

# Profiling Linguistic Knowledge Graphs

Blerina Spahiu<sup>✉</sup>, Renzo Alva Principe<sup>✉</sup>, and Andrea Maurino<sup>✉</sup>

University of Milan-Bicocca, Milan, Italy  
name.surname@unimib.it

**Abstract.** Recently the number of approaches that model and interconnect linguistic data as knowledge graphs has experienced outstanding growth. However, despite the increasing availability of applications that manage such data, little attention has been given to their structural features. In this paper, we propose specific metrics to describe the structural features of knowledge graphs. Such metrics are evaluated on linguistic data and our findings provide a basis for a more efficient understanding of linguistic data.

**Keywords:** Knowledge graphs · data profiling · linguistic data.

## 1 Introduction

Language resources such as dictionaries, terminologies, corpora, etc., are adopting Semantic Web technologies to make their discovery, reuse and integration easy [5]. The Linked Data (LD) paradigm materialises Semantic Web by enabling data belonging to different topics [14] to be interconnected within a data-to-data cloud<sup>1</sup>. The linguistics community has taken advantages of the potential of the LD and has developed the Linguistic Linked Open Data (LLOD) cloud<sup>2</sup> for improving the usability and the discovery of language and linguistic resources.

In this vein, knowledge is represented into graphs using nodes and arcs. Such knowledge is stored and represented in RDF format<sup>3</sup>. The nodes represent entities while arcs represent relations among entities. Entities can have a relation of the form `rdf:type` denoting their types. The sets of possible types and relations are organized into schemas or ontologies, which define the meaning of the terms used in the knowledge graph through logical axioms.

KGs are often large and continuously evolving. As an example we can mention LOD cloud with more than 1,301 as of March 2022. This huge adaption of KGs into applications, is due to the fact that, with respect to relational models, KGs represent a flexible data model (e.g., Google’s Knowledge Graph, Facebook’s Graph API, Wikidata, etc.) where numerous editors are engaged in content creation, where the schema is ever changing, where data are incomplete, and where the connectivity of resources plays a key role. As the number of approaches that model linguistic data as knowledge graphs is increasing rapidly [5],

<sup>1</sup> <https://lod-cloud.net/>

<sup>2</sup> <http://linguistic-lod.org/llod-cloud>

<sup>3</sup> <https://www.w3.org/RDF/>

understanding their structure remains a fundamental step for their reuse. For example, before using a dataset one could be curious of *How types are related to each other?* or *How many triples are used to describe entities?*. In such a scenario, users want to know some structural features of these dataset, but this information is not completely covered in the state-of-the-art.

Even though the use of KGs in different applications is a matter of fact, it has a cost. When a user needs to use a KG for his/her use case, several are the challenges to be faced: (1) No prior knowledge about the data, (2) Missing schema or underspecification, (3) Lack of compliance with respect to the ontology, (4) Scalability challenges of large-scale RDF processing.

Such challenges might be addressed by knowledge graph profiling tools and approaches. Profiling approaches provide insights about the data in form of summaries, statistics or both [15]. Being able to access and explore the profile of a KG, a user can formulate and optimize queries, understand how graphs evolve and change, as well as enable data-management operations, such as compression, indexing, integration, enrichment and so forth. ABSTAT<sup>4</sup> is a data profiling tool proposed to mitigate some of the above challenges and help users understanding the content of a dataset effortlessly.

In this paper we make the following contributions: (i) enrich the profile produced by ABSTAT with more than 24 new statistics; (ii) provide an empirical analysis of the structural features of linguistics datasets, and (iii) provide a short discussion of such features.

The paper is structured as follows: Section 2 discusses approaches and tools used to profile KGs. In Section 3 we provide a brief description of ABSTAT profiles and provide the list of the new statistics added to such tool. Section 4 provides the analysis and findings by applying the enriched profile to LLOD datasets. The discussion analysis is described in Section 5 while conclusions and future work end the paper in Section 6.

## 2 Related Work

RDF graph profiling has been intensively studied, and various approaches and techniques have been proposed to provide a concise and meaningful representation of an RDF KGs. There are different recent surveys that discuss some of the approaches to profile knowledge graphs such as [4], [17] and [13]. In a recent work [15] we have reviewed and categorise profiling approaches. However, in this work we focus only on approaches that aim to produce profiles that quantitatively represent the content of the graph and provide an empirical analysis of structural features of KGs.

Explod [10] is used to summarize a dataset based on a mechanism that combines text labels and bisimulation contractions. It considers four RDF usages that describe interactions between data and metadata, such as class and predicate instantiating, class and predicate usage on which it creates RDF graphs.

<sup>4</sup> <http://abstat.disco.unimib.it/>

It provides also statistics about the number of equivalent entities connected using the `owl:sameAs` predicate to describe the interlinking between datasets. The ExpLOD summaries are extracted using SPARQL queries or algorithms such as partition refinement.

RDFStats generates statistics for data sets behind SPARQL endpoint and RDF documents [11]. These statistics include the number of anonymous subjects and different types of histograms; URIHistogram for URI subject and histograms for each property and the associated range(s). It also uses methods to fetch the total number of instances for a given class, or a set of classes and methods to obtain the URIs of instances.

LODStats is a profiling tool which can be used to obtain 32 different statistical criteria for RDF data sets [3]. These statistics describe the data set and its schema and include statistics about number of triples, triples with blank nodes, labeled subjects, number of `owl:sameAs` links, class and property usage, class hierarchy depth, cardinalities etc. These statistics are then represented using Vocabulary of Interlinked Datasets (VOID) and Data Cube Vocabulary<sup>5</sup>.

LODOP is a framework for executing, optimizing and benchmarking profiling tasks in Linked Data [7]. These tasks include the calculation of: number of triples, average number of triples per resources/ per object URI, number of properties, average number of property values, inverse properties, etc.

ProLOD++ is a web browser tool that implements several algorithms with the aim to compute different profiling, mining or cleansing tasks [1]. The profiling task includes the calculation of: frequencies and distribution of distinct subjects, predicates and objects, range of the predicates etc. ProLOD++ can also identify predicates combinations that contain only unique values as key candidates to distinctly identify entities.

Sansa is a graph processing tool that provides a unified framework for several applications such as link prediction, knowledge base completion, querying, and reasoning [9]. It computes several RDF statistics (such as the number of triples, RDF terms, properties per entity, and usage of vocabularies across datasets), and apply quality assessment in a distributed manner.

The approach most similar to ABSTAT is Loupe [12]. Loupe extracts types, properties and namespaces, along with a rich set of statistics about their use within the data set. It offers a triple inspection functionality, which provides information about triple patterns that appear in the data set and their frequency. Triple patterns have the form `<subjectType, property, objectType>`. Differently from ABSTAT, Loupe does not adapt a minimalization approach thus, Loupe's profiles contains much more triple patterns and are not as concise as ABSTAT profiles.

### 3 Profile Description

ABSTAT is a data profiling framework aiming to help users understanding the content of big data set by exploring its semantic profile [4]. It takes as input a

<sup>5</sup> <http://www.w3.org/TR/vocab-data-cube/>

data set and an ontology (used by the data set) and returns a semantic profile Fig.1. Thanks to the highly distributed architecture, ABSTAT is able to profile very big KGs [2]. The semantic profile produced by ABSTAT consists of a summary of patterns and several statistics Fig. 1. The informative units of ABSTAT’s summaries are Abstract Knowledge Patterns (AKPs), named simply *patterns* in the following, which have the form (subjectType, pred, objectType). Patterns represent the occurrence of triples  $\langle \text{sub}, \text{pred}, \text{obj} \rangle$  in the data, such that **subjectType** is the most specific type of the subject and **objectType** is the most specific type of the object [16]. Despite patterns, ABSTAT extracts also some statistics as the occurrence of types, predicates, patterns and cardinality descriptors (Fig. 1).

Even though ABSTAT profiles provide valuable information about the content of the dataset, it still misses some basic information that could help users in gaining a fast overview of some characteristics that these datasets have.

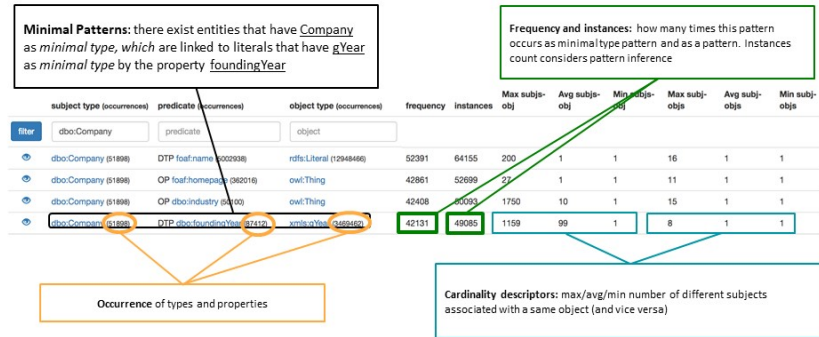


Fig. 1. ABSTAT profile for a sample of DBpedia dataset.

Below we enumerate the list of new statistics that are added to the semantic profile produced by ABSTAT:

- *# triples*: This statistics computes the number of triples in an RDF dataset.
- *# entities*: This statistics computes the number of entities in an RDF dataset.
- *# triples per entity (min, max, average)*: This statistic calculates the minimum, average and the maximum number of triples used to describe an entity.
- *# internal and external concepts*: This statistic computes the number of concepts that are considered to be internal of the dataset (defined in the pay-level domain) and external concepts (not defined in the pay-level domain)<sup>6</sup>.
- *# internal and external properties*: This statistic computes the number of properties that are considered to be internal of the dataset (defined in the

<sup>6</sup> The pay-level domain is defined as the part of a domain name, which can typically be registered by companies, organisations, or private end user [8]

- pay-level domain) and external properties (not defined in the pay-level domain).
- *# blank nodes as subject and # blank nodes as object*: This statistic counts the number of the blank nodes that occur at the subject and at the object position of a triple.
  - *In and out degree*: This statistic counts the number of links coming from the other datasets (in-degree) and the number of links going from the dataset to others (out-degree). The in-degree calculates the number of triples of the form (subject\_inPLD, predicate, object\_notPLD) while the out-degree counts the number of triples of the form (subject\_notD, predicate, object\_inLD).
  - *# owl:sameAs triples*: This statistic counts the number of triples that use (and those that do not use) the predicate `owl:sameAs`.
  - *# rdfs:label triples*: This statistic counts the number of triples that use the predicate `rdfs:label`.
  - *The list of typed and untyped literals*: This statistic gives the list of typed and untyped literals used in a dataset.
  - *The average length of untyped literals*: This statistics calculates the average length of the untyped literals.
  - *# of datatypes and their frequency*: This statistics provides the number and the frequency of use for each datatype used in a dataset.
  - *The list and the occurrence of the used languages*: This statistic enumerates the list with the occurrence of each language used in the dataset.
  - *The list and the occurrence of the used vocabularies*: This statistic enumerates the list with the occurrence of each vocabulary used in the dataset.

All the above statistics are implemented as API calls from the interface of ABSTAT tool.

## 4 Experiments

In this section we provide an analysis of the structural features by applying the above statistics in linguistics datasets from the Linguistic Linked Data Cloud.

### 4.1 Linguistics Datasets

The experiments were run using all the datasets from the Linguistic Linked Open Data Cloud. There are in total 136 datasets belonging to the linguistic domain in the LOD cloud. However, only 72 of them do provide a URL for the dump [6]. During the inspection of the availability of the dump it was possible to download and process the dump for only 48 datasets, while for the other (i) either the URL was not available anymore, or (ii) the dump was available but the dataset had many syntactic errors, or (iii) they were not in RDF.

### 4.2 Empirical Analysis

In this section we analyse the results for each of the above statistics applied to our datasets corpus.

*# triples, # entities, and # triples per entity:* From all the datasets from the LLOD cloud *iate* is the biggest dataset with 74,023,248 triples and 207,263 entities while the smallest datasets with respect to the number of triples is *lemonbuy* with 961 triples and *apertium-rdf-en-es* is the smallest dataset with respect to the number of entities, i.e., 3. Datasets belonging to the *apertium* datasets have from 3 to maximum 7 entities while 47,445 to maximum 156,941 triples. Thus the average number of triples for entities is grater for *apertium* datasets. The datasets that uses in average less triples per entity are *wn-wiki-instances*, *srcmf*, *linked-hypernyms*, *cedict* with around 1 triple per entity.

*# internal and external concepts:* The analysis shows that only three datasets *cedict*, *gwa-ili*, *iso-639-oasis* use 2 internal concepts each to describe entities. As a consequence, the number of external concepts for all the datasets is grater. The dataset with the highest number of external concepts is *linked-hypernyms* with 361. Finally, *wn-wiki-instances* dataset has 0 internal and 0 external concepts.

*# internal and external properties:* Similar analysis as for the concepts is present for the number of internal and external properties. Only 5 datasets use internal properties to describe resources, i.e *cedict* (4), *iso-639-oasis* (3), *lexvo* (13), *saldo-rdf* (2), and *wordnet* (26). All the rest 43 datasets have 0 internal properties but borrow them from external vocabularies. The dataset with the highest number of external properties is *getty-aat* with 196 properties. Around 77% of the datasets have less than 10 properties.

*# blank nodes as subject and # blank nodes as object:* The analysis about the use of blank nodes shows that none of the datasets use blank nodes in the subject position and only 19% of use blank nodes in the object position. The dataset with the highest number of blank nodes is *getty-aat* (5950) and *lemonwiktionary* (332).

*In and out degree:* Datasets in the LLOD are more generally connected from inside to outside, meaning that the object of their triples reside in other datasets. In fact, only 12,5% of the datasets have 0 outgoing links, while 62% have 0 incoming links. *iate* dataset has the highest number of outgoing links with 16,881,770 links while *saldo-rdf* plays the role of a central hub with 320,059 incoming links. Fig. 2 shows the distribution of the number of outgoing and incoming links for each dataset in the LLOD.

*# owl:sameAs triples:* Regarding the type of outgoing and incoming links we further analyse the use of *owl:sameAs* predicate. The distribution of the number of such triples within the LLOD is shown in Fig. 3. The datasets with the highest number of *owl:sameAs* triples is *iate*, which also had the highest number of outgoing links. Around 46% of the datasets have less than 3 sameAs links, while less than 10% have more than 100,000 sameAs links.

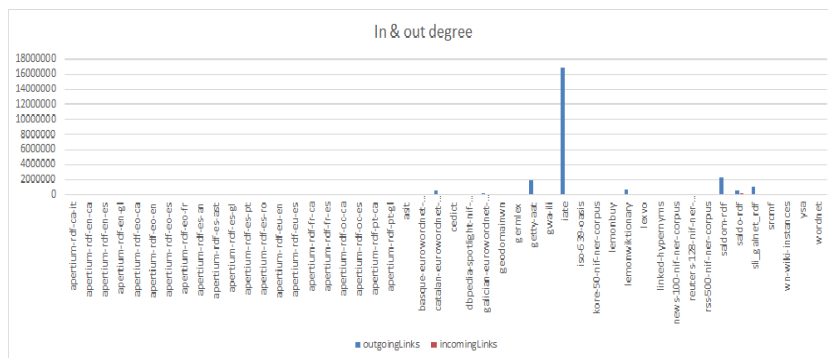


Fig. 2. In & out degree for LLOD datasets.

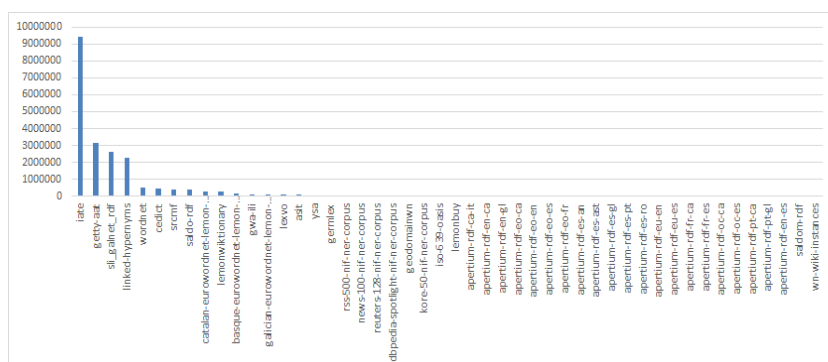


Fig. 3. Distribution of number of owl:sameAs triples per LLOD datasets.

*# rdfs:label triples:* We analysed the use of the predicate *rdfs:label* by the entities of LLOD. Around 77% of the datasets have less than 10 triples with the predicate *rdfs:label*. 4 datasets have more than 100,000 *rdfs:label* triples, i.e., *basque-eurowordnet-lemon-lexicon-3-0* (134,748), *lexvo* (146,530), *catalan-eurowordnet-lemon-lexicon-3-0* (213,787), and *sl\_i\_galnet\_rdf* (723,348) triples.

*# typed and untyped literals:* The graph in Fig. 4 shows the distribution of typed and untyped literals. As from the graph *iate* has most of typed (7,803,650) and untyped (12,922,660) literals. 11 datasets do not have any untyped literals.

*The average length of untyped literals:* Top three datasets that have in average the longest untyped literals are *news-100-nif-ner-corpus* (70), *gwa-ili* (62), and *reuters-128-nif-ner-corpus* (60).

*The list and the occurrence of datatypes:* The most used datatype in the LLOD is <http://www.w3.org/2001/XMLSchema#integer> (8,710,881), followed by <http://www.w3.org/2001/XMLSchema#date> (37347) and

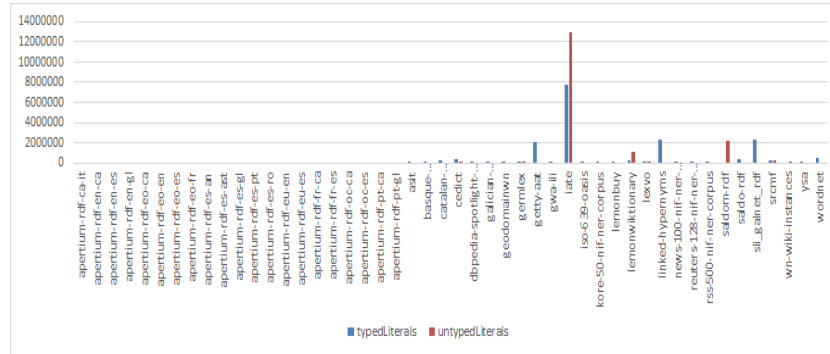


Fig. 4. Distribution of typed and untyped literals per LLOD datasets.

<http://www.w3.org/2001/XMLSchema#dateTime> (36428). The less used datatype instead is <http://www.w3.org/2001/XMLSchema#boolean> (2).

*The list and the occurrence of the used languages:* There are about 176 different language used to tag literals in the LLOD datasets. The dataset with the most languages is *lexvo* with 175 different languages. Around 90% of the datasets have less than five languages. The most used language is English (36), Swedish (6), and French (5).

*The list and the occurrence of the used vocabularies:* The analysis shows that the dataset that uses most vocabularies to describe its data is *lexvo* (626). The distribution of the number of vocabularies per dataset is given in Fig. 5. The most used vocabularies among LLOD datasets are *rdf* (48), *rdfs* (37), and *owl*.

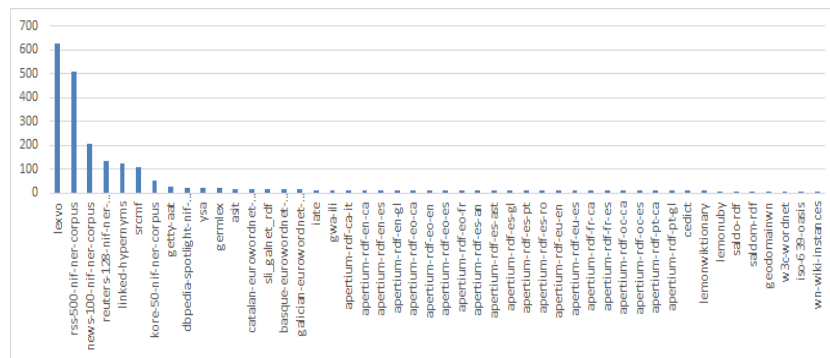


Fig. 5. Number of vocabularies per LLOD datasets.



## 5 Discussion

In this work, we have analysed structural features of Linguistic LOD datasets. All datasets show a skewed structure with respect to the number of internal and external concepts and properties. In fact, almost all the datasets had more external concepts and properties. Complementing the previous finding, our evaluation also revealed that most datasets are extensively typed (more than 99% of datasets have typed entities). Regarding the in & out degree, most of the datasets had more outgoing links. In fact, most of the datasets make use of the *owl:sameAs* predicate. However, our findings are not in line with what is being described in the LLOD website<sup>7</sup>. This for two reasons: (i) we consider the dump of the datasets having the topic linguistic in the metadata of the LOD cloud, and (ii) the version of the LLOD datasets might be different.

With respect to the use of *radfs:label*, we observed that this predicate is not often used as three-quarters of the datasets use it within less than 10 triples. Also the distribution of typed and untyped literals is skewed. While most of the smallest datasets (with respect to the number of triples) do use typed literals, for the biggest ones the number of untyped literals is greater than the typed ones.

The most frequent used language within LLOD dataset is English (99% of the datasets). Moreover, *lexvo* is the dataset with the highest number of languages (175) out of 176 of the languages in total. Regarding the used vocabularies, *rdf* remains the most used vocabulary by most of the datasets in the LLOD.

## 6 Conclusion and Future Work

In this paper we present a first preliminary analysis of structural features of LLOD datasets. We extend the profile built by ABSTAT tool with 24 new statistics in order to have a more detailed view of the content of RDF datasets. Such statistics have been applied to datasets that belong to the linguistics domain of the LOD datasets. We were not able to manage all the datasets belonging to this domain as for many we were not able to find the dump or it had syntactic errors. However, we provide an empirical analysis of the content for 48 datasets.

Currently we are extending the profile with some fine-grained statistics. As future work we plan to integrate all statistics as API calls in the ABSTAT profile. Moreover, we plan to build an interactive interface where users can make more insightful analysis by cross-checking some of the statistics provided by ABSTAT.

## References

1. Z. Abedjan, T. Grütze, A. Jentzsch, and F. Naumann. Profiling and mining rdf data with *prolog++*. In *2014 IEEE 30th International Conference on Data Engineering*, pages 1198–1201. IEEE, 2014.

<sup>7</sup> <https://linguistic-lod.org/>

2. R. A. Alva Principe, A. Maurino, M. Palmonari, M. Ciavotta, and B. Spahiu. Abstat-hd: a scalable tool for profiling very large knowledge graphs. *The VLDB Journal*, pages 1–26, 2021.
3. S. Auer, J. Demter, M. Martin, and J. Lehmann. Lodstats—an extensible framework for high-performance dataset analytics. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 353–362. Springer, 2012.
4. Š. Čebirić, F. Goasdoué, H. Kondylakis, D. Kotzinos, I. Manolescu, G. Troullinou, and M. Zneika. Summarizing semantic graphs: a survey. *The VLDB Journal*, 28(3):295–327, 2019.
5. P. Cimiano, C. Chiarcos, J. P. McCrae, and J. Gracia. *Linguistic Linked Data*. Springer, 2020.
6. M. P. di Buono, H. G. Oliveira, V. B. Mititelu, B. Spahiu, and G. Nolano. Paving the way for enriched metadata of linguistic linked data.
7. B. Forchhammer, A. Jentzsch, and F. Naumann. Lodop-multi-query optimization for linked data profiling queries. In *PROFILES@ESWC*, 2014.
8. T. Gottron, M. Knauf, and A. Scherp. Analysis of schema structures in the linked open data graph based on unique subject uris, pay-level domains, and vocabulary usage. *Distributed and Parallel Databases*, 33(4):515–553, 2015.
9. H. Jabeen, D. Graux, and G. Sejdiu. Scalable knowledge graph processing using sansa. In *Knowledge Graphs and Big Data Processing*, pages 105–121. Springer, 2020.
10. S. Khatchadourian and M. P. Consens. Explod: summary-based exploration of interlinking and rdf usage in the linked open data cloud. In *Extended Semantic Web Conference*, pages 272–287. Springer, 2010.
11. A. Langegger and W. Woss. Rdfstats—an extensible rdf statistics generator and library. In *2009 20th International Workshop on Database and Expert Systems Application*, pages 79–83. IEEE, 2009.
12. N. Mihindukulasooriya, M. Poveda-Villalón, R. García-Castro, and A. Gómez-Pérez. Loupe—an online tool for inspecting datasets in the linked data cloud. In *International Semantic Web Conference (Posters & Demos)*, 2015.
13. Q. Song, Y. Wu, P. Lin, L. X. Dong, and H. Sun. Mining summaries for knowledge graph search. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1887–1900, 2018.
14. B. Spahiu, A. Maurino, and R. Meusel. Topic profiling benchmarks in the linked open data cloud: Issues and lessons learned. *Semantic Web*, 10(2):329–348, 2019.
15. B. Spahiu, M. Palmonari, R. Alva Principe, and A. Rula. Understanding the structure of knowledge graphs with abstat profiles. *Submitted to Semantic Web Journal*.
16. B. Spahiu, R. Porrini, M. Palmonari, A. Rula, and A. Maurino. Abstat: ontology-driven linked data summaries with pattern minimalization. In *European Semantic Web Conference*, pages 381–395. Springer, 2016.
17. M. Zneika, D. Vodislav, and D. Kotzinos. Quality metrics for rdf graph summarization. *Semantic Web*, (Preprint):1–30, 2019.