Linked Papers With Code: The Latest in Machine Learning as an RDF Knowledge Graph

Michael Färber*, David Lamprecht

Karlsruhe Institute of Technology (KIT), Institute AIFB, Germany

Abstract

In this paper, we introduce *Linked Papers With Code* (LPWC), an RDF knowledge graph that provides comprehensive, current information about almost 400,000 machine learning publications. This includes the tasks addressed, the datasets utilized, the methods implemented, and the evaluations conducted, along with their results. Compared to its non-RDF-based counterpart *Papers With Code*, LPWC not only translates the latest advancements in machine learning into RDF format, but also enables novel ways for scientific impact quantification and scholarly key content recommendation. LPWC is openly accessible at https://linkedpaperswithcode.com and is licensed under CC-BY-SA 4.0. As a knowledge graph in the Linked Open Data cloud, we offer LPWC in multiple formats, from RDF dump files to a SPARQL endpoint for direct web queries, as well as a data source with resolvable URIs and links to the data sources SemOpenAlex, Wikidata, and DBLP. Additionally, we supply knowledge graph embeddings, enabling LPWC to be readily applied in machine learning applications.

Keywords

Scholarly Data, Open Science, Ontology Engineering, Machine Learning

1. Introduction

Over the years, several scientific knowledge graphs have emerged, including ORKG [1], MAKG [2], and most recently, SemOpenAlex [3]. However, no knowledge graph exists that explicitly targets the modeling of key content in machine learning on a large scale and always up-to-date. On the other hand, *Papers with Code* (PWC, https://paperswithcode.com) has emerged as a platform for machine learning publications, code, datasets, methods, and evaluation tables that can be updated by anyone. However, PWC is only available as a web page and a JSON dump without semantic modeling and Linked Open Data (LOD) integration.

We introduce *Linked Papers With Code* (LPWC) – the first RDF knowledge graph that comprehensively models the research field of machine learning using an extensive ontology. Our knowledge graph goes beyond simply lifting to RDF format, for example, by resolving complex data formats through graph modeling and linking entities to other LOD sources such as SemOpenAlex, Wikidata, and DBLP. LPWC consists of 7,935,279 RDF triples as of June 2023 and is available at https://linkedpaperswithcode.com. It has multiple applications, from improved management of research data to more effective integration of data across different research

ISWC 2023 Posters and Demos: 22nd International Semantic Web Conference, November 6–10, 2023, Athens, Greece

*Corresponding author.

michael.faerber@kit.edu (M. Färber); david.lamprecht@student.kit.edu (D. Lamprecht)

0000-0001-5458-8645 (M. Färber); 0000-0002-9098-5389 (D. Lamprecht)

CEUR Workshop Proceedings (CEUR-WS.org)

domains. By incorporating FAIR principles that focus on the availability and reuse of research data and artifacts, we expect LPWC to improve the discoverability and applicability of machine learning research results. We make the code used for knowledge graph creation and embedding generation available online (https://github.com/davidlamprecht/linkedpaperswithcode). In the following, we present LPWC in detail.

2. Linked Papers with Code

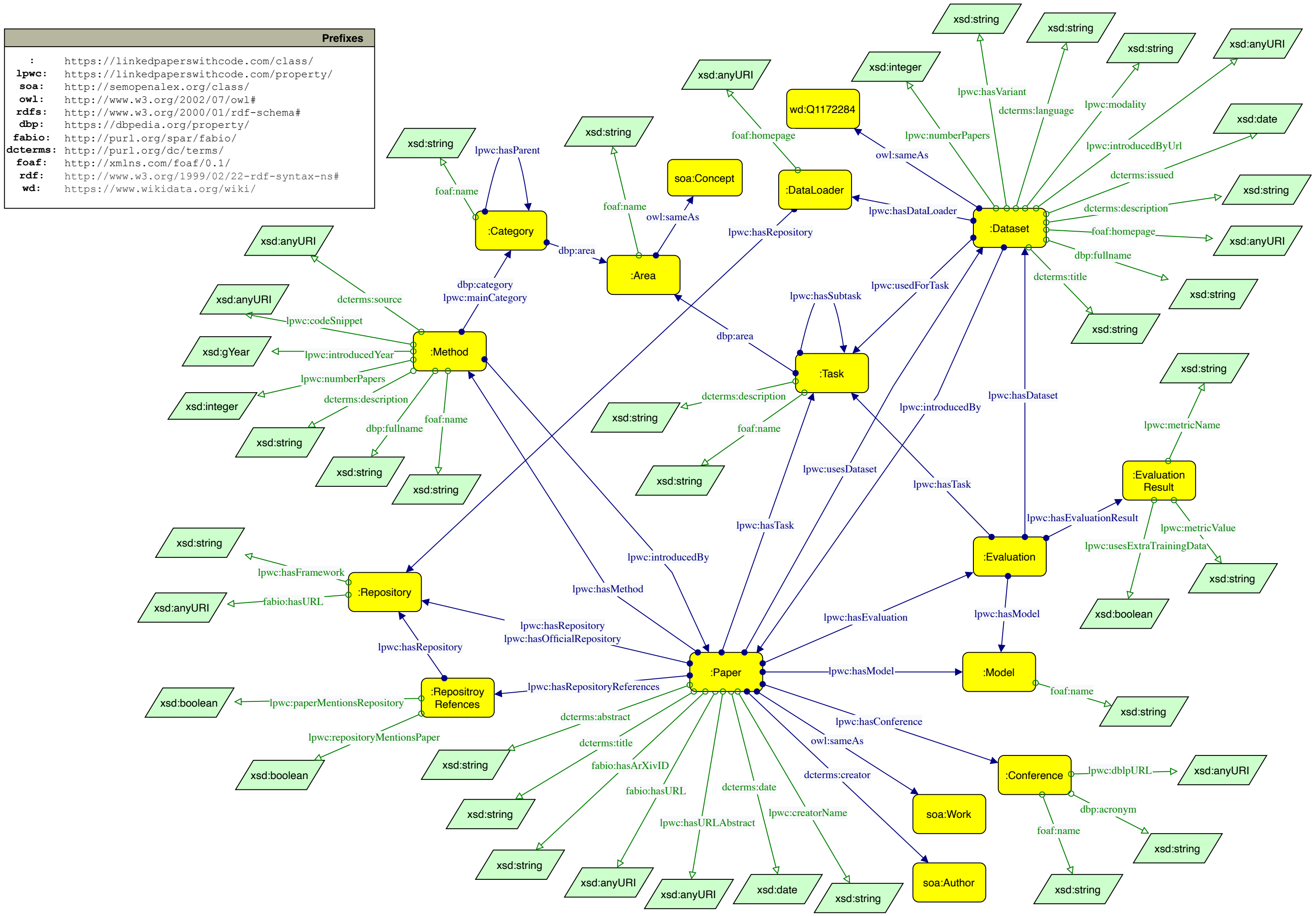
Linked Papers With Code Ontology. First, we develop an ontology that adheres to the best practices of ontology engineering and incorporates as much existing vocabulary as possible. Given that the PWC data dump is sourced directly from the PWC website, thus lacks a standardized schema and comprises diverse JSON objects, it was infeasible to directly model it within an OWL/RDF framework. Consequently, we construct a novel semantic schema to model the data. An overview of the entity types, object properties, and data type properties can be found in Figure 1. The LPWC ontology encompasses *13 entity types* and *47 relationship types.* In addition to the ontology, which is available as an OWL file, we provide a VoID file, following the Linked Open Data good practices to describe our linked dataset.

Linked Papers With Code Knowledge Graph. PWC provides access to its data via a user-friendly, human-readable website. In addition, it offers daily JSON data dumps.[1] However, there are several aspects that currently make using the data difficult: 1. There is a lack of semantic interoperability. Entities, such as authors or AI models, are represented as strings without unique IDs. This prevents effective linking of data and creation of knowledge graphs. 2. Due to the complexity of the data, modeling in JSON format proves difficult, especially when processing or querying the data. This issue becomes particularly apparent with evaluation tables, which are nested within a JSON structure with up to 19 levels in depth. This results in significant data redundancy within the file. In contrast, a graph representation provides a more intuitive and manageable way of modeling. 3. The data, originally designed for a human readable interface, uses markdown for natural language descriptions of entities, which may not be optimal when being processed by NLP methods or displaying it outside of the website.

Data Transformation. To overcome these limitations, we convert the JSON files from the PWC data dump into an RDF knowledge graph based on the developed ontology. This requires major changes in the data formatting and data modeling. In the transformation process we, among other steps, (1) assign unique HTTP URIs to all entities, (2) convert all markdown test to plain text and (3) link the entities to other scholarly data sources in the LOD cloud.

Author Name Disambiguation. The disambiguation of author names given as strings is a crucial step on top of the pure data transformation. Specifically, we develop an efficient two-step method to link the 1,471,006 authors in LPWC to entities in SemOpenAlex, which is a massive RDF dataset modeling the academic landscape with its publications, authors, sources, and institutions, via its public SPARQL endpoint [3]. We leverage LPWC author names and paper titles for the disambiguation. The first step involves exact name matching and publication title substring comparison.

[1] See https://github.com/paperswithcode/paperswithcode-data



Prefixes	
:	https://linkedpaperswithcode.com/class/
lpwc:	https://linkedpaperswithcode.com/property/
soa:	http://semopenalex.org/class/
owl:	http://www.w3.org/2002/07/owl#
rdfs:	http://www.w3.org/2000/01/rdf-schema#
dbp:	https://dbpedia.org/property/
fabio:	http://purl.org/spar/fabio/
dcterms:	http://purl.org/dc/terms/
foaf:	http://xmlns.com/foaf/0.1/
rdf:	http://www.w3.org/1999/02/22-rdf-syntax-ns#
wd:	https://www.wikidata.org/wiki/

Figure 1: *Schema of Linked Papers With Code.*

Entity Type	# Instances
Paper	376,557
Evaluation	52,519
Paper with Evaluations	13,289
Repository	153,476
Model	24,598
Dataset	8,322
Task	4,267
Method	2,101
Conference	1,407

Table 1: Linked Papers With Code entity types and number of instances (as of June 2023).

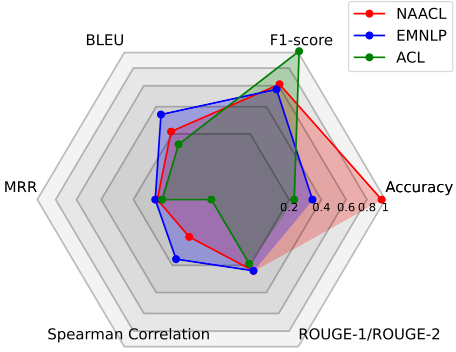


Figure 2: Distribution of evaluation metrics used in NAACL, EMNLP, and ACL conferences.

If no match is found, the second step employs a variant search of LPWC paper titles in SemOpenAlex works, and author matching based on fuzzy similarity techniques. This process yields 947,709 links to SemOpenAlex entities. The remaining 523,297 author names are represented in LPWC using the `lpwc:authorName` property.

Creating owl:sameAs statements. We further link all conferences modeled in LPWC to DBLP. Moreover, we successfully map 267,314 papers (71% of all papers in LPWC) to SemOpenAlex works, utilizing variations of the LPWC paper titles. Lastly, we are able to create 158 mappings (2% of all datasets) between datasets modeled in LPWC and datasets modeled in Wikidata.

Key Statistics. Our knowledge graph's SPARQL endpoint enables the direct computation of interesting statistics. For instance, Table 1 shows the frequency of entities across entity types. Additionally, Figure 2 illustrates how to compare conferences (here: NAACL, EMNLP, ACL) based on the used evaluation metrics of their papers.

Knowledge Graph Embeddings. To enable additional use cases, we compute knowledge graph embeddings for LPWC. Embeddings have proven to be valuable as implicit knowledge representations in various scenarios. We train the embeddings based on state-of-the-art embedding techniques such as TransE, DistMult, ComplEx, and RotatE [4, 5]. The training process involves a maximum of 900 epochs, implementing early stopping based on the mean rank calculated on the validation sets at intervals of 300 epochs. Among the evaluated techniques, TransE shows the best results. Therefore, we provide the TransE-based embedding vectors for all entities and relations online and all our evaluation results in our repository. Notably, our provided embeddings are in line with state-of-the-art results on benchmark datasets with similar characteristics in terms of the number of relations, triples, and entities [4, 5].

Use Case Examples. LPWC can enhance existing use cases while also enabling the development of new ones. In the following, we highlight some potential use cases:

1. *Machine Learning Data Analysis:* LPWC is a novel scientific knowledge graph covering the current field of machine learning. Complex analyses, such as comparing conferences or detecting new research topics, become possible in this way.
2. *Scholarly LOD Cloud Enrichment:* LPWC is highly integrated with the LOD cloud and connected to multiple data sources such as SemOpenAlex, Wikidata, and DBLP. This enables efficient data integration and enhanced research data management in alignment

with the FAIR principles.

3. *Academic Recommender Systems:* Given the information overload in science, scientific recommender systems are becoming increasingly important. LPWC and the provide knowledge graph embeddings can be used directly to build state-of-the-art recommender systems for key scientific content. With LPWC, these systems can recommend also items such as datasets, methods, and conferences.

3. Conclusion

In this paper, we presented *Linked Papers with Code*, the first RDF knowledge graph with detailed information about the machine learning landscape, consisting of close to 8 million RDF triples. We outlined the creation process of this dataset, discussed its characteristics, and examined the procedure for training state-of-the-art knowledge graph embeddings. In future work, we aim to leverage the extensive interconnectivity between LPWC and SemOpenAlex to facilitate large-scale key content extraction from publications.

References

[1] S. Auer, A. Oelen, M. Haris, M. Stocker, J. D'Souza, K. E. Farfar, L. Vogt, M. Prinz, V. Wiens, M. Y. Jaradeh, Improving Access to Scientific Literature with Knowledge Graphs, Bibliothek Forschung und Praxis 44 (2020) 516–529.

[2] M. Färber, The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data, in: Proceedings of the 18th International Semantic Web Conference, ISWC'19, 2019, pp. 113–129.

[3] M. Färber, D. Lamprecht, J. Krause, L. Aung, P. Haase, SemOpenAlex: The Scientific Landscape in 26 Billion RDF Triples, in: Proceedings of the 22nd International Semantic Web Conference, ISWC'23, 2023.

[4] R. Wang, B. Li, S. Hu, W. Du, M. Zhang, Knowledge Graph Embedding via Graph Attenuated Attention Networks, IEEE access 8 (2019) 5212–5224.

[5] C. Demir, A.-C. N. Ngomo, Convolutional Complex Knowledge Graph Embeddings, in: Proceedings of the 18th Extended Semantic Web Conference, ESWC'21, 2021, pp. 409–424.