

Datalens: A Semantic Model and Knowledge Graph for Discovery and Exploration of Machine Learning Resources

Zachari Arnaud¹, Aline Menin^{1,*}, Anaïs Ollagnier¹ and Catherine Faron¹

¹*I3S Laboratory, University Côte d'Azur, CNRS, Inria
2000 Rte des Lucioles, 06900 Sophia Antipolis, France*

Abstract

The rapid proliferation of publicly available machine learning resources complicates the identification of relevant datasets and models. Although numerous repositories provide access to these resources, they rely on heterogeneous metadata schemas, which hinders both intra- and cross-repository discovery. In this paper, we present Datalens, a semantic approach for discovery and exploration of machine learning resources published on the Web. Datalens combines an OWL ontology and a task-oriented SKOS thesaurus to provide a unified semantic layer over machine learning metadata, and an RDF knowledge graph instantiating the ontology that describes datasets and models from the Hugging Face ecosystem. We demonstrate the expressiveness and practical utility of our approach through the implementation of competency questions into SPARQL queries, whose results support intuitive exploration and interpretation of machine learning resources by IT researchers and developers.

Keywords

Knowledge Graphs, Ontology Engineering, Machine Learning Metadata, Dataset Discovery, Semantic Search

1. Introduction

The rapid growth of publicly available machine learning (ML) resources has resulted in an unprecedented volume and diversity of datasets and models hosted across repositories such as Hugging Face. Discovering resources that satisfy specific requirements remains challenging due to heterogeneous metadata schemas, incomplete descriptions, and platform-specific representations. In particular, ML tasks, which represent the primary semantic link between datasets, models, and user goals, are often encoded as flat labels or loosely defined categories, limiting semantic search, multi-criteria exploration, and interoperability across repositories.

Dataset discovery has received considerable attention in recent years, with surveys highlighting advances in search strategies, query processing, ranking, and metadata management [1, 2, 3, 4, 5]. Nevertheless, metadata quality remains a major limitation, as descriptions are often incomplete, inconsistent, and weakly structured [6]. Knowledge Graphs (KG) grounded in Semantic Web standards such as RDF, RDFS, and OWL, provide a natural foundation for addressing these challenges. Established vocabularies such as DCAT [7] and PROV-O [8] support interoperable descriptions of datasets and provenance, while ontologies such as ML-Schema [9] and MLSO [10] model different aspects of the ML lifecycle. In parallel, KG-based approaches such as HuggingKG [11], MLTaskKG [12], and LinkedPapersWithCode [13] have demonstrated the value of structured representations for recommendation and resource discovery.

Existing work has proposed either semantic models for representing ML resources and workflows, or KGs supporting scholarly analysis, recommendation, and graph mining. However, comparatively less attention has been devoted to semantic representations specifically designed to support the discovery, exploration, and analysis of ML resources within large-scale Web repositories. In particular, existing approaches provide limited support for jointly exploring heterogeneous metadata, provenance relationships, usage indicators, and scholarly references through a unified semantic representation. Datalens addresses this gap by introducing a resource-centric KG that integrates these complementary

Posters, Demos, Blue Sky, and Tutorials at SEMANTiCS 2026, Sep 2026, Ghent, Belgium

*Corresponding author.

✉ zachari.arnaud@inria.fr (Z. Arnaud); aline.menin@inria.fr (A. Menin); anais.ollagnier@inria.fr (A. Ollagnier); catherine.faron@inria.fr (C. Faron)

🆔 0000-0002-9345-3994 (A. Menin); 0000-0003-4117-1191 (A. Ollagnier); 0000-0001-5959-5561 (C. Faron)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

dimensions to support semantic querying, exploratory navigation, and repository-level analysis of ML resources. Specifically, our contributions are as follows:

- an OWL and SKOS model to describe ML datasets and models, including provenance and discovery-oriented metadata;
- an RDF KG instantiating the model and describing Hugging Face resources, queryable via a public SPARQL endpoint; and
- a demonstration of semantic search and exploration of ML resources through SPARQL queries and interactive visualizations.

2. From User Needs to Competency Questions

The design of our semantic model was driven by competency questions (CQs), elicited via questionnaire involving six ML practitioners. These CQs represent the requirements that guided our approach. Below, we present four illustrative CQs covering discovery, exploration, and repository analysis.

- CQ1** *Which ML resources satisfy a combination of technical, legal, and contextual constraints?* This CQ captures a multi-criteria resource discovery scenario in which practitioners search for datasets and models by combining heterogeneous metadata dimensions such as task, modality, licensing conditions, language, size, libraries, or usage indicators. Such queries require the ability to semantically combine multiple metadata facets within a unified query framework.
- CQ2** *Which datasets, models, libraries, and publications constitute the ecosystem surrounding a given ML task?* This CQ focuses on the exploration of relationships between datasets, models, tasks, publications, libraries, and other metadata dimensions. It supports understanding how resources are organized around particular application domains or task families and enables navigation across interconnected resources.
- CQ3** *How are ML resources connected through training, derivation, and reuse relationships?* This CQ captures the dependencies and provenance relationships between ML resources, such as which datasets were used to train specific models, how datasets derive from other datasets, or how models evolve through fine-tuning and other transformation processes. Such information provides valuable insights into the origin, evolution, and downstream use of resources.
- CQ4** *What insights can be derived from the structure, metadata, and usage characteristics of a large Web repository?* This CQ captures repository-level analysis scenarios. For instance, identifying the most widely adopted resources, analyzing metadata coverage, studying the distribution of tasks and modalities, or exploring the prevalence of provenance relationships and licensing conditions. Such analyses provide insights into the structure, evolution, and quality of the repository.

3. The Datalens (DL) Semantic Model

Figure 1 illustrates the Datalens semantic model, which combines an OWL ontology (DLO) with a SKOS thesaurus (DLT) to provide an interoperable representation of ML resources and their relationships. Rather than describing the complete ML lifecycle, Datalens focuses on the metadata required for semantic discovery and exploration of datasets and models. The ontology builds upon established vocabularies, including DCAT for dataset description, PROV-O for provenance, Schema.org for software resources, and SKOS for controlled vocabularies. DLO introduces only the concepts and relationships required to represent ML-specific semantics that are absent from these vocabularies.

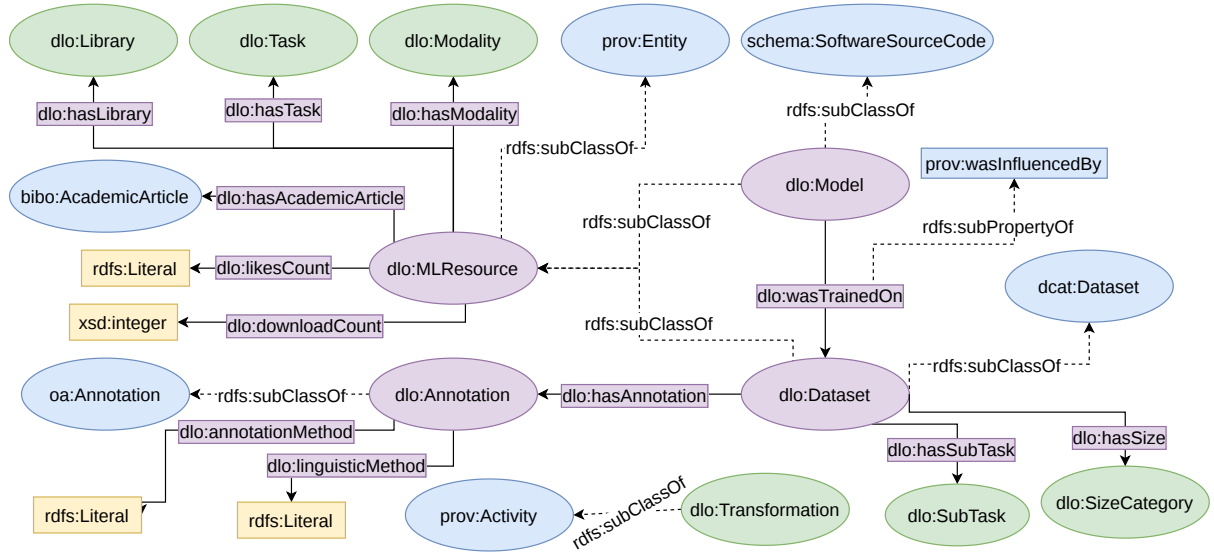


Figure 1: Overview of the Datalens Semantic Model. Green nodes represent DLO classes for which we defined a SKOS concept scheme gathering its instances. Blue nodes represent classes and properties from external vocabularies. Purple nodes represent other DLO classes and properties.

ML Resources Metadata Datasets are represented as instances of class `dcat:Dataset`, and models as instances of class `dlo:Model`, defined as a subclass of `schema:SoftwareSourceCode`. Both resource types share common discovery-oriented metadata, including tasks, modalities, software libraries, popularity indicators, and links to scientific publications. Controlled dimensions such as modalities, libraries, and size categories are represented through SKOS concept schemes to promote consistent annotation and semantic querying across repositories.

Provenance Relationships Building upon PROV-O, Datalens distinguishes two complementary relationships: `dlo:wasTrainedOn`, linking models to the datasets used for training, and `prov:wasDerivedFrom`, describing derivation relationships between datasets or models. Transformation activities are represented as instances of subclasses of `prov:Activity`, enabling explicit representation of operations such as dataset merging, filtering, augmentation, and model fine-tuning. These provenance links support semantic exploration of resource lineage and reuse.

Task Semantics Existing taxonomies do not explicitly model semantic relationships between high-level tasks and finer-grained sub-tasks, limiting semantic search and exploration. Thus, we constructed a SKOS thesaurus from Hugging Face task metadata, enriched with definitions from external knowledge bases, and organized through inferred and manually refined hierarchical relationships. The resulting thesaurus comprises 56 task concepts and 74 sub-task concepts linked through poly-hierarchical `skos:broader` relations. Within DLO, these concepts are instances of the classes `dlo:Task` and `dlo:SubTask`, which are associated with datasets and models through the properties `dlo:hasTask` and `dlo:hasSubTask`, enabling semantic task-oriented discovery beyond repository-specific labels.

4. DL-based KG for Hugging Face

By using the proposed DL semantic model, we built an RDF KG describing datasets and models from the Hugging Face repository. Metadata were collected using the platform’s public API, which resulting JSON contains metadata describing datasets and models, including identifiers, authorship information, timestamps, usage statistics, tags, tasks, modalities, licenses, and external references such as PapersWithCode identifiers. We pre-processed and normalized the metadata to align with the DL semantic model, which includes harmonizing heterogeneous metadata fields, extracting and restructuring task-related information. To ensure interoperability, known entities are aligned with

external standards: languages are mapped to Lexvo (ISO 639), regions to GeoNames, formats to IANA media types, and licenses to SPDX or Creative Commons URIs. Data lifting is done via a rule-based process using xR2RML [14]. The resulting KG comprises 62,148,769 RDF triples¹.

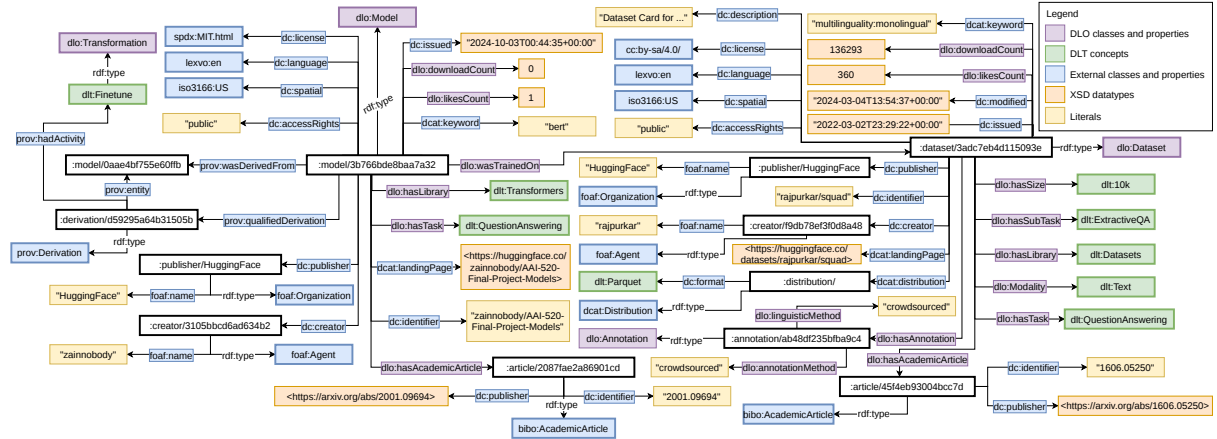


Figure 2: Overview of ML resources description and relationships as represented within the KG.

5. Evaluation and Exploration

Datalens is evaluated through expert validation of the proposed task thesaurus and through the SPARQL implementation of CQs demonstrating the discovery and exploration capabilities of the resulting KG.

Expert Validation. To assess the semantic validity of the inferred task hierarchy, we conducted an expert evaluation involving six IT researchers and developers with varying levels of ML expertise. Participants rated the relevance of each task–sub-task association using a 5-point Likert scale and were invited to suggest alternative associations. Overall, the proposed hierarchy was positively evaluated, with most associations considered relevant or highly relevant. Participants’ feedback also helped identify a small number of associations requiring refinement, leading to improvements of the final thesaurus.

Exploration Scenario. This scenario illustrates how Datalens supports the exploration of provenance relationships between ML resources, based on CQ3. Additional scenarios covering the remaining CQs are available through an interactive visualization dashboard at <https://wimmics.github.io/datalens/>.

Listing 1 presents a SPARQL query retrieving models together with their provenance links. Specifically, it exploits the `prov:wasDerivedFrom` property to capture derivation relationships between models and the `dlo:wasTrainedOn` property to identify the datasets used for model training. By combining these complementary relationships, the query supports the exploration of resource lineage, revealing how models evolve through derivation and how they are connected to the datasets on which they were trained. In this scenario, we focus on highly downloaded models, restricting the results to resources with at least 10,000 downloads. The query returned 931 provenance relationships matching these criteria.

Listing 1: SPARQL query to retrieve provenance relationships between models.

```
SELECT * WHERE {
  ?model a dlo:Model ; dc:identifier ?modelName ; dcat:landingPage ?modelUrl ;
  ?rel ?resource ; dlo:downloadCount ?downloads .

  VALUES ?rel { prov:wasDerivedFrom dlo:wasTrainedOn } # Relationships of interest

  ?rel rdfs:label ?relationship. ?resource dc:identifier ?resourceName .

  FILTER (?downloads > 10000) # Focus on highly downloaded models }
```

Figure 3 presents a graph-based visualization of the query results. Gray nodes represent the queried models, while colored nodes correspond to the resources on which they depend, namely training datasets (blue nodes) and parent models from which they derive (orange nodes). This visual distinction emphasizes the provenance relationships originating from each queried model, facilitating the exploration of

¹The KG is queryable via a SPARQL endpoint exposed at <https://ns.inria.fr/datalens/sparql>

both training dependencies and model derivation chains. The visualization further distinguishes the type of relationship through link colors: red links represent training relationships between models and datasets, while blue links indicate derivation relationships between models (Fig. 3B). The visualization reveals trends concerning training relationships (Fig. 3A), e.g., while few models were trained on multiple datasets, most models appear to be trained on a single dataset.

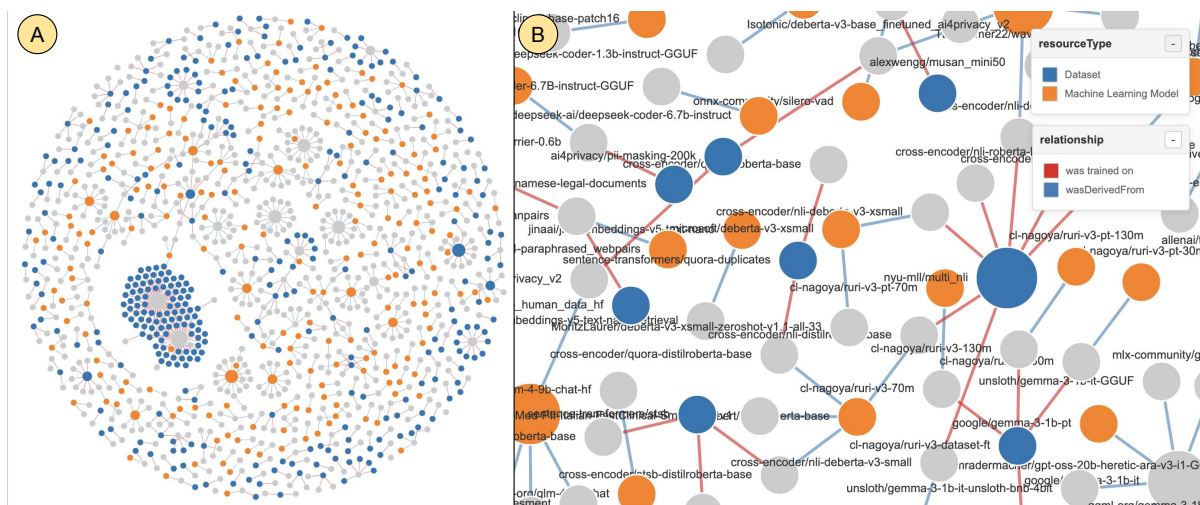


Figure 3: Visualization of CQ3 query results, showing the relationship between models and datasets in the KG. A closer look (right) reveals whether the relationship refers to training or derivation.

6. Discussion and Conclusions

Datalens supports the discovery and exploration of ML resources via a semantic approach combining an OWL ontology, a SKOS thesaurus, and an RDF KG. The proposed approach enables richer resource exploration than repository-specific keyword search by explicitly representing relationships between datasets, models, tasks, and provenance. The expert validation and exploration scenarios implementing the elicited CQs, illustrate the consistency of the proposed ontology and thesaurus, as well as the ability of the semantic model to support expressive discovery scenarios.

The current KG is instantiated from Hugging Face metadata and therefore inherits the incompleteness of the source repository, particularly regarding tasks, provenance information, licensing conditions, and links to scientific publications. Future work will focus on enriching metadata from external sources and extending the KG to additional repositories to improve coverage and interoperability. We also plan to evaluate the proposed approach through user studies investigating the usefulness of semantic exploration for ML resource discovery.

For the sake of reproducibility and reuse by the community, all resources and source code used in this paper are available at <https://github.com/Wimmics/datalens>.

Acknowledgments

We sincerely thank the volunteers who participated in our competency question elicitation and expert validation processes. We acknowledge the IDEX *Initiative d'Excellence* of Université Côte d'Azur for funding Zachari Arnaud's internship.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT for: grammar, spelling check and sentence rephrasing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] A. Chapman, E. Simperl, L. Koesten, G. Konstantinidis, L. Ibáñez, E. Kacprzak, P. Groth, Dataset search: a survey, *The VLDB Journal* (2020). doi:10.1007/s00778-019-00564-x.
- [2] N. W. Paton, J. Chen, Z. Wu, Dataset Discovery and Exploration: A Survey, *ACM Computing Surveys* (2023). doi:10.1145/3626521.
- [3] E. Dimitrakis, K. Sgontzos, Y. Tzitzikas, A survey on question answering systems over linked data and documents, *J. Intell. Inf. Syst.* 55 (2020) 233–259. URL: <https://doi.org/10.1007/s10844-019-00584-7>. doi:10.1007/s10844-019-00584-7.
- [4] M. Mountantonakis, Y. Tzitzikas, Content-based union and complement metrics for dataset search over RDF knowledge graphs, *ACM J. Data Inf. Qual.* 12 (2020) 10:1–10:31. URL: <https://doi.org/10.1145/3372750>. doi:10.1145/3372750.
- [5] S. Castelo, R. Rampin, A. S. R. Santos, A. Bessa, F. Chirigati, J. Freire, Auctus: A dataset search engine for data discovery and augmentation, *Proc. VLDB Endow.* 14 (2021) 2791–2794. URL: <http://www.vldb.org/pvldb/vol14/p2791-castelo.pdf>. doi:10.14778/3476311.3476346.
- [6] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, A. Hanna, Data and its (dis)contents: A survey of dataset development and use in machine learning research, *Patterns* 2 (2021) 100336. URL: <https://doi.org/10.1016/j.patter.2021.100336>. doi:10.1016/J.PATTER.2021.100336.
- [7] W3C, Data catalog vocabulary (dcat) - version 3, <https://www.w3.org/TR/vocab-dcat-3/>, 2024. W3C Recommendation.
- [8] W3C, Prov-overview: An overview of the prov family of documents, <https://www.w3.org/TR/prov-overview/>, 2013. W3C Recommendation.
- [9] W3C Machine Learning Schema Community Group, ML-schema documentation, <https://ml-schema.github.io/documentation/MLSchema.html>, 2016. Accessed: 2026-05-05.
- [10] I. Dasoulas, D. Yang, A. Dimou, MLSea: A Semantic Layer for Discoverable Machine Learning, in: *The Semantic Web*, 2024.
- [11] Q. Chen, K. Huang, X. Zhou, W. Luo, Y. Cui, G. Cheng, Benchmarking Recommendation, Classification, and Tracing Based on Hugging Face Knowledge Graph, in: *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025. doi:10.1145/3726302.3730277.
- [12] M. Liu, C. Zhao, X. Peng, S. Yu, H. Wang, C. Sha, Task-Oriented ML/DL Library Recommendation Based on a Knowledge Graph, *IEEE Transactions on Software Engineering* (2023). doi:10.1109/TSE.2023.3285280.
- [13] M. Färber, D. Lamprecht, Linked Papers With Code: The Latest in Machine Learning as an RDF Knowledge Graph, 2023. doi:10.48550/arXiv.2310.20475, arXiv:2310.20475 [cs.DL].
- [14] F. Priyatna, F. Michel, O. Corcho, frmichel/morph-xr2rml: 1.3.1, 2021. doi:10.5281/zenodo.5148233, software.