that are part of a series.

DCAT Application Profile for describing datasets based on W3C DCAT.

- **DCAT-AP 3.0.0:** fully compatible and aligned with DCAT 3.
- **DCAT-AP for High-Value Datasets:** facilitates adherence to the HVD Implementing Regulation with little additional effort.



DCAT-AP ecosystem

DCAT-AP National Extension for describing national datasets.

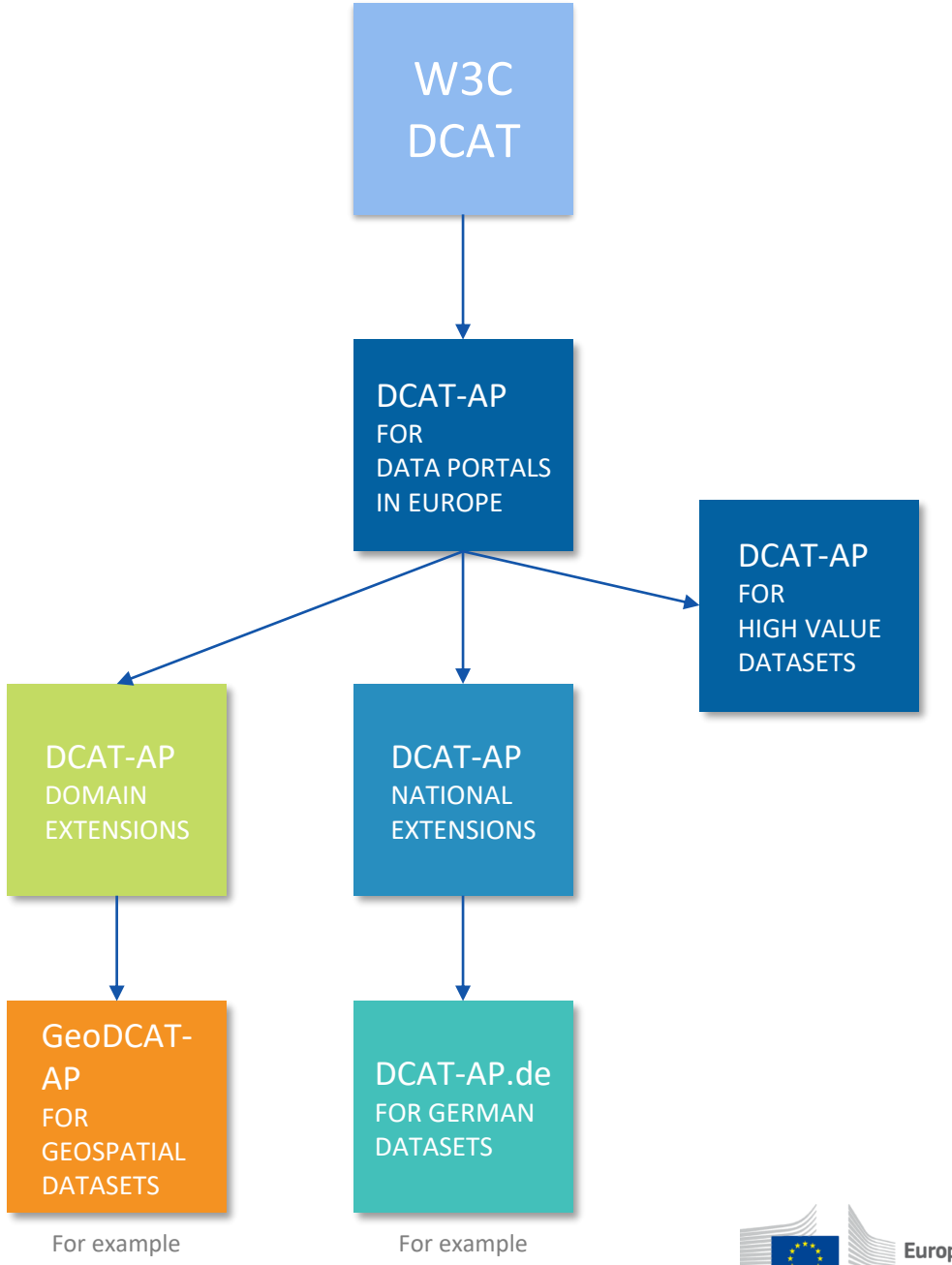
- DCAT-AP.de (Germany)
- DCAT-AP.it (Italy)
- …

DCAT-AP Domain Extension for describing domain specific datasets.

- GeoDCAT-AP for geospatial datasets
- StatDCAT-AP for statistical datasets
- HealthDCAT-AP for datasets in the health industry
- …

The official European Data Portal – data.europa.eu

- Provides access to data from all EU institutions, agencies, and bodies.
- Uses DCAT-AP for own datasets.
- Aggregates datasets from DCAT-AP compliant portals across Europe.



Benefits of the DCAT-AP ecosystem

Enhances the findability and accessibility of data

Comes with a decade of **experience** of documenting, maintaining metadata records; sharing through harvesting, etc.

Provides tooling to **validate the implementation** data.

Enables implementors to make **data catalogues findable**
→ A harvesting network is made possible

Enables implementors to express their **metadata in a standardised way**

Collaborative environment that allows implementors to express their needs and additional requirements (specialisations)

DCAT-AP ecosystem & MLDCAT-AP

Problem statement

Lack of semantic interoperability forbids assets (including machine learning models) to be easily exchanged with other platforms.

Strategy

Define a common data model and enrich existing API with semantics.

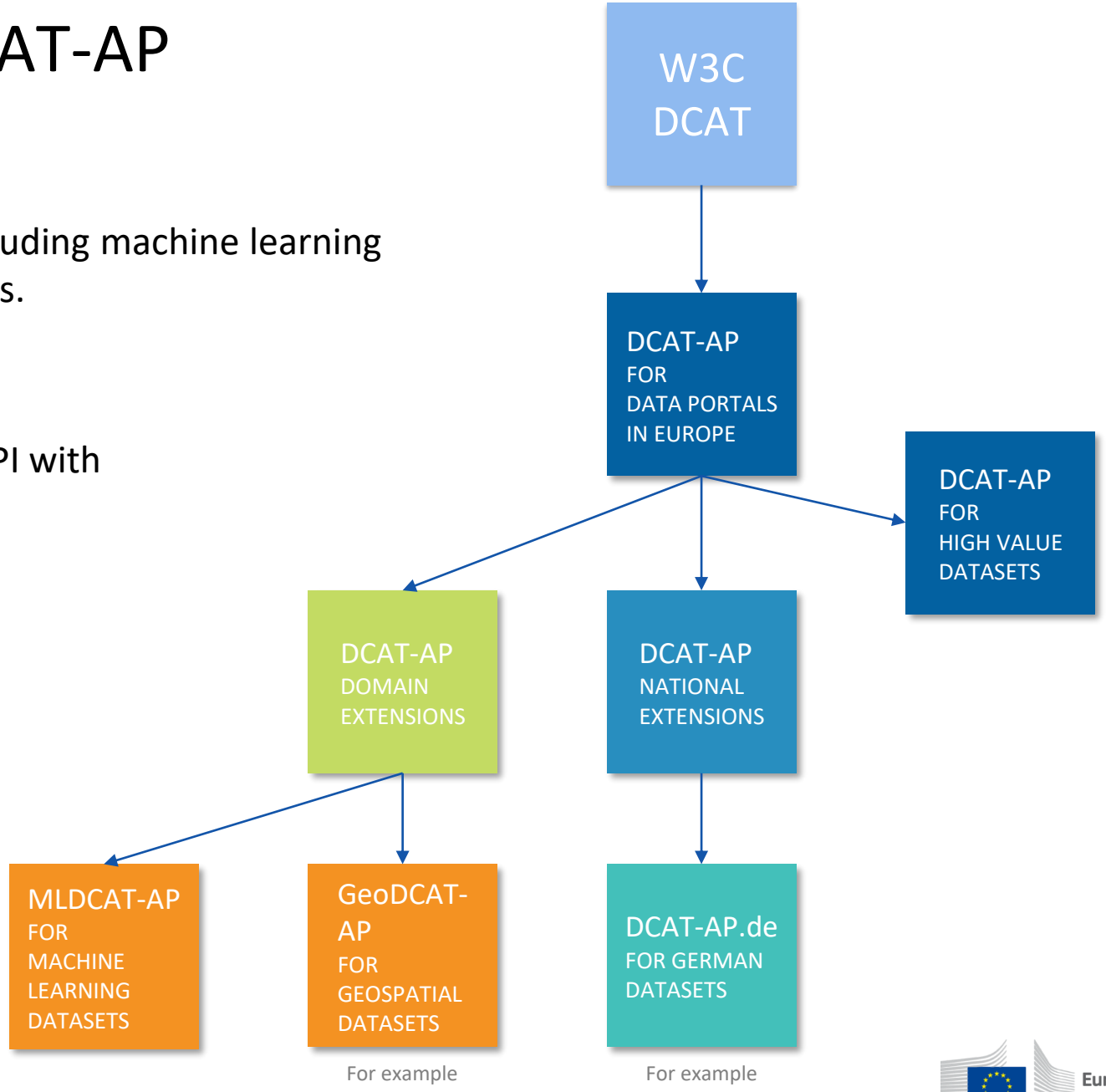
- Compatible with DCAT-AP 3.0.0

Benefits

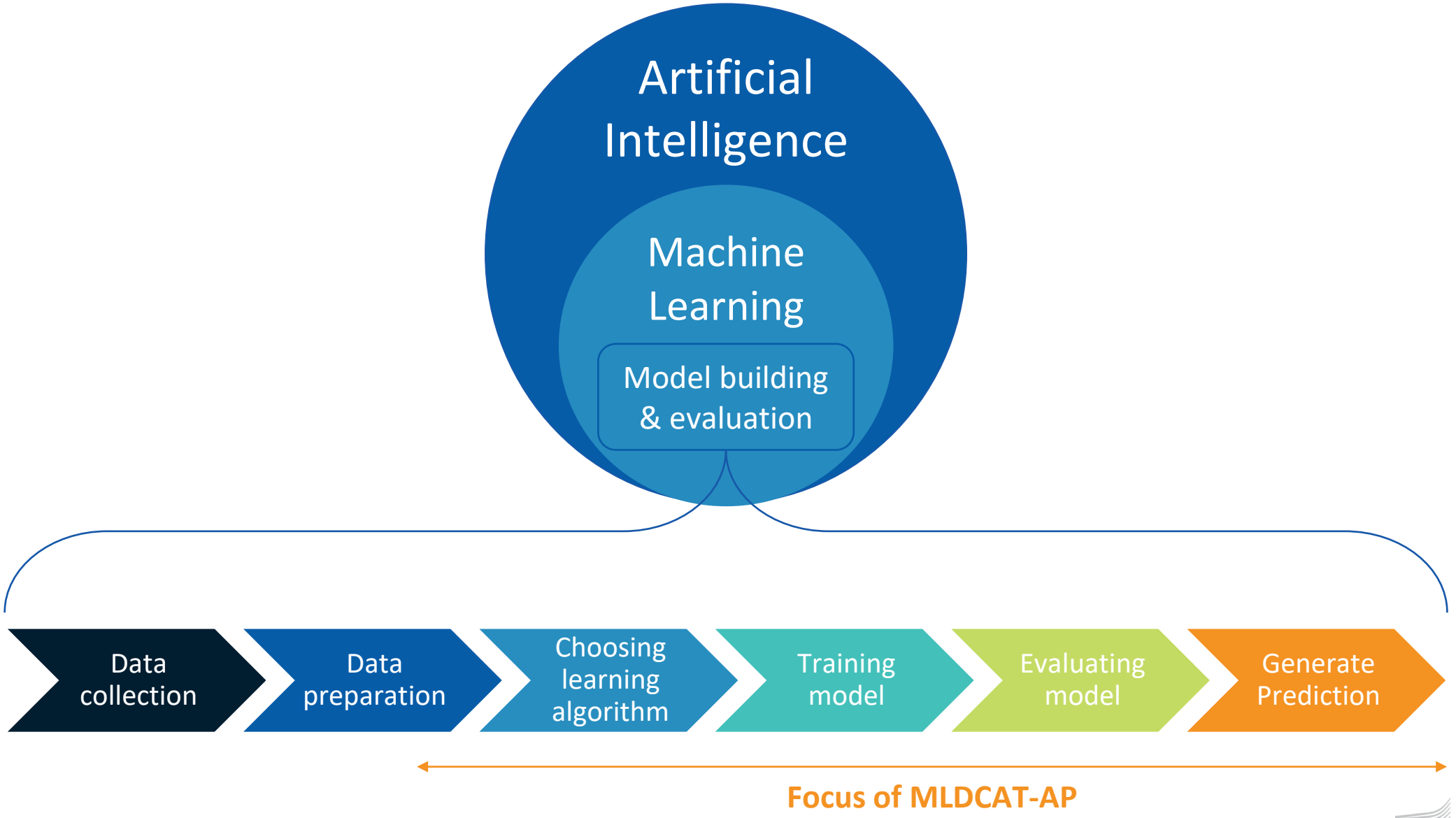
Describing ML Datasets with MLDCAT-AP yields all the advantages of using DCAT-AP.

Additional domain specific benefits are:

- Improved reproducibility
- Integration of RAI principles such as transparency and accountability
- Facilitates adherence to AI Act



Domain of MLDCAT-AP



Guest speaker: OpenML

European Commission

OpenML

MLDCAT-AP pilot

European Commission

Democratize Machine Learning

Democratize Machine Learning

Make ML research easily accessible and reusable…

Democratize Machine Learning

Make ML research easily accessible and reusable…

… by giving frictionless access to all ML experiment data

European Commission

OpenML

Democratize Machine Learning

Make ML research easily accessible and reusable…

… by giving frictionless access to all ML experiment data

Data

European Commission

OpenML

Democratize Machine Learning

Make ML research easily accessible and reusable…

… by giving frictionless access to all ML experiment data

Data

Models

OpenML

Democratize Machine Learning

Make ML research easily accessible and reusable…

… by giving frictionless access to all ML experiment data

Data

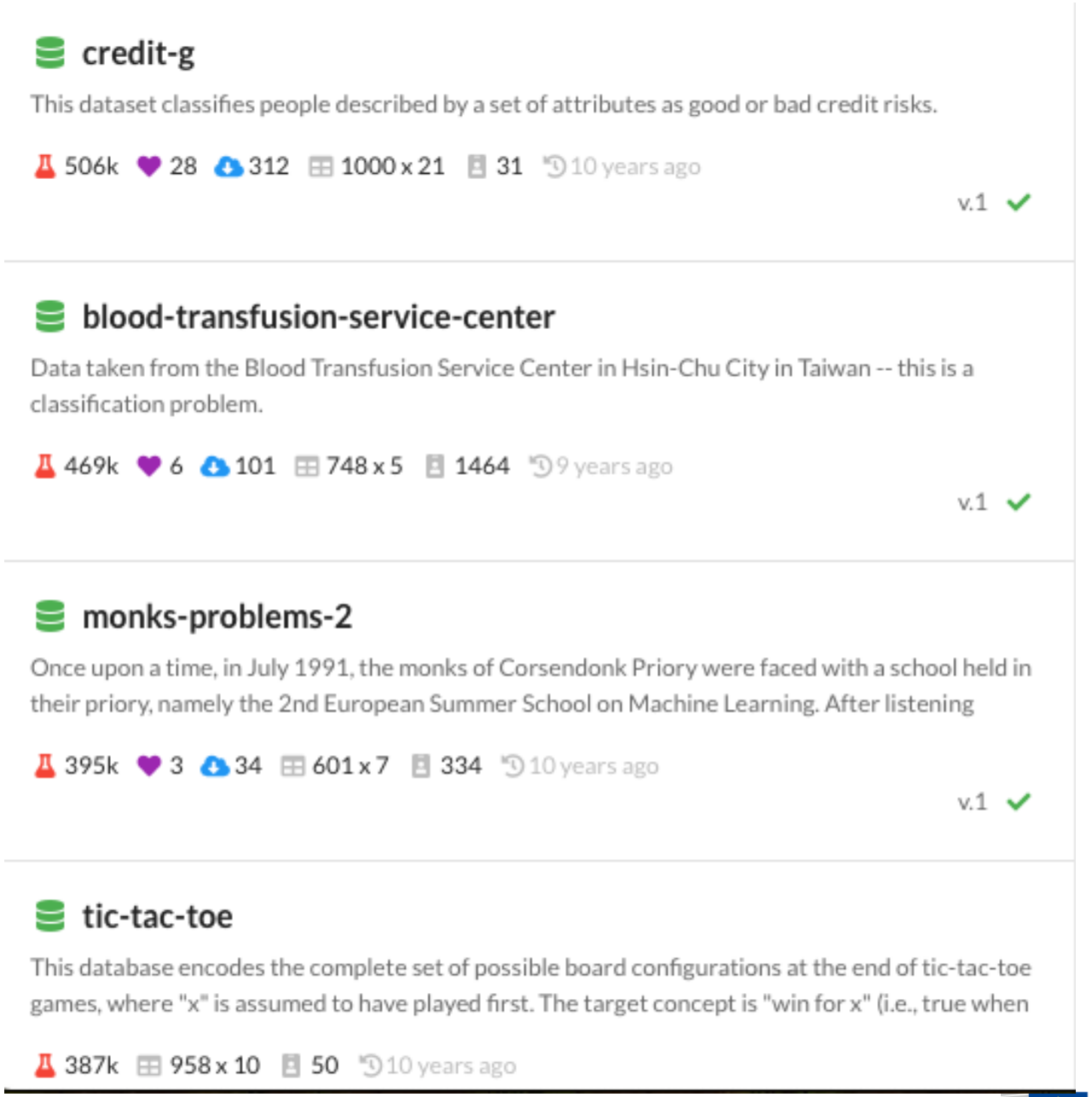
Models

Results

OpenML

Main Concepts:

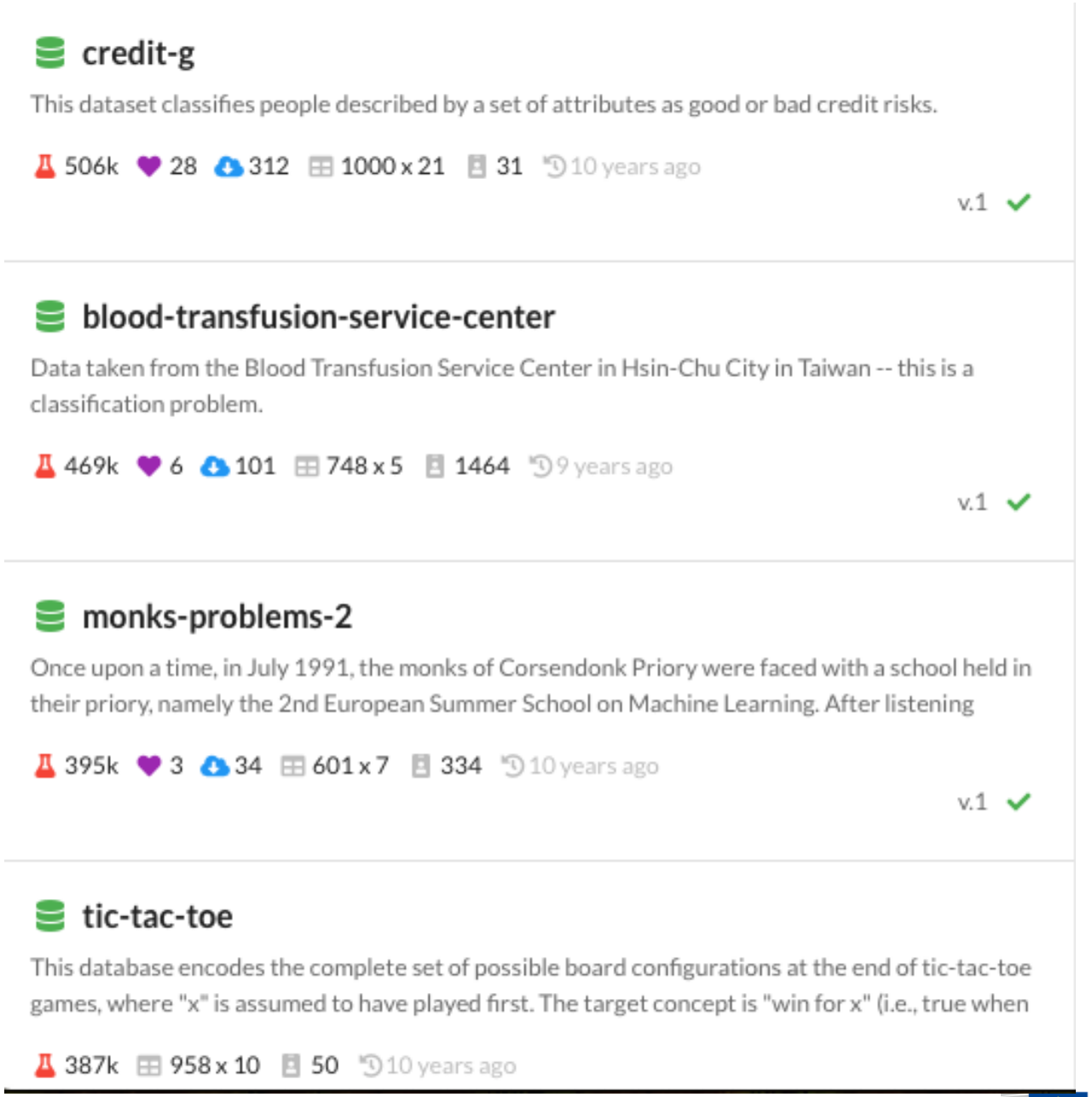
- Datasets
  - Tabular
  - Small-ish (~Gb scale)



OpenML

Main Concepts:

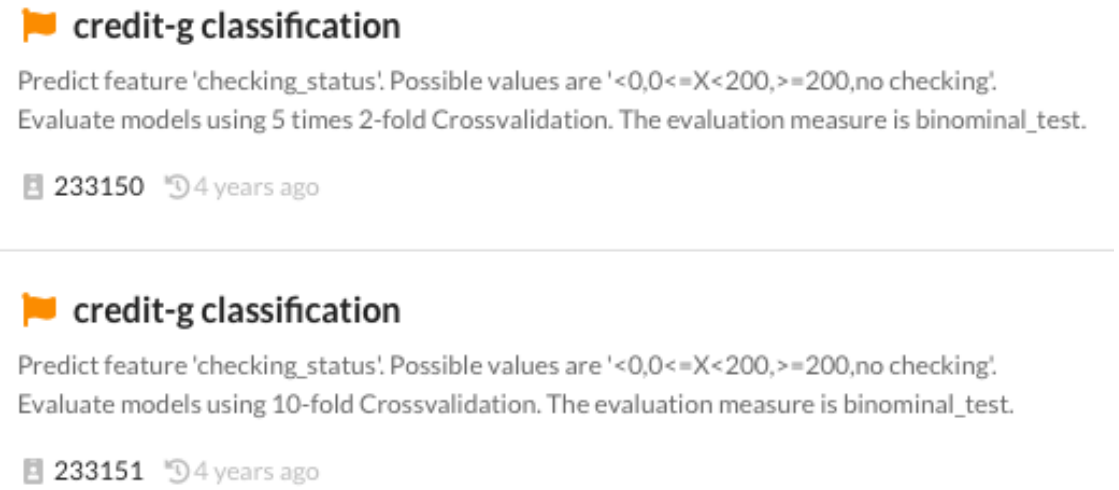
- Datasets
  - Tabular
  - Small-ish (~Gb scale)
  - Working on DL data



OpenML

Main Concepts:

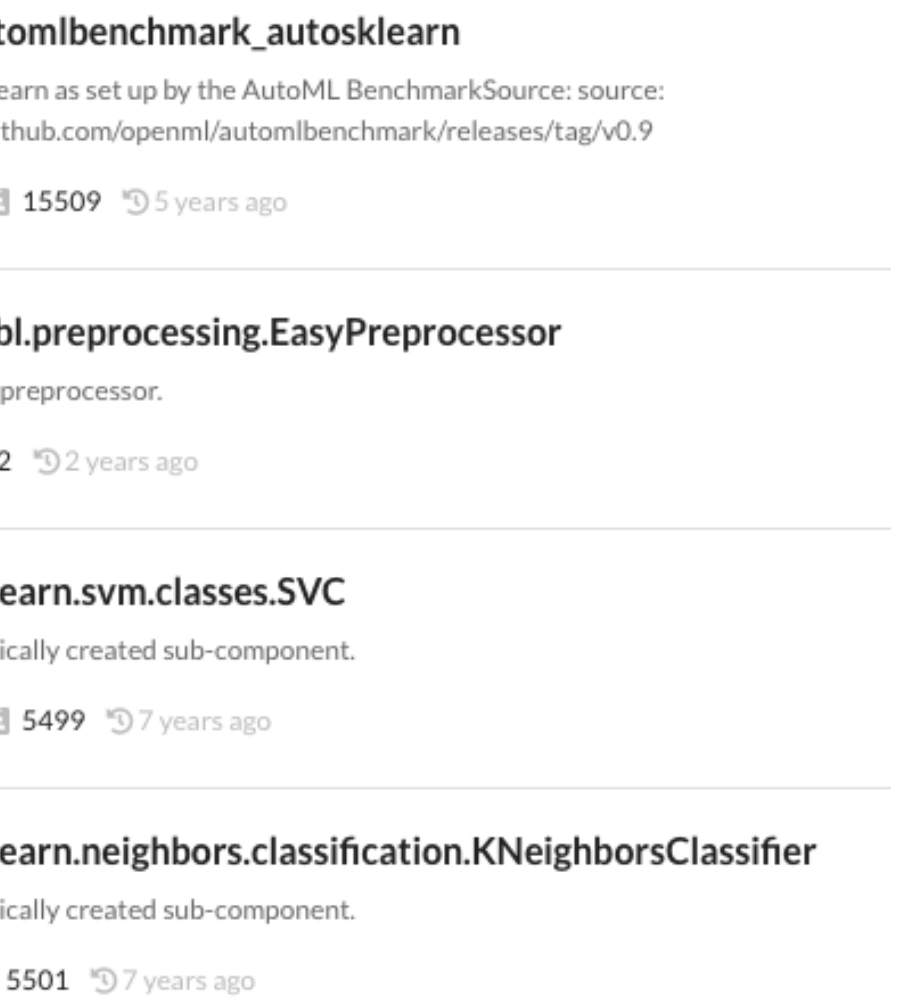
- Datasets
- Tasks
  - Evaluation Procedure
  - Splits
  - Target
  - Metric



OpenML

Main Concepts:

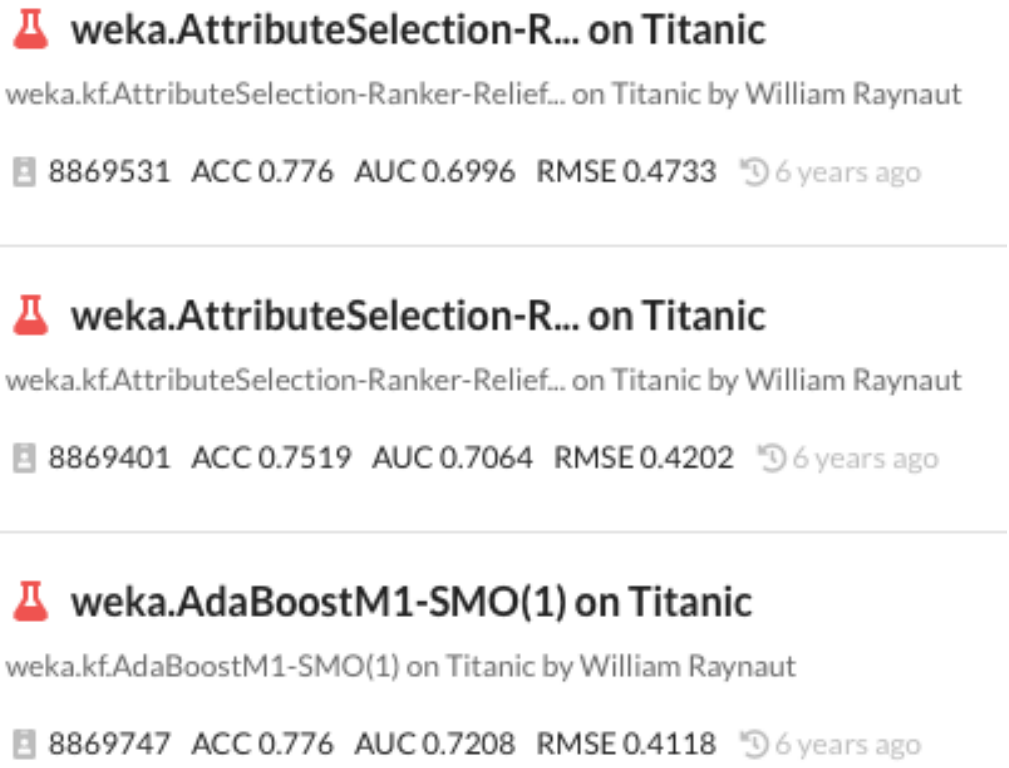
- Datasets
- Tasks
- Flows
  - Algorithm descriptions
  - Hyperparameters



OpenML

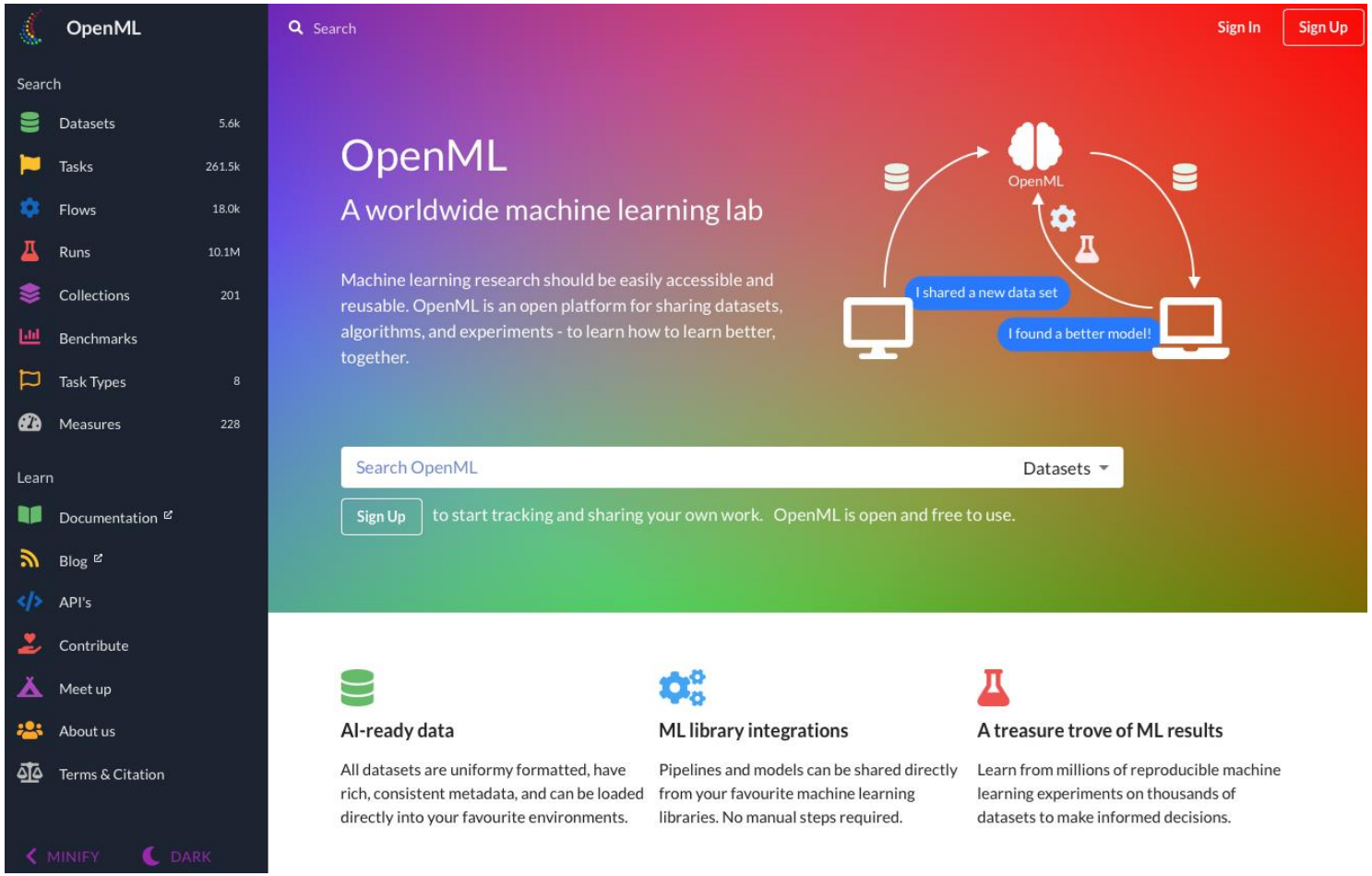
Main Concepts:

- Datasets
- Tasks
- Flows
- Runs
  - evaluation of a Flow on a Task
  - predictions
  - computed metrics

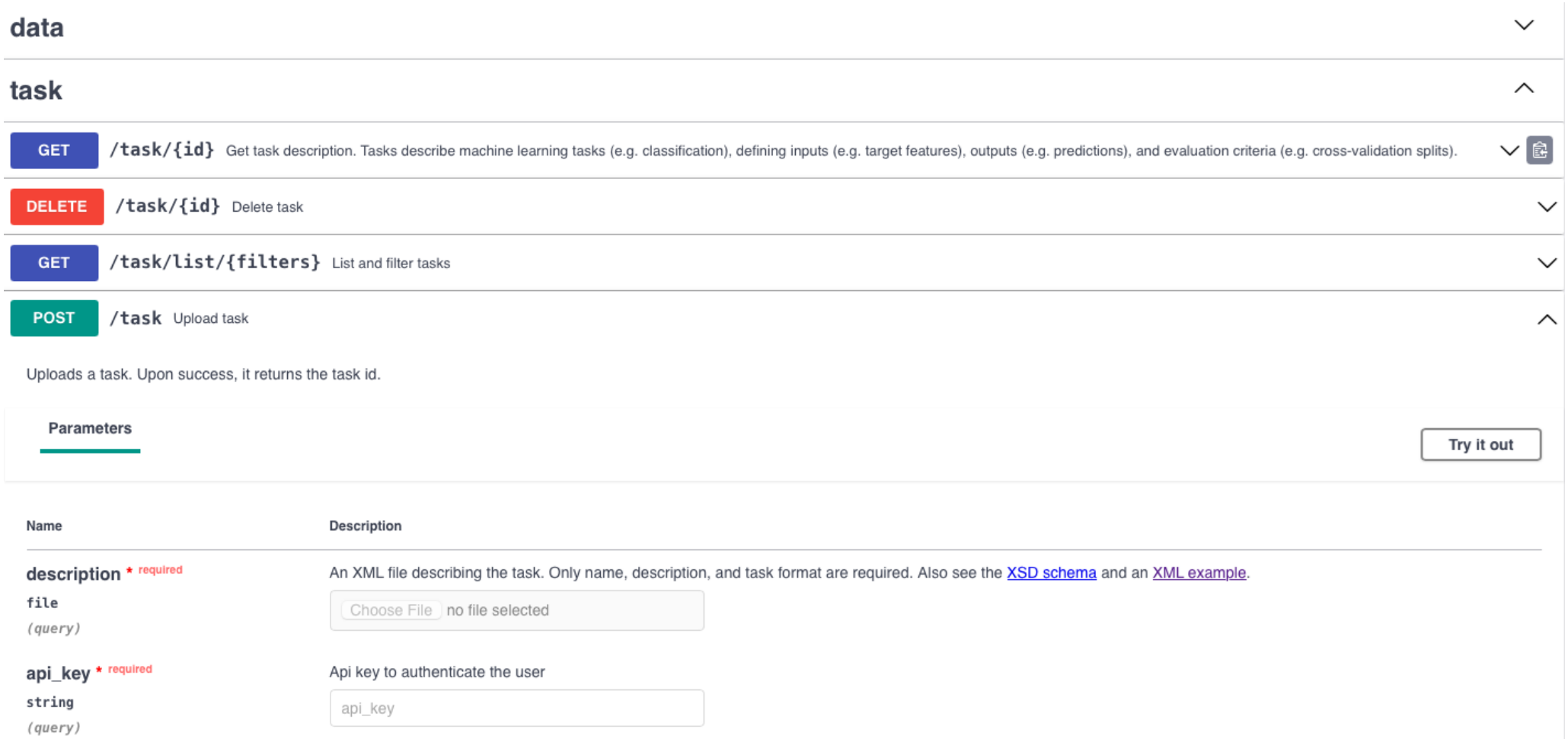


OpenML: Accessing Data

openml.org



OpenML: Accessing Data



OpenML: Accessing Data

```
import openml
iris = openml.datasets.get_dataset("iris")
```

OpenML - Why MLDCAT-AP?

- Custom Format = Learning Curve
- Interoperable Format =
  - No/Smaller Learning Curve

OpenML - Why MLDCAT-AP?

- Custom Format = Learning Curve
- Interoperable Format =
  - No/Smaller Learning Curve
  - Interface with OpenML

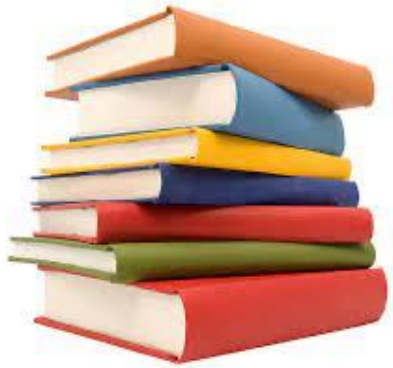
OpenML - Why MLDCAT-AP?

- Custom Format = Learning Curve
- Interoperable Format =
  - No/Smaller Learning Curve
  - Interface with OpenML
  - Complement other platforms

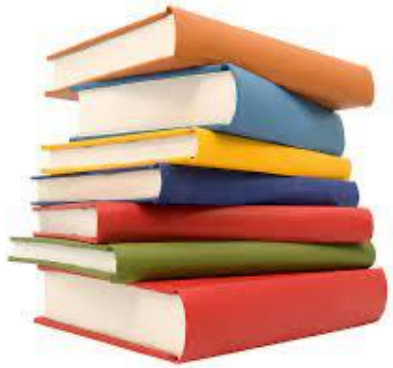
OpenML - MLDCAT-AP Process

European Commission

OpenML - MLDCAT-AP Process



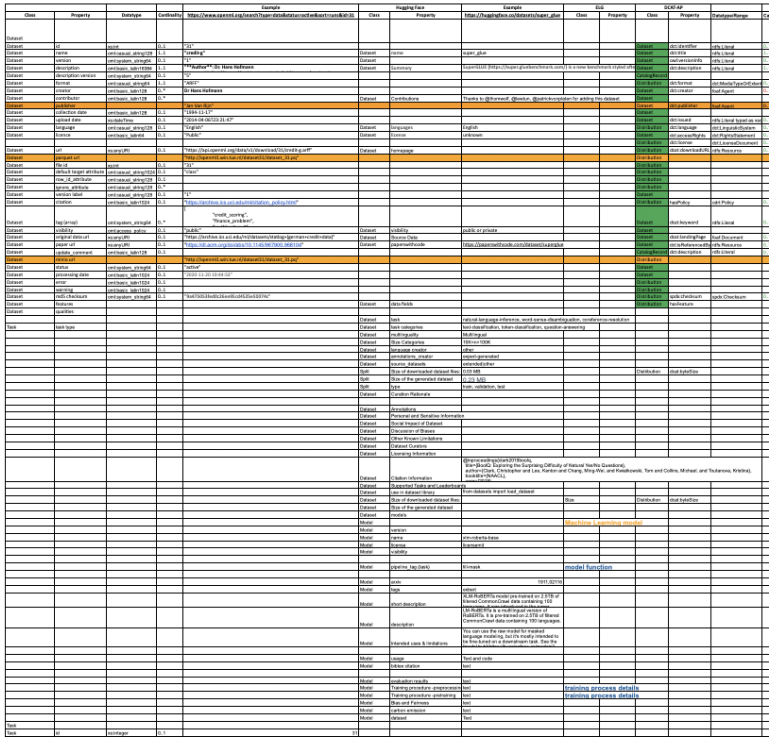
OpenML - MLDCAT-AP Process



{“foo”:“bar”}

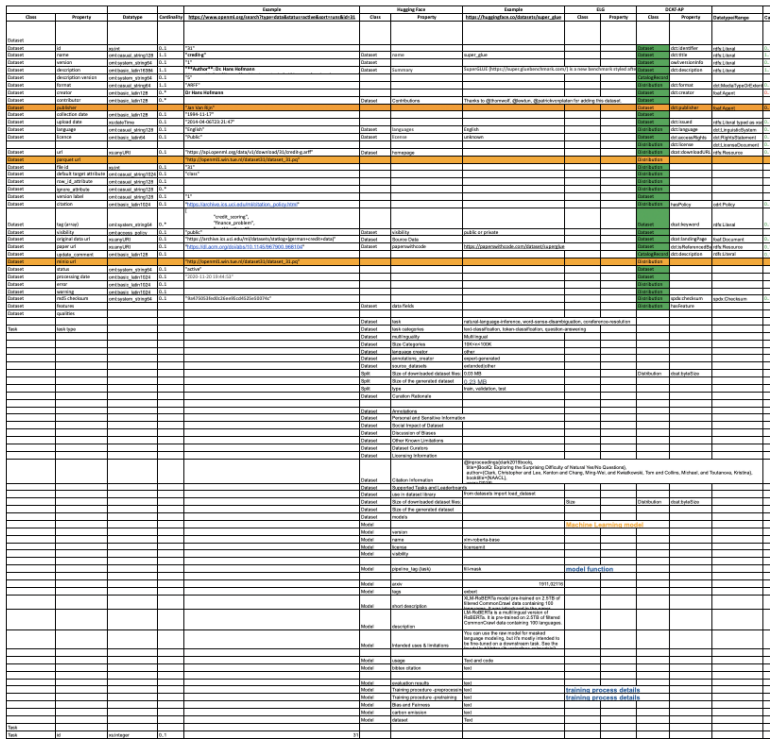
OpenML - MLDCAT-AP Process

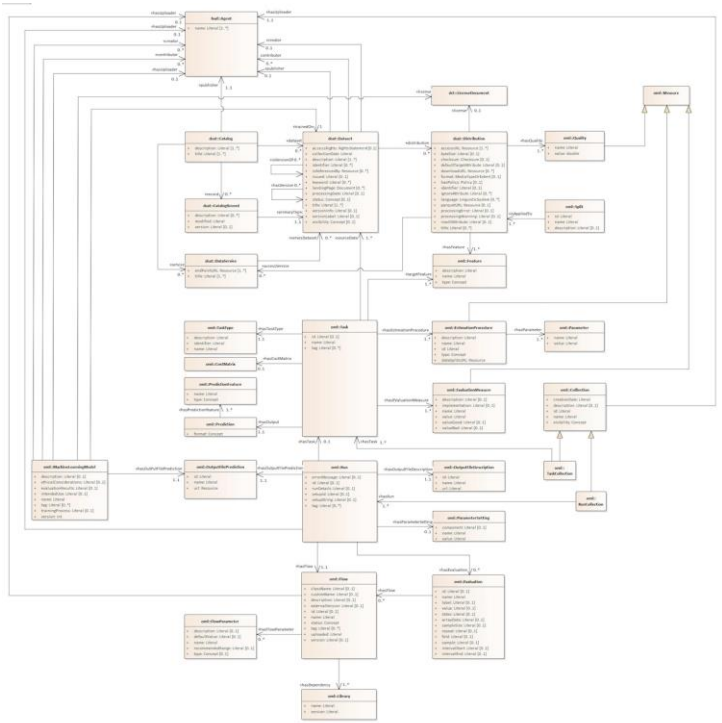
{“foo”:“bar”}



OpenML - MLDCAT-AP Process

{“foo”:“bar”}





MLDCAT-AP

OpenML Dataset	DCAT-AP Class	Property
id	Dataset	identifier
name	Dataset	title
tag	Dataset	keyword
status	Dataset	?
quality	?	?

MLDCAT-AP

OpenML Dataset	DCAT-AP Class	Property
id	Dataset	identifier
name	Dataset	title
tag	Dataset	keyword
status	Dataset	✨ status
qualities	✨ Quality	

MLDCAT-AP

Qualities describe the *data values:*

- simple: number of rows, number of classes, number of nominal features

MLDCAT-AP

Qualities describe the *data values:*

- simple: number of rows, number of classes, number of nominal features
- statistical: mean of means, skewness

MLDCAT-AP

Qualities describe the *data values:*

- simple: number of rows, number of classes, number of nominal features
- statistical: mean of means, skewness
- information theoretic: information gain

MLDCAT-AP

Qualities describe the *data values:*

- simple: number of rows, number of classes, number of nominal features
- statistical: mean of means, skewness
- information theoretic: information gain
- landmarking: based on small ML models

MLDCAT-AP

§ 7.9 Quality Measurement

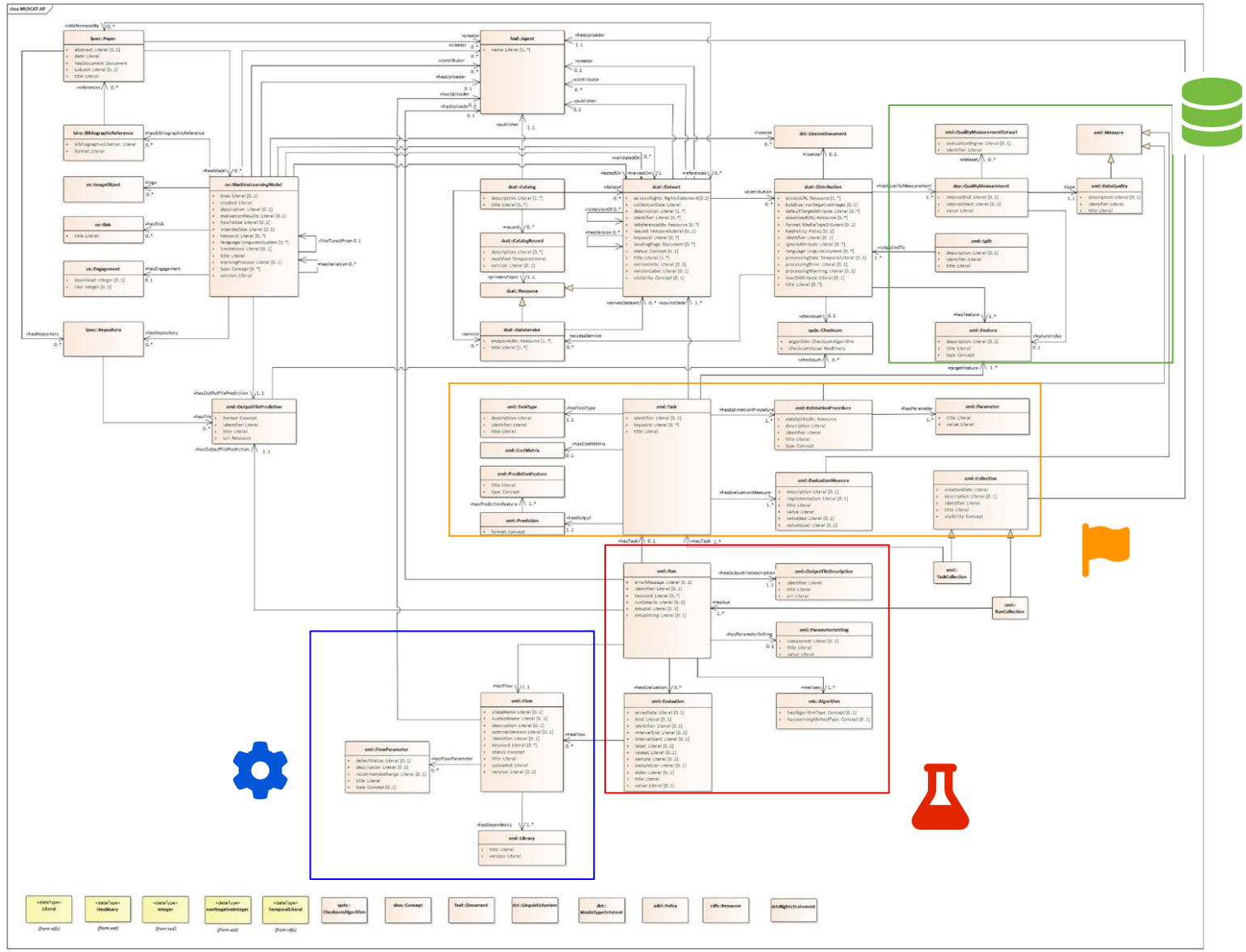
Definition

Represents the evaluation of a given dataset (or dataset distribution) against a specific quality metric.

Properties

For this entity the following properties are defined: dataset , feature index , interval end , interval start , type , value .

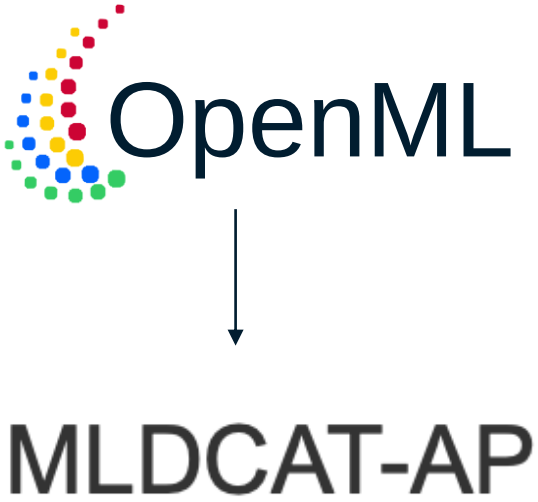
Property	Range	Card	Definition	Usage
dataset	Quality Measurement Dataset	0..*	Indicates the data set of which this observation is a part.	
feature index	Feature	0..1	The feature on which the quality is measure on.	
interval end	Literal	0..1		
interval start	Literal	0..1		
type	Data quality	1	A classification for a quality.	
value	Literal	1	Refers to values computed by metric.	



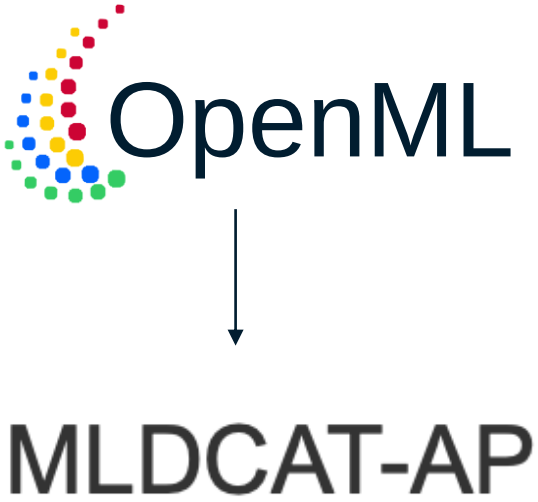
European Commission

Integration

Integration



Integration



Implementation

https://test.openml.org/mldcatap/docs

Roll out together with new API later this year

OpenML pilot

Stakeholders

OpenML
Technical University of Eindhoven

SEMIC community

Input

Hugging Face
ONNX

OpenML API

DCAT-AP FOR DATA PORTALS IN EUROPE

Schema.org

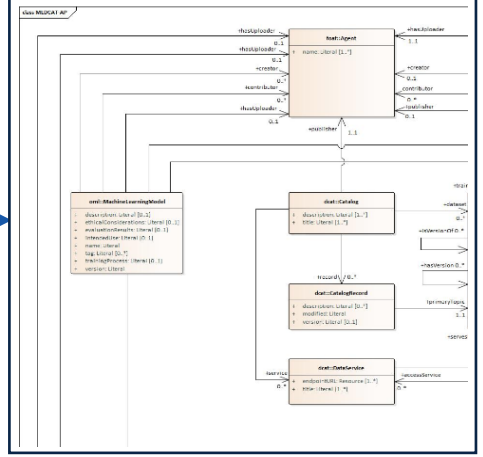
Mapping

OpenML		DCAT-AP	
Class	Property	Class	Property
Dataset			
Dataset	id	Dataset	dct:identifier
Dataset	**name**	Dataset	dct:title
Dataset	version	Dataset	owl:versionInfo
Dataset	**description**	Dataset	dct:description
		Dataset	dct:creator
Dataset	description version	Dataset	
Dataset	format	Distribution	dct:format
Dataset	publisher	Dataset	dct:publisher

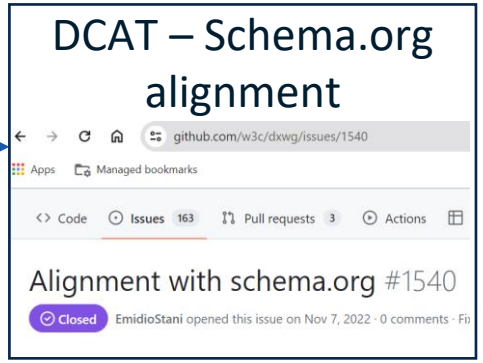
URI	owl:equivalentClass	rdfs:subClassOf
sdo:DataCatalog		dcat:Catalog
sdo:dataset		
sdo:measurementTechnique		
sdo:CreativeWork		
sdo:about		
sdo:abstract		

Output

MLDCAT-AP 1.0



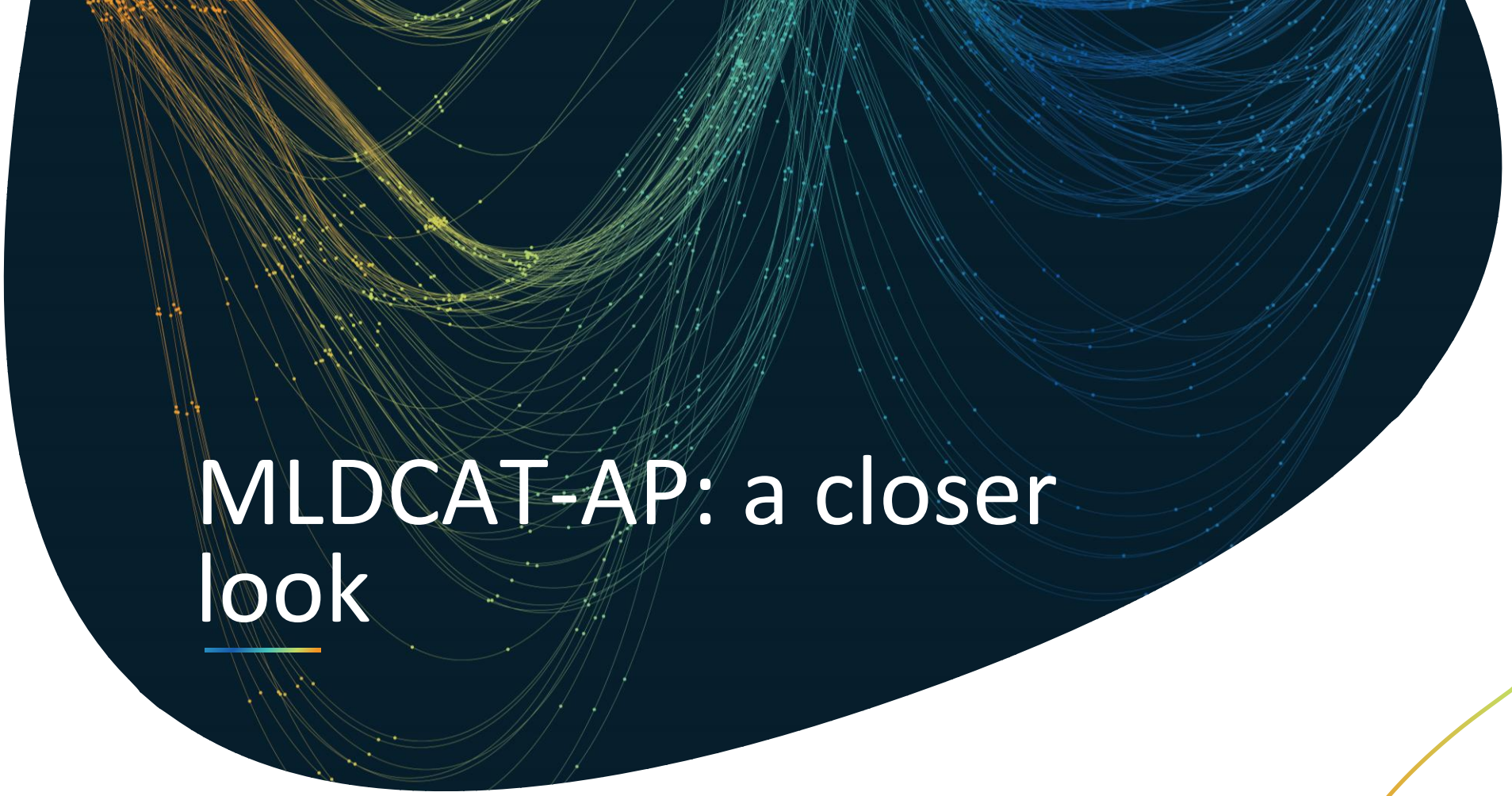
DCAT – Schema.org alignment



Nov 2022

Today

European Commission



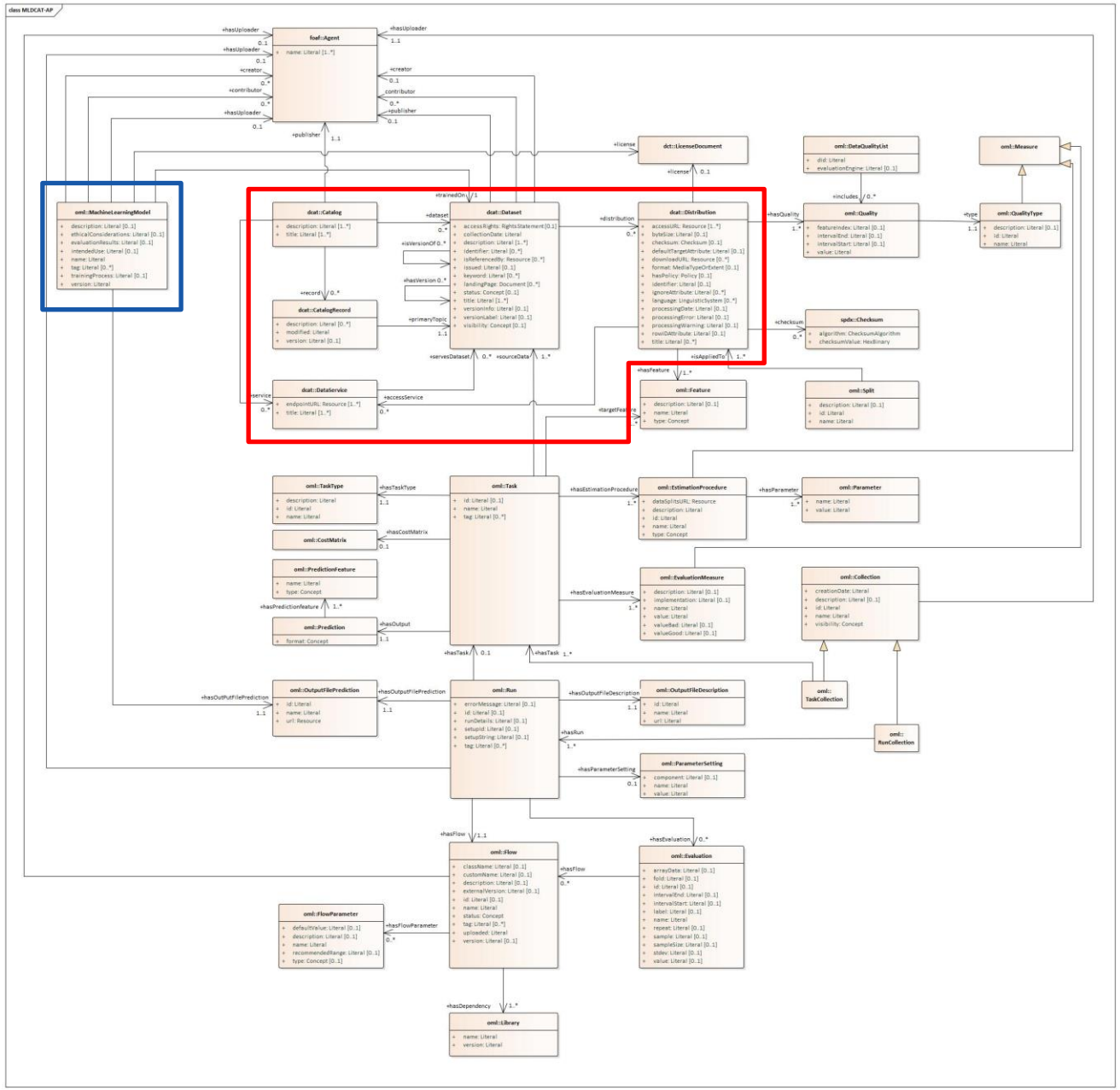
MLDCAT-AP: a closer look

European Commission

MLDCAT-AP 1.0.0

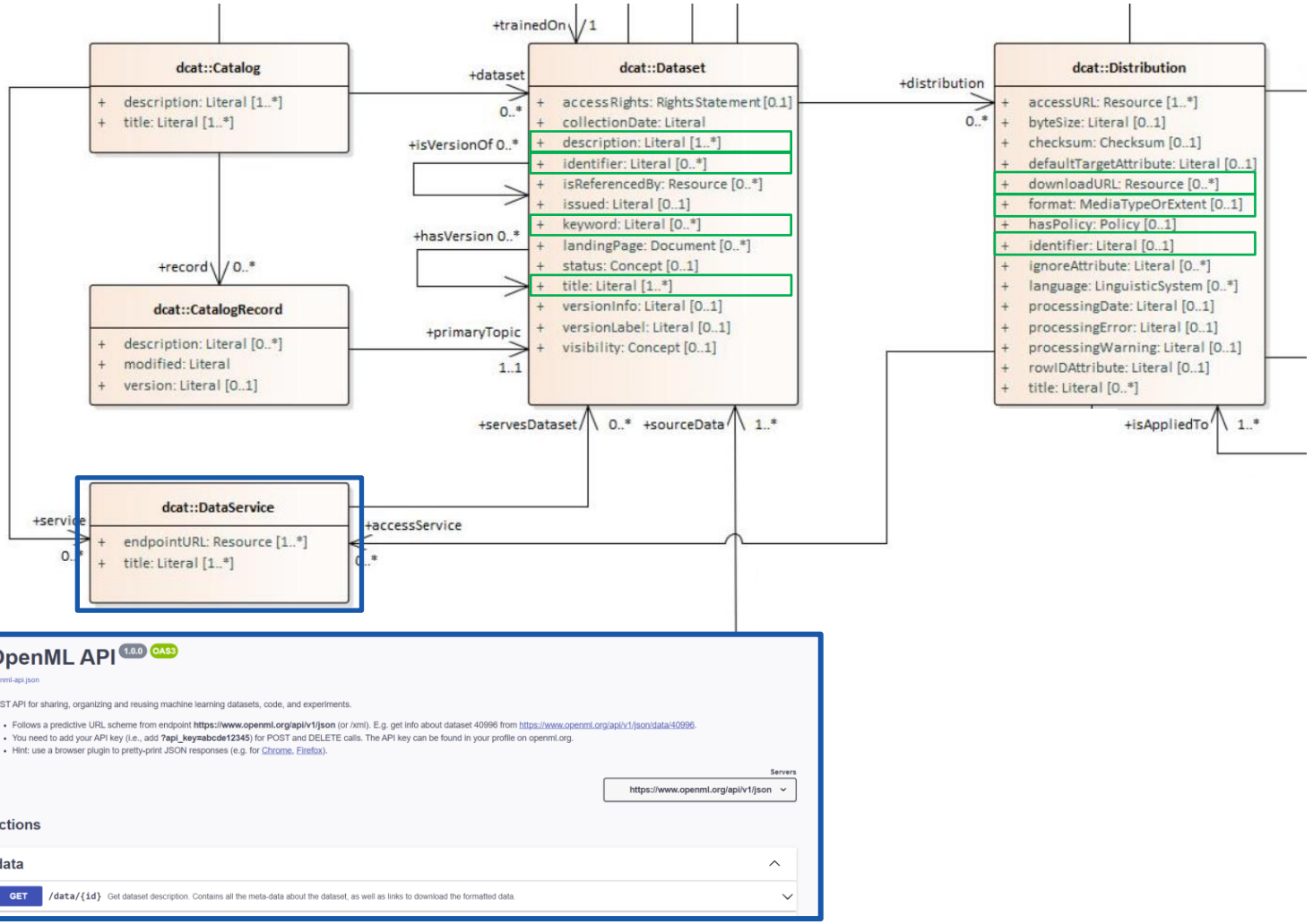
DCAT-AP

MachineLearningModel

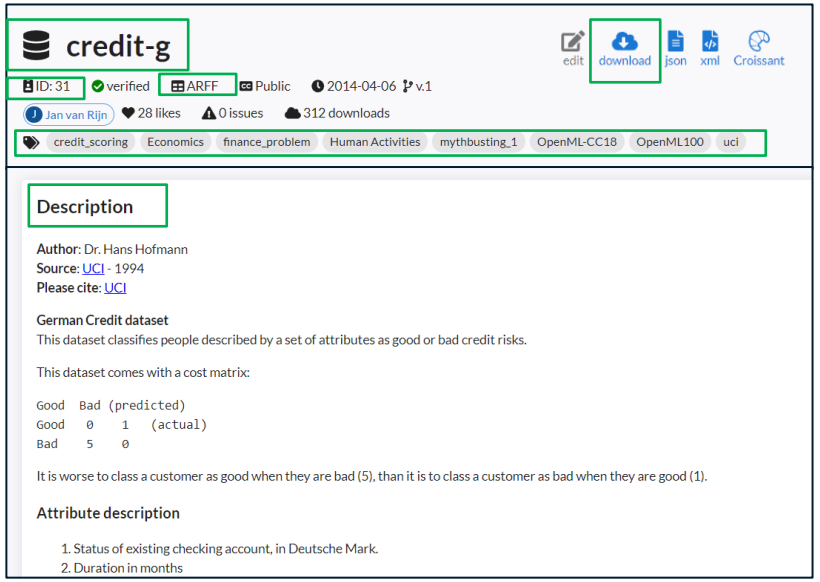


European Commission

MLDCAT-AP 1.0.0 – Reusing DCAT-AP concepts



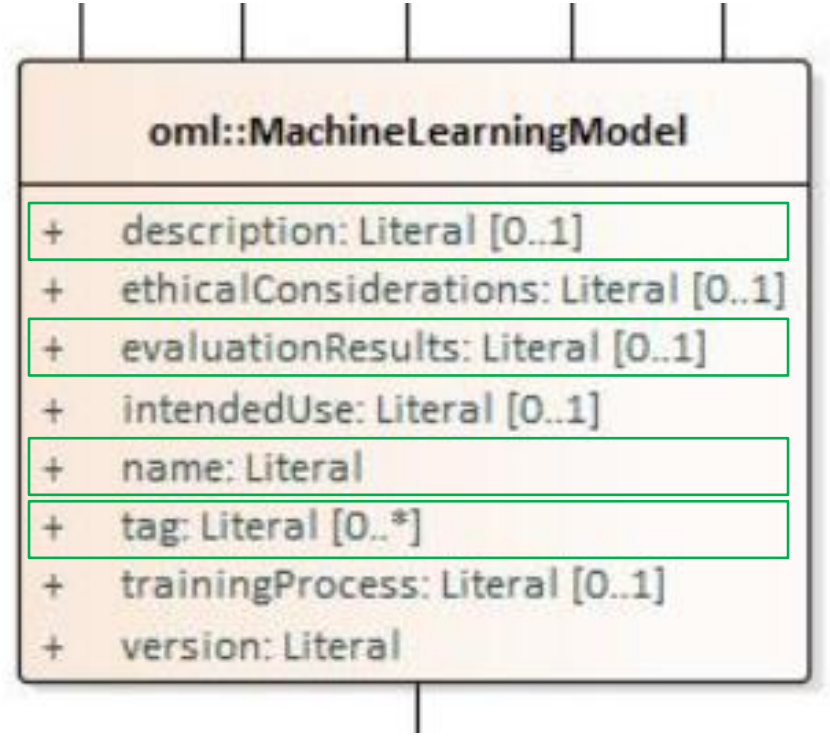
OpenML API

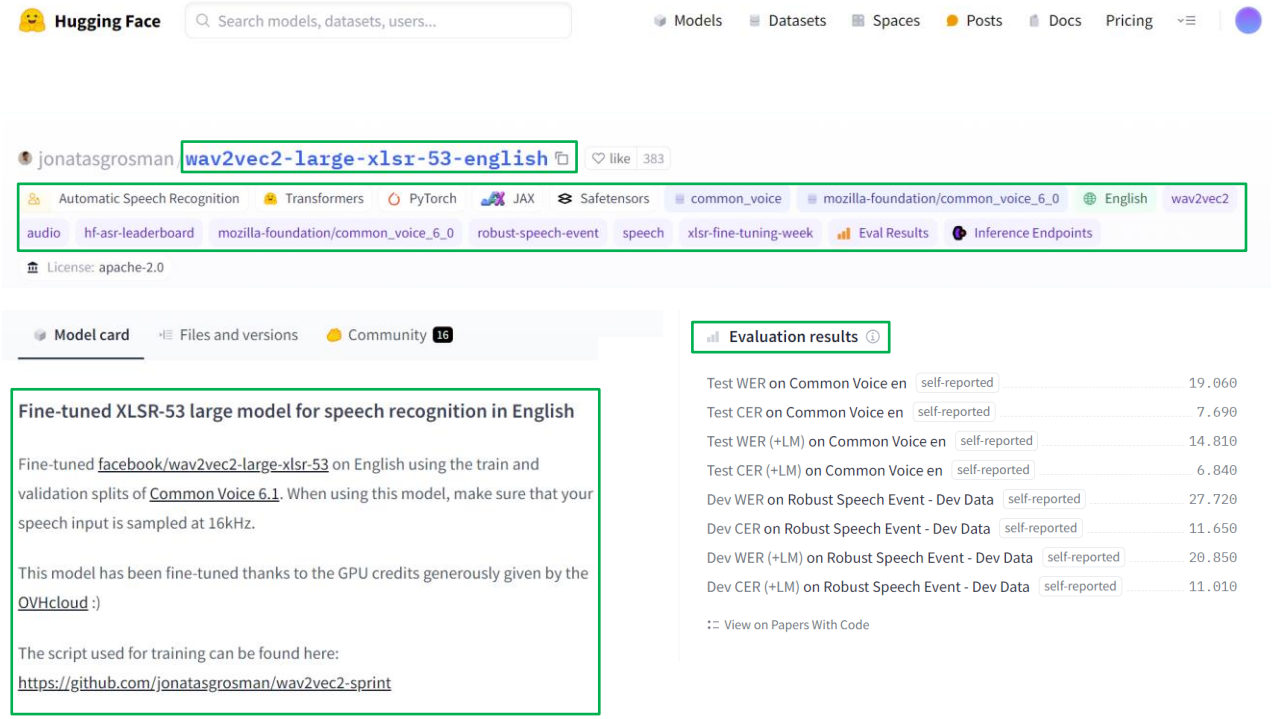


OpenML dataset

MLDCAT-AP 1.0.0 - MachineLearningModel

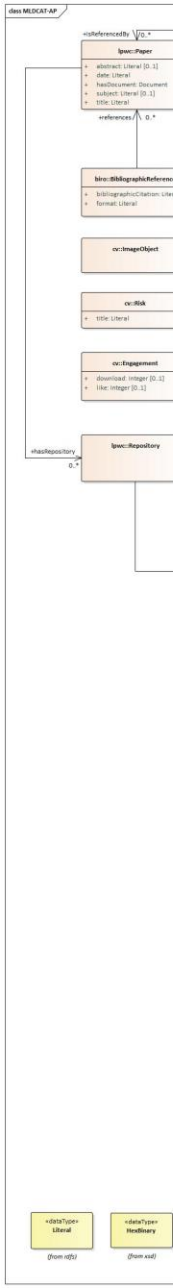
- Based on Hugging Face and ONNX





MLDCAT-AP 2.0.0

- Comparative analysis between repositories of machine learning models
- Inclusion of Papers related to machine learning models
- Inclusion of Algorithm executed during the model building
- Focus on data quality and risk in view of the AI Office and AI Act
- Published in ReSpec HTML format



MLDCAT-AP

14 February 2024

▼ **More details about this document**

Latest published version:
https://semiceu.github.io/MLDCAT-AP/releases/1.0.0/

Latest editor's draft:
https://semiceu.github.io/MLDCAT-AP/releases/2.0.0

History:
Commit history

Editor:
Emidio Stani (PwC EU Services)

Feedback:
GitHub SEMICeu/MLDCAT-AP (pull requests, new issue, open issues)

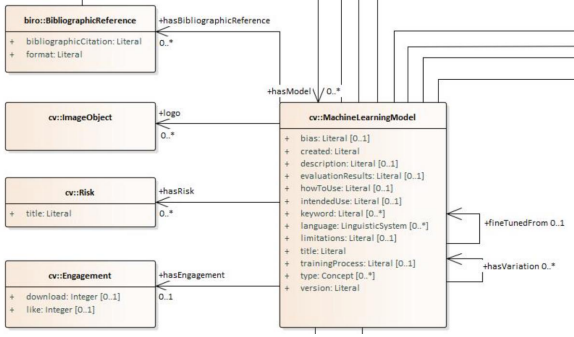
Abstract

MLDCAT-AP aims to describe machine learning models, together with their datasets, quality measured on the datasets and citing papers. It has been originally developed in collaboration with OpenML.

§ 1. Introduction

MLDCAT-AP (Machine Learning DCAT-AP) is an application profile that extends DCAT-AP in the field of machine learning.

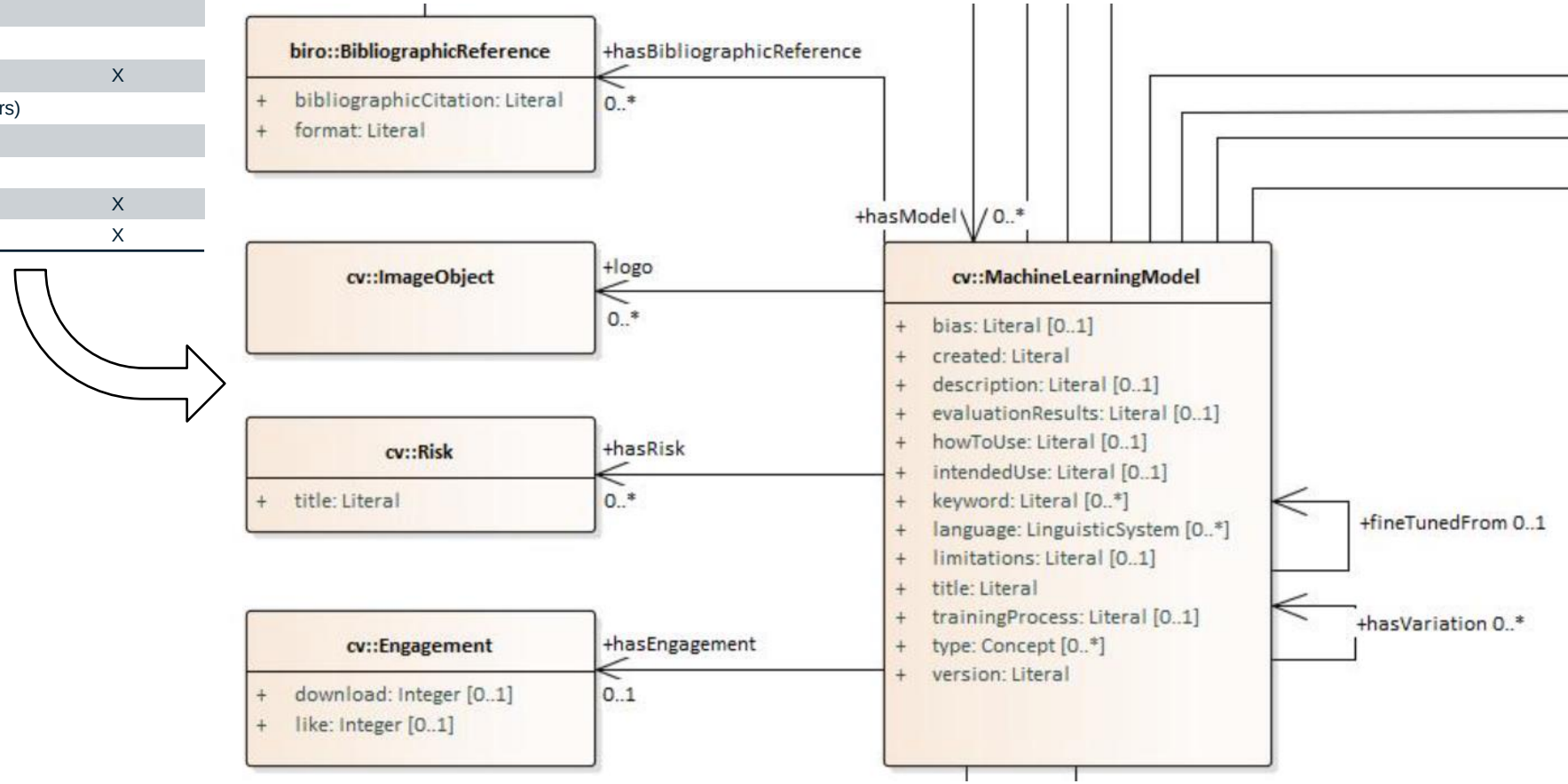
MLDCAT-AP 2.0.0 – MachineLearningModel



Common Properties	Hugging Face 423K models	Kaggle -TensorFlow 278-2K	Pytorch 55	AzureAI 100+ (in org)
Language	X	X	X	X
References to Paper/Code	X	X	X	
How to use	X	X	X	
Logo	X	X	X	
Files	X	X		X
Likes	X	X	X (GitHub stars)	
Downloads	X	X		
Relation	X (fine tuned from)	X (variation)		
Created date	X			X
Checksum	X			X

MLDCAT-AP 2.0.0 – MachineLearningModel

Common Properties	Hugging Face 423K models	Kaggle -TensorFlow 278-2K	Pytorch 55	AzureAI 100+ (in org)
Language	X	X	X	X
References to Paper/Code	X	X	X	
How to use	X	X	X	
Logo	X	X	X	
Files	X	X		X
Likes	X	X	X (GitHub stars)	
Downloads	X	X		
Relation	X (fine tuned from)	X (variation)		
Created date	X			X
Checksum	X			X



MLDCAT-AP 2.0.0 – Algorithm

Commission Decision Establishing the European AI Office

Download

In particular, the Office shall

(a) work with other relevant Directorate-Generals and services of the Commission in the performance of its tasks pursuant to Article 2, notably with the European Centre for Algorithmic Transparency as regards the evaluation and testing of general-purpose AI models and systems;

ANNEX I
ARTIFICIAL INTELLIGENCE TECHNIQUES AND APPROACHES
referred to in Article 3, point 1

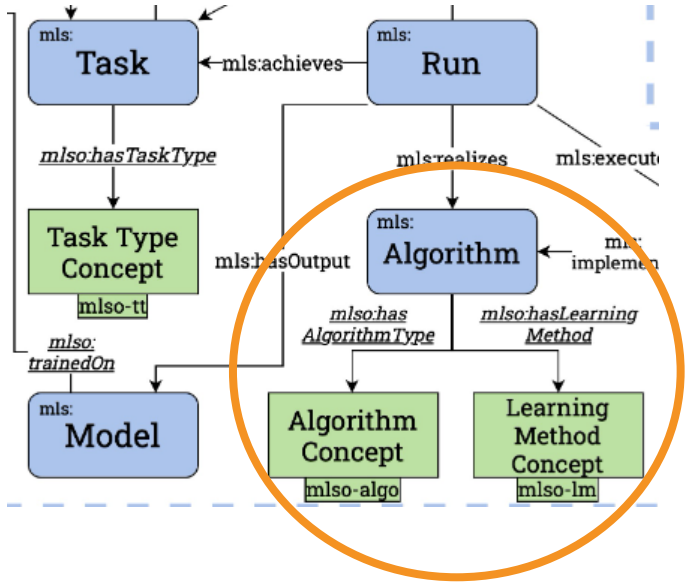
(a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;

(b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;

(c) Statistical approaches, Bayesian estimation, search and optimization methods.

MLDCAT-AP 2.0.0 – Algorithm

MLSO

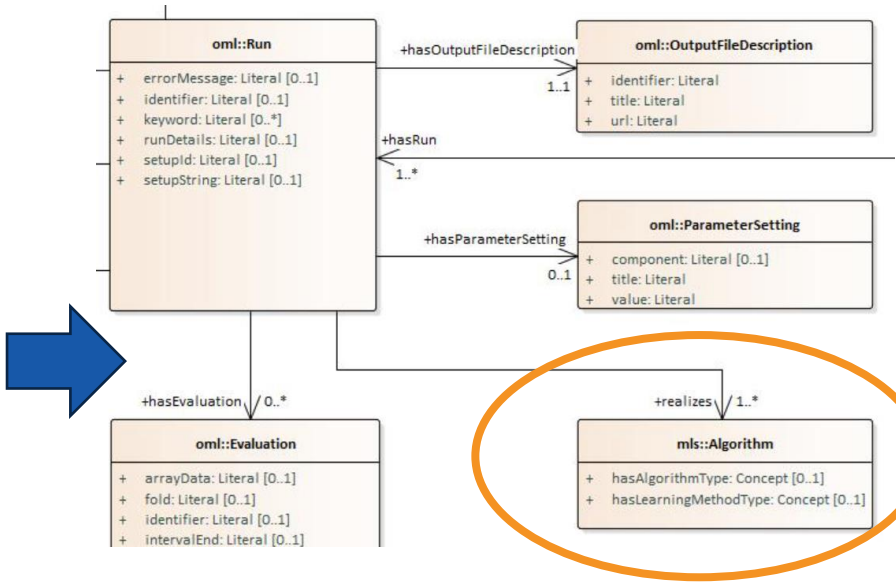


Machine Learning Algorithm

:ArtificialNeuralNetwork
:AssociationRuleLearningAlgorithm
:**Bayesian**
:ClusteringAlgorithm
:DecisionTree
:**DeepLearningAlgorithm**
:DimensionalityReductionAlgorithm
:EnsembleAlgorithm
:InstanceBasedAlgorithm
:RegressionAlgorithm
:RegularizationAlgorithm
:**ReinforcementLearningAlgorithm**
:RuleBased

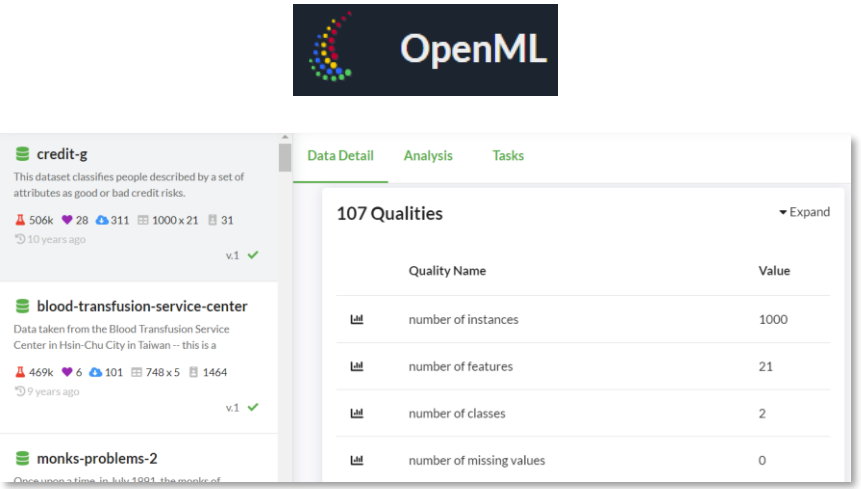
Learning method

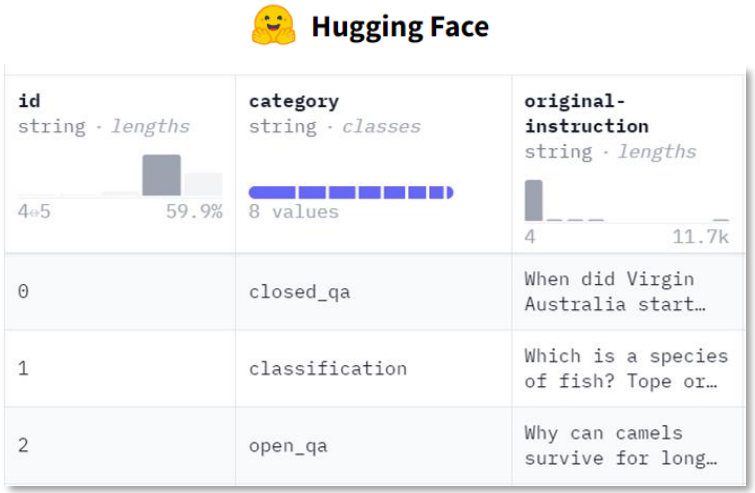
:ActiveLearning
:AdversarialLearning
:AnalyticalLearning
:**BayesianLearning**
:ConceptLearning
:CurriculumLearning
:EnsembleLearning
:FederatedLearning
:FewShotLearning
:IncrementalLearning
:InductiveLearning
:MetaLearning
:MetricLearning
:MultiModalLearning
:MultiTaskLearning
:OnlineLearning
:**Reinforcement_Learning_Algorithm**
:SelfSupervisedLearning
:SelfTaughtLearning
:Semi-supervised_Learning_Algorithm
:SequentialLearning
:**Supervised_Learning_Algorithm**
:TargetedLearning
:TransferLearning
:**Unsupervised_Learning_Algorithm**

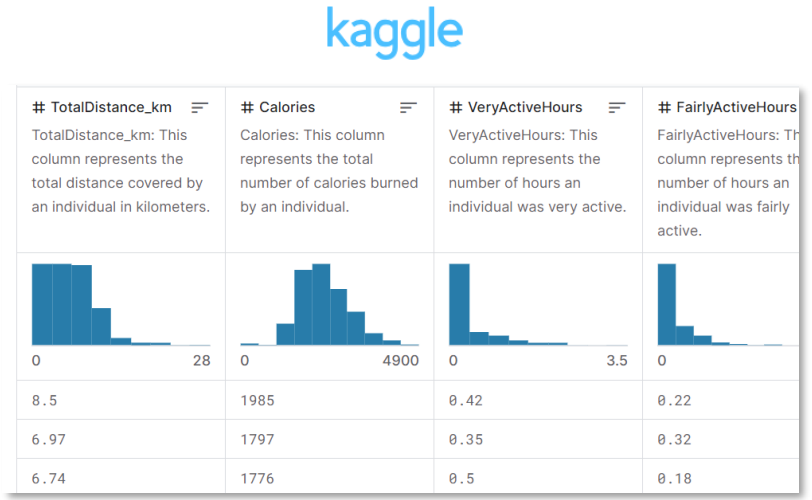


*In **bold** those mentioned by the AI ACT

MLDCAT-AP 2.0.0 – Quality

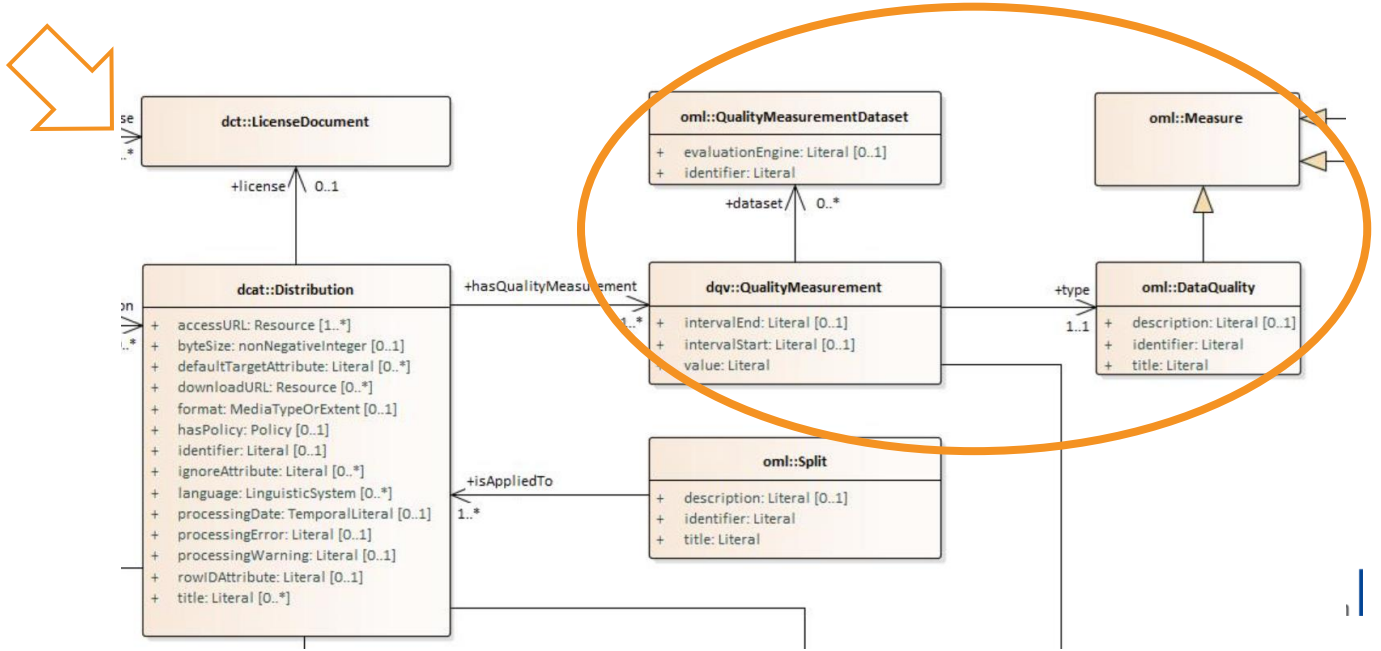






AI Act

For high-risk AI systems, the requirements of **high quality data**, documentation and traceability, transparency, human oversight, accuracy and robustness, are strictly necessary to mitigate the risks to fundamental rights and safety posed by AI and that are not covered by other existing legal frameworks.



MLDCAT-AP 2.0.0 – Paper

376,557
Papers

153,476
Repositories

52,519
Evaluations

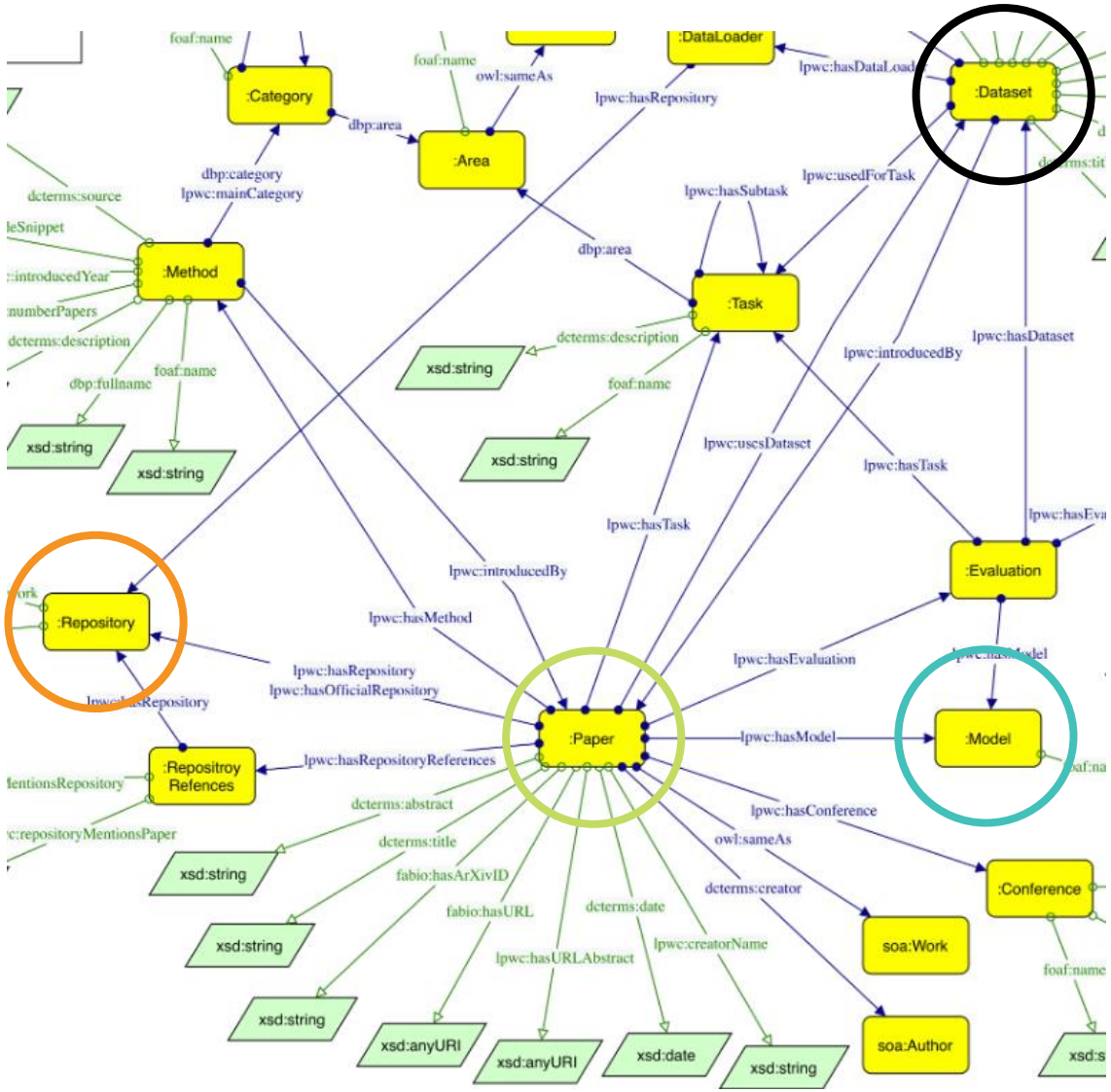
24,598
Models

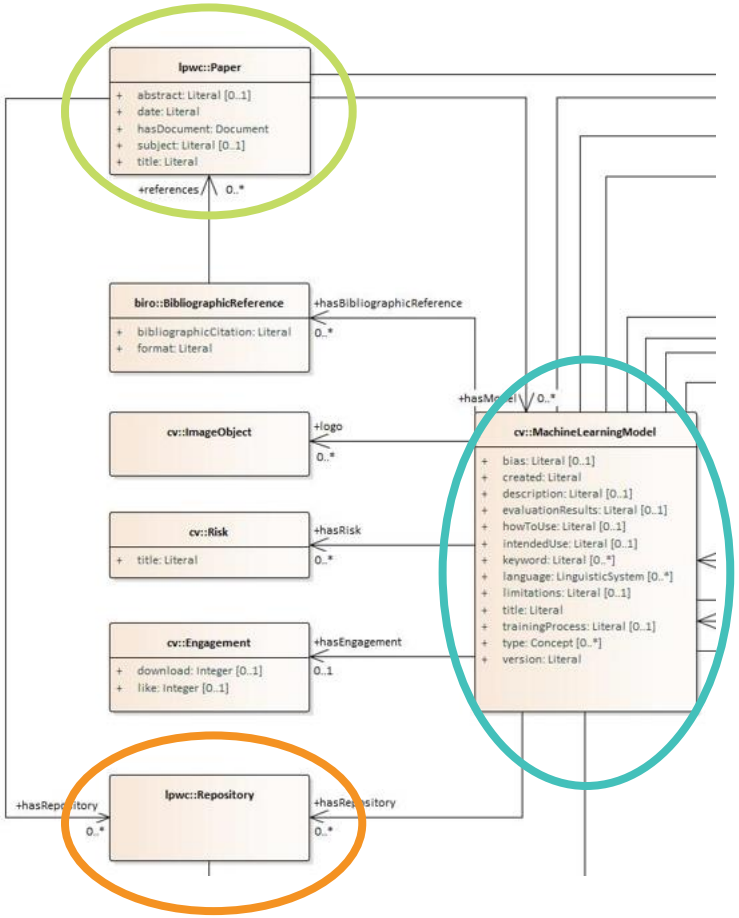
8,322
Datasets

4,267
Tasks

2,101
Methods

1,407
Conferences





LPWC

European Commission

MLDCAT-AP 2.0.0 – Controlled Vocabularies

MLDCAT-AP 1.0.0

Uses 3 Controlled Vocabularies published by OP. For example:

- Dataset.accessRights: Access Rights Named Authority List
- Distribution.format: EU Vocabularies File Type Named Authority List
- Distribution.language: EU Vocabularies Languages Named Authority List

Introduces 9 newly created Controlled Vocabularies. For example:

- Dataset.status: *Active, Deactivated, In Preparation*
- Dataset.visibility: *Public, Private*

MLDCAT-AP 2.0.0

Controlled Vocabularies published by OP can be reused.

Controlled Vocabularies that were newly introduced will have to be recreated and published by OP.

New Controlled Vocabularies necessary for:

- MachineLearningModel.type
- Algorithm.hasAlgorithmType
- Algorithm.hasLearningMethodType

Questions

European Commission

Next steps

European Commission

Next steps

- Monitor AI Act
- Process feedback received
- Continue building the ML community


Thank you

interoperable europe

innovation ∞ govtech ∞ community

Stay in touch

(@InteroperableEU) / Twitter

Interoperable Europe - YouTube

Interoperable Europe | LinkedIn

DIGIT-INTEROPERABILITY@ec.europa.eu

https://joinup.ec.europa.eu/collection/interoperable-europe/interoperable-europe