MLSeascape: Search over Machine Learning Metadata Empowered by Knowledge Graphs

Ioannis Dasoulas[0000−0002−8803−1244], Duo Yang[0009−0008−5942−3397], and Anastasia Dimou[0000−0003−2138−7972]

KU Leuven – Leuven.AI – Flanders Make@KULeuven
{firstname.lastname}@kuleuven.be

Abstract. As Machine Learning (ML) continuously grows in numbers, complexity and components, online ML platforms that gather and serve ML-related knowledge become increasingly important. However, available knowledge is fragmented with each platform representing distinct parts of the ML lifecycle using their own unique representation. In this demo paper, we present MLSeascape, an online application that leverages the MLSea-KG knowledge graph. The MLSea-KG knowledge graph incorporates ML metadata from multiple platforms, such as Kaggle, OpenML and Papers with Code. MLSeascape enables seamless search for ML metadata without needing to be an expert in semantic web technologies.
License: Apache-2.0
MLSeascape: https://w3id.org/mlseascape
Source Code: https://github.com/dtai-kg/MLSeascape
Video: https://youtu.be/jn-GGwm52EM

Keywords: Machine Learning · Knowledge Graphs · Search Systems

1 Introduction

Due to the continuous advancements of Machine Learning (ML), the number and complexity of ML pipelines keep on increasing. ML pipelines incorporate numerous sophisticated components, including datasets, hyper-parameters, algorithms and software, which can hinder their discoverability. Online platforms (e.g., OpenML[1], Kaggle[2] or Papers with Code[3]) play a pivotal role in cataloging ML experiments and metadata. Yet, these platforms vary in their data representations, captured metadata, interfaces and search functionalities, making it challenging for users to efficiently discover and retrieve relevant information.

Recent works demonstrated how ML knowledge discovery can be facilitated by semantically annotating and serving ML metadata from various sources [1,2]. However, existing works require knowledge of semantic web technologies and

[1] https://www.openml.org
[2] https://www.kaggle.com
[3] https://paperswithcode.com

the underlying data model to effectively use them for ML discovery. A user-friendly abstraction layer is needed to make ML search more accessible to a wider audience while leveraging the underlying ML knowledge graph.

In this paper, we introduce **MLSeascape**[4], a knowledge graph-enhanced web application that provides an abstraction layer for discovering ML metadata from online platforms. **MLSeascape** serves diverse ML components and is able to demonstrate their properties and relationships between them, as well as to provide the original sources they are found. By leveraging semantic web technologies, **MLSeascape** aims to streamline the process of accessing and exploring ML resources, facilitating innovation and collaboration in ML.

In the rest, we detail the architecture of **MLSeascape** (Sect. 2), describe the system's functionalities (Sect. 3) and conclude with future work (Sect. 4).

2 MLSeascape as Part of the MLSea Ecosystem

MLSeascape is a web application that aims to facilitate the discovery of ML metadata from various online platforms. It is a search-based platform where users can explore different ML components and view their properties, the relationships between components and their original sources. The data is retrieved from a public SPARQL endpoint[5] hosted by a GraphDB[6] triple store, which is enhanced with a built-in Lucene Connector[7] that accelerates text searches over the triple store while ensuring synchronization with it.

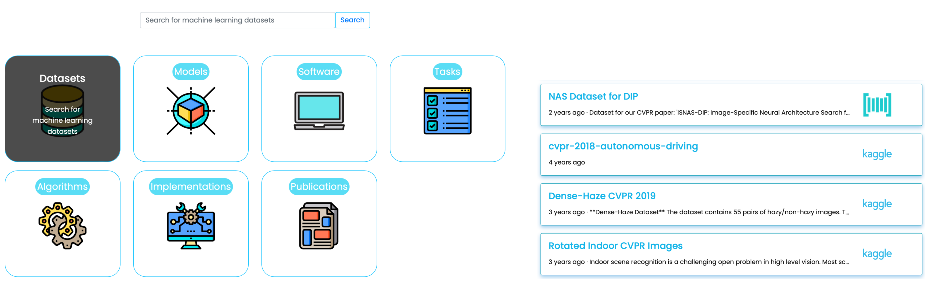
MLSeascape leverages the Machine Learning Knowledge Graph (MLSea-KG)[8] [1] to retrieve ML metadata which are hosted in a GraphDB triple store. MLSea-KG is a comprehensive and regularly updated knowledge graph containing over 1.44 billion RDF triples of ML experiments. It encompasses ML metadata from OpenML, Kaggle and Papers with Code, including datasets used in ML experiments, tasks, implementations, hyper-parameters, experiment executions, configuration settings, evaluation results, code notebooks, algorithms, publications, models, and information about scientists and practitioners.

MLSea-KG is based on the Machine Learning Sailor Ontology (MLSO[9]) and Taxonomies (MLST[10]) [1]. MLSO is an ontology that reuses and extends state-of-the-art ontologies (e.g., MLSchema [6], SDO [3], DCAT [4], FaBiO [5] to describe ML workflows, configurations, experimental results, models, datasets, software implementations and citations. MLST are SKOS taxonomies of ML-related concepts (e.g., task types, evaluation measures).

4 `https://w3id.org/mlseascape`
5 `https://193.190.127.194:7200`
6 `https://graphdb.ontotext.com`
7 `https://graphdb.ontotext.com/documentation/10.6/lucene-graphdb-connector.html`
8 Paper accepted by ESWC 2024 Resource Track
9 `http://w3id.org/mlso`
10 `https://github.com/dtai-kg/MLSO/tree/main/ontology/Taxonomies`

3 Demonstration of MLSeascape

MLSeascape allows users to search for different types of ML artifacts including datasets, models, software, tasks, algorithms, implementations and publications by traversing MLSea-KG. Users first select the type of ML artifact (e.g., datasets) they are interested to search for and input a related keyword (Figure 1a). **MLSeascape** then presents potential matches for their search input in the MLSea-KG. For instance, in (Figure 1b) some results for the search "NAS CVPR" are displayed. When the users select one of the matches, they are led to a new page that presents all generic metadata about their choice (e.g., date published, creators, description, original source) as well as related ML entities (e.g., similar datasets, related software, related ML tasks, publications) for the corresponding artifact. For instance, in (Figure 2), the user selected the "NAS" dataset and MLSeascape retrieved and presented the dataset's metadata as well as relationships with other ML artifacts found online, such as similar datasets, related code repositories, and the publication that introduced the dataset.



(a) Homepage (b) Search Matches

Fig. 1: Artifacts Search

MLSeascape harnesses MLSea-KG's coverage of different ML platforms and its sophisticated relationships between ML artifacts. It emphasizes on revealing interconnections between artifacts that will help users get a more holistic view of the existing works that utilize them. It also focuses on simplicity, allowing non-experts to traverse through MLSea-KG with a single user input.

In the demo, we will demonstrate the full process of searching and browsing ML artifacts. All available types of artifacts and metadata will be demonstrated using prominent ML artifacts, as well as user inputs given during the presentation by the attendees.

4 Conclusion

In this paper, we introduce **MLSeascape**, a user-friendly web application designed to facilitate the exploration of ML artifacts and resources from diverse

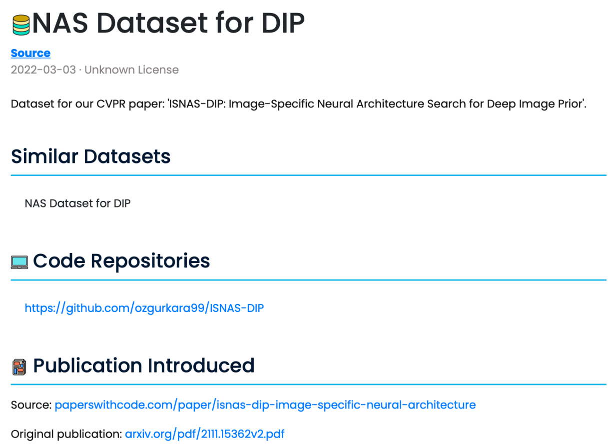


Fig. 2: Dataset Metadata Presentation Snippet

platforms. **MLSeascape's** coverage over ML metadata empowers practitioners and researchers to explore ML more effectively, aiming to promote innovation and advancements in the field. In the future, we aim to expand the coverage of **MLSeascape** by incorporating additional information or platforms that will be available through MLSea-KG. Additionally, we will enhance the search functionalities and improve **MLSeascape** by implementing artifact recommendations.

Acknowledgements This research was partially supported by Flanders Make (REXPEK project), the strategic research centre for the manufacturing industry and the Flanders innovation and entrepreneurship (VLAIO – KG3D project).

References

1. Dasoulas, I., Yang, D., Dimou, A.: MLSea: A Semantic Layer for Discoverable Machine Learning. In: The Semantic Web (2024)
2. Färber, M., Lamprecht, D.: Linked Papers With Code: The Latest in Machine Learning as an RDF Knowledge Graph. In: ISWC 2023 Posters and Demos (2023)
3. Garijo, D., Osorio, M., Khider, D., Ratnakar, V., Gil, Y.: OKG-Soft: An Open Knowledge Graph with Machine Readable Scientific Software Metadata. In: 15^{th} International Conference on eScience (eScience). pp. 349–358 (2019)
4. Maali, F., Erickson, J.: Data Catalog Vocabulary (DCAT). Recommendation, World Wide Web Consortium (W3C) (Jan 2014), `https://www.w3.org/TR/vocab-dcat/`
5. Peroni, S., Shotton, D.: FaBiO and CiTO: ontologies for describing bibliographic resources and citations. Journal of Web Semantics **17**, 33–43 (2012)
6. Publio, G.C., Esteves, D., Ławrynowicz, A., Panov, P., Soldatova, L., Soru, T., Vanschoren, J., Zafar, H.: ML-Schema: Exposing the Semantics of Machine Learning with Schemas and Ontologies. arXiv preprint arXiv:1807.05351 (2018)