Meeting Minutes: Introductory webinar on MLDCAT-AP (SEMIC - A05.07)

Project:	SEMIC	**Date and Time:**	21/03/2024 10:00 - 11:30
Meeting Type:	Webinar	**Location:**	Virtual
Coordinators:	Alexandra Balahur Emidio Stani Nathan Ghesqiuère	**Issue Date:**	29/03/2024

Agenda of the webinar		
10:00 - 10:10	Introduction	Slides 1 - 7
10:10 - 10:30	MLDCAT-AP and the DCAT-AP Ecosystem	Slides 8 - 14
10:30 - 10:45	Guest speaker - OpenML	Slides 15 - 51
10:45 - 11:10	MLDCAT-AP: a closer look	Slides 52 - 63
11:10 - 11:25	Questions & Answers	Slides 64
11:25 - 11:30	Wrap up & next steps	Slides 65 - 68

Meeting Slides
LINK

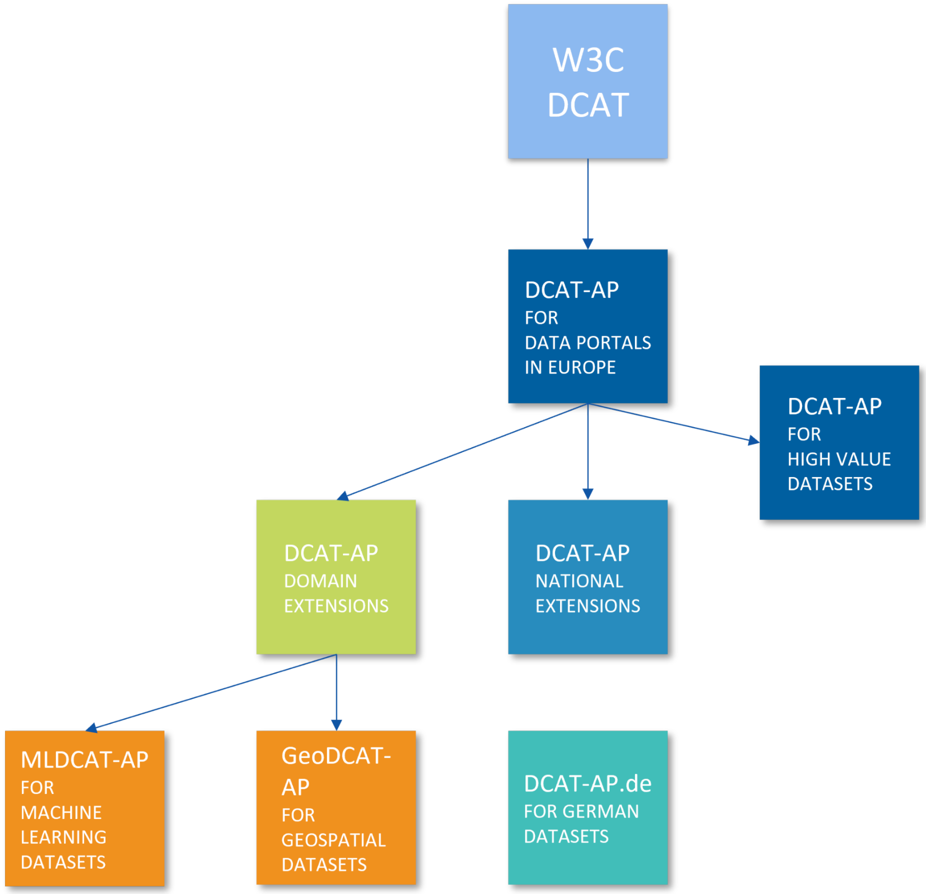
Participants		
Name	**Initials**	**Organisation**
Anastasia Dimou	AD	KU Leuven, Belgium
Anastasia Sofou	AS	DIGIT
Arthur Schiltz	AS	SEMIC Team
Alexandra Balahur	AB	DIGIT
DT	DT	European Commission

Participants		
Name	**Initials**	**Organisation**
Emidio Stani	ES	SEMIC Team
Emilien Caudron	EC	SEMIC Team
Florian Barthelemy	FB	SEMIC Team
Nathan Ghesquière	NG	SEMIC Team
Inês Quintanilha	IQ	European Commission
Ioannis Dasoulas	ID	KU Leuven, Belgium
Jitse De Cock	JDS	SEMIC Team
Juan Alegret	JA	Academic Portals
Maria Martinez Gragera	MMG	Publications Office
Michael Färber	MF	Karlsruhe Institute of Technology, Germany
Mihai Paunescu	MP	Publications Office
Petre Turliu	PT	ESTAT
Pieter Gijsbers	PG	OpenML (Technical University of Eindhoven), Netherlands
Răzvan Radu	RR	Publications Office
Sébastien Albouze	SA	Publications Office
Tom Windels	TW	University of Ghent, Belgium
Vassilis Tzouvaras	VS	European Commission
Zia Alborzi	ZA	Luxembourg National Data Service

Full Meeting Minutes

Welcome & Introduction Slides 1 - 7 **Speaker:** Alexandra Balahur	AB welcomes the audience to the webinar and thanks the audience for their participation. The agenda of the webinar is the following: ● Introduction ● MLDCAT-AP & the DCAT-AP ecosystem ● Guest speaker: OpenML

	● MLDCAT-AP: a closer look ● Next steps Next, the SEMIC context is provided. SEMIC acts as a facilitator of interoperability in Europe through provision of a number of specifications, pilots and a knowledge hub to share documentation. The focus areas of SEMIC are: ● Semantic specifications and extensions ● Catalogue of Services ● Base Registries ● Support in interoperability policy implementation ● AI4interoperability SEMIC specifications enable interoperability in the following ways: ● 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
MLDCAT-AP and the DCAT-AP Ecosystem Slides 8 - 14 **Speaker:** Alexandra Balahur	**The DCAT-AP ecosystem** The objective of DCAT-AP is to support the discovery of/access to (open) data in a cross-border and cross-domain environment, by describing the expression of metadata to be harvested across a distributed network of portals. The DCAT-AP ecosystem consists of W3C DCAT, the model on which DCAT-AP is based. Next is DCAT-AP and its annex DCAT-AP for HVD, followed by the national and domain extensions. MLDCAT-AP finds itself in the latter category, as it is a domain extension specifically for machine learning processes.


W3C DCAT
DCAT-AP FOR DATA PORTALS IN EUROPE
DCAT-AP FOR HIGH VALUE DATASETS
DCAT-AP DOMAIN EXTENSIONS
DCAT-AP NATIONAL EXTENSIONS
MLDCAT-AP FOR MACHINE LEARNING DATASETS
GeoDCAT-AP FOR GEOSPATIAL DATASETS
DCAT-AP.de FOR GERMAN DATASETS

For example

For example

The W3C Data Catalogue Vocabulary is a vocabulary for facilitating interoperability between datasets. It does so through the use of standardised metadata descriptions. The latest version is DCAT 3 which extends DCAT 2 and introduces classes to support offline accessibility of datasets and datasets that are part of a series.

The DCAT Application Profile for describing datasets is based on W3C DCAT. The latest version, DCAT-AP 3.0.0, is fully compatible and aligned with DCAT 3. Additionally, DCAT-AP for High-Value Datasets was recently released. This annex facilitates adherence to the HVD Implementing Regulation with little additional effort.

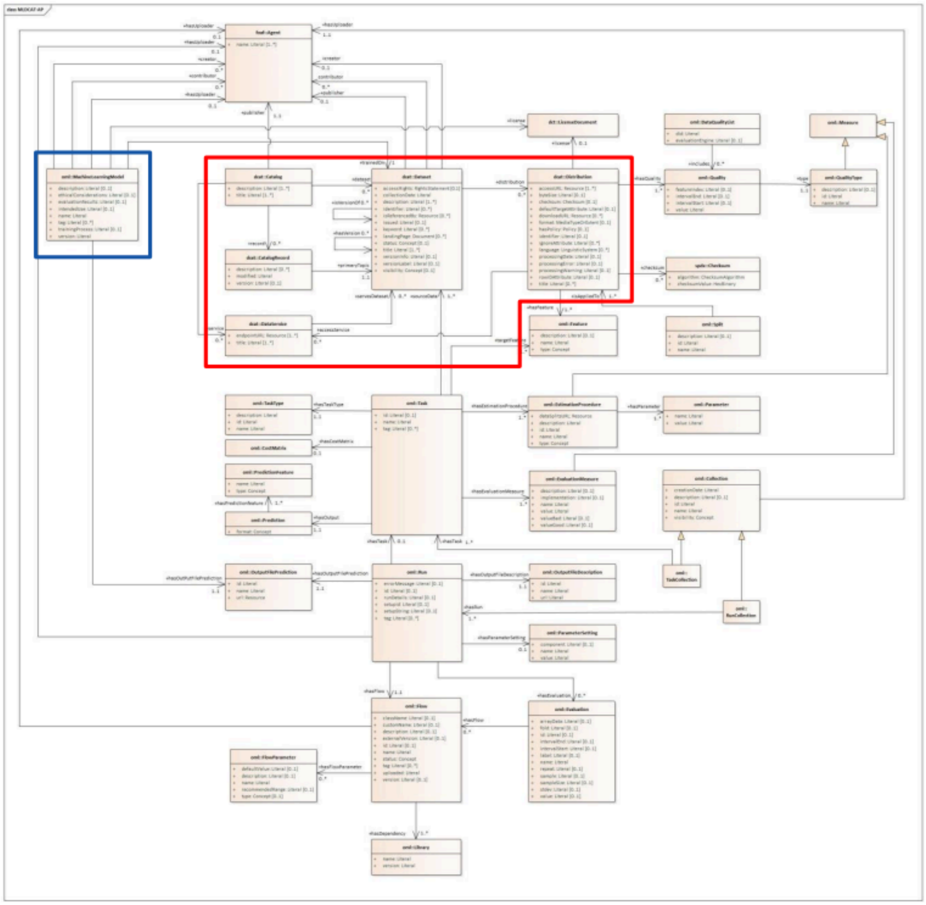
DCAT-AP knows many national extensions that are used for describing national datasets. Some examples are DCAT-AP.de in Germany, and DCAT-AP.it in Italy. However, there are many more.

DCAT-AP also has domain extension for describing domain specific datasets. For example GeoDCAT-AP is used for geospatial datasets, StatDCAT-AP for statistical datasets, and HealthDCAT-AP for datasets in the health industry. In the context of Common European Data

	Spaces these domain extensions are of particular interest as they help Data Spaces achieve interoperability. DCAT-AP is the official standard used on the European Data Portal - data.europa.eu. The European Data Portal provides access to data from all EU institutions, agencies, and bodies. It uses DCAT-AP for its own datasets and aggregates datasets from DCAT-AP compliant portals across Europe. Adoption of DCAT-AP yields multiple benefits: ● 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 implementers to make data catalogues findable. → A harvesting network is made possible. ● Enables implementers to express their metadata in a standardised way. ● Collaborative environment that allows implementers to express their needs and additional requirements (specialisations). **MLDCAT-AP** MLDCAT-AP was developed because a lack of semantic interoperability forbade assets (including machine learning models) to be easily exchanged with other platforms. The aim was to define a common data model and enrich existing API with semantics. The goal is to have a model that is fully compatible with DCAT-AP 3.0.0 In addition to the benefits of DCAT-AP, MLDCAT-AP yields domain specific advantages: ● Improved reproducibility. ● Integration of RAI principles such as transparency and accountability. ● Facilitates adherence to AI Act. MLDCAT-AP finds itself in the domain of machine learning which in itself is a part of the artificial intelligence domain at large. Within the domain of machine learning, the field of model building & evaluation can be found. Within this process, MLDCAT-AP's focus area is on choosing learning algorithms, the training models, evaluating the models and generating predictions.
Guest speaker - OpenML Slides 15 - 51 **Speaker:** Pieter	**The OpenML platform** OpenML is a platform that aims to democratise machine learning (ML). Their goal is to make research on ML accessible and reusable by giving frictionless access to all ML experiment data, including models and results.

Gijsbers	The different concepts in OpenML are: ● Datasets: tabular data with columns and rows with different data types. ○ Support for Deep Learning data, such as audio and video, is being developed ● Tasks: this defines the evaluation procedure, the split being used in the experiment, the target to predict and which metric is being optimised. ● Flows: these contain the algorithm description, dependencies and hyperparameters which affect the way that the model is trained. These flows can be an entire description of how to use the data and build the model with it, or a subset of this process. ● Runs: runs describe the evaluation of a flow on a task, they store the predictions and add metrics calculated by OpenML to assess the quality of the model. The platform can be accessed on the web interface to explore the data and the different concepts. The interface of the REST API can also be used. Additionally, packages in the most common languages such as Python, R and Java are available. Based on this short introduction it becomes clear that making effective use of OpenML comes with a bit of a learning curve. An interoperable format such as MLDCAT-AP reduces this threshold once it would be commonly used in practice. Similar to DCAT-AP for Open Data Portals it would be possible to programmatically interface with OpenML, that is interacting with OpenML using code. Automated harvesting of certain data from OpenML becomes possible with a standardised metadata specification describing the data. This would also work the other way around, when MLDCAT-AP is sufficiently integrated with OpenML, tasks from OpenML could be run on datasets modelled according to (ML)DCAT-AP but are not hosted on OpenML. **Pilot process** In a first step of the pilot process SEMIC was introduced to OpenML and documentation and examples were provided. A mapping was made of OpenML against DCAT-AP, but other platforms such as HuggingFace were also included to assess OpenML specific concepts and more generic ML concepts. A lot of classes and properties are captured within DCAT-AP, however certain specific classes and properties had to be introduced. An example are the elements describing the quality of the dataset. The quality measurement describes the data values. These quality measures can be: ● simple such as number of rows ● statistical such as skewness ● information theoretic such as information gain ● landmarking based on small ML models such as the number of splits in a decision tree.

	An example of a use of quality measures is meta learning by learning relationships between quality of the data and performance of the underlying model. They are also helpful for defining ML benchmarks. **Integration** MLDCAT-AP is implemented with Fast API. The data is converted from the OpenML schema on the OpenML endpoint to the MLDCAT-AP model. In most cases this went smoothly, in some cases optimisations were necessary. SEMIC played an important guiding role for OpenML and helped in the issues they faced regarding the modelling, tooling, implementation, etc. The end result is a number of REST API endpoints that support at minimum the dataset classes of MLDCAT-AP. They are not yet fully developed, but a demo version is online. A full version will be released together with the REST API later this year.
MLDCAT-AP: a closer look Slides 52 - 63 **Speaker:** Emidio Stani	**The Pilot** SEMIC continues with a short summary of the pilot and its outcomes. The input was from the OpenML API and the website, concepts from Hugging Face and ONNX were also used as input. The start was a big mapping exercise between the OpenML schema and DCAT-AP. This led to the first model, MLDCAT-AP 1.0.0. At the same time a mapping with schema.org was done from the DCAT-AP perspective which contributed to the mapping between DCAT and schema.org. **MLDCAT-AP 1.0.0** The classes in red (see image below) are those reused from DCAT-AP. The Machine Learning Model was introduced as a new class. DCAT-AP properties from Dataset, Distribution and Data Service were matched with concepts available on OpenML datasets and the OpenML API. Concrete examples are keywords, download URL, etc.



The ML Model class in blue was based on existing platforms such as Hugging Face and ONNX. Based on the metadata that was available there properties were added such as a description, name, evaluation results, ethical consideration, intended use, etc.

MLDCAT-AP 2.0.0

Moving from version 1.0.0 to version 2.0.0 the model was extended by doing a comparative analysis between repositories of machine learning models. Important classes include: Machine Learning Model, Algorithm, Quality related classes, and Paper.

ML Model class

Four types of repositories were analysed including Hugging Face, Kaggle, Pytorch and AzureAI. Certain properties of these models were taken into account such as language, reference to paper or code, how to use the model, logo's, files, etc.

The output of the comparison is the enrichment of the class ML Model with certain properties and the addition of other classes such a BibliographicReference for paper, an ImageObject for logos and Risk for including risk assessments in light of the AI Act.

	Algorithm class The concept of algorithm is important not only in the ML context but also in the context of the AI Act and the AI Office. This includes the type of algorithm such as supervised or unsupervised algorithms. MLSO (Machine Learning Sailor Ontology) was reused to model the algorithm class. Two controlled lists were reused, the Machine Learning Algorithm and the Learning Method, of which several values were highlighted in the AI Act. Examples of these values are DeepLearningAlgorithm and Bayesian Learning. **Quality related classes** Quality is an important aspect of both Datasets and the ML Model. These quality measures are introduced on OpenML but also Hugging Face and Kaggle. Concrete quality measures are distributions of the data, number of instances and categories. In the AI Act the assessment of an AI system is based on the risk of the AI system. MLDCAT-AP introduces QualityMeasurement, QualityMeasurementDataset, DataQuality and Measure. **Paper class** The BibliographicReference is linked to Paper. The existing resources that are reused originate from Linked Paper With Code (LPWC). The main concept is Paper which is linked with the Dataset, the Repository and the Model. These concepts are reused and integrated in MLDCAT-AP with their respective properties. **Controlled Vocabularies** In MLDCAT-AP 1.0.0 there was the need for Controlled Vocabularies (CV) to describe properties of a Dataset and a Distribution. These CVs are from the Publications Office. However, additional CVs were necessary for MLDCAT-AP specific properties. In total 9 CVs were introduced, for example for the Dataset status which takes values Active, Deactivated or In Preparation. In MLDCAT-AP 2.0.0 new CVs are introduced for the type of the ML Model, the Algorithm type, etc. Newly created CVs will be published by the Publications Office.

Question & Answers Slides 64 **Moderator:** Emidio Stani	**Questions** MF raises a question regarding LPWC and other Linked Data initiatives such as Wikidata, which was mentioned in the beginning. He is interested in how they could be linked and whether there is an intention to do so. ES replies to the first question that currently Wikidata is not yet explored, but the connection with data.europa.eu will be explored in the future because it acts as a registry for datasets. This could be a possible link. A second question from MF is on the future work, for example, how the model would be instantiated. ES notes that the first is to evaluate the model in different use cases as it is quite big. The aim is to keep the model backward compatible. A challenge is on the algorithm side, the CVs are not easy to agree upon. Certain parts of the model will need community agreement. AD adds that the algorithm side has her interest. She raises a question on the model and the pilot with OpenML. It remains unclear on how OpenML reuses the model - is there an actual knowledge graph, or does it implement it in the API? ES mentions that the API uses Linked Data and the JSON-LD context. PG mentions that internally the OpenML schema is used and the conversion happens at a later stage, then the linked data is provided with the endpoints that link the concepts. AD has another question on the quality aspect. What is known from the literature on data quality is different from the quality of a knowledge graph. The data quality vocabulary allows to create different quality measures, but how does this quality fit into potential measurements that can be used, for example for completeness of the data? This works for closed datasets, however in the case of completeness of a knowledge graph the definition of completeness becomes blurry, for example whether it should all be in one endpoint. Let's say you have a restaurant, the owner can assign quality, but the customers can also assign quality through reviews. Self declaration of quality is therefore not always aligned with quality experienced by the user. She wonders whether there is a plan to include user based quality measures as well. ES answers that the quality indicated by the measures is not always clear. The idea of the measures is to indicate that if a certain effort is done to improve the quality of a dataset it becomes more valuable. The idea is not to compare datasets to one another, but to analyse if effort has been taken to improve the quality of the data itself. Another challenge is when a measure is applicable for tabular data but perhaps not for other types of data. PG adds that the measures describe the data and do not indicate the quality directly, but they could be a proxy for for example label leakage in the data which may be problematic for data evaluation. The suspicion of mislabeled data is another example of a quality indicator.

	For OpenML itself the question of what is good quality data is addressed by the concept of benchmarking suites which are a collection of tasks run on a standardised model. The idea is that users can filter these datasets and create a collection of datasets that are regarded as being of good quality. This is not yet fleshed out completely. At this moment these automatically computed measures should serve as a proxy for data quality. AD raises a question on SHACL shapes as the model has a lot of restrictions. She wonders what the vision is on SHACL shapes for MLDCAT-AP. ES mentions that the shapes are used to validate instantiation of models. A focus is on cardinalities and controlled lists. Enlarging these cardinalities can be feasible in terms of backwards compatibility, however restricting is a different matter and could break existing implementations. The shapes are automatically generated in the toolchain. Therefore, cardinalities in the model are automatically reflected in the generated shapes. Validation is provided through a web service for DCAT-AP, if the community raises a need for this it can also be deployed for MLDCAT-AP. PT is looking into how the standard could be useful for their work in ESTAT and statistical datasets. However, they have not yet had the opportunity. FB encourages the audience to provide feedback for potential use cases for MLDCAT-AP that they may see in their domain.
Wrap up & next steps Slides 65 - 68 **Speaker:** Emidio Stani	The next steps with regards to MLDCAT-AP are the monitoring of the publication of the AI Act and the resulting requirements for AI Systems. These requirements from the legislation should be reflected in MLDCAT-AP. Feedback from the community is requested on GitHub and continuing to grow the community of MLDCAT-AP is an important next step. The audience is thanked for their participation.