ML-Schema: Exposing the Semantics of Machine Learning with Schemas and Ontologies

Gustavo Correa Publio
AKSW Group
University of Leipzig, Germany

Diego Esteves
SDA Research
University of Bonn, Germany

Agnieszka Ławrynowicz
Poznan University of Technology
Poland

Panče Panov
Jožef Stefan Institute
Ljubljana, Slovenia

Larisa Soldatova
Brunel University, UK

Tommaso Soru
AKSW Group
University of Leipzig, Germany

Joaquin Vanschoren
Eindhoven
University of Technology
The Netherlands

Hamid Zafar
SDA Research
University of Bonn, Germany

Abstract

The ML-Schema, proposed by the W3C Machine Learning Schema Community Group, is a top-level ontology that provides a set of classes, properties, and restrictions for representing and interchanging information on machine learning algorithms, datasets, and experiments. It can be easily extended and specialized and it is also mapped to other more domain-specific ontologies developed in the area of machine learning and data mining. In this paper we overview existing state-of-the-art machine learning interchange formats and present the first release of ML-Schema, a canonical format resulted of more than seven years of experience among different research institutions. We argue that exposing semantics of machine learning algorithms, models, and experiments through a canonical format may pave the way to better interpretability and to realistically achieve the full interoperability of experiments regardless of platform or adopted workflow solution.

1 Introduction

Complex machine learning models have recently achieved great successes in many predictive tasks. Despite their successes, a major problem is that they are often hard to interpret, which may affect their safeness and the level of trust of their users. The problem of interpretability is one of the key research issues in the area of knowledge engineering and the Semantic Web community, which deals with making the semantics of various phenomena explicit. In this community, the problem of dealing with trust and traceability has gained a major interest last years and resulted in proposals and uptake of models such as the provenance ontology PROV-O [6].

Despite recent efforts to achieve a high level of interoperability of ML experiments through existing workflow systems [9], metadata repositories [11, 7] and schemata [4, 8, 5], we still run into problems created due to the existence of different ML platforms: each of those has a specific conceptualization or schema for representing data and metadata. Figure 1 depicts the complexity of this scenario: (1) and (2) representing the worse scenario to achieve reproducibility of the experiments; (3) and (4) which - although defining a known (local) schema - lack interoperability, since they follow different standards. To reduce this gap, the aforementioned ML vocabularies and ontologies have

2nd Reproducibility in Machine Learning Workshop at ICML 2018, Stockholm, Sweden.

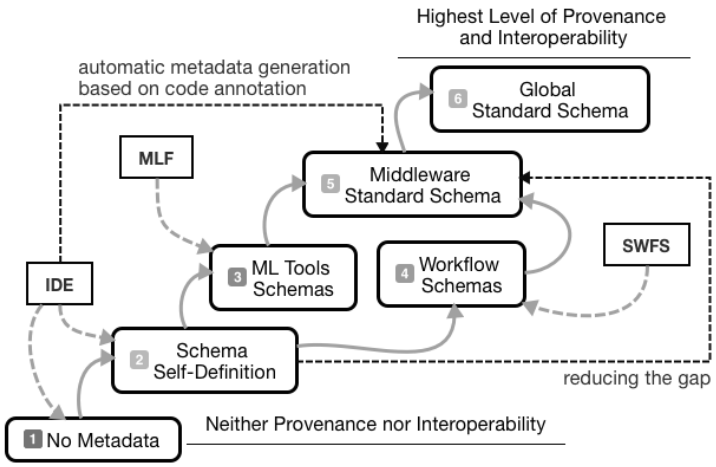

Figure 1: Vertical and Horizontal Interoperability across ML Environments.

been proposed (e.g., MEX and Expose) (5); and finally (6) which encompasses the higher level of interoperability to allow fully reproducible experiments.

In line with those efforts to bridge the gap, in this paper, we present the ML-Schema[1] [3], developed within the W3C Machine Learning Schema Community Group[2]. It is a simply shared schema that provides a set of classes, properties, and restrictions that can be used to represent and interchange information on Machine Learning (ML) algorithms, datasets, and experiments. It can be easily specialized to create new classes and properties and it is also mapped to more specific ontologies and vocabularies on machine learning [4, 10, 8, 5], for instance to represent Deep Learning problems, which are naturally harder to design when compared to supervised approaches. These ontologies, in turn, contain terms for representing more detailed characteristics and properties of ML datasets, algorithms, models, and experiments.

Ultimately, we believe that involving such a canonical and standardized model meta-data descriptors in design of interpretable methods for ML may lead to better insights into the data and the properties of ML algorithms.

The gap can be further significantly reduced by achieving interoperability among state-of-the-art (SOTA) schemata of those resources (Figure 1: item 5), i.e., achieving the horizontal interoperability (Figure 1: item 6). Therefore, different groups of researchers could exchange SOTA metadata files in a transparent manner via web services, e.g.: from OntoDM and MEX (`MLSchema.Schema data = mlschema.convert(‘‘myFile.ttl’’, MLSchema.Ontology.OntoDM, MLSchema.Ontology.MEX)`). Furthermore, the canonical format also directly benefits different environments, such as ML ecosystems (e.g. OpenML [11]) and ML Metadata Repositories (e.g. WASOTA [7]) which can benefit on the mappings of a shared standard.

2 The ML-Schema

In Fig. 2, we depict the classes and the relationships between the classes representing the ML-Schema. The schema contains classes for representing different aspects of machine learning. This includes representations data, datasets and data/dataset characteristics. Next, it includes representations of algorithms, implementations, parameters of implementations and software. Furthermore, the schema contains representations of models, model characteristics and model evaluations. Finally, the schema has the ability to represent machine learning experiments with different granularity. This includes representations of runs of implementations of algorithms on the lowest level and representation of studies at the highest level.

[1]ML-Schema: `http://ml-schema.github.io/documentation/`
[2]W3C ML-Schema Community Group: `https://www.w3.org/community/ml-schema/`

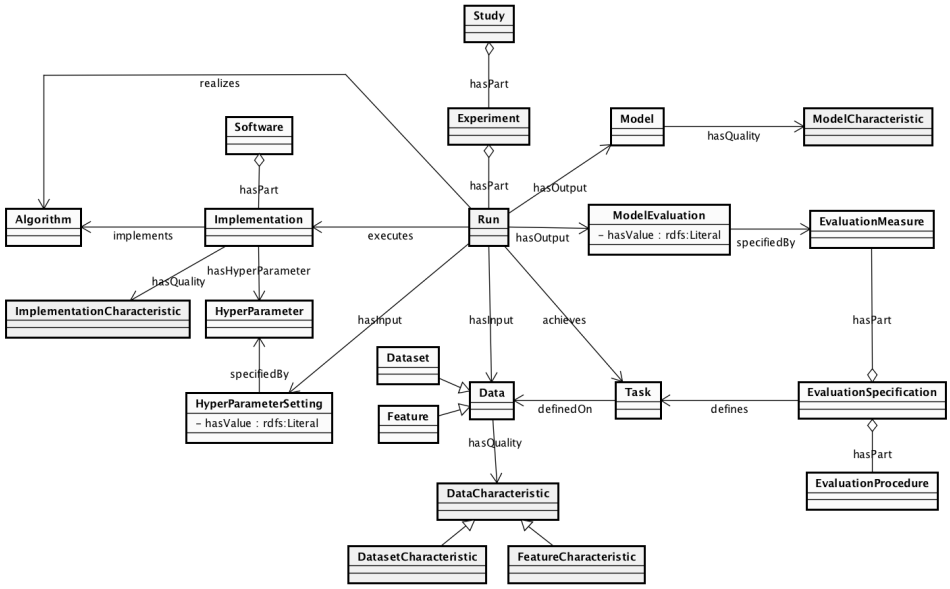

Figure 2: ML-Schema Core classes. Boxes represent classes. Arrows without filled heads represent properties, arrows with empty heads represent subclass relations, and arrows with diamonds represent part-of relations.

3 Machine Learning ontologies

In this section, we present the relationship of the ML-Schema to other proposed ontologies, and vocabularies for the domain of machine learning. The development of ML-Schema was highly influenced from, initially independent, research of several groups on modeling the domain of machine learning. Due to this the classes and relations present in the ML-Schema re-appear in the current ML ontologies and vocabularies. In Table 1, we present the mapping between the terms present in the ML-Schema and the current ML ontologies and vocabularies. Below, we describe each of the mentioned ontology/vocabulary.

The OntoDM-core ontology. The OntoDM-core ontology has been designed to provide generic representations of principle entities in the area of data mining [8]. In one of the preliminary versions of the ontology, the authors decided to align the proposed ontology with the Ontology of Biomedical Investigations (OBI) [2] and consequently with the Basic Formal Ontology (BFO) [1] at the top level, in terms of top-level classes and the set of relations. That was beneficial for structuring the domain in a more elegant way and the basic differentiation of information entities, implementation entities and processual entities. In this context, the authors proposed a horizontal description structure that includes three layers: a specification layer, an implementation layer, and an application layer. The specification layer in general contains information entities. In the domain of data mining, example classes are data mining task and data mining algorithm. The implementation layer in general contains qualities and entities that are realized in a process, such as parameters and implementations of algorithms. The application layer contains processual classes, such as the execution of the data mining algorithm.

The Exposé ontology. The main goal of Exposé [10] is to describe (and reason about) machine learning experiments. It is built on top of OntoDM, as well as top-level ontologies from bio-informatics. It is currently used in OpenML[3], as a way to structure data (e.g. database design) and share data (APIs). MLSchema will be used to export all OpenML data as linked open data (in RDF).

The DMOP ontology. The Data Mining OPtimization Ontology (DMOP) [5] has been developed with a primary use case in meta-mining, that is meta-learning extended to an analysis of full DM

[3]OpenML: `http://www.openml.org/`

Table 1: Mapping between the terms between the ML-Schema and the different ML/DM ontologies and vocabularies

ML-Schema	OntoDM-core	DMOP	Expose	MEX Vocabulary
Task	Data mining task	DM-Task	Task	mexcore:ExperimentConfiguration
Algorithm	Data mining algorithm	DM-Algorithm	Algorithm	mexalgo:Algorithm
Software	Data mining software	DM-Software	N/A	mexalgo:Tool
Implementation	Data mining algorithm	DM-Operator	Algorithm implementation	N/A
HyperParameter	Parameter	Parameter	Parameter	mexalgo:HyperParameter
HyperParameterSetting	Parameter setting	OpParameterSetting	Parameter setting	N/A
Study	Investigation	N/A	N/A	mexcore:Experiment
Experiment	N/A	DM-Experiment	Experiment	N/A
Run	Data mining alg. execution	DM-Operation	Algorithm execution	mexcore:Execution
Data	Data item	DM-Data	N/A	mexcore:Example
Dataset	DM dataset	DataSet	Dataset	mexcore:Dataset
Feature	N/A	Feature	N/A	mexcore:Feature
DataCharacteristic	Data specification	DataCharacteristic	Dataset specification	N/A
DatasetCharacteristic	Dataset specification	DataSetCharacteristic	Data quality	N/A
FeatureCharacteristic	Feature specification	FeatureCharacteristic	N/A	N/A
Model	Generalization	DM-Hypothesis	Model	mexcore:Model
ModelCharacteristic	Generalization quality	HypothesisCharacteristic	Model Structure, Parameter, ...	N/A
ModelEvaluation	Generalization evaluation	ModelPerformance	Evaluation	N/A
EvaluationMeasure	Evaluation datum	ModelEvaluationMeasure	Evaluation measure	mexperf:PerformanceMeasure
EvaluationSpecification	N/A	N/A	N/A	N/A
EvaluationProcedure	Evaluation algorithm	ModelEvaluationAlgorithm	Performance Estimation	N/A

processes. At the level of both single algorithms and more complex workflows, it follows a very similar modeling pattern as described in the MLSchema. To support meta-mining, DMOP contains a taxonomy of algorithms used in DM processes which are described in detail in terms of their underlying assumptions, cost functions, optimization strategies, generated models or pattern sets, and other properties. Such a "glass box" approach which makes explicit internal algorithm characteristics allows meta-learners using DMOP to generalize over algorithms and their properties, including those algorithms which were not used for training meta-learners. DMOP also contains sub-taxonomies of ML models and provides vocabulary to describe their properties and characteristics, e.g. model structures, model complexity measures, parameters.

The MEX vocabulary. The MEX vocabulary [4] has been designed to reuse existing ontologies (e.g., PROV-O [6]) for representing basic machine learning experiment configuration and its outcomes. The aim is not to describe a complete data-mining process, which can be modeled by more complex and semantically refined structures. Instead, MEX is designed to provide a simple and lightweight vocabulary for exchanging basic machine learning metadata in order to achieve a high level of interoperability. Moreover, the schema aims to serve as a basis for data management of ML outcomes in the context of WASOTA [7]. The principal components are: `mex-algo`[4] which describes *algorithms* and *hyperparameters*, `mex-core`[5] which is the basis to describe an *experiment*, its *configurations* and *executions* and `mex-perf`[6], the layer to map the *outcomes* (i.e., performance measures).

4 Conclusions

In this extended abstract, we have presented ML-Schema, a lightweight schema for modeling ML domain. The ML-Schema aligns more fine-grained ontologies and vocabularies, some of which contain detailed vocabulary for representing meta-data on ML models. The vocabulary and axiomatization included in those resources may be used to make explicit the semantics of ML models, making them better interpretable for human users.

Acknowledgments

Gustavo Correa Publio acknowledges the support of the Smart Data Web BMWi project (GA-01MD15010B) and CNPq Foundation (201808/2015-3). Agnieszka Ławrynowicz acknowledges the support from the National Science Centre, Poland, within the grant number 2014/13/D/ST6/02076. Panče Panov acknowledges the support of the Slovenian Research Agency within the grant J2-9230. Hamid Zafar acknowledges the EU H2020 grants for the WDAqua (GA no. 642795) project.

[4] `http://mex.aksw.org/mex-algo`

[5] `http://mex.aksw.org/mex-core`

[6] `http://mex.aksw.org/mex-perf`

References

[1] Robert Arp, Barry Smith, and Andrew D. Spear. *Building Ontologies with Basic Formal Ontology*. The MIT Press, 2015.

[2] Anita Bandrowski, Ryan Brinkman, Mathias Brochhausen, Matthew H. Brush, Bill Bug, Marcus C. Chibucos, Kevin Clancy, Mélanie Courtot, Dirk Derom, Michel Dumontier, Liju Fan, Jennifer Fostel, Gilberto Fragoso, Frank Gibson, Alejandra Gonzalez-Beltran, Melissa A. Haendel, Yongqun He, Mervi Heiskanen, Tina Hernandez-Boussard, Mark Jensen, Yu Lin, Allyson L. Lister, Phillip Lord, James Malone, Elisabetta Manduchi, Monnie McGee, Norman Morrison, James A. Overton, Helen Parkinson, Bjoern Peters, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H. Scheuermann, Daniel Schober, Barry Smith, Larisa N. Soldatova, Christian J. Stoeckert, Jr., Chris F. Taylor, Carlo Torniai, Jessica A. Turner, Randi Vita, Patricia L. Whetzel, and Jie Zheng. The ontology for biomedical investigations. *PLoS ONE*, 11(4):1–19, 04 2016.

[3] Diego Esteves, Agnieszka Lawrynowicz, Pance Panov, Larisa N. Soldatova, Tommaso Soru, and Joaquin Vanschoren. Ml schema core specification. Technical report, W3C, October 2016. http://www.w3.org/2016/10/mls/.

[4] Diego Esteves, Diego Moussallem, Ciro Baron Neto, Tommaso Soru, Ricardo Usbeck, Markus Ackermann, and Jens Lehmann. MEX vocabulary: a lightweight interchange format for machine learning experiments. In *Proceedings of the 11th International Conference on Semantic Systems, SEMANTICS 2015, Vienna, Austria, September 15-17, 2015*, pages 169–176, 2015.

[5] C. Maria Keet, Agnieszka Lawrynowicz, Claudia d'Amato, Alexandros Kalousis, Phong Nguyen, Raúl Palma, Robert Stevens, and Melanie Hilario. The data mining optimization ontology. *J. Web Sem.*, 32:43–53, 2015.

[6] Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. Prov-o: The prov ontology. *W3C Recommendation*, 30, 2013.

[7] Ciro Baron Neto, Diego Esteves, Tommaso Soru, Diego Moussallem, Andre Valdestilhas, and Edgard Marx. Wasota: What are the states of the art? In *SEMANTiCS (Posters, Demos, SuCCESS)*, 2016.

[8] Pance Panov, Larisa N. Soldatova, and Saso Dzeroski. Ontology of core data mining entities. *Data Min. Knowl. Discov.*, 28(5-6):1222–1265, 2014.

[9] Nikolova I et al. Tcheremenskaia O, Benigni R. Opentox predictive toxicology framework: toxicological ontology and semantic media wiki-based opentoxipedia. *Journal of Biomedical Semantics*, 2012.

[10] Joaquin Vanschoren, Hendrik Blockeel, Bernhard Pfahringer, and Geoffrey Holmes. Experiment databases - A new way to share, organize and learn from experiments. *Machine Learning*, 87(2):127–158, 2012.

[11] Joaquin Vanschoren, Jan N Van Rijn, Bernd Bischl, and Luis Torgo. Openml: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.