Karlsruhe Institute of Technology

Michael Färber
michael.faerber@kit.edu

David Lamprecht
david.lamprecht@student.kit.edu

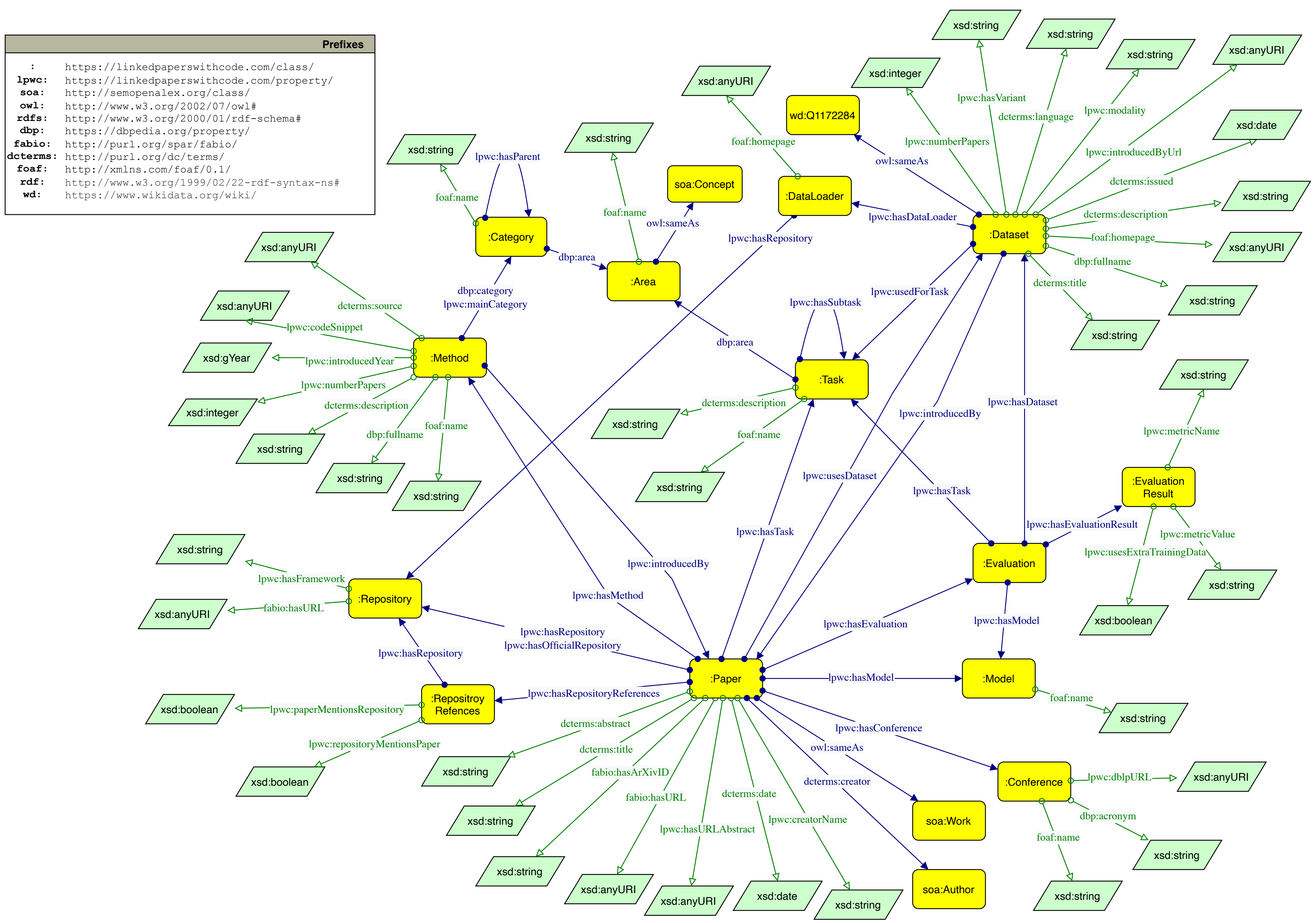
Linked Papers With Code: The Latest in Machine Learning as an RDF Knowledge Graph

Summary

- We transform Paper With Code (PWC, https://paperswithcode.com) into an RDF Knowledge Graph reusing existing vocabularies.
- We disambiguate authors and link resources to SemOpenAlex, Wikidata, and DBLP.
- We provide the data of Linked Papers With Code (LPWC) at **https://linkedpaperswithcode.com**
  - as data dumps (in N-Triples and Turtle format),
  - as resolvable data source in the LOD cloud, and
  - in a triple store with a public SPARQL endpoint.
- We evaluate and provide Knowledge Graph Embeddings for LPWC entities and relations.

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

Linked Papers With Code Knowledge Graph



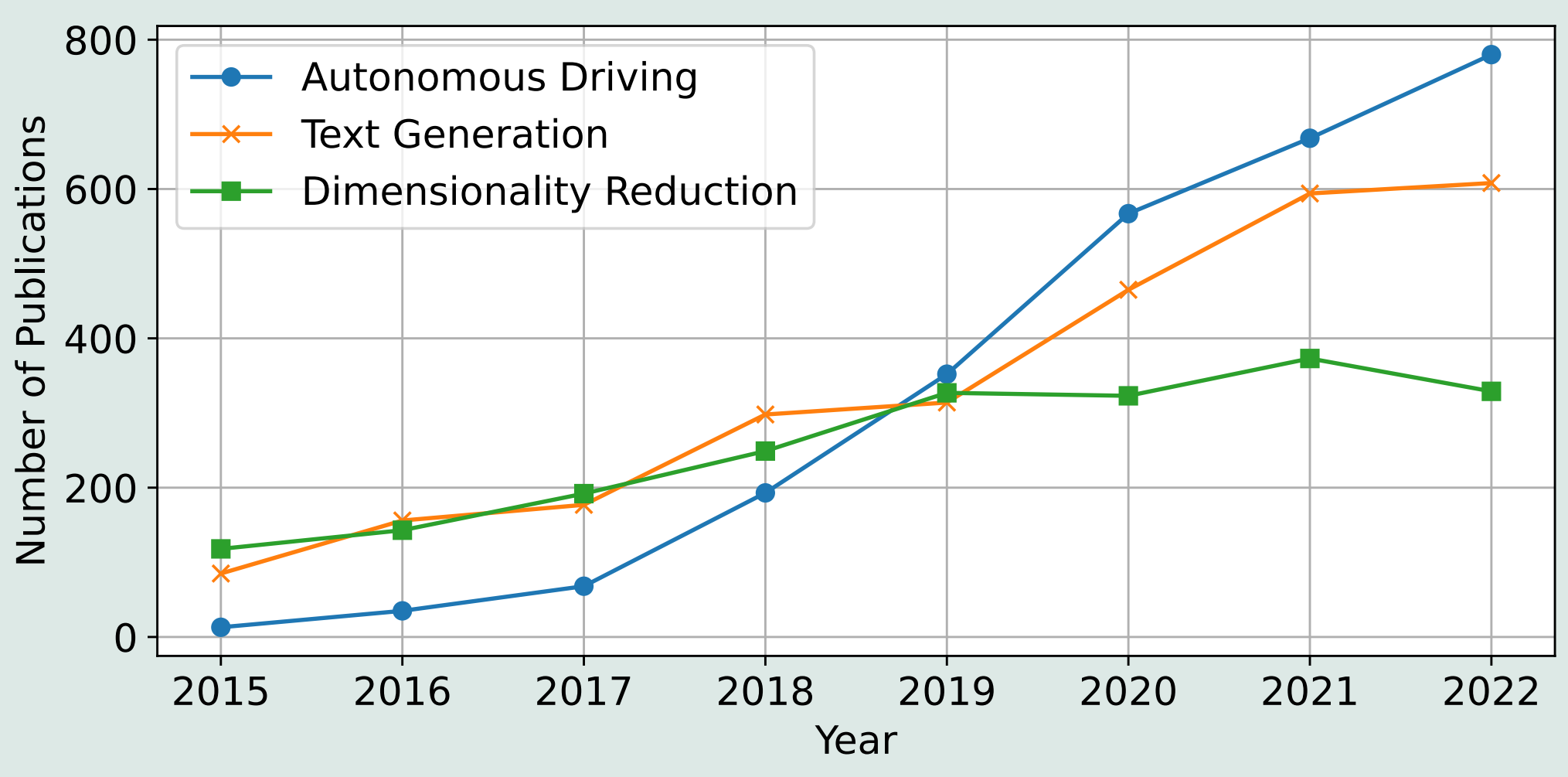
Prefixes

:	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/

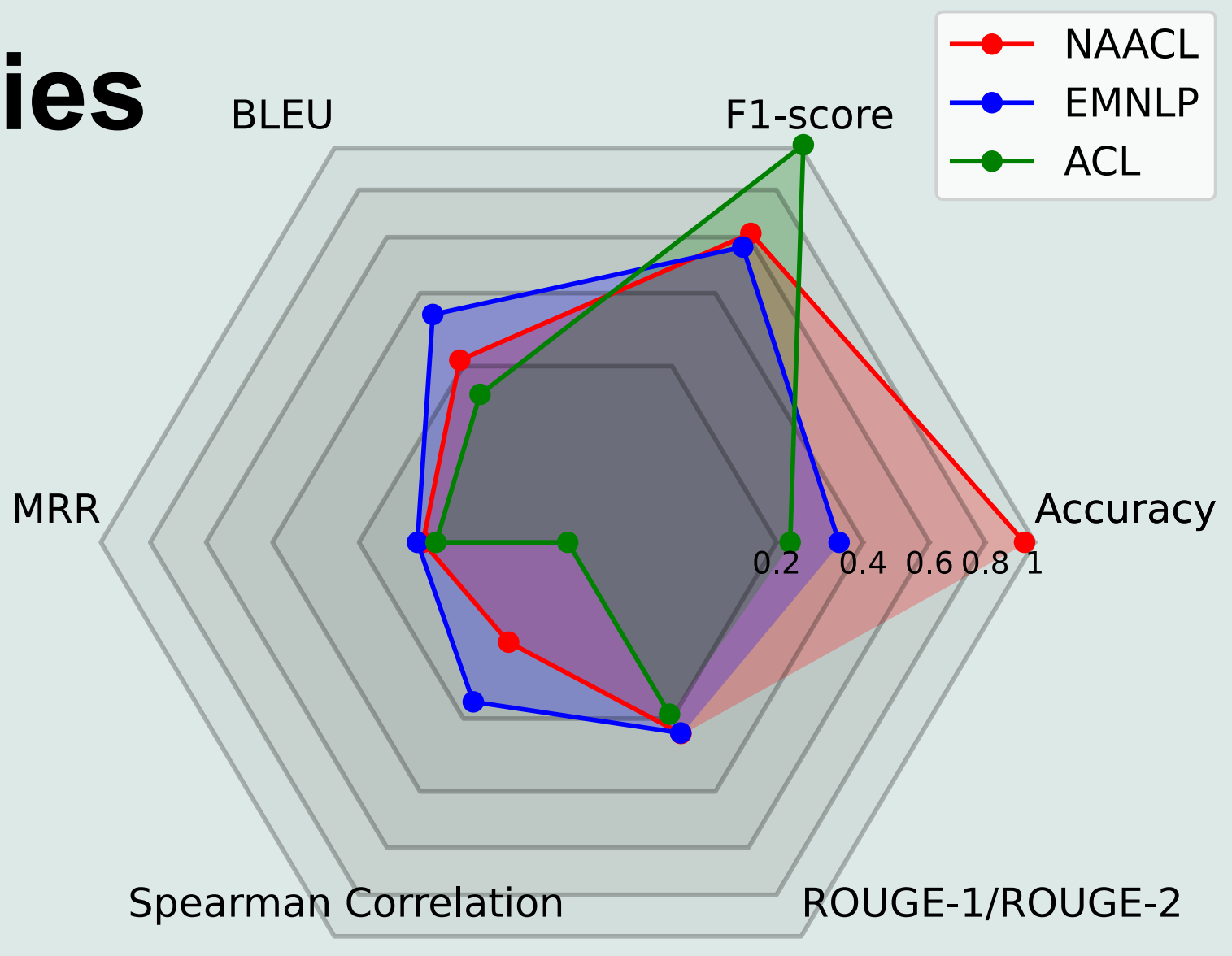
Schema of Linked Papers With Code

- ***Semantic Modeling:*** LPWC models the research field of machine learning in 7,935,279 semantic RDF triples using an extensive ontology which encompasses 13 entity types and 47 relationships.
- ***Transformation:*** To ensure semantic interoperability, the transformation process of the JSON files from the PWC dump to an RDF Knowledge Graph involves:
  1. assigning unique HTTP URIs to all entities,
  2. converting all markdown text to plain text,
  3. linking the entities to other scientific data sources in the LOD cloud, and
  4. an efficient two-step author name disambiguation leveraging PWC author names and paper titles to link the 1,471,006 authors in LPWC to entities in SemOpenAlex.
- ***Creating owl:sameAs statements:*** We link:
  1. all conferences modeled in LPWC to DBLP,
  2. 267,317 papers (71% of all papers in LPWC) to SemOpenAlex works, and
  3. 158 datasets modeled in LPWC to datasets modeled in Wikidata.

Statistics and Example SPARQL Queries



Time-based analyses of research tasks in order to identify research trends.



Comparison of the conferences NAACL, EMNLP and ACL based on the distribution of used evaluation metrics.

```
PREFIX lpwc: <https://linkedpaperswithcode.com/property/>

SELECT ?framework (COUNT(?framework) AS ?count)
WHERE {
?repository a <https://linkedpaperswithcode.com/class/repository>.
?repository lpwc:hasFramework ?framework.
}
GROUP BY ?framework
ORDER BY DESC(?count)
LIMIT 2
```

Number of repositories that use PyTorch and TensorFlow as a framework.

Framework	Count
PyTorch	61,551
TensorFlow	22,503

Example Use Cases

1. Machine Learning Data Analysis
   - E.g., comparing conferences or detecting new research topics.
2. Scholarly LOD Cloud Enrichment
   - Research data management in line with the FAIR principles.
3. Academic Recommender Systems
   - LPWC and the provided embeddings can be used to build state-of- the-art recommender systems for key scientific content.

Knowledge Graph Embeddings

- Data pre-processing for embedding generation leads to a dataset with 1,454,103 triples, 527,817 entities and 15 relations.

Metric	TransE	DistMult	ComplEx	RotatE
MR	**2239.26**	9448.88	25,624.13	8830.03
Hits@1	**0.2395**	0.1931	0.1655	0.1146
Hits@3	**0.3851**	0.3204	0.2814	0.1921
Hits@10	**0.5425**	0.4856	0.4390	0.3133