A Topic Model for the Data Web

Michael Röder, Denis Kuchelev, and Axel-Cyrille Ngonga Ngomo

DICE, Paderborn University, Germany,
michael.roeder@uni-paderborn.de

Abstract. The usage of knowledge graphs in industry and at Web scale has increased steadily within recent years. However, the decentralized approach to data creation which underpins the popularity of knowledge graphs also comes with significant challenges. In particular, gaining an overview of the topics covered by existing datasets manually becomes a gargantuan if not impossible feat. Several dataset catalogs, portals and search engines offer different ways to interact with lists of available datasets. However, these interactions range from keyword searches to manually created tags and none of these solutions offers an easy access to human-interpretable categories. In addition, most of these approaches rely on metadata instead of the dataset itself. We propose to use topic modeling to fill this gap. Our implementation LODCat automatically creates human-interpretable topics and assigns them to RDF datasets. It does not need any metadata and solely relies on the provided RDF dataset. Our evaluation shows that LODCat can be used to identify the topics of hundreds of thousands of RDF datasets. Also, our experiment results suggest that humans agree with the topics that LODCat assigns to RDF datasets. Our code and data are available online.

Keywords: RDF datasets, topic modeling, Data Web

1 Introduction

With the growth of the size and the increase in the number of knowledge graphs available on the Web comes the need to process this data in a scalable way [17]. The large number of datasets that are available online and their sheer size make it costly or even infeasible to handle each of these datasets manually without the support of proper tools. A particularly important issue is that the mere identification of relevant datasets for a particular task (e.g., data integration [24], question answering [32], machine learning [14], etc.) may become challenging. Indeed, domain experts who plan to use knowledge graphs for a task may be able to read Resource Description Framework (RDF) data but will not have the time to read through hundreds of thousands of datasets to determine whether they are relevant. Hence, *we need to be able to characterize RDF datasets so that users can easily find datasets of interest.*

A similar problem is already known from the processing of large amounts of human-readable documents. Most users might be able to read all books within a library. However, they may not have the time to do so just to identify the

books that they are interested in. Although there are search engines that allow the indexing of documents, users would have to know the right keywords to find documents that they are interested in [15]. Hence, "search engines are not the perfect tool to explore the unknown in document collections" [15]. However, today's dataset search engines mainly rely on keyword searches on the dataset's metadata and user-created tags although both suffer from the aforementioned drawback. At the same time, RDF datasets may not have rich metadata that could be used for such a search to improve their findability [37,23]. For example, the Vocabulary of Interlinked Datasets (VoID) vocabulary offers ways to add metadata in form of descriptions that can be indexed by a search engine. However, Paulheim et al. [27] show that a best practice proposed in the VoID specification [2]—i.e., to use the `/.well-known/void` path on a Web server to provide an RDF file with VoID information about datasets hosted on the server—is not adopted on a large scale. Similarly, Schmachtenberg et al. [33] report that only 14.69% of 1014 datasets that they crawled provide VoID metadata.

We tackle this gap with our topic-modeling-based approach LODCat. Topic modeling algorithms can be used to infer latent topics in a given document collection. These topics can be used to structure the document collection and enable users to focus on subsets of the collection, which belong to their area of interest. Our main contribution in this publication is the application of topic modeling to a large set of RDF datasets to support the exploration of the Data Web based on human-interpretable topics. To this end, we tackle the challenge of transforming the RDF datasets into a form that allows the application of a topic modeling algorithm. Our evaluation shows that this approach can be applied to hundreds of thousands of RDF datasets. The results of a questionnaire suggest that humans generally agree with the topics that our approach assigns to a sample of example datasets.

The following section describes related work before Section 3 describes the single steps of our approach. Section 4 describes the setup and results of our evaluation before we conclude with Section 5.

2 Related Work

In their survey of data search engines, Chapman et al. [10] divide these engines into four categories. The first category are database search engines. They are used with structured queries that are executed against a database back end. The second set of search engines are information retrieval engines. These are integrated into data portals like CKAN[1] and offer a keyword-based search on the metadata of datasets. The third category are entity-centric search engines. The query of such an engine comprises entities of interest and the search engine derives additional information about these entities. The last category is named tabular search. A user of such a search engine tries to extend or manipulate one or more existing tables by executing search queries.

[1] https://ckan.org/

The second category represents the most common approach to tackle the search for datasets on the web. Several open data portals exist that offer a list of datasets and a search on the dataset's metadata. Examples are the aforementioned CKAN, kaggle[2] or open government portals like the european data portal[3]. The Google dataset search presented by Brickley et al. [8] works in a similar way but uses the Google crawler to collect the data from different sources. Singhal et al. [34] present DataGopher—a dataset search that is optimized for research datasets. Devaraju et al. [12] propose a personalized recommendation of datasets based on user behavior. Our approach differs to these approaches as we focus on RDF datasets and rely on the dataset itself instead of only using metadata. In addition, we do not rely on a keyword search or user created tags but automatically generated topics that are assigned to the datasets.

Kunze et al. [19] propose an explorative search engine for a set of RDF datasets. This engine is mainly based on filters that work similar to a faceted search. For example, one of these filters is based on the RDF vocabularies that the datasets use. Vandenbussche et al. [39] present a web search for RDF vocabularies.[4] Kopsachilis et al. [18] propose GeoLOD—a dataset catalog that focuses on geographical RDF datasets. LODAtlas [28] combines several features of the previously mentioned systems into a single user interface. These approaches have similar limitations as the generic open data portals described above. While some of them offer additional features, non of them offers human-interpretable categories that go beyond manually created tags.

Topical profiling of RDF datasets [36] is very closely related to our work. The task is defined as a single- or multi-label classification problem. Blerina et al. [36] propose a benchmark that is based on the manually created classes of the Linked Open Data cloud project [22]. In a recent work, Asprino et al. [3] tackle the multi-classification task and extend the benchmark dataset. Some of the features that are used for the classification, like the virtual documents that are generated by the approach of Asprino et al. are similar to the documents our approach creates. However, our approach is unsupervised while the proposed approaches are supervised and rely on training data, that has been created manually.

Several approaches exist to explore RDF datasets. Tzitzikas et al. [38] define a theoretical framework for these explorative search engines and compare several approaches. However, all these approaches focus on exploring a single RDF dataset while our goal is to enable users to derive topically interesting RDF datasets from a set of datasets.

Röder et al. [30] propose the application of topic models to identify RDF datasets as candidates for link prediction. However, their work relies solely on the ontologies of datasets and the topics are used as features for a similarity calculation while we need high-quality topics that can be shown to users.[5] Sleeman et al. [35] use topic modeling to assign topics to single entities of an RDF

[2] https://www.kaggle.com/datasets
[3] https://data.europa.eu/en
[4] https://lov.linkeddata.es/dataset/lov
[5] Chang et al. [9] show that there is a difference between these two usages of topics.

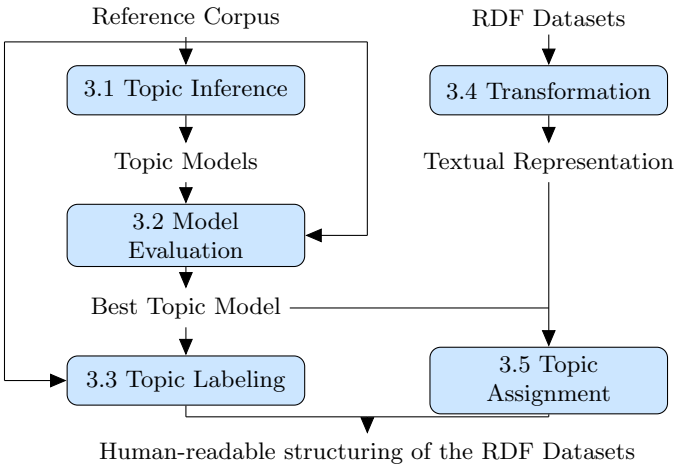

Fig. 1: Overview of the workflow of LODCat.

dataset. Thus, both of the aforementioned approaches use topic modeling with different aims and are not comparable to our approach.

3 LODCat

Figure 1 shows an overview of our proposed approach LODCat. It relies on a reference text corpus (e.g., Wikipedia) as a source of general knowledge and uses topic modeling to assign human-interpretable topics to the single RDF datasets. First, we use the reference corpus to generate several topic models. Thereafter, the single models are evaluated and the best model is chosen for further processing. For each topic of this model, a label is generated to make the complex probability distributions human-readable. In parallel, the RDF datasets are transformed into a textual representation (i.e., documents). Based on the chosen topic model, a topic distribution is assigned to each of the generated documents. At the end, each RDF dataset has a set of topics that are dominant for that dataset and that are described by their labels. This data is used to provide a faceted search, which helps the user to find datasets related to their field of interest. The single steps of our approach are described in more detail in the following.[6]

3.1 Topic Inference

Our current version of LODCat relies on the topic modeling approach Latent Dirichlet Allocation (LDA) [7,6]. This approach assumes that there are latent topics that are defined as distributions over words ϕ, i.e., each word type w has a probability representing the likeliness to encounter this word while reading

[6] LODCat is open source at https://github.com/dice-group/lodcat.

about the topic. The topics are derived from a given corpus D, i.e., a set of documents. Each document d_i has a topic distribution θ_i, i.e., each topic has a probability how likely it is to occur within the document. LDA connects these distributions by defining that each word token w has been created by a single topic and assigning the ID of this topic to the token's z variable. Let $w_{i,j}$ be the j-th word token in the i-th document, let $\mathrm{w}_{i,j}$ be its word type, and let $z_{i,j}$ denote the id of the topic from which the word type of this token has been sampled. Let ϱ be the number of topics and let $Z = \left\{ z_{1,1}, \ldots, z_{|D|,|d_{|D|}|} \right\}$ be the set of the topic indices of all word tokens in the corpus D. Let $\Phi = \{\phi_1, \ldots, \phi_\varrho\}$ be the set of word distributions and $\Theta = \{\theta_1, \ldots, \theta_{|D|}\}$ be the set of topic distributions. LDA is based on the following joint distribution [6]:

$$\mathbb{P}(\Phi, \Theta, Z, D) = \left(\prod_{i=1}^{\varrho} \mathbb{P}(\phi_i) \right) \left(\prod_{i=1}^{|D|} \mathbb{P}(\theta_i) \left(\prod_{j=1}^{|d_i|} \mathbb{P}(z_{i,j} | \theta_i) \, \mathbb{P}\left(\mathrm{w}_{i,j} \middle| \phi_{z_{i,j}} \right) \right) \right) . \tag{1}$$

We use the LDA inference algorithm proposed by Hoffman et al. [16] that takes a corpus and the number of topics ϱ as input. The output is a topic model comprising the topics' distributions Φ.[7] Since the best number of topics ϱ is unknown we generate several models with different ϱ values.

3.2 Model Evaluation

In this step, we choose the best model from the set of generated topic models. To this end, we use topic coherence measures to evaluate the human-readability and interpretability of the model's topics. We represent each topic by its 10 top words $\bar{W}$, i.e., the 10 words that have the highest probability in the topic's word distribution Φ. We use these top words as input for two coherence measures proposed by Röder et al. [31], namely C_P and a variant of the C_V measure that we call C_{V2}.[8] Hence, we get two coherence values for each topic. For each measure, we sort all models based on the average coherence value of their topics. The model that achieves the best rank on average for both coherence measures is the model that will be used for further processing.

In addition, we use the coherence values to identify low-quality topics. These are topics that should not be shown to the user. We define a topic to be of low-quality if its C_{V2} or C_P value is below 0.125 or 0.25, respectively.

3.3 Topic Labeling

For each topic of the chosen model, we assign label that can be used to present the topic to users. For our current implementation, we use the Neural Embedding

[7] Due to space limitations, we refer the interested reader to Blei et al. [6] and Hoffman et al. [16] for further details about LDA and the used inference algorithm.

[8] The C_{V2} measure has the same definition as the C_V measure but uses the S^{one}_{all} segmentation [31]. The variant showed a better performance in our experiments.

Topic Labelling (NETL) approach of Bathia et al. [5] since 1) their evaluation shows that NETL outperforms the approach of Lau et al. [20] and 2) the approach is available as open-source project.[9] NETL generates label candidates for a given topic from a reference corpus and ranks them according to a trained model. Following Bathia et al. [5], we use the English Wikipedia as reference corpus and use their pre-trained support vector regression model to rank the label candidates. We also use the topics top words as additional topic descriptions.

3.4 RDF Dataset Transformation

Our goal in this step is to transform given RDF datasets into a textual representation that can be used in combination with the generated topic model. This step relies on the Internationalized Resource Identifiers (IRIs) that occur in the datasets. We determine the frequency f of each IRI in the dataset (either as subject, predicate or object of a triple). IRIs of well-known namespaces that do not have any topical value like `rdf`, `rdfs` and `owl` are filtered out. After that, the labels of each IRI are retrieved. This label retrieval is based on the list of IRIs that have been identified as label-defining properties by Ell et al. [13]. Additionally, we treat values of `rdfs:comment` as additional labels. If there are no labels available, the namespace of the IRI is removed and the remaining part is used as label. If this generated label is written in camel case or contains symbols like underscores, it is split into multiple words. The derived labels are further preprocessed using a tokenizer and a lemmatizer [21]. The derived words inherit the counts f of their IRI. If IRIs share the same word their counts are summed up to derive the count of this word.

However, we do not use the counts directly for generating a document since some IRIs may occur hundreds of thousand times within a dataset. Their words would dominate the generated document and marginalize the influence of other words. In addition, large count values could lead to very long documents that may create further problems with respect to the resource consumption in later steps. To reduce the influence of words with very high f values we determine the frequency ψ of word type w for the document d'_i of the i-th dataset as follows:

$$\psi_{i,\mathrm{w}} = r(\log_2(f_{\mathrm{w}}) + 1)\,, \tag{2}$$

where r is the rounding function which returns the closest integer value preferring the higher value in case of a tie [1].[10] The result of this step is a bag of words representation of one document for each RDF dataset.

3.5 Topic Assignment

The last step is the assignment of topics to the documents that represent the RDF datasets. For each created synthetic document d'_i, we use the chosen topic

[9] https://github.com/sb1992/NETL-Automatic-Topic-Labelling-

[10] The transformation of counts into occurrences in a synthetic document is similar to the logarithmic variant of the approach described by Röder et al. [30].

model to infer a topic distribution θ'_i. This distribution is used to derive the dataset's top topics, i.e., the topics with the highest probabilities in the distribution. The labels and top words of these topics are used as a human-readable representation of the RDF dataset.

4 Evaluation of LODCat

We evaluate LODCat in a setup close to a real-world scenario.[11] We start with the English Wikipedia as corpus and process more than 600 thousand RDF datasets using LODCat following the steps described in Section 3. The evaluation can be separated into the following three consecutive experiments:

1. The generation and selection of the topic model,
2. The RDF dataset transformation and the topic assignment, and
3. The evaluation of the assigned topics based on a user study.

4.1 Datasets

For our evaluation, we use two types of data—a reference corpus to generate the topic model and the set of RDF datasets, which should be represented in a human-interpretable way. We use the English Wikipedia as reference corpus.[12] We preprocess the dump file by removing Wikimedia markup, removing redirect articles and handling each remaining article as an own document. Each document is preprocessed as described in Section 3.1. From the created set of documents, we derive all word types and count their occurrence. Then, we filter the word types by removing 1) common English terms based on a stop word list, and 2) all word types that occur in more than 50% of the documents or 3) in less than 20 documents.[13]From the remaining word types, we select the 100 000 word types with the highest occurrence counts and remove all other from the documents. After that, we remove empty documents and randomly sample 10% of the remaining documents. Finally, we get a corpus with 619 475 documents and 190 million word tokens.

We gather 623 927 real-world RDF datasets from the LOD Laundromat project [4].[14] These datasets should be represented in a human-interpretable way. Note that these datasets have been stored without any metadata that could be used. Figure 2 shows the size of the RDF datasets. The largest dataset has 43 million triples while the majority has between 100 and 10 000 triples. In total, the datasets comprise 3.7 billion triples.[15]

[11] The RDF datasets we use are available at https://figshare.com/s/af7f18a7f3307cc86bdd while the results as well as the Wikipedia-based corpus can be found at https://figshare.com/s/9c7670579c969cfeac05.

[12] We use the dump of the English Wikipedia from September 1st 2021.

[13] The stop word list can be found online. We will add the link after the review phase.

[14] We downloaded the datasets in January 2018.

[15] Note that we do not deduplicate the triples across the datasets.

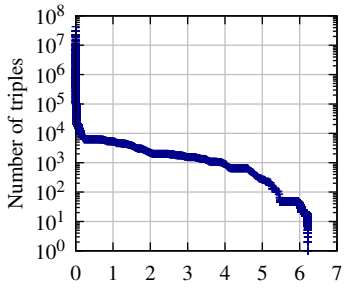

Fig. 2: The sizes of the RDF datasets (the x-axis is the dataset ID in 10^5).

4.2 Setup

Experiment I. In the first experiment, we infer the topic models based on the English Wikipedia corpus. We infer several models with different numbers of topics.[16] For this evaluation, we use $\varrho = \{80, 90, 100, 105, 110, 115, 120, 125, 135\}$ and generate three models for each number of topics. We choose the best topic model as described in Section 3.2, analyze this model with respect to the model's coherence values and show example topics.

Experiment II. Based on the best topic model created in the first experiment, we process each of the 623 927 RDF datasets by LODCat. During this step, we remove datasets that lead to an empty document. The result comprises a topic distribution for each dataset based on the used topic model. We analyze these distributions by looking at their top topics.

Experiment III. Finally, we evaluate the assignment of the topics to the datasets. Chang et al. [9] propose the topic intruder experiment to evaluate the assignment of topics to documents. They determine the top topics of a document and insert a randomly chosen topic from the same topic model that is not one of the document's top topics. This randomly chosen topic is called intruder topic. After that, volunteers are given the created list of topics and the document, and are asked to identify the intruder. The more often the intruder is successfully identified, the better is the topic assignment of the topic model. We use the same approach to evaluate whether a topic model can assign meaningful topics to an RDF dataset. We sample 60 datasets that have more than 100 and less than 10 000 triples. For each of the sampled datasets, we derive the three topics with the highest probability. Based on the dataset content and the quality of their top topics, we choose 10 datasets that 1) have at least two high-quality topics among the top three topics, 2) have a high-quality topic as highest ranked topic, 3) have a content that can be understood without accessing further sources, and 4) have not exactly the same top topics as the already chosen datasets. For each chosen dataset, we sample an intruder topic from the set of high-quality topics that are not within the top three topics of the dataset.

[16] We use the Gensim library [29] with hyper parameter optimization. https://radimrehurek.com/gensim/index.html

We create a questionnaire with 10 questions. Each question gives the link to one of the chosen datasets and a list of topics comprising the top topics of the dataset and the intruder topic in a random order. 5 chosen datasets have three high-quality topics while the other 5 datasets have one top topic with a low coherence value. We remove the topics with the low values. Hence, 5 question comprise 4 topics and the other 5 questions have 3 topics from which a user should choose the intruder topic. For the questionnaire, the topics are represented in the human-readable way described in Section 3.3, i.e., with their label and their top words. The participants of the questionnaire are encouraged to look into the RDF dataset. However, they should not include further material. We sent this questionnaire to several mailing lists to encourage experts and experienced users of the Semantic Web to participate.

Following Chang et al. [9], we calculate the topic log odds to measure the agreement between the topic model and the human judgments that we gather with our questionnaire. Let θ'_i be the topic distribution of the i-th document d'_i. Let $\theta'_{i,k}$ be the probability of the k-th topic for document d'_i. Let $Y_i = \{y_{i,1}, \ldots\}$ be the bag of all user answers for document d'_i, i.e., the j-th element is the id of the topic that the j-th user has chosen as intruder topic for this document. Let x_i be the id of the real intruder topic for document d'_i. Chang et al. [9] define the topic log odds $\mathfrak{o}$ for the i-th document as the average difference between the probabilities of the chosen intruder topics compared to the real intruder topic:

$$\mathfrak{o}(\theta'_i, Y_i, x_i) = \frac{1}{|Y_i|} \sum_{j=1}^{|Y_i|} \left(\log(\theta'_{i,x_i}) - \log(\theta'_{i,y_{i,j}}) \right) . \tag{3}$$

A perfect agreement between the human participants and the topic model would lead to $\mathfrak{o} = 0$. In practice, this is only reached if all participating volunteers find the correct intruder topic.

4.3 Results

Experiment I. From the 27 generated topic models, the model that received the best average ranks according to both topic coherence measures is a model with 115 topics. Figure 3 shows the coherence values of this model's topics for both coherence measures. The dashed line shows the threshold used to distinguish between high and low-quality topics. Based on the two thresholds, 74 topics are marked as high-quality topics while the remaining 41 topics are treated as low-quality topics. Table 1 shows the model's topics with the 5 highest and the 5 lowest C_{V2} coherence values. While the first 5 topics seem to focus on a single topic the topics with the low coherence scores comprise words that seem to have no strong relation to each other.

Experiment II. LODCat is able to assign topics to 561 944 of the given 623 927 RDF datasets, which are $> 90\%$. 61 983 datasets lead to the creation of empty documents and, hence, cannot get any topics assigned. These are mainly small datasets with IRIs that cannot be transformed into meaningful words, i.e., words that are not removed by our stop word removal step.

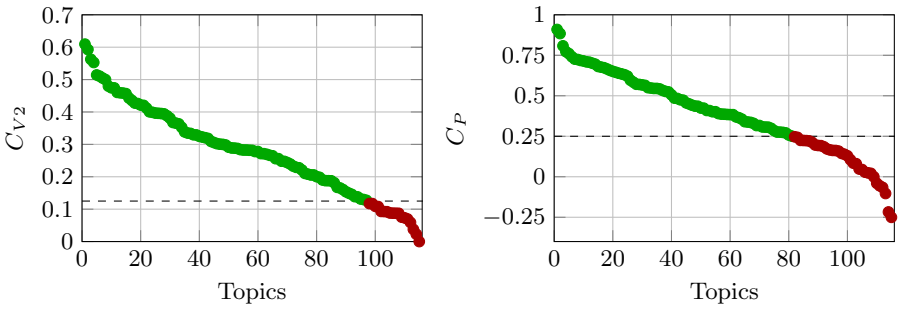

Fig. 3: Topics of the best performing model sorted by their C_{V2} and C_P coherence values, respectively. The dashed line shows the threshold used to separate high-quality topics (green) from low-quality topics (red).

Table 1: The top words of the 5 topics of the chosen topic model with the highest and lowest C_{V2} values, respectively.

C_{V2}	$\bar{W}$
0.60942	canadian, canada, quebec, ontario, montreal, toronto, ottawa, nova, scotia, alberta
0.59273	album, song, release, band, music, chart, record, single, track, records
0.56269	age, population, household, female, city, male, family, census, average, year
0.55282	chinese, china, singapore, li, wang, shanghai, chen, beijing, hong, zhang
0.51424	league, club, player, football, season, cup, play, goal, team, first
0.06969	rank, time, men, advance, event, final, result, athlete, heat, emperor
0.05900	use, language, word, name, form, one, english, see, greek, two
0.03767	use, system, one, number, two, function, set, space, model, time
0.02269	use, health, may, child, include, provide, would, act, make, public
0.00000	j., a., m., c., r., s., l., e., p., d.

After generating the topic distributions for the documents created from the RDF datasets, we analyze these distributions. For each dataset, we determine its main topic, i.e., the high-quality topic with the highest probability for this dataset. Figure 4 shows the number of datasets for which each topic is the main topic. The figure shows that a single topic covers more than 508 thousand datasets. Table 2 shows the 5 topics that have the highest values in Figure 4. We can see that a weather-related topic covers roughly 90% of the datasets to which LODCat could assign topics. The next biggest topics are transportation- and car-related topics and each of them covers nearly 10 thousand datasets. They are followed by a computer- and a travel-related topic.

We further analyze the RDF datasets with respect to the claim that the majority of them is related to weather. We analyze the namespaces that are used within the RDF datasets and count the number of datasets in which they occur. Figure 5 shows the result of this analysis for all 623 thousand RDF datasets. In the lower right corner of the figure, we can see that there is only a small number of namespaces that are used in many datasets. Table 3 shows the 12 namespaces that occur in more than 100 thousand datasets. The most often used

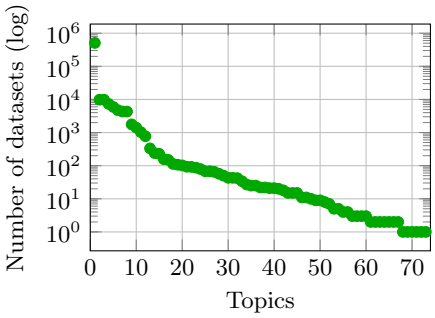

Fig. 4: Number of datasets per topic for which this topic has the highest probability sorted in descending order. Topics with no datasets have been left out.

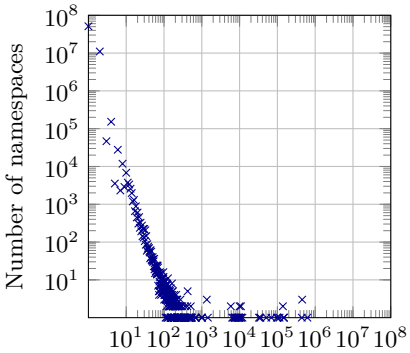

Fig. 5: The number of namespaces that occur in a number of datasets.

Table 2: Top topics with the highest number of datasets.

Datasets	$\bar{W}$
508 095	water, storm, wind, tropical, nuclear, temperature, hurricane, damage, cause, system
9 828	station, road, route, line, street, bridge, railway, city, highway, east
9 794	car, engine, model, vehicle, first, use, point, motor, design, safe
7 328	use, system, software, user, datum, computer, include, information, support, service
5 946	airport, international, brazil, portuguese, são, romanian, portugal, brazilian, language, romania

namespace is the `rdf` namespace, which is expected. The namespaces on position 2–4 occur in more than 450 thousand RDF datasets and belong to datasets with sensor data described by Patni et al. [26]. A further search revealed that the data comprises hurricane and blizzard observations from weather stations [25]. These datasets also make use of the fifth namespace from Table 3. The sixth namespace is the Data Cube namespace, which is used to described statistical data in RDF [11]. This namespace occurs often together with the remaining namespaces (7–12). They occur in datasets that origin from the Climate Change Knowledge Portal of the World Bank Group.[17] These datasets contain climate data, e.g., the temperature for single countries and their forecast with respect to different climate change scenarios. We summarize that our analysis shows that the majority of the datasets contain sensor data, and statistical data that are related to weather. This is in line with the results returned by our approach LODCat.

Experiment III. Our questionnaire received 225 answers from 65 participants.[18] Figure 6 shows the results. The left side of the figure summarizes the

[17] https://climateknowledgeportal.worldbank.org/

[18] We used LimeSurvey for the questionnaire (https://www.limesurvey.org/). The questionnaire allowed users to skip questions. These skipped questions are not taken into account for the number of answers.

Table 3: The namespaces that occur in more than 100 000 datasets.

ID	Namespace IRI	Datasets
1	http://www.w3.org/1999/02/22-rdf-syntax-ns\#	620 653
2	http://knoesis.wright.edu/ssw/ont/weather.owl\#	452 453
3	http://knoesis.wright.edu/ssw/ont/sensor-observation.owl\#	452 453
4	http://knoesis.wright.edu/ssw/	452 453
5	http://www.w3.org/2006/time\#	442 719
6	http://purl.org/linked-data/cube\#	147 731
7	http://worldbank.270a.info/property/	147 348
8	http://purl.org/linked-data/sdmx/2009/dimension\#	147 305
9	http://worldbank.270a.info/dataset/world-bank-climates/	139 865
10	http://worldbank.270a.info/classification/variable/	139 865
11	http://worldbank.270a.info/classification/scenario/	114 064
12	http://worldbank.270a.info/classification/percentile/	103 202

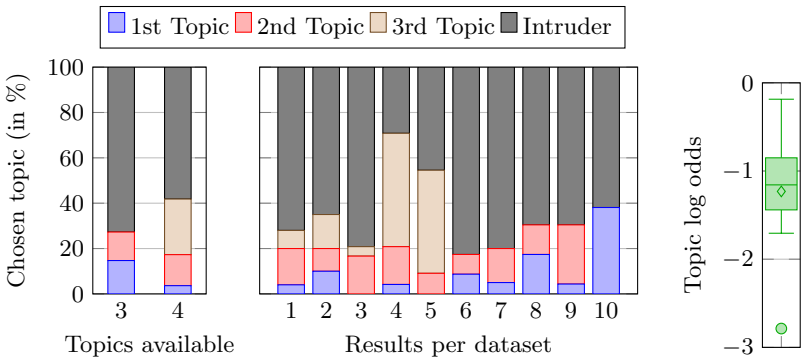

Fig. 6: Questionnaire results. Left: Average amount of topics chosen as intruder. Center: Amount of topics chosen as intruder for the single datasets. Right: The topic log odds o per dataset. The diamond marks the arithmetic mean.

results for the two groups of questions—those with 3 and 4 topics, respectively. The center of the figure shows the detailed results for each of the questions. The plot shows that in the majority of cases, the intruder was successfully identified by the participants. The results look slightly different for datasets 4 and 5. In both cases, the third topic is not strongly related to the dataset and has been chosen quite often as intruder. However, since the first and second topic have been chosen much less often for these datasets, the result shows that the ranking of the topics make sense, i.e., the participants were able to identify the first two topics as related to the given dataset.

On the right of Figure 6, there is a box plot for the topic log odd values that have been measured for the single documents. The average value across the 10 datasets is -1.23 with dataset 4 getting the worst value. This value is visible as an outlier in the lower part of the plot. This result is comparable to the results Chang et al. [9] present for various topic modeling models on two

different corpora. This confirms our finding that the human-readable topics fit to the RDF datasets to which they have been assigned. However, the experiment setup comes with two restrictions. First, we manually chose the RDF datasets for this experiment with the requirement that the participants of the questionnaire have to be able to easily understand the content of the chosen datasets. This may have introduced a bias. However, it can be assumed that the results would be less reliable if the datasets would have been selected randomly since the experiment setup suggested by Chang et al. [9] relies on the assumption that the participants understand the target object to which the topics have been assigned (in our case, the RDF dataset). Second, we made use of topic coherence measures to filter low-quality topics and we chose datasets that have at least two high-quality topics within their top-3 topics. It can be assumed that the topic log odd values would be lower if we would have included low-quality topics, since they are less likely interpretable by humans. However, a dataset that has mainly low-quality topics assigned could cause problems in a user application since no human-interpretable description of the dataset could be provided. We find that out of the 561 944 RDF datasets, to which LODCat could assign topics, only 220 datasets have not a single high-quality topic within their top-3 topics. Hence, the filtering of low-quality topics seems to have a minor impact on the number of RDF datasets for which LODCat is applicable.

5 Conclusion

Within this paper, we presented LODCat—an approach to support the exploration of the Data Web based on human-interpretable topics. With this approach, we ease the identification of RDF datasets that might be interesting to a user since they neither have to go through all available datasets nor do they need to read through the single RDF triples of a dataset. Instead, LODCat provides the user with human-interpretable topics that are automatically derived from a reference corpus and give the user an impression of a dataset's content. Our evaluation showed that LODCat was able to assign topics to 90% of a large, real-world set of datasets. The results of our questionnaire showed that humans agree with the topics that LODCat assigned to these RDF datasets. At the same time, our approach does neither need metadata of a dataset nor does it rely on manually created tags or classification systems. However, it can be easily combined with existing explorative search engines or integrated into dataset portals. Our future work includes the application of other topic modeling inference algorithms to create a hierarchy of topics.

Acknowledgements

This work has been supported by the Ministry of Culture and Science of North Rhine-Westphalia (MKW NRW) within the project SAIL under the grant no NW21-059D.

References

1. Java Platform Standard Ed. 8: Class Math. Website (2014), https://docs.oracle.com/javase/8/docs/api/java/lang/Math.html, last time accessed, May 18th, 2022.
2. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets with the void vocabulary. W3C Note, W3C (March 2011), http://www.w3.org/TR/2011/NOTE-void-20110303/
3. Asprino, L., Presutti, V.: Observing lod: Its knowledge domains and the varying behavior of ontologies across them. IEEE Access 11 (2023)
4. Beek, W., Rietveld, L., Bazoobandi, H.R., Wielemaker, J., Schlobach, S.: Lod laundromat: A uniform way of publishing other people's dirty data. In: The Semantic Web – ISWC 2014. pp. 213–228. Springer International Publishing, Cham (2014)
5. Bhatia, S., Lau, J.H., Baldwin, T.: Automatic labelling of topics with neural embeddings. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 953–963. The COLING 2016 Organizing Committee, Osaka, Japan (Dec 2016)
6. Blei, D.M.: Probabilistic topic models. Commun. ACM 55(4), 77–84 (Apr 2012)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (Mar 2003)
8. Brickley, D., Burgess, M., Noy, N.: Google dataset search: Building a search engine for datasets in an open web ecosystem. In: The World Wide Web Conference. p. 1365–1375. WWW '19, Association for Computing Machinery (2019)
9. Chang, J., Gerrish, S., Wang, C., Boyd-graber, J.L., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Advances in Neural Information Processing Systems 22, pp. 288–296. Curran Associates, Inc. (2009)
10. Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.D., Kacprzak, E., Groth, P.: Dataset search: a survey. The International Journal on Very Large Data Bases 29, 251–272 (2020)
11. Cyganiak, R., Reynolds, D.: The rdf data cube vocabulary. W3c recommendation (January 2014), http://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/
12. Devaraju, A., Berkovsky, S.: A hybrid recommendation approach for open research datasets. In: Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization. p. 207–211. UMAP '18, ACM (2018)
13. Ell, B., Vrandečić, D., Simperl, E.: Labels in the web of data. In: The Semantic Web – ISWC 2011. pp. 162–176. Springer Berlin Heidelberg (2011)
14. Heindorf, S., Blübaum, L., Düsterhus, N., Werner, T., Golani, V.N., Demir, C., Ngonga Ngomo, A.C.: Evolearner: Learning description logics with evolutionary algorithms. In: Proceedings of the ACM Web Conference 2022. pp. 818–828 (2022)
15. Hinneburg, A., Preiss, R., Schröder, R.: Topicexplorer: Exploring document collections with topic models. In: ECML PKDD 2012. Springer (2012)
16. Hoffman, M., Bach, F., Blei, D.: Online Learning for Latent Dirichlet Allocation. In: Advances in Neural Information Processing Systems. Curran Associates (2010)
17. Ji, S., Pan, S., Cambria, E., Marttinen, P., Philip, S.Y.: A survey on knowledge graphs: Representation, acquisition, and applications. IEEE Transactions on Neural Networks and Learning Systems (2021)
18. Kopsachilis, V., Vaitis, M.: Geolod: A spatial linked data catalog and recommender. Big Data and Cognitive Computing 5(2) (2021)
19. Kunze, S., Auer, S.: Dataset retrieval. In: Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on. pp. 1–8 (Sept 2013)

20. Lau, J.H., Grieser, K., Newman, D., Baldwin, T.: Automatic labelling of topic models. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. p. 1536–1545. HLT '11, Association for Computational Linguistics, USA (2011)
21. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 55–60. Association for Computational Linguistics (2014)
22. McCrae, J.P.: The Linked Open Data Cloud. Website (May 2021), https://www.lod-cloud.net/, last time accessed, August 24th 2021
23. Mohammadi, M.: (semi-) automatic construction of knowledge graph metadata. In: The Semantic Web: ESWC 2022 Satellite Events. pp. 171–178 (2022)
24. Ngomo, A.C.N., Sherif, M.A., Georgala, K., Hassan, M.M., Dreßler, K., Lyko, K., Obraczka, D., Soru, T.: Limes-a framework for link discovery on the semantic web. Journal of Web Semantics (2018)
25. Patni, H.: Linkedsensordata. Website in the web archive (September 2010), https://web.archive.org/web/20190816202119/http://wiki.knoesis.org/index.php/SSW_Datasets, last time accessed, May 11th, 2022.
26. Patni, H., Henson, C., Sheth, A.: Linked sensor data. In: 2010 International Symposium on Collaborative Technologies and Systems. pp. 362–370 (2010)
27. Paulheim, H., Hertling, S.: Discoverability of SPARQL Endpoints in Linked Open Data. In: Proceedings of the ISWC 2013 Posters & Demonstrations Track. vol. 1035, pp. 245–248. CEUR-WS.org, Aachen, Germany, Germany (2013)
28. Pietriga, E., Gözükan, H., Appert, C., Destandau, M., Čebirić, Š., Goasdoué, F., Manolescu, I.: Browsing Linked Data Catalogs with LODAtlas. In: The Semantic Web – ISWC 2018. pp. 137–153. Springer International Publishing, Cham (2018)
29. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA (May 2010)
30. Röder, M., Ngonga Ngomo, A.C., Ermilov, I., Both, A.: Detecting similar linked datasets using topic modelling. In: ESWC (2016)
31. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the WSDM (2015)
32. Saxena, A., Tripathi, A., Talukdar, P.: Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In: Proceedings of the 58th annual meeting of the association for computational linguistics (2020)
33. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the linked data best practices in different topical domains. In: The Semantic Web – ISWC 2014 (2014)
34. Singhal, A., Kasturi, R., Srivastava, J.: Datagopher: Context-based search for research datasets. In: Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014). pp. 749–756 (2014)
35. Sleeman, J., Finin, T., Joshi, A.: Topic modeling for rdf graphs. In: ISWC (2015)
36. Spahiu, B., Maurino, A., Meusel, R.: Topic profiling benchmarks in the linked open data cloud: Issues and lessons learned. Semantic Web 10(2), 329–348 (2019)
37. Spahiu, B., Porrini, R., Palmonari, M., Rula, A., Maurino, A.: Abstat: Ontology-driven linked data summaries with pattern minimalization. In: ESWC (2016)
38. Tzitzikas, Y., Manolis, N., Papadakos, P.: Faceted exploration of rdf/s datasets: A survey. Journal of Intelligent Information Systems 48(2), 329–364 (Apr 2017)
39. Vandenbussche, P.Y., Atemezing, G.A., Poveda-Villalón, M., Vatant, B.: Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web. Semantic Web 8(3), 437–452 (2017)