Journal of Computer Languages 76 (2023) 101209

Contents lists available at ScienceDirect

Journal of Computer Languages

journal homepage: www.elsevier.com/locate/cola

A domain-specific language for describing machine learning datasets

Joan Giner-Miguelez [a,*], Abel Gómez [a], Jordi Cabot [b]

[a] Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya (UOC), Rambla del Poblenou, 156, Barcelona, 08018, Spain
[b] Luxembourg Institute of Science and Technology, 5, Av. des Hauts-Forneaux, Esch-sur-Alzette, 4362, Luxembourg

ARTICLE INFO

Keywords:
Datasets
Machine learning
MDE
Domain-specific languages
Fairness

ABSTRACT

Datasets are essential for training and evaluating machine learning (ML) models. However, they are also at the root of many undesirable model behaviors, such as biased predictions. To address this issue, the machine learning community is proposing a *data-centric cultural shift*, where data issues are given the attention they deserve and more standard practices for gathering and describing datasets are discussed and established.

So far, these proposals are mostly high-level guidelines described in natural language and, as such, they are difficult to formalize and apply to particular datasets. In this sense, and inspired by these proposals, we define a new domain-specific language (DSL) to precisely describe machine learning datasets in terms of their structure, provenance, and social concerns. We believe this DSL will facilitate any ML initiative to leverage and benefit from this data-centric shift in ML (e.g., selecting the most appropriate dataset for a new project or better replicating other ML results). The DSL is implemented as a Visual Studio Code plugin, and it has been published under an open-source license.

1. Introduction

Due to the centrality of data in machine learning (ML) applications, the processes involved in creating datasets are becoming more complex [1]. Dataset creation involves different teams at different stages, and comprises complex tasks such as data collection, labeling, and design. Despite this increasing complexity, recent studies have pointed out the lack of standard practices around the datasets used to train ML models [2,3]. For instance, they have detected a lack of formal documentation and fine-grained requirements as some of the main difficulties in complex data development processes.

Meanwhile, recent studies have revealed unintended consequences and negative downstream effects in the entire machine learning pipeline due to data issues [4,5]. For example, facial analysis datasets with a low proportion of darker-skinned faces may reduce the accuracy of face analysis models for that group, causing social harm [6]. As another example, because of the differences in language accents and styles, a natural language dataset gathered from Australian speakers may reduce the accuracy of models trained to support users in the United States [7]. In both cases, there is a need to save information about provenance or high-level analysis, such as the social impact on specific groups.

This situation has triggered growing concerns and generated new discussions within the research community about a *data-centric cultural shift* in the field of machine learning.[1] Standardization of data creation processes, formal documentation, and mature tools to adopt best practices are all common demands within the research community. As a result, recent works such as *Datasheets for datasets* [7–11], among others, have proposed the main guidelines for the creation of standard dataset documentation. In these proposals, the authors identify data aspects that may influence how the dataset is used or the quality of the ML models trained with it. Nevertheless, these proposals rely on textual descriptions in natural language, which presents clear challenges when it comes to automatically compute and analyze them, hampering their benefits.

We propose a domain-specific language (DSL) to precisely describe datasets according to the dimensions demanded by the aforementioned proposals. Our approach enables the standardization of dataset description, providing a structured format. Moreover, once the dataset is modeled using our DSL, it can then be manipulated with any of the existing model-driven engineering tools and techniques, such as model management [12] and model transformation [13] tools, opening the door to a number of (semi)automated application scenarios. To mention a few of them, we could: *(i)* check the quality and completeness of existing datasets – e.g., the quality of their labeling processes – ; *(ii)* compare datasets targeting the same domain to highlight their differences – e.g., comparing the infrastructure used to obtain the data –; *(iii)* search the most suitable dataset based on the requirements of the ML projects – e.g., searching for a dataset compliant with specific social concerns, such as specific demographic – starting what, in the

* Corresponding author.
E-mail addresses: jginermi@uoc.edu (J. Giner-Miguelez), agomezlla@uoc.edu (A. Gómez), jordi.cabot@list.lu (J. Cabot).

[1] https://datacentricai.org/

https://doi.org/10.1016/j.cola.2023.101209
Received 28 December 2022; Received in revised form 20 April 2023; Accepted 24 April 2023
Available online 2 May 2023

Table 1
Mapping of the contributions of the documentation proposals to each part of the DSL.

Documentation proposals	DSL parts								
	Metadata				Composition		Provenance		
	Description	Application	Authoring	Distribution	DataInstace	Attribute	Gathering	Labeling	Social Issues
Datasheets for datasets [8]	✓	✓	✓	✓			✓	✓	✓
Dataset Nutrition Labels [11]			✓		✓	✓			✓
Data Statements [7]	✓						✓	✓	
Data Readiness Report [15]	✓				✓	✓			
HF dataset cards [9]		✓	✓	✓			✓	✓	✓
GEM benchmark [10]		✓	✓	✓			✓	✓	✓
Data Cards [16]				✓	✓				
Montreal data license [17]				✓					
Deprecating datasets [18]				✓					
CrowdWorkSheets [19]								✓	

future, could become a dataset marketplace; *(iv)* generate other artifacts – documentation, code, etc. – from the dataset description; or *(v)* facilitate the replication of ML research results by better mimicking the conditions of the datasets used in the experiment when the same ones are unavailable — e.g., medical experiments where data has privacy concerns.

To support our approach, we present a Visual Studio Code plugin [14] that implements the DSL. The plugin enables users to import and annotate existing datasets while supporting all standard modern language features, such as autocompletion, syntax highlights, and code snippets, to name a few. To evaluate the feasibility and completeness of our proposal, we conducted a case study on three well-known datasets that served as examples in the aforementioned documentation proposals. Finally, to validate our approach, we conducted an empirical experiment with 17 practitioners both from research and industry working in the machine learning field.

The rest of the paper is organized as follows. Section 2 reviews the current dataset definition of the ML community. Section 3 presents the abstract syntax of the proposed DSL, while Section 4 presents the concrete syntax, and Section 5 the tool support. Section 6 presents the case study, and Section 7 the empirical evaluation we performed. Finally, Section 8 presents the related work, and Section 9 wraps up the conclusions and presents a set of detected challenges.

2. State of the art: Data documentation practices for ML

The need for proper documentation of datasets to be used in ML processes is clearly defined in the well-known paper *Datasheets for Datasets* [8] by Gebru et al. This work gets the idea of datasheets from the electronics field where every component has an associated datasheet as documentation. A key point of this proposal is the datasheet document structure. For each phase of a dataset description process – such as data design, gathering, and labeling – the authors pinpoint data aspects that could affect how the dataset should be used or the quality of ML models trained with it. They also ask for a discussion about bias and potential harms of the data contained in the dataset as part of its description. Gebru et al.'s work has been adopted by benchmark datasets in specific tasks like pedestrian detection [20], question answering [21], and machine translation [22], adapted to specific domains like healthcare [23,24], and proposed as an artifact to promote transparency in ML systems [25].

Complementing Gebru et al.'s work, other proposals zoom in on specific aspects of the dataset, such as the internal dataset composition and its relevant statistical properties. In particular, the *Dataset Nutrition Label* [11] presents a modular framework to provide an exploratory statistic analysis of the data. With it, dataset creators can signal relevant properties of the data using probabilistic models and ground truth correlations between attributes. This information facilitates the evaluation of the suitability of a dataset by data scientists for specific tasks. The *Data Readiness Report* [15] presents a similar proposal, deriving its design from the data readiness framework [26]. On top of the statistical analysis, it also defines a set of quality metrics for evaluating datasets' composition.

Focusing on specific aspects of the datasets, Sasha et al. [18] highlight the potential technical, legal, and ethical issues raised by deprecated datasets that continue to circulate, and proposes a dataset deprecation report to mitigate it. Geiger et al. [27] and Diaz [19], on the other hand, propose specific documentation for data annotators to facilitate the transparency of critical decision points at various stages of the data annotation pipeline. This study, like many others – as Vaughan et al. [28] – emphasizes the significance of considering crowd-worker contexts such as labor conditions, demographics, and potential biases among labeling tasks. Finally, in terms of data uses and distribution, Zhang et al. [29] focus on attribute privacy, offering definitions and methods to define it, while studies as Contractor et al. [30] and Benjamin et al. [17] are more concerned with data-specific licenses for machine learning.

Discussions regarding the quality of datasets for ML are also taking place in the natural language processing (NLP) field. For instance, *Data Statements* [7] emphasizes the need to annotate natural language datasets with additional metadata such as the demographics of data gatherers and data annotators – those labeling the data to prepare it for the training phase – and the specific context of the text in the dataset. Also, in this NLP field, we find other proposals such as *Dataset Accountability* [1], *Data Cards* [16], dataset documentation [9] implemented in the Huggingface's dataset repository,[2] and *GEM Benchmark* [10], that can be regarded as slight variations and simplifications of those already mentioned above.

3. A domain-specific language for describing ML datasets: Abstract syntax

This section describes our proposal for a DSL for describing machine learning datasets. The DSL comprises a set of modeling primitives that enable dataset creators to easily express all the relevant aspects of their datasets. Once described with our DSL, the annotated dataset can be automatically processed (e.g., search and comparison, analysis, documentation generation, etc.). This DSL has been inspired by the discussions and requirements presented in Section 2, and it is intended to be a superset of them. Therefore, Table 1 depicts the mapping between each section of the DSL and the aforementioned proposals.

The DSL is structured into three main parts. The *Metadata* part covers the description, applications, distribution, and authoring information about a dataset. The *Composition* part focuses on the structure

[2] Dataset cards documentation page: https://huggingface.co/docs/hub/datasets-cards.

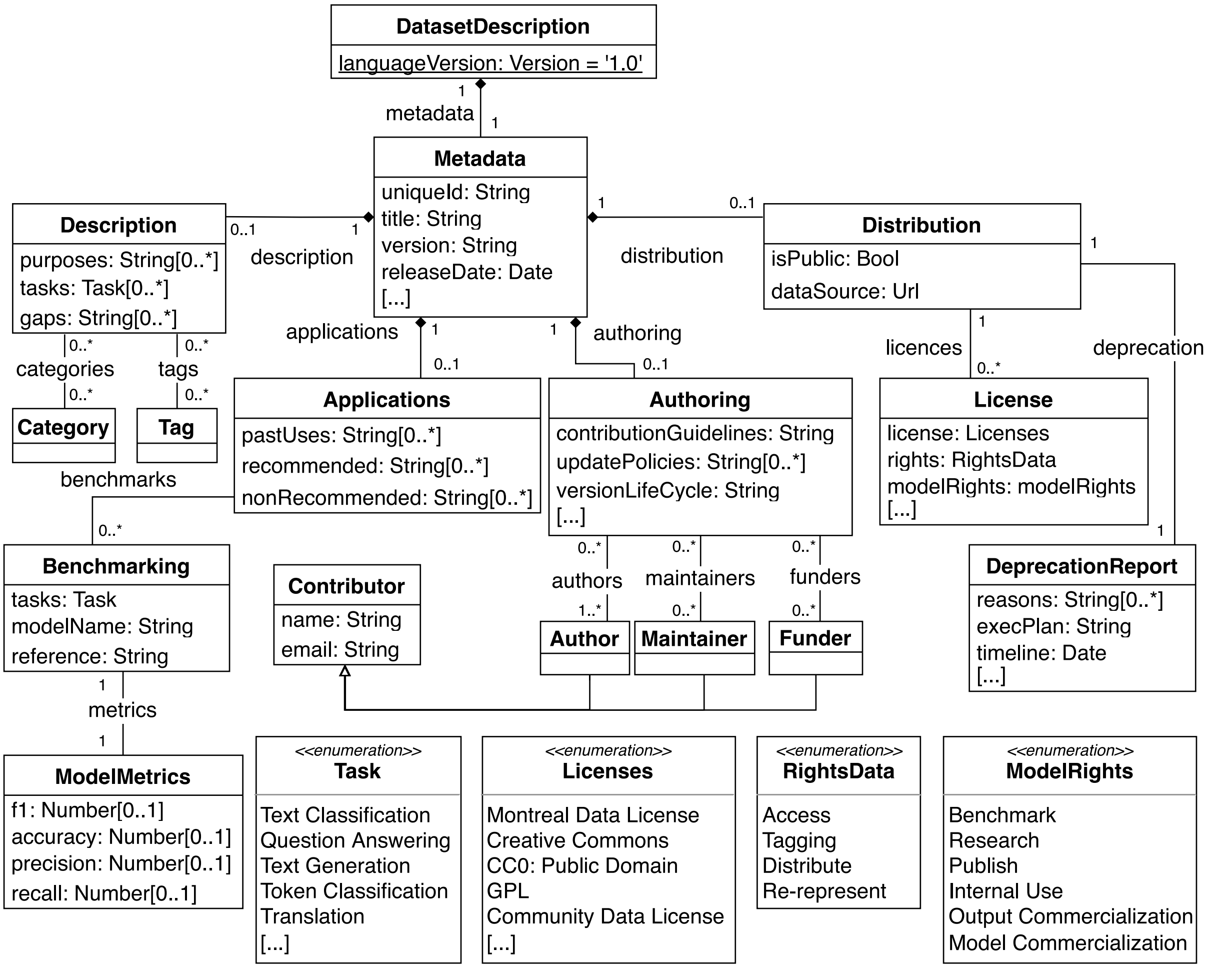


Fig. 1. Metadata model excerpt.

of the data instances and each specific attribute by describing relevant statistical concepts, quality metrics, and data consistency rules. Finally, the *Provenance and Social Concerns* part describes the gathering and labeling process conducted to build the dataset and its potential social biases when used to train ML models. Each mentioned part is independent and optional to annotate, so authors can annotate the parts that suit better their use case. Moreover, the DSL also includes the language version used to perform the description.

In the following, we go over these aspects and present the abstract syntax – i.e., metamodel – of the DSL, while in Section 4 we discuss the concrete syntax of the DSL, illustrated with examples.

3.1. Metadata

In the *Metadata* part, we have general information about the dataset. In Fig. 1, we can see that *Metadata* has attributes such as *uniqueId*, *title*, or the specific *version* number, to name a few. Additionally, *Metadata* is related to a *Description*, a set of *Applications* for the dataset, a *Distribution* setting the legal and licensing terms of a dataset, and finally to the *Authoring* part.

The *Description* part is composed of three attributes: *purposes*, *tasks*, and *gaps* — similarly to the *Datasheet for Datasets* proposal [8]. Using these attributes, creators can express the specific purposes the dataset was created for, the gaps it wants to fill, and the specific ML tasks this dataset is intended for. In addition, we have a *Categories* and *Tags* that allow classifying the dataset.

The *Applications* part expresses past usages of the data and recommends – or discourages – its use in specific scenarios. For example, creators can dis-recommend specific applications due to the potential social impact of the data, as Cao et al. [31] do regarding gender research. On the other side, the *Applications* part allows creators to annotate the results obtained with specific models trained with the dataset. These annotations, adopted by popular projects such as *Papers With Code*,[3] become relevant in the case of datasets used as a reference for specific tasks – such as sentiment analysis – in the ML field.

The *Distribution* part expresses the legal terms and the condition where the dataset can be used, is distributed, and will be deprecated in the future. In terms of licensing, the DSL supports some common licenses that software can be released under, and also supports the Montreal data license format [17], where the user can express, in addition to the license, the stand-alone rights of the data and the rights regarding its use in ML models. Finally, the DSL implements the deprecation report proposed by Sasha et al. [18] to indicate which are the plans to deprecate the dataset and the timeline associated with the deprecation process.

Finally, the *Authoring* part describes the *Contributors* of the dataset, such as the dataset *Authors*, the *Funders*, and current *Maintainers*. Regarding funders, creators can define—for example—the funders' type (public, private, or mixed) or the grants they have received—not shown in the figure for brevity purposes. In addition, creators can define maintenance policies. For instance, they can annotate how a user can collaborate with the dataset with the contributing guidelines, which are the dataset's intended lifecycle, the data update policies, or the place where the authors will report the detected errors.

[3] Example of a benchmark dataset: https://paperswithcode.com/dataset/cifar-10.

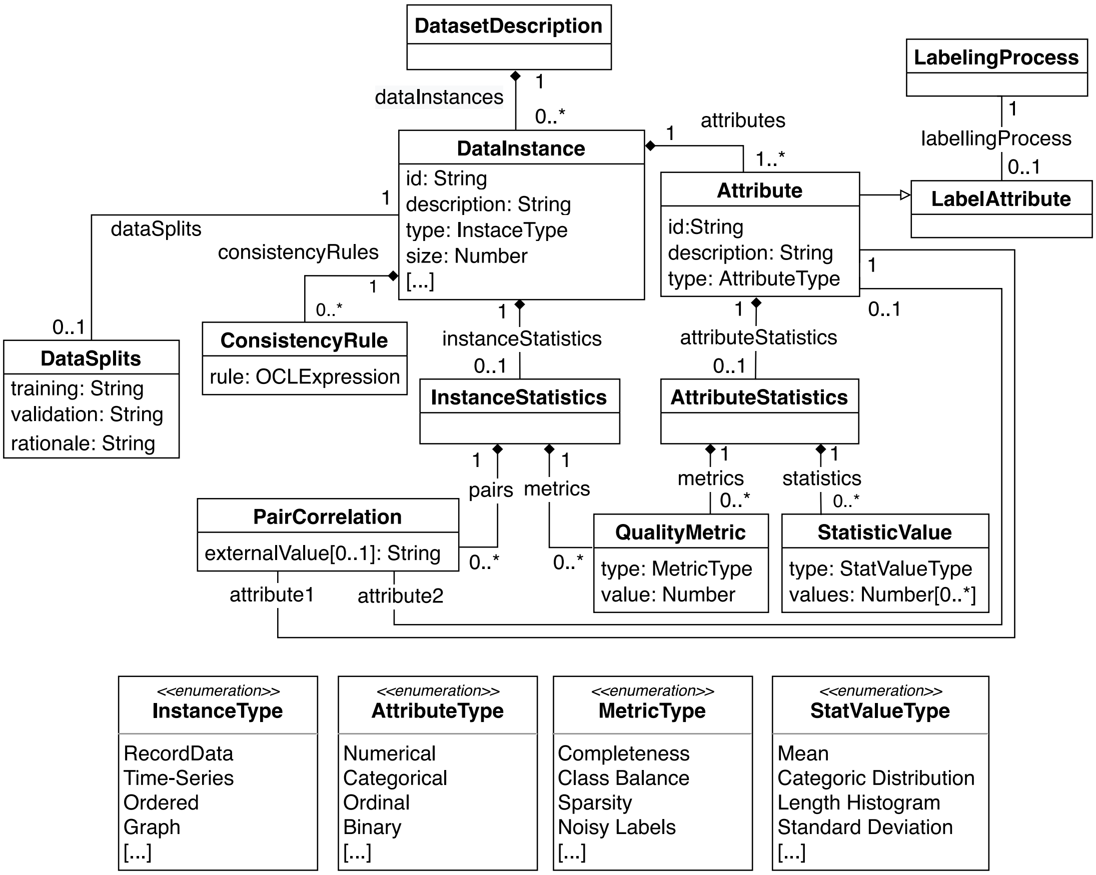


Fig. 2. Composition model excerpt.

3.2. *Composition*

In the *Composition* part, see Fig. 2, we can express aspects concerning the data structure, statistical description values, quality metrics, and the consistency rules that the dataset satisfies. This part is mostly inspired by the *Dataset Nutrition Label* and the *Data Readiness Report* proposals.

With the *Composition* modeling constructs, creators can define a set of data instances[4] and the *Attributes* composing these instances. At *DataInstance* level, creators can provide a general description of each instance, defining the *size* of the instance and its general *type* structure, such as record, time-series, or linked data. Besides, creators can use *InstanceStatistics* to express statistical information either by defining *pairCorrelations* between two attributes – or between one attribute and an external source of truth, such as national statistical records – or by expressing relevant quality metrics, such as class–category–balance, noisy labels, outliers, etc.

For each *Attribute*, creators can provide a description and specify the *type*, such as numerical or categorical. Then, if the attribute is the result of a labeling process (*LabelAttribute*), it can be linked with the details of the labeling process as shown later in the *Provenance* section. To express statistical information specific to a particular attribute, creators can use the *AttributeStatistics*. Creators can define *StatisticValues* such as *mode*, *mean*, and *standard deviation*, and a set of *QualityMetrics*, such as the completeness of the attribute, or its sparsity — i.e., number of values equal to 0.

Finally, a collection of *ConsistencyRules* can be attached to a *DataInstance*. These rules allow creators to express statements on the consistency of the data. As we could have a large variety of statements, we have adopted Object Constraint Language (OCL) [32] – widely used to express consistency rules [33] –, in particular, the *OCLExpression* class, for this purpose. This way, consistency rules could contain all the predefined functions and types available in OCL.

Not all of the information for each attribute should be included. It is up to the dataset authors to choose whether the information is important enough to warrant annotation. For example, statistical metrics such as the mean are meaningless for a gender attribute. However, expressing its categorical distribution may be extremely significant to know whether the dataset is gender-balanced so that ML developers may determine whether to utilize it in their models. Sometimes they may be looking for a balanced dataset, while others may want an unbalanced one if they are training a model for a specific community. On the other hand, the level and detail of information for each property, will vary depending on the domain, as some attributes may be more significant than others. For example, a melanoma patient's age group may be more critical than their civil status and should be annotated in more detail.

3.3. *Provenance and social concerns*

In the *Provenance and Social Concerns* part, we focus on the datasets gathering and labeling processes, the applied data preprocess, and the potential social impact of the data. From the *Data Statements* [7] and *Crowdworksheet* [19] proposals, we have included several aspects relevant to the gathering and labeling process, such as the demographics of the process and the teams, the validation mechanism of the annotation process or the requirements of both. From the *HF dataset cards* [9] and *GEM benchmark* [10], we have included the *DataPreprocess* part, and

[4] Notice that, in the data science field, an *instance* is understood as the group of attributes of an entity in the real world, similarly to the concept of *class* in the modeling community and therefore radically different from our typical understanding of the word *instance* in object-oriented programming.

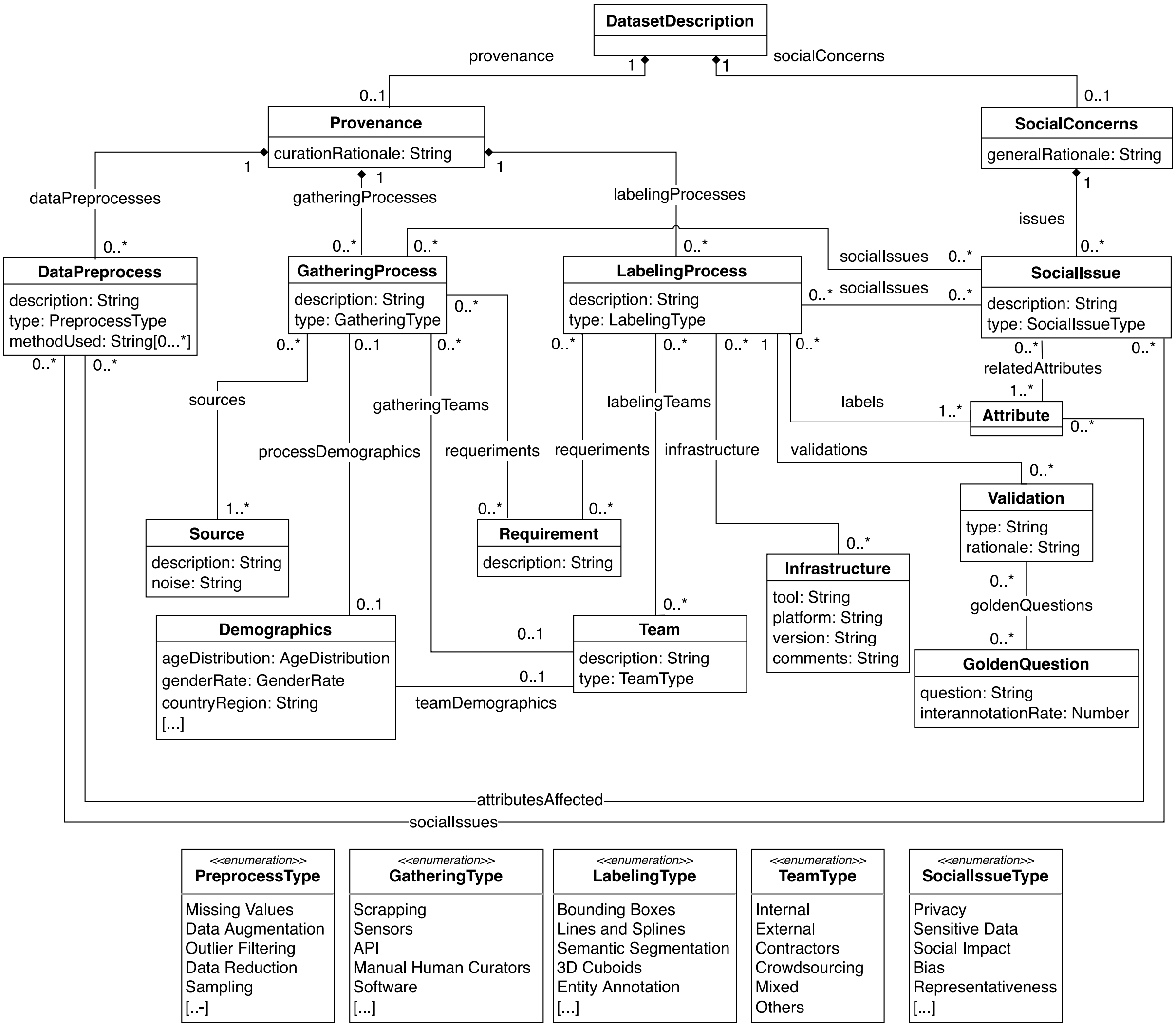


Fig. 3. Provenance and Social Concerns model excerpt.

finally, from *Datasheets for Datasets* [8] proposal, we have taken the description of the social aspects. In Fig. 3, we see an excerpt of the *Provenance* and *Social Concerns* part of our proposal.

Provenance has a *curationRationale* that allows creators to describe the general process and rationale to build the dataset. Moreover, a set of details on the *GatheringProcess* and *LabelingProcess* can be defined. Both processes have similarities, such as they both include information on the *Team* contributing, the process *Requirements*, and the *SocialIssues* that could result from them. Regarding the *Team*, we can describe and define the team's type and demographics. In terms of *Requirements*, we can annotate the guidance and requests given to the teams to collect or annotate the dataset, and finally, we can trigger each process with a particular social issue.

As such, these social issues can be defined in the *SocialConcerns* class. For instance, a gathering process done in only one country may lead to a geographical bias issue, and therefore, this process could be related to a *Social Issue* of type *Bias*. In addition, we can map this *Social Issue* to a specific *Attribute*, indicating the particular attributes that could be affected.

On the other hand, specific to the *GatheringProcess*, we can also define a set of data *Sources*. For example, a dataset built from IoT sensors could have different sensors with different noise characteristics (such as tolerance). Finally, and specific to the *LabelingProcess*, we have the concrete list of *labels*, which relates to the specific *Attributes* that are the result of this process. On the other hand, we can define the *Infrastructure* – the set of tools – used to perform the annotations and the *Validation* mechanisms for the annotation process. Inside the *Validation* mechanisms, the *GoldenQuestions* refer to the questions that have been sent to annotators, such as "Is there any dog in the image?", together with the agreement rate between annotators for each question (inter-annotation agreement). For instance, a question with a low annotation rate agreement could express low confidence in the annotations.

Finally, the *dataPreprocess* component allows creators to document all the processes performed on the data before publication. Cleaning missing values, feature construction processes, or augmenting data can raise social concerns. Consequently, together with describing the type and methods used in each process, creators can relate them to specific social issues.

4. A domain-specific language for describing ML datasets: Concrete syntax

In this section, we present our DSL's concrete syntax. To do so, we present the grammar implementation, together with an example describing the *ISIC Melanoma Classification Challenge Dataset* [34] — from now on, *Melanoma dataset*. This dataset, intended to detect melanoma from pictures of skin patients, can be considered as a benchmark for dataset documentation since some of the proposals presented in Section 2, such as the *Dataset Nutrition Labels* [11], use it as an example. The concrete syntax full implementation can be found at the published language reference guide.[5] The *Melanoma dataset* example shown in

[5] http://hdl.handle.net/20.500.12004/1/A/DML/002

Listing 1: DSL grammar excerpt

```
 1 Metadata:
 2     'Description:'
 3         (description=STRING |
 4         (('Purposes:' purposes=STRING)? ('Tasks:''[' tasks+=MLTasks ((','tasks+=MLTasks)*']')?)?)
 5         'Tags:' tags+=Tag ((','tags+=Tag)*)?
 6     ('Citation:' citation=Citation)?
 7 Distribution:
 8     ('Licences:' (licence=CommonLicences | licence=Other))?
 9                                    [...]
10 Other: "Other" name=STRING;
11                                    [...]
12 Composition:
13     'Rationale:' compodesc=STRING
14     'Instances:' (instances = DataInstances)
15 DataInstance: name=ID
16     ('Attributes:' (instanceAttributes+=Attribute)*)?
17 Attribute: name=ID
18     'Description:' attdesc=STRING
19     ('OfType:' ((attType=Categorical)|(attType=Numeric|attType=Ordinal|attType=Binary)))?;
20                                    [...]
21 Provenance:
22     'Curation Rationale:' curation=STRING
23     'Gathering processes:' (gatheringProc+=GatheringProcess)*)?
24                                    [...]
25 GatheringProcess: name=ID
26     'Source:' source=DataSource
27     ('Related Instance:' mapInstance=[DataInstance])?
28     ('Social Issues:' labelSocialIssues=[SocialIssue])?
29                                    [...]
30 SocialIssue:
31     'IssueType:' iType=SocialIssueType
32     ('Related Attributes:'.('relAttributes:'rAtt+=[Attribute])*)?
33                                    [...]
34 MLTasks returns string: "Text-classification"|"Question-answering"|  //..
35 SocialIssueType returns string: "Privacy" | "Bias" | "Sensitive Data"//..
```

this section has been created using the tool described in Section 5, and its complete version can be found in a public repository.[6]

4.1. The grammar

The grammar has been defined using the extended Backus–Naur form (EBNF), and in Listing 1 we can see an excerpt of the implementation. Looking at the listing, we defined the OR operator using the "|" symbol. For instance, in lines 3–5 of the listing, we see that the *Description* attribute can be either a single string, or a set of *Purposes*, *MLTasks* and *Tags*. Another example can be seen in lines 17–19, where depending on the Attribute's type, the *Attribute* can be annotated with different statistical values. For instance, annotating the mean for a categorical value, for example, may not make much sense.

Furthermore, we use the "?" symbol to describe the optional attributes, such as *Citation* on line 6, where we can optionally choose to annotate the desired citation of the dataset. In contrast, we use the "*" symbol to denote zero to any multiplicity relationship. For example, in line 5, a set of Tags (separated by commas) can be annotated, or in line 14, a set of *Attributes* in a *DataInstance* can be annotated. It could be noted that these *Attributes* are stored as an array in the *instaceAttributes* variable using the "+=" symbol.

On the other hand, the grammar defines semantic rules, allowing the user to choose between a list of concepts. For instance, in line 4, users can choose between the *MLtasks* present in line 34, and in line 31, users can choose between the *IssueTypes* of line 35. Finally, we have used brackets to describe cross-references, such as in line 27, where we can associate the *GatheringProcess* with a given *Social Issue*, or in line 32, where we can associate *Attribute* with a specific *Social Issue*. As an example, an attribute identifying people's "gender" may be associated with a social issue of gender representativeness.

4.2. Syntax example

To illustrate the concrete syntax of the DSL, in Listing 2, we present an example of the *Melanoma dataset* description. This dataset is composed of images of the patient's skin lesion annotated by dermoscopy experts intended to detect melanoma. To explain the example, we will go over the different parts of the DSL – metadata, composition, provenance and social concerns – explained in the last section. As mentioned, the complete example can be found in the public repository.

Metadata: Looking at the listing, we can see an excerpt of the Metadata part from lines 1 to 11. In this part, we can see the *Description* presented in the previous section where the purpose is to *advance in the medical image innovation*, the task of what dataset was designed for is *classification*, and have a set of annotated *Tags*. In line 7, we see the *Distribution* part, where the dataset is licensed with the "CC BY-NC 4.0" license and indicates that the data can be distributed freely, but the models trained with the data can only be used by research purposes following the Montreal data license [17] format. Finally, in line 11, we can see how the dataset is not recommended to train ML Models that work on real patients due to representativeness issues.

Composition: On the other hand, from lines 13 to 28, we have an excerpt of the Composition part. The *Melanoma dataset* is composed of a *DataInstance* called *skinImages* (line 16), which contains attributes such as *benignant_malignant* (line 19) and *ageGroup* (line 18). We see that *benignant_malignant* is of type categorical, and it is a label, so it is associated with a *LabelingProcess* called *DiagnosisLabel* (see line 44). In

[6] http://hdl.handle.net/20.500.12004/1/A/DML/003

Listing 2: Example excerpt of the SIIM-ISIC Melanoma classification dataset description

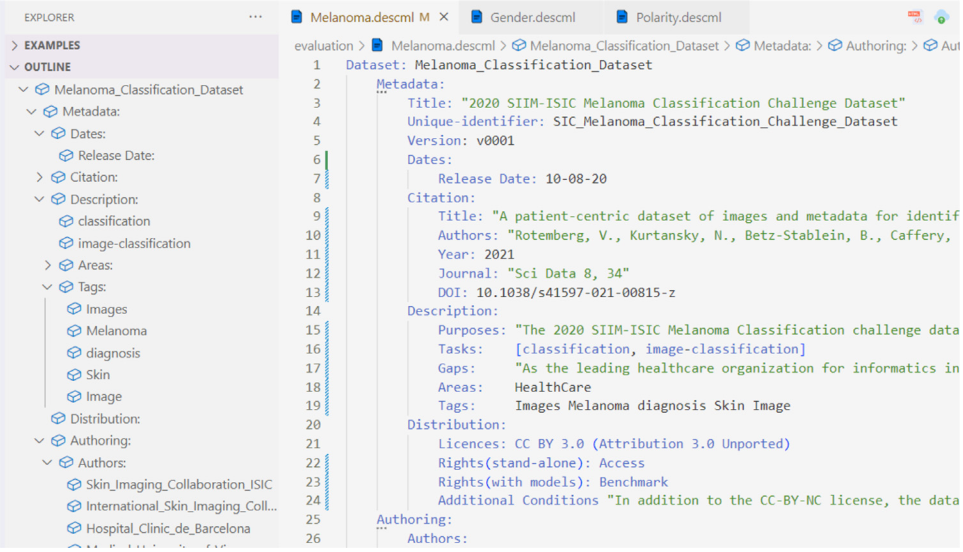
```
 1 Metadata:
 2   Title: "2020 SIIM-ISIC Melanoma Classification Challenge Dataset"
 3   Description:
 4     Purposes: "Advance medical image innovation ..."
 5     Tasks:   [Image-classification]
 6     Tags:    Images, Melanoma, Diagnosis, Skin Image
 7   Distribution:
 8     License: CC BY-NC 4.0 (Attribution-NonCommercial 4.0 International)
 9     Rights:  Distribute  ModelRights: Research
10   Applications:
11     Non-Recommended:"Do not use over real patient as it is not representative of melanomas..."
12 [...]
13 Composition:
14   Rationale: "The dataset is composed of lesions patients images annotated with 8 attributes..."
15   Data Instances:
16     Instance: skinImages
17       Attributes:
18         Attribute: ageGroup [...]
19         Attribute: benignant_malignant
20           Description: "Medical diagnosis of the patient"
21           Labeling process: DiagnosisLabel
22           OfType: Categorical
23           Categorical Distribution: [ "beningnant": 45%, "malignant": 55% ]
24     Statistics:
25       Pair Correlation:
26         Between ageGroup and external source
27           From: "Official population indicator..." Rationale: "Similar age distributions"
28       Consistency Rules: Inv skinImages: (ageGroup >= 0)
29 [...]
30 Data Provenance:
31   Curation Rationale: "Different gathering process has been done by each healthcare..."
32   Gathering Processes:
33     Process: Melanoma_Institute_Australia
34       Description: "Practitioners taking pictures from ..."
35       Type: Manual Human Curators
36       Timeframe: From 1998 to 2019
37       Source: imagePictures
38         Description: "Melanoma Institute Australia and the Sydney Melanoma Diagnosis ..."
39         Noise: "Pictures were taken using cameras..."
40       Social Issues: skinColorRepresentative
41       Process Demographics: Countries: Australia
42 [...]
43   Labeling Processes:
44     Process: DiagnosisLabel
45       Description: "Medical staff visualizing images..."
46       Type: Image & video annotations
47       Labels: skinImages.benignant_malignant
48       Infrastructure: Tool: Tagger
49 [...]
50 Social Concerns:
51   Social Issue: skinColorRepresentative
52     Description: "Dataset is not representative in terms of skin colors for darker skins..."
53     IssueType: Bias Related Attributes: ImageId
```

line 23, we can see the *Categorical Distribution* ratio of benignant and malignant diagnoses in the dataset.

Moreover, in line 24, we describe a set of statistics regarding the whole *DataInstance*. More specifically, we express a *Pair Correlation* between an attribute and an external source of truth, inspired in the *Dataset Nutrition Labels* proposal. In line 26, we relate the *ageGroup* distribution of the dataset with a hypothetical official population indicator, arguing that the dataset is representative of age groups. Finally, we have defined one *Consistency Rule*, indicating that the *ageGroup* is always equal to or higher than 0. These rules implemented with an OCL expression give flexibility to dataset creators to express a diverse set of consistency rules.

Provenance and Social Concerns: In lines 30 to 53, we have an excerpt of the Provenance and Social Concerns part of the *Melanoma dataset*. In line 31, we describe the general *Curation Rationale*, which specifies that the dataset has been built thanks to the collaboration of different hospitals. In lines 32–41, we present an excerpt of the gathering *Process* for one of those hospitals, the *Melanoma Institute of Australia*. In this process, we provide a description, we define the type (in this case, *Manual Human Curators*), the time frame where data was collected, the data source and its potential noise, the *Social Issues* related with this process (in this case, the *skinColorRepresentative* issue), and finally, the *Process Demographics*.

On the other hand, in lines 43–49, we partially describe the Labeling *Process* by describing the type (in this case, image and video annotations) and mapping the labels to the specific attribute in the dataset. The attribute in our example is the *benignant_malignant* of the instance *skinImages*. Then, we describe the *infrastructure* used to annotate the



```
Dataset: Melanoma_Classification_Dataset
    Metadata:
        Title: "2020 SIIM-ISIC Melanoma Classification Challenge Dataset"
        Unique-identifier: SIC_Melanoma_Classification_Challenge_Dataset
        Version: v0001
        Dates:
            Release Date: 10-08-20
        Citation:
            Title: "A patient-centric dataset of images and metadata for identif
            Authors: "Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery,
            Year: 2021
            Journal: "Sci Data 8, 34"
            DOI: 10.1038/s41597-021-00815-z
        Description:
            Purposes: "The 2020 SIIM-ISIC Melanoma Classification challenge data
            Tasks:    [classification, image-classification]
            Gaps:     "As the leading healthcare organization for informatics in
            Areas:    HealthCare
            Tags:     Images Melanoma diagnosis Skin Image
        Distribution:
            Licences: CC BY 3.0 (Attribution 3.0 Unported)
            Rights(stand-alone): Access
            Rights(with models): Benchmark
            Additional Conditions "In addition to the CC-BY-NC license, the data
    Authoring:
        Authors:
```

Fig. 4. Overview of the tool UI.

data. In the example, the hospital has used an internally developed tool called Tagger. Finally, in lines 50–53, we have defined a *Social Issue* of type *Bias* that indicates that the dataset may not be representative in terms of skin color. This issue is similar to the facial analysis issue presented in the introduction of this work, and shows how this type of bias can be annotated using the language.

5. Tool support

To support our approach, we created *DescribeML* [14], a tool that implements the presented DSL. The tool is a Visual Studio Code (VSCode) plugin that guides practitioners through the documentation process by providing standard modern language features as well as a set of VSCode extension services to facilitate the dataset documentation process. In Fig. 4 we see an overview of the tool UI. The plugin was created with Langium [35], a low-code language engineering toolkit for creating textual DSLs, has been released under an open-source[7] license, and is available on the Visual Studio Code Market.[8]

As language features, the plugin provides syntactic and semantic highlights, autocompletion of the enumerations of the DSL, and a set of predefined code snippets. In parallel, the tool provides hints to dataset creators extracted from the presented works in Section 2. These hints aim to flatten the DSL's learning curve and ensure the proper usage of the DSL sections. In addition, the tools implement custom validations to ensure the correct usage of the attributes, such as the statistical values.

Aside from language features, the tool includes a set of VSCode extensions to help with the documentation process. The main goal of these extensions is to automate several parts of the documentation process and provide a structured workflow. In this sense, in Fig. 5, we can see the tool's usage workflow. The first step of the workflow is to manually create a *.descml* file to let the IDE detect that we are using our language. Then, users can use the data preloader service that generates a draft description file from existing data. With this service, practitioners do not need to start from scratch.

When the description is complete, practitioners can use the generator service to generate HTML documentation. In addition, this service can also generate a Schema.org [36] notation from a valid description, allowing search engines, such as *Google Dataset Search*,[9] to index the dataset. This service is a practical example of how we can compute a description and shows the benefits of describing a dataset in a structured format. Moreover, there is room to extend this generation to other data description initiatives, such as the *Data Documentation Initiative* [37] or the *Data Catalog Vocabulary* [38]. Fig. 6 shows an example of the HTML generated from the *Melanoma dataset* implementing Schema.org.

Regarding the developer experience, we chose the Visual Studio Code (VSCode) environment because it is one of the most popular development environments in the machine learning field. Moreover, we decided to keep the user interface similar to developing code with VSCode as we believe that the usage experience is then already familiar for developers of the ML community.

6. Case study

To validate the expressiveness of the DSL, we performed a case study by modeling three well-known datasets in the ML space. The datasets were chosen because they have a diverse provenance and composition, and have been already used as examples in the ML community discussions described in Section 2. This case study aimed to evaluate whether the DSL can express the concepts mentioned in the aforementioned works. Next, we present each selected dataset with some evaluation conclusions. The full descriptions of the datasets can be found in our open repository.[10]

6.1. The Gender Inclusive Coreference dataset

The *Gender Inclusive Coreference* [31] is a dataset for testing the performance of coreference resolution systems[11] on texts that discuss non-binary and binary transgender. The dataset is composed of natural text labeled with the mentioned coreference by the authors and gathered from a variety of sources from the web. It is documented using the *Datasheets for dataset* [8] format, and the documentation can be found in an open repository.[12]

We find that we were able to express all the terms expressed in the documentation. However, during the documentation, we detected specific relevant statistical values that are not directly supported by the DSL because the dataset has been built using the CONLL-U format.[13]

[7] https://github.com/SOM-Research/DescribeML

[8] http://hdl.handle.net/20.500.12004/1/A/DML/005

[9] Project's homepage: https://datasetsearch.research.google.com.

[10] http://hdl.handle.net/20.500.12004/1/A/DML/003

[11] A coreference system is a system that aims to automatically detect who is mentioned in a transcription, for instance, of a public parliament, to label data

[12] https://github.com/TristaCao/into_inclusivecoref/blob/master/GICoref/datasheet-gicoref.md

[13] CONLL-U is a format used to annotate data at the sentence level and the word/token level. Homepage: https://universaldependencies.org/format.html.

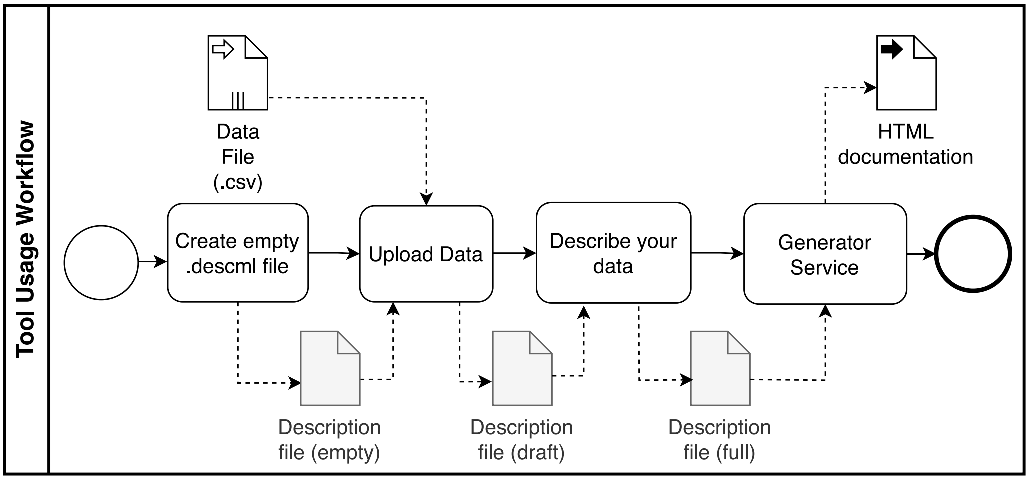
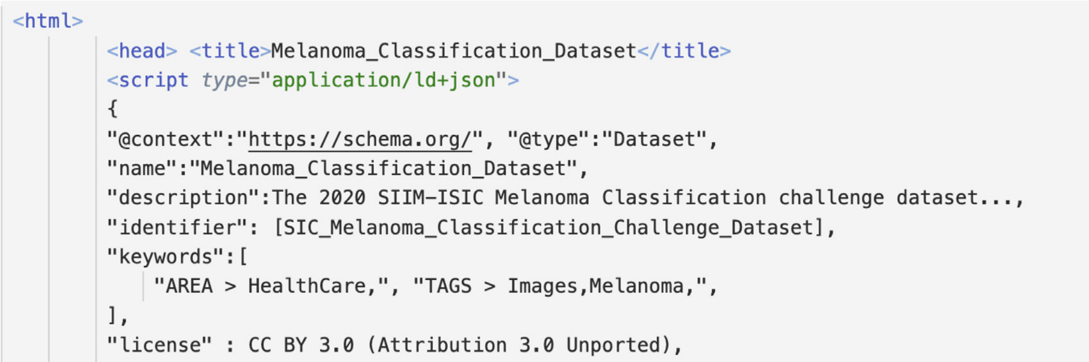


Fig. 5. Tool usage workflow.



```
<html>
        <head> <title>Melanoma_Classification_Dataset</title>
        <script type="application/ld+json">
        {
        "@context":"https://schema.org/", "@type":"Dataset",
        "name":"Melanoma_Classification_Dataset",
        "description":The 2020 SIIM-ISIC Melanoma Classification challenge dataset...,
        "identifier": [SIC_Melanoma_Classification_Challenge_Dataset],
        "keywords":[
            "AREA > HealthCare,", "TAGS > Images,Melanoma,",
        ],
        "license" : CC BY 3.0 (Attribution 3.0 Unported),
```

Fig. 6. Example of HTML code implementing Schema.org generated by the generation service.

For example, as shown in Listing 3, we used the *sparsity* over attribute *mentions* to express the ratio of tokens per mention — whether a token mentions a person or not. In CONLL-U format, this attribute is 0 when the token does not mention a person, so expressing the ratio of values besides 0 (as sparsity does) allows us to define the ratio of tokens per mention.

In this dataset, we found that the provided documentation is very extensive regarding provenance and social concerns. The authors point out several issues from the sources that may affect people. For instance, they point out that the data gathered from Wikipedia could identify individuals. Finally, we did not find any relevant issues in terms of structure.

6.2. Movie Reviews Polarity dataset

The *Movie Reviews Polarity* [39] is a benchmark dataset for sentimental analysis tasks and is composed of a set of movie reviews tagged with a sentimental flag – positive or negative – by a group of reviewers. The data is gathered through web scrapping from the website, and its composition is the extracted text with the mentioned labels. The dataset is documented using the *Datasheets for dataset* [8] format, and the documentation can be found together with the mentioned paper.

In terms of expressiveness, the data is formatted as ordinary tabular data, and we can describe all the documentation's relevant terms using our DSL. This dataset contains inappropriate content and private personal information, despite being widely used. In Listing 4, we can see how we have expressed this in the description. On the other side, we discovered a detailed description of the labeling requirements but no information regarding the dataset's annotators and their demographics. Finally, we see that the dataset documentation uses the gathering rationale to express essential details about the data composition.

6.3. SIIM-ISIC Melanoma Classification dataset

The *SIIM-ISIC Melanoma Classification* [34] dataset, used as an example in the previous section, is composed of a set of images and patient information tagged with a diagnosis label of melanoma. The dataset has been gathered by a set of different health institutions across

Listing 3: An excerpt of the sparsity attribute of the *Gender Inclusive Coreference* dataset description

```
Instance: Wikipedia
   Description: "Each instance consists of text that has been sentence separated, tokenized ..."
   Attributes:
      attribute: tokens
      attribute: mentions
         Quality Metrics:
            Sparsity: 7.5   [...]
```

Listing 4: Excerpt of the social concerns part of the *Movie Reviews Polarity* dataset description

```
Social Issue: inappropiateContent
 IssueType: Social Impact
 Description: "Some movie reviews might contain moderately inappropriate..."
 Related Attributes: text_body

Social Issue: personalInformation
 IssueType: Privacy
 Description: "Some personal information is retained..."
```

the world using different infrastructures and has been annotated by internal health practitioners of each institution. The dataset has been documented by the *Nutrition Label Project*, and its documentation can be found on the project homepage [11].

In terms of expressiveness, we could express all the relevant concepts of the documentation using our DSL. Also, we have seen an issue similar to the *Movie Review Polarity* dataset, where the gathering section is used to express relevant composition details. In addition, we miss relevant information regarding the maintenance policies of the dataset and ethical issues such as the individual patient's consent.

However, the relevant point here is the extensive explanation of the social concerns that may arise from the dataset. The authors express a racial representativeness issue because the dataset is not balanced in terms of skin color and, therefore, is not representative of the general incidence of melanoma. Even though it is a clear warning telling us not to use this dataset to train models intended to work with real patients, it has been difficult for us to detect it across the documentation. In contrast, as shown in the *Social Concerns* part of Listing 2, we have been able to express this issue more clearly.

6.4. Discussion

We can state that we were able to express all the relevant concepts present in the original datasets documentation using our DSL. But the opposite is not true, every dataset was missing important information, such as annotator's information or ethical concerns. Missing pieces of information were not obvious until an in-depth analysis of the original documentation was done. So far, we found that beyond uncovering and formalizing the information available in the datasets, our DSL helps highlighting the missing parts of the datasets documentation, prompting the authors to complete them.

On one hand, the structure of the information was not always clear in the analyzed dataset documentation. For instance, two out of the three modeled datasets have relevant composition details inside the explanation of the gathering process. This issue emphasizes the need of structuring the documentation creation process, aligning it with the dataset development process. As a consequence, we see that modeling the datasets with our proposed structured format helps in this task.

Finally, dataset-specific formats, such as CONLL-U.[14] in conference systems, have specific domain concepts that our DSL's expressiveness not directly support (such as the ratio of tokens per mention). During this case study, we see that our DSL could express these specific concepts indirectly, in that case using the sparsity of the attribute. Nonetheless, this situation opens the door to improving the expressiveness of the DSL for specific formats. However, there has been no research on the documentation requirements of these specific formats, indicating that this is a clear research path in the dataset documentation practices field.

7. Evaluation

In this section, we present the empirical evaluation we designed to validate the DSL's usability. The study sample was composed of 17 participants from 5 companies' data science departments and researchers from 3 groups working on machine learning issues. We designed the study both from the perspective of a data creator—analyzing the experience while creating the description—and from the perspective of a data consumer—studying the experience while understanding an already completed description. In consequence, the research questions this evaluation process aims to answer are:

RQ1 — How is the experience of a dataset consumer reading a previously created description?

RQ2 — How is the experience of a dataset creator writing a description using the DSL?

RQ3 — What is the experience while using the provided tool support?

The design of the evaluation was based in the methodology for conducting usability studies of Rubin et al. [40]. To assess the quality of the study, we evaluated a set of threats to validity, and assessed the post-interview quality using the *Empirical Standards for Software Engineering Research* [41]. In this section, we present the study's design, the research results, and the potential threats to validity, together with the mitigation strategies. All the documentation relevant to the empirical study, together with the anonymized raw data, can be found in our repository.[15]

7.1. Study design

The study is composed of 4 parts: *(i)* a previous screening test of the participants to evaluate the participant's suitability, two on-site exercises to evaluate *(ii)* the reading and *(iii)* writing of dataset description, and *(iv)* a semi-structured interview to be completed after the evaluation. Next, we describe each part in detail.

Screening test — The first part is a screening test intended to ensure participants have a minimum level of knowledge in the machine learning field and a minimum expertise in using and managing datasets for machine learning. This part has been done asynchronously, sending a set of questions to participants together with a video presenting the tool. Then, we collected the answers to check the participant's suitability.

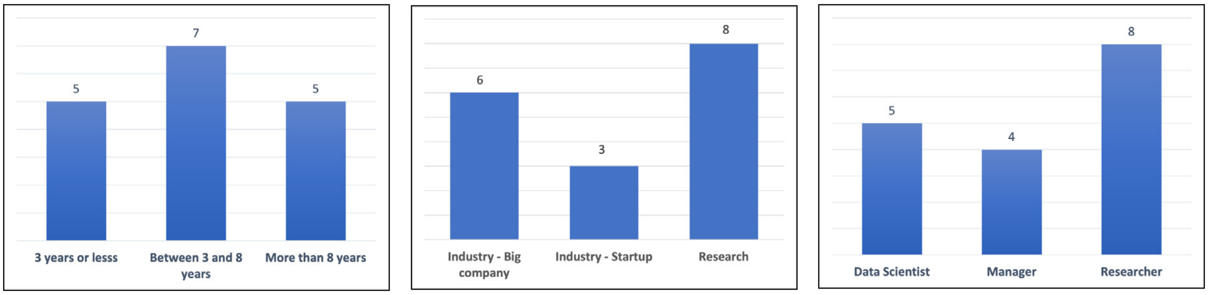
Reading exercise — The second part is a synchronous exercise aiming to answer the research question RQ1. A complete description of a complex dataset is given to the participants. Participants are asked to answer a set of questions regarding the dataset. The goal of this study is to evaluate the readability of the DSL and to determine if users and can easily find complex concepts, such as the distribution of a specific attribute, who annotate the data, or the methods used to collect it, in the definition.

[14] See footnote 13.

[15] http://hdl.handle.net/20.500.12004/1/A/DML/003

Table 2
Questions of the post-experiment interview.

Question #	Question	Possible answers
Q1	Is the tool easy to install and set up?	Very easy, easy, normal, hard or very hard
Q2	How the installation process can be improved?	Open answer
Q3	How easy was reading and understanding the description file?	Very easy, easy, normal, hard or very hard
Q4	How was your experience on reading/understanding the description file?	Open answer
Q5	How easy was to write a description of a dataset using describeML?	Very easy, easy, normal, hard or very hard
Q6	How was your experience in writing a description of a dataset using describeML?	Open answer
Q7	Do you think language like describeML is a useful approach?	Open answer
Q8	How useful do you think the tool is?	1 to 5
Q9	Do you have any final comments?	Open answer



(a) Participants experience (b) Participants affiliation (c) Participants job/position

Fig. 7. Evaluations participant's profile.

Writing exercise — The third part is a synchronous exercise aiming to answer the research question RQ2. A dataset published in a popular repository (Kaggle) is presented to the participants with its published metadata. Then, participants are asked to use the tool to describe the dataset using the data and the metadata present in the dataset online repository in 30 minutes. This exercise aims to evaluate the usability of our DSL for dataset creators in their documentation processes.

Post-evaluation interview — The final part is a semi-structured interview presented to the participants after the exercises. In Table 2, we can see the questions together with the format of the answers. Answers were collected asynchronously by the same person who conducted the study, and the goal was to extract qualitative insights from the participant's experiences.

In terms of the sample quality, 17 people accepted our invitation to participate in the evaluation and qualified for it. Following the rule proposed by Alroobaea et al. [42] that suggests a minimum of 16 +/−4 participants, we conclude that we have a good sample to evaluate our DSL. The sample was from two countries (Spain and France) and was divided into 53% of researchers and 47% of practitioners from the private sector. Fig. 7 depicts the profiles of the participants showing their job titles, years of expertise, and type (research or private companies).

The evaluations were conducted from September 30 to October 18, 2022. We organized 5 sessions of synchronous group evaluations with the selected participants during this time lapse. Participants were asked to send back the resulting description file of the *Writing* exercise to gain insight into any errors or misconceptions that occurred during the exercise. An author was present during the sessions, recording the time spent by each participant in the *Reading* and *Writing* exercises. We have analyzed the mentioned elements, and together with the evaluation answers, we present the results in the following section.

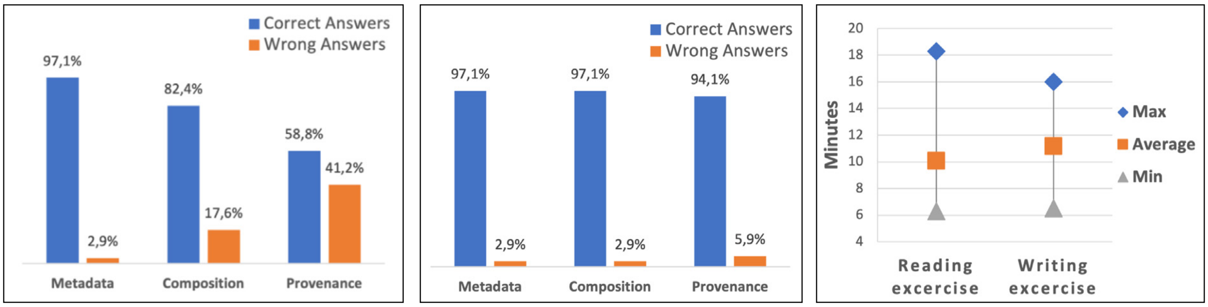
7.2. Empirical evaluation results

In Fig. 8, we can see the results of the reading and writing exercises of the evaluation together with the time spent by participants in each exercise. To illustrate the results, we grouped the answers by the DSL parts. For instance, all the questions in the metadata part were grouped, and the figure depicts the average of the correct and incorrect responses. We show the time spent divided by exercise, showing the minimum, the maximum, and the average. Finally, Fig. 9 shows the results of the post-evaluation interview questions that have finite values. We used a scale from 1 to 5 in question 8, and an ordered categorical scale – from very hard to very easy – in questions 1, 3, and 5.

RQ1 (Reading experience) — To answer the *RQ1*, we analyzed the answers to the reading exercise, the time spent in it, and the answers to questions Q3 and Q4 of the post-evaluation interview. As practically all the exercise questions were solved correctly and the time spent was approximately 10 minutes on average, we can state that participants had a good experience while reading and comprehending a description file. In addition, we see no significant difference between managers — more likely to be unfamiliar with code–and developers, and no significant difference between those with more and less experience.

However, we found a usability trade-off between a participant with manager and developer profiles. While the former shows an equivalent performance to developers during the reading exercise, some developers found it hard to read the document, expecting a "*more JSON style format*". However, we choose to design a specific grammar close to natural language to help managers to understand and use the language. Despite this trade-off, looking at the overall participants' performance, and at the time spent (managers spent 9:20 minutes on average, and developers 12:10 minutes on average), we can state that we found a good equilibrium in the understandability of our DSL for both profiles

On the other hand, some participants struggle to understand attributes with similar names. For instance, the two *Description Rationale* of the provenance and the *Description* of the gathering process were mentioned several times as "*hard to understand their difference*" during the reading exercise. In that sense, we found it crucial to develop the tool support by providing more hints to avoid these difficulties and flatten the learning curve.



(a) Results of the writing excercise

(b) Results of the reading exercise

(c) Participants time

Fig. 8. Evaluation results.

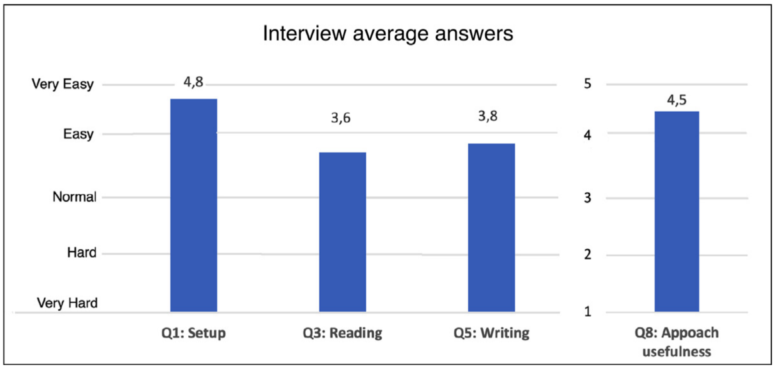


Fig. 9. Interview average answers.

RQ2 (Writing experience) — To answer the *RQ2*, we analyzed the answers regarding the writing exercise, the time spent in it, and the answers to questions Q5 and Q6 of the post-evaluation interview. In contrast with the reading exercise, we found more difficulties from the participants filling the provenance part (41,2% answered this part wrong).

Despite this, the time spent by the participants was equivalent to the reading exercise, and the writing experience was evaluated slightly better (3.8 over 5) than the reading experience. Because this exercise came in second place and was more difficult – it required participants to write sections of a description proactively – we may conclude that individuals improved their performance over the previous exercise. This improvement shows that participants learned about the DSL during the first exercise, demonstrating that the DSL is easy and quick to learn. Similarly to the reading exercise, there was no significant difference between managers and developers, where managers get the 85% of correct answer and spend 12:40 minutes in average, and developers get the 83.6% of correct answers and spend 12:32 minutes in average.

On the other hand, the template provided by the tool has been mentioned positively several times. Showing that one of the main contributions of the language is the structure it provides. However, some participants had expected more intelligent interactivity from the structure, such as automatic domain adaptation of the structure or prospective recommendations during the documentation process. This issue is also mentioned in recent empirical studies regarding dataset documentation practices [43] and shows the need for better tool support for documenting datasets.

Additionally, most participants expressed concerns about analyzing the dataset metadata in the Kaggle repository. The lack of information and clarity of some concepts in the original documentation has been a continuous demand during the different evaluation waves. In addition, several participants have expressed that this was the first time they realized this situation.

RQ3 (Tool support) — Finally, in terms of tool support, most participants stated that the tool was easy to install and appreciated very positively the setup process (see Fig. 9, Q1). However, we find a disparity of opinions regarding the Visual Studio Code platform. The participants that were familiar with that IDE expressed a very positive experience. In contrast, the participants unfamiliar with the IDE pointed it as the main usage limitation. Therefore, developing tool support independent of the VSCode IDE arises as a suggestion from the evaluation.

7.3. Lesson learned

A design trade-off between user profiles. Our language is designed to be used by a broad scope of users involved in different data stages. In this study, however, we discovered a design trade-off between developers who expected a "*more JSON style*" and managers who performed well in a more natural language style. In this compromise, we believe we have found a good balance between both profiles, since both have shown a positive performance and similar learning capabilities during the evaluation.

Raising the community awareness about data good practices. The evaluation also raised the participants' attention to the lack of good documentation in public dataset repositories such as Kaggle. Participants frequently mentioned this issue during the evaluation, and most mentioned it was the first time they realized it. Therefore, we can conclude that our DSL could help raise the community's awareness about good data practices.

Evolving the tool support. On the other hand, participants have asked for more intelligent interactivity from the tool support. For instance, an automatic adaptation of the structure depending on the informed fields by the user and prospective recommendations (such as a value of an informed field could lead to a potential bias) during the documentation process. In addition, some of them find the usage of the IDE (VSCode) as one of the main limitations. In conclusion, despite being well evaluated by participants, there is a path to improve the tool support by developing intelligent interactivity features and by building it independently of the VSCode platform.

7.4. Threats to validity

One of the fundamentals of every empirical study is the validity of the results. To assess the quality of our evaluation processes, we have identified a set of validity threats. Next, we present each one classified into four categories: internal, external, construct, and conclusions validity [44]. Together with each identified threat, we explain our mitigation approaches.

In terms of *internal validity*, we detected that all the participants were volunteers. This can conduct in bias due to the higher motivation of the participants, as volunteers can tend to have better performance. We do not consider this issue critical, as our proposal is intended for users working in the ML field instead of the general population. However, we have considered it during the work's conclusion. On the other hand, if participants know about previous session evaluations (e.g. feedback or tips from other participants), their behavior can be affected. In our case, we kept each group's results anonymous, and we distributed the participants working in the same place in the same study session.

Finally, the usage of the Integrated Development Environment (IDE) tied to the tool support (DescribeML) has been a limitation for some participants. Participants without experience with this IDE can show a limitation in their performance, and this could lead to a bias in its behavior. To mitigate this, we have shared explicit manual installations and set up instructions with participants. On the other hand, we can understand the IDE setup as a natural factor in DSL adoption. Therefore, we will take this issue into account in the study's conclusions.

Regarding *external validity* threats, having a not representative population for the study could lead to bad results. In our case, to minimize this threat, we have selected data engineers and managers in the data science departments, as well as researchers in software engineering and ML fields.

In terms of *construct validity* threats, some people may fear being evaluated, and others may have different expectancies about the experience of participating in a scientific evaluation. To mitigate this, we tried to create a comfortable environment during the different sessions, responding to the participant's questions and leaving space for informal conversation between them. In addition, to avoid different treatments between groups, we developed, published, and shared with the participants a detailed protocol that can be found in the tool repository.

Finally, as a *conclusion's validity* threats, we detected that our study suffer from a high heterogeneity group, as we choose practitioners from different fields (research, private), different company types (startup, big companies), and different roles (managers and engineers). To avoid this, we have performed a screening study, asking a set of questions to the participants to ensure their roles and a minimum shared experience in the field. This information was taken into account in the evaluation conclusions. Lastly, in terms of sample size, the number of participants in the study may be insufficient to be considered representative. To mitigate this, we compared with the rule that Alroobaea et al. [42] that suggest a minimum of 16 +/−4 participants. Since the number of participants has been 17, we can consider the sample size satisfactory.

8. Related work

In Section 2, we looked at data documenting initiatives within the ML community, and we concluded that there is a need to formalize and structure these initiatives, being these the motivation of our proposal. This section reviews the initiatives to describe data outside the ML community, and examines the model-driven practices applied in the ML field, specifically, the DSL proposals intended to facilitate ML and dataset-related tasks. After the review, we can conclude that our proposal is the first DSL aimed to facilitate the description of ML datasets covering the ML community's description needs.

8.1. Data description initiatives beyond the ML field

Several initiatives have been growing to describe data in the last few years. In terms of linked data over the Web, initiatives such as the *Dublin Core Metadata Initiative* (DCMI) [45], and the *Resource Description Framework* (RDF) [46] developed by the *World Wide Web Consortium* (W3C), have led the basis of the development of structured metadata to support resource discovery on the Web. As an example of a work based on this basis, we have the *Data Catalog Vocabulary* [38]. This proposal is a specific data format intended for data catalogs published over the Web and is widely used in the case of open data portals. Furthermore, public administrations, such as the European Union, published its open data catalogs following a custom specification of this format.[16] As another example of a specific RDF application, we found *Schema.org* [36], a data vocabulary designed to work for search engines. Compared to our proposal, these works have a broader scope, focusing on the discoverability of the linked data, while ours, focus on the particular needs of the ML field stated in Section 2.

Of particular interest for the data provenance component of our DLS is the W3C Prov [47] initiative. W3C Prov aims to support the interchange of provenance information on the Web and defines provenance as the information about entities, activities, and people involved in producing data. From a reusability perspective, works as *Sciunits* [48] also proposes to add standardized provenance information, in this case, to enable the reusability of research objects. In addition, beyond the field of computer science, initiatives as the *Data Documentation Initiative* [37] also work on data provenance. In that case, providing a standard for describing the data produced by surveys and other observational methods in social, behavioral, economic, and health sciences. Concerning our work, we share a common definition of provenance, but we go further than that, also focusing on the different steps of a data creation pipeline, such as preprocessing, annotation, and gathering, and on the social concerns that may arise if the data is used in ML models.

To sum up, although each of these solutions has unique characteristics and is intended for different scenarios, the root of all is the same: defining data. Therefore, it will be definitely interesting to create bridges that help to exchange and interoperate data among these different proposals to enable all of them to include more ML-centric descriptions as part of their schemas.

8.2. DSLs in the ML field

Beyond the data-centric approaches described above, we have started to see works presenting some kind of DSL to help in several ML tasks. We have proposals aimed at facilitating DevOps approaches for ML pipelines such as *OptiML* [49], *ScalOps* [50] or *StreamBrain* [51]; proposals targeting the creation of ML components such as *DeepDSL* [52], *DEFine* [53], *AIDsl* [54] and *MD4DSPRR* [55] for describing deep neural networks and cross-platform ML applications; or proposals like *ThingML2* [56] that look to integrate IoT components in ML pipelines, and *LEV4REC* [57] that aims to facilitate the building of recommender systems.

Additionally, there are works tied to particular tools or techniques for engineering ML, such as *TensorFlow Eager* [58], a DSL built on top of *TensorFlow* to help practitioners in the developments processes of ML

[16] EU DCAT specification homepage: https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/11.

artifacts, and Hartmann et al. [59], that propose a meta-model for the meta-learning technique for building ML artifacts. Graphical modeling tools themselves have been also extended, to a certain extent, with ML units to define workflows involving the execution of some type of ML task (*Knime*[17] would be a representative example in this category). More on the dimension of social concerns, *Arbiter* [60] is a DSL for expressing ethical requirements in ML training processes together with annotations that enable ML experts to describe the training process itself.

Focusing on datasets for ML, *SEMKIS-DSL* [61] is designed to simplify the specification of dataset requirements. These specifications are then used to augment the datasets, resulting in better data for the ML components. Celms et al. [62], on the other hand, present DSLs to ease the engineering (versioning and storage) of data management in deep learning projects, and finally, de la Vega et al. [63] present a DSL to facilitate the dataset selection and formatting processes. While these works focus on the requirement elicitation stage or data management aspects, our work captures a broader scope of data aspects, considering the data's provenance, composition, and social concerns.

9. Conclusions and further work

In this work, we have presented a DSL for describing datasets for machine learning together with a Visual Studio Code plugin to assist practitioners during the dataset description process. We have assessed the expressiveness of our approach by performing a case study against the documentation of state-of-the-art works presented in Section 2. In addition, we performed an empirical experiment to validate the usability of our DSL from a data consumer and data creator perspectives.

We believe our DSL is a step forward towards the standardization of dataset descriptions and its future impact in achieving higher quality ML models, especially from a social perspective (fairness, diversity, absence of bias, etc.). However, we have identified as future work a set of challenges that the research community should face before completely achieving these goals. Some of the identified challenges are:

Uncertainty in datasets descriptions. Dataset authors may not always be completely sure about some aspects of the dataset — e.g., the provenance, the quality of some attributes, or the confidence in specific labels. We envision the need to express uncertainties in dataset description models – see Muñoz et al. [64] for instance – to enable the annotation of our DSL elements with uncertain values and expressions.

Expressing commercial usage and distribution aspects. Not all datasets need to be free. Indeed, data collection and curation are time-intensive tasks. Therefore, beyond licensing information [17], we envision additional DSL primitives to express more complex usage rights based on a variety of business models — e.g., royalties derived from the applications of the ML models trained with the dataset.

Describing ML models. Beyond datasets, there is also interest in the community to describe ML models and other elements of an ML pipeline. Describing models and the different steps of the ML pipeline will help us analyze potential root causes of undesired behaviors from an end-to-end point of view of the ML applications. As such, we envision the extension of our DSL to cover as well the dimensions proposed by existing model documentation proposals embracing the complete ML lifecycle, such as Mitchell et al. [65,66] or *FactSheets* from Sokol et al. [67].

Dataset reverse engineering. Relevant datasets in the field are starting to be published along with accompanying documentation. This documentation, however, is in the form of natural text, and is contained in technical reports, or scientific papers published alongside the data. In that regard, there is room to investigate NLP techniques to leverage such documentation to extract from the text some of the dimensions of the DSL and used them to pre-populate the dataset description file.

[17] https://www.knime.com/

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The dataset descriptions examples and presented tool's are published in an open repository.

Acknowledgments

The research described in this paper has been partially supported by the Spanish government (LOCOSS - PID2020-114615RB-I00), the AIDOaRt project, which has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 101007350. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Sweden, Austria, Czech Republic, Finland, France, Italy, and Spain. Jordi Cabot is supported by the Luxembourg National Research Fund (FNR) PEARL program, grant agreement 16544475.

References

[1] B. Hutchinson, A. Smart, A. Hanna, E. Denton, C. Greer, O. Kjartansson, P. Barnes, M. Mitchell, Towards accountability for machine learning datasets: Practices from software engineering and infrastructure, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 560–575.

[2] N. Nahar, S. Zhou, G. Lewis, C. Kästner, Collaboration challenges in building ML-enabled systems: Communication, documentation, engineering, and process, in: 44th International Conference on Software Engineering (ICSE '22), Vol. 1, 2022, p. 3.

[3] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, L.M. Aroyo, "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–15.

[4] A. Paullada, I.D. Raji, E.M. Bender, E. Denton, A. Hanna, Data and its (dis) contents: A survey of dataset development and use in machine learning research, Patterns 2 (11) (2021) 100336.

[5] C. Renggli, L. Rimanic, N.M. Gürel, B. Karlaš, W. Wu, C. Zhang, A data quality-driven view of mlops, Data Eng. (2021) 11.

[6] A. Khalil, S.G. Ahmed, A.M. Khattak, N. Al-Qirim, Investigating bias in facial analysis systems: A systematic review, IEEE Access 8 (2020) 130751–130761.

[7] E.M. Bender, B. Friedman, Data statements for natural language processing: Toward mitigating system bias and enabling better science, Trans. Assoc. Comput. Linguist. 6 (2018) 587–604.

[8] T. Gebru, J. Morgenstern, B. Vecchione, J.W. Vaughan, H. Wallach, H.D. Iii, K. Crawford, Datasheets for datasets, Commun. ACM 64 (12) (2021) 86–92.

[9] A. McMillan-Major, S. Osei, J.D. Rodriguez, P.S. Ammanamanchi, S. Gehrmann, Y. Jernite, Reusable templates and guides for documenting datasets and models for natural language processing and generation: A case study of the HuggingFace and GEM data and model cards, in: Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics, ACM, Online, 2021, pp. 121–135.

[10] S. Gehrmann, T. Adewumi, J. Zhou, The gem benchmark: natural language generation, its evaluation and metrics, in: 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021), Bangkok, Thailand (online), August 5-6, 2021, Association for Computational Linguistics, 2021, pp. 96–120.

[11] S. Holland, A. Hosny, S. Newman, J. Joseph, K. Chmielinski, The dataset nutrition label, in: Data Protection and Privacy, Volume 12: Data Protection and Democracy, 12, Bloomsbury Publishing, 2020, p. 1.

[12] A. Boronat, J.A. Carsí, I. Ramos, Exogenous model merging by means of model management operators, Electron. Commun. Eur. Assoc. Softw. Sci. Technol. 3 (2006).

[13] F. Jouault, F. Allilaire, J. Bézivin, I. Kurtev, ATL: A model transformation tool, Sci. Comput. Program. 72 (1–2) (2008) 31–39.

[14] J. Giner-Miguelez, A. Gómez, J. Cabot, DescribeML: a tool for describing machine learning datasets, in: Proceedings of the 25th International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings, 2022, pp. 22–26.

[15] S. Afzal, C. Rajmohan, M. Kesarwani, S. Mehta, H. Patel, Data readiness report, in: 2021 IEEE International Conference on Smart Data Services, SMDS, IEEE, 2021, pp. 42–51.

[16] M. Pushkarna, A. Zaldivar, Data cards: Purposeful and transparent documentation for responsible AI, in: 35th Conference on Neural Information Processing Systems, 2021.

[17] M. Benjamin, P. Gagnon, N. Rostamzadeh, C. Pal, Y. Bengio, A. Shee, Towards standardization of data licenses: The montreal data license, 2019, arXiv preprint arXiv:1903.12262.

[18] A.S. Luccioni, F. Corry, H. Sridharan, M. Ananny, J. Schultz, K. Crawford, A framework for deprecating datasets: Standardizing documentation, identification, and communication, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 199–212.

[19] M. Díaz, I. Kivlichan, R. Rosen, D. Baker, R. Amironesei, V. Prabhakaran, E. Denton, CrowdWorkSheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 2342–2351.

[20] I. Seck, K. Dahmane, P. Duthon, G. Loosli, Baselines and a datasheet for the cerema AWP dataset, 2018, arXiv preprint arXiv:1806.04016.

[21] E. Choi, H. He, M. Iyyer, M. Yatskar, W.-t. Yih, Y. Choi, P. Liang, L. Zettlemoyer, QuAC: Question answering in context, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2174–2184, URL https://aclanthology.org/D18-1241.

[22] M.R. Costa-jussà, R. Creus, O. Domingo, A. Domínguez, M. Escobar, C. López, M. Garcia, M. Geleta, MT-adapted datasheets for datasets: Template and repository, 2020, arXiv e-prints arXiv–2005.

[23] N. Rostamzadeh, D. Mincu, S. Roy, A. Smart, L. Wilcox, M. Pushkarna, J. Schrouff, R. Amironesei, N. Moorosi, K. Heller, Healthsheet: Development of a transparency artifact for health datasets, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 1943–1961.

[24] C. Garbin, P. Rajpurkar, J. Irvin, M.P. Lungren, O. Marques, Structured dataset documentation: a datasheet for CheXpert, 2021, arXiv preprint arXiv:2105.03020.

[25] A.I. Anik, A. Bunt, Data-centric explanations: explaining training data of machine learning systems to promote transparency, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–13.

[26] L.A. Castelijns, Y. Maas, J. Vanschoren, The abc of data: A classifying framework for data readiness, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2019, pp. 3–16.

[27] R.S. Geiger, K. Yu, Y. Yang, M. Dai, J. Qiu, R. Tang, J. Huang, Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, in: FAT* '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 325–336.

[28] J.W. Vaughan, Making better use of the crowd: How crowdsourcing can advance machine learning research, J. Mach. Learn. Res. 18 (1) (2017) 7026–7071.

[29] W. Zhang, O. Ohrimenko, R. Cummings, Attribute privacy: Framework and mechanisms, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 757–766.

[30] D. Contractor, D. McDuff, J.K. Haines, J. Lee, C. Hines, B. Hecht, N. Vincent, H. Li, Behavioral use licensing for responsible AI, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 778–788.

[31] Y.T. Cao, H. Daumé III, Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle, Comput. Linguist. 47 (3) (2021) 615–661.

[32] J. Cabot, M. Gogolla, Object constraint language (OCL): a definitive guide, in: International School on Formal Methods for the Design of Computer, Communication and Software Systems, Springer, 2012, pp. 58–90.

[33] D. Torre, Y. Labiche, M. Genero, UML consistency rules: a systematic mapping study, in: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, 2014, pp. 1–10.

[34] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, A. Halpern, B. Helba, H. Kittler, K. Kose, S. Langer, K. Lioprys, J. Malvehy, S. Musthaq, J. Nanda, O. Reiter, G. Shih, A. Stratigos, P. Tschandl, J. Weber, H.P. Soyer, A patient-centric dataset of images and metadata for identifying melanomas using clinical context, Sci. Data 8 (1) (2021) 34.

[35] TypeFox, Langium - Home, 2022, https://langium.org/, last accessed April 2022.

[36] P.F. Patel-Schneider, Analyzing schema. org, in: International Semantic Web Conference, Springer, 2014, pp. 261–276.

[37] K.B. Rasmussen, G. Blank, The data documentation initiative: a preservation standard for research, Arch. Sci. 7 (1) (2007) 55–71.

[38] W3 Consortium, et al., Data catalog vocabulary (DCAT), 2014.

[39] B. Pang, L. Lee, A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts, in: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 2004, pp. 271–es.

[40] J. Rubin, D. Chisnell, Handbook of Usability Testing: How to Plan, Design and Conduct Effective Tests, John Wiley & Sons, 2008.

[41] P. Ralph, ACM SIGSOFT empirical standards released, ACM SIGSOFT Softw. Eng. Notes 46 (1) (2021) 19.

[42] R. Alroobaea, P.J. Mayhew, How many participants are really enough for usability studies? in: 2014 Science and Information Conference, IEEE, 2014, pp. 48–56.

[43] A.K. Heger, L.B. Marquis, M. Vorvoreanu, H. Wallach, J. Wortman Vaughan, Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata, Proceedings of the ACM on Human-Computer Interaction 6 (CSCW2) (2022) 1–29.

[44] C. Wohlin, M. Höst, K. Henningsson, Empirical research methods in software engineering, in: Empirical Methods and Studies in Software Engineering, Springer, 2003, pp. 7–23.

[45] S.L. Weibel, T. Koch, The Dublin core metadata initiative, D-Lib Mag. 6 (12) (2000) 1082–9873.

[46] K.S. Candan, H. Liu, R. Suvarna, Resource description framework: metadata and its applications, Acm Sigkdd Explor. Newslett. 3 (1) (2001) 6–19.

[47] W3 Consortium, W3C prov working group, 2022.

[48] D.H. Ton That, G. Fils, Z. Yuan, T. Malik, Sciunits: Reusable research objects, in: 2017 IEEE 13th International Conference on e-Science (e-Science), 2017, pp. 374–383, http://dx.doi.org/10.1109/eScience.2017.51.

[49] A.K. Sujeeth, H. Lee, K.J. Brown, H. Chafi, M. Wu, A.R. Atreya, K. Olukotun, T. Rompf, M. Odersky, Optiml: an implicitly parallel domain-specific language for machine learning, in: Proceedings of the 28th International Conference on International Conference on Machine Learning, 2011, pp. 609–616.

[50] M. Weimer, T. Condie, R. Ramakrishnan, et al., Machine learning in ScalOps, a higher order cloud computing language, in: NIPS 2011 Workshop on Parallel and Large-Scale Machine Learning (BigLearn), Vol. 9, Citeseer, 2011, pp. 389–396.

[51] A. Podobas, M. Svedin, S.W. Chien, I.B. Peng, N.B. Ravichandran, P. Herman, A. Lansner, S. Markidis, Streambrain: an hpc framework for brain-like neural networks on cpus, gpus and fpgas, in: Proceedings of the 11th International Symposium on Highly Efficient Accelerators and Reconfigurable Technologies, 2021, pp. 1–6.

[52] T. Zhao, X. Huang, Design and implementation of DeepDSL: A DSL for deep learning, Comput. Lang., Syst. Struct. 54 (2018) 39–70.

[53] N. Dethlefs, K. Hawick, Define: A fluent interface dsl for deep learning applications, in: Proceedings of the 2nd International Workshop on Real World Domain Specific Languages, 2017, pp. 1–10.

[54] V. García-Díaz, J. Pascual Espada, B.C. Pelayo G-Bustelo, J.M. Cueva Lovelle, Towards a standard-based domain-specific platform to solve machine learning-based problems, Int. J. Interact. Multimedia Artif. Intell. 3 (5) (2015).

[55] F. Melchor, R. Rodriguez-Echeverria, J.M. Conejero, Á.E. Prieto, J.D. Gutiérrez, A model-driven approach for systematic reproducibility and replicability of data science projects, in: International Conference on Advanced Information Systems Engineering, Springer, 2022, pp. 147–163.

[56] A. Moin, S. Rössler, M. Sayih, S. Günnemann, From things' modeling language (ThingML) to things' machine learning (ThingML2), in: Proceedings of the 23rd ACM/IEEE International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings, 2020, pp. 1–2.

[57] C. Di Sipio, J. Di Rocco, D. Di Ruscio, D.P.T. Nguyen, A low-code tool supporting the development of recommender systems, in: Fifteenth ACM Conference on Recommender Systems, 2021, pp. 741–744.

[58] A. Agrawal, A. Modi, A. Passos, A. Lavoie, A. Agarwal, A. Shankar, I. Ganichev, J. Levenberg, M. Hong, R. Monga, et al., TensorFlow eager: A multi-stage, python-embedded DSL for machine learning, in: Proceedings of Machine Learning and Systems, Vol. 1, 2019, pp. 178–189.

[59] T. Hartmann, A. Moawad, C. Schockaert, F. Fouquet, Y. Le Traon, Meta-modelling meta-learning, in: 2019 ACM/IEEE 22nd International Conference on Model Driven Engineering Languages and Systems, MODELS, IEEE, 2019, pp. 300–305.

[60] J. Zucker, M. d'Leeuwen, Arbiter: A domain-specific language for ethical machine learning, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 421–425.

[61] B. Ries, N. Guelfi, B. Jahic, An mde method for improving deep learning dataset requirements engineering using alloy and uml, in: Proceedings of the 9th International Conference on Model-Driven Engineering and Software Development, SCITEPRESS, 2021, pp. 41–52.

[62] E. Celms, J. Barzdins, A. Kalnins, P. Barzdins, A. Sprogis, M. Grasmanis, S. Rikacovs, DSL approach to deep learning lifecycle data management, Baltic J. Mod. Comput. 8 (4) (2020) 597–617.

[63] A. de la Vega, D. García-Saiz, M. Zorrilla, P. Sánchez, Lavoisier: A DSL for increasing the level of abstraction of data selection and formatting in data mining, J. Comput. Lang. 60 (2020) 100987.

[64] P. Muñoz, P. Karkhanis, M. van den Brand, A. Vallecillo, Modeling objects with uncertain behaviors, in: Proc. of ECMFA'21. Journal of Object Technology, (3) 2020, pp. 1–24.

[65] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I.D. Raji, T. Gebru, Model cards for model reporting, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 220–229.

[66] J. Tagliabue, V. Tuulos, C. Greco, V. Dave, DAG card is the new model card, 2021, arXiv preprint arXiv:2110.13601.

[67] K. Sokol, P. Flach, Explainability fact sheets: a framework for systematic assessment of explainable approaches, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 56–67.

Joan Giner is a Ph.D. candidate at the Open University of Catalonia, Barcelona, Spain. He graduated in Telecommunication Engineering from the UAB, and he worked as a software architect in several private companies. His research focuses on improving the quality of the data used to train machine learning models. Contact him at jginermi@uoc.edu or visit https://som-research.uoc.edu/joan-giner-miguelez/.

Abel Gómez is a senior researcher of the Internet Interdisciplinary Institute, a research center of the Universitat Oberta de Catalunya, Spain. Previously, he has hold different positions at the Universidad de Zaragoza, the Ecole des Mines de Nantes & Inria, and the Universitat Politecnica de Valencia where he obtained his Ph.D. degree; His research interests fall in the broad field of Model-Driven Engineering (MDE), and his research lines have evolved in two complementary directions: the development of core technologies to support MDE activities; and the application of MDE techniques to solve Software Engineering problems. More information is available at https://abel.gomez.llana.me.

Jordi Cabot is the head of the Software Engineering RDI Unit at the Luxembourg Institute of Science and Technology. His research interests include software and systems modeling, pragmatic formal model verification techniques, analysis of software communities and the role AI can play in software development (and vice versa). For more information, visit https://jordicabot.com/.