The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data

Michael Färber(✉)

Institute AIFB, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
michael.faerber@kit.edu

Abstract. In this paper, we present the *Microsoft Academic Knowledge Graph* (MAKG), a large RDF data set with over eight billion triples with information about scientific publications and related entities, such as authors, institutions, journals, and fields of study. The data set is licensed under the Open Data Commons Attribution License (ODC-By). By providing the data as RDF dump files as well as a data source in the Linked Open Data cloud with resolvable URIs and links to other data sources, we bring a vast amount of scholarly data to the Web of Data. Furthermore, we provide entity embeddings for all 210 million represented publications. We facilitate a number of use case scenarios, particularly in the field of digital libraries, such as (1) entity-centric exploration of papers, researchers, affiliations, etc.; (2) data integration tasks using RDF as a common data model and links to other data sources; and (3) data analysis and knowledge discovery of scholarly data.

Keywords: Scholarly data · Knowledge graph · Digital libraries

1 Introduction

A vast number of scientific publications are published every year. In total, we can count over 81 million scientific journal articles and over 4 million conference papers that have been published across the scientific fields so far.[1] The availability of the metadata about all these publications (and also the publications themselves) enables development of new systems and approaches in the field of digital libraries. For instance, relevant papers can be recommended to users for further reading (i.e., *paper recommendation*) or for citing (i.e., *citation recommendation*). Also, other kinds of entities in the academic field (e.g., venues or reviewers) can be recommended.

However, obtaining large data sets about scientific publications, researchers, institutes, and venues is often nontrivial (see Sect. 2). Only very few data providers provide data according to W3C standards and linked data principles

[1] The values are based on SPARQL queries executed against our data set presented in Sect. 3.

C. Ghidini et al. (Eds.): ISWC 2019, LNCS 11779, pp. 113–129, 2019.
https://doi.org/10.1007/978-3-030-30796-7_8

(i.e., model the data in RDF, enable use of SPARQL as a query language, use resolvable URIs, and link resources to other data sources). The existing RDF data sets are limited in that (1) they are rather small, (2) they cover only a few entity types, (2) they only cover specific scientific domains, (3) they cover data primarily from a single publisher, or (4) they are outdated (see Sect. 2).

In this paper, we present a large RDF data set with over eight billion triples containing information about scientific publications and entities of related entity types, such as authors, institutions, journals, conferences, and fields of study. This data set is based on the *Microsoft Academic Graph* (MAG)[2] [1], which is available with a subscription.[3] Contrarily to what the word "graph" suggests, Microsoft does not provide this data in the form of a (knowledge) graph, although the data is amenable to be modeled in such a structure. Instead, large database dumps (text files, overall about 350 GB in size) are provided every couple of weeks. Although the data seems to be relevant for a variety of disciplines and institutions (e.g., libraries) and various use cases (e.g., evaluating new metrics for the scientific impact of papers and researchers), storing and processing this data set would require overcoming considerable obstacles. In particular, researchers in nontechnical research disciplines, such as digital libraries, digital humanities, and social sciences, might lack the necessary skills and infrastructure to work with the dump files. Going one step further and having the data set available in RDF might appear even more utopian. In addition, IT experts and practitioners might be interested in just using an existing SPARQL endpoint, in getting resource descriptions via URI resolution, or in using pretrained entity embeddings.

By enriching the MAG data and providing this data as an RDF knowledge graph (both in the form of RDF files and as a data source on the Web with HTTP-resolvable URIs) and pretrained entity embeddings of it, potential data consumers of the MAG can get rid of these obstacles. We facilitate a number of scenarios concerning data consumption and data analytics: (1) entity-centric exploration of papers (even time-aware, as we provide updates every few months); (2) easier data integration through the use of RDF and by linking resources to other data sources; and (3) data analysis and knowledge discovery (e.g., measuring the popularity of papers and authors; recommending papers, researchers, and venues; and analyzing the evolution of topics over time).

Overall, we make the following contributions in this paper:

- We transform all data of the MAG, available as text files with a subscription, into RDF, while reusing common vocabularies and serializing the data in the N-Triples format.[4] This leads to a knowledge graph with over 8 billion triples.

[2] See https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/.

[3] Both the initial MAG data set and the MAKG provided by us are licensed under the Open Data Commons Attribution License (ODC-By; https://opendatacommons.org/licenses/by/1-0/index.html; last access: April 9, 2019).

[4] The source code is available online at https://github.com/michaelfaerber/MAG2RDF.

- We link resources to other data sources on the Web, such as DBpedia, Wikidata, OpenCitations [2], and the Global Research Identifier Database (GRID).[5]
- We provide the *Microsoft Academic Knowledge Graph* (MAKG), hosted at http://ma-graph.org,[6] in the following ways:
  1. Every few months, we provide NT-files at http://ma-graph.org, Zenodo,[7] and Amazon S3[8] to the public (1+ TB per version).
  2. We make the URIs of the MAKG resolvable, allowing the MAKG to be part of the Linked Open Data cloud.[9]
  3. We index all MAKG data in a triple store and make it publicly available via a SPARQL endpoint (see http://ma-graph.org/sparql).
- We provide entity embeddings for all 210 million publications represented in the MAKG.

The rest of this paper is structured as follows: First, we discuss related work (see Sect. 2). Then, we describe the process of generating the MAKG RDF data and its characteristics (see Sect. 3), before presenting the MAKG entity embeddings (see Sect. 4). Subsequently, we outline use case scenarios (see Sect. 5), before we conclude the paper (see Sect. 6).

2 Related Work

First of all, the modeling of the computer science bibliography DBLP in RDF [3], Springer's SciGraph,[10] and OpenCitations [2] are noteworthy projects. These projects are restricted to a single discipline (e.g., DBLP), to publications derived from one publisher (e.g., SciGraph), or to the pure modeling of papers and their citation relations without considering other entity types, such as venues and fields of study (e.g., OpenCitations).

Based on initiatives such as WikiCite,[11] Wikidata contains a considerable amount of bibliographic metadata about publications and their authors. Note, however, that the MAKG contains significantly more bibliographic information than Wikidata (e.g., 209,792,741 papers in the MAKG vs. 16,324,110 in Wikidata; see Table 1). The MAKG encompasses 1,380,196,397 references[12] between papers. This is almost eight times the number of references in Wikidata. Note also that in Wikidata, most of the papers are written in English, while in the MAKG, only 65% of the papers are in English.

[5] See https://www.grid.ac/.

[6] The MAKG is also available at the persistent URI https://w3id.org/makg/.

[7] See http://doi.org/10.5281/zenodo.2159723.

[8] See the S3 bucket `arn:aws:s3:::ma-kg`.

[9] See, e.g., `curl -H"Accept:text/n3"` http://ma-graph.org/entity/2826592117 and `curl -H "Accept:text/ttl"` http://ma-graph.org/entity/2826592117.

[10] See https://www.springernature.com/de/researchers/scigraph.

[11] See http://wikicite.org/.

[12] In our paper, the term "citations" refers to in-text citations while "references" refers to links on the document level.

Table 1. Statistical comparison of scholarly RDF data sets (the MAKG as of Nov. 2018, the Open Citations Corpus (OOC) as of Sept. 2017, the OpenCitations Index of Crossref open DOI-to-DOI citations (COCI) as of Nov. 2018, Wikidata as of Dec. 2018 based on http://wikicite.org/, and the AceKG as of 2018).

	MAKG	OOC	COCI	Wikidata	AceKG
# Publications	209,792,741	326,743	46,534,705	21,783,796	61,704,089
# References	1,380,196,397	12,652,601	445,826,118	174,259,894	479,648,000

Among the most similar works to our work is the AceKG [4], a database with 3 billion triples of academic facts about papers, authors, fields of study, venues, and institutes. AceKG data is modeled in RDF. However, contrary to our work, no significant existing vocabularies are reused, and no publicly available triple store or host for resolving URIs via HTTP can be expected. Moreover, all data is gained from the database of the startup Acemap, and no continuous updates of this knowledge graph are provided.

SPedia [5] is a knowledge graph with information about 9 million papers gained from the platform SpringerLink. With over 300 million RDF triples, this data set is a rich source of bibliographic information in the RDF format. Still, it is considerably smaller than the MAKG. Although no SPARQL endpoint or URI-resolving host is available online, data is available upon request. Furthermore, no mappings to other Linked Open Data sources are provided.

Nuzzolese et al. [6] focus on refactoring the Semantic Web Conference ontology. They propose a new ontology [7] and an RDF data set [8] based on it. However, the data set only covers Semantic Web conferences [6]. It is thus only suitable for rather specific usage scenarios compared to our MAKG.

Konstantinou et al. [9] introduce a transformation process for converting an institutional repository into Linked Open Data. This includes the process of creating mappings between domain vocabularies.

3 The Microsoft Academic Knowledge Graph

Based on a push service of Microsoft, we are able to obtain a fresh version of the MAG every few weeks in the form of tab-separated plaintext files. All relevant data to be modeled in RDF takes about 350 GB of disk space. Because the data is in the form of a relational database dump, the data needs to be transformed to obtain the MAKG in the graph structure as outlined in Fig. 1. Furthermore, creating links to other data sources, such as to DBpedia, Wikidata, OpenCitations, and GRID, is another important step to integrating the MAKG into the Linked Open Data cloud.

Overall, we cover 10 entity types in the MAKG, including papers, authors, and affiliations. An overview of the entity types, the object properties, and the data type properties is provided in Fig. 1. Concerning the used knowledge graph properties, our goal was to reuse as much existing vocabulary as possible. Because the data in the MAKG is about publications, researchers, institutions,

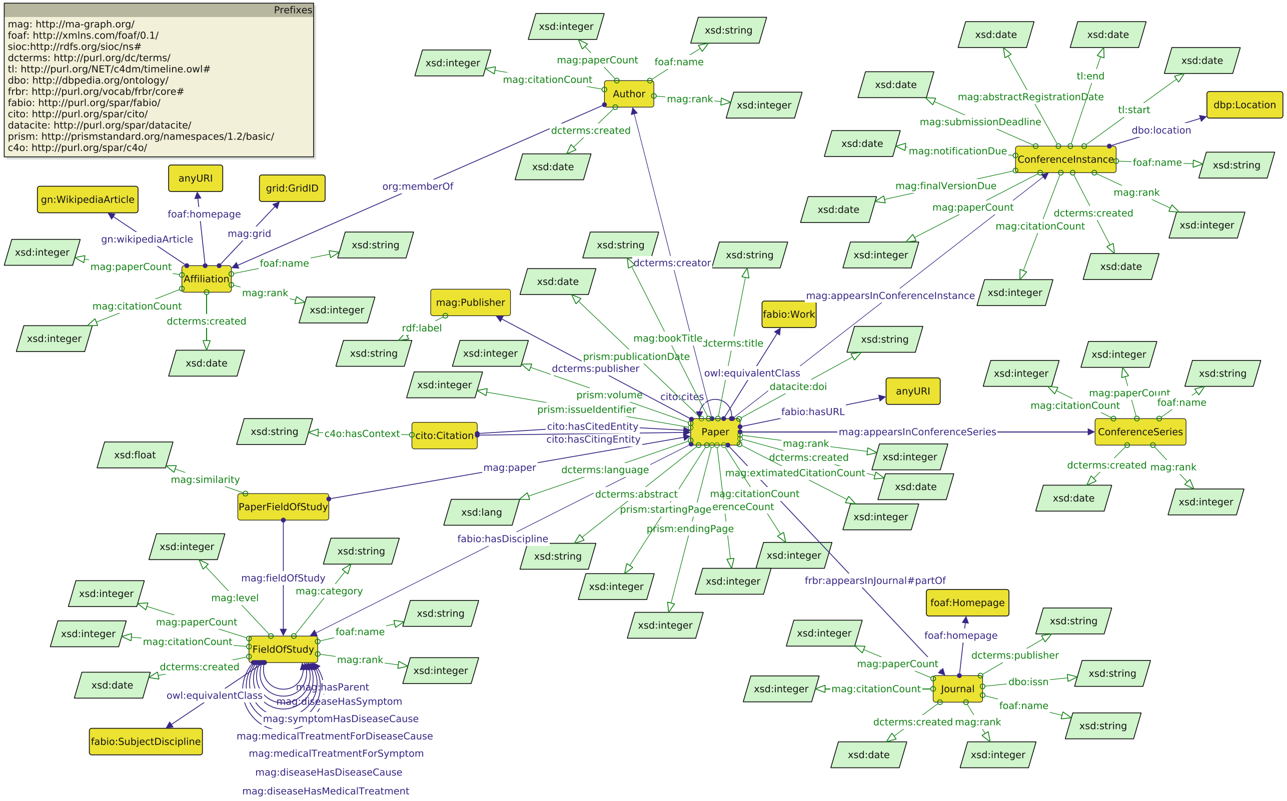


Fig. 1. Schema of the Microsoft Academic Knowledge Graph.

Table 2. Used vocabularies and corresponding prefixes.

Prefix	Associated URI
`mag`	http://ma-graph.org/
`foaf`	http://xmlns.com/foaf/0.1/
`sioc`	http://rdfs.org/sioc/ns
`dcterms`	http://purl.org/dc/terms/
`tl`	http://purl.org/NET/c4dm/timeline.owl
`dbo`	http://dbpedia.org/ontology/
`frbr`	http://purl.org/vocab/frbr/core
`fabio`	http://purl.org/spar/fabio/
`cito`	http://url.org/spar/cito/
`datacite`	http://purl.org/spar/datacite/
`prism`	http://prismstandard.org/namespaces/1.2/basic/
`c4o`	http://purl.org/spar/c4o/

and similar items, we were mainly able to orient ourselves on the existing Semantic Publishing and Referencing (SPAR) ontologies [10], such as FaBiO, CiTo, PRISM, and C4O. In the end, we reused the vocabularies listed in Table 2.

3.1 The Creation Process

The original MAG data dump is presumably designed primarily for data processing (e.g., abstracts are pre-tokenized and not provided as one string). To create an RDF knowledge graph based on these dump files, major changes in the data formatting and the data modeling are necessary. In the following, we outline the most crucial steps of this transformation process.

Papers. The metadata about scientific papers is the core of the MAG data set. The file `Papers.txt` of the initial MAG data dump contains information directly related to papers, such as the paper's title, the publication date, the publisher, the link to the conference at which it appeared, and the reference and citation counts (in total, 21 attributes per paper). We model the represented document type of each publication according to the document types covered in the FaBiO ontology (see Fig. 4). Furthermore, we represent the information about the paper's associated journal, conference series, and conference instance in the form of URIs to provide facts about those entities later on. Note that we skip some information, such as the paper's normalized title (in lower case) and the publication year, from the initial dump for the RDF creation, because this information is already provided in the form of other facts.

Further information about papers represented in the MAKG originate from the following dump files:

- PaperAbstractInvertedIndex.txt: For a fraction of the papers stored in the MAG, the abstracts are available. However, the abstracts are only provided as JSON objects in which the key represents the token position and the value the token string (i.e., it is an inverted index). Because our knowledge graph is designed for providing the data in a more natural, non-data mining fashion, we reverted the index and added the papers' abstracts as literal information to the papers in our knowledge graph.
- PaperLanguages.txt: Each paper usually has one assigned language in which it is written. We follow the MAG's initial language encoding and use ISO 639-1 for the language code and ISO 3166 for the region code as necessary (e.g., "en" for English and "zh_chs" for simplified Chinese).
- PaperUrls.txt: We include the URL at which each paper is available online as an attribute of each paper in our knowledge graph. Note that the URLs provided in the MAG dump often do not link to the papers directly but to the landing pages provided by the papers' publishers.

Authors. Providing information about papers' authors is an obvious next step. Given PaperAuthorAffiliations.txt, we can derive which authors wrote which paper (and, in theory, in which author position), and model this information as facts in our knowledge graph, thereby connecting papers with authors. Author entities themselves are enriched by the attribute information provided in Authors.txt. Specifically, we store, among other things, the authors' names, their last-known affiliations (linking to affiliation entities using the memberOf property), their paper counts, and their citation counts.

Affiliations. In our knowledge graph, we also provide information about the affiliations of the papers' authors based on the Affiliations.txt file. Among others, we include the affiliation's name as literal, a link to the institution's GRID identifier, a link to the institution's official homepage, a link to the English Wikipedia article describing this institution, and the number of papers and citations of the institution so far, given the reference and citation information in the MAG. Similar to the other file conversions, we transformed the data into RDF statements using appropriate data types in the case of literals. As far as possible and appropriate, we also transformed string values into URIs in accordance with the linked data principles of having entities represented as URIs. In particular, the links between the affiliations and the Global Research Identifier Database (GRID) identifiers are noteworthy. Because the GRID is part of the Linked Open Data cloud and because GRID URIs of the form http://www.grid.ac/institutes/grid.446382.f are resolvable via HTTP, we transformed the pure GRID identifiers into URIs by adding the URI prefix.

Venues. The MAG data dump provides us with information about conferences (given ConferenceInstances.txt and ConferenceSeries.txt) and journals (given Journals.txt).

- *Conference instances* represent single events at which papers are presented. In addition to the conference name (given in abbreviated form, such as "ECIR 2015"), we represent various attributes of each conference instance in the MAKG, such as the location, the website, temporal information (the start and end date of the conference and deadlines, such as the abstract submission deadline, the paper submission deadline, the notification date, and the final version due date), the number of papers published at this conference, and the number of citations of this conference's papers. For a better integration of the MAKG as a data source into the Linked Open Data cloud, we transform the strings with the conference location (typically city names with their country, such as "Oslo, Norway") into DBpedia URIs. In order to ensure a well performing word-sense-disambiguation, we use the state-of-the-art text annotation tool x-LiSA [11]. Because DBpedia is very rich in terms of cities, we obtained URIs for almost all locations (namely 15,530). Given the conference series identifier, we link each conference instance to the corresponding conference series (e.g., "SIGMOD 2015" to all SIGMOD conferences).
- *Conference series* are represented as URIs with facts about their names (e.g., "SIGMOD"), their paper counts, their citation counts, and their ranks (according to the MAG data set).
- *Journals* are modeled in RDF with facts about the name (e.g., "Scandinavian Journal of Forest Research"), the ISSN number, the publisher, the homepage, the paper count, the citation count, and the rank within the MAKG, among other things.

Taxonomy of Scientific Concepts. Papers in the MAG are assigned to specific research fields and concepts, called the *fields of study* (given in `FieldsOfStudy.txt`). Each field of study is associated with an abstraction level, ranging from 1 to 5. For the MAKG, we model the fields of study as entities of the entity type `FieldOfStudy`. We also store the association of each paper with at least one field of study (given in `PaperFieldsOfStudy.txt`). In this way, the RDF data can be used to categorize papers. We use parent-child relationships between the fields of study (given in `FieldsOfStudyChildren.txt`) to form a taxonomy of scientific concepts within the MAKG. Note, however, that this taxonomy is not a tree. Fields of study can have multiple parents (e.g., "Graph theory" is assigned to computer science and mathematics).

For specific fields of study, the original MAG data contains in addition so-called "main type" information about papers. Main types are primarily given in the field of biology (e.g., "biology.organism_classification"), because this field is well-represented. Relatively few fields of study have such a main type. Nevertheless, we store this additional information if available, along with the general information about the fields of study (e.g., the field of study's name, paper count, citation count, and hierarchy level).

Citations and References. The information about which papers reference which other papers is available as `PaperReferences.txt` and can be directly transformed to RDF triples. Note that if reference information is given, it is ensured that both the referencing paper and the referenced paper are covered by the MAKG. Thus, the issue of various other corpora containing scientific papers (e.g., arXiv CS [12], unarXiv [13], the ACL Anthology Network,[13] and the Scholarly Dataset 2[14]) that reference papers "outside" the data set (i.e., detailed metadata is not available, leading to issues regarding developing approaches for recommending papers, citations, or references) is not a problem.

In addition to the references, which are links on the document level, for a fraction of all papers, the MAG also contains the sentences in which the citations occur, i.e., the so-called *citation contexts* (as a string and with the identifiers of the citing paper and cited paper; see `PaperCitationContexts.txt`). Note that these citation contexts have been automatically extracted by Microsoft. Thus, the citation contexts are to some extent noisy. Because we have to deal with a ternary relationship (which paper cites which other paper in which context) here, we decided to model the citation information separately from the reference information using the class `cito:Citation` from the CiTo ontology as the entity type. Although we do not have the citation context for each reference, it is a valuable information source for tasks such as citation recommendation and citation-based paper summarization.

Summary. Overall, the MAKG, based on the MAG dump of November 2018, contains 8,272,187,245 RDF triples. About 1.2 TB of disk space need to be allocated for the uncompressed RDF files. Indexing the data in Virtuoso requires about 514 GB of disk space and takes about 10 h. Although a lower assignment would be possible, we configured the RDF triple store to use 256 GB of RAM. As a consequence, our SPARQL endpoint can be queried by many users simultaneously and real-world queries can be executed without timeouts.

On the schema level, the MAKG contains 47 properties and 13 entity types (with 8 entity types being in the namespace http://ma-graph.org). As outlined previously, we were able to link 6,706 institute representations to the corresponding DBpedia concepts, 15,530 conference instances to the corresponding Wikipedia articles, and 18,673 affiliations to the corresponding GRID URIs.

3.2 Creating `owl:sameAs` Statements

In addition to the MAKG core data set outlined so far, we linked instances of the MAKG to instances of OpenCitations and Wikidata.[15] The mappings were created by matching the papers' digital object identifiers (DOIs).

[13] See http://clair.eecs.umich.edu/aan/index.php.

[14] See https://www.comp.nus.edu.sg/~sugiyama/Dataset2.html.

[15] The source code is online available at https://github.com/michaelfaerber/makg-linking. The mappings are available as `nt` files with `owl:sameAs` statements on our website.

Table 3. MAKG's entity types and number of instances (as of 2018-11-09).

Entity type	# Instances
Author	253,641,783
Paper	209,792,741
Citation	146,257,535
Field of study	229,716
Journal	48,650
Affiliation	25,431
Conference instance	15,704
Conference series	4,337

1. *OpenCitations.* We were able to create 15,666,233 mappings between papers modeled in the OpenCitations Corpus and papers modeled in the MAKG. This corresponds to 7.5% of all MAKG's papers and 3.5% of OpenCitations' papers. For the mapping, the papers' URIs in OpenCitations were used because they contain the DOI (cf. `http://dx.doi.org/<DOI>`). Of the papers having `owl:sameAs` links to the MAKG, 97.3% are written in English.
2. *Wikidata.* We were able to create 5,472,038 mappings between papers modeled in Wikidata and papers modeled in the MAKG. This corresponds to 2.6% of the MAKG's papers and 33.5% of Wikidata's papers. Note that only those Wikidata's papers were candidates for interlinking that provide DOIs. Of the Wikidata papers having `owl:sameAs` links to the MAKG, 99% are written in English.

3.3 Key Statistics of the Microsoft Academic Knowledge Graph and Example SPARQL Queries

Table 3 shows the distribution of the entities among the different entity types.[16] The MAKG surprisingly contains more authors than papers and more papers than citations. Also, the number of affiliations (25,431) is relatively low given that all research institutions in all fields should be represented. This explains why we have an affiliation in the MAKG only for a fraction of the papers (namely, for 20,928,914 papers). On average, according to this data version, 2.45 authors write a paper together and an author writes 2.94 papers.

Compared to a previous analysis of the MAG [14], the number of instances for all entity types has increased, except for the number of conference instances, which has dropped from 50,202 to 15,704. An obvious reason for this reduction is a data cleaning process. While the number of journals, authors, and papers

[16] Note that only the number of citations is listed and not the number of references, because references are modeled in the MAKG via a relation (`cito:cites`). There are 1,380,196,397 references in the MAKG.

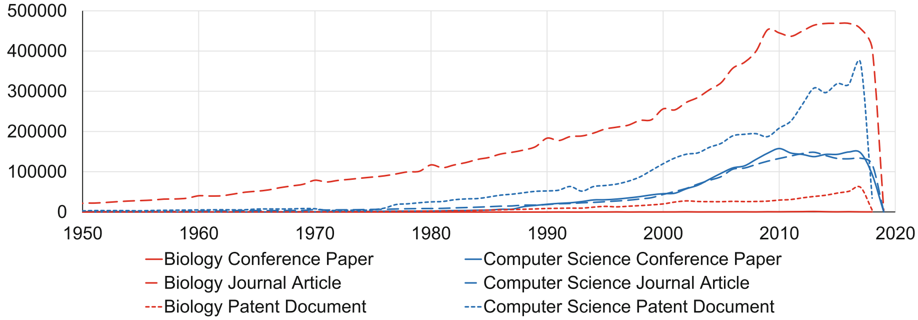
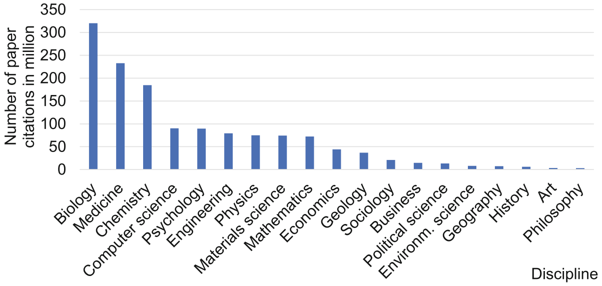


Fig. 2. Number of publications per publication year, per discipline (computer science and biology), and per publication type since 1950.



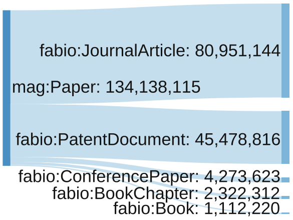


Fig. 3. Number of paper citations per discipline (over all publication types).

Fig. 4. Distribution of `mag:Paper` instances.

have doubled in size compared to the 2016 version [14], the number of conference series and fields of study have increased (almost) four times.

Figure 2 shows how many publications have been published per year in the field of biology and computer science (as example disciplines) according to our data. We can observe that journal articles in biology are published most frequently, followed by patent documents in computer science. Not surprisingly, the number of conference papers in biology is marginal.

Figure 3 displays the number of paper citations per discipline according to the MAKG. As expected, biology, medicine, and chemistry papers are cited the most, while history, art, and philosophy papers are cited the least. Figure 4 shows the frequency of instances per subclass of `mag:Paper`. Note that Figs. 2, 3, and 4 were generated by means of SPARQL queries using our SPARQL endpoint. Listings 1 and 2 show examples of how the MAKG can be queried with SPARQL.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX magp: <http://ma-graph.org/property/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX fabio: <http://purl.org/spar/fabio/>
PREFIX org: <http://www.w3.org/ns/org#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?affilName ?citCountAffil
WHERE {
?field rdf:type <http://ma-graph.org/class/FieldOfStudy> .
?field foaf:name "Machine␣learning"^^xsd:string .
?paper fabio:hasDiscipline ?field .
?paper dcterms:creator ?author .
?author org:memberOf ?affiliation .
?affiliation foaf:name ?affilName .
?affiliation magp:citationCount ?citCountAffil . }
GROUP BY ?affilName ?citCountAffil
ORDER BY DESC(?citCountAffil)
LIMIT 100
```

List. 1. Querying the top 100 institutions in the area of machine learning according to their overall number of citations.

```
PREFIX magp: <http://ma-graph.org/property/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX dbr: <http://dbpedia.org/resource/>

SELECT ?authorName (COUNT(?confInstance) AS ?freq)
WHERE {
 ?paper dcterms:creator ?author .
 ?author foaf:name ?authorName .
 ?paper magp:appearsInConferenceInstance ?confInstance .
 ?confInstance dbo:location dbr:Honolulu . }
GROUP BY ?authorName
ORDER BY DESC(?freq)
LIMIT 100
```

List. 2. Querying the top 100 researchers who have been most frequently to conferences in Honolulu, Hawaii.

3.4 Linked Data Set Descriptions and Ratings

The initial MAG data was provided by Microsoft under the *Open Data Commons license* (ODC-BY),[17] which grants users the right to add values and redistribute the derivatives based on the terms of the Open Data Commons license. All MAKG resources produced by us are published under ODC-BY.

Aside from the MAG RDF documents, we provide the following linked data set descriptions (all available at http://ma-graph.org/):

- *OWL*: We provide our ontology as an OWL file describing the used classes, object properties, and data type properties.

[17] See https://docs.microsoft.com/en-us/academic-services/graph/get-started-setup-provisioning#open-data-license-odc-by.

- *VOAF*: We enrich our ontology with Vocabulary-of-a-Friend (VOAF)[18] descriptors. VOAF is an extension of VoID[19] for linking the ontology to other vocabularies and for introducing the vocabulary to the Linked Open Data community.
- *VoID*: We provide a VoID file to describe our linked data set with an RDF schema vocabulary.

Furthermore, we can categorize the MAKG 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:*[20] Our MAKG RDF data set is a 5-star data set according to this scheme, because we provide our data set in RDF (leading to 4 stars) and link (1) entity URIs to DBpedia, Wikidata, OpenCitations, and GRID, and (2) our vocabulary URIs to other vocabularies (leading to 5 stars).
- *Linked Data vocabulary star rating* [15]: This rating is intended to rate the use of vocabulary within Linked (Open) Data. By providing an OWL file, by linking our vocabulary to other vocabularies (see the SPAR ontologies), and by creating a VOAF file, we are able to provide the vocabulary with 4 stars.

Due to our subscription to the Microsoft Academic services, we periodically obtain fresh versions of the MAG dump files. The transformation process described in Sect. 3.1 runs periodically in a semi-automated fashion. Because we are on the mailing list of the Microsoft Academic team, we are notified of any changes to the MAG data and of the data provisioning. In the past, this process has ensured updates of the RDF generation step according to changed data formatting and data provisioning (from Azure Data Lake to Azure Storage).

4 The Microsoft Academic Knowledge Graph Entity Embeddings

Apart from creating and providing the MAKG data set and services (e.g., the SPARQL endpoint), we computed embeddings for the MAKG entities. Entity embeddings have proven to be useful as implicit knowledge representations in a variety of scenarios, as indicated in Sect. 5. Because the MAKG is available in RDF, we applied RDF2Vec [16] to the MAKG using the skip-gram model, a windows size of 5, 128 dimensions, and 10 epochs of training. The training was performed on a machine with 500 GB of RAM and 64 cores. The resulting embedding vectors for all 210 million papers in the MAKG (uncompressed using 310 GB and compressed using 93 GB of storage) are linked on our website.

[18] See http://lov.okfn.org/vocommons/voaf.

[19] See http://www.w3.org/TR/void/.

[20] See http://5stardata.info/.

5 Use Cases of the Microsoft Academic Knowledge Graph

In the past, the MAG has been used in various contexts. This is reflected in the high number of citations of the publication that describes the original data set [1].[21] Also, the MAKG has been recognized and adopted by the community. Considering only Zenodo, the MAKG data has been viewed 1000+ times, downloaded 100+ times, and seen by 75+ Twitter users so far.[22] On average, 50+ unique visitors reach the MAKG website each day.[23] In the following, we outline typical use cases of the original MAG data and of the MAKG.

Using the MAKG as a Linked Data Source in the Linked Open Data Cloud. Because the MAKG is part of the Linked Open Data cloud and contains links to other data sources (see Sect. 3.2), it contributes significantly to the use of linked data in the digital libraries context [17]. Particularly, by using our SPARQL endpoint, users and machines can perform queries that are often associated with fewer burdens than when using the original MAG data dump consisting of raw text files [17] (see also Sect. 3.3). The MAKG can be considered a central data hub for credibility in the linked data context, because it contains metadata about papers (and their authors) that state claims. Claims and crucial concepts mentioned in text documents (e.g., papers' full texts) can be linked to papers and authors in the MAKG to substantiate them [18].

Using the MAKG for Natural Language Processing Tasks. We can mention two examples here:

1. Citation-based tasks, such as citation recommendation, often depend on natural language processing and require implicit or explicit representations of papers, researchers, and institutions. In the case of the MAKG, embeddings for papers and other entities can easily be generated using existing methods for RDF graph embeddings, as demonstrated in Sect. 4.
2. Entity linking describes the task of linking phrases in a text to knowledge graph entities. It has shown several advantages compared to traditional text mining and information retrieval approaches. Consequently, MAKG entities, such as the fields of study and the authors, can be used as the basis for annotating texts (e.g., annotating scientific texts with scientific concepts [19]). Furthermore, using the MAKG, semantic search systems can be developed [20] that are superior to bag-of-words models.

Using the MAKG for Digital Library Tasks. So far, the MAG has been used, among other ways, for citation analysis [21] and for impact analysis of

[21] Sinha et al. [1] have obtained 187 citations as of March 29, 2019, according to Google Scholar.

[22] See https://doi.org/10.5281/zenodo.2159723 (as of April 10, 2019). Note that the data set is also available at http://ma-graph.org/ and on Amazon S3.

[23] See http://ma-graph.org/usage-statistics/ for usage statistics concerning the website and the SPARQL endpoint.

papers and researchers [22,23]. The original MAG data has also been combined with AMiner data to form the Open Citation Graph.[24] In the future, Linked Open Data-based recommender systems that recommend papers or citations can use the MAKG as an underlying database. Furthermore, one can envision that the working style of researchers will considerably change in the next few decades [24,25]. 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. The MAKG can be combined with such new structured data sets easily due to its RDF data format.

Using the MAKG for Benchmarking. Because the MAKG is large in size (over 1 TB in N-Triples format), contains various kinds of information (e.g., papers, authors, institutions, and venues as well as various data types), has uncertainty in the data, and is updated periodically, the MAKG data fulfills the "4 V's" of big data very well. Thus, the MAKG may also be suitable for evaluating methods and benchmarking systems.

6 Conclusions

In this paper, we proposed a large RDF data set with over eight billion triples that covers scholarly data in all scientific disciplines. We described the creation process based on the Microsoft Academic Graph data and the characteristics of our data set. We showed that querying the data set based on SPARQL enables performing complex analyses. By making the SPARQL endpoint publicly available and the URIs resolvable, we enriched the Linked Open Data cloud with a rich data source in the field of digital libraries. We provide RDF dumps, linked data set descriptions, a SPARQL endpoint, and trained entity embeddings online at http://ma-graph.org. In the future, we will use our data set for social analysis studies, because complex information needs can be answered by single SPARQL queries, and for recommending citations in scientific texts.

References

1. 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 2015, pp. 243–246 (2015)
2. Peroni, S., Dutton, A., Gray, T., Shotton, D.M.: Setting our bibliographic references free: towards open citation data. J. Doc. **71**(2), 253–277 (2015)
3. Aleman-Meza, B., Hakimpour, F., Arpinar, I.B., Sheth, A.P.: SwetoDblp ontology of computer science publications. J. Web Semant. **5**(3), 151–155 (2007)
4. 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 2018, pp. 1487–1490 (2018)

[24] See https://www.openacademic.ai/oag/.

5. Aslam, M.A., Aljohani, N.R.: SPedia: a central hub for the linked open data of scientific publications. Int. J. Semant. Web Inf. Syst. **13**(1), 128–146 (2017)
6. Nuzzolese, A.G., Gentile, A.L., Presutti, V., Gangemi, A.: Conference linked data: the scholarlydata project. In: Groth, P., et al. (eds.) ISWC 2016. LNCS, vol. 9982, pp. 150–158. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46547-0_16
7. Nuzzolese, A.G., Gentile, A.L., Presutti, V., Gangemi, A.: Semantic web conference ontology - a refactoring solution. In: Sack, H., Rizzo, G., Steinmetz, N., Mladenić, D., Auer, S., Lange, C. (eds.) ESWC 2016. LNCS, vol. 9989, pp. 84–87. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47602-5_18
8. Gentile, A.L., Acosta, M., Costabello, L., Nuzzolese, A.G., Presutti, V., Recupero, D.R.: Conference live: accessible and sociable conference semantic data. In: Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, pp. 1007–1012 (2015)
9. Konstantinou, N., Spanos, D., Houssos, N., Mitrou, N.: Exposing scholarly information as Linked Open Data: RDFizing DSpace contents. Electron. Libr. **32**(6), 834–851 (2014)
10. Peroni, S., Shotton, D.: The SPAR ontologies. In: Vrandečić, D., et al. (eds.) ISWC 2018. LNCS, vol. 11137, pp. 119–136. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00668-6_8
11. Zhang, L., Rettinger, A.: X-LiSA: cross-lingual semantic annotation. PVLDB **7**(13), 1693–1696 (2014)
12. Färber, M., Thiemann, A., Jatowt, A.: A high-quality gold standard for citation-based tasks. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, pp. 1885–1889 (2018)
13. Saier, T., Färber, M.: Bibliometric-enhanced arXiv: a data set for paper-based and citation-based tasks. In: Proceedings of the 8th International Workshop on Bibliometric-enhanced Information Retrieval, BIR 2019, pp. 14–26 (2019)
14. Herrmannova, D., Knoth, P.: An analysis of the Microsoft academic graph. D-Lib Mag. **22**(9/10) (2016)
15. Janowicz, K., Hitzler, P., Adams, B., Kolas, D., Vardeman, C.: Five stars of linked data vocabulary use. Semant. Web **5**(3), 173–176 (2014)
16. Ristoski, P., Paulheim, H.: RDF2Vec: RDF graph embeddings for data mining. In: Groth, P., et al. (eds.) ISWC 2016. LNCS, vol. 9981, pp. 498–514. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46523-4_30
17. 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)
18. Fathalla, S., Vahdati, S., Auer, S., Lange, C.: Towards a knowledge graph representing research findings by semantifying survey articles. In: Kamps, J., Tsakonas, G., Manolopoulos, Y., Iliadis, L., Karydis, I. (eds.) TPDL 2017. LNCS, vol. 10450, pp. 315–327. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67008-9_25
19. 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 2019, pp. 436–437 (2019)
20. Färber, M., Sampath, A., Jatowt, A.: *PaperHunter*: a system for exploring papers and citation contexts. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) ECIR 2019. LNCS, vol. 11438, pp. 246–250. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-15719-7_33
21. Hug, S.E., Ochsner, M., Brändle, M.P.: Citation analysis with Microsoft academic. Scientometrics **111**(1), 371–378 (2017)

22. Mohapatra, D., Maiti, A., Bhatia, S., Chakraborty, T.: Go wide, go deep: quantifying the impact of scientific papers through influence dispersion trees. In: Proceedings of the 19th ACM/IEEE Joint Conference on Digital Libraries, JCDL 2019, pp. 305–314 (2019)
23. Fire, M., Guestrin, C.: Over-optimization of academic publishing metrics: observing Goodhart's law in action. CoRR abs/1809.07841 (2018)
24. Hoffman, M.R., Ibáñez, L.-D., Fryer, H., Simperl, E.: Smart papers: dynamic publications on the blockchain. In: Gangemi, A., et al. (eds.) ESWC 2018. LNCS, vol. 10843, pp. 304–318. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93417-4_20
25. 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)