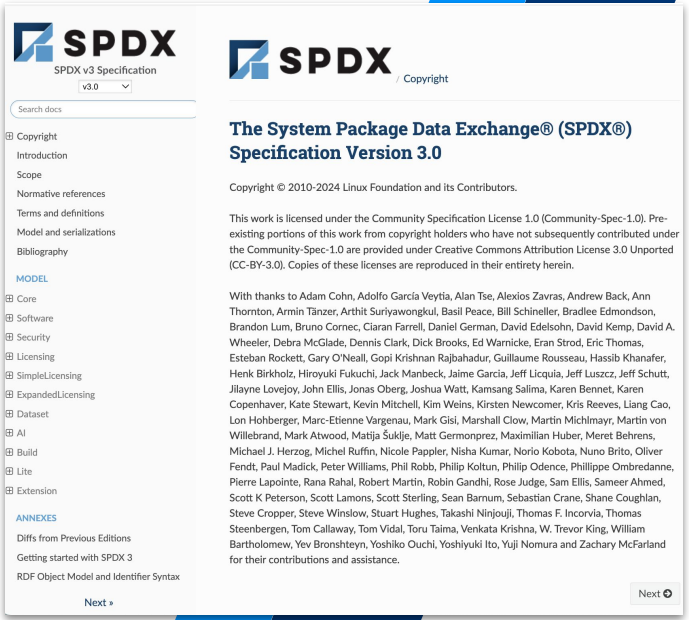
SPDX 3.0: Evolving to Represent System Bill of Materials Efficiently



Kate Stewart (SPDX Technical Team Lead)
Karen Bennet (SPDX AI Profile Lead)

https://spdx.dev/engage/participate/technical-team/

Timeline of SPDX Evolution - Use Case by Use Case

I AM THE Cavalry

Legislation: proposed software transparency, updatability & bill of material as reqts in safety critical sectors (automotive & healthcare)

NTIA: Software Transparency begins

3T-SBOM: OMG/CISQ begins w/ CONOPs for Tool-to-Tool SBOM

SPDX 3.0 initial draft
3T-SBOM initial draft

Executive Order 14028

Transition of SBOM work to DHS

EU Cyber Resilience Act

2010 SPDX begins

Standardized Single Package Information: Machine and human readable formats

2011 SPDX 1.0

Compliance Use Cases: Additional Project and License Information

2013 SPDX 1.2

Package Relations: 30+ additional use cases supported for complex packaging relationships and distribution scenarios

2015 SPDX 2.0

Security use cases: External references for vulnerabilities and product identification

2018

2019

2020 SPDX 2.2

SPDX & 3T-SBOM efforts merge: SPDX revises charter as an SDO

2021

Free ISO Standard: ISO/IEC 5962 SPDX available

2022 SPDX 2.3

Format Interoperability

CDx SPDX

2023 SPDX 3rc1

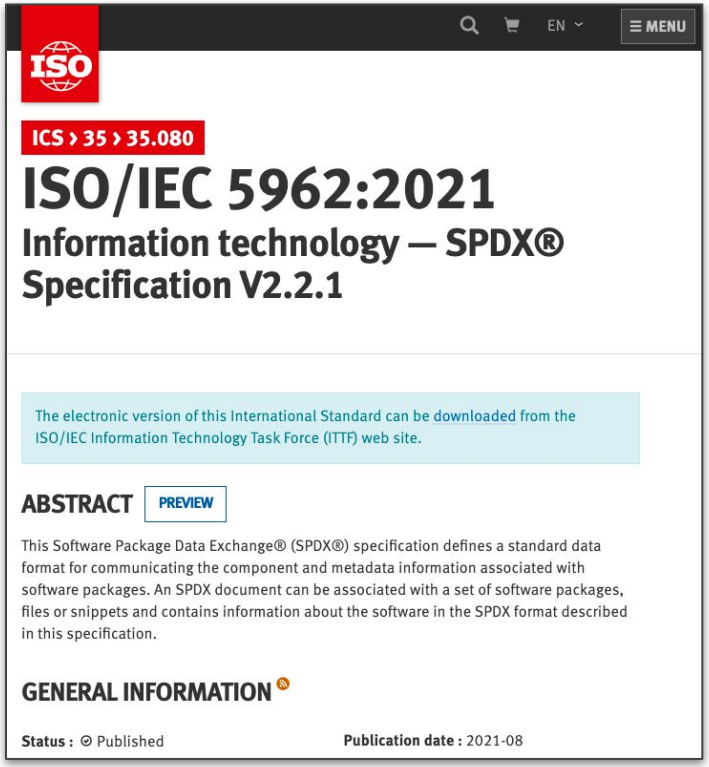
Profiles: New areas of use cases:
- Build
- Data
- Security
- AI
- Lite

2024+ SPDX 3.0

Profiles: New areas of use cases:
- Services
- Hardware
- Safety
- Operations

ISO/IEC 5962:2021

- Able to represent SBOMs from binary images and track back to the source files and snippets.
- Specification is freely available from ISO site.
- Future updates are live tracked at: https://spdx.github.io/spdx-spec and work on satisfying safety requirements is being included
- More information at spdx.dev



SPDX

NTIA Software Bill Of Materials (SBOM) Guidance - Minimum Elements

Data Field	Description
Supplier Name	The name of an entity that creates, defines, and identifies components.
Component Name	Designation assigned to a unit of software defined by the original supplier.
Version of the Component	Identifier used by the supplier to specify a change in software from a previously identified version.
Other Unique Identifiers	Other identifiers that are used to identify a component, or serve as a look-up key for relevant databases.
Dependency Relationship	Characterizing the relationship that an upstream component X is included in software Y.
Author of SBOM Data	The name of the entity that creates the SBOM data for this component.
Timestamp	Record of the date and time of the SBOM data assembly.

Source: https://www.ntia.gov/files/ntia/publications/sbom_minimum_elements_report.pdf

SPDX 2.2 + (ISO/IEC 5962:2021) **supports all required minimum elements** (as well the optional that are mentioned in report) and **many more use cases**

Checker available at: https://github.com/spdx/ntia-conformance-checker

SPDX

SPDX Governance

Working groups	Technical Team	Steering Committee

- Open
  - We are open to contributions from anyone interested - all team and working group meetings are open
  - Team leads nominated by any participant and steering committee composed of team leads
- Transparent
  - Most of our work is maintained in the SPDX Github organization including minutes
- Inclusive
  - Leads and steering committee consists of individuals from diverse backgrounds and industries
  - Over 40 organizations contributed directly to the SPDX spec representing a majority of industry segments (Technology, Critical Infrastructure, Financial, Gov't) and geographies

Additional Standards Support

- Support for NTIA Minimum SBOM and tracking OpenSSF and CISA working groups (e.g. support for VEX and vulnerability data)
- OMG - release 3.0 has been approved by the OMG Architecture Board
- IANA assigned Mime types for SPDX serialization formats (JSON and Tag/Value)
- Standard schema support for easier tooling and validation
  - Supported RDF since release 1.0
    - Provides support for linked data and semantic web reasoning
  - JSON Schema available since 2.1
  - OWL Web Ontology Language since 1.0
  - 3.0 will add support for the Shapes Constraint Language (SHACL)

SPDX

Who is Using SPDX today?

- Wide range of users - from sophisticated security software analysts to non-technical procurement and compliance executives
  - Requires a wide range of “serialization” formats ranging from RDF/XML, to JSON, to YAML, to Spreadsheets
- Software and Software as a Service providers - from the very large to to the very small including many software tooling vendors
  - Many providing or utilizing open source libraries which support overall adoption
- Open source package maintainers across a wide variety of ecosystems (e.g. Python Poetry, Maven, Gradle, NPM)

SPDX

SPDX 3.0

Why SPDX 3.0?

- **Vulnerabilities occur in Systems** of which Software is only a part.
- Interest in SPDX for additional scenarios and use cases
  - Supporting **security** and safety critical application compliance requirements
  - **AI/ML** and **Datasets** increasing need **for system transparency**
  - Software and Dataset **build provenance**
- Simplify
  - Profiles
  - Remove confusing names
  - Reorganize and enable general SBOM use cases with minimum overhead
- Flexibility
  - Designed for **online access**
  - Support **optional inclusion** properties for specific profiles
  - Enhanced **relationship** structure to enable metadata knowledge graph

SPDX

Profiles Overview

SPDX SECURITY	Security information - vulnerability details related to software
SPDX BUILD	Build related information - provenance and reproducible builds
SPDX AI	Information about AI models - ethical, security, and model data
SPDX DATA	Information about datasets - AI and other data use cases
SPDX LITE	Minimal subset to support industry supply chain workflows
SPDX LICENSING	Information about copyrights and licenses - supports compliance
SPDX SOFTWARE	Information specific to software
SPDX CORE	Information used across all profiles

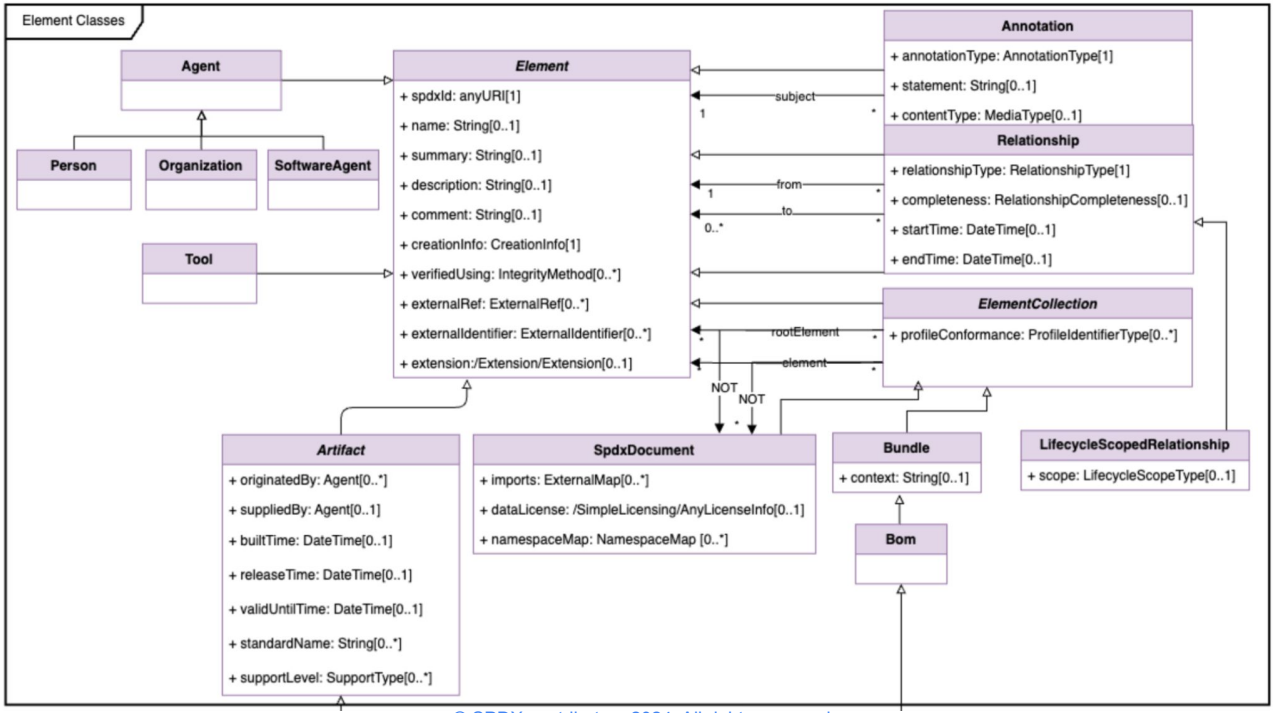
SPDX 3.0 Specification Infrastructure

- Specification expressed in markdown describing
  - Classes, Properties, Enumerations
  - Metadata (type and cardinality) and description for each element
  - Will be able to automatically generate schema from this version (for JSON, YAML, RDF, XML, tag-value, etc) and reduce errors
- Profiles can add their own Classes and Properties and may also restrict other profiles (e.g. values, cardinalities, etc)
- See https://github.com/spdx/spdx-3-model for model diagrams
- See https://spdx.github.io/spdx-spec/v3.0/ to review the specification

SPDX

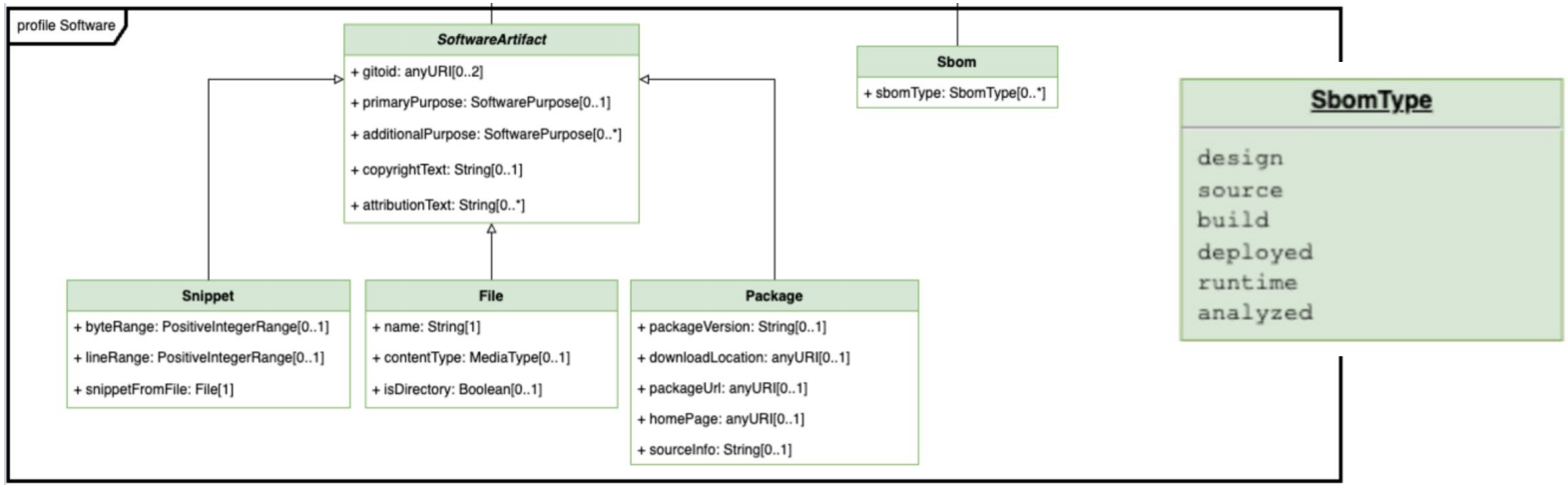
Core Profile Overview

- Defines foundational concepts which are the basis for **all SPDX 3.0 profiles**



Software Profile Overview

- Defines concepts related to software artifacts, augments core profile. AI & Data Profiles augment on top of these classes.
- Introduces and supports CISA SBOM Types*
  - Design/Source/Build/Deployed/Runtime/Analyzed



* https://www.cisa.gov/sites/default/files/2023-04/sbom-types-document-508c.pdf

Initial AI BOM Transparency Survey: What's Required?

Datasheets

arXiv:1803.09010v3 [cs.DB] 9 Jul 2018

Datasheets for Datasets

Timnit Gebru [1] Jamie Morgenstern [2] Briana Vecchione [3] Jennifer Wortman Vaughan [1] Hanna Wallach [1]
Hal Daumé III [1 4] Kate Crawford [1 5]

Abstract

The machine learning community has no standardized way to document how and why a dataset was created, what information it contains, what tasks it should and should not be used for, and whether it might raise any ethical or legal concerns. To address this gap, we propose the concept of datasheets for datasets. In the electronics industry, it is standard to accompany every component with a datasheet providing standard operating characteristics, test results, recommended usage, and other information. Similarly, we recommend that every dataset be accompanied with a datasheet documenting its creation, composition, intended uses, maintenance, and other properties. Datasheets for datasets will facilitate better communication between dataset creators and users, and encourage the machine learning community to prioritize transparency and accountability.

1. Introduction

Machine learning is no longer a purely academic discipline. Domains such as criminal justice (Garvie et al., 2016; Systems, 2017; Andrews et al., 2006), hiring and employment (Mann & O'Neil, 2016), critical infrastructure (O'Connor, 2017; Chui, 2017), and finance (Lin, 2012) all increasingly depend on machine learning methods.

By definition, machine learning models are trained using data; the choice of data fundamentally influences a model's behavior. However, there is no standardized way to document how and why a dataset was created, what information it contains, what tasks it should and shouldn't be used for, and whether it might raise any ethical or legal concerns. This lack of documentation is especially problematic when datasets are used to train models for high-stakes applications.

We therefore propose the concept of datasheets for datasets. In the electronics industry, every component is accompanied by a datasheet describing standard operating characteristics, test results, and recommended usage. By analogy, we recommend that every dataset be accompanied with a datasheet documenting its motivation, creation, composition, intended uses, distribution, maintenance, and other information. We anticipate that such datasheets will increase transparency and accountability in the machine learning community.

Section 2 provides context for our proposal. Section 3 discusses the evolution of safety standards in other industries, and outlines the concept of datasheets in electronics. We give examples of questions that should be answered in datasheets for datasets in Section 4, and discuss challenges and future work in Section 5. The appendix includes a more complete proposal along with prototype datasheets for two well-known datasets: Labeled Faces in the Wild (Huang et al., 2007) and Pang and Lee's polarity dataset (2004).

2. Context

A foundational challenge in the use of machine learning is the risk of deploying systems in unsuitable environments. A model's behavior on some benchmark may say very little about its performance in the wild. Of particular concern are recent examples showing that machine learning systems can amplify existing societal biases. For example, Buolamwini & Gebru (2018) showed that commercial gender classification APIs have near perfect performance for lighter-skinned males, while error rates for darker-skinned females can be as high as 33%.[1] Bolukbasi et al. (2016) showed that word embeddings trained on news articles exhibit gender biases, finishing the analogy "man is to computer programmer as woman is to X" with "homemaker," a stereotypical role for women. Caliskan et al. (2017) showed these embeddings also contain racial biases: traditional European-American names are closer to positive words like "joy," while African-American names are closer to words like "agony."

These biases can have dire consequences that might not be easily discovered. Much like a faulty resistor or a capacitor in a circuit, the effects of a biased machine learning

[1]Microsoft Research, New York, NY [2]Georgia Institute of Technology, Atlanta, GA [3]Cornell University, Ithaca, NY [4]University of Maryland, College Park, MD [5]AI Now Institute, New York, NY. Correspondence to: Timnit Gebru <tgebru@gmail.com>.

Proceedings of the 5[th] Workshop on Fairness, Accountability, and Transparency in Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

[1]The evaluated APIs also provided the labels of female and male, failing to address the complexities of gender beyond binary.

Source:https://arxiv.org/pdf/1810.03993.pdf

Model Cards

arXiv:1810.03993v2 [cs.LG] 14 Jan 2019

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca

ABSTRACT

Trained machine learning models are increasingly used to perform high-impact tasks in areas such as law enforcement, medicine, education, and employment. In order to clarify the intended use cases of machine learning models and minimize their usage in contexts for which they are not well suited, we recommend that released models be accompanied by documentation detailing their performance characteristics. In this paper, we propose a framework that we call model cards, to encourage such transparent model reporting. Model cards are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups (e.g., race, geographic location, sex, Fitzpatrick skin type [15]) and intersectional groups (e.g., age and race, or sex and Fitzpatrick skin type) that are relevant to the intended application domains. Model cards also disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information. While we focus primarily on human-centered machine learning models in the application fields of computer vision and natural language processing, this framework can be used to document any trained machine learning model. To solidify the concept, we provide cards for two supervised models: One trained to detect smiling faces in images, and one trained to detect toxic comments in text. We propose model cards as a step towards the responsible democratization of machine learning and related artificial intelligence technology, increasing transparency into how well artificial intelligence technology works. We hope this work encourages those releasing trained machine learning models to accompany model releases with similar detailed evaluation numbers and other relevant documentation.

CCS CONCEPTS

• **General and reference** → **Evaluation;** • **Social and professional topics** → *User characteristics*; • **Software and its engineering** → **Use cases**; *Documentation*; *Software evolution*; • **Human-centered computing** → *Walkthrough evaluations*;

FAT '19, January 29–31, 2019, Atlanta, GA, USA*
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6125-5/19/01.
https://doi.org/10.1145/3287560.3287596

KEYWORDS

datasheets, model cards, documentation, disaggregated evaluation, fairness evaluation, ML model evaluation, ethical considerations

ACM Reference Format:
Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru. 2019. Model Cards for Model Reporting. In *FAT* '19: Conference on Fairness, Accountability, and Transparency, January 29–31, 2019, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3287560.3287596

1 INTRODUCTION

Currently, there are no standardized documentation procedures to communicate the performance characteristics of trained machine learning (ML) and artificial intelligence (AI) models. This lack of documentation is especially problematic when models are used in applications that have serious impacts on people's lives, such as in health care [14, 42, 44], employment [1, 13, 29], education [23, 45] and law enforcement [2, 7, 20, 34].

Researchers have discovered systematic biases in commercial machine learning models used for face detection and tracking [4, 9, 49], attribute detection [5], criminal justice [10], toxic comment detection [11], and other applications. However, these systematic errors were only exposed after models were put into use, and negatively affected users reported their experiences. For example, after MIT Media Lab graduate student Joy Buolamwini found that commercial face recognition systems failed to detect her face [4], she collaborated with other researchers to demonstrate the disproportionate errors of computer vision systems on historically marginalized groups in the United States, such as darker-skinned women [5, 41]. In spite of the potential negative effects of such reported biases, documentation accompanying trained machine learning models (if supplied) provides very little information regarding model performance characteristics, intended use cases, potential pitfalls, or other information to help users evaluate the suitability of these systems to their context. This highlights the need to have detailed documentation accompanying trained machine learning models, including metrics that capture bias, fairness and inclusion considerations.

As a step towards this goal, we propose that released machine learning models be accompanied by short (one to two page) records we call model cards. Model cards (for model reporting) are complements to "Datasheets for Datasets" [21] and similar recently proposed documentation paradigms [3, 28] that report details of the datasets used to train and test machine learning models. Model cards are also similar to the `nutrition` statement proposal in medicine [25]. We provide two example model cards in Section 5. A smiling detection model trained on the CelebA dataset [36] (Figure 2), and a public toxicity detection model [32] (Figure 3). Where Datasheets highlight characteristics of the data feeding into the model, we

Source:https://www.microsoft.com/en-us/research/uploads/prod/2019/01/1803.09010.pdf

FactSheets

arXiv:1808.07261v2 [cs.CY] 7 Feb 2019

FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity

M. Arnold,[1] R. K. E. Bellamy,[1] M. Hind,[1] S. Houde,[1] S. Mehta,[2] A. Mojsilović,[1] R. Nair,[1] K. Natesan Ramamurthy,[1] D. Reimer,[1] A. Olteanu,* D. Piorkowski,[1] J. Tsay,[1] and K. R. Varshney[1]
IBM Research
[1]Yorktown Heights, New York, [2]Bengaluru, Karnataka

Abstract

Accuracy is an important concern for suppliers of artificial intelligence (AI) services, but considerations beyond accuracy, such as safety (which includes fairness and explainability), security, and provenance, are also critical elements to engender consumers' trust in a service. Many industries use transparent, standardized, but often not legally required documents called supplier's declarations of conformity (SDoCs) to describe the lineage of a product along with the safety and performance testing it has undergone. SDoCs may be considered multi-dimensional fact sheets that capture and quantify various aspects of the product and its development to make it worthy of consumers' trust. Inspired by this practice, we propose FactSheets to help increase trust in AI services. We envision such documents to contain purpose, performance, safety, security, and provenance information to be completed by AI service providers for examination by consumers. We suggest a comprehensive set of declaration items tailored to AI and provide examples for two fictitious AI services in the appendix of the paper.

1 Introduction

Artificial intelligence (AI) services, such as those containing predictive models trained through machine learning, are increasingly key pieces of products and decision-making workflows. A service is a function or application accessed by a customer via a cloud infrastructure, typically by means of an application programming interface (API). For example, an AI service could take an audio waveform as input and return a transcript of what was spoken as output, with all complexity hidden from the user, all computation done in the cloud, and all models used to produce the output pre-trained by the supplier of the service. A second more complex example would provide an audio waveform translated into a different language as output. The second example illustrates that a service can be made up of many different models (speech recognition, language translation, possibly sentiment or tone analysis, and speech synthesis) and is thus a distinct concept from a single pre-trained machine learning model or library.

In many different application domains today, AI services are achieving impressive accuracy. In certain areas, high accuracy alone may be sufficient, but deployments of AI in high-stakes decisions, such as credit applications, judicial decisions, and medical recommendations, require greater trust in AI services. Although there is no scholarly consensus on the specific traits that imbue trustworthiness in people or algorithms [1, 2], fairness, explainability, general safety, security, and transparency are some of the issues that have raised public concern about trusting AI and threatened the further adoption of AI beyond low-stakes uses [3, 4]. Despite active research and development to address these issues, there is no mechanism yet for the creator of an AI service to communicate how they are addressed in a deployed version. This is a major impediment to broad AI adoption.

Toward transparency for developing trust, we propose a *FactSheet* for AI Services. A FactSheet will contain sections on all relevant attributes of an AI service, such as intended use, performance, safety, and security. Performance will include appropriate accuracy or risk measures along with timing information. Safety, discussed in [5, 3] as the minimiza-

*A. Olteanu's work was done while at IBM Research. Author is currently affiliated with Microsoft Research.

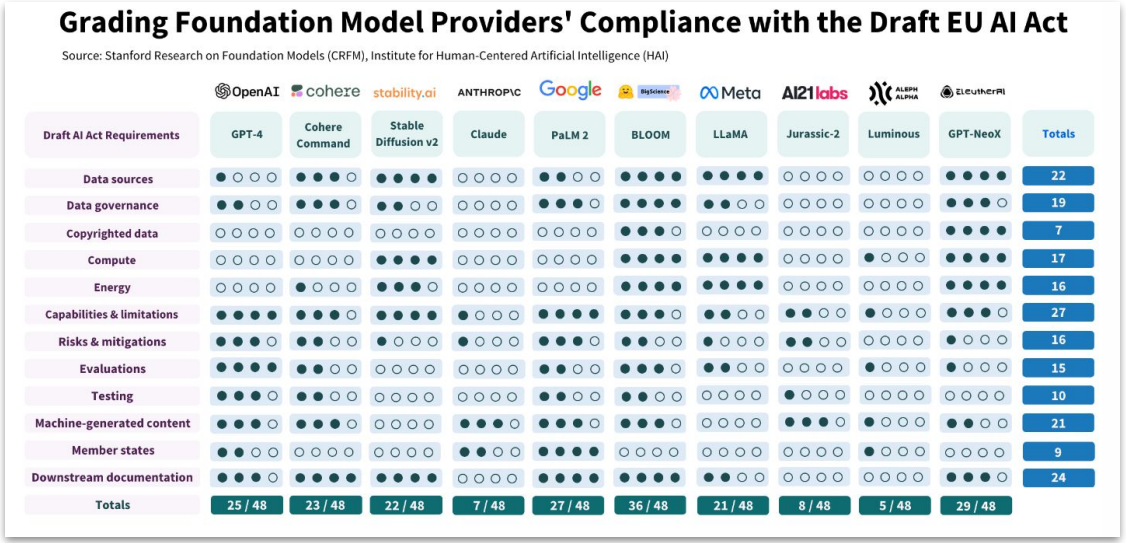
Source: https://arxiv.org/pdf/1808.07261.pdf

Build on other Standards & Regulatory Efforts

SPDX Fields	IEEE (AI)	ISO (AI)	EU Act
autonomyType	X		
buildTime	X		
createdBy	X	X	X
comment	X	X	X
domain	X		X
downloadLocation	X	X	
energyConsumption	X		
description	X	X	X
hyperparameter	X	X	
informationAboutTraining	X	X	X
informationAboutApplication	X	X	X
limitation	X	X	X
modelDataPreprocessing	X	X	X
modelExplainability	X		X
metric.	X	X	X
metricDecisionThreshold	X	X	
releaseTime	X		
safetyRiskAssessment	X	X	X
sensitivePersonalInformation	X	X	
standardCompliance	X	X	X

SPDX Fields	IEEE (AI)	ISO (AI)	EU Act
sourceInfo	X	X	X
typeOfModel	X	X	X
releaseTime	X		
validUntilTime	X		
confidentialityLevel	X		X
dataCollectionProcess	X		X
dataPreprocessing	X	X	X
anonymizationMethodUsed	X		X
datasetAvailability	X	X	X
datasetNoise	X	X	X
datasetSize	X	X	X
datasetType	X	X	X
datasetUpdateMechanism			
intendedUse	X	X	X
knownBias	X	X	X
sensitivePersonalInformation	X	X	X
sensor	X		
copyrighted data			X
Version	X		
primaryPurpose	X	X	X

Automating Issue Identification Requires Machine Readable AI BOM



Grading Foundation Model Providers' Compliance with the Draft EU AI Act

Source: Stanford Research on Foundation Models (CRFM), Institute for Human-Centered Artificial Intelligence (HAI)

Draft AI Act Requirements	GPT-4	Cohere Command	Stable Diffusion v2	Claude	PaLM 2	BLOOM	LLaMA	Jurassic-2	Luminous	GPT-NeoX	Totals
Data sources	● ○ ○ ○	● ● ● ○	● ● ● ●	○ ○ ○ ○	● ● ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	22
Data governance	● ● ○ ○	● ● ● ○	● ● ○ ○	○ ○ ○ ○	● ● ● ○	● ● ● ●	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	19
Copyrighted data	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	7
Compute	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	● ○ ○ ○	● ● ● ●	17
Energy	○ ○ ○ ○	● ○ ○ ○	● ● ● ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	16
Capabilities & limitations	● ● ● ●	● ● ● ○	● ● ● ●	● ○ ○ ○	● ● ● ●	● ● ● ○	● ● ○ ○	● ● ○ ○	● ○ ○ ○	● ● ● ○	27
Risks & mitigations	● ● ● ○	● ● ○ ○	● ○ ○ ○	● ○ ○ ○	● ● ● ○	● ● ○ ○	● ○ ○ ○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	16
Evaluations	● ● ● ●	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ○ ○	● ● ● ○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	● ○ ○ ○	15
Testing	● ● ● ○	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ○ ○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	10
Machine-generated content	● ● ● ○	● ● ● ○	○ ○ ○ ○	● ● ● ○	● ● ● ○	● ● ● ○	○ ○ ○ ○	● ● ● ○	● ○ ○ ○	● ● ○ ○	21
Member states	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ○ ○	● ● ● ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ○ ○ ○	○ ○ ○ ○	9
Downstream documentation	● ● ● ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	● ● ● ●	● ● ● ●	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	24
Totals	25 / 48	23 / 48	22 / 48	7 / 48	27 / 48	36 / 48	21 / 48	8 / 48	5 / 48	29 / 48	

- Many issues exist even when comparing to one standard (EU AI Act) across most of the existing state-of-the-art AI models
- In order to be able to establish compliance to several regulations around the world and identify potential risks associated with the use and adaption of AI software an automated standard, that captures all the required data and is machine readable is pivotal.

AI Profile

SPDX AI Profile

A profile that adds on top of the core-software profile to describe the AI specific elements that will enable transparency and traceability of both components and process that enables the creation of an AI software. It is important to note that, special consideration is given to capture process (in addition to just capturing the components) as process introduces a lot of risks and uncertainties that make up the non-deterministic system that AI is.

SPDX AI Profile

Various details that we want to capture as a part of the profile

Properties

External property restrictions

Properties that are borrowed from other SPDX profiles

Used to express connections between other profiles and elements

Relationships

Required and optional fields

We propose a minimal and expansive list of properties to ensure easy adaptability

Why not just use Model cards?

arXiv:1810.03993v2 [cs.LG] 14 Jan 2019

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca

ABSTRACT

Trained machine learning models are increasingly used to perform high-impact tasks in areas such as law enforcement, medicine, education, and employment. In order to clarify the intended use cases of machine learning models and minimize their usage in contexts for which they are not well suited, we recommend that released models be accompanied by documentation detailing their performance characteristics. In this paper, we propose a framework that we call model cards, to encourage such transparent model reporting. Model cards are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups (e.g., race, geographic location, sex, Fitzpatrick skin type [15]) and intersectional groups (e.g., age and race, or sex and Fitzpatrick skin type) that are relevant to the intended application domains. Model cards also disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information. While we focus primarily on human-centered machine learning models in the application fields of computer vision and natural language processing, this framework can be used to document any trained machine learning model. To solidify the concept, we provide cards for two supervised models: One trained to detect smiling faces in images, and one trained to detect toxic comments in text. We propose model cards as a step towards the responsible democratization of machine learning and related artificial intelligence technology, increasing transparency into how well artificial intelligence technology works. We hope this work encourages those releasing trained machine learning models to accompany model releases with similar detailed evaluation numbers and other relevant documentation.

CCS CONCEPTS

• **General and reference** → **Evaluation**; • **Social and professional topics** → *User characteristics*; • **Software and its engineering** → **Use cases**; *Documentation*; *Software evolution*; • **Human-centered computing** → *Walkthrough evaluations*;

FAT '19, January 29–31, 2019, Atlanta, GA, USA*

ACM ISBN 978-1-4503-6125-5/19/01.
https://doi.org/10.1145/3287560.3287596

KEYWORDS

datasheets, model cards, documentation, disaggregated evaluation, fairness evaluation, ML model evaluation, ethical considerations

ACM Reference Format:
Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru. 2019. Model Cards for Model Reporting. In *FAT* '19: Conference on Fairness, Accountability, and Transparency, January 29–31, 2019, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3287560.3287596

1 INTRODUCTION

Currently, there are no standardized documentation procedures to communicate the performance characteristics of trained machine learning (ML) and artificial intelligence (AI) models. This lack of documentation is especially problematic when models are used in applications that have serious impacts on people's lives, such as in health care [14, 42, 44], employment [1, 13, 29], education [23, 45] and law enforcement [2, 7, 20, 34].

Researchers have discovered systematic biases in commercial machine learning models used for face detection and tracking [4, 9, 49], attribute detection [5], criminal justice [10], toxic comment detection [11], and other applications. However, these systematic errors were only exposed after models were put into use, and negatively affected users reported their experiences. For example, after MIT Media Lab graduate student Joy Buolamwini found that commercial face recognition systems failed to detect her face [4], she collaborated with other researchers to demonstrate the disproportionate errors of computer vision systems on historically marginalized groups in the United States, such as darker-skinned women [5, 41]. In spite of the potential negative effects of such reported biases, documentation accompanying trained machine learning models (if supplied) provide very little information regarding model performance characteristics, intended use cases, potential pitfalls, or other information to help users evaluate the suitability of these systems to their context. This highlights the need to have detailed documentation accompanying trained machine learning models, including metrics that capture bias, fairness and inclusion considerations.

As a step towards this goal, we propose that released machine learning models be accompanied by short (one to two page) records we call model cards. Model cards (for model reporting) are complements to "Datasheets for Datasets" [21] and similar recently proposed documentation paradigms [3, 28] that report details of the datasets used to train and test machine learning models. Model cards are also similar to the TRIPOD statement proposal in medicine [25]. We provide two example model cards in Section 5: A smiling detection model trained on the CelebA dataset [36] (Figure 2), and a public toxicity detection model [32] (Figure 3). Where Datasheets highlight characteristics of the data feeding into the model, we

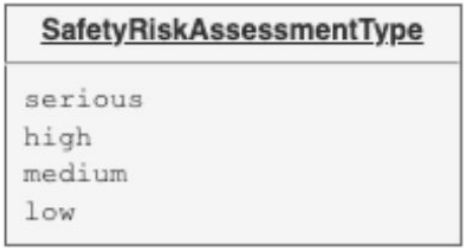
Limitations

- Limited Scope
- Oversimplification
- Interdependencies
- Data Flow
- Security and compliance
- Versioning and update
- Environmental impact
- Operational and maintenance details are lacking
- Ethical and Social implications

AI/ML Elements

Additional Artificial Intelligence and Machine Learning information information about Applications/Models that can be **optionally** included :

- **Characteristics**
  - Type of Model
  - Standard Compliance
  - Domain
  - Autonomy Type
- **Transparency**:
  - Limitation
  - Information About Training
  - Information About Application
  - Hyperparameter
  - Model Data Preprocessing
  - Model Explainability
  - Metric
  - Metric Decision Threshold
- Safety Risk Assessment
- Sensitive Personal Information
- Energy Consumption



* Using categorization according to the EU general risk assessment methodology which implements Article 20 of Regulation (EC) No 765/2008 and is intended to assist authorities when they assess general product safety compliance. It is important to note that this categorization **differs from the one proposed in the EU AI Act**'s **provisional agreement**.

Data Profile

SPDX Dataset Profile

A profile that adds on top of the core-software profile to describe the dataset that is used to train or test an AI software. These datasets could also be used for other purposes. Similar, to AI profile, we take special care to ensure that process of forming a dataset is captured. In addition, we also make it a point to capture the provenance and the lineage associated with the dataset. We introduce new fields to do so as provenance and lineage of a dataset has more facets to it compared to traditional software.

SPDX Dataset Profile

Various details that we want to capture as a part of the profile

Properties

External property restrictions

Properties that are borrowed from other SPDX profiles

Used to express connections between other profiles and elements

Relationships

Required and optional fields

We propose a minimal and expansive list of properties to ensure easy adaptability

Why not Datasheets?

Limitations

- Lack of detailed metadata
- Data provenance and lineage details are missing
- Versioning details
- Data dependencies
- Data processing pipeline details
- Data privacy and security
- Bias information

arXiv:1803.09010v8 [cs.DB] 1 Dec 2021

Datasheets for Datasets

TIMNIT GEBRU, Black in AI
JAMIE MORGENSTERN, University of Washington
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research

1 Introduction

Data plays a critical role in machine learning. Every machine learning model is trained and evaluated using data, quite often in the form of static datasets. The characteristics of these datasets fundamentally influence a model's behavior: a model is unlikely to perform well in the wild if its deployment context does not match its training or evaluation datasets, or if these datasets reflect unwanted societal biases. Mismatches like this can have especially severe consequences when machine learning models are used in high-stakes domains, such as criminal justice [1, 13, 24], hiring [19], critical infrastructure [11, 21], and finance [18]. Even in other domains, mismatches may lead to loss of revenue or public relations setbacks. Of particular concern are recent examples showing that machine learning models can reproduce or amplify unwanted societal biases reflected in training datasets [4, 5, 12]. For these and other reasons, the World Economic Forum suggests that all entities should document the provenance, creation, and use of machine learning datasets in order to avoid discriminatory outcomes [25].

Although data provenance has been studied extensively in the databases community [3, 8], it is rarely discussed in the machine learning community. Documenting the creation and use of datasets has received even less attention. Despite the importance of data to machine learning, there is currently no standardized process for documenting machine learning datasets.

To address this gap, we propose *datasheets for datasets*. In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet describing its operating characteristics, test results, recommended usage, and other information. By analogy, we propose that every

Authors' addresses: Timnit Gebru, Black in AI; Jamie Morgenstern, University of Washington; Briana Vecchione, Cornell University; Jennifer Wortman Vaughan, Microsoft Research; Hanna Wallach, Microsoft Research; Hal Daumé III, Microsoft Research; University of Maryland; Kate Crawford, Microsoft Research.

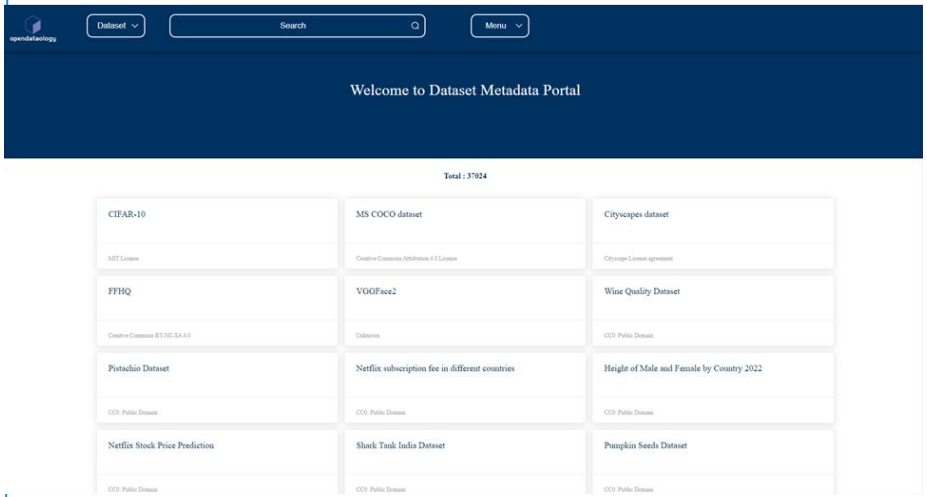
SPDX Dataset Profile Properties

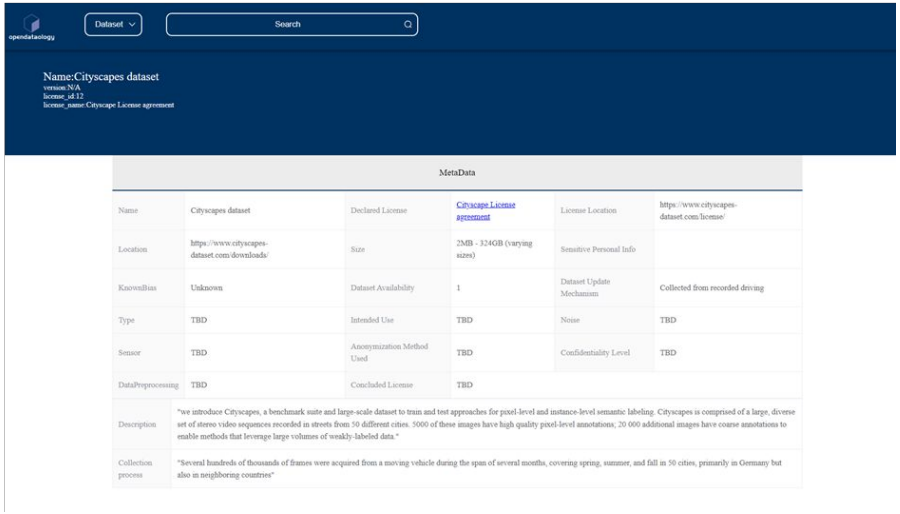
Datasets have “Supply Chains” as well. It is important to be able to describe **dataset provenance** in order to have confidence when using with AI/ML. **Additional** dataset information that can be **optionally** included :

- Characteristics:
  - Dataset Type
  - Dataset Size
  - Dataset Noise
  - Dataset Availability
  - Confidentiality Level
  - Sensitive Personal Information
- Provenance:
  - Dataset Collection Process
  - Dataset Update Mechanism
  - Intended Use
  - Known Bias
  - Data Preprocessing
  - Anonymization Method Used
  - Sensor (used for collection & calibration data)

Dataset Profile – OpenDataology Example

Over 37,000 datasets metadata are captured using the SPDX dataset profile in order to conduct license compliance analysis as a part of the OpenDataology project





Thank you to all SPDX Supporters!

Logos used with written permission

Questions?

https://spdx.dev/engage/participate/technical-team/