



# Agenda







# Workshop practicalities

#### Audio

Click on 'connect audio' but please mute your





#### Chat

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



#### Recording

The workshop will be recorded

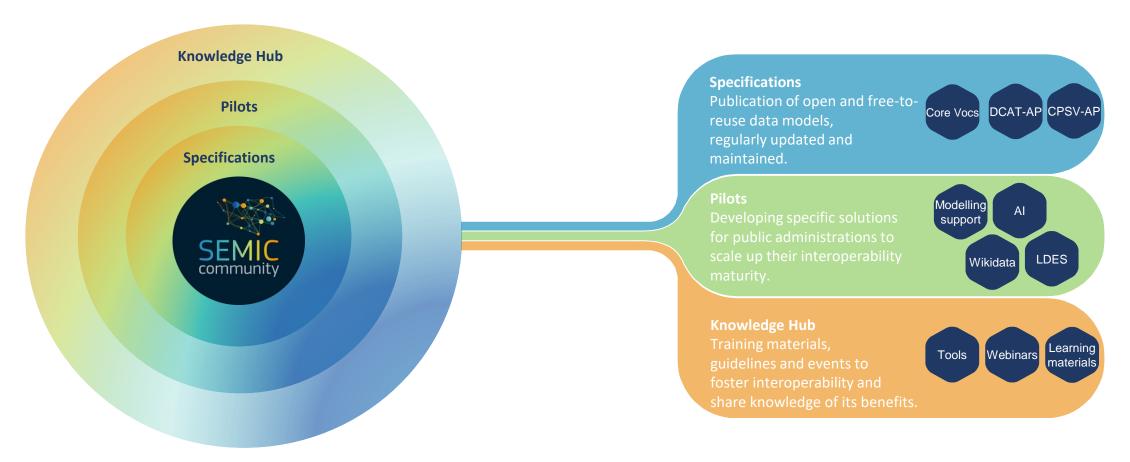






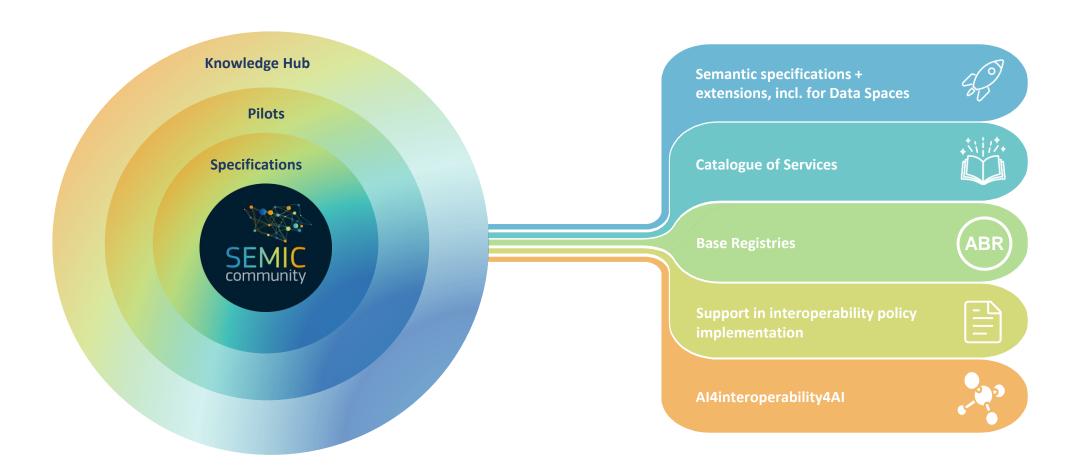
### **SEMIC**

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.



# MLDCAT-AP and the DCAT-AP ecosystem



# 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.





- NAPCORE-Mobility
- GeoDCAT-AP
- HealthDCAT-AP
- ...





# DCAT-AP ecosystem



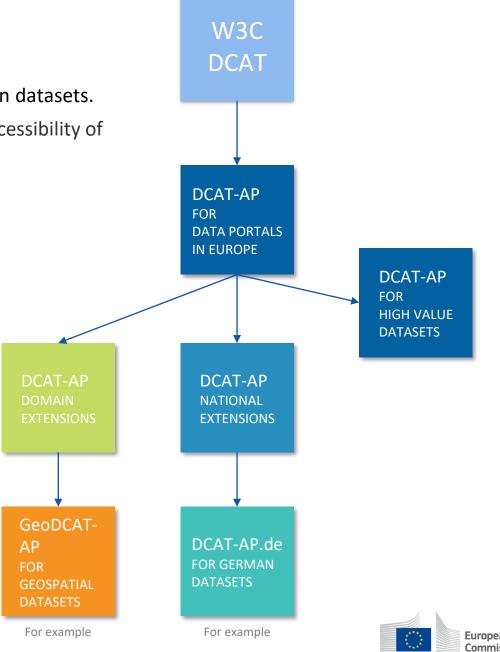
**W3C Data Catalogue Vocabulary** for facilitating interoperability between datasets.

 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)
- ...



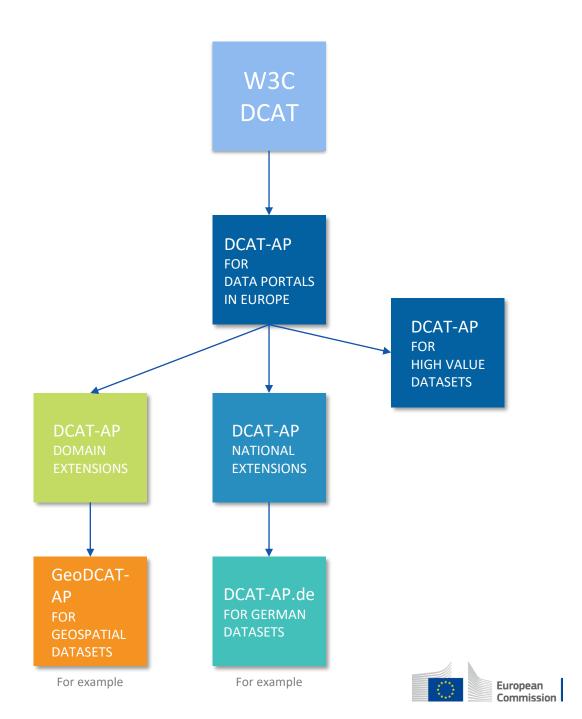
**DCAT-AP Domain Extension** for describing domain specific datasets.

- 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.

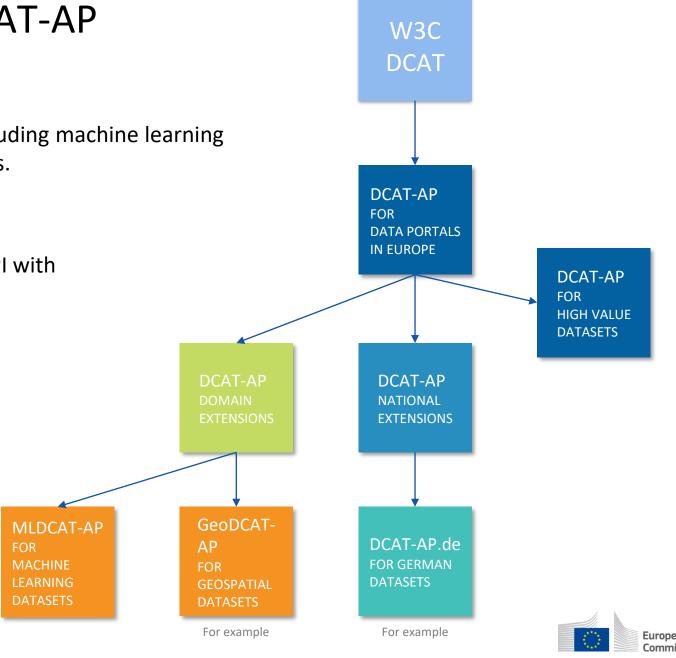
Compatible with DCAT-AP 3.0.0

#### **Benefits**

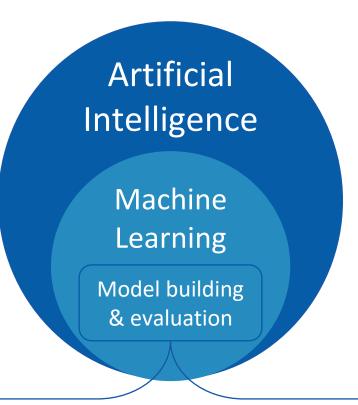
Describing ML Datasets with MLDCAT-AP yields all the advantages of using DCAT-AP.

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



Data Choosing learning model Training model Generate Prediction

Choosing learning model Fvaluating model Generate Prediction







MLDCAT-AP pilot









Make ML research easily accessible and reusable...

European Commission



Make ML research easily accessible and reusable...





Make ML research easily accessible and reusable...







Make ML research easily accessible and reusable...





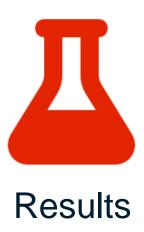




Make ML research easily accessible and reusable...













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



This dataset classifies people described by a set of attributes as good or bad credit risks.

▲ 506k ♥ 28 ▲ 312 

■ 1000 x 21 
■ 31 ⑤ 10 years ago

v.1 🗸



#### blood-transfusion-service-center

Data taken from the Blood Transfusion Service Center in Hsin-Chu City in Taiwan -- this is a classification problem.

▲ 469k ♥ 6 ♣ 101 748 x 5 1464 9 years ago





#### monks-problems-2

Once upon a time, in July 1991, the monks of Corsendonk Priory were faced with a school held in their priory, namely the 2nd European Summer School on Machine Learning. After listening

v.1 🗸



#### tic-tac-toe

This database encodes the complete set of possible board configurations at the end of tic-tac-toe games, where "x" is assumed to have played first. The target concept is "win for x" (i.e., true when

▲ 387k ⊕ 958 x 10 □ 50 ⑤ 10 years ago







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



This dataset classifies people described by a set of attributes as good or bad credit risks.



#### v.1 🗸

#### blood-transfusion-service-center

Data taken from the Blood Transfusion Service Center in Hsin-Chu City in Taiwan -- this is a classification problem.







#### monks-problems-2

Once upon a time, in July 1991, the monks of Corsendonk Priory were faced with a school held in their priory, namely the 2nd European Summer School on Machine Learning. After listening





#### tic-tac-toe

This database encodes the complete set of possible board configurations at the end of tic-tac-toe games, where "x" is assumed to have played first. The target concept is "win for x" (i.e., true when









- Evaluation Procedure
- Splits
- Target
- Metric

#### credit-g classification

Predict feature 'checking\_status'. Possible values are '<0,0<=X<200,>=200,no checking'. Evaluate models using 5 times 2-fold Crossvalidation. The evaluation measure is binominal\_test.

233150 3 4 years ago

#### credit-g classification

Predict feature 'checking\_status'. Possible values are '<0,0<=X<200,>=200,no checking'. Evaluate models using 10-fold Crossvalidation. The evaluation measure is binominal test.

233151 5 4 years ago











- Algorithm descriptions
- Hyperparameters

#### automlbenchmark\_autosklearn

Auto-sklearn as set up by the AutoML BenchmarkSource: source: https://github.com/openml/automlbenchmark/releases/tag/v0.9

26 15509 5 years ago

#### dabl.preprocessing.EasyPreprocessor

A simple preprocessor.

19112 3 2 years ago

#### sklearn.svm.classes.SVC

Automatically created sub-component.

▲ 58 5499 7 years ago

#### sklearn.neighbors.classification.KNeighborsClassifier

Automatically created sub-component.







- evaluation of a Flow on a Task
- predictions
- computed metrics

#### weka.AttributeSelection-R... on Titanic

weka.kf.AttributeSelection-Ranker-Relief... on Titanic by William Raynaut

■ 8869531 ACC 0.776 AUC 0.6996 RMSE 0.4733 ⑤ 6 years ago

#### weka.AttributeSelection-R... on Titanic

weka.kf.AttributeSelection-Ranker-Relief... on Titanic by William Raynaut

8869401 ACC 0.7519 AUC 0.7064 RMSE 0.4202 5 6 years ago

#### weka.AdaBoostM1-SMO(1) on Titanic

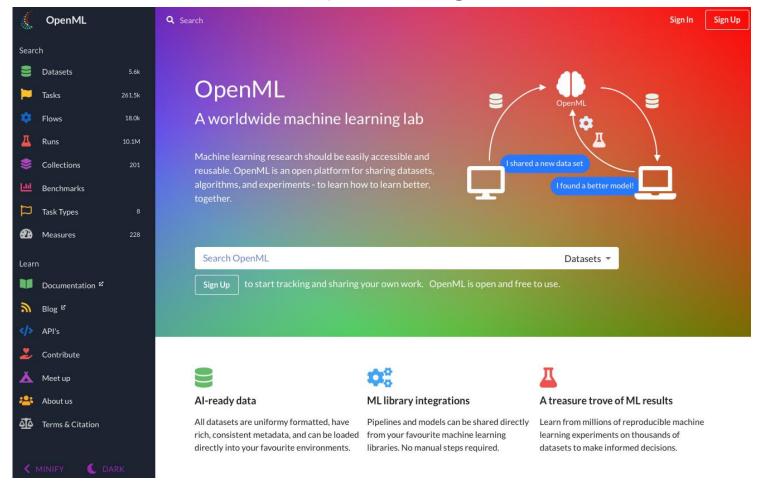
weka.kf.AdaBoostM1-SMO(1) on Titanic by William Raynaut

8869747 ACC 0.776 AUC 0.7208 RMSE 0.4118 5 6 years ago



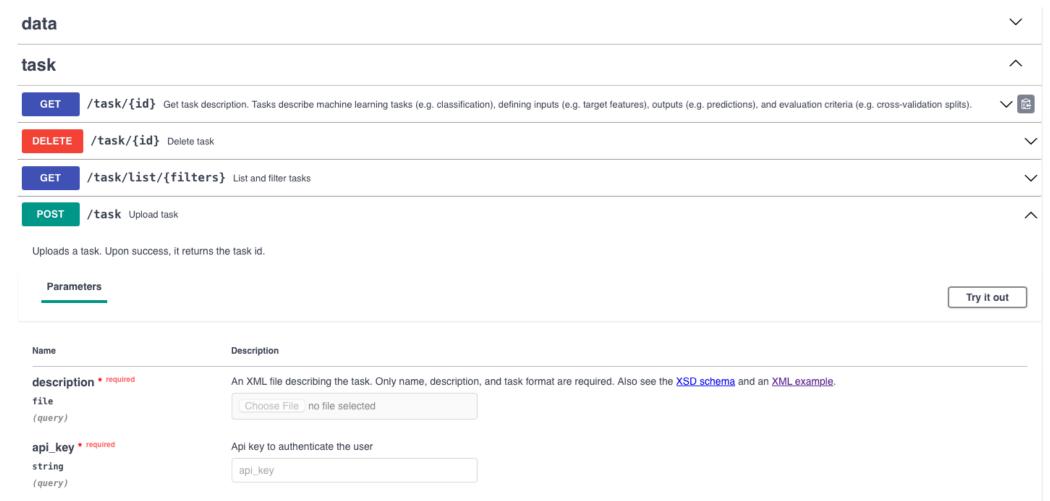


### openml.org











# 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





# OpenML - MLDCAT-AP Process







# OpenML - MLDCAT-AP Process





{"foo":"bar"}



# OpenML - MLDCAT-AP Process





{"foo":"bar"}

Own				Europie		Hugging Face	Example	Gass E	16		CKE-AP	
Class	Property	Datatype	Cartinolity	https://www.openni.org/search?typer-data@status-sect-w@scettmans@shill	Class	Property	https://huggingface.co/datasets/super_glue	Class	Property	Class	Property	Owtotype/Renge
										l		
Dutaset										l		
Corporet	id	2025	0.1	1201							datidentifier	reflectional
Comment	rare	petcasus_strights	1.4	"creditg"		78 FIG.	super_glue				dostrix	refultional refultional
Cotavet	senion description	onksystem_string64	0.1	""Author" (In Hare Helmann	Cutacet	Sammary	SaparSLUI Petga; Mages glasborchmask com	// h		Chesset	pulvenioninfo desdescription	
Corport	description description version	certibasic lider16384 certisystom_string54		"Matter" DE Hare Holleann	Ontaout	Sammary	paparatus prapaginaper guarante con	/] it a new pinc	STAIR IT PRODUCTS	Constant	dictidescription.	rdfs Libral
Company	format	cercosum streets		'ABIT'						Conchesion	deferred	
Corport	envetor	ontbasic_lidn128	0.*	Or Hors Hateans						Comment	detervator	del Madia Type CHE stand
Consuet	contributor	certbasic lidn128	0.*		Dataset	Certifictions	Trans to @horwof, @loster, @astrokoople	danifer adding th	is delicant.	Consum		
Codwood	publisher			"Ass You Right"						Corport	(kit palakidar	Shid Faces) C
Cotavet	collection date	ceribasic_lidn128 eurlateTens		"1994-11-17" "2004-04-0572-21-47"						Contracet	decount	
Consent	(produces	enitosousi shino139		Statist'	Cutavet	languages	Enrish		_	Conduction	dicional	oth Libral typed as not det.LinguisticBystem
Company	lorsor	cethoric lidrot		"N46c"		Tores	unknown			Comment		of Reinstabreet
											ditilicense	dd.Rights/Submert
Consect	urt	essary UPD	0.1	"Https://spi.opentvi.org/data/vii/download/2ii/oredit-g.orff"	Dataset	Гетераре					dot download./Fo	rd's Resource
Corport	parquet uri			"Mity/(opennil.win.tue.nl/detainthi/detaint_hi.pg"						Distantes		
Consect	file id default target attribute	soirt.		"21" "(last"					_	Distribution		$\overline{}$
Corport		entossusi sking135		1000						Diorbation		
Corport	grave_athribute	ced casual shirty 128								Charketer		
Corport	version label	percessual string(28)	0.1	12*						Commont		
Cotacet	distion	certhasis_lidn1004	0.1	"App. Nachive ics uni edulministation, policy htm"							Ten/Yoloy	odd Policy
				"treats serve".								
Contract	tag (annox)	ontsystem_string64	0.*	"feasce_problem",	l	l	1	l	1	Different .	dut boyward	other trees
Consec	sisbilitz	ontacons policy		'ouble'	Cutwet	visblita	public or private			Commen		
Codword	original data url	essary UPD		"Petge://beshive.ics.aci.edu/vij/detsweis/htmlog-igenman-coods-data/"	Deterror	Source Date					datiseding?sgs	Youl Document
Corport	paper uri	ssanyURI	0.1	*Mpc.Nd.aom.org/tolabs*10.1145/967900.968104*	Dataset	paperswithcode	https://paperseithende.com/dateset/upergice			Comment	data Reference dit	rd's Resource
Drisset	updata_comment	cerchosis_lidn128	0.1							CartalogRecard	det description.	refe Shwal
Corport	reinio urt		0.1	"http://opensit.wiv.tue.rl/datasetts/dataset_https:/						Character .		
Cotaset	processing date	certisystem_string\$4 certisasis lidn1504	0.1	"2000-11-20-18-84-53"			<del>                                     </del>	_		Comment		$\vdash$
Drawer	teath.	cethosic lide/1024	0.1				1			Distribution	_	
Corpore	NAME OF TAXABLE PARTY.	cetbox's lidn1604	0.1									
Corport	radii checkura	cercsystom_stringt4	0.1	"ks4750526e00c26ee06c04525e59376c"							spdschedisum	spdx:Drocksum
Codword	features		_		Dataset	data fields				Distribution	terfesture	
Corport	qualities		_							_		
Tera	Nasik Sapo		-			Nest cologenes	ratural longuage inflanence, word same cleants tool dagsification, token dagsification, question-	guater, confin	eco-cosolulian	-		$\overline{}$
	m= 500		<del></del>			hask categories realtinguality	tori dassification, tokan dassification, question- Multilegual		<del></del>		-	
	_		_		Detend	Str Crissian	100 mm 100K			_		
					Detaset	longuage crostor	other					
					Dataset	annotations_meater	expert-gonerated					
					Dataset	source_datasets	orlanded)citier					
			_		Sell.	Size of downloaded dataset lies:	0.00 MB		_	Dankstan	clouf byteSize	$\overline{}$
					Spill.	Size of the generated dataset byte	0.23 MB train, validation, test					
					Dataset	bye Curation Rationals	Iner, velidation, test		_		_	$\overline{}$
					CHRIST	Camping						
					Dataset	Arrelators						
					Delayet	Personal and Sensitive Information	ce .					
			_		Detailed	Social Impact of Dataset						
						Discussion of Siases						
			_		Delated	Other Known Limitedons Delayed Custom			_	_		$\overline{}$
					Dataset	Licensing Information						
			_		URRING.	Coording House	(Singroceedings/stark2018socks,					
1	1		ı	I	1	l	@mgnoceedings/inte/2018/sooks, Star-(Book): Exploring the Surplaing Officially authors (Clark, Christopher and Lee, Kenton an book Star-(SAACL).	of Natural Year'N	Cuestions).	or Ton set C	office Michael and	Sedannia Kalaina
1	1		ı	I	Detanet	Citation Information	BOOKSTO-DEAACEL					
					Detacet	Supported Tasks and Loaderhoan	For datases ingest load_dataset					
			_		Detailed	use in dataset library	From-distances import load_clatenet.			_		
			-					Size	_	Databution	dout.byte5ice	
			_	-	Dataset	Size of the generated distance models	-		-	-		-
			<del>                                     </del>		Mone		1	Machine 1	namina ma	44		_
						version						
					Model	rate	sim-roberta-base					
					Model	loense	Icensevil					
			-		Model	visibility	1		_	_		_
			_	-	Model	pipeline (ap (lask)	Elimisk	model fund	etiese.	_		
	_		<del>                                     </del>		reace:	Description (	and the same of th	moudi Tuni	111111	_		
					Model	ande	1911,02116					
						high						
							XLM-RudER/is model pre-trained on 2.515 of Starred CommonDrawi data containing 100					
					Model	short-description	Instrument Section Control of the Co					
1	1		ı	I	1	l	LM-Russettle is a multiregual ventor of Russettle. It is pre-trained on 2.5TB of Staned	1	1	ı	ı	1 1
1			l	I	Model	desciplier		l	1	l	1	
							You can use the raw model for masked language modeling, but it's modily intended to be then bined on a downshoon last. See the					
1	1		ı	I	Model	Intended uses & limitations	tanguage modeling, but it's mostly intended to be time based on a dependence task. See the	1	1	ı	ı	1 1
			-			THE COLUMN STREET			-	-		
	_			l	More	Leage	Yest and code					
						biblios oligition	tool .					
					Model	enakadon results	Novi .					
						Training precedure -proprocessin	ton0	training pr	ocess detai	8		
	-		-			Training precedure -proteoring Blas and Fairness	Next .	training pr	ocess detai	-		
			-			Class and Painness carbon arribotor	Next Seed		-	_		
			_			detent	Seed.					
Task												
Task	ed .	sainteger	0.1	31								



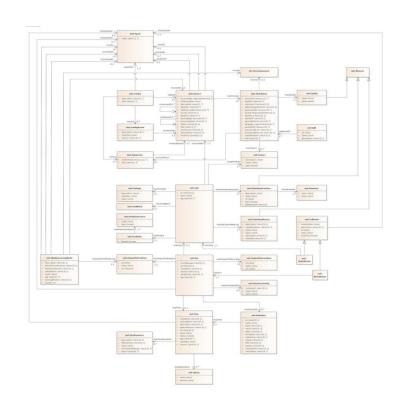
# OpenML - MLDCAT-AP Process





{"foo":"bar"}

Closes				Europie		Hugging Face	Euroje	- '	16		CKE-AP	
	Property	Datatype	Cartinolity	https://www.openni.org/search?tago-data&statu-nctive&sort-nans&id-25	Class	Property	https://huggingface.co/datasets/super_glue	Сами	Fraparty	Class	Property	Ownerspecifiers
	1	1	1		ĺ	I	l	1	1	1	1	1
et					1	l				ı		
65	id	eniod.	0.1	-30.						Commont	decidentifier	refort bond
101	rere	ontosausi, shing128	1.1	"creditg"	Dataset	name	uper_glue				destrik	refle Librari
105	venion	ontossion stringt4	0.1	41	Dataset						partversioninfo	refultional
HAT.	description	cetbasic.lidn16384	1.1	**** Author** - Dr. Hare Hofmone	Dataset	Sammary	SuperSLUE (https://huper.gluebonchreark.com	/) it a new brood	Demand of place of the		desdescription	refor Literal
105	description venion	ontsystem stringt4	0.1	9								-
885	format	cedossusi shinged	1.1	'ARCC'						Charbeton	detformut	det.Modia?yy
2005	creator	certbasic lidn128	0.*	Or Hors Hatmans						Common	determater	Soul Agent
200	contributor	cetbasic lidn128	0.*		Dataset	Certifictors	Thanks to @thorwood, @looker, @outrickscopic	dander arbline ti	No risking at	Common		
266	publisher			"Jan Yan Ron"						Comment	(ACL) and Aller	Solf-Roset
995	collectice date	ceribasic_lafn128	0.1	"1996-11-17"								
	upliced date	exidado Tena	0.1	*2004-04-06723:21:47*							decimand	reforcional by
200	prulimites	cercoscusi_sking128	0.1	"English"	Cotsost	languages	English					det.Linguistict
885	SORECE.	certhasis lide/04	0.1	"Nake"	Distaurt	Eceese	unknown				difuscess/flights	dd.Rightstw
845	uri	essare/LPU	0.0	"Https://spi.operenl.org/data/vil/download/20/oredit-g.seff"	Dataset	Ретизаре					dat download.Fi	rd's Reserve
7900	perquet un			"Mip //opennil.wiv.nar/d/datasethi/dataset_hi pq"						Charbeton		
190	fileid	raint	0.1	-31,								
945	default target attribute	certosacut stright24	0.1	"class"								
200	row_id_attribute	cercossusi shing128	0.1									
	grave_athrove	certosauxi Márqi128	6.*									
200	venion label	cercossus shring138	0.1	·Pr						Commont		
140	dation	certibasis, lidentitida	0.1	"App. Nachiva ka uci adulmiti Mini, polisy htm"	1						hes/Yoloy	odd Policy
net	L	I	L.	"oveit seering", "feance_problem",	1	I	l	1	i .			I
	tag (annay)	prilipator_string64	0."						-		datheyward	rdfc Libral
260	cliability original diess unt	ontacees policy	0.1	"public" "Prips //brithins ice and aduly mildeteening testing-ligen man-condit-dates"	Cotsout	visibility	public or private		_	CONTRACT.	dottlerding?ops	
985		essary/LP0	0.1	"https://arthins.ics.aci.edu/h	Delanet	Source Date capemy/thcode	American de la companya del companya del companya de la companya d		_	(nesset		
260	paper uri	assany LPU	0.1	TKSN-19-AUT-OFFEDERS TO 1145 TK 7500 168104"	AMERICAN STREET	paymout/Icode	https://paperowithcode.com/datmet/ispergice			1002005	dittisReferenced5 dittidescription	grafs Resource
1000	ipdida_comment.	cetbosis_lidn128	0.1	har december of the office of the order	_				_	Latered Recent	extranscription.	press schwal
-	re=0 tm		0.1	AND TODAY OF SALAN AND SAL						and Substitute		
net net	status	cedsystem_string54	0.1	"actiw" "2000-11-20 19:64 53"					-	COMMON TO SERVICE		_
H#1	processing date error	pethasis juln1624	0.1	AND TO SELECT STATE USE.	-		l		-	Conchest.		-
1945 1945		cetbasic_lidn1024	0.1		-		<del> </del>		-			-
H45	radii cheduum	onthosis lidn1504 ontosion shing54	0.1	*ks475053fed0c26ee86c4625e50374c*	_				_		sods shedisam	spdr.Chocks
eret	features	CERCSASSORS 33 m304	B1.18	ANTANIA MANAGAMBANA	Curtainet	deta-fields					herioghers	1909.UNICES
HHC	qualities		_		Lintagent	CIPTA TIERON			_	U-0164701	raineature	_
HP.	sparses.		_		Detend	tiest		_		_		_
	Nacid Suppo		_				rational longuage inflarence, word sames cleanly	iguation, confirm	0700-70936,631			_
	NA SO		_		Dataset	hask categories multimountly	teró-dassification, token-classification, question-	areverny	_	_	_	_
					Detased		Multilegual 186 vol 1006			_		_
			_		Dataset	Size Categories longuists creator			-	_		_
			_		Delanet	onvolutions reader	other			_		_
			_		Dataset	source datasets	expert-gonerated extended inflor		_	_	_	_
	-				Self	Size of downloaded dataset line:	OTO MR		_	Danitudes	chal byta5ica	_
			_		Solit Solit	Size of the generaled dataset	0.23 MB			CHINCEST	con chesses	_
							0.25 MB Fain, validation, test		_	_	_	_
			_		Owtened	byse Conston Retionals	Files, Velicesce, drie.			_		_
	-				Dataset	Caration Mationals			_	_	_	_
			_		Dataset	Arrelatore				_		_
	-				Dataset	Personal and Sensitive Information			_			_
			_		Dataset	Social Impact of Dataset						
					Dataset	Discussion of Blases			_	_		_
			_			Other Known Limitations			_			
			_		Dataset	Owlanet Curations			_	_		_
					Detend	Licensing Information			_	_	_	_
			_		Ownset	Ceoraing more size	Biographic Mark Williams					
		1			l	l	@renx endings/slat/2018/sods, Star-Book): Exploring the Surprising Officely author-Dirak, Christopher and Lee, Kenton or book Exp-SAACL;	of Natural Year'N	a Questions).			
					1	l	author (Clark, Christopher and Lee, Kenton or	d Charg, Mag-V	Nel, and Kwistopy	esili, Tors and C	olins, Mitheel, and	Toutamore, Kris
					Dataset	Citation Information						
					Detased	Supported Tasks and Loaderhoa	for-datasets import load, dataset					
		_	-		Dataset	Use in dataset throny Size of Asserbaded dataset New	romounds report bed_datest.		-	Danis dan	rhot hutofilm	-
	1		_				l	OUZE		ueltidulen	coal.byle5ice	-
	-		_		Delanet	Size of the generated dataset			_	-	_	-
	1		_		Dataset	models			1			_
					Model			Machine L	earning mor	del	_	-
					Model	vanion				-		-
										_	_	_
					Model	rame	xinvobeta-base				-	-
					More		Kminderfaltese Kemenii					
					More	reme Scenie Visibility						
					Model Model	Scenie visibility	tomavi					
					More			model fund	ction			
					Model Model Model	domine visibility pipeline_lap(lask)	Scenarii Sirask	model fund	ction			
					Model Model Model	Icome vability pipeline_tap (lask) aco:	tomavi	model fund	ction			
					Model Model Model	domine visibility pipeline_lap(lask)	Scenaria  Nilmask  1911,02/16	model fund	ction			
					Model Model Model Model Model	Scenie Valbilly pipeline_tap (lask) activ tags	Stemank  Silesask  1991,92216  solver  ALM BySERIo recklipts dozumed or 2.515 of	model fund	ction			
					Model Model Model	Icome vability pipeline_tap (lask) aco:	Stemank  Silesask  1991,92216  solver  ALM BySERIo recklipts dozumed or 2.515 of	model fund	ction			
					Model Model Model Model Model	Scenie Valbilly pipeline_tap (lask) activ tags	Stemank  Silesask  1991,92216  solver  ALM BySERIo recklipts dozumed or 2.515 of	model fund	ction			
					Model Model Model Model Model Model Model	homme Vability operine_lap@mik) andv hom short-description	Element  Element  1911,02100	model fund	ction			
					Model Model Model Model Model	Scenie Valbilly pipeline_tap (lask) activ tags	Element  Element  1911,02100	model fund	ction			
					Model Model Model Model Model Model Model Model Model	frome vahelig  opeline Jap (bek)  oneline Jap (bek)  oneline Jap (bek)  short description  description	Element  Element  1911,02100	model fun	ction			
					Model Model Model Model Model Model Model	homme Vability operine_lap@mik) andv hom short-description	Stemank  Silesask  1991,92216  solver  ALM BySERIo recklipts dozumed or 2.515 of	model fun	ction			
					Model	Scenae visibility populine_Lap (best)  andix begs short description clescription briannied same & limitalizes	Stimmark  Stimma	model fund	ction			
					Model	Season variability operative Lag (Seat) or on the Control of Contr	Element  Element  1911,02100	model fund	ction			
					Model	Scenae visibility populine_Lap (best)  andix begs short description clescription briannied same & limitalizes	Stimmark  Stimma	model fun	clion			
					Model	Season variability operative Lag (Seat) or on the Control of Contr	Stimmark  Stimma	model fund	ction			
					Model	foreman ventrally operation (say (best) or only (best) or only (best) or only (best) or only (best) of only (be	Stimmark  Stimma	model fund	ction			
					Model	foreman ventrally operation (say (best) or only (best) or only (best) or only (best) or only (best) of only (be	Stimmark  Stimma		ction			
					Model	Security Auditory  updation (Aug (MAX)) populine, May (MAX) populine, MAX popu	Stimmark  Stimma	training pr	ocess detail			
					Model	Security Auditory  updation (Aug (MAX)) populine, May (MAX) populine, MAX popu	Stimmark  Stimma	training pr				
					Model	Secure ventral programs (against segulari)  seen seen seen segulari)  seen seen seen seen seen seen seen see	Tomassel  Tomass	training pr	ocess detail			
					Model	Security Auditory  updation (Aug (MAX)) populine, May (MAX) populine, MAX popu	Stimmark  Stimma	training pr	ocess detail			
					Model	Section Section 1 (1994) (1994	Command  The Agents  1918 Agent	training pr	ocess detail			





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



OpenML Dataset	DCAT-AP Class	Property
id	Dataset	identifier
name	Dataset	title
tag	Dataset	keyword
status	Dataset	→ status
qualities	<b>☼</b> Quality	



Qualities describe the data values:

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



### Qualities describe the data values:

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



### 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



### 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



#### § 7.9 Quality Measurement

#### Definition

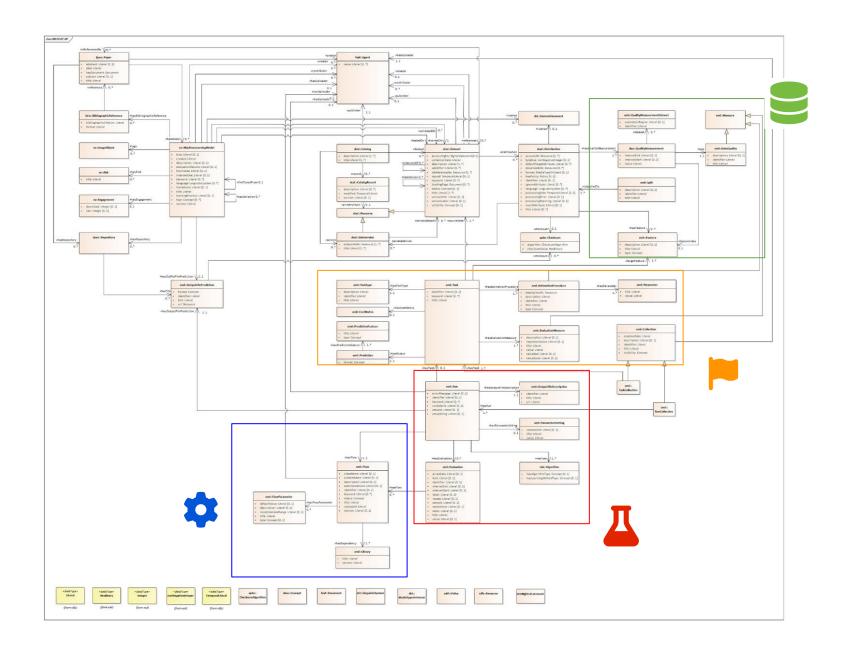
Represents the evaluation of a given dataset (or dataset distribution) against a specific quality metric.

#### **Properties**

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

Property	Range	Card	Definition	Usage
dataset	Quality Measurement Dataset	0*	Indicates the data set of which this observation is a part.	
feature index	<u>Feature</u>	01	The feature on which the quality is measure on.	
interval end	Literal	01		
interval start	Literal	01		
type	Data quality	1	A classification for a quality.	
value	<u>Literal</u>	1	Refers to values computed by metric.	







# Integration





# Integration

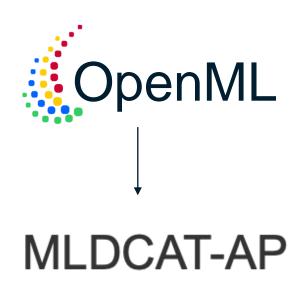






# Integration









# Implementation

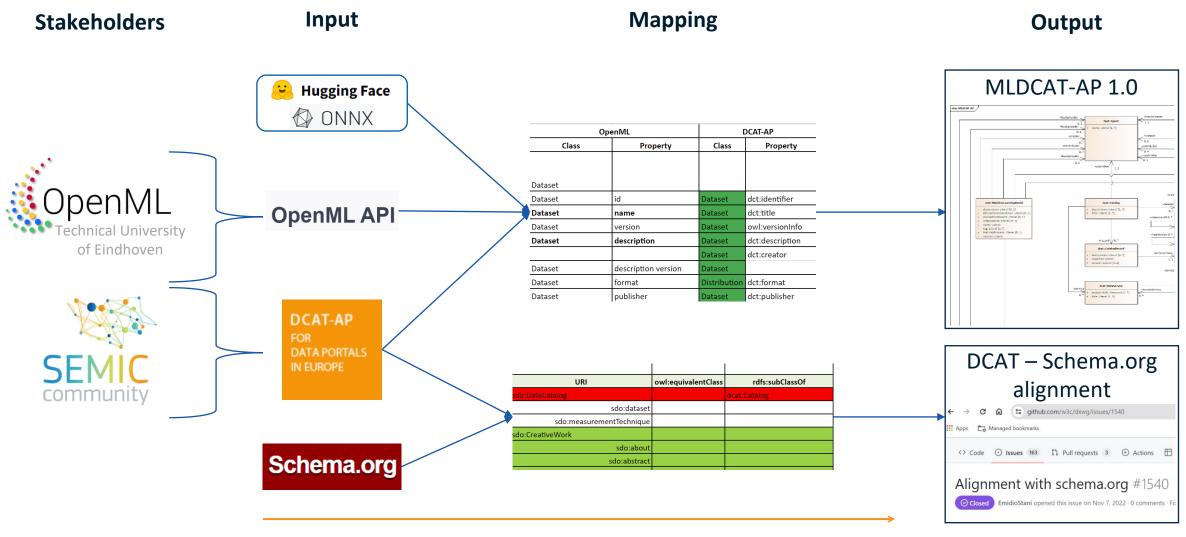
https://test.openml.org/mldcatap/docs



Roll out together with new API later this year



# OpenML pilot



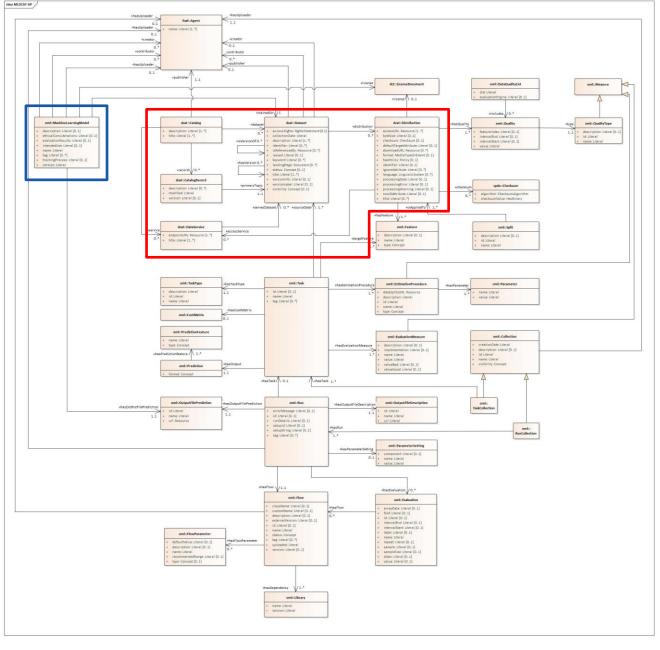
European Commission



# MLDCAT-AP 1.0.0

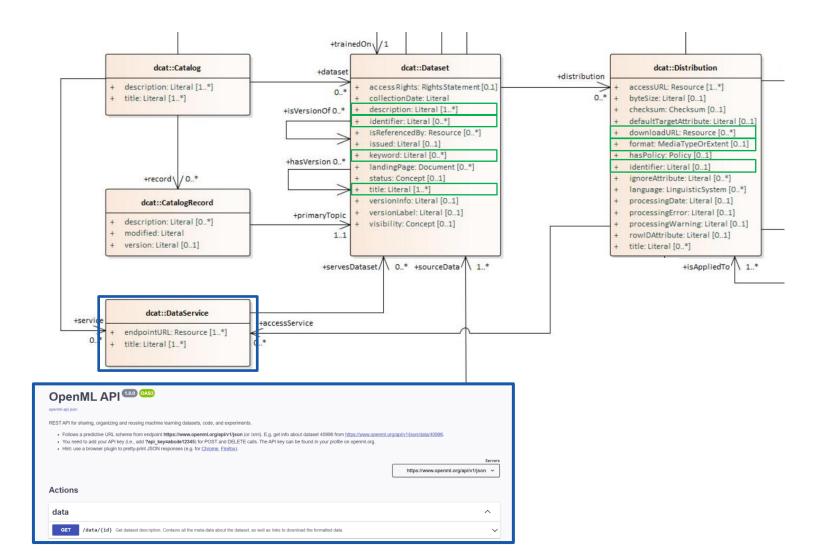
DCAT-AP

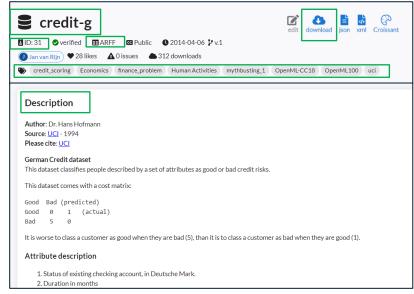
MachineLearningModel





## MLDCAT-AP 1.0.0 – Reusing DCAT-AP concepts



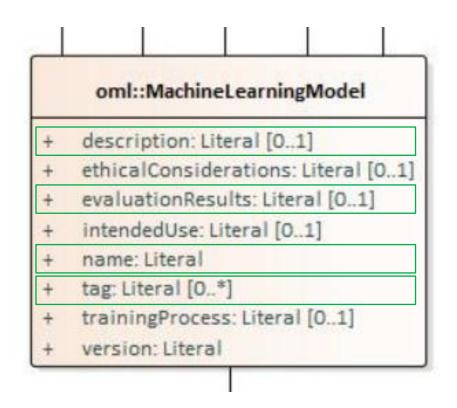


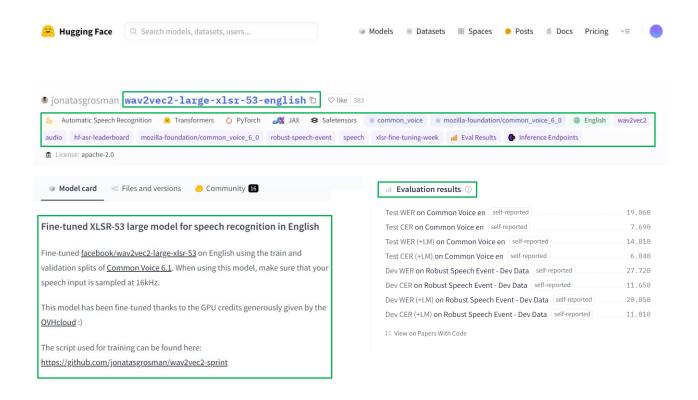
OpenML dataset



## 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)

Copyright © 2024

#### 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 <a href="OpenML">OpenML</a>.

#### § 1. Introduction

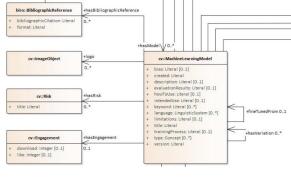


edstaTypen Hextlinory MLDCAT-AP (Machine Learning DCAT-AP) is an application profile that extends DCAT-AP in the field of machine learning.

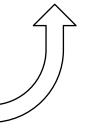




# MLDCAT-AP 2.0.0 – MachineLearningModel



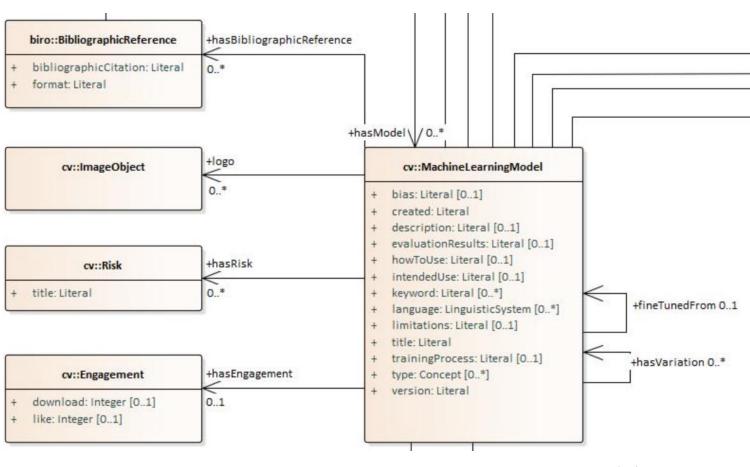
Common Properties	Hugging Face 423K models	Kaggle -TensorFlow 278-2K	Pytorch 55	AzureAl 100+ (in org)
Language	X	X	X	X
References to Paper/Code	X	X	X	
How to use	X	X	X	
Logo	X	X	X	
Files	X	X		X
Likes	X	X	X (GitHub stars)	
Downloads	X	X		
Relation	X (fine tuned from)	X (variation)		
Created date	X			X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





# MLDCAT-AP 2.0.0 – MachineLearningModel

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





### MLDCAT-AP 2.0.0 – Algorithm

Commission Decision Establishing the European Al 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;



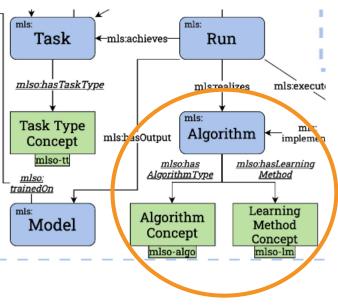
# ANNEX I ARTIFICIAL INTELLIGENCE TECHNIQUES AND APPROACHES referred to in Article 3, point 1

- (a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;
- Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;
- Statistical approaches, Bayesian estimation, search and optimization methods.



# MLDCAT-AP 2.0.0 – Algorithm

#### **MLSO**



#### Machine Learning Algorithm Learning method

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

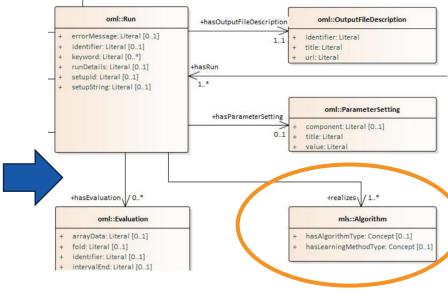
- :ActiveLearning
- :AnalyticalLearning

:AdversarialLearning

- :BayesianLearning
- :ConceptLearning
- :CurriculumLearning
- :EnsembleLearning
- :FederatedLearning
- :FewShotLearning
- :IncrementalLearning
- :InductiveLearning
- :MetaLearning
- :MetricLearning
- :MultiModalLearning
- :MultiTaskLearning
- :OnlineLearning

#### :Reinforcement Learning Algorithm

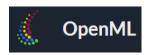
- :SelfSupervisedLearning
- :SelfTaughtLearning
- :Semi-supervised Learning Algorithm
- :SequentialLearning
- :Supervised Learning Algorithm
- :TargetedLearning
- :TransferLearning
- :Unsupervised Learning Algorithm

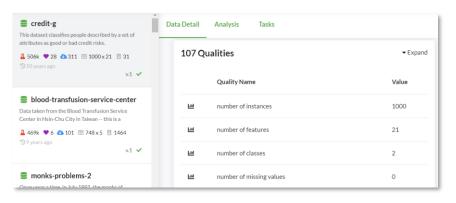


\*In **bold** those mentioned by the AI ACT



# MLDCAT-AP 2.0.0 – Quality





#### 😕 Hugging Face

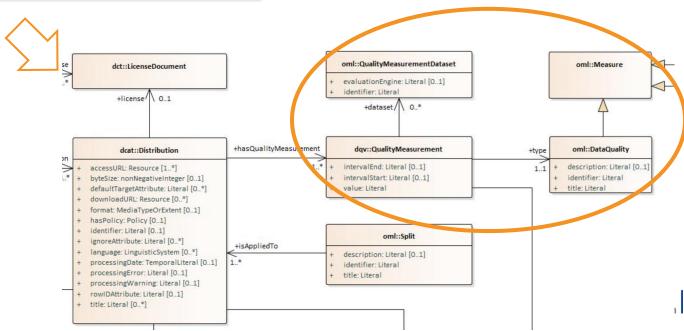


#### kaggle

# TotalDistance_km = TotalDistance_km: This column represents the total distance covered by an individual in kilometers.	# Calories  Calories: This column represents the total number of calories burned by an individual.	# VeryActiveHours   VeryActiveHours: This column represents the number of hours an individual was very active.	# FairlyActiveHours FairlyActiveHours: The column represents the number of hours and individual was fairly active.
0 28	0 4900	0 3.5	0
8.5	1985	0.42	0.22
6.97	1797	0.35	0.32
6.74	1776	0.5	0.18

#### Al 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

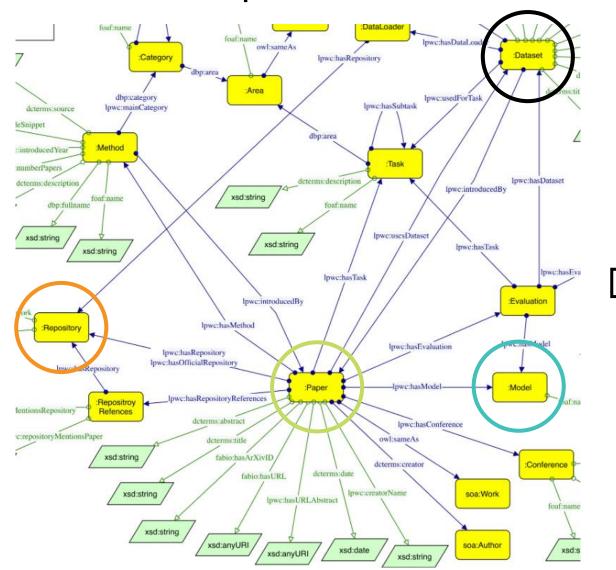


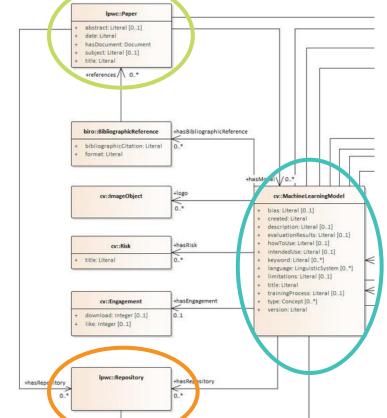
Methods

2,101



1,407
Conferences









### MLDCAT-AP 2.0.0 – Controlled Vocabularies

#### MLDCAT-AP 1.0.0

Uses 3 Controlled Vocabularies published by OP. For example:

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

Introduces 9 newly created Controlled Vocabularies. For example:

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



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





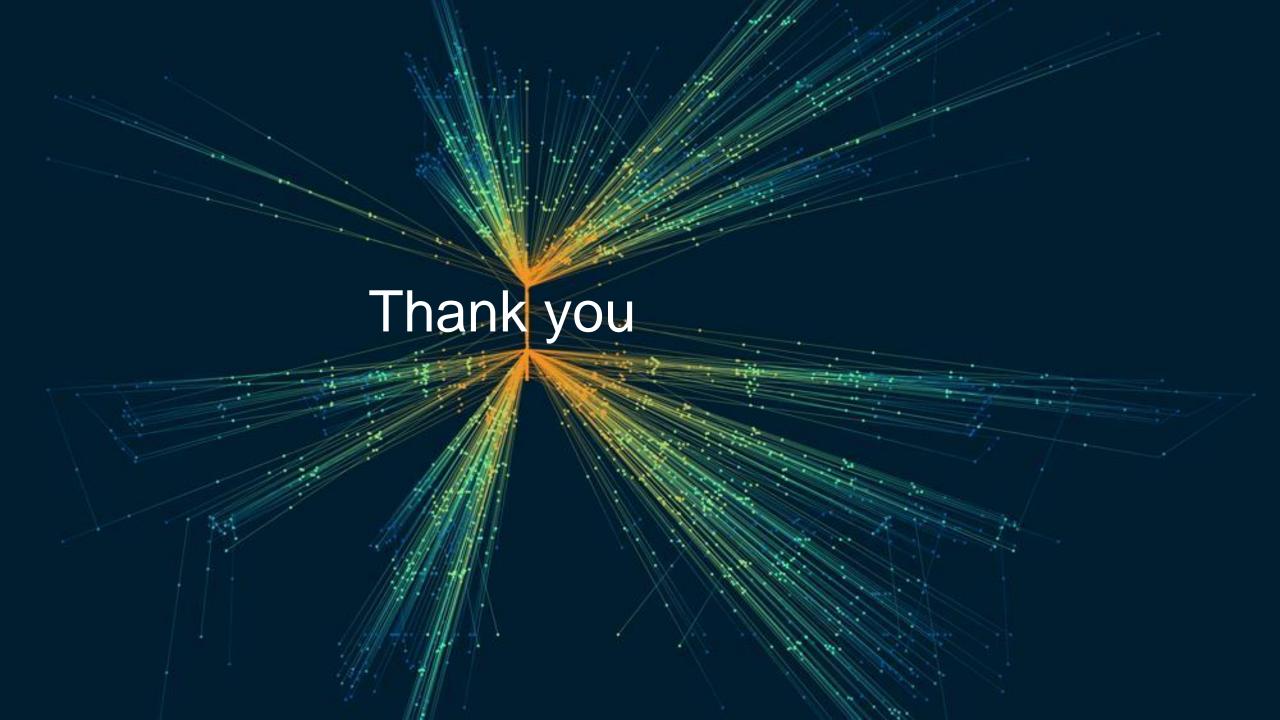




## Next steps

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







# intercoerable europe

innovation ∞ govtech ∞ community

Stay in touch



(@InteroperableEU) / Twitter



<u>Interoperable Europe - YouTube</u>



<u>Interoperable Europe | LinkedIn</u>



DIGIT-INTEROPERABILITY@ec.europa.eu



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