RESEARCH ARTICLE

The data set knowledge graph: Creating a linked open data source for data sets

Michael Färber and David Lamprecht

Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Keywords: data sets, linked open data, scholarly knowledge graph

ABSTRACT

Several scholarly knowledge graphs have been proposed to model and analyze the academic landscape. However, although the number of data sets has increased remarkably in recent years, these knowledge graphs do not primarily focus on data sets but rather on associated entities such as publications. Moreover, publicly available data set knowledge graphs do not systematically contain links to the publications in which the data sets are mentioned. In this paper, we present an approach for constructing an RDF knowledge graph that fulfills these mentioned criteria. Our *data set knowledge graph*, DSKG, is publicly available at http://dskg.org and contains metadata of data sets for all scientific disciplines. To ensure high data quality of the DSKG, we first identify suitable raw data set collections for creating the DSKG. We then establish links between the data sets and publications modeled in the Microsoft Academic Knowledge Graph that mention these data sets. As the author names of data sets can be ambiguous, we develop and evaluate a method for author name disambiguation and enrich the knowledge graph with links to ORCID. Overall, our knowledge graph contains more than 2,000 data sets with associated properties, as well as 814,000 links to 635,000 scientific publications. It can be used for a variety of scenarios, facilitating advanced data set search systems and new ways of measuring and awarding the provisioning of data sets.

1. INTRODUCTION

The number of data sets available on the web has increased steadily. Google Dataset Search (Brickley, Burgess, & Noy, 2019), for instance, covered more than 6 million data sets in September 2018 but over 28 million data sets by March 2020 (Benjelloun, Chen, & Noy, 2020). In addition, data portals and registration services, such as OpenAIRE with Zenodo (http://zenodo.org) as well as re3data (https://re3data.org/), have seen a sharp increase in the number of indexed data sets. Furthermore, scientific communities increasingly demand researchers to publish their research data according to the FAIR principles (Wilkinson, Dumontier et al., 2016) fostering the provision and reuse of data sets and their metadata. Having access to and using high-quality, rich, and interoperable metadata of data sets is therefore essential in many scenarios and will continue to gain in importance.

At present, metadata about data sets are collected in diverse ways: Web crawlers exist that search the web for data sets (Brickley et al., 2019); there are open data portals with collections or catalogs that index metadata and refer to the data set files; and there are freely accessible databases that were created and expanded jointly by users (Neumaier, Polleres et al., 2017). Aside from the data sources, the metadata about data sets are modeled by means of various

an open access journal

Citation: Färber, M., & Lamprecht, D. (2021). The data set knowledge graph: Creating a linked open data source for data sets. *Quantitative Science Studies*, *2*(4), 1324–1355. https://doi.org/10.1162/qss_a_00161

DOI: https://doi.org/10.1162/qss_a_00161

Corresponding Author: Michael Färber
michael.faerber@kit.edu

The MIT Press

standards (e.g., Schema.org and DCAT (Brickley et al., 2019), the DataCite metadata schema (Manghi, Bardi et al., 2019), and the CKAN and Socrata metadata schemas (Neumaier, Umbrich, & Polleres, 2017) and with varying degrees of quality (see Section 3). However, using Semantic Web technologies, such as the Resource Description Framework (RDF) (W3C, 2014), allowing the creation of knowledge graphs based on a standardized data model and format, has turned out to be particularly helpful for modeling metadata and linking it to existing data sources on the web (Latif, Limani, & Tochtermann, 2021; Neumaier et al., 2017; Vahdati, Karim et al., 2015). Specifically, in the academic field, several large knowledge graphs have been proposed and are freely available. For instance, the Microsoft Academic Knowledge Graph (MAKG) (Färber, 2019) contains 8 billion triples about publications and associated entities, such as authors, venues, and affiliations. Wikidata (https://wikidata.org), OpenCitations (Peroni & Shotton, 2020), and the Open Research Knowledge Graph (ORKG) (Jaradeh, Oelen et al., 2019) are further noteworthy knowledge graphs.

Although the existing scholarly knowledge graphs model data sets to some degree, they do not primarily focus on data sets but rather associated entities such as publications (Färber, 2019) (see Section 2). Moreover, they are often not publicly available (Brickley et al., 2019) and do not contain links to the publications in which the data sets are mentioned. We argue that a knowledge graph that fulfills the following criteria is highly beneficial for scholarly data mining:

1. The knowledge graph is publicly available and integrated into the Linked Open Data cloud. This means that the knowledge graph is based on RDF (W3C, 2014) as a widely used data model, facilitating data interoperability and data integration efforts, and that it is interlinked to other data sources, following the FAIR data principles (Wilkinson et al., 2016).
2. The knowledge graph is of high quality with respect to the accuracy and coverage (i.e., high number of provided properties).
3. All data sets modeled in the knowledge graph are linked to the scientific publications in which they are mentioned. Linking data sets to publications enables novel ways of knowledge discovery and scientific impact quantification (Baglioni, Manghi, & Mannocci, 2020). Specifically, the MAKG (Färber, 2019) modeling rich metadata about millions of publications from all scientific fields can be used as a link target.

In this paper, we present an approach for constructing an RDF knowledge graph that fulfills these mentioned criteria. Our *data set knowledge graph*, DSKG, is publicly available at http://dskg.org and https://doi.org/10.5281/zenodo.4478921, and contains metadata of data sets for the various scientific disciplines. To ensure high data quality (e.g., high accuracy of statements and high coverage of used properties) of the final knowledge graph, we first analyzed existing data set metadata collections and identified suitable data set collections that are particularly suitable for building a data set knowledge graph. Furthermore, we only considered data sets with their metadata that are mentioned in scientific publications. To this end, we parsed all 146 million publications' abstracts and all 241.5 million citation contexts available in the MAKG (Färber, 2019). Data sets mentioned in the abstracts or citation contexts of these publications are an essential aspect of these papers and we therefore link the data sets to the publications. As the author names of data sets can be ambiguous and our knowledge graph requires unique identifiers (URIs) for each entity, we developed and evaluated a method for author name disambiguation—the first one considering data set authors, to the best of our knowledge. To ensure that the knowledge graph is well integrated into the Linked Open Data cloud, we enrich the knowledge graph with links to ORCID, Wikidata, and the MAKG. Last

but not least, we provide data set entity embeddings for machine learning tasks. The embeddings were created by applying RDF2Vec (Ristoski, Rosati et al., 2019) to our knowledge graph. Overall, our knowledge graph contains 2,208 data sets and 813,551 links to scientific publications. We will update the knowledge graph quarterly via a semiautomatic process.

The DSKG can be used for a variety of scenarios regarding data consumption and data analysis:

1. The DSKG can be used as a database and evaluation basis for new applications, particularly in the context of data set search. For instance, based on our preliminary online system http://datasetsearch.net we show how data sets can be retrieved based on scientific problem descriptions. To this end, we utilized the interlinkage between data sets and publications.
2. The DSKG allows for easier data integration through the use of standard RDF vocabulary and by linking resources to other data sources.
3. The DSKG facilitates new ways of scholarly data analysis, such as determining the scientific influence and impact of data sets (Färber, Albers, & Schüber, 2021) ("h-index of data sets"), authors (Yi, Ludo, & Yong, 2021), and affiliations (Lin, Zhu et al., 2021).

Overall, our main contributions can be summarized as follows:

- We analyze the metadata about data sets from several sources with respect to data quality aspects.
- We link data sets to scientific publications in which they are mentioned. This results in 813,551 links to 634,803 publications in the MAKG.
- We implement and evaluate a method for author name disambiguation based on our DSKG.
- We link our DSKG to other Linked Open Data sources (namely, ORCID, Wikidata, and MAKG).
- We provide our DSKG with a SPARQL endpoint, resolvable URIs, and entity embeddings at http://dskg.org to the public and also share it at https://doi.org/10.5281/zenodo.4478921. Our source code for generating the DSKG is available at https://github.com/michaelfaerber/data-set-knowledge-graph.

Our paper is structured as follows: We first examine related work and delimit it from our work (see Section 2). In Section 3, we analyze existing data set metadata collections, based on which we select the sources for our DSKG. Section 4 presents the approach for creating the knowledge graph for data sets. We provide statistical key figures of the knowledge graph and an evaluation of our performed author name disambiguation in Section 5. In Section 6, we show possible application scenarios of the knowledge graph. Finally, in Section 7, we summarize our work.

2. RELATED WORK

In recent years a field of research has arisen around the search for data sets (Chapman, Simperl et al., 2020). Knowledge graphs play a central role here, as they facilitate semantic search and recommender systems. In the following, we first outline schemas for modeling data sets' metadata. We then describe existing approaches for data set knowledge graph creation and conclude with an overview of scholarly knowledge graphs in general.

2.1. Standards for Describing Data Set Metadata

To fully exploit the potential of knowledge graphs, interoperability between the knowledge graphs needs to be established (Manola, Mutschke et al., 2019). To achieve this, already widespread RDF standard vocabularies for data set metadata and mappings between them exist. For example, there is a recommended mapping between the vocabularies DCAT and Schema.org (W3C, 2020).

The most important vocabularies in this paper are the following:

- **VoID**. The VoID is an RDF vocabulary for the representation of metadata concerning linked RDF data sets. It is therefore not entirely suitable for modeling metadata from open data portals, as these usually provide resources in different formats (Assaf, Troncy, & Senart, 2015; Neumaier et al., 2017).
- **DCAT**. The DCAT is an RDF vocabulary for the representation of metadata of data sets and data services. The DCAT—Version 2 was published on February 4, 2020 as a W3C recommendation (W3C, 2020). The aim of the W3C is to use the DCAT to solve the problem of heterogeneous metadata schemes in the data portals (Neumaier et al., 2017). The use of DCAT facilitates the interoperability of data set metadata from different data portals. This should make it easier for applications such as search engines to use metadata from different sources. The data set metadata can be published decentrally on the web and still be used for a common search (Assaf et al., 2015).
- **Schema.org**. Schema.org is a collection of schemas for providing structured data on the web. Schema.org's vocabulary can be used with many different encodings, including RDFa, Microdata, and JSON-LD. This structured data enables many applications, such as search engines, to understand the information contained in the web pages. This improves the display of search results in web browsers and makes it easier to find relevant information (Assaf et al., 2015). Schema.org covers many areas. In the context of this work, the data set schema (https://schema.org/Dataset) created on the basis of W3C DCAT is particularly relevant (Brickley et al., 2019).

2.2. Existing Data Set Knowledge Graphs

Table 1 provides an overview of existing knowledge graphs modeling data sets. Overall, one peculiarity of our knowledge graph is that each data set is linked to at least one scientific publication. As a result, the knowledge graph can be used as a data and evaluation basis for new research approaches that have this requirement facilitating sophisticated data set search.

2.2.1. Existing data set knowledge graphs which combine data from several sources

In 2018, Google launched *Google Dataset Search* (https://datasetsearch.research.google.com/) (Brickley et al., 2019), which is based on collecting metadata descriptions of data sets in Schema.org or W3C DCAT from the web. The standardized metadata are processed into a common graph data model, which essentially corresponds to RDF triples. By crawling metadata from the full web, it is inevitable that the metadata corpus will give a significant number of data set representations with incorrect metadata. For example, there are websites that use https://schema.org/Dataset but actually do not contain any metadata for a data set. In contrast to the DSKG, the Google Dataset Knowledge Graph is not part of the Linked Open Data Cloud and not publicly available (Benjelloun et al., 2020; Brickley et al., 2019; Canino, 2019).

Table 1. Overview of related data set knowledge graphs

Knowledge graph with data set metadata	Use-Case	Metadata			Linked Data		Number of data sets	Providing the data
		Data basis	Mapping	Used vocabulary	Data sets linked to each other	Links to external data		
Google Dataset Search knowledge graph[1]	Knowledge graph for the Google Data Set Search	Google Crawler crawls structured metadata	×	Schema.org DCAT	✓	✓	28 million	Knowledge graph is not openly available
Open Data Portal Watch[2]	Add metadata to the web of linked data	261 open data portals	✓	DCAT Schema .org	(✓) only partially	×	854,000	SPARQL endpoint Data dumps (Turtle files)
DataMed[3]	Knowledge graph for DataMed Search	76 life-sciences data portals	✓	DATS	✓	✓	2.3 million	Web API
Ojo and Sennaike (2020)[4]	Improve the recommendation of related data sets	DubLinked	✓	No standard vocabulary	✓	×	205	Not specified
K4OGD[5]	Semantically model government data to enable smart applications	Data portal of the Ministry of Health of Morocco	✓	GovDataset Ontology	✓	×	Not specified	Not specified
Research Graph Data[6]	Increase interoperability and accessibility of data and support linked data	Research Graph database	✓	Schema.org	✓	✓	Not specified	Not specified
DSKG (our knowledge graph)	Knowledge graph for development of semantic search engines and as a linked data source	OpenAIRE and Wikidata	✓	DCAT	✓	✓	2,208	SPARQL endpoint Data dumps URI resolution

[1]Brickley et al. (2019), [2]Neumaier et al. (2016), [3]Ohno-Machado, Sansone et al. (2017), [4]Ojo and Sennaike (2020), [5]Younsi Dahbi, Lamharhar, and Chiadmi (2020), [6]Wang, Aryani et al. (2017).

Open Data Portal Watch (Neumaier, Umbrich, & Polleres, 2016) collects data set metadata from more than 260 freely accessible data portals and focuses on open government data (OGD). The Open Data Portal Watch Framework maps the metadata to the standard DCAT vocabulary. The mapping according to DCAT takes place according to a fixed scheme for the metadata standards of the individual data portals. This creates a uniform representation of the metadata. The Open Data Portal Watch Framework also performs a quality assessment of the metadata. The unified DCAT metadata and their respective quality assessments are available via a SPARQL endpoint. However, the knowledge graph contains hardly any links to external knowledge graphs.

DataMed (Ohno-Machado et al., 2017) contains collected metadata from the field of life sciences from 76 data portals. The collected metadata is transformed into the uniform DATS schema (Sansone, Gonzalez-Beltran et al., 2017) and used for the DataMed search for data sets. The DATS core schema contains core elements that can be applied to any type of data set, as well as advanced elements specifically designed for the field of life sciences. The modeled data sets are linked to publications, software, and data portals. DATS can be mapped to Schema.org elements (Sansone et al., 2017). In contrast to the DSKG, DataMed does not cover all scientific disciplines but is limited to the life sciences.

2.2.2. *Existing data set knowledge graphs which use data from one source*

Ojo and Sennaike (2020) propose an approach to constructing a knowledge graph based on metadata of an open data catalogue. The edges between the data sets of the knowledge graph represent the similarity of the data sets. The similarities between the data sets are constructed using their metadata and the SOM algorithm (Sennaike, Waqar et al., 2017). The knowledge graph is used to enhance the search and recommendation for data sets within a portal. It contains only the 205 Dublin City Council (DubLinked) instance of the CKAN platform and has no links to other data on the web. As no common standard vocabulary is used, the interoperability of the knowledge graph is poor.

Younsi Dahbi et al. (2020) present an approach to constructing a knowledge graph on the basis of freely accessible data from the public health sector. The metadata is transformed into RDF. Established vocabularies and schemes, such as DCAT, are reused and expanded with new properties. In addition, the authors interpret the contents of the data sets and generate RDF data from them, which are expressed in a scheme specially adapted to the public health sector. Thus, both the metadata and the contents of the data sets are represented as a knowledge graph using a domain-specific schema.

Wang et al. (2017) use Schema.org to model research graph data. In particular, the knowledge graph contain data sets, researchers, and scientific publications. The original data of the research graph is not described in a uniform vocabulary ensuring interoperability of the data. By using Schema.org, the data can be made available semantically as linked data.

2.3. Dataset Metadata Collections

Aside from (RDF) knowledge graphs, metadata about data sets have been modeled and provided in various ways. First of all, we can mention initiatives for research data management (RDM), such as *DataCite* with *re3data* (https://datacite.org/re3data.html) and the *Research Data Alliance* (https://rd-alliance.org/) promoting the exchange and reuse of research data sets on an international level. The DataCite Metadata Working Group has published the DataCite metadata schema for the publication and citation of research data (DataCite Metadata Working Group, 2017). Several projects at the national and EU level complement the RDM landscape.

Noteworthy is in this context in particular the *German National Research Data Infrastructure* (NFDI), which is an effort to fund consortia regarding research data management with up to €85 million per year. Furthermore, we can refer to the *Generic Research Data Infrastructure* (GeRDI) (https://www.gerdi-project.eu/), a project funded from 2016–2019 to provide a generic open software platform connecting heterogeneous research data repositories facilitating interdisciplinary and FAIR research data management. Last but not least, during the time of writing this paper, Google Research has published a data set on Kaggle with data set metadata derived from Schema.org (https://www.kaggle.com/googleai/dataset-search-metadata-for-datasets). A similar metadata data set based on Schema.org annotations is considered in our analysis in Section 3.

3. ANALYSIS OF THE DATA SETS

In this section, we analyze data set collections that contain metadata about data sets. We thereby only consider data sets that have values for the basic properties title and description (Benjelloun et al., 2020). Our analysis results are used to assess which data sets are suitable for building a data set knowledge graph that can be used for a variety of use cases, such as data set recommendation.

We came up with the following data sources as being available and relevant for building the knowledge graph.

1. **Wikidata**. Wikidata is a widely used, cross-domain knowledge graph edited by the crowd. It contains instances of data sets modeled by various classes. A list of all relevant classes which represent data sets can be found in our online repository https://github.com/michaelfaerber/data-set-knowledge-graph. The instances of the classes and their properties can be accessed via semantic queries and Wikidata's publicly available SPARQL endpoint. At the time of writing, Wikidata contains 4,209 data set instances.
2. **OpenAIRE**. We also consider a subset of the OpenAIRE Research Graph Dump (Manghi et al., 2019) (as of December 2019). OpenAIRE is an open data portal indexing data sets registered at widely used platforms, such as Zenodo. The used OpenAIRE metadata dump contains the metadata of 23,401 data sets. The OpenAIRE Research Graph Dump is available in several XML files that conform to the OpenAIRE data model.
3. **Schema.org**. Finally, we consider metadata of data sets provided on websites. To this end, we use a subset of the Data Commons project (Web Data Commons, 2018) that contains extracted structured data from crawled websites given in the Common Crawl (https://commoncrawl.org/). It contains all resource descriptions that have the attribute https://schema.org/Dataset. This collection contains 16,962 data set instances in total.

3.1. Degree of Filling of the Attributes

The data set collections contain data set metadata in different quality. In the following, the metadata are evaluated according to the coverage of different information domains. We evaluate the coverage of the information domains based on the Dublin Core Metadata Element Set (DCMES) because the DCAT vocabulary makes extensive use of terms from Dublin Core and the DCMES information domains provide important information about data sets (W3C, 2020). DCMES describes 15 core fields which provide essential metadata about resources (NISO, 2004).

The 15 core fields can be summarized in six overarching information domains:

- **Date**: For each data set, we store the creation date, modification date, blocking period, and similar dates. In the DCMES this information is stored in the core field `date`.
- **People and Organizations**: We can model the name of the persons or organizations who were involved in the creation or publication of the resource. In DCMES this information is represented as the core fields `creator`, `publisher`, and `contributor`.
- **Description of the content**: In the DCMES, the information about the resource and its content is modeled by means of the core fields `title`, `description`, `subject`, and `coverage`.
- **Technical data**: In the DCMES, the information concerning the technical nature of the data is modeled in the core fields `format`, `type`, and `language`.
- **ID**: In DCMES, unique identifiers for resources and web links are stored in the core field `identifier`.
- **Rights**: In the DCMES, information regarding the property rights related to the resources is stored in the core field `rights`.

We evaluate the three data set collections with respect to the availability and degree of filling of attributes that contain information on these information areas. Table 2 shows the assignment of the attributes of the data sets to the information domains and the degree of filling of the attributes. It serves as an overview of to what extent the data sets cover the respective information domains. More detailed information regarding the coverage of the information domains by single properties can be found in our repository online.

We can observe that OpenAIRE has attributes with the highest filling degree to cover the individual information domains. Only when describing the data sets' content is specific information missing, such as the spatial coverage of the data set. The attributes of the Wikidata subset have the second highest filling degree to cover the information domains. As with the OpenAIRE subset, Wikidata does not contain information about the spatial coverage of the data sets. However, Wikidata covers all other information areas better than the Schema.org data set collection. Although the Schema.org collection contains information on the spatial coverage of the data sets, it has the lowest degree of filling, averaged over all metadata, and therefore covers the information domains the least.

3.2. Qualitative Evaluation

We also carried out a manual evaluation to determine which scientific discipline a data set is to be assigned to and whether the metadata entries are actually valid data sets. For this purpose, 100 randomly selected data sets are assessed manually as to whether they are valid data sets and to which scientific discipline they belong. According to the definition of DCAT, a data set is a collection of data that is published or managed by a single agent and is available in one or more representations for access or download (W3C, 2020). Thus, in our assessment, metadata entries (i.e., items describing data sets) were judged as nonvalid data sets if they are pure data portals, pure software, or only websites containing information but not offering a data set download or other kinds of data set access.

3.2.1. Proportion of valid data sets

The results of the manual evaluation are as follows. Out of 100 data set representations, Wikidata contained 86, OpenAIRE 100, and Schema.org 68 valid data sets. In the case of Schema.org, a resource can belong to the class https://schema.org/Dataset even though it is not a

Table 2. Coverage of data sets (with title and description) regarding the information domains

	Wikidata	OpenAIRE	Schema.org
Date	date Modified (100%)	dateofacceptance (76.98%)	datePublished (22.32%)
	publication date (28.15%)	storagedate (76.97%)	dateModified (18.95%)
	inception (17.68%)	relevantdate (69.60%)	temporalDataCoverage (13.92%)
		lastmetadataupdate (51.85%)	dateCreated (10.97%)
		embargoenddate (9.77%)	
People and Organizations	author (24.85%)	creator (99.95%)	author (16.34%)
	author name string (8.32%)	publisher (76.81%)	creator (10.97%)
	maintained by (29.46%)	contributor (14.10%)	publisher (16.25%)
	sponsor (24.59%)	contactperson (10.33%)	provider (13.55%)
	operator (8.34%)		includedInDataCatalog (6.68%)
	owned by (7.75%)		
Content description	label (100%)	title (100%)	name (100%)
	description (100%)	description (100%)	description (100%)
	main subject (13.02%)	subject (98.21%)	keywords (24.40%)
			spatialDataCoverage (14.19%)
			spatialData (10.52%)
Technical data	language of work or name (18.63%)	format (62.45%)	inLanguage (7.88%)
		language (34.25%)	
ID	official website (59.92%)	doi (77.01%)	url (44.18%)
	full work available at URL (26.99%)	originalId (44.83%)	identifier (21.82%)
	Freebase ID (7.03%)	documentationUrl (8.12%)	accessURL (9.50%)
Rights	copyright license (7.60%)	bestaccessright (25.23%)	license (11.99%)

data set. Such incorrect entries occur, for example, when a website contains a https://schema.org/Dataset description, although it does not describe any data set metadata. It can therefore not be guaranteed that all metadata entries actually describe data sets. The resulting problem of many incorrect entries is an unsolved problem and known in research (Benjelloun et al., 2020).

3.2.2. Scientific disciplines

We manually assigned scientific disciplines to each of the 100 randomly selected data sets in order to determine the discipline coverage. For this purpose, we reused the set of disciplines

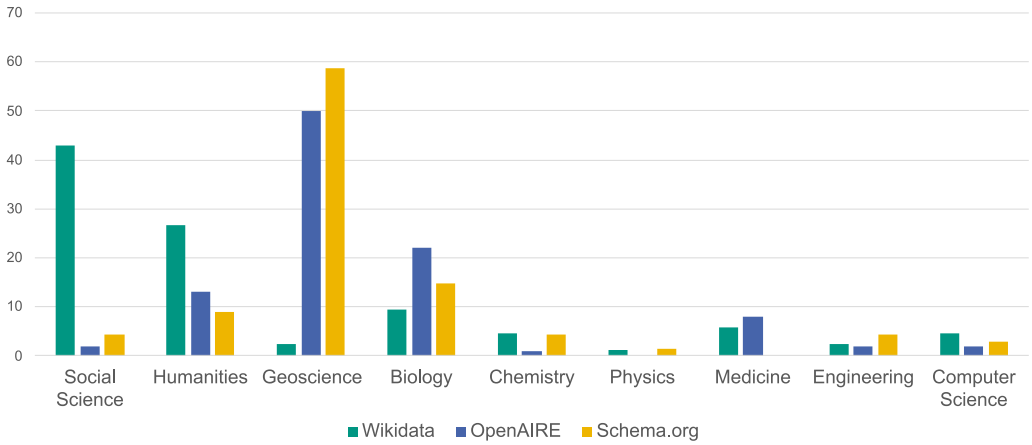

Figure 1. Coverage of the scientific disciplines.

used by Benjelloun et al. (2020) for the analysis of the Google Dataset Search data corpus. To determine the scientific discipline of a data set, we used not only the metadata entries available but also, as far as available, the web pages on which the data sets are available online. To be able to compare the data sets as intuitively as possible, we assigned each data set only to its main discipline and omitted possible double assignments.

The results of our analysis are shown in Figure 1. We can see that the resources of the three considered data set collections cover the scientific disciplines to varying extents. The data sets of Wikidata largely cover the disciplines of the humanities and social sciences. The data sets of OpenAIRE and Schema.org, however, mainly cover the natural sciences. In particular, the geosciences as well as biology and agricultural science are disproportionately represented. Such an imbalance in the domains can be observed for existing cross-domain knowledge graphs as well (Färber, Bartscherer et al., 2018).

3.3. Quality of the Metadata Entries

We also analyzed the data set titles and descriptions in the data set collections. The average number of words of the data set descriptions is shown in Table 3. Table 4 shows the average number of words in the data set title.

We observe that the titles and descriptions of entries in Wikidata are significantly shorter than the entries in OpenAIRE and Schema.org. In Wikidata it is intended that descriptions of resources are kept short. The descriptions are mainly used to disambiguate resources (Vrandecic, 2019). If a longer description is required for an application, the descriptive section of the Wikipedia article or the official website of the resource can be used. Of the 4,209 data sets from Wikidata, 901 (21.4 %) have an English Wikipedia entry. The listed websites of the data sets is available for 2,522 (59.9%) data sets.

3.4. Bottom Line

We can summarize our analysis results as follows:

- The **Wikidata** data set collection contains the fewest resources and the data set titles, like the data set descriptions, are kept short. However, it can be assumed that many more data sets will be added to Wikidata in the future. The metadata of the data sets can be called up via the freely accessible SPARQL endpoint.

Table 3. Number of words in the data set description

	Wikidata	OpenAIRE	Schema.org
Average	5.1	117.2	126.6
Median	4	48	66
0.25 quantile	1	17	26
0.75 quantile	7	156	221
Maximum value	68	9311	1533
Minimal value	1	1	1

- The **OpenAIRE** data set collection covers to the highest degree the information areas examined. Furthermore, no incorrect entries were found in the manual evaluation for this data set collection.
- The **Schema.org** data set collection covers the examined information domains the least and contains many entries that are not data sets.

To provide a good data basis for novel research approaches, it is important that the knowledge graph contains as few incorrect entries as possible. The many incorrect entries in the Schema.org data set would reduce the quality of the knowledge graph and represent a problem that should not be neglected. Thus, we decided to use Wikidata and OpenAIRE to build a knowledge graph with the lowest possible proportion of incorrect entries. The knowledge graph is therefore particularly suitable as a data basis for research approaches that *focus on high precision and less on high recall*.

4. APPROACH TO CREATING THE DATA SET KNOWLEDGE GRAPH

In the following section, we show our approach to creating a knowledge graph for data sets. The overall approach to the structure of the knowledge graph is sketched in Figure 2. We can differentiate between the following steps:

1. The data set metadata used are originally in tabular form. First, we link the data sets to the publications they reference using a string-matching algorithm. The procedure is described in detail in Section 4.3.

Table 4. Number of words in the data set title

	Wikidata	OpenAIRE	Schema.org
Average	4.6	10.7	8.6
Median	4	9	7
0.25 quantile	2	6	4
0.75 quantile	6	14	12
Maximum value	24	485	103
Minimal value	1	1	1

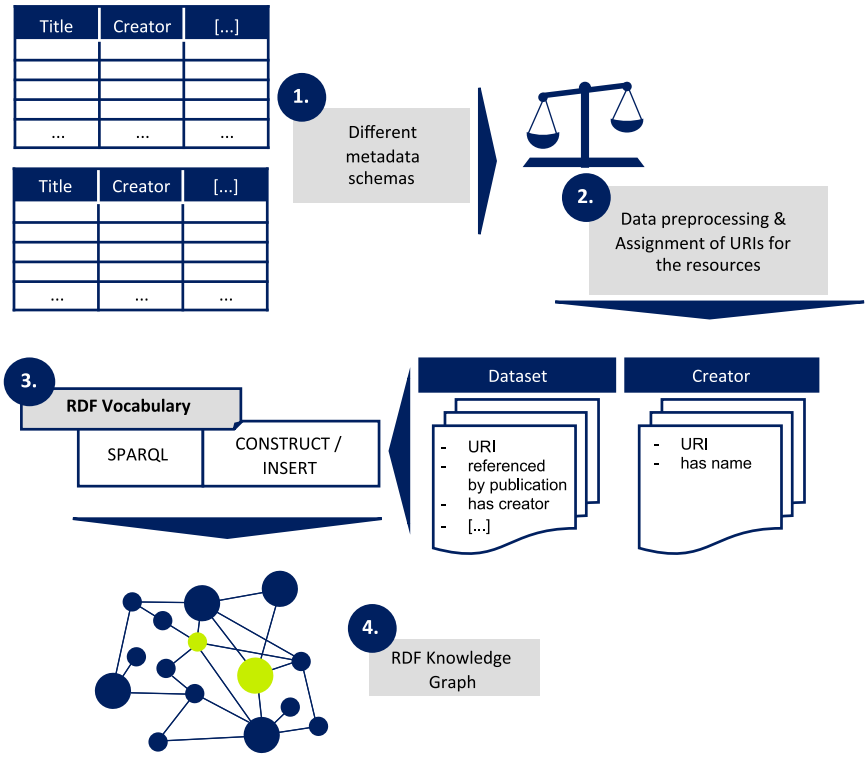

Figure 2. Approach to create our data set knowledge graph.

2. Next, the metadata entries are prepared and cleaned up so that they meet the requirements of the RDF target vocabulary. This includes, among other things, a classification of the resources contained in the metadata and an extensive author disambiguation. This step is described in Section 4.4.
3. We map the processed metadata in the RDF standard vocabulary DCAT, with which the knowledge graph is created.
4. The result of running our approach is an RDF knowledge graph based on the four design principles of linked data (Heath & Bizer, 2011). By mapping the metadata to an RDF vocabulary and linking to other data sources on the web, the knowledge graph created is a five-star data set as defined by Tim Berners-Lee (Heath & Bizer, 2011).

In the following, we describe the knowledge graph schema and the single steps in more detail.

4.1. Knowledge Graph Schema

We developed a schema of our knowledge graph as depicted in Figure 3. The figure shows the entity types present in the knowledge graph and their properties. The reused vocabularies and their corresponding prefixes are also given. The elements that are literals and their corresponding data type are indicated in a node with a green background. Elements that are clearly identified by a URI are indicated in a node with a yellow background. The labeled edges between two nodes indicate the relationship between the nodes.

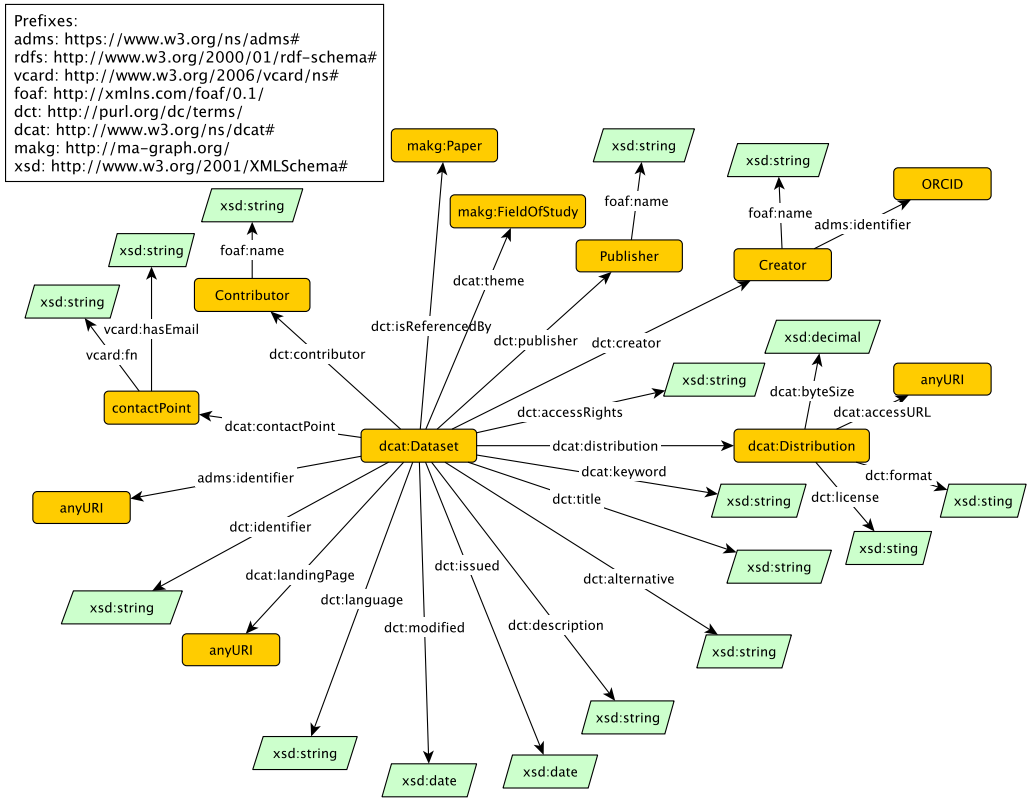

Figure 3. Schema of the data set Knowledge Graph.

We reused the W3C DCAT vocabulary to build the knowledge graph. In this way, we are reusing a widely used standard vocabulary, which allows us to model metadata from different sources in a standardized way (see Section 2.1). The knowledge graph schema covers six different entity types, such as data sets and authors. In our knowledge graph, a data set describes a metadata instance that contains a data set distribution and a number of other properties for describing the data set. A data set distribution is a specific representation of a data set. The data set distribution thus represents the accessible form of the data set. This can be a downloadable CSV file, for instance. It also contains properties for license information and general information describing the resource (Neumaier, 2019; W3C, 2020). Listing 1 shows the metadata of an example data set in DCAT.

4.2. Mapping of the Metadata According to DCAT

Our DSKG is created based on the data set collections as outlined and analyzed in Section 3. As the initial data set collections contain the metadata in tabular form, the first step in our framework is to transform the data into RDF using the knowledge graph schema (i.e., ontology) outlined above.

4.2.1. Mapping of the metadata properties

As the next step, we define mapping rules to map the metadata properties and the metadata is mapped to the RDF vocabulary DCAT. The mapping rules for the metadata properties are

```
:dataset-588
    a dcat:Dataset ;
    dct:title "Arabic␣Handwritten␣Digits␣data␣set"@en ;
    dcat:keyword "ComputingMethodologies"@en ;
    dct:isReferencedBy <http://ma-graph.org/entity/1990665784> , <http://ma-graph.org/
        entity/2553946018> ;
    dct:creator :creator-10257 ;
    dcat:distribution :distribution-100588 .
:distribution-100588
    a dcat:Distribution ;
    dcat:accessURL <https://www.kaggle.com/mloey1/ahdd1> ;
    dct:format "csv" ;
    dcat:byteSize "2103733.0"xsd:decimal .
:creator-10257
    a foaf:Person ;
    foaf:name "Mohamed␣Loey"@en .
```

Listing 1. Example of an RDF serialization of a represented data set.

shown in Table 5. Note that Schema.org is not included as it is not considered as a data source based on our data analysis in Section 3. We generate the mapping rules based on the following observations:

- **Wikidata data model**: The project *WikiProject data sets* (https://www.wikidata.org/wiki/Wikidata:WikiProject_Datasets) deals with the coordination and improvement of data set descriptions in Wikidata. As a result, mapping rules between the Wikidata data model and DCAT were drawn up (*WikiProject Datasets/Data Structure/DCAT – Wikidata – Schema.org mapping*, 2018). Denny Vrandecic presents mapping rules between the Wikidata data model and Schema.org, which are equivalent to DCAT (Vrandecic, 2019). We use these existing mappings of the Wikidata data model to DCAT. Because the above-mentioned documents were working drafts at the time this work was being carried out, not all classes and properties have been finalized. For example, the drafts do not contain a mapping for the Wikidata property Also known as. For missing mappings, we analyzed whether two properties with the same name exist in the data models and whether they can be mapped onto one another.
- **OpenAIRE data model**: The OpenAIRE Research Graph data model is inspired by several existing metadata standards. In particular, the *DataCite metadata schema* is reused to describe data sets (Manghi et al., 2019). A draft exists that describes how DataCite metadata can be mapped in a DCAT-compliant representation (Perego, Austin et al., 2020). The illustrations of the OpenAIRE Research Graph data model according to DCAT used by us are based on the assignments in this draft.

4.2.2. Preprocessing of metadata entries for the DSKG

We need to adapt the metadata entries according to the requirements of DCAT. DCAT defines the data type and the data format (W3C, 2020) for literals. Thus, if necessary, we adapt the metadata entries to the prescribed data formats. The size of a data set distribution is specified in DCAT in bytes. We therefore convert the size specifications of the metadata entries into bytes. Furthermore, the data available in the metadata are converted uniformly into the ISO 8601 standard used by DCAT for the representation of date and time information.

4.2.3. Mapping of the data set instances

In order not to have duplicate entries of data sets in the knowledge graph, the intersection of duplicate data sets of the OpenAIRE data set and the Wikidata data set is determined. We

Table 5. Mapping of the metadata schema according to DCAT

DCAT URI	Representation in Wikidata	Representation in OpenAIRE
dct:description	description	description
dct:title	label	title
dcat:keyword	mainSubject (P921)	subject
dct:alternative	altLabel	–
dct:issued	publication date (P577)	storagedate
dct:modified	dateModified	relevantDate
dct:language	language (P407)	language
dcat:landingPage	official website (P856) workURL (P963) wikipedia article	originalId
dct:identifier	Catalog code (P528)	objIdentifier
adms:identifier	wikidata item identifier	doi
dct:accessRights	–	bestaccessright
dct:publisher	publisher (P123)	publisher
dct:creator	author (P50) author name string (P2093)	creator
dct:contributor	contributor to the creative work or subject (P767)	contributor
dcat:contactPoint	–	contactPerson
dct:license	licence (P275)	–
dct:format	file format (P2701)	format
dcat:accessURL	url (P2699)	originalId
dcat:byteSize	data size (P3575)	size

considered two data sets as duplicates if the term frequency of their titles has a cosine similarity of over 0.9. Data sets that are in the intersection are only added to the knowledge graph once. Due to the relatively low number of data sets that are linked to a publication (see Section 4.3), only one data set is duplicated in the two sources considered.

4.2.4. *Data transformation*

The transformation of the metadata in tabular form into a knowledge graph takes place with the help of operations from SPARQL (W3C, 2013a) and SPARQL 1.1 Update (W3C, 2013b).

The built knowledge graph is described with the help of SPARQL CONSTRUCT clauses. In the WHERE clause, the metadata is extracted from the tables and assigned to the variables in the CONSTRUCT clause. The assignment takes place according to the mapping rules defined in Table 5. To comply with the W3C standards and the design principles of linked data, the resources contained in the metadata are designated with URIs. The implementation of the semantic modeling of the resources is given in our repository.

4.3. Linking the Data Sets to Scientific Publications

Adding links to other data sources is another important step in integrating the knowledge graph into the Linked Open Data Cloud and facilitating important novel use cases, such as data set recommendation for given publications. In many scientific publications, data sets are mentioned that scientists used or created for their research (Gregory, Cousijn et al., 2020; Henderson & Kotz, 2015). Therefore, we link the data sets to scientific publications in which they are mentioned. In our knowledge graph, the data sets refer to the publications (W3C, 2020) via the property dct:isReferencedBy. The MAKG is used as the data basis for the scientific publications because the MAKG covers all scientific disciplines and is one of the largest freely available scholarly knowledge graphs. It contains 210 million publications (Färber, 2019).

The MAKG contains 146 million publications' abstracts and 241.5 million citation contexts (i.e., sentences in which other publications are cited via citation markers). We searched for mentions of data sets in these publications' abstracts and citation contexts to create links between data sets and publications. We use a string-based algorithm for that purpose (see our GitHub repository). The titles of the data sets, alternative titles and the data set IDs are used to identify data sets in the scientific publications. For data sets listed in OpenAIRE, we used the attributes `title`, `doi`, and `originalId`. For data sets listed in Wikidata, we used `label`, `altLabel`, `official website`, `work URL`, and `url`.

To minimize the number of incorrect links, some preprocessing steps are carried out before applying the matching algorithm. In order not to distort the comparison by meaningless data set titles, such as `Dataset`, `Language`, or `README`, data set titles only are used for the comparison if they are not among frequently used English words. We use the English Word Frequency data set (Tatman, 2017) to filter out such titles. To consider the different nature of the metadata entries from the different sources (see Table 4) in the comparison, a case-sensitive comparison is performed for the data sets listed in Wikidata for the data set titles and the alternate titles. In addition, only data set titles that have a minimum length of four letters are considered. Alternative data set titles are only considered if they have a minimum length of five letters.

We link 2,208 data sets to 634,803 scientific publications in the MAKG. The data sets are mentioned in the abstracts or citation contexts of the linked publications and are therefore an essential aspect of these publications. Some 588 data sets originate from OpenAIRE and 1,620 data sets from Wikidata. More links are found to publications for data sets from Wikidata, because on the one hand, alternative titles are available for the comparison in addition to the data set title. Furthermore, the data set titles in Wikidata are shorter (see Table 4). As Wikidata enables its users to provide metadata in a decentralized manner, many known data sets from third parties have been registered in Wikidata (Vrandecic, 2019). These known data sets are often mentioned in publications. For example, the 10 data sets with the most linked publications are all from Wikidata. Table 6 provides an overview of the 10 data sets with the most linked publications.

Table 6. The 10 data sets with the most linked publications

Data set URL	Data set title	# Linked publications
http://dskg.org/entity/dataset/1051	Embase	101,139
http://dskg.org/entity/dataset/733	GenBank	63,946
http://dskg.org/entity/dataset/777	Web of Science	54,878
http://dskg.org/entity/dataset/667	Scopus	41,084
http://dskg.org/entity/dataset/758	Cochrane Libary	39,573
http://dskg.org/entity/dataset/1018	CINAHL	33,523
http://dskg.org/entity/dataset/1079	PsycINFO	32,001
http://dskg.org/entity/dataset/1510	ImageNet	25,520
http://dskg.org/entity/dataset/1276	Scope	19,186
http://dskg.org/entity/dataset/621	Gene Ontology	17,852

The distribution of the number of links to scientific publications is shown in Figure 4. We can clearly see that the number of linked publications of the data sets is very unbalanced and that most of the data sets are associated with around 10 to 1,000 publications. We argue that, in particular, this long tail of data sets makes it difficult for researchers and data scientists to be aware of all relevant data sets for a given field and that data set recommendation based on our knowledge graph is a promising avenue for future work. The distribution of the number of data sets to the number of publications in which they are mentioned is shown in Table 7. As we can see, the vast majority of publications only mention very few data sets. The publications that are included in the DSKG mention on average 1.28 data sets.

In total, linking the data sets to scientific publications results in 813,551 links from the 2,208 DSKG data sets to 634,803 unique publications represented in the MAKG.

4.3.1. Fields of application of the data sets

The fields of application of the data sets are useful for domain-specific data set search and recommendation. The fields of application can be derived from the linked scientific

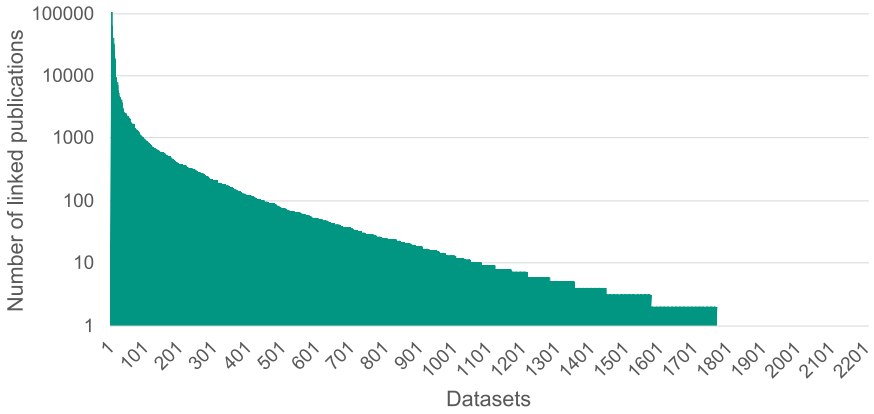

Figure 4. Distribution of the number of linked publications using a log scale.

Table 7. Distribution of the number of data sets to the number of publications in which they are mentioned

# Datasets	1	2	3	4	5	6	7	8	9	10–20
# Publications	510,524	85,537	27,572	8,175	2,102	584	165	69	28	47

publications. The publications in the MAKG have the property https://purl.org/spar/fabio/hasDiscipline, which has a resource of the class http://ma-graph.org/class/FieldOfStudy as range. The class contains 19 different subject areas. The specialist areas of the publications that reference the data sets are therefore known and can be added to the metadata of the data sets. The dcat:theme property is used to model the application areas, as DCAT does not provide any property that specifically describes the fields of application area of a data set. This property specifies a subject area of a data set. A data set can be assigned to several fields. A data set is assigned to an application area if at least 25% of its linked publications are assigned to this subject area. The distribution of the fields of application of the data sets in the knowledge graph is described in Section 5.

4.4. Semantic Representation of the Metadata

A fundamental idea of linked data is the use of URIs to name resources (W3C, 2014). This enables resources to be linked to other things on the web. To provide the metadata according to the Linked Data principles, we therefore transform the strings representing resources as URIs. This is achieved by classifying strings and resolving word ambiguities.

- **Classification of resources**: The resources from the OpenAIRE data set are only available as a character string but not as URIs and not with associated entity type. Also, the Wikidata data set contains resources that are not semantically represented (e.g., the author name string). To transform the character strings into semantic resources in the knowledge graph, we automatically determine the class of each resource in the first step. We use the Python library spaCy (https://spacy.io) to determine whether a character string represents a person (`foaf:Person`, e.g., "Menghan Hu") or an organization (`foaf:Organization`, e.g., "US Department of Health and Human Services"). If the class of a character string remains unknown, it is modeled as `foaf:Agent`.
- **Resolving ambiguities**: The next step is to resolve ambiguities (i.e., identifying strings representing the same real-world object) so that we can assign unique URIs to the resources. Overall, our knowledge graph contains 1,169 resources of type `foaf:Person`, 246 resources of type `foaf:Organization`, 102 resources of type `foaf:Agent`, and 19 resources of type `vcard:Kind`. Due to the negligibly small number of ambiguous names for entities of the classes `foaf:Organization`, `foaf:Agent`, and `vCard:Kind`, no automated disambiguation and resolution of ambiguities is necessary in these cases. The resources of the class foaf:Person are called authors in the following. Due to the large number of entries, an author name disambiguation is necessary.
- **Author name disambiguation**: The aim of the author name disambiguation is to resolve the ambiguities of the author names and to clearly name authors using URIs. In our case, an author should only be assigned the data sets that they have actually published. The challenge is that different authors can have the same name on the one hand (homonyms) and on the other hand an author can publish under different names (synonyms) (Tekles &

Bornmann, 2019). Figure 5 highlights that it is often nontrivial to disambiguate authors mentioned in different data set collections.

Tekles and Bornmann (2019) compare in their work several approaches of unsupervised learning for author name disambiguation in literature databases. The result of their investigation shows that the approach of Caron and van Eck (2014) based on rule-based evaluations and clustering delivers the best results. It is based on the assumption that the likelihood that two publications are written by the same author is higher the number of shared bibliographic elements between the two publications. To reduce the number of pairwise comparisons, only authors of publications whose names have a certain similarity are compared. If the compared authors of two publications exceed a certain threshold, the authors are clustered using single linkage clustering. It is then assumed that the publications were written by the same author (i.e., in our case, that data sets were published by the same person).

Based on the approach of Caron and van Eck (2014), we have developed a rule-based approach for author name disambiguation that is adapted to the metadata of data sets. To calculate the similarity between two author names and, thus, to know the candidates for author name disambiguation, we use the Jaro-Winkler similarity. This is often used to calculate the similarity of short strings, especially for personal names (Bilenko, Mooney et al., 2003). In our implementation, two authors are compared with each other if their names have a Jaro-Winkler

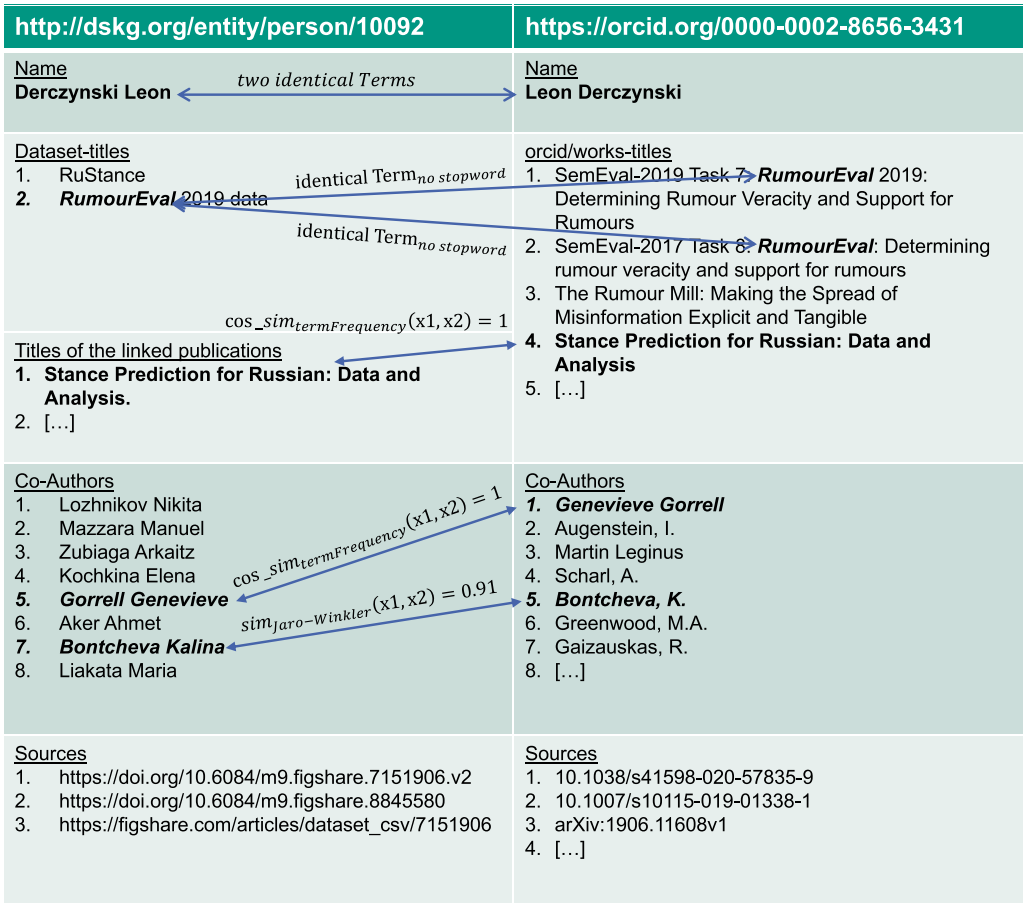

Figure 5. Comparing two author profiles.

similarity of at least 0.9 following Donner (2014) and Hajra, Radevski, and Tochtermann (2015). Concerning the rules for author name disambiguation, we use factors that have already proven to be reliable in the literature and which are rated according to their (Caron & van Eck, 2014; Cen, Dragut et al., 2013; Dendek, Bolikowski, & Lukasik, 2012; Protasiewicz & Dadas, 2016).

The rules used for author name disambiguation are based on different types of confirmation. They can be divided into two categories. On the one hand, we can use explicit information that comes directly from the data set metadata. On the other hand, we can use implicit evidence derived from the data set metadata (Ferreira, Gonçalves, & Laender, 2012; Protasiewicz & Dadas, 2016). One technique for finding implicit evidence is to identify latent topics of data sets shared by the data set authors. LDA (Blei, Ng, & Jordan, 2003) is among the most popular techniques to obtain latent topics. We create an LDA model for the data sets from the combined titles, descriptions, and keywords. We use 10 topics for our LDA model because it has been shown in the literature that an LDA model as an implicit attribute for author disambiguation with 10 topics usually achieves the best results (Song, Huang et al., 2007). When calculating the similarity of two authors, the cosine similarity between the LDA vectors of the underlying data sets is considered as one of the criteria. This counteracts the problem that there are data sets that only have values for a few properties and thus only a few bibliographic elements can be used for the comparison.

Table 8 provides an overview of the rules used in our approach, together with their weighting and the threshold value. Overall, we use the following rules:

- The first rule checks the initials of the authors of two data sets.
- The second rule checks that the adjusted first names are equal. It is assumed that the authors' names are in the form `first name last name`.
- The third rule evaluates the number of joint coauthors between two data sets.
- The fourth and fifth rules check whether the data sets have common publishers and contributors.
- The sixth rule checks the titles of the data sets for common words.
- With the application of the seventh rule, the years of publication of the data sets are considered.
- In the eighth rule, the fields of application of the data sets are compared.
- The ninth rule checks the cosine similarity of the LDA vectors of the data sets calculated using the LDA model.

Two compared authors names are rated as being identical if they exceed the given threshold ($\theta >= 11$, following Caron and van Eck (2014)).

The results and evaluation of the performed author disambiguation are described in Section 5.3.

4.5. Linking of the Data Set Authors to ORCID

The transformation of the metadata in RDF as linked data opens up new possibilities to link the resources in the knowledge graph to other resources on the web. In particular, the use of the ORCID record of researchers offers added value. ORCID provides a register with persistent and unique identifiers for the unique identification of scientific authors (Haak, Fenner et al., 2012). It was designed to solve the problem of author name disambiguation for scientific publications (Caron & van Eck, 2014). We can thus use ORCIDs as ground truth for evaluating our author

Table 8. Criteria for author name disambiguation

Rule	Meta-information	Criteria	Score
1a	Author initials	Two identical initials	3
1b		More than two identical initials	7
1c		Different initials	−5
2a	Authors first names	General first names	2
2b		Nongeneral first names	5
3a	Same coauthors	One	4
3b		Two	7
3c		More than two	10
4	Same publishers	At least one	1
5	Same contributors	At least one	3
6a	Similarity title	One identical word	2
6b		Two identical words	4
6c		More than two identical words	6
7	Year of publication	Not more than 10 years difference	1
8	Fields of application	At least one identical field of application	1
9a	$cos - sim_{LDA}\ (d_1, d_2)$	>0.75	2
9b		>0.95	3
9c		>0.99	4
		Threshold	11

name disambiguation approach. As of December 2020, ORCID (https://orcid.org) has issued more than 10 million author identifications. The authors in our knowledge graph are enriched by links to their ORCID record.

To determine whether an author in our knowledge graph is identical to a registered author at ORCID, we compare the authors by means of a variety of features, such as the author's name, the titles of the author's publications, and the coauthorship of authors, as well as the identifiers of publications (Hajra et al., 2015; Radevski, Hajra, & Limani, n.d.). A challenge here is that scientific authors usually only list their scientific *publications* on their public ORCID record, but not published *data sets* (see Table 13). Therefore, in addition to the titles of the published data sets by the authors, we consider the titles of the linked publications of the data sets. Because there are over 10 million records in ORCID, of which many records are incomplete, it is a nontrivial task to identify the correct ORCID record for an author. Therefore, we use the following rules for comparing the authors:

1. **Similarity between the names of the author**: Two author profiles are only compared with each other if the names of the author profiles have at least two identical terms (strings).

2. **Similarity of the titles of the publications**: The titles of the data sets of the author profile from the knowledge graph are compared with the titles of the published works of the author profile by ORCID. To enable comparison between data set titles and publication titles, the number of identical words, which are not stop words, in the titles is counted. For one identical word a rating of 1 is given, for two a rating of 2 and for more than two a rating of 4.
   The titles of the linked publications of the data sets are compared with the titles of the published works by ORCID. There is a match for two compared titles if, according to the cosine similarity, the following applies:

$$sim_{\cos,TF}(titel_1, titel_2) > 0.75 \tag{1}$$

   If there is at least one match, a rating of 4 is given.
3. **Matches in the lists of coauthors**: We automatically compare the list of coauthors of the author profiles. There is a positive match for two names from the lists if at least one of the two similarities applies:

$$sim_{\cos,TF}(name_1, name_2) = 1 \tag{2}$$

$$sim_{Jaro-Winkler}(name_1, name_2) > 0.9 \tag{3}$$

   If there is a match, a rating of 3 is given. For two or more matches, a rating of 4.
4. **Similarity of the identifiers of the publications**: The identifiers of the author profiles are compared with one another. If at least one identical identifier is available, a rating of 4 is given.

If two author profiles achieve a rating of at least 4 points, it is assumed that they are the same author and a link to the author's ORCID record is added in the knowledge graph. This rating to be achieved is determined experimentally. If the rating is too low, we obtain many false positive hits. If the value is too high, many true positive hits will not be recognized. In accordance with DCAT, we use the adms:identifier property to integrate the links to ORCID.

Figure 5 shows an example of linking a DSKG author to a modeled author in ORCID. In this case, the threshold value is exceeded and a corresponding link to the ORCID record is set.

4.6. Data Set Entity Embeddings

Apart from creating and providing the DSKG via data dumps, URI resolution, and a public SPARQL endpoint, we computed embeddings for all data sets modeled in the DSKG. Entity embeddings have proven to be useful as implicit knowledge representations in a variety of scenarios (see Section 6). Because the DSKG is available in RDF, we applied RDF2Vec (Ristoski et al., 2019) to the DSKG using the skip-gram model, a random walking strategy, a window size of 5, 128 dimensions, and 10 epochs of training. The resulting embedding vectors for all data sets in the DSKG are available at http://dskg.org.

5. STATISTICAL ANALYSIS AND EVALUATION

In this section, we present statistical key figures of the knowledge graph as well as the results of the evaluation of the author name disambiguation method and the semantic enrichment of data set authors.

5.1. Key Statistics

The created knowledge graph contains 2,208 data set instances and 813,551 links to 634,803 scientific publications in the MAKG. The knowledge graph consists of 850,288 RDF triples. Table 9 provides an overview of the proportion of data sets and data set distributions with certain properties.

The data set representations in the knowledge graph are enriched via the properties `dct:isReferencedBy` and `dcat:theme`. In case of the DSKG, all data sets have at least one link to a publication. In contrast, the data set collection underlying the Google Dataset Search

Table 9. Share of data sets and data set distributions with certain properties

Class	Property	Total	Percentage
dcat:Dataset	dct:title	2,208	100%
	dct:isReferencedBy	**2,208**	**100%**
	dcat:distribution	2,208	100.0%
	dct:description	2,207	99.9%
	adms:identifier	1,989	90.1%
	dcat:theme	**1,922**	**87.0%**
	dcat:landingPage	1,837	83.2%
	dct:modified	1,620	73.4%
	dct:alternative	999	45.2%
	dcat:keyword	813	36.8%
	dct:identifier	588	26.6%
	dct:creator	518	23.5%
	dct:language	449	20.3%
	dct:issued	432	19.6%
	dct:publisher	393	17.8%
	dct:accessRights	246	11.1%
	dcat:contactPoint	21	0.9%
	dct:contributor	11	0.5%
dcat:Distribution	dcat:byteSize	543	24.6%
	dcat:accessURL	387	17.5%
	dct:format	316	14.3%
	dct:licence	171	7.7%

contains 28 million metadata entries and is the largest snapshot of data sets on the web. However, less than 1% of the data sets in this corpus have a property that links the data sets to scientific publications. Only 20.9 % of the data sets have a property that indicates the subject area or something similar of the data set (Benjelloun et al., 2020).

Another feature of our knowledge graph is the high proportion of modeled data sets with the property `dct:alternative`. Although 45.2% of the data sets in the knowledge graph have this property, it is only 3.4% in the data set corpus of Google Dataset Search. The reason for this is that the Also known as (altLabel) property is widely used in Wikidata. Especially for string-based algorithms, such as the linking of the data sets to scientific publications (see Section 4.3), alternative data set titles offer added value. In the data set corpus of Google Dataset Search, only the publisher and provider of the data set are given for the data sets. A total of 84.59% of the data set descriptions contain this information. The authors of the data sets, however, are only given for 14.12% of the data sets. The share of data sets in the knowledge graph for which an author is specified is, at 23.5%, significantly larger than in the data set corpus of the Google Dataset Search. This is due to the fact that the specification of the authors is standard in OpenAIRE, whereas the publisher is specified less often. In Wikidata, too, the publisher of the data sets is rarely given. As a result, only 17.8% of the data sets have the `dct:publisher` property.

5.2. Areas of Application of the Data Sets

The areas of application of the data sets are determined on the basis of their linked scientific publications in the MAKG using the MAKG's fields of study. No application area can be determined for 286 data sets of our knowledge graph. Given these data sets, at least 25% of the linked publications are not assigned to any subject area (i.e., field of study). One application area can be determined for 1,492 data sets, two application areas for 383 data sets, three application areas for 41 data sets, and four application areas for six data sets.

Figure 6 gives an overview of the distribution of the application areas of the data sets based on the associated publications' fields of study. The relatively high coverage of computer

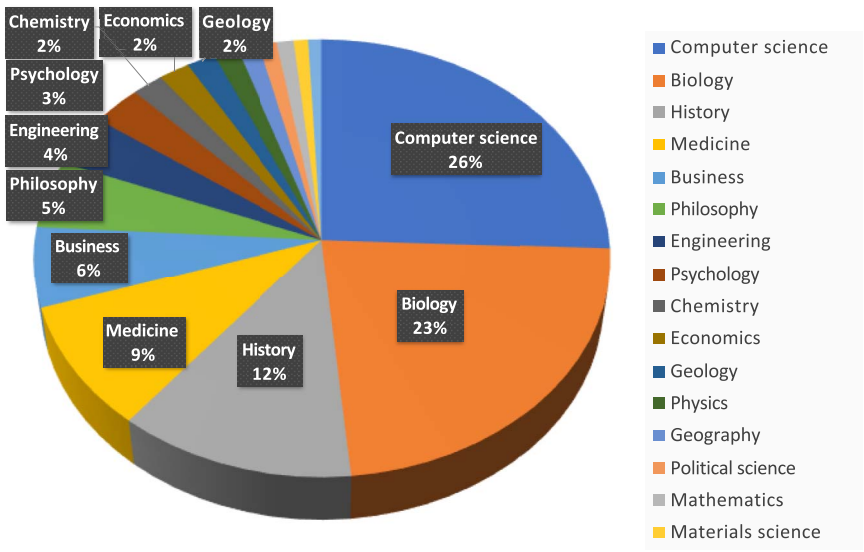

Figure 6. Distribution of application areas of data sets in the DSKG.

Table 10. Confusion matrix with LDA model

	Actually identical authors	Actually nonidentical authors
Predicted identical authors	True Positive 129	False Positive 0
Predicted nonidentical authors	False Negative 2	True Negative 16

science and biology—compared to disciplines such as engineering and medicine—can be traced back to the fact that the number of publications using data sets in these disciplines is higher and therefore we can add more links to publications in these disciplines. Furthermore, Wikidata and OpenAIRE contain a relatively high number of biological data sets (see Figure 1). The fact that many biological data sets are published is also reflected in the corpus of the Google Dataset Search. A total of 15.2% of the data sets of the corpus belong to this discipline (Benjelloun et al., 2020).

5.3. Evaluation of Author Name Disambiguation

Our DSKG without author name disambiguation contains 1,298 author names. Of these, 1,190 originate from OpenAIRE and 108 from Wikidata. Only 56 authors from Wikidata are semantically represented and have a URI. The remaining 52 authors from Wikidata are only specified as a character string, as in OpenAIRE (attributes of the property https://www.wikidata.org/prop/P2093 – `author name string`). When performing the author name disambiguation, 147 comparisons are made. For these 147 compared author names, we carried out a manual evaluation to determine author names referring to the same real-world entity. We evaluated the results of the author disambiguation both with the added LDA model and without the LDA model. The other criteria and their evaluations as well as the threshold value were retained unchanged, as defined in Table 8. The results of the evaluation are shown in the confusion matrices of Tables 10 and 11. It turns out that considerably more true positive hits can be achieved by adding the LDA model. Table 12 shows the differences between precision, recall, and the F1 score. The precision is always at 100%. With the addition of the LDA model, we obtain a higher recall (increase from 94.66% to 98.47%). The F1 score increases from 97.26% to 99.23%.

Overall, we can state that adding the LDA model resulted in higher evaluation scores. Using the cosine similarity of the LDA vectors provides an added value in particular if data sets have only a few properties and therefore only a few elements are available for comparison. All in all, we obtain satisfactory evaluation scores for disambiguating author names of data sets.

Our final knowledge graph contains 1,169 disambiguated authors or resources of the class `foaf:Person`.

Table 11. Confusion matrix without LDA model

	Actually identical authors	Actually nonidentical authors
Predicted identical authors	True Positive 124	False Positive 0
Predicted nonidentical authors	False Negative 7	True Negative 16

Table 12. Results of the evaluation metrics of the author disambiguation

Method	Precision	Recall	F1
Author disambiguation without LDA model	100%	94.66%	97.26%
Author disambiguation with LDA model	100%	98.47%	99.23%

5.4. Evaluation of the Linking to ORCID

When the data set authors are linked to their ORCID record, the name of each author in the knowledge graph is queried via the public API of ORCID. The first 25 results that this query returns are evaluated. A detailed comparison between two author profiles is only carried out if the author names of the author profiles have at least two identical terms. We compare 3,802 author profiles and find 214 matches. Of these, 208 are true positives and six are false positives. The developed comparison thus achieves a precision of 97.2%. Overall, we can add a link to their ORCID record for 17.8% of the authors in the knowledge graph.

Table 13 shows an overview of how often the individual criteria outlined in Section 4.5 were positive. We can observe that the coauthor information is a relatively clear signal for setting a correct link to ORCID. Authors often publish data sets with the same authors with whom they publish publications. The coauthor network is therefore an effective criterion for determining whether two authors are identical.

Scientists generally only state their publications and their associated identifiers on their ORCID records. Data sets are rarely given. Therefore, when comparing the various author profiles, only six matches are found for the provided identifiers. This problem is counteracted by the fact that the titles of the linked publications of the data set of an author are also considered for the comparison. It is assumed that the authors mention their data sets in their own publications. The comparison of the titles of the linked publications with the specified titles of the works on the ORCID records resulted in 37 matches. The clear identification of the authors of data sets can be made easier in the future if authors include their published data sets in their publication lists or, if possible, link their data sets to their own publications.

Table 13. Number of matching criteria of the performed linking to ORCID

Metadata	Criteria	Number of positive matches
Titles of the works	One identical word	17
	Two identical words	1
	More than two identical words	0
	Match with at least one linked publication title	37
Coauthors	One common coauthor	52
	More than one common coauthor	210
Identifiers	At least one identical identifier	6

6. APPLICATION SCENARIOS

In this section, we outline how the presented DSKG can be used, particularly in the context of new application scenarios.

- **Linked open data**: Because the DSKG is part of the Linked Open Data cloud and contains links to other data sources, it contributes to the use of linked data in the context of data sets on the web. By using the SPARQL endpoint available at http://dskg.org/sparql or the URI resolution, both users and programs can query the data. The reusability of the knowledge graph by third parties is simplified by the reuse of the DCAT vocabulary (Neumaier et al., 2017). As a result, the knowledge graph can be used as a data source for new or existing applications that are related to data sets on the web (Hallo, Luján-Mora et al., 2016). In particular, the high number of links indicating in which publications data sets were mentioned can be highly beneficial. Applications that previously only used the metadata concerning publications (e.g., as provided by the MAKG) can now use the data set representations to have a better understanding of the publications' key content.
- **Scholarly search and recommender systems**: Our knowledge graph can be used as a data and evaluation basis for innovative search engines for data sets. The online demonstration system http://datasetsearch.net, for instance, illustrates how users can search for data sets based on scientific problem descriptions as the user's input and the DSKG as the database. Having the DSKG in RDF allows us to compute data set entity embeddings and to use these embeddings in the context of state-of-the-art neural network-based search and recommender systems. In addition, representing the data in RDF enables developers to deploy semantic search systems that combine the data set metadata with metadata concerning publications, authors, venues, and institutions (Chapman et al., 2020; Färber, 2019) and allows users to search across scholarly data, such as data sets, publications, citations, authors, and venues (Baglioni et al., 2020). Due to the careful selection of the data sources when creating the DSKG (see Section 3), the DSKG exhibits a high level of accuracy of the data set metadata. Thus, we avoid the situations in which the users first need to identify and filter out malicious metadata (e.g., items that are not data sets or have incorrect attributes).
- **Scholarly data analysis and trend detection**: Using SPARQL queries, we can determine statistical key figures regarding modeled data sets and authors. For example, SPARQL queries can be used to determine authors of data sets in a specific application area whose data sets are linked to a large number of publications. Listing 2 shows a possible SPARQL query to identify the most influential authors in the field of biology. The SPARQL query determines the 50 authors with the most frequently referenced data sets from the

```
SELECT ?authorName ?orcid
       (COUNT(DISTINCT ?isReferencedBy) as ?NumberReferences)
       (GROUP_CONCAT(DISTINCT ?dataset; SEPARATOR=",␣") as ?dataset)
WHERE {
    ?dataset rdf:type  dcat:Dataset ;
             dct:isReferencedBy ?isReferencedBy ;
             dcat:theme <http://ma-graph.org/entity/86803240> ;
             dct:creator ?author .
    ?author rdf:type foaf:Person ;
            foaf:name  ?authorName .
    OPTIONAL { ?author adms:identifier ?orcid }
} GROUP BY ?authorName ?orcid
ORDER BY DESC (?NumberReferences)
LIMIT 50
```

Listing 2. SPARQL: The most influential authors in biology.

```
SELECT ?dataset (COUNT(DISTINCT ?isReferencedBy) as ?NumberReferences)
WHERE {
    ?dataset rdf:type  dcat:Dataset ;
             dct:isReferencedBy ?isReferencedBy ;
             dcat:theme <http://ma-graph.org/entity/41008148> .
} GROUP BY ?dataset
ORDER BY DESC (?NumberReferences)
LIMIT 10
```

Listing 3. SPARQL: The most influential data sets in computer science.

422 authors in this area. The authors are returned in descending order according to their number of references. In addition, the data sets of the authors and—if available—their ORCID ID are provided. Listing 3 shows a possible SPARQL query to determine the most scientifically influential data sets in the field of computer science. The SPARQL query returns the 10 data sets in this area that are referenced by most publications.

- **Scientific impact quantification and research evaluation**: Scientific impact quantification and research evaluation (Bornmann, 2013) are important parts of the fields digital libraries and science of science (Fortunato, Bergstrom et al., 2018; Wang & Barabási, 2021). In the past, a plethora of work has been proposed to measure the quality and impact of researchers and researchers' artifacts beyond plain citation-based metrics, such as the citation count or the *h*-index. Noteworthy are in particular altmetrics (Sugimoto, Work et al., 2017) defined as social web metrics for publications and approaches to measure the scientific impact beyond academia, such as the impact on the economy, society, health, and legislation (Ravenscroft, Liakata et al., 2017).
  Although existing approaches to scientific impact quantification and research evaluation are mainly dedicated to researchers and publications, relatively little attention has been paid to the assessment and rewarding of data sets. However, given the enormous growth in published data sets, disruptive approaches to data set evaluation and interlinking are more valuable than ever and reliable insights into data set provisioning and usage are required. In addition, creating, providing, and maintaining data sets as scientific artifacts is labor- and time-intensive, and these tasks are typically not sufficiently rewarded by stakeholders in the scientific landscape (Schöpfel & Azeroual, 2021).
  In the context of research data management, various efforts on the national and international level have been proposed to implement the FAIR principles (Wilkinson et al., 2016) and, as such, to promote the provisioning of data sets and their metadata. Noteworthy are, for instance, the national research data infrastructure in Germany and projects such as OpenAIRE with Zenodo in the EU. However, in the light of open science with the vision to model various kinds of scholarly data on the web (e.g., publications' metadata as well as key content, such as data sets, methods, and research findings), no knowledge graph specifically covering data sets and being compatible with linked open data has been proposed so far. Using RDF (W3C, 2014) as our data model and links to other linked open data sources, the DSKG facilitates data interoperability and data integration efforts, following the FAIR data principles (Wilkinson et al., 2016), and closes this gap.
  We argue that the DSKG can be used as a database for measuring the impact of data sets, as well as for establishing reward opportunities for data set providers in the future. For instance, in Färber et al. (2021), we indicated how an *h*-index for data sets can be created using links between data sets and publications in which the authors used the data sets. The frequencies of how often data sets are mentioned in a paper collection—as modeled in the DSKG—can be used as a first step for measuring the scientific impact of data sets (Belter, 2014; Konkiel, 2020). Furthermore, thanks to the rich metadata about scholarly

entities in the MAKG (Färber, 2019), which is interlinked with the DSKG, such as information about publications, authors, affiliations, conferences, journals, and fields of study, this information provides a great basis for advanced scientific impact quantification and research evaluation. For the measurements, SPARQL queries (see Listings 2 and 3 as examples) can be executed without any data dump downloading and processing by the user.

Overall, the DSKG provides the basis for realizing the vision that, by establishing and using links to other entities, such as publications, data sets are no longer seen as isolated marginal products of research but as research artifacts in their own right that can be reused and make a major contribution to science.

7. CONCLUSION

In this paper, we presented the DSKG, a knowledge graph for data sets with a corresponding schema. Based on an analysis of several data set collections, OpenAIRE and Wikidata were selected as the data basis for the DSKG. The metadata of the data sets which were mentioned in at least one publication of the Microsoft Academic Knowledge Graph (Färber, 2019) provided the basis for the DSKG.

To resolve the ambiguities of the author names, we adapted an author name disambiguation approach to data set metadata. In addition to explicit evidence, implicit evidence was taken into account in the form of latent topic modeling. When evaluating the author name disambiguation method, we obtained a satisfactory F1 score of 99.2%.

In addition, we developed a method for linking the data set authors given in the DSKG to their ORCIDs. Using this method, a link to their ORCID record could be added for 17.8% of the authors in the DSKG.

Our knowledge graph is available as Linked Open Data at http://dskg.org. Besides resolving URIs via HTTP content negotiation and providing RDF dump files on Zenodo, we provide a public SPARQL endpoint for querying. The knowledge graph comprises 2,208 data set instances and 813,551 links to scientific publications.

We outlined potential use cases of the created knowledge graph and showed that the DSKG can be used in particular in the context of search and recommender systems, as well as for scientific impact quantification.

We can assume that the number of published data sets will continue to increase in the coming years (Benjelloun et al., 2020), not least because of the increasing implementation of the FAIR principles (Wilkinson et al., 2016). Therefore, the need for knowledge graphs covering data sets will continue to increase. We will tackle this challenge by periodically updating the DSKG and linking the DSKG to future scholarly knowledge graphs that will cover the key content of scientific publications in a fine-grained manner (Jaradeh et al., 2019).

ACKNOWLEDGMENTS

We acknowledge support by the KIT-Publication Fund of the Karlsruhe Institute of Technology.

AUTHOR CONTRIBUTIONS

Michael Färber: Conceptualization, Data curation, Methodology, Resources, Supervision, Visualization, Writing—original draft, Writing—review & editing. David Lamprecht:

Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing—original draft, Writing—review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

No funding has been received for this research.

DATA AVAILABILITY

We provide our DSKG with a SPARQL endpoint, resolvable URIs, and entity embeddings at http://dskg.org to the public and also share it at https://doi.org/10.5281/zenodo.4478921. Our source code for generating the data set knowledge graph is available at https://github.com/michaelfaerber/data-set-knowledge-graph.

REFERENCES

Assaf, A., Troncy, R., & Senart, A. (2015). HDL – Towards a Harmonized Dataset Model for Open Data Portals. In *Proceedings of the 5th International Workshop on Using the Web in the Age of Data (USE-WOD'15) and the 2nd International Workshop on Dataset PROFIling and fEderated Search for Linked Data (PROFILES '15) @ ESWC'15* (pp. 62–74).

Baglioni, M., Manghi, P., & Mannocci, A. (2020). Context-driven discoverability of research data. In *International Conference on Theory and Practice of Digital Libraries* (pp. 197–211). https://doi.org/10.1007/978-3-030-54956-5_15

Belter, C. W. (2014). Measuring the value of research data: A citation analysis of oceanographic data sets. *PLOS ONE*, *9*(3), e92590. https://doi.org/10.1371/journal.pone.0092590, PubMed: 24671177

Benjelloun, O., Chen, S., & Noy, N. F. (2020). Google Dataset Search by the Numbers. In *Proceedings of the 19th International Semantic Web Conference* (pp. 667–682). https://doi.org/10.1007/978-3-030-62466-8_41

Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., & Fienberg, S. (2003). Adaptive name matching in information integration. *IEEE Intelligent Systems*, *18*(5), 16–23. https://doi.org/10.1109/MIS.2003.1234765

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Bornmann, L. (2013). What is societal impact of research and how can it be assessed? A literature survey. *Journal of the American Society for Information Science and Technology*, *64*(2), 217–233. https://doi.org/10.1002/asi.22803

Brickley, D., Burgess, M., & Noy, N. F. (2019). Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In *Proceedings of the World Wide Web Conference* (pp. 1365–1375). https://doi.org/10.1145/3308558.3313685

Canino, A. (2019). Deconstructing Google Dataset search. *Public Services Quarterly*, *15*(3), 248–255. https://doi.org/10.1080/15228959.2019.1621793

Caron, E., & van Eck, N. J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. In *Proceedings of the 19th International Conference on Science and Technology Indicators* (pp. 79–86).

Cen, L., Dragut, E. C., Si, L., & Ouzzani, M. (2013). Author disambiguation by hierarchical agglomerative clustering with adaptive stopping criterion. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 741–744). https://doi.org/10.1145/2484028.2484157

Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.-D., … Groth, P. (2020). Dataset search: A survey. *The VLDB Journal*, *29*(1), 251–272. https://doi.org/10.1007/s00778-019-00564-x

DataCite Metadata Working Group. (2017). *Datacite metadata schema documentation for the publication and citation of research data. Version 4.1*. https://doi.org/10.5438/0014

Dendek, P. J., Bolikowski, L., & Lukasik, M. (2012). Evaluation of features for author name disambiguation using linear support vector machines. In *Proceedings of the 10th IAPR International Workshop on Document Analysis Systems* (pp. 440–444). https://doi.org/10.1109/DAS.2012.36

Donner, P. (2014). Enhanced self-citation detection by fuzzy author name matching. *STI 2014 Leiden* (p. 178).

Färber, M. (2019). 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* (pp. 113–129). https://doi.org/10.1007/978-3-030-30796-7_8

Färber, M., Albers, A., & Schüber, F. (2021). Identifying used methods and datasets in scientific publications. In *Proceedings of the AAAI-21 Workshop on Scientific Document Understanding*.

Färber, M., Bartscherer, F., Menne, C., & Rettinger, A. (2018). Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web*, *9*(1), 77–129. https://doi.org/10.3233/SW-170275

Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. F. (2012). A brief survey of automatic methods for author name disambiguation. *SIGMOD Rec.*, *41*(2), 15–26. https://doi.org/10.1145/2350036.2350040

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., … Barabási, A.-L. (2018). Science of science. *Science*, *359*(6379). https://doi.org/10.1126/science.aao0185, PubMed: 29496846

Gregory, K. M., Cousijn, H., Groth, P., Scharnhorst, A., & Wyatt, S. (2020). Understanding data search as a socio-technical practice.

Journal of Information Science, *46*(4), 459–475. https://doi.org/10.1177/0165551519837182

Haak, L. L., Fenner, M., Paglione, L., Pentz, E., & Ratner, H. (2012). ORCID: A system to uniquely identify researchers. *Learned Publishing*, *25*(4), 259–264. https://doi.org/10.1087/20120404

Hajra, A., Radevski, V., & Tochtermann, K. (2015). Author profile enrichment for cross-linking digital libraries. In *Proceedings of the 19th International Conference on Theory and Practice of Digital Libraries* (pp. 124–136). https://doi.org/10.1007/978-3-319-24592-8_10

Hallo, M., Luján-Mora, S., Maté, A., & Trujillo, J. (2016). Current state of linked data in digital libraries. *Journal of Information Science*, *42*(2), 117–127. https://doi.org/10.1177/0165551515594729

Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, *1*(1), 1–136. https://doi.org/10.2200/S00334ED1V01Y201102WBE001

Henderson, T., & Kotz, D. (2015). Data citation practices in the CRAWDAD wireless network data archive. *D-Lib Magazine*. https://doi.org/10.1045/january2015-henderson

Jaradeh, M. Y., Oelen, A., Farfar, K. E., Prinz, M., D'Souza, J., ... Auer, S. (2019). Open Research Knowledge Graph: Next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture* (pp. 243–246). https://doi.org/10.1145/3360901.3364435

Konkiel, S. (2020). Assessing the impact and quality of research data using altmetrics and other indicators. *Scholarly Assessment Reports*, *2*(1). https://doi.org/10.29024/sar.13

Latif, A., Limani, F., & Tochtermann, K. (2021). On the complexities of federating research data infrastructures. *Data Intelligence*, 1–8. https://doi.org/10.1162/dint_a_00080

Lin, Q., Zhu, Y., Lu, H., Shi, K., & Niu, Z. (2021). Improving university faculty evaluations via multi-view knowledge graph. *Future Generation Computer Systems*, *117*, 181–192. https://doi.org/10.1016/j.future.2020.11.021

Manghi, P., Atzori, C., Bardi, A., Schirrwagen, J., Dimitropoulos, H., ... Summann, F. (2019). *OpenAIRE Research Graph Dump*. Zenodo. https://doi.org/10.5281/zenodo.3516918

Manghi, P., Bardi, A., Atzori, C., Baglioni, M., Manola, N., ... Principe, P. (2019). *The OpenAIRE Research Graph Data Model*. Zenodo. https://doi.org/10.5281/zenodo.2643199

Manola, N., Mutschke, P., Scherp, G., Tochtermann, K., Wittenburg, P., ... Van Uytvanck, D. (2019). Implementing fair data infrastructures. In *Dagstuhl Perspectives Workshop 18472: "Implementing Fair Data Infrastructures"* (pp. 16–38).

Neumaier, S. (2019). *Semantic Enrichment of Open Data on the Web* (Unpublished doctoral dissertation). TU Wien.

Neumaier, S., Polleres, A., Steyskal, S., & Umbrich, J. (2017). Data integration for open data on the web. In *Proceedings of the 13th Reasoning Web International Summer School* (pp. 1–28). https://doi.org/10.1007/978-3-319-61033-7_1

Neumaier, S., Umbrich, J., & Polleres, A. (2016). Automated quality assessment of metadata across open data portals. *ACM Journal of Data and Information Quality*, *8*(1), 2:1–2:29. https://doi.org/10.1145/2964909

Neumaier, S., Umbrich, J., & Polleres, A. (2017). Lifting data portals to the web of data. In *Proceeding of the Workshop on Linked Data on the Web (LDOW'17) @ WWW'17*. CEUR-WS.org.

NISO. (2004). Understanding metadata. *National Information Standards*.

Ohno-Machado, L., Sansone, S.-A., Alter, G., Fore, I., Grethe, J., ... Kim, H.-E. (2017). Finding useful data across multiple biomedical data repositories using Datamed. *Nature Genetics*, *49*(6), 816–819. https://doi.org/10.1038/ng.3864, PubMed: 28546571

Ojo, A., & Sennaike, O. (2020). Constructing knowledge graphs from data catalogues. In *Proceedings of the 16th International Conference on Distributed Computing and Internet Technology* (pp. 94–107). https://doi.org/10.1007/978-3-030-36987-3_6

Perego, A., Austin, T., Friis-Christensen, A., Vaccari, L., & Tsinaraki, C. (2020). *DataCite to DCAT-AP Mapping*. https://ec-jrc.github.io/datacite-to-dcat-ap/ (accessed: November 18, 2020)

Peroni, S., & Shotton, D. M. (2020). Opencitations, an infrastructure organization for open scholarship. *Quantitative Science Studies*, *1*(1), 428–444. https://doi.org/10.1162/qss_a_00023

Protasiewicz, J., & Dadas, S. (2016). A hybrid knowledge-based framework for author name disambiguation. In *Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 594–600). IEEE. https://doi.org/10.1109/SMC.2016.7844305

Radevski, V., Hajra, A., & Limani, F. (n.d.). Semantically related data as technology-enhanced support for research assistive and quality tools. In *UNESCO International Workshop* (p. 37).

Ravenscroft, J., Liakata, M., Clare, A., & Duma, D. (2017). Measuring scientific impact beyond academia: An assessment of existing impact metrics and proposed improvements. *PLOS ONE*, *12*(3), e0173152. https://doi.org/10.1371/journal.pone.0173152, PubMed: 28278243

Ristoski, P., Rosati, J., Di Noia, T., De Leone, R., & Paulheim, H. (2019). RDF2Vec: RDF graph embeddings and their applications. *Semantic Web*, *10*(4), 721–752. https://doi.org/10.3233/SW-180317

Sansone, S.-A., Gonzalez-Beltran, A., Rocca-Serra, P., Alter, G., Grethe, J. S., ... Ohno-Machado, L. (2017). DATS, the data tag suite to enable discoverability of datasets. *Scientific Data*, *4*, 170059. https://doi.org/10.1038/sdata.2017.59, PubMed: 28585923

Schöpfel, J., & Azeroual, O. (2021). Rewarding research data management. In *Companion of the Web Conference 2021, Virtual Event, Ljubljana, Slovenia, April 19–23, 2021* (pp. 446–450). ACM/IW3C2. https://doi.org/10.1145/3442442.3451367

Sennaike, O. A., Waqar, M., Osagie, E., Hassan, I., Stasiewicz, A., ... Ojo, A. (2017). Towards intelligent open data platforms: Discovering relatedness in datasets. In *Proceedings of the 2017 Intelligent Systems Conference* (pp. 414–421). https://doi.org/10.1109/IntelliSys.2017.8324327

Song, Y., Huang, J., Councill, I. G., Li, J., & Giles, C. L. (2007). Efficient topic-based unsupervised name disambiguation. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries* (pp. 342–351). https://doi.org/10.1145/1255175.1255243

Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology*, *68*(9), 2037–2062. https://doi.org/10.1002/asi.23833

Tatman, R. (2017). *English Word Frequency*. https://www.kaggle.com/rtatman/english-word-frequency (accessed: December 12, 2020)

Tekles, A., & Bornmann, L. (2019). Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches. In *Proceedings of the 17th International Conference on Scientometrics and Informetrics* (pp. 1548–1559).

Vahdati, S., Karim, F., Huang, J.-Y., & Lange, C. (2015). Mapping large scale research metadata to linked data: A performance comparison of HBase, CSV and XML. In *Research Conference on Metadata and Semantics Research* (pp. 261–273). https://doi.org/10.1007/978-3-319-24129-6_23

Vrandecic, D. (2019). Describing datasets in Wikidata. In *Proceedings of the 15th International Conference on eScience* (pp. 528–529). https://doi.org/10.1109/eScience.2019.00070

W3C. (2013a). *SPARQL 1.1 Query Language*. https://www.w3.org/TR/sparql11-query/ (accessed: November 19, 2020)

W3C. (2013b). *SPARQL 1.1 Update*. https://www.w3.org/TR/2013/REC-sparql11-update-20130321/#graphUpdate (accessed: November 19, 2020)

W3C. (2014). *RDF 1.1 Concepts and Abstract Syntax*. https://www.w3.org/TR/rdf11-concepts/ (accessed: November 4, 2020)

W3C. (2020). *Data Catalog Vocabulary (DCAT) – Version 2*. https://www.w3.org/TR/vocab-dcat-2/ (accessed: November 15, 2020)

Wang, D., & Barabási, A. (2021). *The science of science*. Cambridge University Press. https://doi.org/10.1017/9781108610834

Wang, J., Aryani, A., Wyborn, L., & Evans, B. (2017). Providing research graph data in JSON-LD using Schema.org. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 1213–1218). https://doi.org/10.1145/3041021.3053052

Web Data Commons. (2018). *Download Instructions for the WDC RDFa, Microdata, Embedded JSON-LD, and Microformats Data Sets (November 2018)*. Retrieved from https://webdatacommons.org/structureddata/2018-12/stats/how_to_get_the_data.html

WikiProject Datasets/Data Structure/DCAT – Wikidata – Schema.org mapping. (2018). https://www.wikidata.org/wiki/Wikidata:WikiProject_Datasets/Data_Structure/DCAT_-_Wikidata_-_Schema.org_mapping (accessed: November 18, 2020)

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9. https://doi.org/10.1038/sdata.2016.18, PubMed: 26978244

Yi, B., Ludo, W., & Yong, H. (2021). A multi-dimensional framework for characterizing the citation impact of scientific publications. *Quantitative Science Studies*, *2*(1), 155–183. https://doi.org/10.1162/qss_a_00109

Younsi Dahbi, K., Lamharhar, H., & Chiadmi, D. (2020). Towards a knowledge graph for open healthcare data. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(4), 5654–5662. https://doi.org/10.30534/ijatcse/2020/216942020