SemOpenAlex: The Scientific Landscape in 26 Billion RDF Triples

Michael Färber[1(✉)], David Lamprecht[1], Johan Krause[1], Linn Aung[2], and Peter Haase[2]

[1] Institute AIFB, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
michael.faerber@kit.edu, {david.lamprecht,johan.krause}@student.kit.edu
[2] metaphacts GmbH, Walldorf, Germany
{la,ph}@metaphacts.com

Abstract. We present *SemOpenAlex*, an extensive RDF knowledge graph that contains over 26 billion triples about scientific publications and their associated entities, such as authors, institutions, journals, and concepts. SemOpenAlex is licensed under CC0, providing free and open access to the data. We offer the data through multiple channels, including RDF dump files, a SPARQL endpoint, and as a data source in the Linked Open Data cloud, complete with resolvable URIs and links to other data sources. Moreover, we provide embeddings for knowledge graph entities using high-performance computing. SemOpenAlex enables a broad range of use-case scenarios, such as exploratory semantic search via our website, large-scale scientific impact quantification, and other forms of scholarly big data analytics within and across scientific disciplines. Additionally, it enables academic recommender systems, such as recommending collaborators, publications, and venues, including explainability capabilities. Finally, SemOpenAlex can serve for RDF query optimization benchmarks, creating scholarly knowledge-guided language models, and as a hub for semantic scientific publishing.

Data and Services:	https://semopenalex.org https://w3id.org/SemOpenAlex
Code:	https://github.com/metaphacts/semopenalex/
Data License:	Creative Commons Zero (CC0)
Code License:	MIT License

Keywords: Scholarly Data · Open Science · Digital Libraries

1 Introduction

With the increasing number of scientific publications, staying up-to-date with current research presents a significant challenge. For instance, in 2022 alone, more than 8 million scientific publications were registered [1]. To explore related scholarly entities such as authors and institutions, researchers rely on a range of methods from search interfaces to recommendation systems [2,3]. One effective way to model the underlying scholarly data is to represent it as an RDF

T. R. Payne et al. (Eds.): ISWC 2023, LNCS 14266, pp. 94–112, 2023.
https://doi.org/10.1007/978-3-031-47243-5_6

knowledge graph (KG). Doing so facilitates standardization, visualization, and interlinking with Linked Data resources [4]. Consequently, scholarly KGs play a pivotal role in transforming document-centric scholarly data into interconnected and machine-actionable knowledge structures [2].

However, available scholarly KGs have one or several of the following limitations. Firstly, they rarely contain an exhaustive catalog of publications across all disciplines [5]. Secondly, they often cover only certain disciplines, such as computer science [6]. Thirdly, they are not regularly updated, rendering many analyses and business models obsolete [7]. Fourthly, they often contain usage restrictions [8]. Lastly, even if they fulfill these requirements, they are not available according to W3C standards such as RDF [1,9]. These issues hinder the application of scientific KGs on a broad scale, such as in comprehensive search and recommender systems, or for scientific impact quantification. For instance, the Microsoft Academic Graph was discontinued in 2021 [10], which hinders further updates to its derivative in RDF, the Microsoft Academic Knowledge Graph (MAKG) [7]. This leaves a gap that the novel OpenAlex dataset aims to fill [1]. However, the data in OpenAlex is not available in RDF and does not comply with Linked Data Principles [11]. Consequently, OpenAlex cannot be considered a KG, which makes semantic queries, integration into existing applications, or linking to additional resources non-trivial. At first glance, integrating scholarly data about scientific papers into Wikidata and thus contributing to the WikiCite initiative may seem like an obvious solution. However, apart from the dedicated schema, the volume of the data is already so large that the Blazegraph triplestore which is used in the Wikidata Query Service reaches its capacity limit, preventing any integration [12] (see Sect. 2).

In this paper, we introduce *SemOpenAlex*, an extremely large RDF dataset of the academic landscape with its publications, authors, sources, institutions, concepts, and publishers. SemOpenAlex consists of more than 26 billion semantic triples and includes over 249 million publications from all academic disciplines. It is based on our rich ontology (see Sect. 3.1) and includes links to other LOD sources such as Wikidata, Wikipedia, and the MAKG. To ensure easy and efficient use of SemOpenAlex's integration with the LOD cloud, we provide a public SPARQL endpoint. In addition, we provide a sophisticated semantic search interface that allows users to retrieve real-time information about contained entities and their semantic relationships (e.g., displaying co-authors or an author's top concepts – information, which is not directly contained in the database but obtained through semantic reasoning). We also provide the full RDF data snapshots to enable big data analysis. Due to the large size of SemOpenAlex and the ever-increasing number of scientific publications being integrated into SemOpenAlex, we have established a pipeline using AWS for regularly updating SemOpenAlex entirely without any service interruptions. Additionally, to use SemOpenAlex in downstream applications, we trained state-of-the-art knowledge graph entity embeddings. By reusing existing ontologies whenever possible, we ensure system interoperability in accordance with FAIR principles [13] and pave the way for the integration of SemOpenAlex into the Linked Open

Data Cloud. We fill the gap left by the discontinuation of MAKG by providing monthly updates that facilitate ongoing monitoring of an author's scientific impact, tracking of award-winning research, and other use cases using our data [14,15]. By making SemOpenAlex free and unrestricted, we empower research communities across all disciplines to use the data it contains and integrate it into their projects. Initial use cases and production systems that use SemOpenAlex already exist (see Sect. 5).

Overall, we make the following contributions:

1. We create an *ontology* for SemOpenAlex reusing common vocabularies.
2. We create the SemOpenAlex *knowledge graph* in RDF, covering 26 billion triples, and provide all *SemOpenAlex* data, code, and services for public access at https://semopenalex.org/:
   (a) We provide monthly updated RDF data snapshots free of charge on AWS S3 at `s3://semopenalex` (via browser: https://semopenalex.s3.amazonaws.com/browse.html), accepted as AWS Open Data project.[1]
   (b) We make all URIs of SemOpenAlex resolvable, allowing SemOpenAlex to be part of the Linked Open Data cloud.[2]
   (c) We index all data in a triple store and make it publicly available via a SPARQL endpoint (https://semopenalex.org/sparql).
   (d) We provide a semantic search interface including entity disambiguation to access, search, and visualize the knowledge graph and its statistical key figures in real time.
3. We provide state-of-the-art knowledge graph embeddings for the entities represented in SemOpenAlex using high-performance computing.

In the following, we first discuss related work (see Sect. 2) and describe the SemOpenAlex ontology and RDF data (see Sect. 3), before presenting the SemOpenAlex entity embeddings (see Sect. 4). Subsequently, we outline existing and potential use cases (see Sect. 5), before we conclude the paper (see Sect. 6).

2 Related Work

A comparison of scholarly RDF datasets is presented in Table 1. It is obvious from the table that SemOpenAlex (1) is the only RDF KG that follows the Linked Data Principles, (2) is fully open, (3) contains a vast amount of bibliographic information from all scientific disciplines, and (4) is regularly updated, making it a valuable resource in various contexts (see Sect. 5).

The OpenAIRE Research Graph provides open and free access to metadata of 145 million publications, datasets, and software via an API, a SPARQL endpoint,

[1] The AWS Open Data Sponsorship program covers the cost of storing and retrieving all SemOpenAlex data, ensuring the long-term sustainability of our project. Upon request, it was confirmed that Zenodo does not support the provision of SemOpenAlex data due to its size.

[2] See, e.g. `curl -H "Accept:text/n3"` https://semopenalex.org/work/W4239696231.

Table 1. Statistical comparison of scholarly RDF datasets.

	OpenAIRE	AceKG	Wikidata	COCI	MAKG	SemOpenAlex
# Works	145M	62M	42M	76M	239M	249M
# Triples	1.4B	3.13B	–	1.4B	8B	26.4B
# References	0	480M	288M	1.4B	1.4B	1.7B
Snapshot size	100 GB	113 GB	120 GB	1.5 TB	1.4 TB	1.7 TB
Regular updates			(✓)	✓		✓
SPARQL endpoint			✓	✓	✓	✓
Entity embeddings					✓	✓

OpenAIRE as of March 2021, AceKG as of 2018, Wikidata as of Dec. 2022, the OpenCitations Index of Crossref open DOI-to-DOI citations (COCI) as of Oct. 2022, the MAKG as of March 2021, and SemOpenAlex as of March 2023

and database dumps [16]. However, not only is the number of publications significantly lower than in SemOpenAlex but on May 8, 2023, OpenAIRE stopped its LOD services and closed the SPARQL endpoint.[3]

WikiCite[4] has incorporated bibliographic metadata into Wikidata, but SemOpenAlex covers considerably more metadata (e.g., 249M papers vs. 42M), including additional properties such as papers' abstracts. While using Wikidata as a central KG and regularly importing SemOpenAlex information seems logical, the scalability of the Blazegraph triplestore backend which hosts the Wikidata Query Service is limited, and Wikimedia has announced a plan to delete scholarly articles in case of bulk imports.[5]

AceKG [17] is a database containing 62 million publications, along with academic details related to authors, fields of study, venues, and institutes. AceKG data is modeled in RDF. However, unlike our approach, it does not use existing vocabularies, lacks a publicly available triple store, and does not offer continuous updates. All data is sourced from a company's database.

OpenCitations focuses on publications and their citation relationships [18]. Specifically, it covers metadata about publications and their citations, but not descriptions of affiliated organizations (institutions) or hosting conferences and journals (venues). OpenCitations includes several datasets, including the OpenCitations Index of Crossref Open DOI-to-DOI Citations (COCI) with 76 million items to date, and smaller datasets such as the OpenCitations Corpus (OCC) and OpenCitations in Context Corpus (CCC) [5].

The Microsoft Academic Knowledge Graph (MAKG) is based on the Microsoft Academic Graph (MAG), containing information on publications, authors, institutions, venues, and concepts [7,19]. The MAKG has high coverage

[3] See https://www.openaire.eu/pausing-our-lod-services.
[4] See http://wikicite.org/.
[5] See https://m.wikidata.org/wiki/Wikidata:SPARQL_query_service/WDQS_backend_update/Blazegraph_failure_playbook.

Table 2. SemOpenAlex entity types and number of instances (as of March 2023).

Entity Type	# Instances
Work	249,450,604
Author	135,360,159
Source	226,413
Institution	108,618
Concept	65,073
Publisher	7,017

across scientific domains and has enabled novel use cases. However, it will no longer be updated due to lack of source data [10]. Several analyses have assessed the MAG and MAKG, revealing the need for improvements in areas such as citation accuracy, concept assignment, and disambiguation [20–23]. Compared to MAKG, SemOpenAlex provides a similar schema, but provides fresh data that is in addition cleaned by an author name disambiguation provided by OpenAlex and a neater mapping of concepts to papers using the Simple Knowledge Organization System (SKOS) ontology [24,25].

Further notable scholarly KGs are the DBLP KG[6] and the Open Research Knowledge Graph (ORKG) [26]. DBLP provides only high-quality metadata about computer science publications, resulting in a coverage of roughly 6 million publications [6]. ORKG is a project that aims to provide a KG infrastructure for semantically capturing and representing the content of research papers [2, 27]. ORKG contains a relatively small set of more than 25,000 publications, however, with many RDF statements, indicating considerable semantic richness. Due to their different focuses, SemOpenAlex can complement ORKG as an LOD data source: while SemOpenAlex provides a broad basis of metadata about a massive amount of publications and related entities in RDF (with a focus on high coverage, see Table 2), ORKG focuses on modeling scientific contributions as well as methodology aspects, which are manually curated (with a focus on high data quality and key insights of papers).

3 SemOpenAlex

In the following, we describe the design of the SemOpenAlex ontology (Sect. 3.1) and the process of generating SemOpenAlex data (Sect. 3.2). We also explain how we publish and enable user interaction with the data (Sect. 3.3), and present key statistics of the KG (Sect. 3.4). Furthermore, we evaluate to what extent SemOpenAlex meets linked data set descriptions and rankings (Sect. 3.5).

[6] See https://www.dagstuhl.de/en/institute/news/2022/dblp-in-rdf.

3.1 Ontology of SemOpenAlex

We developed an ontology following the best practices of ontology engineering reusing as much existing vocabulary as possible. An overview of the entity types, the object properties, and the data type properties is provided in Fig. 1. Overall, the ontology of SemOpenAlex covers *13 entity types*, including the main entity types *works, authors, institutions, sources, publishers and concepts*, as well as *87 relation types*.

Table 3. Used ontologies, their corresponding prefixes and namespace.

Ontology	Prefix	Associated URI
SemOpenAlex	:	https://semopenalex.org/class/
SemOpenAlex	soa:	https://semopenalex.org/property/
OpenAlex	oa:	http://openalex.org/
XML Schema	xsd:	http://www.w3.org/2001/XMLSchema#
OWL	owl:	http://www.w3.org/2002/07/owl#
RDF	rdf:	http://www.w3.org/1999/02/22-rdf-syntax-ns#
RDF Schema	rdfs:	http://www.w3.org/2000/01/rdf-schema#
Dubin Core	dcterms:	http://purl.org/dc/terms/
CiTO	cito:	http://purl.org/spar/cito/
FaBiO	fabio:	http://purl.org/spar/fabio/
BiDO	bido:	http://purl.org/spar/bido/
DataCite	datacite:	http://purl.org/spar/datacite
PRISM	prism:	http://prismstandard.org/namespaces/basic/2.0/
DBpedia	dbo:	https://dbpedia.org/ontology/
DBpedia	dbp:	https://dbpedia.org/property/
FOAF	foaf:	http://xmlns.com/foaf/0.1/
W3 ORG	org:	http://www.w3.org/ns/org#
GeoNames	gn:	https://www.geonames.org/ontology#
SKOS	skos:	http://www.w3.org/2004/02/skos/core#

We reused the vocabularies listed in Table 3. To describe publications, researchers, and institutions, we leveraged established Semantic Publishing and Referencing (SPAR) ontologies [28], such as FaBiO and CiTO. FaBiO is used to describe specific identifiers such as a work's PubMedID, while CiTO represents citing relationships between works. For bibliographic metadata, such as a work's publication date and abstract, we used the Dublin Core ontology (DCterms). To represent more generic features and relations, we relied on cross-domain

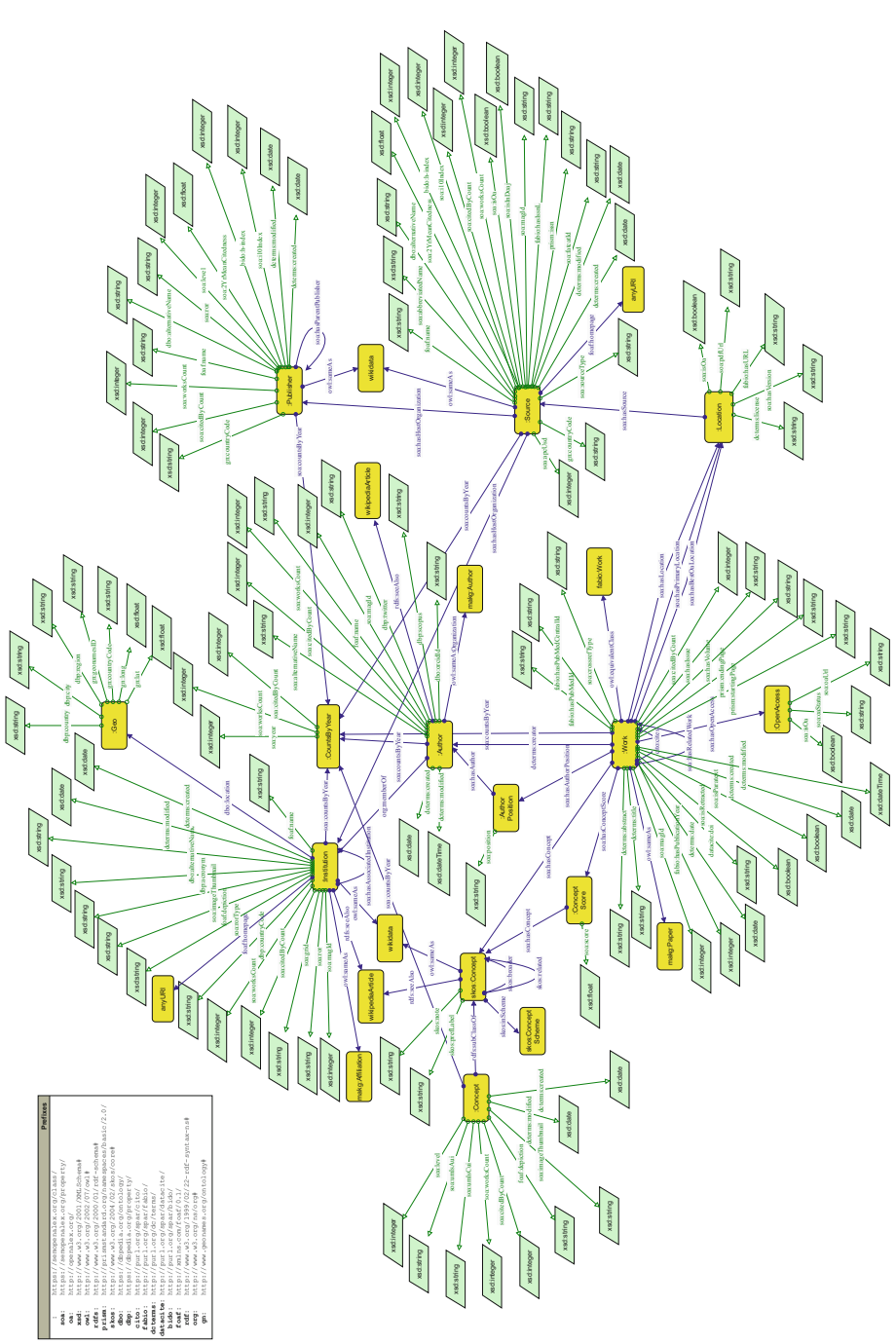

Fig. 1. Ontology of SemOpenAlex.

ontologies such as DBpedia and the W3 Organization Ontology (W3 ORG). The works are classified using a concept hierarchy, which we represented in a SKOS vocabulary of 65k SKOS concepts and semantic relations (`skos:broader` and `skos:related`). The concepts are further linked with Wikidata entities, allowing for additional interoperability and providing multi-lingual labels.

3.2 Knowledge Graph Creation Process

The raw OpenAlex data was presumably designed for data processing (e.g., abstracts are provided as inverted index and not provided as one string). To create an RDF KG based on the OpenAlex dump files, major changes in the data formatting and the data modeling are necessary. In the following, we outline the essential steps of this transformation process.

Transformation. We carry out a number of distinct steps for the transformation that can be reproduced via the code in our GitHub repository.[7]

1. *Data Preprocessing:* We download the OpenAlex snapshot in compressed `.jsonl` format from its AWS S3 bucket and use the Python multiprocessing package for efficient parallel processing of the large amount of data. To ensure valid triple generation according to the *W3C RDF 1.1 Concepts and Abstract Syntax*[8] later, we remove problematic characters from literal values, such as non-escaped backslashes in URLs or newlines in publication titles. Additionally, we convert the abstracts, which are included in OpenAlex as an inverted index, to plain text to improve accessibility.
2. *RDF Generation:* We transform the preprocessed data from JSON into RDF according to the ontology shown in Fig. 1. For the generation of the triples, we draw on the rdflib Python package,[9] which offers functionality to handle, process and validate RDF data. During triple serialization, we create a buffer subgraph that is written once a fixed number of statements is reached to reduce the number of I/O operations. In total, we generate 26,401,183,867 RDF triples given the data snapshot as of 2023-03-28.
3. *Compression and Deployment:* The RDF data generated for SemOpenAlex takes up 1.7 TB in the TriG format[10] when uncompressed. To make the data more manageable, we compress it into .gz archives, resulting in a reduction of over 80% in file size to 232 GB. These compressed files are then imported into the GraphDB triple store and made available for download as an open snapshot. Additionally, we provide a data sample on GitHub.

[7] See https://github.com/metaphacts/semopenalex.
[8] See https://www.w3.org/TR/rdf11-concepts/#section-Graph-Literal.
[9] See https://github.com/RDFLib/rdflib/.
[10] TriG is an extension of Turtle, extended to support representing a complete RDF dataset (see https://www.w3.org/TR/trig/).

Update Mechanism. To ensure that SemOpenAlex remains up-to-date, we perform the transformation process described earlier on a monthly basis, which involves downloading the latest OpenAlex snapshot. This enables us to observe temporal dynamics in the data, and ensures that SemOpenAlex provides the most recent information available. The updated version of the data is available through all three access points (RDF dump, SPARQL endpoint, and visual interface). The update process is semi-automated and takes approximately five days to complete on an external server instance. We use one AWS instance to provide SemOpenAlex services and one instance to process the next SemOpenAlex release. Changes to SemOpenAlex data resulting from changes in the raw OpenAlex files are tracked using announcements via the OpenAlex mailing list. Several adaptations have been performed in this way in the past.

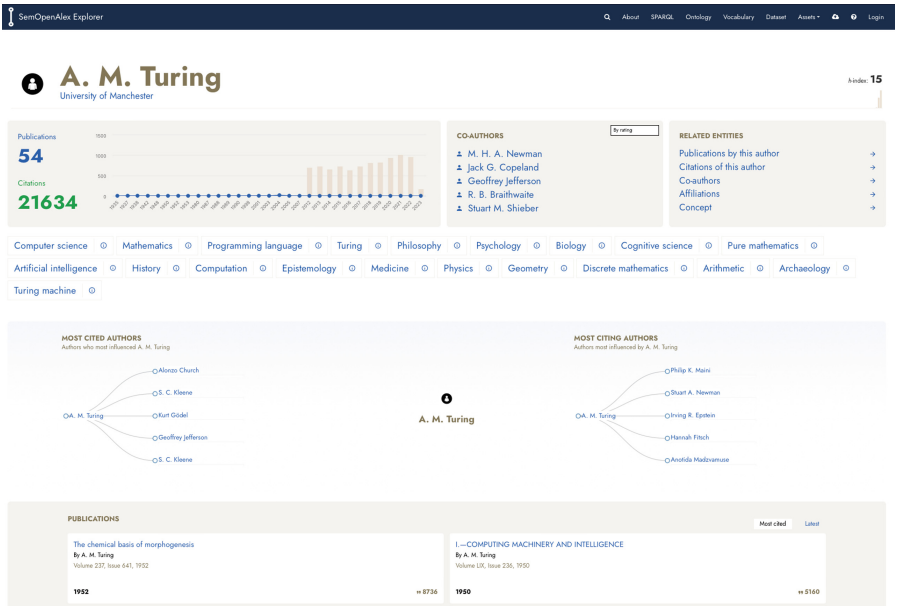

Fig. 2. Author overview page for A.M. Turing, accessible at https://semopenalex.org/author/A2430569270.

3.3 Data Publishing and User Interaction

Our KG is publicly accessible at https://semopenalex.org/. We utilize the metaphactory knowledge graph platform [29] on top of a GraphDB triple store to deploy the KG. metaphactory serves as a Linked Data publication platform and ensures that the URIs of SemOpenAlex are fully resolvable. The data is published in machine-readable RDF formats as well as human-readable HTML-based templates using content negotiation. Figure 2 displays the page for the URI https://semopenalex.org/author/A2430569270.

Among other features, the interface provided for SemOpenAlex enables users to: (1) access SemOpenAlex through a search interface with filtering options; (2) visualize arbitrarily large sub-graphs for objects and relations of interest; (3) formulate and execute SPARQL queries to retrieve objects from the graph using a provided SPARQL endpoint; (4) examine the ontology of SemOpenAlex; (5) obtain key statistics for each object in SemOpenAlex in a dashboard, as shown in the screenshot in Fig. 2; (6) assess the underlying multi-level concept hierarchy; and (7) interact with further linked entities such as co-authors or concepts and access external resources such as links to Wikidata.

3.4 Key Statistics of SemOpenAlex and Example SPARQL Queries

In this subsection, we present several statistics that we generated based on queries using our SPARQL endpoint. We provide the queries on GitHub.

Figure 3 shows the number of papers published in the field of machine learning and natural language processing by researchers from Karlsruhe Institute of Technology from 2000 to 2021. While the number of machine learning papers

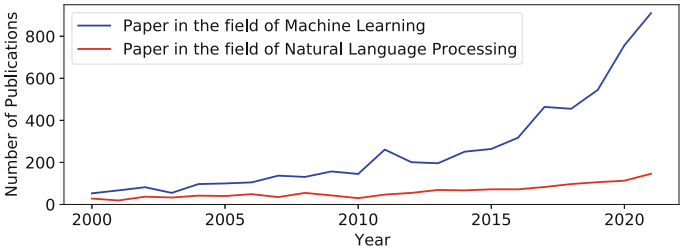

Fig. 3. Number of publications published in machine learning and natural language processing by researchers from Karlsruhe Institute of Technology.

Table 4. Number of institution for the countries with the most institutions.

Country	# Institutions
US	32,814
GB	7,743
DE	5,096
CN	4,856
JP	4,031
FR	3,965
IN	3,731
CA	3,498

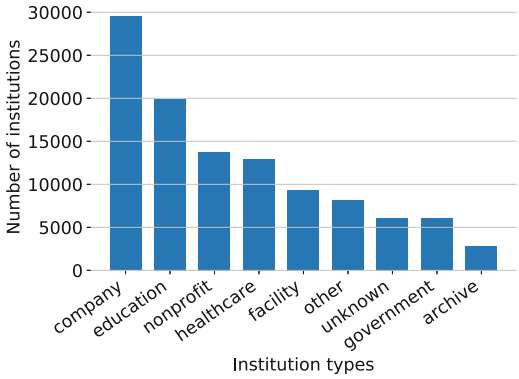

Fig. 4. Distribution of institution types.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX soa: <https://semopenalex.org/property/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

SELECT DISTINCT ?paperTitle ?citedByCount ?firstAuthorName
WHERE {
  ?paper dcterms:title ?paperTitle .
  ?paper soa:hasConcept ?Concept .
  ?Concept skos:prefLabel "Semantic␣Web"^^xsd:string .
  ?paper soa:citedByCount ?citedByCount .
  ?paper soa:hasAuthorPosition ?authorPosition .
  ?authorPosition soa:position "first"^^xsd:string .
  ?authorPosition soa:hasAuthor ?firstAuthor .
  ?firstAuthor foaf:name ?firstAuthorName .
}
ORDER BY DESC(?citedByCount)
LIMIT 100
```

List. 1. Querying the top 100 most cited papers with the concept "Semantic Web" as well as their citation count and first author.

received a sharp increase from 2015, the number of papers in the field of natural language processing increased at a rather constant rate. SemOpenAlex enables institutions to create such relevant key figures and trends in the context of strategic controlling in a simple and cost-free way.

SemOpenAlex covers the worldwide scientific landscape and contains publications from institutions around the globe. In total, institutions from 225 different countries are included. The 8 countries with the highest number of institutions are shown in Table 4.

By distinguishing between eight types of institutions, SemOpenAlex enables differentiated data analyses. In Fig. 4, we can see the distribution of the 108,618 unique institutions across the different types. We can see that the majority of the organizations are companies, followed by educational and nonprofit institutions.

Listing 1 shows an example of how SemOpenAlex can be queried with SPARQL. This query retrieves the top 100 most cited papers in the field of semantic web, along with their citation counts and first authors. It is worth noting that this query cannot be executed on other scholarly KGs like the MAKG, as they do not cover information about the author's position for a given paper.

3.5 Linked Data Set Descriptions and Ratings

Following the licensing model of the underlying OpenAlex data,[11] we provide all SemOpenAlex data under the *CC0 license*, which grants users the right to freely build upon, enhance, and reuse the works for any purpose without restriction, paving the way for other researchers and software engineers to build upon

[11] See https://openalex.org/about.

SemOpenAlex in any context. The RDF data files are available for unrestricted and free download as they are hosted with the AWS Open Data program.[12]

We can categorize SemOpenAlex according to the two kinds of 5-star rating schemes in the Linked Data context:

- *Tim Berners-Lee's 5-star deployment scheme for Open Data*[13]: Our SemOpenAlex RDF dataset is a 5-star data set according to this scheme, because we provide our data in RDF (leading to 4 stars) and the (1) entity URIs are linked to Wikidata, Wikipedia and the MAKG and (2) our vocabulary URIs to other vocabularies (leading to 5 stars).
- *Linked Data vocabulary star rating* [30]: This rating is intended to rate the use of vocabulary within Linked (Open) Data. By providing a turtle file, by linking our vocabulary to other vocabularies (see the SPAR ontologies), we are able to provide the vocabulary with 4 stars.

Aside from the SemOpenAlex RDF documents, we provide the following linked data set descriptions (all available at https://semopenalex.org/):

- *Turtle:* We provide our ontology as a Turtle file describing the used classes, object properties, and data type properties.
- *VoID:* We provide a VoID file to describe our linked data set with an RDF schema vocabulary.

4 Graph Embeddings for SemOpenAlex

Apart from creating and providing the SemOpenAlex data set and services (e.g., the SPARQL endpoint), we computed embeddings for all SemOpenAlex entities. Entity embeddings have proven to be useful as implicit knowledge representations in a variety of scenarios, as we describe in Sect. 5. Based on the SemOpenAlex data in RDF, we trained entity embeddings based on several state-of-the-art embedding techniques and compared the performance of the respective results with regard to link prediction tasks. Specifically, we applied the following approaches: TransE [31], DistMult [32], ComplEx [33], a GraphSAGE neural network [34], and a graph attention network [35]. To address the nontrivial challenges associated with training on SemOpenAlex as a very large knowledge graph, we employed the Marius framework [36]. Marius[14] is designed to optimize resource utilization by pipelining hard disk, CPU, and GPU memory during training, thereby reducing idle times. In our evaluation, we opted for a configuration of 100 embedding dimensions, a batch size of 16,000, and trained for 3 epochs on a high-performance computing system (bwUniCluster 2.0) using Python 3.7, Marius 0.0.2, PyTorch 1.9.1, and CUDA 11.2.2. These parameters are in line with previous research on large-scale entity embeddings [24].

[12] See https://aws.amazon.com/opendata/open-data-sponsorship-program/.
[13] See http://5stardata.info/.
[14] See https://marius-project.org.

The computational effort required for the different embedding techniques varied, with the GraphSAGE and the graph attention network approaches requiring the most memory. These methods used up to 716 GB of CPU RAM and took the longest time to train, with each epoch taking roughly 24 h. Despite the resource-intensive nature of the GraphSAGE and graph attention network approaches, DistMult yielded the highest mean reciprocal rank (MRR) score in our link prediction evaluation (see all evaluation results on GitHub). Therefore, we provide the DistMult-based embedding vectors for all entities online.[15]

5 Use Cases of SemOpenAlex

Scholarly KGs have proven to be a valuable data basis for various use cases and scenarios, such as analyzing research dynamics between academia and industry [37], scientific impact quantification [14,38], and linking research data sets to publications [39]. This is also reflected in the high number of citations of the reference publications of the MAG [40] and MAKG [7].[16] In the following, we focus on existing and potential use cases of SemOpenAlex.

Scholarly Big Data Analytics and Large-Scale Scientific Impact Quantification. SemOpenAlex can serve for scientific impact quantification and innovation management. For instance, OpenAlex has been utilized as a comprehensive and reliable data source to rank researchers and institutions worldwide on `research.com`.[17] InnoGraph is a new project that leverages OpenAlex to represent innovation ecosystems as a KG for innovation management and forecasting [41]. By using SemOpenAlex as underlying database for such projects and efforts, the need to deal with cumbersome data integration issues can be reduced. Currently, universities such as KIT rely on paid scholarly services like those from Springer Nature for measuring their performance and ranking as a university [42]. However, in the future, these institutions can use SemOpenAlex as a free database to run analytics and evaluations on all relevant publications and associated entities.

Scholarly Search and Recommender Systems. Recommendation systems – both content-based and collaborative filtering-based – have become increasingly important in academia to help scientists navigate the overwhelming amount of available information resulting from the exponential increase in the number of publications. In this paper, we provide entity embeddings for nearly all existing entities in the scientific landscape, which can be used directly to build state-of-the-art recommender systems. These systems can recommend items such as papers to read and cite, as well as venues and collaborators [43]. SemOpenAlex

[15] See https://doi.org/10.5281/zenodo.7912776.

[16] Sinha et al. [40] have obtained 1,041 citations and Färber [7] has obtained 115 citations as of April 28, 2023, according to Google Scholar.

[17] See https://research.com/university/materials-science/humboldt-university-of-berlin.

can be utilized to make these recommendations explainable, as symbolic information from the KG can be shown to the user. Due to SemOpenAlex's rich ontology, including various entity types, SemOpenAlex can serve as a realistic dataset for training and evaluating state-of-the-art graph neural networks designed for heterogeneous information networks and with a specific focus on scalability and semantics. Moreover, our rich KG can be utilized to provide recommendations in complex scenarios, such as finding the optimal consortium for large, possibly interdisciplinary research projects. In the context of semantic search, SemOpenAlex can be used for entity linking, annotating scientific texts [44] or tables [45] for enhanced search capabilities.

Semantic Scientific Publishing. SemOpenAlex is a part of the Linked Open Data Cloud and contains links to other data sources such as Wikidata, Wikipedia, and MAKG. As a result, it significantly contributes to the use of linked data in areas such as digital libraries and information retrieval [46]. SemOpenAlex has a unique selling point among available scientific knowledge graphs, with its coverage of publications worldwide and across all scientific disciplines, totaling around 250 million publications (see Table 2), and its regular updates. SemOpenAlex can serve as a central catalog for publications, researchers, and research artifacts, to which other data repositories and KGs can link. This creates an opportunity to use SemOpenAlex as a basis for modeling scientific artifacts, such as datasets, scientific methods, and AI models, and thus beyond SemOpenAlex' current scope. This information may be modeled in separate, interlinked KGs or as part of SemOpenAlex in the future. For instance, the Data Set Knowledge Graph [39], which currently links 600,000 publications in which datasets are mentioned to the MAKG, can now link datasets to papers in SemOpenAlex. Similarly, semantic representations of datasets and scientific methods [47], as well as representations of scientific facts and claims mentioned in full-text articles [48], can be linked to publications and authors in SemOpenAlex to provide rich context information as explanations of academic recommender systems. Furthermore, links between SemOpenAlex and KGs modeling AI models and their energy consumption, such as the Green AI Knowledge Graph [49], can be used to combine previously isolated data for performing complex analytics. In this way, questions of strategic controlling, such as "How green are the AI models developed at my institution?" [49], can be automatically answered. Finally, it makes sense to link full-text paper collections to SemOpenAlex, for instance, to leverage its concept schema, since SemOpenAlex applies concept tags to all its papers published globally and across all scientific fields. An excellent example of an existing paper collection linked to SemOpenAlex is unarXive 2022 [50], sourced from two million arXiv papers.

Research Project Management and Modeling. KGs have become increasingly important in supporting research projects by providing a structured representation of various research entities and their relationships [51]. These project-specific KGs encapsulate a diverse range of research entities, such as topics, methods, tasks, materials, organizations, researchers, their skills, interests, and activities, as well as research outputs and project outcomes. To facilitate the

development and support of KGs for research projects, SemOpenAlex serves as a knowledge hub by providing existing data on project participants and relevant research. Researchers can use tools and vocabularies provided by the Competency Management Ontology [52] to seamlessly describe their skills, current research interests, and activities in terms of the entities already contained in SemOpenAlex. Moreover, SemOpenAlex's concept hierarchy allows for the construction of ontologies for specific research domains, streamlining research tasks such as performing a state-of-the-art analysis for a research area. Existing resources from SemOpenAlex can be integrated into KG-based project bibliographies, enhancing collaboration between researchers through resource sharing.

SemOpenAlex has already been used to provide a comprehensive and structured overview of research projects. In particular, personalized dashboards have been created by metaphacts that display recently added publications from SemOpenAlex that are relevant to the current research context. Newly created resources within a project, such as research papers and datasets, can also be described and linked to SemOpenAlex. Ultimately, published results become a valuable part of SemOpenAlex.

Groundwork for Scientific Publishing in the Future. One can envision that the working style of researchers will considerably change in the next few decades [53,54]. For instance, publications might not be published in PDF format any more, but in either an annotated version of it (with information about the claims, the used methods, the data sets, the evaluation results, and so on) or in the form of a flexible publication form, in which authors can change the content and, in particular, citations, over time. SemOpenAlex can be easily combined with new such data sets due to its structure in RDF. Furthermore, ORKG is an ongoing effort that targets the semantic representation of papers and their scientific contributions. We argue that SemOpenAlex can be used as data basis for ORKG in the sense that with SemOpenAlex, users do not need to take care of first creating papers and authors in the ORKG, but to directly import or link the corresponding information from SemOpenAlex, which has its focus on being a comprehensive KG covering all scientific publications worldwide.

Knowledge-Guided Language Models. Large language models, including ChatGPT and GPT-4, have been criticized for their lack of explainability and their failure to provide reliable in-text citations to reference literature. Often, when citations are provided, they are incorrect and reflect "hallucinations". In this context, SemOpenAlex represents a valuable repository for guiding language models in providing reliable references to scientific literature and as a basis for text-editing generative models. With metadata of 250 million scientific works, SemOpenAlex can serve as a valuable resource for source attribution and improving the accuracy and quality of scientific writing generated by these models.

Benchmarking. SemOpenAlex is a prime example of big data, fulfilling the "4 V's" criteria: it is very large, with a wide variety of information types (including papers, authors, institutions, venues, and various data formats), contains uncertainties, and is updated periodically. This makes it suitable for bench-

marking systems and approaches, particularly in the context of querying large, realistic KGs [55]. In fact, the MAKG has already been used for this purpose [56] and we expect SemOpenAlex to follow suit.

6 Conclusions

In this paper, we presented a comprehensive RDF dataset with over 26 billion triples covering scholarly data across all scientific disciplines. We outlined the creation process of this dataset and discussed its characteristics. Our dataset supports complex analyses through SPARQL querying. By making the SPARQL endpoint publicly available and the URIs resolvable, we enriched the Linked Open Data cloud with a valuable source of information in the field of academic publishing. We offer RDF dumps, linked dataset descriptions, a SPARQL endpoint, and trained entity embeddings online at https://semopenalex.org/. In the future, we plan to incorporate metadata about funding programs to enable in-depth and comprehensive evaluations of funding lines of governments and institutions [51,57,58].

Acknowledgments. This work was partially supported by the German Federal Ministry of Education and Research (BMBF) as part of the project IIDI (01IS21026D). The authors acknowledge support by the state of Baden-Württemberg through bwHPC.

References

1. Priem, J., Piwowar, H., Orr, R.: OpenAlex: a fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv preprint arXiv:2205.01833 (2022)
2. Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M., Vidal, M.E.: Towards a knowledge graph for science. In: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics. WIMS'18, June 2018, pp. 1–6 (2018)
3. Christensen, A.: Wissenschaftliche Literatur entdecken: Was bibliothekarische Discovery-Systeme von der Konkurrenz lernen und was sie ihr zeigen können. LIBREAS, Library Ideas (2022)
4. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., et al.: Knowledge graphs. Synth. Lect. Data Semant. Knowl. **12**(2), 1–257 (2021)
5. Peroni, S., Shotton, D.: OpenCitations, an infrastructure organization for open scholarship. Quant. Sci. Stud. **1**(1), 428–444 (2020)
6. Aleman-Meza, B., Hakimpour, F., Budak Arpinar, I., Sheth, A.P.: SwetoDblp ontology of Computer Science publications. J. Web Semant. **5**(3), 151–155 (2007)
7. Färber, M.: 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, pp. 113–129 (2019)
8. Waltman, L., Larivière, V.: Special issue on bibliographic data sources. Quant. Sci. Stud. **1**(1), 360–362 (2020)
9. Manghi, P., Mannocci, A., Osborne, F., Sacharidis, D., Salatino, A., Vergoulis, T.: New trends in scientific knowledge graphs and research impact assessment. Quant. Sci. Stud. **2**(4), 1296–1300 (2021)

10. Microsoft Research: Next Steps for Microsoft Academic - Expanding into New Horizons, May 2021. https://www.microsoft.com/en-us/research/project/academic/articles/microsoft-academic-to-expand-horizons-with-community-driven-approach/
11. Berners-Lee, T.: Linked Data - Design Issues, July 2006. https://www.w3.org/DesignIssues/LinkedData.html
12. WDQS Search Team: WDQS Backend Alternatives Working Paper (2022). Version 1.1, 29 March 2022. Wikimedia Foundation, San Franciscio, CA, USA. https://www.wikidata.org/wiki/File:WDQS_Backend_Alternatives_working_paper.pdf
13. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., et al.: The FAIR guiding principles for scientific data management and stewardship. Sci. Data **3**(1) (2016)
14. Huang, Y., Lu, W., Liu, J., Cheng, Q., Bu, Y.: Towards transdisciplinary impact of scientific publications: a longitudinal, comprehensive, and large-scale analysis on Microsoft Academic Graph. Inf. Process. Manag. **59**(2) (2022)
15. Wagner, C.S., Horlings, E., Whetsell, T.A., Mattsson, P., Nordqvist, K.: Do nobel laureates create prize-winning networks? An analysis of collaborative research in physiology or medicine. PLOS ONE **10**(7) (2015)
16. Manghi, P., et al.: OpenAIRE Research Graph Dump (June 2022) Version Number: 4.1 https://doi.org/10.5281/zenodo.6616871
17. Wang, R., et al.: AceKG: a large-scale knowledge graph for academic data mining. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. CIKM'18, pp. 1487–1490 (2018)
18. Peroni, S., Dutton, A., Gray, T., Shotton, D.: Setting our bibliographic references free: towards open citation data. J. Doc. **71**(2), 253–277 (2015)
19. Sinha, A., et al.: An overview of microsoft academic service (MAS) and applications. In: Proceedings of the 24th International Conference on World Wide Web, Florence Italy, ACM, pp. 243–246, May 2015
20. Herrmannova, D., Knoth, P.: An analysis of the Microsoft Academic graph. D-Lib Mag. **22**(9/10) (2016)
21. Visser, M., van Eck, N.J., Waltman, L.: Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. Quant. Sci. Stud. **2**(1), 20–41 (2021)
22. Chen, C.: A glimpse of the first eight months of the COVID-19 literature on Microsoft Academic graph: themes, citation contexts, and uncertainties. Front. Res. Metrics Anal. **5** (2020)
23. Wang, K., Shen, Z., Huang, C., Wu, C.H., Dong, Y., Kanakia, A.: Microsoft Academic Graph: when experts are not enough. Quant. Sci. Stud. **1**(1), 396–413 (2020)
24. Färber, M., Ao, L.: The Microsoft Academic knowledge graph enhanced: author name disambiguation, publication classification, and embeddings. Quant. Sci. Stud. **3**(1), 51–98 (2022)
25. Tay, A., Martín-Martín, A., Hug, S.E.: Goodbye, Microsoft Academic - hello, open research infrastructure? May 2021. https://blogs.lse.ac.uk/impactofsocialsciences/2021/05/27/goodbye-microsoft-academic-hello-open-research-infrastructure/
26. Auer, S., et al.: Improving access to scientific literature with knowledge graphs. Bibliothek Forschung und Praxis **44**(3), 516–529 (2020)
27. Jaradeh, M.Y., et al.: Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge. In: Proceedings of the 10th International Conference on Knowledge Capture. K-CAP'19, Marina Del Rey, CA, USA, pp. 243–246 (2019)
28. Peroni, S., Shotton, D.: The SPAR ontologies. In: Proceedings of the 17th International Semantic Web Conference. ISWC'18, pp. 119–136 (2018)

29. Haase, P., Herzig, D.M., Kozlov, A., Nikolov, A., Trame, J.: Metaphactory: a platform for knowledge graph management. Semant. Web **10**(6), 1109–1125 (2019)
30. Janowicz, K., Hitzler, P., Adams, B., Kolas, D., Vardeman, C.: Five stars of linked data vocabulary use. Semant. Web **5**(3), 173–176 (2014)
31. Bordes, A., Usunier, N., Garcia-Durán, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. NIPS'13, Red Hook, NY, USA, pp. 2787–2795. Curran Associates Inc. (2013)
32. Yang, B., Yih, W.t., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575 (2014)
33. Trouillon, T., Welbl, J., Riedel, S., Gaussier, E., Bouchard, G.: Complex Embeddings for Simple Link Prediction, June 2016
34. Hamilton, W.L., Ying, R., Leskovec, J.: Inductive representation learning on large graphs, September 2018. arXiv:1706.02216
35. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
36. Waleffe, R., Mohoney, J., Rekatsinas, T., Venkataraman, S.: MariusGNN: resource-efficient out-of-core training of graph neural networks (2022)
37. Angioni, S., Salatino, A., Osborne, F., Recupero, D.R., Motta, E.: AIDA: a knowledge graph about research dynamics in academia and industry. Quant. Sci. Stud. **2**(4), 1356–1398 (2021)
38. Schindler, D., Zapilko, B., Krüger, F.: Investigating software usage in the social sciences: a knowledge graph approach. In: Proceedings of the Extended Semantic Web Conference. ESWC'20, pp. 271–286 (2020)
39. Färber, M., Lamprecht, D.: The data set knowledge graph: creating a linked open data source for data sets. Quant. Sci. Stud. **2**(4), 1324–1355 (2021)
40. Sinha, A., et al.: An overview of Microsoft Academic Service (MAS) and applications. In: Proceedings of the 24th International Conference on World Wide Web Companion. WWW'15, pp. 243–246 (2015)
41. Massri, M.B., Spahiu, B., Grobelnik, M., Alexiev, V., Palmonari, M., Roman, D.: Towards innograph: a knowledge graph for AI innovation. In: Companion Proceedings of the ACM Web Conference, pp. 843–849 (2023)
42. Marginson, S.: University rankings and social science. Eur. J. Educ. **49**(1), 45–59 (2014)
43. Hu, Z., Dong, Y., Wang, K., Sun, Y.: Heterogeneous graph transformer. In: Proceedings of the Web Conference, pp. 2704–2710 (2020)
44. Färber, M., Nishioka, C., Jatowt, A.: ScholarSight: visualizing temporal trends of scientific concepts. In: Proceedings of the 19th ACM/IEEE on Joint Conference on Digital Libraries. JCDL'19, pp. 436–437 (2019)
45. Lou, Y., Kuehl, B., Bransom, E., Feldman, S., Naik, A., Downey, D.: S2abEL: a dataset for entity linking from scientific tables. arXiv preprint arXiv:2305.00366 (2023)
46. Carrasco, M.H., Luján-Mora, S., Maté, A., Trujillo, J.: Current state of linked data in digital libraries. J. Inf. Sci. **42**(2), 117–127 (2016)
47. Färber, M., Albers, A., Schüber, F.: Identifying used methods and datasets in scientific publications. In: Proceedings of the Workshop on Scientific Document Understanding Co-located with 35th AAAI Conference on Artificial Intelligence. SDU@AAAI'21 (2021)
48. Fathalla, S., Vahdati, S., Auer, S., Lange, C.: Towards a knowledge graph representing research findings by semantifying survey articles. In: Proceedings of the 21st

International Conference on Theory and Practice of Digital Libraries. TPDL'17, pp. 315–327 (2017)
49. Färber, M., Lamprecht, D.: The green AI ontology: an ontology for modeling the energy consumption of AI models. In: Proceedings of the 21st International Semantic Web Conference. ISWC'22 (2022)
50. Saier, T., Krause, J., Färber, M.: unarxive 2022: All arXiv publications preprocessed for NLP, including structured full-text and citation network. In: Proceedings of the 2023 Joint Conference on Digital Libraries. JCDL'23 (2023)
51. Diefenbach, D., Wilde, M.D., Alipio, S.: Wikibase as an infrastructure for knowledge graphs: the EU knowledge graph. In: Hotho, A., et al. (eds.) ISWC 2021. LNCS, vol. 12922, pp. 631–647. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-88361-4_37
52. Heist, N., Haase, P.: Flexible and extensible competency management with knowledge graphs. In: Proceedings of the 20th International Semantic Web Conference. ISWC'21 (2021)
53. Hoffman, M.R., Ibáñez, L.D., Fryer, H., Simperl, E.: Smart papers: dynamic publications on the blockchain. In: Proceedings of the 15th Extended Semantic Web Conference. ESWC'18, pp. 304–318 (2018)
54. Jaradeh, M.Y., Auer, S., Prinz, M., Kovtun, V., Kismihók, G., Stocker, M.: Open research knowledge graph: towards machine actionability in scholarly communication. CoRR abs/1901.10816 (2019)
55. Cossu, M., Färber, M., Lausen, G.: Prost: distributed execution of SPARQL queries using mixed partitioning strategies. In: Proceedings of the 21st International Conference on Extending Database Technology, EDBT 2018, Vienna, Austria, 26–29 March 2018, OpenProceedings.org, pp. 469–472 (2018)
56. Bassani, E., Kasela, P., Raganato, A., Pasi, G.: A multi-domain benchmark for personalized search evaluation. In: Proceedings of the 31st ACM International Conference on Information and Knowledge Management, pp. 3822–3827 (2022)
57. Dzieżyc, M., Kazienko, P.: Effectiveness of research grants funded by European Research Council and Polish National Science Centre. J. Informetrics **16**(1) (2022)
58. Jonkers, K., Zacharewicz, T., et al.: Research Performance Based Funding Systems: A Comparative Assessment. Publications Office of the European Union, Luxembourg (2016)