A Survey on Metadata for Machine Learning Models and Datasets: Standards, Practices, and Harmonization Challenges

Genet-Asefa Gesese[1,2], Zongxiong Chen[3], Oussama Zoubia[4], Fidan Limani[5], Kanishka Silva[6], Muhammad Asif Suryani[6], Benjamin Zapilko[6], Leyla Jael Castro[7], Ekaterina Kutafina[4], Dhwani Solanki[7], Heike Fliegl[1], Sonja Schimmler[3,8], Zeyd Boukhers[4] and Harald Sack[1,2]

[1]*FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany*
[2]*Karlsruhe Institute of Technology, Institute AIFB, Germany*
[3]*Fraunhofer FOKUS, Germany*
[4]*Institute for Biomedical Informatics, Medical Faculty, University of Cologne, Germany*
[5]*ZBW – Leibniz Information Centre for Economics, Germany*
[6]*GESIS – Leibniz Institute for the Social Sciences, Germany*
[7]*ZBMED – Information Centre for Life Sciences*
[8]*TU Berlin, Germany*

Abstract

The growing availability of machine learning (ML) models, datasets, and related artifacts across platforms, such as Hugging Face, GitHub, and Zenodo, has amplified the need for structured and standardized metadata. However, metadata practices remain highly heterogeneous, differing in schema design, vocabulary usage, and semantic expressiveness, posing significant challenges for tasks such as representation, extraction, alignment, and integration. This fragmentation impedes the development of infrastructures that depend on machine-actionable metadata to support discovery, provenance tracking, or cross-platform interoperability. While metadata is also foundational to enabling FAIR (Findable, Accessible, Interoperable, and Reusable) principles in ML, there is a lack of consolidated understanding of how existing standards support interoperability and alignment across platforms. In this survey, we review and compare a range of general-purpose and ML-specific metadata standards, evaluating their suitability for cross-platform alignment, discoverability, extensibility, and interoperability. We assess these standards based on defined criteria and analyze their potential to support unified, FAIR-compliant metadata infrastructures for ML, laying the groundwork for scalable and interoperable tooling in future ML ecosystems.

Keywords

Metadata, Machine Learning, Datasets, FAIR, Research Artifacts Harmonization

1. Introduction and Motivation

The rapid growth of machine learning (ML) research has led to an explosion in the availability of ML artifacts, such as models, datasets, and training code, which are now shared across a wide range of platforms, including GitHub[1], Hugging Face[2], Zenodo[3], or OpenML[4]. These platforms have become

5th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment, Nov 2025, Nara, Japan
genet-asefa.gesese@fiz-karlsruhe.de (G. Gesese); zongxiong.chen@fokus.fraunhofer.de (Z. Chen); oussama.zoubia@uk-koeln.de (O. Zoubia); f.limani@zbw.eu (F. Limani); kanishka.silva@gesis.org (K. Silva); asif.suryani@gesis.org (M. A. Suryani); benjamin.zapilko@gesis.org (B. Zapilko); ljgarcia@zbmed.de (L. J. Castro); ekaterina.kutafina@uni-koeln.de (E. Kutafina); solanki@zbmed.de (D. Solanki); Heike.Fliegl@fiz-Karlsruhe.de (H. Fliegl); sonja.schimmler@fokus.fraunhofer.de (S. Schimmler); zeyd.boukhers@fit.fraunhofer.de (Z. Boukhers); harald.sack@fiz-karlsruhe.de (H. Sack)
0000-0003-3807-7145 (G. Gesese); 0000-0003-2452-0572 (Z. Chen); 0000-0002-7930-7157 (O. Zoubia); 0000-0002-5835-2784 (F. Limani); 0000-0003-2958-9552 (K. Silva); 0000-0003-1669-5524 (M. A. Suryani); 0000-0001-9495-040X (B. Zapilko); 0000-0003-3986-0510 (L. J. Castro); 0000-0002-3430-5123 (E. Kutafina); 0000-0002-7541-115X (H. Fliegl); 0000-0002-8786-7250 (S. Schimmler); 0000-0001-7069-9804 (H. Sack)

[1]https://github.com/
[2]https://huggingface.co/
[3]https://zenodo.org/
[4]https://www.openml.org/

an essential infrastructure for disseminating pre-trained models, experimental results, datasets, and reproducible workflows. However, the scale and diversity of these artifacts have outpaced any consistent metadata practices, resulting in fragmented, incompatible, and semantically shallow descriptions across platforms [1, 2, 3].

Metadata provides structured descriptions of digital objects such as datasets, software, and models, supporting both human understanding and machine interoperability. It also carries critical supplementary information, including provenance, quality, licensing, versioning, and usage constraints that bring additional context to a resource. However, significant heterogeneity exists in how metadata is designed and applied across platforms. This includes differences in schema design, vocabulary usage, expressiveness, and machine readability. This lack of metadata standardization limits machine-actionability and affects automated workflows, making it difficult, for instance, to discover models, link them to related publications, or incorporate them into knowledge-driven systems [4]. Figure 1 illustrates this progression from fragmentation toward semantic integration and FAIR (Findable, Accessible, Interoperable, and Reusable)[5] infrastructures. These difficulties become more evident in workflows that rely on structured knowledge representations, such as Data Science (DS) and Artificial Intelligence (AI) pipelines or Knowledge Graph (KG)-powered discovery tools.

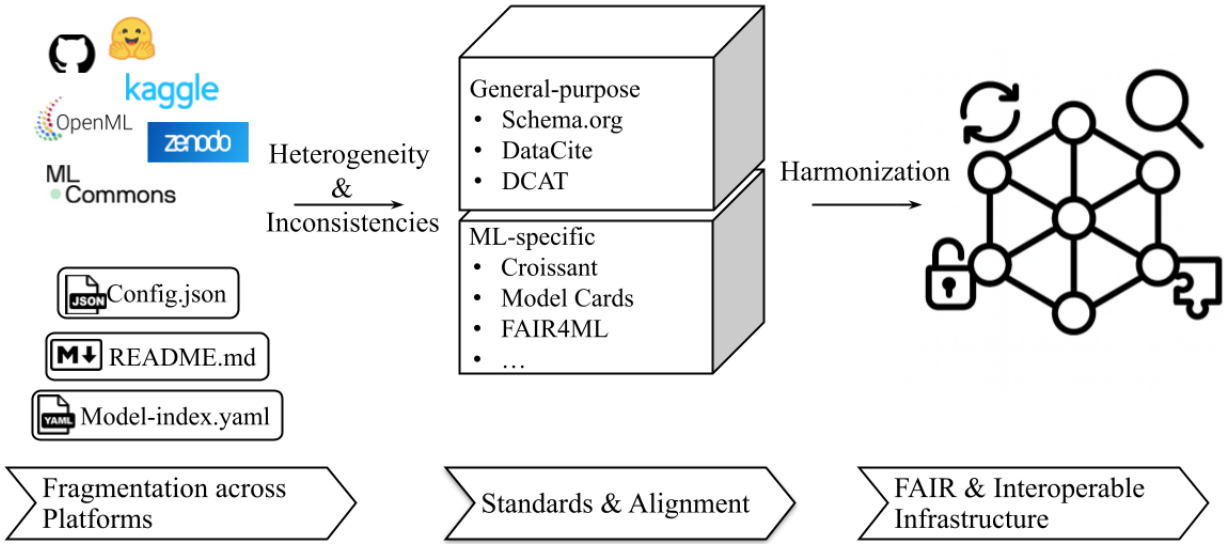

Figure 1: Metadata integration pipeline for ML. The figure illustrates the progression from fragmented metadata across ML platforms, through standardization and semantic alignment, toward FAIR and interoperable metadata infrastructures.

In ML contexts, metadata encompasses descriptive, administrative, structural, provenance, evaluation, and ethical dimensions, each crucial for the reuse and interpretability of ML artifacts. Without such contextual detail, those artifacts cannot be reliably discovered, interpreted, or reused. In this sense, metadata is a cornerstone of FAIR infrastructures: in line with the FAIR principles, metadata alongside data and infrastructure plays an important role, and its relevance extends beyond ML artifacts themselves [5]. In scientific research, particularly in the fields of DS & AI, metadata serves as the connecting layer that describes and contextualizes digital artifacts. In order to address these foundational needs, efforts to improve metadata quality and standardization in ML have emerged, often linked with FAIR research, responsible AI, and data-centric workflows. In this paper, we use the term "standards" in a broad sense, encompassing formal specifications, vocabularies and conceptual models, as well as community practices. Notable initiatives include model cards[6], dataset documentation frameworks, benchmark metadata formats such as the MLCommons Model Index[7], and the adoption of general-

[5] https://www.go-fair.org/fair-principles/
[6] https://huggingface.co/docs/hub/en/model-cards
[7] https://github.com/mlcommons

purpose metadata standards, such as Schema.org[8], DCAT[9], or DataCite[10]. Despite these advances, these initiatives remain largely siloed, often designed for specific platforms. These efforts, while valuable, tend to address specific phases or artifact types and often neglect or have a limited focus on alignment and integration challenges across the full ML life cycle. As a result, these platform-specific metadata silos limit their interoperability in multi-repository or multi-domain applications, which is an essential requirement for scalable scientific infrastructures.

Adding to the issue, the wide range of metadata standards and schema – both general-purpose and domain-specific, as well as the conceptual models, vocabulary terms, and representation formats they adopt, create significant barriers to semantic interoperability, complicating integration efforts across repositories. This leads to (terminology) incompatibility and fragmentation between different platforms. Without systematic alignment and mappings between these heterogeneous metadata standards, it becomes difficult to construct unified metadata layers that can support scenarios like reasoning, querying, or KG construction across platforms. Despite isolated efforts to bridge metadata silos, the field still lacks a shared framework or a consolidated understanding of how existing standards compare in terms of alignment potential, extensibility, and machine-actionability.

Multiple lines of research have focused on formalizing ML metadata. Li et al. [4], for example, proposed a unified representation to query model repositories. Other works survey scientific metadata standards [6], data provenance in computational workflows [7], while Samuel, Löffler & König-Ries [8] and Limani et al. [9] focus on FAIRification of ML pipelines and that of ML models, respectively. However, there is a need to address the ML artifacts from a metadata perspective, such as the suitability of metadata frameworks for cross-platform alignment, semantic interoperability, and integration based on KGs, linked to practical challenges such as dataset search [10] and metadata inconsistencies [11].

The main contributions of this paper are summarized as follows:

- A review and comparative analysis of metadata practices in major ML platforms.
- A review of the existing ecosystem of metadata standards for ML artifacts and their suitability for semantic integration.
- Identification and a detailed discussion of the challenges inherent in mapping, aligning, and integrating heterogeneous ML metadata.
- Identification of key gaps, limitations, and research opportunities in the field of ML metadata management and semantic integration.

2. Metadata Practices in Prominent ML Platforms

This section reviews metadata practices in major ML platforms, focusing on their structure, granularity, and machine-actionability.

2.1. Criteria for Selecting Platforms (CSP)

To ensure a representative and practical comparison, the following six criteria were used to select ML platforms.

- **CSP1: Popularity, Adoption, and Influence.** The platform is widely used in the ML community, as demonstrated by active contributors, hosted artifacts, GitHub metrics, or citations in academic literature, or integration into major workflows in academia or industry.

[8] https://schema.org/
[9] https://www.w3.org/TR/vocab-dcat-3/
[10] https://schema.datacite.org/

- **CSP2: Metadata Accessibility and Machine-Actionability.** The platform exposes metadata in structured or semi-structured formats (e.g., JSON, YAML, XML, or RDF), and provides programmatic access via APIs for retrieving or exporting metadata. Preference is given to platforms whose metadata supports parsing, automated extraction, and reuse without manual intervention.
- **CSP3: ML Artifact Coverage.** The platform supports multiple ML-specific artifact types, including models, datasets, training code, and optionally notebooks or experiment traces, as well-described and retrievable entities.
- **CSP4: Open Access and Licensing Transparency.** Metadata and artifacts are publicly accessible without restrictive authentication or institutional barriers. Clear licensing supports metadata reuse, redistribution, and integration into downstream applications.
- **CSP5: Interoperability Potential.** The platform is relevant to metadata alignment efforts and provides metadata that can be mapped to standard schemas with minimal transformation effort.
- **CSP6: Participation in Standards and Community-Driven Practices.** The platform is involved in or adopts emerging metadata standards, such as the MLCommons' Model Index, or Hugging Face's model-index.yaml, or contributes to open science infrastructure through community initiatives.

2.2. Relevant Machine Learning Platforms

Hugging Face Model & Dataset Hubs is widely adopted for sharing pretrained models and datasets, particularly in natural language processing (NLP) and computer vision (CSP1, CSP3). Metadata is primarily provided through semi-structured `README.md` files, often containing structured YAML headers and markdown descriptions, alongside auxiliary files such as `config.json` and `dataset_infos.json` (CSP2). Additional effort has been undertaken to describe datasets using the Croissant ML extension to Schema.org [12] (CSP2). These elements support common fields like license, task, and language, and often include links to publications such as arXiv papers (CSP5). However, schema adherence varies across entries, and content inconsistencies hinder machine-actionability (CSP2). While the Hugging Face Hub API supports metadata access, it does not enforce schema constraints (CSP6). Nevertheless, community-driven practices such as the use of `model-index.yaml` and cross-publishing on platforms like Zenodo reflect emerging support for structured metadata (CSP4, CSP6). FAIR-oriented tools like MLentory [13] demonstrate ongoing efforts to extract structured metadata for integration into knowledge infrastructures (CSP5).

Zenodo is a general-purpose repository used to archive research artifacts, including datasets and ML models (CSP1, CSP3). It adopts the well-established DataCite schema and assigns persistent DOIs, ensuring long-term preservation and citation capabilities (CSP2,CSP6). Metadata is accessible in XML and JSON formats, with public APIs and license declarations (CSP4). Although metadata for general scientific artifacts is rich, Zenodo lacks expressiveness for ML-specific features, like model architecture or training metrics (CSP5). Thus, integration with ML-specific standards requires supplementary metadata or schema extensions. Despite this, its stability, openness, and alignment with FAIR principles make it a valuable component in cross-platform metadata workflows involving Hugging Face and other platforms (CSP5).

GitHub is a ubiquitous platform for hosting ML-related content, such as serialized models, datasets, and training code, often serving as the origin point for Hugging Face model repositories and Zenodo archives (CSP1, CSP3). Metadata on GitHub is primarily unstructured, embeded within *README.md*, *LICENSE*, or commit messages, without adherence to any formal schema (CSP2). While GitHub provides version control and open access features supportive of reproducibility (CSP4), extracting structured metadata often requires NLP-based or code-based analysis pipelines, limiting machine-actionability and integration.

Kaggle[11] is a platform for hosting ML datasets and notebooks, widely used for competitions, and educational purposes (CSP1, CSP3). Metadata for datasets include column descriptions, file formats, and data types. For notebooks, execution environment, associated datasets, and runtime information are captured in a structured form (CSP2). It provides a public API for accessing metadata, which is generally well-structured and machine-actionable within its ecosystem (CSP2, CSP5). Despite its strong internal schema, metadata remain tightly platform-specific and lack standardized vocabulary reuse, limiting external interoperability (CSP5). Nevertheless, there is an ongoing effort to align metadata for datasets to Croissant ML (CSP2).

MLCommons is a collaborative initiative focused on improving reproducibility, benchmarking, and metadata standardization in ML research and engineering (CSP1, CSP6). It provides a YAML-based schema (*model_index.yaml*) describing model domains, evaluation metrics, and usage context in a structured, machine-readable form (CSP2). Though the scope is currently limited to benchmarking, the quality of structured metadata and community involvement make MLCommons a key actor in ML metadata standardization (CSP5, CSP6). Artifacts are publicly available through repositories or GitHub (CSP4). ML Commons is also the main driver behind Croissant ML (CSP2).

Hugging Face Trending Papers succeeds the now-retired Papers with Code [12], continuing the mission of bridging academic publications and code repositories. It is a discovery interface that highlights recent and popular ML research papers, ranked based on community engagement and GitHub star activity (CSP1, CSP3). While the interface supports paper-code linkage and improves research visibility, the metadata is minimally structured and lacks alignment with formal schemas or standardized vocabularies (CSP2). Consequently, its integration into structured metadata pipelines remains limited. The feature functions primarily as a community-curated signal layer for research discovery rather than a source of machine-actionable metadata (CSP6).

OpenML is a collaborative platform designed for sharing datasets, ML tasks, and experiment results, with a strong focus on traceability and reproducibility (CSP1, CSP3). OpenML provides a REST API and a Python client library for programmatic access, offering excellent machine actionability and semantic transparency (CSP2, CSP4). It aligns closely with FAIR principles and supports metadata standards, including the use of standardized vocabularies and experiment tracking formats (CSP5). It also integrates with platforms like scikit-learn and supports schema extensions such as Croissant ML, reinforcing its role in interoperable ML metadata ecosystems (CSP6).

Summary. ML platforms vary widely in their metadata practices, reflecting differing goals, communities, and technical architectures. OpenML and MLCommons support structured, FAIR-aligned metadata, while GitHub and Hugging Face rely on unstructured formats, limiting interoperability and automation without additional enrichment. Zenodo offers openness and persistent identifiers but lacks ML-specific schema support. Kaggle provides structured metadata within its ecosystem, though with limited external integration. Hugging Face Trending Papers serves as a lightweight discovery interface for recent ML research, but lacks the metadata depth and structure needed for integration into interoperable systems. Overall, while foundational infrastructure exists, metadata practices across platforms remain fragmented. Increased adoption of common schemas, shared vocabularies, and standardized metadata pipelines is essential to enable reproducibility, discoverability, and cross-platform alignment in the ML research ecosystem.

3. Existing Metadata Standards Relevant to ML

Having examined how metadata is currently structured and exposed across major ML platforms, we now turn to the metadata standards that underpin or could enhance these practices. In this section, both general-purpose and ML-specific metadata standards are provided, assessing their applicability to

[11] https://kaggle.com/
[12] https://paperswithcode.com/

describing ML models and datasets, especially in terms of semantic interoperability and cross-platform alignment. This dual focus matters in reconciling the maturity of general-purpose standards with the specific needs of ML metadata. The selection of standards and initiatives presented here followed a structured but pragmatic approach: we considered those (i) explicitly referenced or adopted by major ML platforms, (ii) widely recognized in the broader research data management ecosystem, or (iii) discussed in the existing literature and community initiatives on FAIR and reproducible ML.

3.1. Criteria for Selecting Standards (CSS)

To evaluate metadata standards for their suitability to describe ML artifacts, five criteria are defined. These criteria reflect key technical and conceptual requirements for supporting structured, interoperable, and FAIR-compliant metadata infrastructures in ML.[13]

- **CSS1: Relevance to ML Artifacts** refers to whether the standard is directly applicable to describing ML models, datasets, or experimental workflows.
- **CSS2: Adoption in Research Platforms** considers whether the standard is integrated into widely used repositories, infrastructures, or policy frameworks.
- **CSS3: Semantic Expressiveness** reflects the degree to which the standard supports formal semantics such as RDF, OWL, or Linked Data principles.
- **CSS4: FAIR Alignment** evaluates the extent to which the standard contributes to the achievement of FAIRness, through persistent identifiers, license fields, access protocols, or reuse of vocabularies.
- **CSS5: Maturity and Stability** examines whether the standard is well-specified, actively maintained, and supported by an established community. Considerations include specification completeness, release frequency, and ecosystem support.

These criteria are consistently applied when analyzing standards grouped under general-purpose and ML-specific metadata categories.

3.2. General-Purpose Metadata Standards

These standards are widely used across disciplines and provide essential scaffolding for metadata representation.

Schema.org is a widely adopted for annotating datasets, software, and publications on the web (CSS2, CSS5). The use of JSON-LD enables integration with the Semantic Web (CSS3), and its popularity supports interoperability across systems (CSS4) [14]. While Schema.org does not provide native support for ML-specific entities, e.g., model architectures, training configurations, or evaluation results, it is partially applicable to ML contexts due to its flexible class structure and extensibility (≈CSS1; cf. Figure 2). To address its ML limitations, there are extensions that aim to improve the coverage of software (e.g., CodeMeta [15], maSMPs [16], Croissant ML for datasets [12], and FAIR4ML for ML models [17].

DataCite is a prominent metadata schema designed for the citation and identification of research artifacts, including datasets and software (CSS2, CSS5). It captures administrative and provenance metadata such as `creator`, `publisher`, `DOI` and `publication date` (CSS4), but lacks technical or ML-aware descriptors (¬CSS1). Its semantic depth is limited as it relies on key-value pairs (¬CSS3), and its rigid schema complicates extensions for ML purposes. However, its integration with persistent identifier infrastructures makes it essential for ensuring the citability and long-term preservation of research artifacts.

[13] Notation: CSSi indicates full support for criterion i; ≈ CSSi indicates partial support; ¬CSSi indicates lack of support.

DCAT is a widely implemented vocabulary for describing digital resources and dataset catalogs (CSS2, CSS5). It defines core classes such as `dcat:Dataset`, `dcat:Catalog`, and `dcat:Distribution`, supporting discoverability and interoperability across data catalogs (CSS3, CSS4). However, it is not designed for ML artifacts (¬CSS1), and do not accommodate extensions for tasks, models, or evaluation. Therefore, its function better as bridging vocabularies than as standalone solutions for ML-specific metadata integration.

3.3. ML-Specific Metadata Standards

The limitations of general-purpose standards in describing ML-specific artifacts have led to the emergence of dedicated metadata standards designed to capture the unique semantics of ML, offering greater granularity, semantic expressiveness, and task-specific coverage.

Croissant ML [12] is a JSON-LD metadata specification developed by Google to describe ML datasets (CSS1). It provides detailed structure descriptions for dataset components, including features, files, licences, and schema types, enabling machine-actionable metadata and alignment with FAIR priciples (CSS3, CSS4). Though still emerging (CSS2), it demonstrates strong extensibility and is built upon mature vocabularies and schemas like Schema.org (≈CSS5; see Figure 2).

Model Cards [18] are semi-structured documents designed to communicate the usage scenarios, limitations, evaluation metrics, and ethical considerations of ML models (CSS1, CSS4). They have moderate adoption in applied ML communities (CSS2), especially in platforms like Hugging Face. However, Model Cards are primarily designed for human interpretation and lack a formalized schema or machine-actionable structure, limiting their semantic expressiveness and integration (¬CSS3, ¬CSS5).

FAIR4ML[14] **[17]** is an ontology-based extension of Schema.org designed to enhance the FAIRness of ML model documentation (CSS1, CSS3, CSS4). It introduces semantically precise terms for modeling evaluation metrics, intended applications, and tasks, thereby enabling machine-actionable metadata across ML workflows. Despite its rich semantic expressiveness, it is relatively new with limited adoption in mainstream ML platforms (¬CSS2), and its tooling ecosystem is still developing (¬CSS5).

ML Schema[15] is a lightweight, extensible vocabulary for describing ML experiments, models, algorithms, and metrics (CSS1, CSS3). It enables semantic annotation of ML workflows and supports integration with linked data infrastructures, aligning well with FAIR principles (CSS4). While semantically expressive and conceptually accessible, ML Schema has seen limited adoption in major ML platforms (¬CSS2), and its tooling ecosystem is still underdeveloped, with minimal support (¬ CSS5).

OntoDM-core [19] is a foundational ontology for the data mining domain that provides a formal representation of core concepts such as data, tasks, algorithms, models, and results (CSS1, CSS3, CSS5). It supports logical inference and reuse across domains. However, it has limited uptake in mainstream ML platforms (¬ CSS2) and only partial alignment with FAIR practices (≈CSS4).

Expose Ontology [20] focuses on modeling the experimental design and execution of data analysis processes, with a particular emphasis on provenance (CSS1, CSS3). It is semantically rich and complements PROV-O [16], but has seen little adoption outside specific projects (¬ CSS2) and lacks broader tooling support (¬ CSS5). Its alignment with FAIR is partial (≈ CSS4).

DMOP [21] addresses the optimization dimension of data mining and ML workflows, especially in relation to algorithm selection, hyperparameter tuning, and performance evaluation (CSS1, CSS3, CSS5). DMOP is semantically expressive and extensible, making it suitable for modeling complex ML pipelines and adaptations in domains like AutoML or meta-learning. However, its adoption is niche (¬ CSS2), and its alignment with FAIR principles lacks sufficient support (¬ CSS4).

[14] https://w3id.org/fair4ml
[15] https://ml-schema.github.io
[16] https://www.w3.org/TR/prov-o/

MEX [22] offers a lightweight vocabulary for describing ML experiments, including datasets, algorithms, hyperparameters, and results (CSS1, CSS3,CSS4). MEX has not seen significant adoption in mainstream ML repositories or tools (¬CSS2). Its formal specification is available, but the ecosystem around implementation, tooling, and maintenance remains limited (¬CSS5).

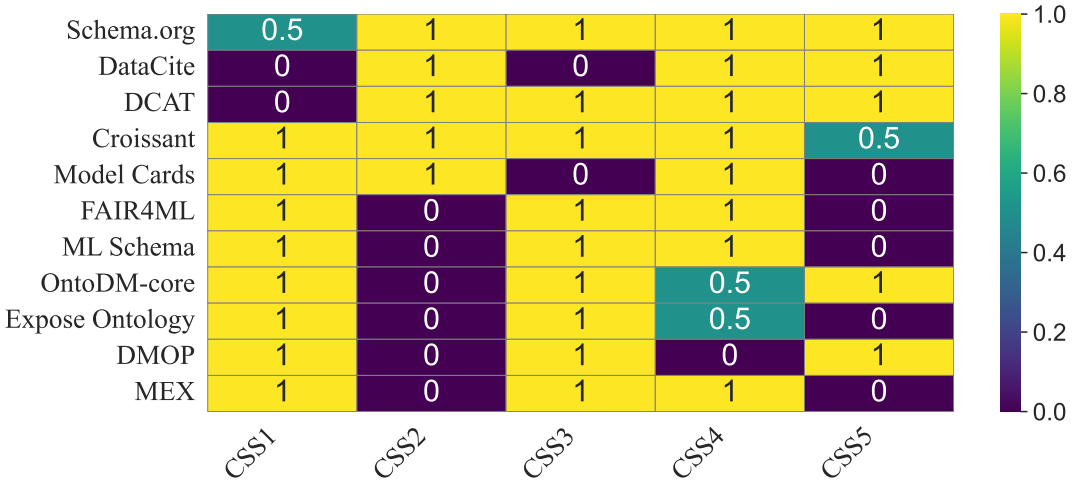

Figure 2: Heatmap of metadata standards evaluated against five ML-relevant criteria. Each row represents a metadata standard, and each column corresponds to a specific evaluation criterion: CSS1: Relevance to ML Artifacts, CSS2: Adoption in Research Platform, CSS3: Semantic Expressiveness, CSS4: : FAIR Alignment, and CSS5: Maturity and Stability. Scores are normalized as follows: 1 indicates full support for the criterion, 0.5 indicates partial support, and 0 indicates lack of support. This visualization highlights which standards most comprehensively fulfill the requirements for interoperable, machine-actionable metadata in ML ecosystems.

Summary. This section outlines the distinction between general-purpose and ML-specific metadata standards. Mature and widely adopted standards such as Schema.org, DataCite, and DCAT support discoverability and citation but lack the semantic richness and extensibility required to describe ML-specific artifacts, including models, training configurations, and evaluations. In contrast, ML-specific standards, such as Croissant, FAIR4ML, ML Schema, and Model Cards address these gaps but remain in early stages of adoption, with limited tooling. Ontology-driven approaches like OntoDM-core, Expose, DMOP, and MEX provide formal representations and reasoning capabilities but vary in scope and maturity. As expected, no single standard fully satisfies all requirements. Figure 2 presents a comparative evaluation based on criteria defined in Section 3.1, *showing that general-purpose standards score well on adoption and maturity (CSS2, CSS5) but poorly on ML relevance (CSS1), while ML-specific and ontology-driven approaches provide higher semantic expressiveness (CSS3) yet limited adoption (CSS2).* Moreover, a summary of these standards covering primary focus, semantic expressiveness, machine-actionability, support for provenance, support for ethical/bias information, and extensibility, is provided in Table 1 in Appendix.

4. Metadata Harmonization

The heterogeneity and fragmentation of metadata practices across ML platforms and standards introduce substantial barriers to interoperability. Even when a structured metadata is used, it is often implemented with divergent schemas, inconsistent vocabularies, and varying levels of detail. For example, platforms such as GitHub, Zenodo, Hugging Face, and OpenML differ in how they represent authorship, tasks, licensing, and performance metrics, making direct alignment difficult without additional normalization or transformation. These inconsistencies hinder the seamless integration and reuse of metadata across systems, complicate downstream applications that rely on unified metadata, and ultimately reduce the discoverability, traceability, and reproducibility of ML artifacts. Addressing these issues requires

effective strategies for metadata extraction, mapping, and harmonization across heterogeneous sources.

4.1. Challenges in Metadata Harmonization

Harmonizing metadata from heterogeneous platforms introduces several challenges:

- **Schema Heterogeneity:** Different platforms adopt different data models, property labels, and data types to describe the same concepts. For instance, the property referring to the creator of a model may be labeled as `author` in Schema.org, `creator` in DataCite.
- **Vocabulary Inconsistencies:** Even when schemas are conceptually aligned or share a common vocabulary, communities may interpret the same terms differently, leading to semantic drift. For example, the same ML task may be labeled as *classification*, *categorization*, or *label prediction*, or conversely, the term *accuracy* may refer to different evaluation protocols depending on context. These inconsistencies complicate cross-platform querying, semantic alignment.
- **Granularity Mismatch:** Metadata varies in depth and detail across platforms. Some platforms provide high-level descriptors (e.g., model family and task domain), while others offer fine-grained specifications such as hyperparameters, environment settings, or evaluation protocol. For example, a large language model may be referred to simply as *LLaMA*, or more precisely as *LLaMA-7B* or *LLaMA-13B*, each with distinct architectures and training settings. Aligning across these granularity levels requires careful abstraction or enrichment strategies.
- **Semantic Ambiguity:** Some commonly used terms are themselves semantically vague or overloaded. For example, *accuracy* may refer to different evaluation metrics (e.g., Top-1 vs. Top-5), data splits (e.g., test vs. validation), or output settings (e.g., single-label vs. multi-label), unless explicitly defined.
- **Unstructured Metadata:** Crucial metadata is often embedded in free-text sources such as `README.md` files, model cards, or publications. While information extraction (IE) techniques can be applied, the process does not always yield structured outputs that are complete or reliable enough for downstream integration tasks.
- **Provenance Gaps:** Many metadata records lack information about the origin and transformation history of datasets and models, limiting trust and traceability.

4.2. Techniques for Metadata Harmonization

Metadata Mapping and Crosswalks. A common approach to dealing with metadata heterogeneity is via *mapping* concepts from one standard to another. Metadata mapping can be applied to a broad set of cases, – from less semantic approaches (e.g., DataCite), to non-semantic ones (e.g., model cards or schemas used internally in a specific platform), to semantic ones (ontologies). In line with this idea, *metadata crosswalks* [23] are nowadays used to map metadata. Usually manually defined, they map properties in different schemas, and are often managed and structured as spreadsheets. Metadata crosswalks provide interpretability and flexibility, but they are labor intensive, error-prone, and difficult to scale. Moreover, they lack formal semantics, limiting their utility in Linked Data and automated environments, which limits their applicability.

To improve traditional methods, structured mapping frameworks like the Simple Standard for Sharing Ontology Mappings (SSSOM) [24] have been developed. SSSOM enables formal, machine-readable mappings with metadata for provenance, alignment type (e.g., exact, broader), and confidence scores. For instance, the concept `DataCite:creator` can be mapped as `skos:exactMatch` to the concept `schema:author`. SSSOM-style mappings are being piloted in tools like the NFDI4DS QA[25] service [17]

[17] https://nfdi-search.nliwod.org/

to enable structured integration across metadata schemes. A modified version is being used for the update of CodeMeta [18] crosswalks.

Despite these advances, large-scale adoption of semantic crosswalks remains limited. The creation of high-quality semantically precise mappings still relies heavily on expert curation, which is a bottleneck for scalability.

Automated Extraction and Harmonization. Automated techniques enhance scalability and consistency in metadata harmonization by reducing manual effort. NLP methods, such as named entity recognition, relation extraction, and classification, are widely used to extract structured metadata from sources like README files, model cards, and research papers [26]. These approaches identify entities (e.g., models, datasets), relationships (e.g., "trained on"), and attributes (e.g., license, metrics). Schema matching and ontology alignment aim to reconcile concepts across metadata schemas using lexical, structural, or instance-based similarity. Recent approaches incorporate embeddings and LLMs to improve semantic alignment [27, 28]. Following extraction and alignment, validation and normalization ensure semantic consistency across metadata sources [29], using rule-based constraints or ontology-aware checks. While automation improves efficiency, expert oversight remains essential for accuracy and explainability in integrated metadata pipelines.

Summary. Metadata harmonization is a persistent challenge in ML due to fragmented schemas, inconsistent vocabularies, and missing provenance. Manual approaches like crosswalks and SSSOM offer structured mappings but require expert effort. Automated methods, such as NLP-based extraction, schema matching, and validation, enhance scalability but still face issues of accuracy and explainability. A hybrid approach that combines semantic precision, automation, and expert oversight is essential to build interoperable and reusable metadata infrastructures.

5. Gaps, Limitations, and Research Opportunities

Despite ongoing efforts to formalize metadata practices in ML, the current landscape remains fragmented and underdeveloped in several critical areas, opening compelling directions for future research.

5.1. Current Gaps and Limitations

A major limitation is the absence of a unified, comprehensive metadata standard tailored to the full lifecycle of ML artifacts. While efforts such as FAIR4ML and ML Schema have made notable progress, no widely adopted standard exists yet that captures the full range of ML entities, including models, datasets, evaluation metrics, workflows, and ethical dimensions, in an integrated and expressive way. Another persistent challenge lies in the representation of dynamic metadata. ML models and datasets are inherently mutable, often updated through processes such as fine-tuning, retraining, or automated CI/CD pipelines. Existing metadata standards are largely static in design and tend to focus on fixed snapshots of artifacts. As a result, they provide limited support for describing evolving provenance, behavioral shifts, or version histories in a machine-actionable and consistent manner.

Scalability presents an additional concern. Current integration strategies often rely on manual curation or semi-automated tools that do not scale effectively to the growing volume and diversity of ML artifacts across platforms. Furthermore, automated and robust metadata integration pipelines remain in an early stage of development. Finally, while bias documentation is becoming more standardized, broader aspects of ethical metadata, including privacy, safety, explainability, and societal impact, are not yet consistently represented across standards or platforms. At this point, a general-purpose ethical metadata vocabulary for ML is still lacking.

[18] https://codemeta.github.io/

5.2. Future Research Directions

Addressing these limitations calls for several targeted lines of investigation. First, there is a growing need for methods to automate metadata extraction and generation. Future research should leverage advances in NLP, code analysis, and LLMs to infer structured metadata from documentation, code repositories, and execution traces, thereby reducing reliance on manual input. Second, the development of semantic interoperability frameworks for ML is crucial. Such frameworks should combine extensible terminology solutions (ontologies, vocabularies, etc.), possibly with Linked Data principles for machine readability, to enable automated alignment and querying across heterogeneous platforms. Research is particularly needed in automated ontology matching and mapping tailored to the ML domain, which remains underexplored. Third, standardization efforts around ethical AI metadata should be expanded. This includes formalizing descriptors for fairness, transparency, accountability, and explainability, potentially as extensions to existing efforts like Model Cards. These standards should aim to be both human-interpretable and machine-actionable. Lastly, future work must address the management of dynamic and versioned metadata. Novel models and infrastructures are required to track the evolution of ML artifacts over time, capturing temporal and contextual changes in training data, hyperparameters, and performance outcomes.

Together, these research directions represent a roadmap toward more robust, scalable, and ethically grounded metadata ecosystems for ML.

6. Conclusion

The growing availability of ML models and datasets highlights the compelling need for structured, standardized metadata supporting research artifacts across repositories. This survey analyzed metadata practices and standards for ML artifacts, identifying challenges such as schema heterogeneity, inconsistent granularity, and unstructured documentation that impede integration and semantic interoperability. Strategies such as schema alignments, crosswalks, and the use of shared conceptual models enable integration and semantic interoperability across platforms, supporting consistent interpretation of metadata despite underlying heterogeneity. While initiatives like Model Cards and FAIR4ML show promise, gaps remain in unified ontologies, dynamic metadata management, and automated tooling. Addressing these limitations requires sustained community effort, not only in developing and adopting robust metadata standards, but in establishing a uniform conceptualization of the ML life cycle. Such a shared foundation would facilitate consistent mappings between standards and improve best practices in applying them. This collective effort is essential for building scalable, FAIR-compliant metadata infrastructures that support discovery, traceability, and reuse across the ML ecosystem.

As a continuation of this survey, future work will focus on exploring the practical integration of standardized metadata into downstream ML applications, such as KGs for semantic search and discovery. This step will further enhance the insights provided by this survey by expanding the exploration of how FAIR metadata supports automation, reproducibility, and knowledge discovery across the ML ecosystem. This work will be addressed in a forthcoming version of the study with an extended scope.

7. Acknowledgment

This work has been partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), NFDI4DS (Grant number 460234259). Authors acknowledge the data sources and also thank the individuals involved in this research.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check. After using this tool/service, the authors reviewed and edited the content as needed and take

full responsibility for the publication's content.

References

[1] M. Schlegel, K.-U. Sattler, Management of machine learning lifecycle artifacts: A survey, ACM SIGMOD Record 51 (2023) 18–35.

[2] Z. Li, R. Hai, A. Bozzon, A. Katsifodimos, Metadata representations for queryable ml model zoos, arXiv preprint arXiv:2207.09315 (2022).

[3] S. G. Labou, A. Pennington, H. J. S. Yoo, M. Baluja, Toward enhanced reusability: A comparative analysis of metadata for machine learning objects and their characteristics in generalist and specialist repositories, Journal of eScience Librarianship 13 (2024).

[4] Z. Li, H. Kant, R. Hai, A. Katsifodimos, M. Brambilla, A. Bozzon, Metadata representations for queryable repositories of machine learning models, IEEE Access 11 (2023) 125616–125630. doi:10.1109/ACCESS.2023.3330647.

[5] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (2016) 1–9.

[6] D. I. Hillmann, R. Marker, C. Brady, Metadata standards and applications, The Serials Librarian 54 (2008) 7–21.

[7] J. Freire, D. Koop, E. Santos, C. T. Silva, Provenance for computational tasks: A survey, Computing in science & engineering 10 (2008) 11–21.

[8] S. Samuel, F. Löffler, B. König-Ries, Machine learning pipelines: Provenance, reproducibility and fair data principles, in: B. Glavic, V. Braganholo, D. Koop (Eds.), Provenance and Annotation of Data and Processes, Springer International Publishing, Cham, 2021, pp. 226–230.

[9] F. Limani, L. Tofik, A. Latif, K. Tochtermann, Fair for machine learning model: Principles and assessment metrics, 2024. URL: https://doi.org/10.5281/zenodo.13835105. doi:10.5281/zenodo.13835105.

[10] M. Hulsebos, W. Lin, S. Shankar, A. Parameswaran, It took longer than i was expecting: Why is dataset search still so hard?, in: Proceedings of the 2024 Workshop on Human-In-the-Loop Data Analytics, HILDA 24, Association for Computing Machinery, New York, NY, USA, 2024, p. 1–4. URL: https://doi.org/10.1145/3665939.3665959. doi:10.1145/3665939.3665959.

[11] K. Ito, S. Matsubara, Estimating metadata of research artifacts to enhance their findability, in: 2024 IEEE 20th International Conference on e-Science (e-Science), 2024, pp. 1–2. doi:10.1109/e-Science62913.2024.10678684.

[12] M. Akhtar, O. Benjelloun, C. Conforti, L. Foschini, J. Giner-Miguelez, P. Gijsbers, S. Goswami, N. Jain, M. Karamousadakis, M. Kuchnik, et al., Croissant: A metadata format for ml-ready datasets, Advances in Neural Information Processing Systems 37 (2024) 82133–82148.

[13] D. Solanki, N. Quiñones, D. Rebholz-Schuhmann, L. Jael, Mlentory, an fdo registry for machine learning models (2024).

[14] K. Payne, C. Verhey, Schema. org for research data managers: a primer, International Journal of Big Data Management 2 (2022) 95–116.

[15] M. B. Jones, C. Boettiger, A. C. Mayes, A. Smith, P. Slaughter, K. Niemeyer, Y. Gil, M. Fenner, K. Nowak, M. Hahnel, L. Coy, A. Allen, M. Crosas, A. Sands, N. C. Hong, P. Cruse, D. Katz, C. Goble, CodeMeta: an exchange schema for software metadata. KNB Data Repository, 2016. URL: https://raw.githubusercontent.com/codemeta/codemeta/1.0/codemeta.jsonld. doi:10.5063/SCHEMA/CODEMETA-1.0, medium: application/ld+json.

[16] L. J. Castro, O. Giraldo, L. Geist, N. Quiñones, D. Solanki, D. Rebholz-Schuhmann, machine-actionable Software Management Plan Ontology (maSMP Ontology), 2024. URL: https://zenodo.org/records/10582073. doi:10.5281/zenodo.10582073, publisher: Zenodo.

[17] L. J. Castro, D. Garijo, D. Rebholz-Schuhmann, D. Solanki, J. T. Ciuciu-Kiss, D. Katz, L. Eklund, G. Bharathy, R. D. A. F. Task, FAIR4ML-schema, 2024. URL: https://zenodo.org/records/14002310. doi:10.5281/zenodo.14002310, publisher: Zenodo.

[18] 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.
[19] P. Panov, L. Soldatova, S. Džeroski, Ontology of core data mining entities, Data Mining and Knowledge Discovery 28 (2014) 1222–1265.
[20] J. Vanschoren, L. Soldatova, Exposé: An ontology for data mining experiments, in: International workshop on third generation data mining: Towards service-oriented knowledge discovery (SoKD-2010), 2010, pp. 31–46.
[21] C. M. Keet, A. Ławrynowicz, C. d'Amato, A. Kalousis, P. Nguyen, R. Palma, R. Stevens, M. Hilario, The data mining optimization ontology, Journal of web semantics 32 (2015) 43–53.
[22] D. Esteves, D. Moussallem, C. B. Neto, T. Soru, R. Usbeck, M. Ackermann, J. Lehmann, Mex vocabulary: a lightweight interchange format for machine learning experiments, in: Proceedings of the 11th International Conference on Semantic Systems, ACM, 2015, pp. 169–176.
[23] J. Martínková, N. Juty, A. Gonzalez-Beltran, C. Goble, Y. Le Franc, Moving towards fair mappings and crosswalks., in: FAIR principles for Ontologies and Metadata in Knowledge Management, 2024.
[24] N. Matentzoglu, J. P. Balhoff, S. M. Bello, C. Bizon, M. Brush, T. J. Callahan, C. G. Chute, W. D. Duncan, C. T. Evelo, D. Gabriel, et al., A simple standard for sharing ontological mappings (sssom), Database 2022 (2022) baac035.
[25] H. B. Giglou, T. A. Taffa, R. Abdullah, A. Usmanova, R. Usbeck, J. D'Souza, S. Auer, Scholarly question answering using large language models in the nfdi4datascience gateway, Natural Scientific Language Processing and Research Knowledge Graphs (2024) 3.
[26] J. L. Martinez-Rodriguez, A. Hogan, I. Lopez-Arevalo, Information extraction meets the semantic web: a survey, Semantic Web 11 (2020) 255–335.
[27] Y. Xiang, Z. Zhang, J. Chen, X. Chen, Z. Lin, Y. Zheng, Ontoea: Ontology-guided entity alignment via joint knowledge graph embedding, arXiv preprint arXiv:2105.07688 (2021).
[28] K. Higashi, Z. Nakagawa, T. Yamada, H. Mori, Automated harmonization and large-scale integration of heterogeneous biomedical sample metadata using large language models, bioRxiv (2024) 2024–10.
[29] M. Hosseini, S. P. Horbach, K. Holmes, T. Ross-Hellauer, Open science at the generative ai turn: An exploratory analysis of challenges and opportunities, Quantitative science studies 6 (2025) 22–45.

A. A summary of Existing Standards

Table 1
Overview of existing standards.

Standard	Primary Focus	Semantic Expressiveness	Machine-actionability	Support for Provenance	Support for Ethical/Bias Info	Extensibility
Schema.org	generic web content	low to moderate (broad coverage no deep semantics)	high	yes	no	high
DataCite	research outputs	bibliographic and administrative metadata	moderate	yes	no	low, very strictly defined
DCAT	datasets in web catalogs	moderate (RDF vocabulary, supports linking datasets...)	high	yes	no	high
Croissant	ML datasets	moderate (Structured metadata using JSON-LD)	high	yes	yes	high
Model Cards	ML models	low (no formal semantics)	low	no	yes	moderate
FAIR4ML	FAIR metadata for ML models	moderate to high	high	yes	yes	high
ML Schema	ML experiments, models, algorithms, metrics	high (OWL-based schema)	high	no	no	high
OntoDM-core	data mining and ML ontology	moderate	high	yes	no	high
Expose Ontology	ML experiment design, provenance focused	high (OWL-based with terms for ML configurations and executions.)	high	yes	no	high
DMOP	ML workflows	high (Highly expressive OWL ontology)	high	yes	no	high
MEX	ML experiments	moderate (light OWL-based schema)	high	yes	no	moderate