FAIR multi-label classification

FAIR MLC data representation and repositories (Part II)

Kocev Dragi, Bogatinovski Jasmin, **Kostovska Ana**,

Panov Pance

ECML PKDD 2021 Tutorial

17.09.2021

Outline

- Overview of existing MLC repositories
- Semantic web technologies in the context of FAIR data
- Semantic annotation of MLC datasets
- Web-based semantic repository of MLC datasets
- Evaluation of FAIRness

MLC repositories

Dataset	Domain	m	d	q	Card	Dens	Div	avgIR	rDep	m×q×d	Original dataset / Mulan
20NG	Text	19300	1006	20	1.029	0.051	0.003	1.007	0.984	3.88E+08	
3s-bbc1000	Text	352	1000	6	1.125	0.188	0.234	1.718	0.733	2.11E+06	
3s-guardian1000	Text	302	1000	6	1.126	0.188	0.219	1.773	0.667	1.81E+06	
3s-inter3000	Text	169	3000	6	1.142	0.190	0.172	1.766	0.400	3.04E+06	

➔ Multi-Label Classification Dataset Repository by KDIS Research Group of the University of Córdoba (http://www.uco.es/kdis/mllresources/)

COMETA | Home | Browse datasets | Submit a dataset | Source code

Browse datasets

Find a dataset — 74 datasets

Click the name of a dataset to access information and downloads. Sort the table by clicking the corresponding header.

Name	Instances	Attributes	Inputs	Labels	Labelsets	Single	Max freq	Card	Dens	Mean IR	Scumble	TCS
bibtex	7395	1995	1836	159	2856	2199	471	2.4019	0.0151	12.4983	0.0938	20.5414
birds	645	279	260	19	133	73	294	1.014	0.0534	5.407	0.033	13.3955
bookmarks	87856	2358	2150	208	18716	14971	6087	2.0281	0.0098	12.308	0.0597	22.8479

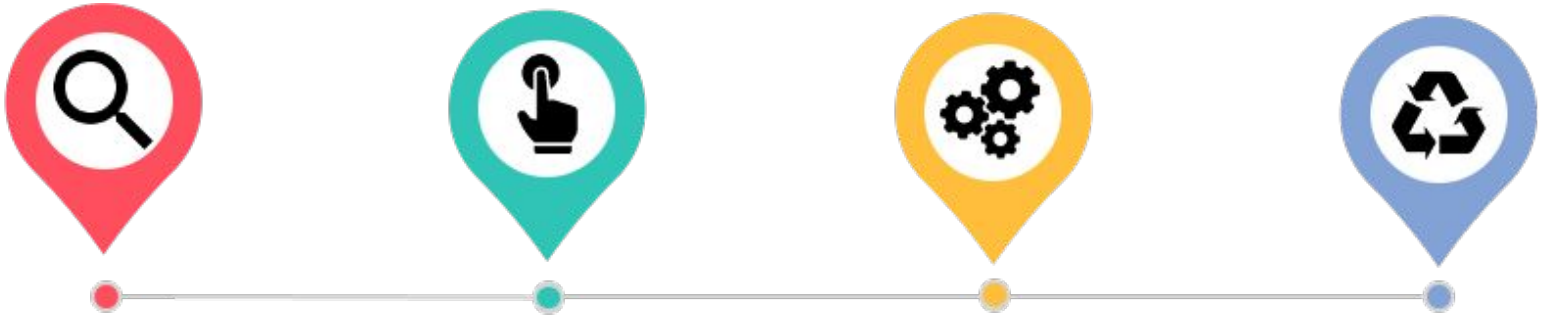
➔ COMETA (https://cometa.ujaen.es/datasets/)

Dataset statistics & download

Dataset	Download	BoW Feature Dimensionality	Number of Labels	Number of Train Points	Number of Test Points	Avg. Points per Label	Avg. Labels per Point	Original Source
LF-AmazonTitles-131K	BoW Features Raw text	40,000	131,073	294,805	134,835	5.15	2.29	[28]
LF-Amazon-131K	BoW Features Raw text	80,000	131,073	294,805	134,835	5.15	2.29	[28]
LF-WikiSeeAlsoTitles-320K	BoW Features Raw text	40,000	312,330	693,082	177,515	4.67	2.11	-
LF-WikiSeeAlso-320K	BoW Features Raw text	80,000	312,330	693,082	177,515	4.67	2.11	-
LF-WikiTitles-500K	BoW Features Raw text	80,000	501,070	1,813,391	783,743	17.15	4.74	-
LF-AmazonTitles-1.3M	BoW Features Raw text	128,000	1,305,265	2,248,619	970,237	38.24	22.20	[29] + [30]

➔ The Extreme Classification Repository: Multi-label Datasets & Code (http://manikvarma.org/downloads/XC/XMLRepository.html)

FAIR data principles



Findable

Data is described with rich **metadata.** (Meta)data are assigned **URIs** and are easy to find for both humans and machines.

Accessible

(Meta)data are retrievable using **standard communication protocols.** Metadata are **always accessible.**

Interoperable

Metadata is based on **formal vocabularies** for knowledge representation.

Reusable

Data have a clear **usage license** and accurate **provenance information.**

FAIRness in existing MLC data repositories?

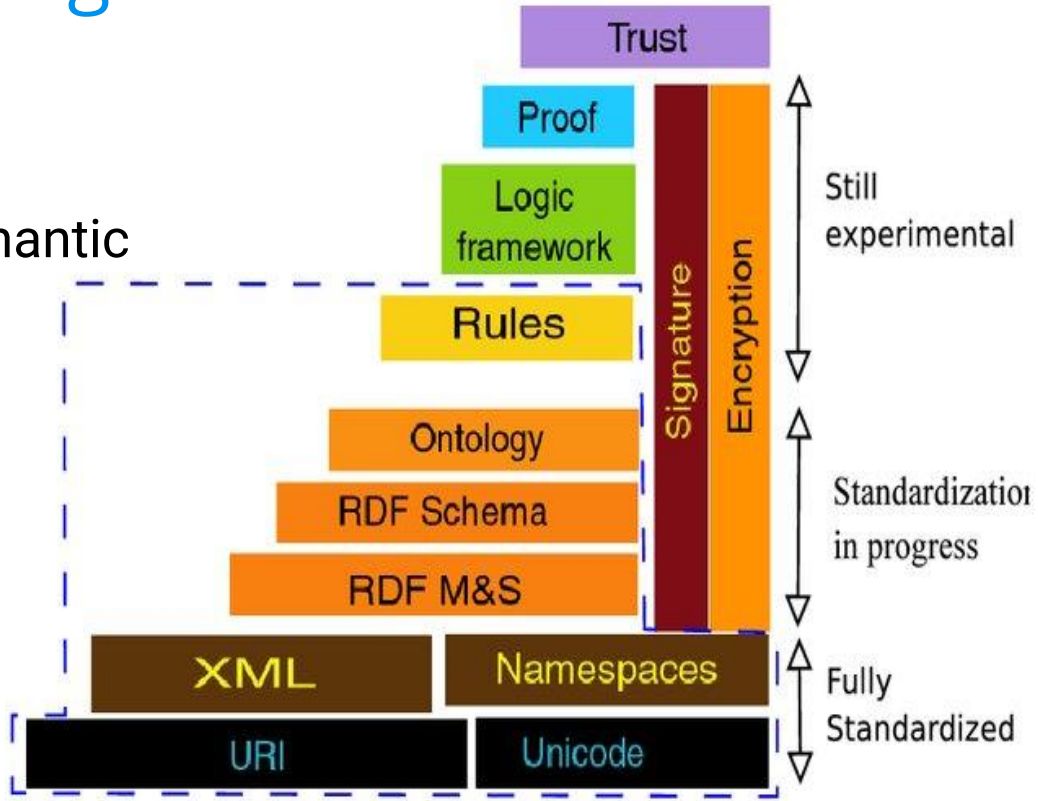
- Data descriptors are not based on formal ontologies and vocabularies.
- Limited interoperability.
- No clear information about the licence and terms of usage.

How do we make MLC data FAIR?

Semantic web technologies

- Enhance data on the web with structured, self-describing, semantic data.
- Semantic Web Technologies:

  URI, RDF, RDFS, SPARQL, etc.



*The **Semantic Web** is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation.*

-Tim Berners-Lee

Semantic annotation of MLC datasets

- Provenance information annotation

- DM-specific annotations

Provenance Information annotation

What?

- Describes the origin of a dataset.

Why?

- Helps better understand, easily cite and reuse the data.
- Search engines can index the datasets and improve discoverability.

How?

- Annotate datasets with the **Schema.org** vocabulary.


```
{
  "@context":"https://schema.org/",
  "@type":"Dataset",
  "name":"Forestry_Kras_LiDAR_Lansat",
  "description":"Remotely sensed data from the Karst region, Slovenia.",
  "url":"http://semantichub.ijs.si/ontodm",
  "creator":{
     "@type":"Person",
     "url": "https://www.researchgate.net/profile/Daniela_Stojanova2",
     "name":"Daniela Stojanova"
  },
  "temporalCoverage":"2001-08-03, 2002-05-18, 2002-11-10, 2003-03-18",
  "spatialCoverage":{
     "@type":"Place",
     "geo": {
        "@type": "GeoCoordinates",
        "latitude": 45.3818,
        "longitude": 13.4815
     }
  },
  "citation": {
     "@type": "ScholarlyArticle",
     "name": "Estimating vegetation height and canopy cover from remotely
       sensed data with machine learning",
     "identifier": "https://doi.org/10.1016/j.ecoinf.2010.03.004"
  },
  "license": "https://creativecommons.org/licenses/by/4.0/"
}
```

Semantic annotation of MLC datasets

- Provenance information annotation

- **DM-specific annotations**

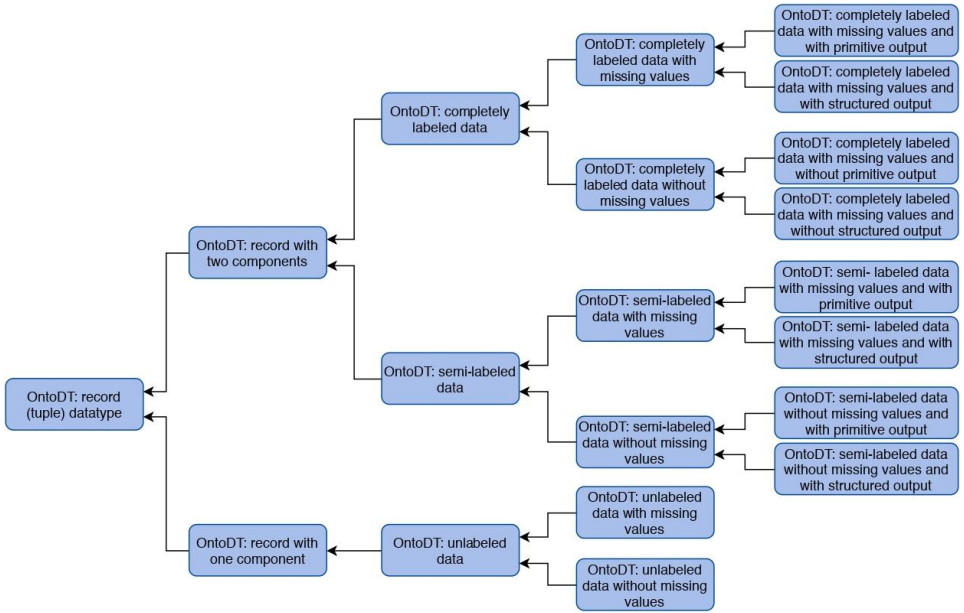
ML-specific annotations

- Annotation of:
  - datatypes
  - data specification
  - ML task
  - MLC related meta features

- Annotations based on community standard ontologies for machine learning.

OntoDT

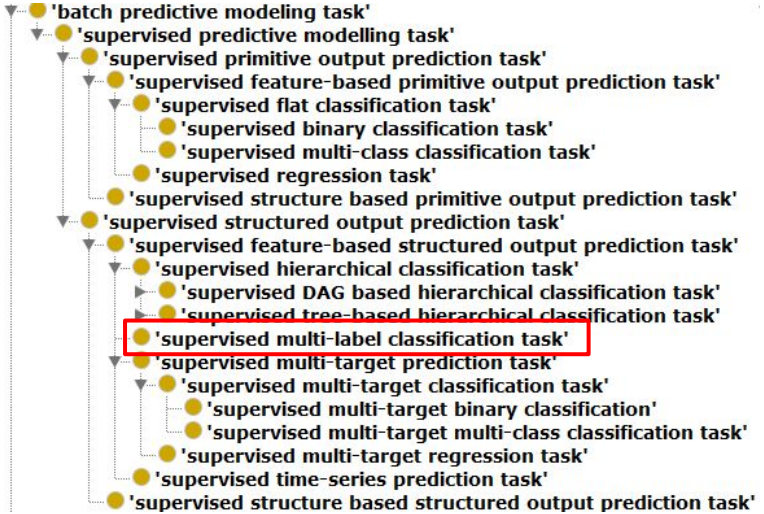
- The OntoDT ontology is a generic ontology for representation of knowledge about datatypes.
- It was initially designed for generating descriptors of datatypes for data from the domain of data mining and using them to determine the data mining task and the set of applicable algorithms
- Its usage is not restricted and it can apply to a variety of domains.



Source: Panče Panov, Larisa Soldatova, Sašo Džeroski. Generic Ontology of Datatypes. Information Sciences 329 (2016) 900–920

OntoDM-core

- OntoDM-core is an ontology of core data mining entities.
- OntoDM-core provides a framework for describing the key DM entities, i.e., data, DM task, generalizations, DM algorithms, implementations of algorithms, DM software.
- OntoDM-core defines several taxonomies such as taxonomy of datasets, data mining tasks, and data mining algorithms.



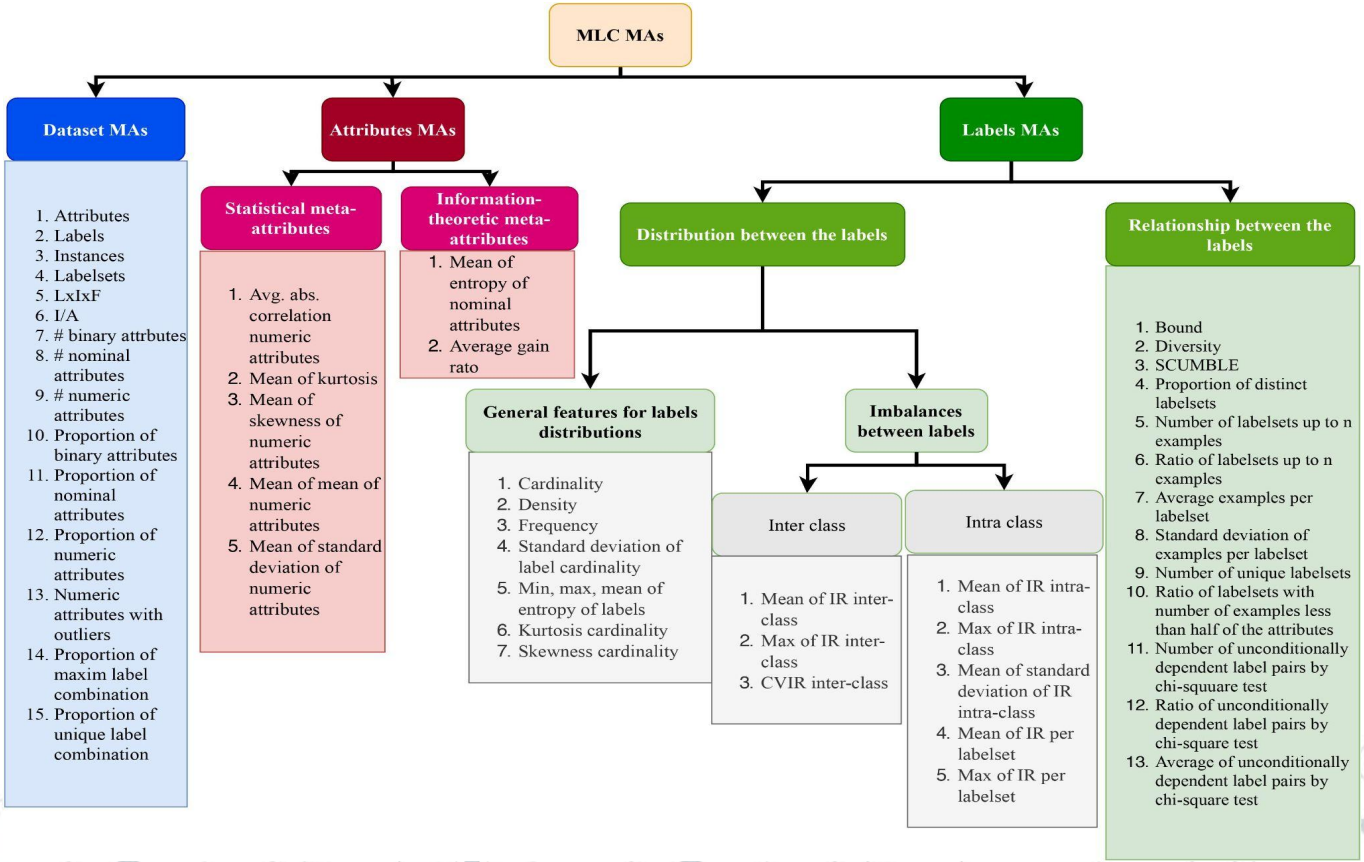
Panov P., Soldatova L., Džeroski S. Ontology of core data mining entities. Data Mining and Knowledge Discovery 28(5-6):1222-1265, 2014
DOI 10.1007/s10618-014-0363-0

Meta learning for the task of MLC

- Meta-learning is a machine learning sub-field and an approach of learning from past experiences or from data of past machine learning experiments commonly referred to as meta-data .
- The meta-dataset usually consists of a set of dataset characteristics (or meta-features)..
- Meta features describe the datasets from a task level. This means that they encode knowledge for the task properties. As such, they can be used to make an empirical study for the properties of the learning task or to guide practitioners to better understand the expected challenges for solving their problem.
- Moyano et al. (2017) defined a list of meta-features specific for multi-label classification datasets, categorized into 5 groups:
    - dimensionality, e.g., number of features, number of labels, number of instances
    - label distribution, e.g., frequency, cardinality and density of labels
    - label imbalance, e.g., mean of inter-class imbalance ratio
    - labels relationship, e.g., proportion of distinct labelsets
    - attribute metrics, e.g., number of binary attributes
- There is an open-source Java-based tool for calculation of the meta-features of multi-label datasets.

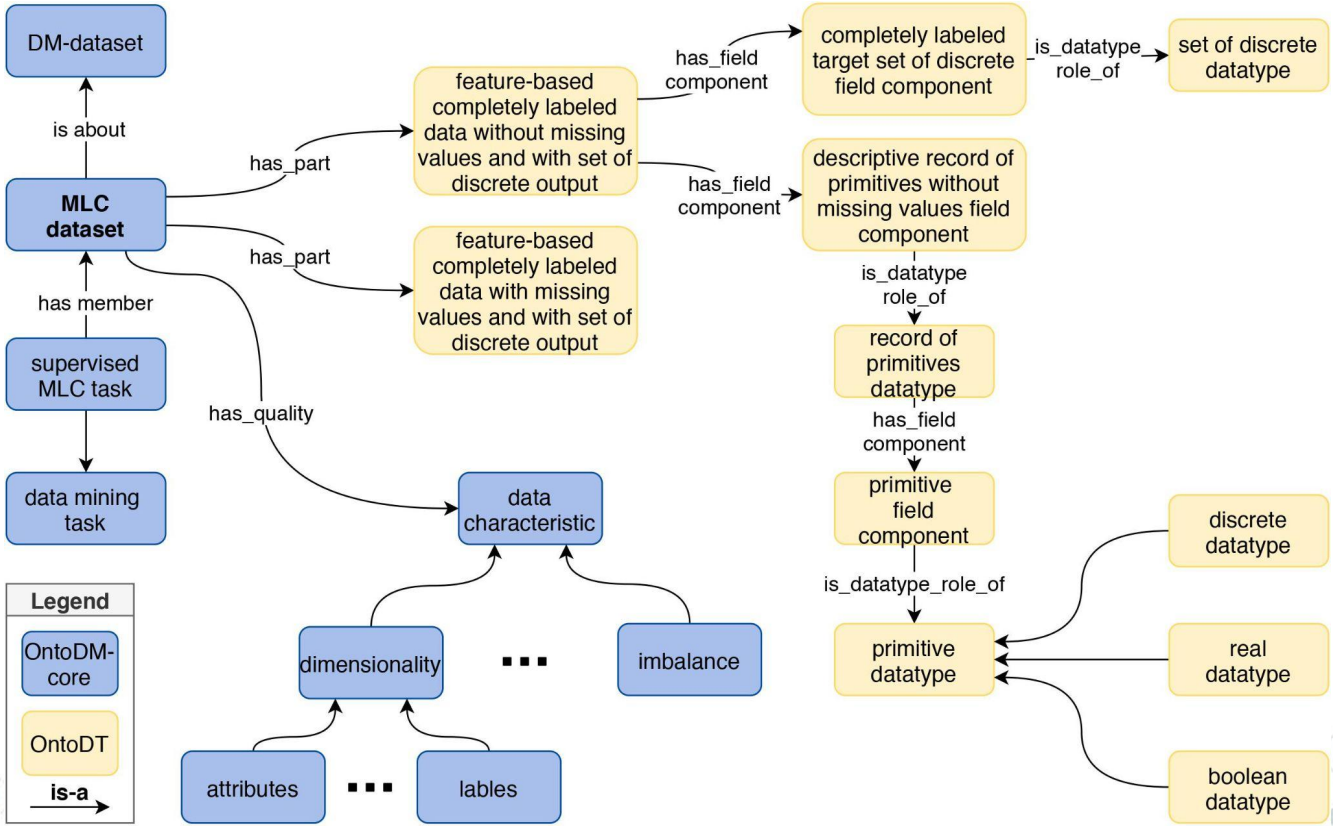
Moyano, Jose M., Eva L. Gibaja, and Sebastián Ventura. "MLDA: A tool for analyzing multi-label datasets." *Knowledge-Based Systems* 121 (2017): 1-3.

MLC meta-features taxonomy

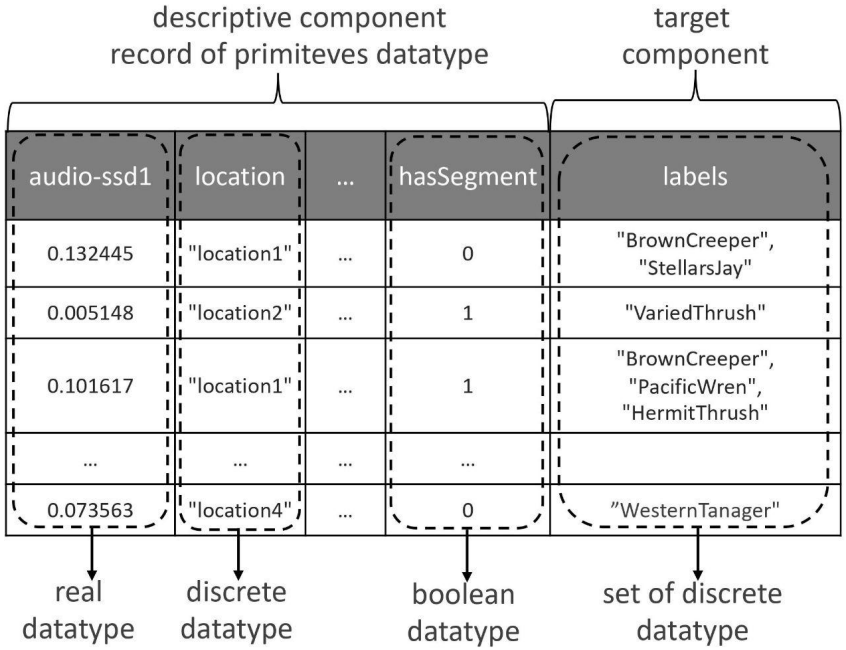


16

DM-specific annotation schema for MLC datasets



Example annotations of MLC dataset



Provenance information

Name: "Birds"
Description: "Birds is a dataset representing the problem of bird species classification from acoustic recordings...."
Same as: [http://mulan.sourceforge.net/datasets-mlc.html, https://cometa.ujaen.es/datasets/birds, http://www.uco.es/kdis/mllresources/#BirdsDesc]
Identifiers: [10.1121/1.4707424, https://www.ncbi.nlm.nih.gov/pubmed/22712937]
Creator(s): [{
name: "Forrest Briggs",
url: https://scholar.google.com/citations?user=WmyVSLQAAAAJ&hl=en}]
Licence: https://creativecommons.org/licenses/by/4.0/

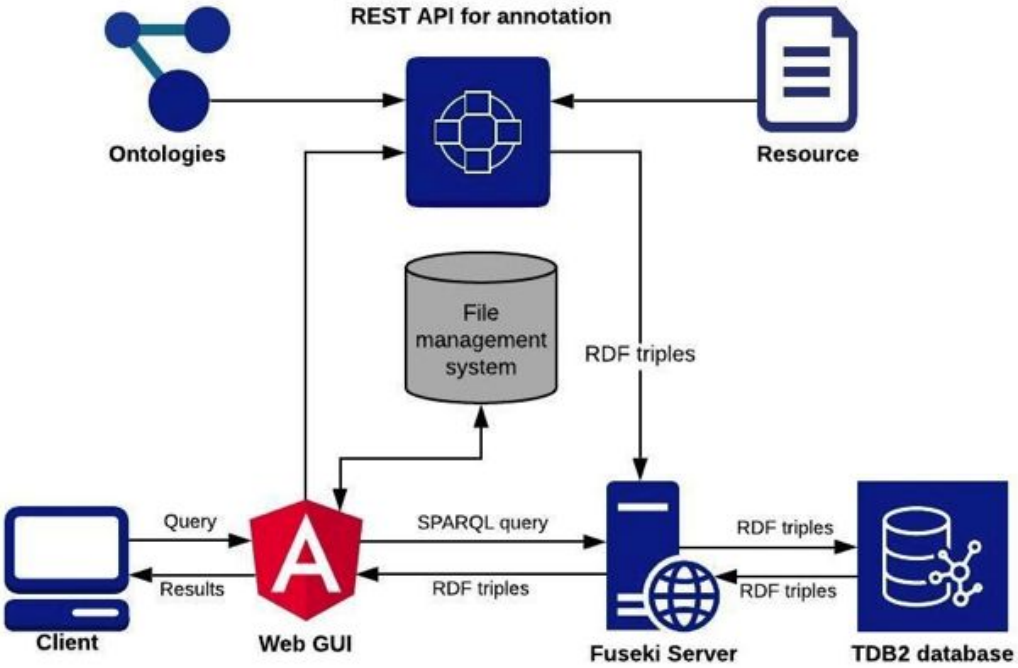
MLC meta-features

Instances: 290
Default accuracy: 0.46552
Standard defiation of label cardinality: 15.487
Kurtosis cardinality: 0.653
Proportion of maxim label combination (PMax): 0.465
Ratio of number of labelsets up to 10 examples: 0.972
.....
Mean of entropies of labels: 0.269

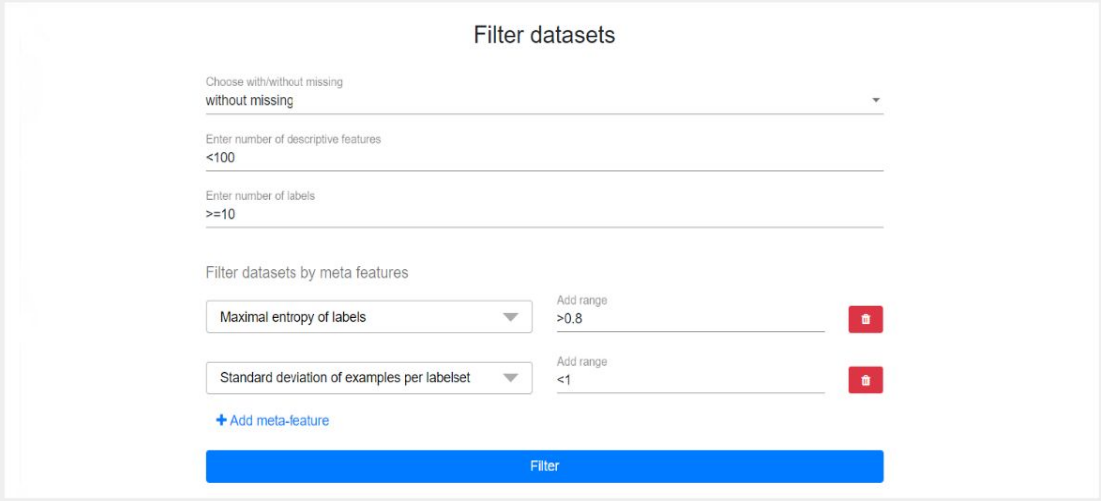
Web-based semantic repository of MLC datasets

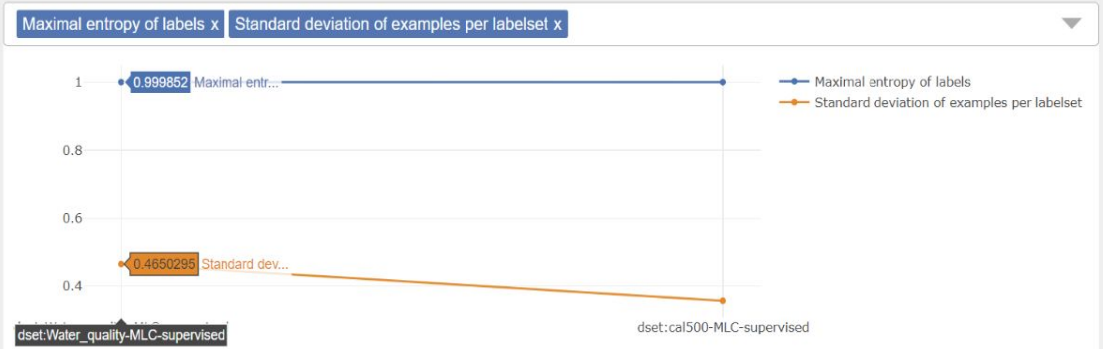
- Hosts metadata for **89 MLC datasets** in total.
- The main contribution of our repository is the **semantic** layer for standardized, formal description of datasets through the application of ontologies and vocabularies.
- The rich semantic annotations provide the repository with advanced querying capabilities that employ the reasoning power of ontologies.
- The explicit inclusion of semantics further broadens the range of applications of the available datasets, as they help practitioners better understand, reuse and augment the data in an automatic fashion.

System Architecture and Design



http://semantichub.ijs.si/MLCdatasets/





Dataset name	No. of instances	No. of descriptive features	No. of targets	Missing values
Water_quality-MLC-supervised	711	16	14	false
cal500-MLC-supervised	362	68	174	false

Items per page: 10 1 - 2 of 2

cal500

Keywords: Audio annotation and retrieval, music information retrieval, semantic music analysis

Description: Cal 500 is a dataset from the multimedia domain, from the subcategory of audio. Each feature is calculated by analyzing a short-time series of the audio signal using various time-series generated features from the audio signal, obtained by human annotators. The targets represent various aspects of music composition such as the emotional level of the song, the music genre, the instruments present in the recording etc.

Same as: https://cometa.ujaen.es/dataset/cal500, http://www.uco.es/kdis/mllresources/#CAL500Desc

Identifiers: https://ieeexplore.ieee.org/document/4432852, 10.1109/TASL.2007.913750

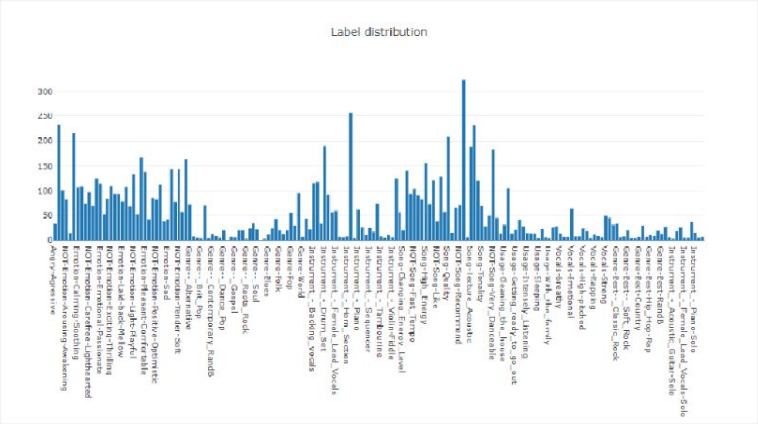
Creator(s): Douglas Turnbull

Download meta-features calculated on the train dataset: cal500_train.json

Download meta-features calculated on the test dataset: cal500_test.json

Download the dataset: cal500.arff

Download annotations: cal500.rdf



Feature	Type	Distribution (train set)
Mean_Acc1000_Std_Mem40_Rolloff_Power_powerFFT_WinHamming_HopSize512_WinSize512_AudioCh0	numeric	
Mean_Acc1000_Mean_Mem40_MFCC1_Power_powerFFT_WinHamming_HopSize512_WinSize512_AudioCh0	numeric	
Std_Acc1000_Std_Mem40_MFCC12_Power_powerFFT_WinHamming_HopSize512_WinSize512_AudioCh0	numeric	

Is our data FAIR enough?

SATIFYD

- Self-Assessment Tool to Improve the FAIRness of Your Dataset.
- The tool shows you how FAIR your dataset is and will provide you with tips to score (even) higher on FAIRness.
- The list is accessible at: https://satifyd.dans.knaw.nl/

Data Archiving and Networked Services
DANS

DANS Checklist for FAIRness

	Question	Points
	Is the data repository you have chosen trustworthy?	2/4
F	Will your dataset have a Persistent Identifier after deposit?	1/1
	Did you provide enough information (metadata) about your data for others to understand and reuse your data?	1/1
	Did you provide rich additional documentation?	1/1
A	Is the metadata publicly accessible?	1/1
I	Are the data stored and archived in preferred archival formats?	1/1
	Did you use standardized vocabulary?	1/1
R	Did you give detailed provenance information for the data?	1/1
	Does the dataset have a usage license?	1/1
	Do you make use of relevant community standards?	0/1
	Total:	10/13 points

CoreTrustSeal certificate

- CoreTrustSeal is an international, community based, non-governmental, and non-profit organization promoting sustainable and trustworthy data infrastructures.

CoreTrustSeal Trustworthy Data Repositories Requirements:

https://www.coretrustseal.org/why-certification/requirements/

Conclusions

- One of the grand challenges of data-intensive science is to facilitate knowledge discovery by assisting humans and machines in their discovery of, access to, integration and analysis of, task-appropriate scientific data and their associated algorithms and workflows.
- *MLC data, and research data in general, will not become nor stay FAIR by magic.*
- *We need people skilled to work with interoperable semantic technologies.*
- *Community collaboration to build and maintain research data infrastructure is needed.*

Thanks!

Any questions?

You can contact me at:

ana.kostovska@ijs.si