Profiling Linguistic Knowledge Graphs

Blerina Spahiu
University of Milan-Bicocca
Milan, Italy

Renzo Alva Principe
University of Milan-Bicocca
Milan, Italy

Andrea Maurino
University of Milan-Bicocca
Milan, Italy

{blerina.spahiu | renzo.alvaprincipe | andrea.maurino} @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 ba-sis for a more efficient understanding of 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 (Cimiano et al., 2020). The Linked Data (LD) paradigm materialises Semantic Web by enabling data belonging to different topics (Spahiu et al., 2019) to be interconnected within a data-to-data cloud[1]. The linguistics community has taken advantages of the potential of the LD and has developed the Linguistic Linked Open Data (LLOD) cloud[2] for improving the usabil-ity 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[3]. The nodes represent entities while arcs rep-resent relations among entities. Entities can have a relation of the form rdf:type denoting their types. The sets of possible types and re-lations are organized into schemas or ontolo-gies, 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 data sets as of March 2022. This huge adoption 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 (Cimiano et al., 2020), understand-ing 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 datasets, but this information is not completely covered in the state-of-the-art tools and approaches.

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) Miss-ing 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 (Spahiu et al., 2023). Be-ing able to access and explore the profile of a

[1] https://lod-cloud.net/
[2] http://linguistic-lod.org/llod-cloud
[3] https://www.w3.org/RDF/

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[4] 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 24 new statistics; (ii) provide a list of applications where such statistics are useful; (iii) provide an empirical analysis of the structural features of linguistics datasets, and (iv) 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. Sec-tion 4 provides the analysis and findings by ap-plying the enriched profile to LLOD datasets. The discussion analysis is described in Section 5 while conclusions and future work end the paper in Section 7.

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 KG. There are different recent surveys that discuss some of the approaches to profile knowledge graphs such as (Čebirić et al., 2019), (Zneika et al., 2019) and (Song et al., 2018). In a recent work (Spahiu et al., 2023) we have reviewed and categorise profiling approaches. However, in this work, we focus only on approaches that aim to produce profiles that quantita-tively represent the content of the graph and provide an empirical analysis of the structural features of KGs.

ExpLOD (Khatchadourian and Consens, 2010) 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 be-tween data and metadata, such as class and predicate instantiating, and class and predi-cate usage on which it creates RDF graphs. It provides also statistics about the num-ber of equivalent entities connected using the owl:sameAs predicate to describe the inter-linking between datasets. The ExpLOD summaries are extracted using SPARQL queries or algorithms such as partition refinement.

RDFStats generates statistics for datasets behind SPARQL endpoint and RDF documents (Langegger and Woss, 2009). 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 to-tal 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 that can be used to obtain 32 different statistical crite-ria for RDF datasets (Auer et al., 2012). These statistics describe the dataset and its schema and include statistics about the num-ber 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[5].

Sansa is a graph processing tool that provides a unified framework for several applications such as link prediction, knowledge base completion, querying, and reasoning (Jabeen et al., 2020). It computes several RDF statis-tics (such as the number of triples, RDF terms, properties per entity, and usage of vocabular-ies across datasets), and applies quality assess-ment in a distributed manner.

The approach most similar to ABSTAT is Loupe (Mihindukulasooriya et al., 2015). Loupe extracts types, properties and namespaces, along with a rich set of statistics about their use within the dataset. It offers a triple inspection functionality, which provides information about triple patterns that appear in the dataset and their frequency. Triple patterns have the form <subjectType, prop-erty, objectType>. Differently from ABSTAT, Loupe does not adapt a minimalization ap-proach thus, Loupe's profiles contain much

[4] http://abstat.disco.unimib.it/

[5] http://www.w3.org/TR/vocab-data-cube/

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 datasets by exploring its semantic pro-file (Spahiu et al., 2023). It takes as in-put a data set and an ontology (used by the data set) and returns a semantic profile (Fig.1). Thanks to the highly distributed ar-chitecture, ABSTAT is able to profile very big KGs (Alva Principe et al., 2021). The semantic profile produced by ABSTAT con-sists of a summary of patterns and several statistics (Fig. 1). The informative units of ABSTAT's summaries are Abstract Knowl-edge Patterns (AKPs), named simply pat-terns in the following, which have the form (subjectType, pred, objectType). Pat-terns represent the occurrence of triples <sub, pred, obj> in the data, such that subjectType is the most specific type of the subject and objectType is the most specific type of the object (Spahiu et al., 2016). De-spite patterns, ABSTAT extracts also some statistics as the occurrence of types, pred-icates, patterns and cardinality descriptiors (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.

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

- # triples: This statistic computes the number of triples in an RDF dataset.
- # entities: This statistic 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)[6].
- # internal and external properties: This statistic computes the number of properties that are considered to be internal of the dataset (defined in the 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 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, pred-icate, 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.

[6]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 (Gottron et al., 2015)

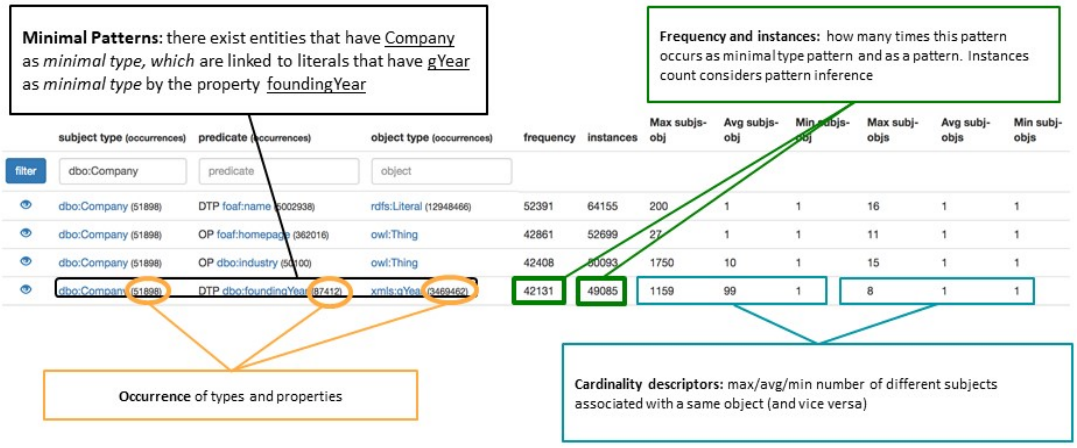

Figure 1: ABSTAT profile for a sample of DBpedia dataset.

- 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 vocabu-lary 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 (di Buono et al., 2022). 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 the biggest dataset is iate with 74, 023, 248 triples and 20, 726, 310 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., 2. Datasets belonging to the apertium datasets have from 2 to maximum 6 entities while 47, 445 to maximum 156, 941 triples. Thus the average number of triples for entities is greater 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 ce-dict, 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 greater. 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: Sim-ilar analysis for the concepts is present for the number of internal and external proper-ties. Only 5 datasets use internal properties to describe resources, i.e cedict (4), iso-639-oasis (3), lexvo (13), saldo-rdf (2), and word-net (26). All the rest 43 datasets have 0 in-ternal properties but borrow them from exter-

nal 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 only lemonbuy uses blank nodes in the subject position while 52% of datasets use blank nodes in the object position. The dataset with the highest number of blank nodes is cedict (554367) and wordnet (423986).

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.

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 predi-cate rdfs:label. 4 datasets have more than 100, 000 radfs:label triples, i.e., basque-eurowordnet-lemon-lexicon-3-0 (134, 748), lexvo (146, 530), catalan-eurowordnet-lemon-lexicon-3-0 (213, 787), and sli_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 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 176 languages used to tag literals in the LLOD datasets. The dataset with 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.

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 outgo-ing links. In fact, most of the datasets make use of the owl:sameAs predicate. However, our finding are not in line with what is be-ing described in the LLOD website[7]. This is for two reasons: (i) we consider the dump of the datasets having the topic linguistic in the

[7] https://linguistic-lod.org/

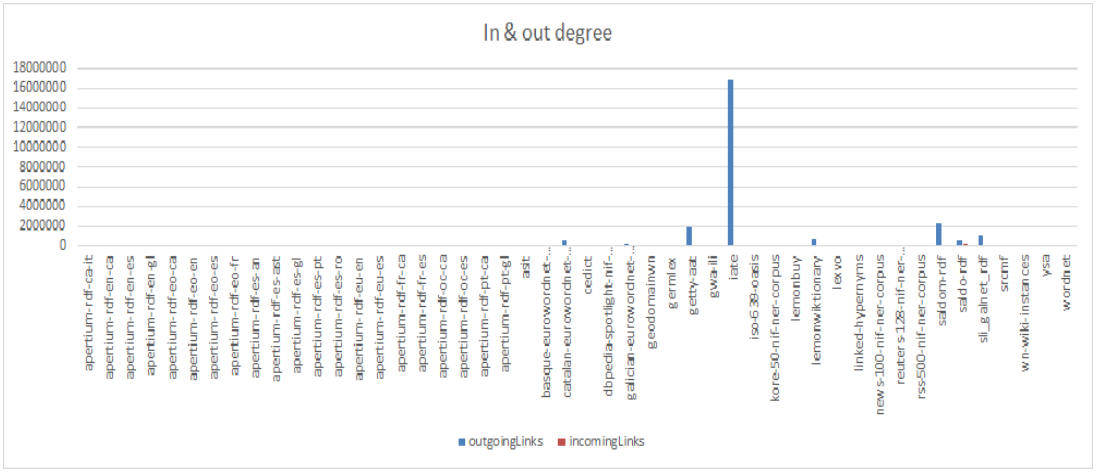

Figure 2: In & out degree for LLOD datasets.

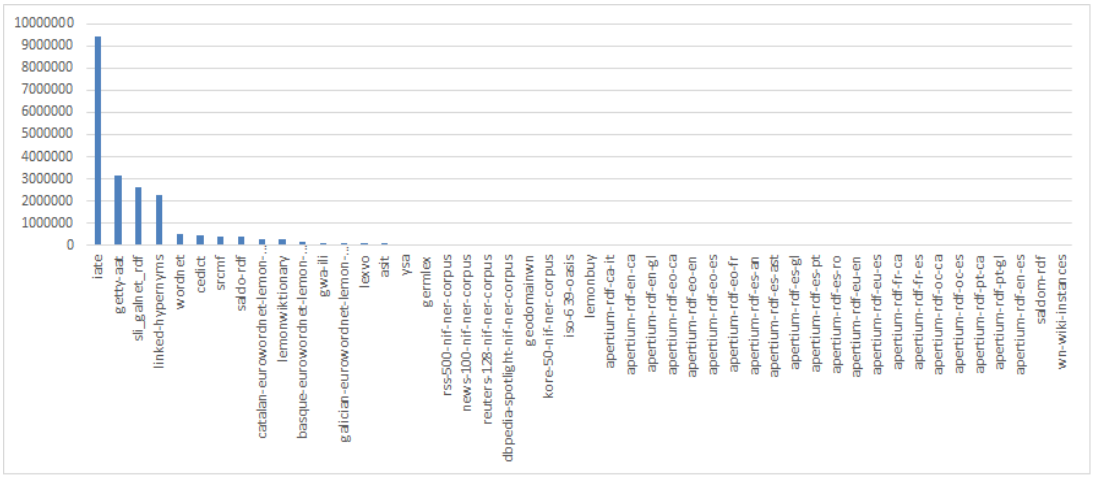

Figure 3: Distribution of number of owl:sameAs triples per LLOD datasets.

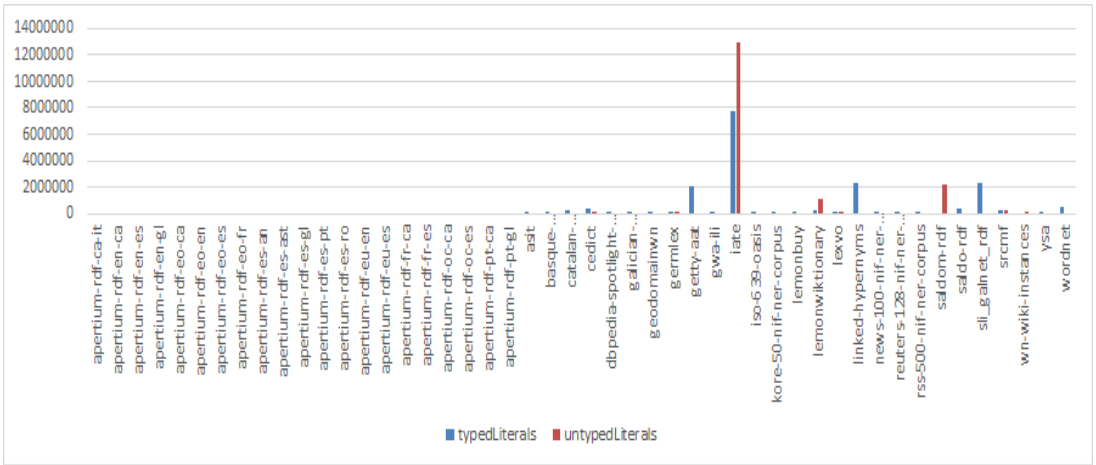

Figure 4: Distribution of typed and untyped literals per LLOD datasets.

metadata of the LOD cloud, and (ii) the version of the LLOD datasets might be different.

We observed that rdfs:label 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

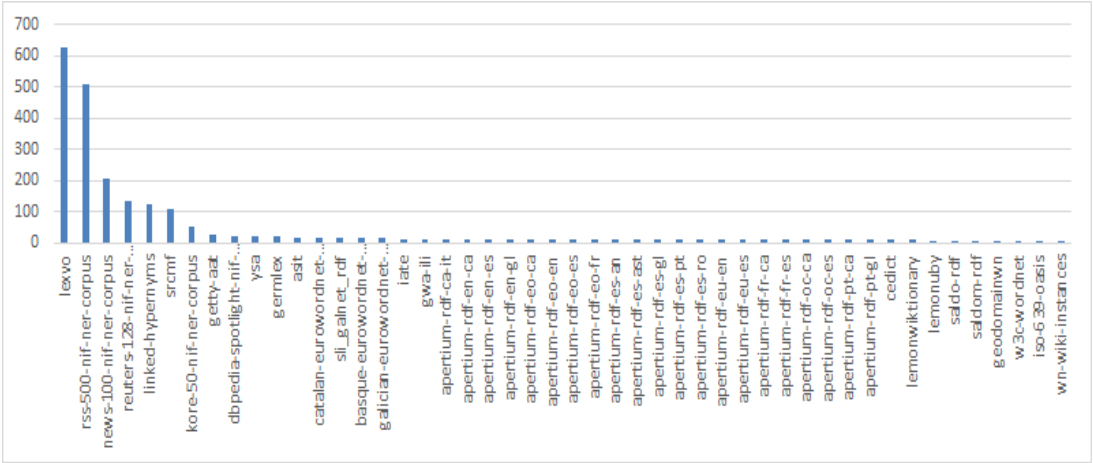

Figure 5: Number of vocabularies per LLOD datasets.

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 Analysis of key statistics and their application significance

In this section, we group the above statistics in regard to their application.

Entity Summarization: For this application scenario, the statistics in Table 1 provide (i) a quantitative understanding of the size and density of the knowledge graph, allowing for efficient summarization techniques; (ii) help identifying the source and coverage of concepts and properties used in the graph, aiding in accurate and comprehensive entity summarization; and (iii) providing information about entity equivalence and human-readable labels, enabling improved entity summarization and labeling.

Recommendation Systems: The statistics useful for recommendation systems (i) identify the level of information available for each entity, enabling more informed and personalized recommendations, (ii) analyse the connectiv-ity of the dataset with external sources help-ing in incorporating relevant information from external sources for more accurate recommendations, and (iii) assists in identifying equivalent entities, which can enhance recommendation algorithms by considering similar or related items.

Question Answering: For this downstream application these statistics provide (i) a sense of the knowledge graph's size and coverage, aiding in understanding the scope and poten-tial for answering a wide range of questions;(ii) identify the level of detail available for each entity, assisting in generating comprehensive and informative answers, (iii) provide human-readable labels for entities, improving the clar-ity and understandability of question answer-ing results.

Information Extraction: The statistics for this application offer (i) insights into the over-all scope and coverage of the knowledge graph, helping in identifying relevant entities and relationships for extraction tasks, and (ii) assist in identifying instances where entities are represented as blank nodes, allowing for appropriate handling during information extraction processes.

Link Prediction: Link prediction is supported by (i) providing information about the richness of entity descriptions, aiding in more accurate link prediction by considering entities with more detailed representations, (ii) analyzing the connectivity of the dataset with external sources helps in predicting links between the knowledge graph and external entities; and (iii) in identifying equivalent entities, supporting link prediction across different datasets or ontologies.

Table 1: Application-specific metric

	Entity Summarisation	Recommendation Systems	Question Answering	Information Extraction	Link Prediction	Anomaly Detection	Semantic Search	Data Integration and Fusion
# triples	x		x	x			x	x
# entities	x		x	x			x	x
# triples per entity	x	x	x		x	x		x
# internal and external concepts	x							x
# internal and external properties	x							x
# blank nodes as subject				x		x		
# blank nodes as object				x		x		
in and out degree		x			x			
# owl:sameAs triples	x	x			x			x
# rdfs:label triples	x		x				x	
list of typed and untyped literals							x	
average length of untyped literals							x	
# of datatypes and their frequency						x		
list and the occurrence of the used languages							x	
list and the occurrence of the used vocabularies							x	

Anomaly Detection: The statistics for this application scenario (i) identify entities with abnormal numbers of triples, aiding in anomaly detection by flagging entities with unusual representations or relationships, and (ii) assist in identifying instances where blank nodes are involved in triples, which can be in-dicative of potential anomalies or incomplete information.

Semantic Search: These statistics for Semantic Search offer: (i) they indicate the knowledge graph's size and coverage, ensuring comprehensive and accurate semantic search results; (ii) they provide human-readable labels for entities, enhancing the relevance and presentation of search results; (iii) they include textual information linked to entities, thereby improving the retrieval of relevant results.

Data Integration and Fusion: Such statis-tics help understanding the size and scope of the knowledge graph, supporting data integra-tion efforts by assessing the compatibility and overlap with external datasets and assist in identifying concepts and properties shared.

7 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

Renzo Arturo Alva Principe, Andrea Maurino, Matteo Palmonari, Michele Ciavotta, and Blerina Spahiu. 2021. Abstat-hd: a scalable tool for profiling very large knowledge graphs. The VLDB Journal, pages 1–26.

Sören Auer, Jan Demter, Michael Martin, and Jens Lehmann. 2012. Lodstats–an extensible framework for high-performance dataset analyt-ics. In International Conference on Knowledge Engineering and Knowledge Management, pages 353–362. Springer.

ŠejlaČebirić, François Goasdoué, Haridimos Kondylakis, Dimitris Kotzinos, Ioana Manolescu, Georgia Troullinou, and Mussab Zneika. 2019. Summarizing semantic graphs: a survey. The VLDB Journal, 28(3):295–327.

Philipp Cimiano, Christian Chiarcos, John P McCrae, and Jorge Gracia. 2020. Linguistic Linked Data. Springer.

Maria Pia di Buono, Hugo Gonçalo Oliveira, Verginica Barbu Mititelu, Blerina Spahiu, and

Gennaro Nolano. 2022. Paving the way for enriched metadata of linguistic linked data. Semantic Web, 13(6):1133–1157.

Thomas Gottron, Malte Knauf, and Ansgar Scherp. 2015. Analysis of schema structures in the linked open data graph based on unique subject uris, pay-level domains, and vocabu-lary usage. Distributed and Parallel Databases, 33(4):515–553.

Hajira Jabeen, Damien Graux, and Gezim Sejdiu. 2020. Scalable knowledge graph processing us-ing sansa. In Knowledge Graphs and Big Data Processing, pages 105–121. Springer.

1. Shahan Khatchadourian and Mariano P Consens. 2010. Explod: summary-based exploration of interlinking and rdf usage in the linked open data cloud. In Extended Semantic Web Con-ference, pages 272–287. Springer.

Andreas Langegger and Wolfram Woss. 2009. Rdfstats-an extensible rdf statistics generator and library. In 2009 20th International Work-shop on Database and Expert Systems Applica-tion, pages 79–83. IEEE.

Nandana Mihindukulasooriya, María Poveda-Villalón, Raúl García-Castro, and Asunción Gómez-P érez. 2015. Loupe-an online tool for in-specting datasets in the linked data cloud. In In-ternational Semantic Web Conference (Posters & Demos).

Qi Song, Yinghui Wu, Peng Lin, Luna Xin Dong, and Hui Sun. 2018. Mining summaries for knowledge graph search. IEEE Transactions on Knowledge and Data Engineering, 30(10):1887–1900.

Blerina Spahiu, Andrea Maurino, and Robert Meusel. 2019. Topic profiling benchmarks in the linked open data cloud: Issues and lessons learned. Semantic Web, 10(2):329–348.

Blerina Spahiu, Matteo Palmonari, Renzo Alva Principe, and Anisa Rula. 2023. Understanding the structure of knowledge graphs with abstat profiles. Submitted to Semantic Web Journal.

Blerina Spahiu, Riccardo Porrini, Matteo Palmonari, Anisa Rula, and Andrea Maurino. 2016. Abstat: ontology-driven linked data sum-maries with pattern minimalization. In Euro-pean Semantic Web Conference, pages 381–395. Springer.

Mussab Zneika, Dan Vodislav, and Dimitris Kotzinos. 2019. Quality metrics for rdf graph summarization. Semantic Web, (Preprint):1–30.