VOYAGE: A Large Collection of Vocabulary Usage in Open RDF Datasets

Qing Shi[1], Junrui Wang[1], Jeff Z. Pan[2], and Gong Cheng[1][0000−0003−3539−7776]

[1] State Key Laboratory for Novel Software Technology, Nanjing University, China
{qingshi,181840223}@smail.nju.edu.cn, gcheng@nju.edu.cn
[2] School of Informatics, University of Edinburgh, UK
http://knowledge-representation.org/j.z.pan/

Abstract. Shared vocabularies facilitate data integration and application interoperability on the Semantic Web. An investigation of how vocabularies are practically used in open RDF data, particularly with the increasing number of RDF datasets registered in open data portals, is expected to provide a measurement for the adoption of shared vocabularies and an indicator of the state of the Semantic Web. To support this investigation, we constructed and published VOYAGE, a large collection of vocabulary usage in open RDF datasets. We built it by collecting 68,312 RDF datasets from 517 pay-level domains via 577 open data portals, and we extracted 50,976 vocabularies used in the data. We analyzed the extracted usage data and revealed the distributions of frequency and diversity in vocabulary usage. We particularly characterized the patterns of term co-occurrence, and leveraged them to cluster vocabularies and RDF datasets as a potential application of VOYAGE. Our data is available from Zenodo at https://zenodo.org/record/7902675. Our code is available from GitHub at https://github.com/nju-websoft/VOYAGE.

Keywords: Open RDF data · Vocabulary usage · Term co-occurrence.

1 Introduction

The Semantic Web has entered its third decade. Driven by the ambitious vision of creating a Web where applications reach agreement on common vocabularies (i.e., sets of terms including classes and properties) to facilitate data integration and establish interoperability, we have witnessed the global adoption of vocabularies like schema.org [12] for annotating webpages to enhance Web search. Analyzing the practical usage of vocabularies could provide metrics and insights that are useful for measuring and understanding the adoption of vocabularies, as well as the state of the Semantic Web from the perspective of vocabulary usage.

Motivations. While the usage of a few exceptional vocabularies, such as schema.org, has been extensively analyzed, e.g., [21], such analyses are yet to be generalized to the majority of vocabularies used in open RDF data, mainly due to the lack of a large, representative, and timely data collection for this purpose. Note that our notion of open RDF data [25] goes beyond the conventional and

relatively small RDF documents, which are the main sources of previous data collections such as Billion Triple Challenge [14] and WebDataCommons [22]. Indeed, the increasing number of large *RDF datasets registered in open data portals* (ODPs), including but not limited to the Linked Open Data (LOD) Cloud, deserve more attention. It motivates us to particularly collect such RDF datasets for analyzing their vocabulary usage.

Furthermore, existing analyses of vocabulary usage are predominantly limited to simple metrics such as the frequency of occurrence of each individual vocabulary [23,7,31,29,24,14,17,13]. While this kind of elementary analysis is useful as it provides a basis, it has been desirable to further look into more advanced and useful indicators. In particular, investigating how terms are jointly used to describe an entity may exhibit notable *patterns of term co-occurrence*, and understanding such patterns is important to a wide range of Semantic Web research tasks and applications. For example, they have already played a vital role in RDF store optimization [28] and RDF data sampling [34], which rely on the "emergent schema" these patterns represent. It motivates us to extend our mining and analysis of vocabulary usage along this direction. We believe such analysis is valuable for vocabulary reuse when constructing knowledge graphs [27].

Resource. With the above motivations, we construct VOYAGE, short for VOcabularY usAGE, a large collection for analyzing vocabulary usage in open RDF datasets. Our data sources are 68,312 RDF datasets registered in 577 ODPs we collected. From the crawled and deduplicated RDF datasets, we extracted 50,976 vocabularies containing 62,864 classes and 842,745 properties that are actually used in the data, and we extracted their 767,976 patterns of co-occurrence in entity descriptions. We published the extracted usage data with provenance information. VOYAGE meets the following quality and availability criteria.

- It is publicly available and findable as a Zenodo dataset[3] with documentation explaining the structure of each JSON file, which is also summarized in the Resource Availability Statement at the end of the paper.
- It has metadata description available in multiple formats (e.g., DCAT).
- It is published at a persistent DOI URI.[4]
- It is associated with a canonical citation [30].
- It is open under the CC BY 4.0 license.

Applications. We ate our own dog food by analyzing the usage data provided by VOYAGE from multiple angles. Specifically, for both individual vocabularies and their patterns of co-occurrence, we characterized their usage by analyzing their frequencies across RDF datasets and their diversity in each RDF dataset, and obtained a set of new findings. Besides, as another potential application of our resource, we employed the patterns of co-occurrence to simultaneously cluster vocabularies and RDF datasets, and we found that the resulting clusters provided a reasonable complement to the conventional topic-based clustering, thus showing their value for downstream applications such as vocabulary recommender systems [6] and exploratory dataset search engines [5,26].

[3] https://zenodo.org/record/7902675
[4] https://doi.org/10.5281/zenodo.7902675

Table 1: Statistics about Data Collection (Notes: ODP catalogues may overlap. Inaccessible ODPs/datasets and non-RDF datasets are not counted.)

ODP catalogue	#ODP	(%)	#dataset	(%)	#triple	(%)
CKAN	109	(18.89%)	15,858	(22.00%)	397,404,207	(40.96%)
DataPortals.org	110	(19.06%)	25,341	(35.15%)	555,050,054	(57.21%)
DKAN	36	(6.24%)	3,007	(4.17%)	14,345,698	(1.48%)
Open Data Portal Watch	135	(37.40%)	37,407	(51.89%)	689,106,507	(71.02%)
Socrata	398	(68.98%)	55,653	(77.20%)	427,739,164	(44.09%)
LOD Cloud	1	(0.17%)	308	(0.43%)	128,902,453	(13.29%)
Total	577	(100.00%)	72,088	(100.00%)	970,258,378	(100.00%)
After deduplication			68,312		920,501,102	

Outline. We describe data collection in Section 2, analyze the usage of vocabularies in Section 3, extract and analyze their patterns of co-occurrence in Section 4, based on which we co-cluster vocabularies and RDF datasets in Section 5. Related work is discussed in Section 6 before we conclude in Section 7.

2 Data Collection

To construct VOYAGE, we collected RDF datasets from ODPs, deduplicated the crawled datasets, and extracted vocabularies used in the crawled RDF data.

2.1 RDF Dataset Crawling

To find as many ODPs as possible, we used five large catalogues of ODPs: CKAN,[5] DataPortals.org,[6] DKAN,[7] Open Data Portal Watch,[8] and Socrata.[9] They collectively listed 1,207 distinct ODPs where 576 ODPs were accessible at the time of crawling (i.e., Q1 2022). We manually submitted the LOD Cloud as an ODP to our crawler, resulting in a total of 577 ODPs to be accessed.

For each ODP, we invoked its API to retrieve the metadata of all the *datasets* registered in this ODP. Datasets providing at least one dump file in an RDF format (e.g., RDF/XML, Turtle, N-Triples) were identified as *RDF datasets*. We successfully downloaded and parsed the dump files of 72,088 RDF datasets using Apache Jena,[10] and extracted a total of 970,258,378 RDF triples.

Table 1 summarizes the data sources of VOYAGE.

[5] https://ckan.org/
[6] http://dataportals.org/
[7] https://getdkan.org/
[8] https://data.wu.ac.at/portalwatch/
[9] https://dev.socrata.com/
[10] https://jena.apache.org/

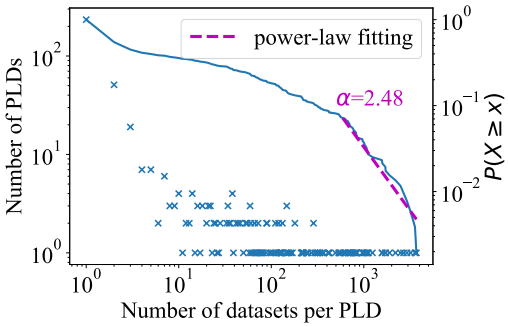


Fig. 1: Distribution (crosses) and cumulative probability distribution (curve) of the number of RDF datasets crawled from a PLD.

Table 2: Top-Ranked PLDs

PLD	#dataset	(%)
datos.gov.co	3,703	(5.42%)
cityofnewyork.us	3,575	(5.23%)
socrata.com	3,209	(4.70%)
smcgov.org	2,926	(4.28%)
dati.lombardia.it	2,683	(3.93%)
utah.gov	2,497	(3.66%)
wa.gov	2,199	(3.22%)
edmonton.ca	1,894	(2.77%)
ny.gov	1,721	(2.52%)
seattle.gov	1,711	(2.50%)

2.2 RDF Dataset Deduplication

We observed that the same RDF dataset might have been registered in multiple ODPs. However, we found it difficult to accurately identify duplicate datasets only based on their metadata, e.g., a certain dataset had different titles, different descriptions, and different download URLs in its metadata registered in different ODPs. Therefore, we employed the actual RDF data to detect duplicates.

Specifically, we regarded two crawled RDF datasets as duplicates if they were crawled from the same pay-level domain (PLD) and their dump files were parsed into two isomorphic RDF graphs. We used the BLabel algorithm [15] to test RDF graph isomorphism, and we followed [32] to decompose each RDF graph which may have a large size into a unique set of practically very small subgraphs to accelerate isomorphism testing.

After deduplication, among the 72,088 RDF datasets we kept 68,312 distinct ones, containing a total of 920,501,102 RDF triples. They were crawled from 517 PLDs. Figure 1 plots the distribution of the number of RDF datasets crawled from a PLD. The distribution appears uneven: while 285 PLDs (55.13%) contribute at most 2 RDF datasets, some PLDs contribute several thousand RDF datasets. Motivated by its highly skewed shape, we tried to fit a power law using powerlaw.[11] According to the Kolmogorov-Smirnov test, the null hypothesis that the tail of the distribution ($X \geq 593$) fits a power law with $\alpha = 2.48$ is accepted ($p = 0.89$). However, no single PLD can dominate: as shown in Table 2, a single PLD contributes at most 5.42% of all the RDF datasets we crawled, which is important as it shows the diversity of our data sources.

2.3 Vocabulary Extraction

We extracted vocabularies that are actually used (i.e., instantiated) in the crawled RDF data. For example, a class is used in an RDF dataset if its IRI appears

[11] https://github.com/jeffalstott/powerlaw

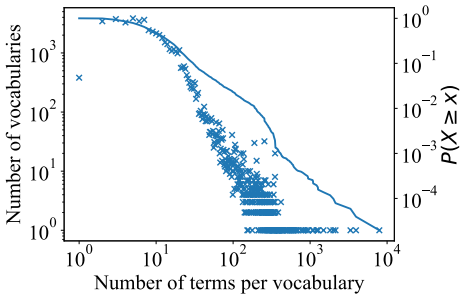


Fig. 2: Distribution (crosses) and cumulative probability distribution (curve) of the number of used terms in a vocabulary.

as the object of an RDF triple in this dataset where the predicate is `rdf:type`, and a property is used if its IRI appears as the predicate of an RDF triple in this dataset. Classes and properties are collectively called *terms*. A *vocabulary* is a set of terms denoted by IRIs starting with a common namespace IRI. A vocabulary is used in an RDF dataset if any of its terms is used in this dataset.

From the 68,312 RDF datasets we crawled, we extracted 62,864 distinct classes and 842,745 distinct properties, belonging to 50,976 distinct vocabularies. Figure 2 plots the distribution of the number of used terms in a vocabulary, with a median of 9 and a rejected power-law fitting ($p = 1.98\text{E}{-13}$). While most vocabularies are small, the largest vocabulary contains 7,930 terms.

3 Frequency and Diversity in Vocabulary Usage

To characterize the usage of vocabularies provided by VOYAGE, we firstly analyzed their frequencies across all the crawled RDF datasets and their diversity in each dataset, providing a basis for the subsequent experiments. In this analysis, we excluded five language-level vocabularies since they were found to be trivially used in many RDF datasets, i.e., `xsd`,[12] `rdf`,[13] `rdfs`,[14] `owl`,[15] and `skos`.[16]

3.1 Frequency Analysis

We analyzed to what extent vocabularies have been shared among open RDF datasets by calculating their dataset frequencies. Figure 3 plots the distribution of the number of RDF datasets using a vocabulary. Fitting the tail of the distribution ($X \geq 837$) to a power law with $\alpha = 2.58$ is accepted ($p = 0.99$). Most vocabularies (87.41%) are only used in a single RDF dataset, but there are also

[12] http://www.w3.org/2001/XMLSchema#
[13] http://www.w3.org/1999/02/22-rdf-syntax-ns#
[14] http://www.w3.org/2000/01/rdf-schema#
[15] http://www.w3.org/2002/07/owl#
[16] http://www.w3.org/2004/02/skos/core#

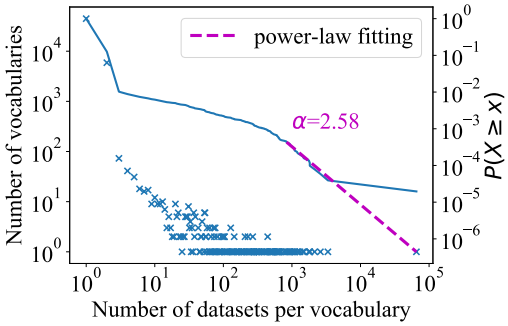


Fig. 3: Distribution (crosses) and cumulative probability distribution (curve) of the number of RDF datasets using a vocabulary.

Table 3: Top-Ranked Vocabularies

Vocabulary	#PLD	(%)
`foaf`	329	(63.64%)
`dcterms`	183	(35.40%)
`socrata`	162	(31.33%)
`dc`	117	(22.63%)
`geo`	58	(11.22%)
`void`	58	(11.22%)
`admin`	39	(7.54%)
`schema`	35	(6.77%)
`dcat`	34	(6.58%)
`cc`	27	(5.22%)

317 vocabularies used in at least ten RDF datasets. As shown in Table 3, four vocabularies are very popular and are used in RDF datasets from more than one hundred PLDs, i.e., `foaf`,[17] `dcterms`,[18] `socrata`,[19] and `dc`.[20] These observations suggest that *vocabulary sharing is common among open RDF datasets, although only a small proportion of vocabularies are widely shared.*

The Semantic Web community may not be very familiar with `socrata`. Although this vocabulary is not dereferenceable, it is used in RDF datasets from 162 PLDs, which represent an important part of RDF data in the real world.

3.2 Diversity Analysis

We analyzed to what extent a multiplicity of terms and vocabularies have been used in an open RDF dataset by calculating their diversity in each dataset. Figure 4 plots the distribution of the number of terms used in an RDF dataset, with a median of 11 and a rejected power-law fitting ($p = 1.60\text{E}{-10}$). In particular, four RDF datasets exhibit a complex constitution of schema where more than one thousand terms are used. Figure 5 plots the distribution of the number of vocabularies used in an RDF dataset, with a median of 3. Fitting the tail of the distribution ($X \geq 10$) to a power law with $\alpha = 3.04$ is accepted ($p = 0.43$). One RDF dataset entitled "TaxonConcept Knowledge Base" notably uses 458 vocabularies, being the largest number among all the crawled RDF datasets. These observations suggest that *it is common for an open RDF dataset to use multiple vocabularies and diverse terms*, which motivated the following analysis.

[17] http://xmlns.com/foaf/0.1/
[18] http://purl.org/dc/terms/
[19] http://www.socrata.com/rdf/terms#
[20] http://purl.org/dc/elements/1.1/

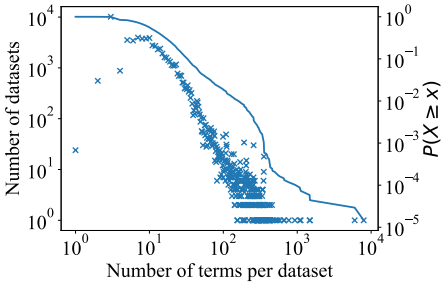


Fig. 4: Distribution (crosses) and cumulative probability distribution (curve) of the number of terms used in an RDF dataset.

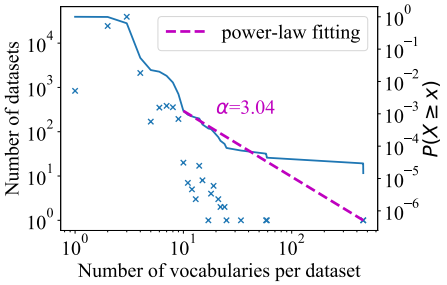


Fig. 5: Distribution (crosses) and cumulative probability distribution (curve) of the number of vocabularies used in an RDF dataset.

4 Patterns of Term Co-Occurrence

The observations obtained in Section 3 indicate the possibility that some terms and vocabularies have been *jointly used* in many RDF datasets. In particular, terms may have been jointly used to describe many entities and hence exhibit a notable *pattern of term co-occurrence* in entity descriptions. Such patterns represent an "emergent schema" and are of particular interest. Indeed, they have been exploited in a variety of research tasks including RDF store optimization [28] and RDF data sampling [34]. Therefore, we extracted them from the crawled RDF data as part of VOYAGE, and characterized their usage by analyzing their frequencies across all the crawled RDF datasets and their diversity in each dataset.

4.1 Term Co-Occurrence Extraction

Following [34], we refer to a pattern of term co-occurrence as an *entity description pattern*, or *EDP* for short. Specifically, in an RDF dataset T which contains a set of RDF triples, the EDP of an entity e consists of the sets of all the classes (`C`), forward properties (`FP`), and backward properties (`BP`) used in T to describe e:

$$\begin{aligned} \texttt{EDP}(e) &= \langle \texttt{C}(e),\ \texttt{FP}(e),\ \texttt{BP}(e) \rangle\,, \\ \texttt{C}(e) &= \{c : \exists \langle e,\ \texttt{rdf:type},\ c \rangle \in T\}\,, \\ \texttt{FP}(e) &= \{p : \exists \langle e,\ p,\ o \rangle \in T,\ p \neq \texttt{rdf:type}\}\,, \\ \texttt{BP}(e) &= \{p : \exists \langle s,\ p,\ e \rangle \in T\}\,. \end{aligned} \tag{1}$$

For example, Figure 6 illustrates the description of an entity in an RDF dataset, and Table 4 shows the EDP of this entity.

From the descriptions of 58,777,001 entities in the 68,312 RDF datasets we crawled, we extracted 767,976 distinct EDPs. On average, an EDP consists of 1.14 classes, 63.31 forward properties, and 0.15 backward property, i.e., most properties used in open RDF datasets are literal-valued. For example, from one

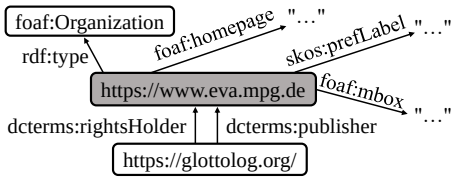


Fig. 6: An entity denoted by `https://www.eva.mpg.de` described in an RDF dataset.

Table 4: An Example of EDP

C	`foaf:Organization`

FP	`foaf:homepage`
	`foaf:mbox`
	`skos:prefLabel`

BP	`dcterms:publisher` `dcterms:rightsHolder`

RDF dataset entitled "Higher Education Cost Data From IPEDS Utah 2000-2010", we extracted an EDP consisting of 890 terms, most of which are literal-valued properties for describing various statistic data for a year, being the largest number among all the extracted EDPs.

We particularly examined the EDPs extracted from the RDF datasets in the LOD Cloud as they might be of special interest to the Semantic Web community. These EDPs consist of relatively more classes ($2.45 > 1.14$) and more backward properties ($0.55 > 0.15$) but fewer forward properties ($20.80 < 63.31$), i.e., the entities in the LOD Cloud are better typed and better interlinked with each other. Since the presence of types and the completeness of interlinks are important metrics for assessing the syntactic validity and completeness of RDF data, respectively [36], these observations suggest that *the RDF datasets in the LOD Cloud exhibit relatively high data quality in terms of typing and interlinking.*

4.2 Frequency Analysis

We analyzed to what extent EDPs have been shared among open RDF datasets by calculating their dataset frequencies. Figure 7 plots the distribution of the number of RDF datasets using an EDP. Fitting the tail of the distribution ($X \geq$

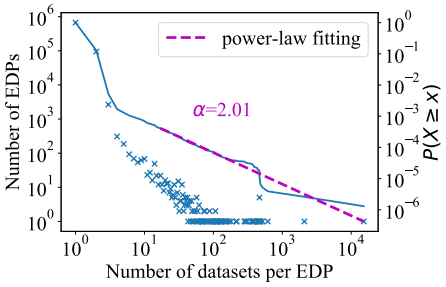


Fig. 7: Distribution (crosses) and cumulative probability distribution (curve) of the number of RDF datasets using an EDP.

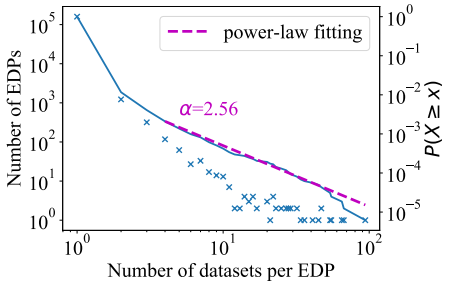


Fig. 8: Distribution (crosses) and cumulative probability distribution (curve) of the number of RDF datasets in the LOD Cloud using an EDP.

Table 5: Top-Ranked Singleton (Top) and Non-Singleton (Bottom) EDPs

EDP	#PLD
C = ∅, FP = ∅, BP = {foaf:homepage}	128
C = ∅, FP = ∅, BP = {foaf:document}	89
C = ∅, FP = ∅, BP = {foaf:depiction}	65
C = ∅, FP = ∅, BP = {dcterms:license}	62
C = ∅, FP = ∅, BP = {foaf:workplaceHomepage}	57
C = ∅, FP = {socrata:rowID, rdfs:member}, BP = ∅	65
C = {foaf:PersonalProfileDocument}, FP = {admin:errorReportsTo, admin:generatorAgent, foaf:maker, foaf:primaryTopic}, BP = ∅	27
C = {foaf:PersonalProfileDocument}, FP = {foaf:maker, foaf:primaryTopic}, BP = ∅	17
C = {foaf:Document}, FP = {dcterms:hasFormat, foaf:primaryTopic, foaf:topic}, BP = ∅	11
C = {foaf:Document}, FP = {dc:format, rdfs:label}, BP = {dcterms:hasFormat}	11

Table 6: Top-Ranked Singleton (Top) and Non-Singleton (Bottom) EDPs in the LOD Cloud

EDP	#PLD
C = ∅, FP = ∅, BP = {dcterms:license}	42
C = ∅, FP = ∅, BP = {dcterms:subject}	32
C = ∅, FP = ∅, BP = {foaf:homepage}	31
C = ∅, FP = ∅, BP = {dcterms:creator}	31
C = ∅, FP = ∅, BP = {void:feature}	23
C = {foaf:Organization}, FP = {foaf:mbox, foaf:homepage, skos:prefLabel}, BP = {dcterms:rightsHolder, dcterms:publisher}	7
C = {dcmit:Software}, FP = {dcterms:identifier}, BP = ∅	7
C = {void:Dataset}, FP = {skos:example, skos:hiddenLabel, void:rootResource, skos:prefLabel}, BP = {void:subset, void:rootResource}	6
C = ∅, FP = ∅, BP = {dcterms:creator, dcterms:publisher}	5
C = {owl:Thing}, FP = ∅, BP = {dcterms:conformsTo}	5

17) to a power law with $\alpha = 2.01$ is accepted ($p = 0.95$). Most EDPs (87.14%) are only used in a single RDF dataset, but there are also 464 EDPs used in more than ten RDF datasets, and 53 EDPs used in more than one hundred RDF datasets. These observations suggest that despite the decentralized nature of the Semantic Web, *a few patterns of term co-occurrence for describing entities have emerged and are shared among open RDF datasets*. Table 5 illustrates the most popular singleton (i.e., consisting of a single term) and non-singleton EDPs used in RDF datasets from tens to hundreds of PLDs.[21]

We particularly restricted the above distribution to the RDF datasets in the LOD Cloud. As shown in Figure 8, fitting the tail of the distribution ($X \geq 4$) to a power law with $\alpha = 2.56$ is also accepted ($p = 0.95$). There are 55 EDPs used in more than ten RDF datasets in the LOD Cloud. These observations suggest that *the RDF datasets in the LOD Cloud also share a few patterns of co-occurrence for describing entities*. However, the most popular EDPs used in the LOD Cloud illustrated in Table 6 differ from those in Table 5. There are

[21] EDPs that solely consist of terms in the five language-level vocabularies (i.e., `xsd`, `rdf`, `rdfs`, `owl`, and `skos`) are excluded from Table 5 and Table 6.

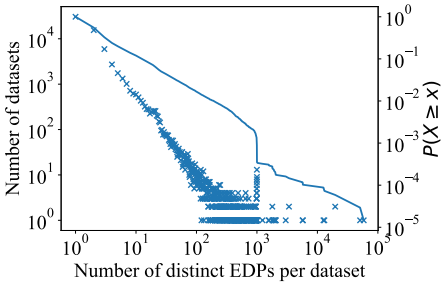


Fig. 9: Distribution (crosses) and cumulative probability distribution (curve) of the number of distinct EDPs used in an RDF dataset.

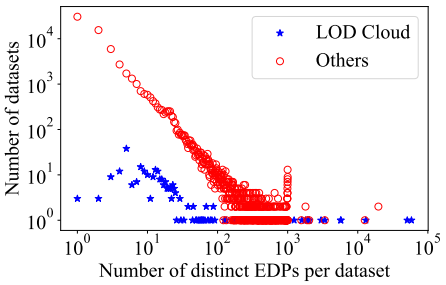


Fig. 10: Distribution of the number of distinct EDPs used in an RDF dataset in or outside the LOD Cloud.

descriptions of organizations and software in Table 6, not limited to descriptions of documents in Table 5.

4.3 Diversity Analysis

We analyzed to what extent a multiplicity of distinct EDPs have been used in an open RDF dataset by calculating their diversity in each dataset. Figure 9 plots the distribution of the number of distinct EDPs used in an RDF dataset, with a median number of 2 and a rejected power-law fitting ($p = 7.92\text{E}{-4}$). Nearly half of the RDF datasets (44.70%) use only a single EDP that describes all the entities, i.e., each of these RDF datasets describes all the entities in a homogeneous manner. There are also 24 RDF datasets using more than ten thousand distinct EDPs. For example, one RDF dataset entitled "Open Food Facts" describes entities in a highly heterogeneous manner, using 19,693 distinct EDPs which represent different combinations of nutrition facts about food products.

We particularly divided the distribution in Figure 9 into two distributions in Figure 10: one over all the RDF datasets in the LOD Cloud, and the other over those outside. The two distributions are noticeably different. The RDF datasets in the LOD Cloud use relatively more distinct EDPs in terms of median ($14 > 2$). The distribution over the LOD Cloud peaks at 5 EDPs, while most RDF datasets outside the LOD Cloud (67.68%) use at most 2 distinct EDPs. These observations suggest that *most RDF datasets outside the LOD Cloud contain nearly homogeneous entity descriptions, and the RDF datasets in the LOD Cloud describe entities in a relatively heterogeneous manner.*

One potential application of this kind of analysis is for choosing a suitable RDF store. For example, among RDF store solutions [1], a property table stores the description of each entity in a row, and each column stores the values of a distinct property, thereby allowing to retrieve entities having multiple specified property values without join operations. Property table is suitable for storing an RDF dataset using one or a few distinct EDPs, since otherwise there will be

many null values which waste space. By contrast, vertical partitioning separately stores the values of different properties in different tables. A triple table stores all the RDF triples in a single table. Vertical partitioning and triple table are more suitable for storing an RDF dataset using a large number of distinct EDPs, since there will be no null values despite more joins at query time.

5 Clusters of Vocabularies Based on Co-Occurrence

In this section, we exemplify another potential application of the extracted EDPs. We leveraged them to cluster vocabularies and RDF datasets. The generated clusters can be used in recommendation to support serendipitous discovery of vocabularies and RDF datasets. As we will see in this section, such EDP-based clusters are complementary to the conventional topic-based clusters.

5.1 Clustering Method

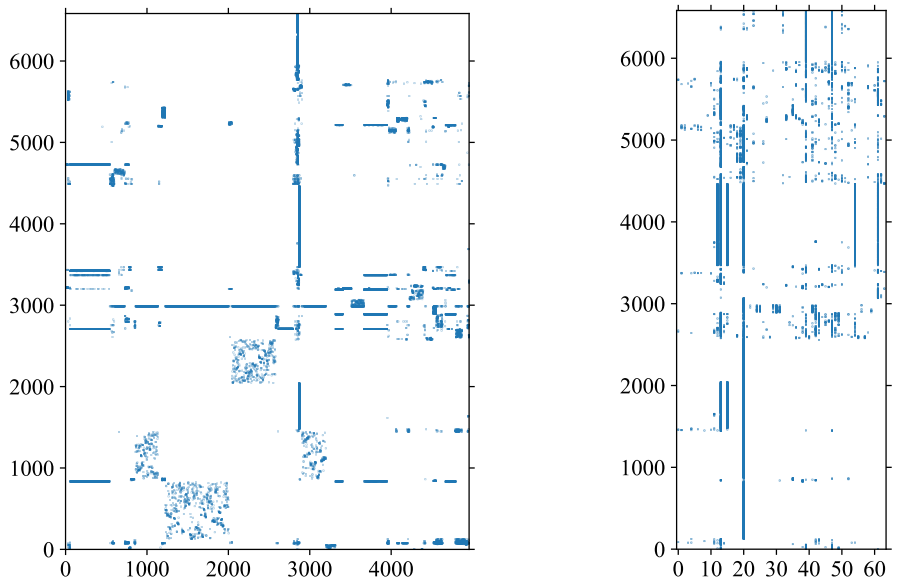
Graph Construction. Our clustering method relies on the following tripartite relation between RDF datasets, EDPs, and vocabularies: RDF datasets use EDPs which consist of terms belonging to vocabularies. To represent this relation, we constructed a tripartite *dataset-EDP-vocabulary graph*, where nodes represent RDF datasets, EDPs, and vocabularies; edges connect each EDP with all the RDF datasets using it, and with all the vocabularies its constituent terms belong to. For example, the dataset illustrated in Figure 6 and the EDP illustrated in Table 4 are represented by two adjacent nodes; the EDP node is also adjacent with three vocabulary nodes representing `foaf`, `skos`, and `dcterms`.

Graph Clustering. Our idea is to exploit vocabulary co-occurrence in RDF data to simultaneously cluster RDF datasets and vocabularies via their connections with EDPs. We converted it into the problem of finding two co-clusterings on the dataset-EDP-vocabulary graph: one between EDPs and RDF datasets, and the other between EDPs and vocabularies, subject to that the consensus between the clusters of EDPs in the two co-clusterings should be maximized.

We solved this problem by using MV-ITCC [35], which is a multi-view co-clustering algorithm. Specifically, we treated EDPs as the main items to be clustered by the algorithm, and treated RDF datasets and vocabularies as items' features in two different views. We used MV-ITCC to compute a two-sided two-view clustering, where the items (i.e., EDPs) were clustered by exploiting the agreement and disagreement between different views, and the features in each view (i.e., RDF datasets or vocabularies) were simultaneously clustered.

5.2 Implementation Details

Preprocessing. We constructed a dataset-EDP-vocabulary graph from VOYAGE. Subject to the scalability of the MV-ITCC algorithm, we performed the following preprocessing to reduce the size of the graph and extract its core structure. First, we removed all the infrequent vocabularies used in RDF datasets



(a) EDP-dataset matrix.

(b) EDP-vocabulary matrix.

Fig. 11: Adjacency matrix representation of the dataset-EDP-vocabulary graph.

from less than five PLDs, and we removed five language-level vocabularies as described in Section 3 as well as the `socrata` vocabulary since they have been trivially used in many RDF datasets. We also removed the RDF datasets that only use these vocabularies, and removed the EDPs whose constituent terms only belong to these vocabularies. Second, we merged the nodes representing EDPs whose constituent terms belong to exactly the same set of vocabularies, i.e., adjacent with the same set of vocabulary nodes in the graph. It actually generalized EDPs from the term level to the more coarse-grained vocabulary level. Finally, we removed all the isolated nodes from the graph.

Parameter Selection. Applying the MV-ITCC algorithm to the constructed dataset-EDP-vocabulary graph required specifying the expected numbers of clusters of RDF datasets, EDPs, and vocabularies, denoted by k_d, k_e, and k_v, respectively. To find their optimal setting, we heuristically searched each parameter from $\lceil 0.5\sqrt{\frac{n}{2}} \rceil$ to $\lceil 1.5\sqrt{\frac{n}{2}} \rceil$ in $\lceil 0.1\sqrt{\frac{n}{2}} \rceil$ increments, where n denotes the number of items to be clustered. Specifically, for each k_e, we found an optimal setting of k_d and k_v as follows. For each k_d-k_v combination, we employed the ITCC algorithm to compute a co-clustering of the bipartite EDP-dataset subgraph, and a co-clustering of the bipartite EDP-vocabulary subgraph. We measured the similarity between the clusters of EDPs in the two co-clusterings by calculating their adjusted rand index (ARI), and chose the k_d-k_v combination featuring the best

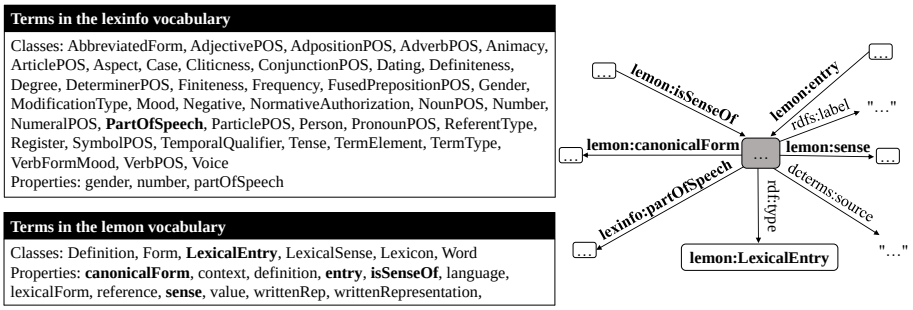


Fig. 12: Left: two vocabularies with few overlaps in the names of their constituent terms. Right: the description of an entity using terms in both vocabularies.

ARI to form a k_e-k_d-k_v combination. Finally, we chose the k_e-k_d-k_v combination featuring the highest quality of co-clustering measured by Silhouette Coefficient.

5.3 Cluster Analysis

After preprocessing and parameter selection, our dataset-EDP-vocabulary graph constructed from VOYAGE was reduced to 4,958 RDF dataset nodes, 6,584 merged EDP nodes, and 64 vocabulary nodes, which were grouped into 45 clusters of RDF datasets and 6 clusters of vocabularies intermediated by 52 clusters of merged EDPs. Figure 11 visualizes the adjacency matrix representation of its two bipartite subgraphs, where rows and columns are rearranged according to the clusters. Both matrices contain many noticeable dense sub-matrices representing subgraphs where nodes are densely connected, i.e., they represent cohesive clusters. This observation suggests that *open RDF datasets and vocabularies both exhibit distinguishable clusters based on the patterns of vocabulary co-occurrence.*

We compared our co-occurrence-based clusters with conventional topic-based clusters of vocabularies generated by Latent Dirichlet Allocation (LDA), which was fed with a set of pseudo documents each consisting of the names of all the terms in a vocabulary. We found 678 pairs of vocabularies that were clustered by our approach but not by LDA. For example, Figure 12 illustrates two vocabularies, `lexinfo`[22] and `lemon`,[23] having few overlaps in the names of their constituent terms, thus not clustered by LDA. By contrast, their constituent terms co-occur in 4 EDPs, thus clustered by our approach. This result is reasonable because `lexinfo` is exactly a vocabulary created to be used with `lemon`. These observations suggest that *our co-occurrence-based clusters of vocabularies provide a useful complement to the conventional topic-based clusters.*

[22] http://www.lexinfo.net/ontology/2.0/lexinfo#
[23] http://lemon-model.net/lemon#

Table 7: Existing Analyses of Vocabulary Usage

	Data collection	Vocabulary usage analysis
LDOW'12 [23]	crawled RDF documents	frequency
JoWS'12 [16]	crawled RDF documents	frequency, co-occurrence
CSWS'12 [7]	crawled RDF documents	frequency
CSWS'12 [31]	crawled RDF documents	frequency
WI'12 [2]	crawled RDF documents	frequency, co-occurrence
ESWC'13 [10]	crawled RDF documents	co-occurrence
COLD'13 [9]	crawled RDF documents	co-occurrence, dynamics
ISWC'13 [4]	crawled RDF documents	frequency, co-occurrence
JoWS'13 [8]	crawled RDF documents	frequency, co-occurrence
ISWC'14 [29]	crawled RDF documents	frequency
ISWC'14 [22]	crawled RDF documents	frequency, co-occurrence
DPD'15 [11]	crawled RDF documents	co-occurrence
OIR'17 [24]	crawled RDF documents	frequency
ISWC'19 [14]	crawled RDF documents	frequency
ISWC'19 [17]	EuroDP datasets	frequency
JDIQ'20 [13]	LOD datasets	frequency

6 Related Work

We are among the first to construct and publish a large collection that is specifically for analyzing vocabulary usage in open RDF datasets. Our analysis offers new measures that differ from previous analyses of vocabulary usage in Table 7.

Our analysis is focused on the patterns of term co-occurrence represented by EDPs, which have also been considered in previous analyses. For example, Dividino et al. [9] investigated the dynamics of EDPs. Gottron et al. [10,11] compared the informativeness of class sets and property sets. However, their analyses treated all the crawled RDF documents as a whole, whereas we separately analyzed each RDF dataset and obtained new findings, e.g., we characterized the diversity of vocabularies and EDPs used in each RDF dataset. In [16,2,4,8,22], RDF datasets were also separately analyzed, but these analyses were relatively coarse-grained—characterizing vocabulary co-occurrence in RDF datasets, whereas our more fine-grained analysis characterizes term co-occurrence in entity descriptions to provide a more accurate measurement.

Many analyses of vocabulary usage did not address co-occurrence as ours but they only reported frequencies [23,7,31,29,24,14,17,13]. A few researches were focused on the usage of a particular vocabulary such as GoodRelations [18] or schema.org [21]. Instead of vocabulary usage, some works analyzed vocabulary definitions [20,19] and inter-vocabulary links derived from their definitions [33,3]. All these analyses are considered orthogonal to our analysis of co-occurrence.

Another distinguishing feature of our VOYAGE is that we collected RDF datasets from ODPs. By contrast, previous analyses mostly crawled RDF documents from the Web and then heuristically grouped RDF documents into RDF datasets by PLD [16,29]. Such heuristic construction of pseudo RDF datasets

may suffer from inaccuracy. A recent study gave attention to the RDF datasets in the LOD Cloud [13]. We further extended the scope by also crawling RDF datasets registered in many other ODPs. Our comparative analysis of the RDF datasets in and outside the LOD Cloud revealed their large differences.

7 Conclusion

We have constructed and published VOYAGE, a large collection of vocabulary usage from a diverse set of open RDF datasets, with a particular focus on the patterns of term co-occurrence in entity descriptions. We conclude the paper with a discussion of its impact, reusability, and future plans.

Impact. Different from previous data collections, our VOYAGE is sourced from RDF datasets registered in ODPs, and provides the usage of vocabularies, terms, and their patterns of co-occurrence extracted from each RDF dataset. It facilitates measuring the adoption of vocabularies and reviewing the state of the Semantic Web from a new angle. Indeed, our analysis of frequency and diversity in vocabulary usage has revealed some new findings of interest to the Semantic Web community and the open data community. Our observations collectively reflect that the Semantic Web is not too far away from establishing interoperability via shared vocabularies. This result is expected to encourage the continued adoption of Semantic Web technologies.

Reusability. Our presented analysis of VOYAGE is not exhaustive, and VOYAGE has the potential to be used in further analyses. For example, in some experiments we ablated the LOD Cloud to be specifically analyzed, and one may partition the RDF datasets and/or vocabularies in VOYAGE in a different way to perform comparative analysis. VOYAGE can also be applied in other scenarios. For example, we have showed its usefulness in vocabulary clustering, and one may explore its value for other tasks. Reusing and extending VOYAGE is easy since we have documented the structure of its JSON files.

Plans for the Future. VOYAGE is sourced from RDF datasets registered in ODPs, which is complementary to the sources of other existing data collections such as WebDataCommons. Therefore, we plan to extend VOYAGE with vocabulary usage extracted from the latest version of WebDataCommons[24] and from a re-crawl of the Billion Triple Challenge dataset.[25] As for long-term maintenance, we plan to periodically (i.e., yearly or more frequently) recollect all the data sources and publish updated vocabulary usage.

Resource Availability Statement: VOYAGE is available from Zenodo at https://zenodo.org/record/7902675. For each of the accessed 577 ODPs, its name, URL, API type, API URL, and the IDs of RDF datasets collected from it are given in `odps.json`. For each of the crawled 72,088 RDF datasets, its ID, title, description, author, license, dump file URLs, and PLDs are given in `datasets.json`. The IDs of the deduplicated 68,312 RDF datasets and whether they are in the

[24] http://webdatacommons.org/structureddata/#results-2022-1
[25] https://zenodo.org/record/2634588

LOD Cloud are given in `deduplicated_datasets.json`. The extracted 62,864 classes, 842,745 properties, and the IDs of RDF datasets using each term are given in `terms.json`. The extracted 50,976 vocabularies, the classes and properties in each vocabulary, and the IDs of RDF datasets using each vocabulary are given in `vocabularies.json`. The extracted 767,976 distinct EDPs and the IDs of RDF datasets using each EDP are given in `edps.json`. The clusters of vocabularies generated by MV-ITCC and LDA are given in `clusters.json`. All the experiments presented in the paper can be reproduced from the above files, for which some helpful scripts are available from GitHub at https://github.com/nju-websoft/VOYAGE.

Acknowledgements This work was supported by the NSFC (62072224) and the Chang Jiang Scholars Program (J2019032).

References

1. Ali, W., Saleem, M., Yao, B., Hogan, A., Ngomo, A.N.: A survey of RDF stores & SPARQL engines for querying knowledge graphs. VLDB J. **31**(3), 1–26 (2022). https://doi.org/10.1007/s00778-021-00711-3
2. Ashraf, J., Hussain, O.K.: Analysing the use of ontologies based on usage network. In: WI 2012. pp. 540–544 (2012). https://doi.org/10.1109/WI-IAT.2012.203
3. Asprino, L., Beek, W., Ciancarini, P., van Harmelen, F., Presutti, V.: Observing LOD using equivalent set graphs: It is mostly flat and sparsely linked. In: ISWC 2019, Part I. pp. 57–74 (2019). https://doi.org/10.1007/978-3-030-30793-6_4
4. Bizer, C., Eckert, K., Meusel, R., Mühleisen, H., Schuhmacher, M., Völker, J.: Deployment of RDFa, Microdata, and Microformats on the Web - A quantitative analysis. In: ISWC 2013, Part II. pp. 17–32 (2013). https://doi.org/10.1007/978-3-642-41338-4_2
5. Brickley, D., Burgess, M., Noy, N.F.: Google dataset search: Building a search engine for datasets in an open Web ecosystem. In: WWW 2019. pp. 1365–1375 (2019). https://doi.org/10.1145/3308558.3313685
6. Cheng, G., Gong, S., Qu, Y.: An empirical study of vocabulary relatedness and its application to recommender systems. In: ISWC 2011, Part I. pp. 98–113 (2011). https://doi.org/10.1007/978-3-642-25073-6_7
7. Cheng, G., Liu, M., Qu, Y.: NJVR: the NanJing vocabulary repository. In: CSWS 2012. pp. 265–272 (2012). https://doi.org/10.1007/978-1-4614-6880-6_23
8. Cheng, G., Qu, Y.: Relatedness between vocabularies on the Web of data: A taxonomy and an empirical study. J. Web Semant. **20**, 1–17 (2013). https://doi.org/10.1016/j.websem.2013.02.001
9. Dividino, R.Q., Scherp, A., Gröner, G., Grotton, T.: Change-a-LOD: does the schema on the Linked Data Cloud change or not? In: COLD 2013 (2013)
10. Gottron, T., Knauf, M., Scheglmann, S., Scherp, A.: A systematic investigation of explicit and implicit schema information on the Linked Open Data Cloud. In: ESWC 2013. pp. 228–242 (2013). https://doi.org/10.1007/978-3-642-38288-8_16
11. Gottron, T., Knauf, M., Scherp, A.: Analysis of schema structures in the Linked Open Data graph based on unique subject URIs, pay-level domains, and vocabulary usage. Distributed Parallel Databases **33**(4), 515–553 (2015). https://doi.org/10.1007/s10619-014-7143-0

12. Guha, R.V., Brickley, D., Macbeth, S.: Schema.org: evolution of structured data on the Web. Commun. ACM **59**(2), 44–51 (2016). https://doi.org/10.1145/2844544
13. Haller, A., Fernández, J.D., Kamdar, M.R., Polleres, A.: What are links in Linked Open Data? A characterization and evaluation of links between knowledge graphs on the Web. ACM J. Data Inf. Qual. **12**(2), 9:1–9:34 (2020). https://doi.org/10.1145/3369875
14. Herrera, J., Hogan, A., Käfer, T.: BTC-2019: the 2019 Billion Triple Challenge dataset. In: ISWC 2019, Part II. pp. 163–180 (2019). https://doi.org/10.1007/978-3-030-30796-7_11
15. Hogan, A.: Canonical forms for isomorphic and equivalent RDF graphs: Algorithms for leaning and labelling blank nodes. ACM Trans. Web **11**(4), 22:1–22:62 (2017). https://doi.org/10.1145/3068333
16. Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., Decker, S.: An empirical survey of Linked Data conformance. J. Web Semant. **14**, 14–44 (2012). https://doi.org/10.1016/j.websem.2012.02.001
17. Ibáñez, L., Millard, I., Glaser, H., Simperl, E.: An assessment of adoption and quality of Linked Data in European open government data. In: ISWC 2019, Part II. pp. 436–453 (2019). https://doi.org/10.1007/978-3-030-30796-7_27
18. Kowalczuk, E., Potoniec, J., Lawrynowicz, A.: Extracting usage patterns of ontologies on the Web: a case study on GoodRelations vocabulary in RDFa. In: OWLED 2014. pp. 139–144 (2014)
19. Manaf, N.A.A., Bechhofer, S., Stevens, R.: The current state of SKOS vocabularies on the Web. In: ESWC 2012. pp. 270–284 (2012). https://doi.org/10.1007/978-3-642-30284-8_25
20. Matentzoglu, N., Bail, S., Parsia, B.: A corpus of OWL DL ontologies. In: DL 2013. pp. 829–841 (2013)
21. Meusel, R., Bizer, C., Paulheim, H.: A Web-scale study of the adoption and evolution of the schema.org vocabulary over time. In: WIMS 2015. p. 15 (2015). https://doi.org/10.1145/2797115.2797124
22. Meusel, R., Petrovski, P., Bizer, C.: The WebDataCommons Microdata, RDFa and Microformat dataset series. In: ISWC 2014, Part I. pp. 277–292 (2014). https://doi.org/10.1007/978-3-319-11964-9_18
23. Mika, P., Potter, T.: Metadata statistics for a large Web corpus. In: LDOW 2012 (2012)
24. Nogales, A., Urbán, M.Á.S., Barriocanal, E.G.: Measuring vocabulary use in the Linked Data Cloud. Online Inf. Rev. **41**(2), 252–271 (2017). https://doi.org/10.1108/OIR-06-2015-0183
25. Pan, J.Z.: Resource Description Framework. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, pp. 71–90. Springer (2009). https://doi.org/10.1007/978-3-540-92673-3_3
26. Pan, J.Z., Thomas, E., Sleeman, D.: ONTOSEARCH2: searching and querying Web ontologies. In: WWW/Internet 2006. pp. 211–218 (2006)
27. Pan, J.Z., Vetere, G., Gómez-Pérez, J.M., Wu, H. (eds.): Exploiting Linked Data and Knowledge Graphs in Large Organisations. Springer (2017). https://doi.org/10.1007/978-3-319-45654-6
28. Pham, M., Boncz, P.A.: Exploiting emergent schemas to make RDF systems more efficient. In: ISWC 2016, Part I. pp. 463–479 (2016). https://doi.org/10.1007/978-3-319-46523-4_28
29. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the Linked Data best practices in different topical domains. In: ISWC 2014, Part I. pp. 245–260 (2014). https://doi.org/10.1007/978-3-319-11964-9_16

30. Shi, Q., Wang, J., Pan, J.Z., Cheng, G.: VOYAGE: A large collection of vocabulary usage in open RDF datasets (2023), https://doi.org/10.5281/zenodo.7902675
31. Stadtmüller, S., Harth, A., Grobelnik, M.: Accessing information about Linked Data vocabularies with vocab.cc. In: CSWS 2012. pp. 391–396 (2012). https://doi.org/10.1007/978-1-4614-6880-6_34
32. Tummarello, G., Morbidoni, C., Bachmann-Gmür, R., Erling, O.: RDFSync: efficient remote synchronization of RDF models. In: ISWC 2007 + ASWC 2007. pp. 537–551 (2007). https://doi.org/10.1007/978-3-540-76298-0_39
33. Vandenbussche, P., Atemezing, G., 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). https://doi.org/10.3233/SW-160213
34. Wang, X., Cheng, G., Lin, T., Xu, J., Pan, J.Z., Kharlamov, E., Qu, Y.: PCSG: pattern-coverage snippet generation for RDF datasets. In: ISWC 2021. pp. 3–20 (2021). https://doi.org/10.1007/978-3-030-88361-4_1
35. Xu, P., Deng, Z., Choi, K., Cao, L., Wang, S.: Multi-view information-theoretic co-clustering for co-occurrence data. In: AAAI 2019. pp. 379–386 (2019). https://doi.org/10.1609/aaai.v33i01.3301379
36. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for Linked Data: A survey. Semantic Web **7**(1), 63–93 (2016). https://doi.org/10.3233/SW-150175