Analyzing the Evolution of Vocabulary Terms and their Impact on the LOD Cloud

Mohammad Abdel-Qader[1,2], Ansgar Scherp[1,2], and Iacopo Vagliano[2]

[1] Christian-Albrechts University, Kiel, Germany
{stu120798,asc}@informatik.uni-kiel.de
[2] ZBW – Leibniz Information Centre for Economics, Kiel, Germany
{m.abdel-qader,a.scherp,I.Vagliano}@zbw.eu

Abstract. Vocabularies are used for modeling data in Knowledge Graphs (KGs) like the Linked Open Data Cloud and Wikidata. During their lifetime, vocabularies are subject to changes. New terms are coined, while existing terms are modified or deprecated. We first quantify the amount and frequency of changes in vocabularies. Subsequently, we investigate to which extend and when the changes are adopted in the evolution of KGs. We conduct our experiments on three large-scale KGs: the Billion Triples Challenge datasets, the Dynamic Linked Data Observatory dataset, and Wikidata. Our results show that the change frequency of terms is rather low, but can have high impact due to the large amount of distributed graph data on the web. Furthermore, not all coined terms are used and most of the deprecated terms are still used by data publishers. The adoption time of terms coming from different vocabularies ranges from very fast (few days) to very slow (few years). Surprisingly, we could observe some adoptions before the vocabulary changes were published. Understanding the evolution of vocabulary terms is important to avoid wrong assumptions about the modeling status of data published on the web, which may result in difficulties when querying the data from distributed sources.

1 Introduction

Vocabulary terms define the schema of Knowledge Graphs (KGs) such as the Linked Open Data (LOD) cloud or Wikidata. After ontology engineers publish the first version of a vocabulary, the terms are subject to changes to reflect new requirements or shifts in the domains the vocabularies model. So far it is unknown how such vocabulary changes are reflected by the KGs that are using their terms. Data publishers may not be aware that changes in the vocabulary terms happened since it occurs rather rarely [7]. Explicitly triggering data publishers to update their model is also challenging due to the distributed nature of KGs such as the LOD cloud. Data publishers may be interested in being notified when certain vocabulary term changes happen, but they lack proper tools and services to track whether and what kind of changes on vocabulary terms happened. Likewise, ontology engineers are not aware of who uses their vocabularies

and lack a tool that reflects the adoption status of their ontologies and changes on the defined terms. In this paper, we study the evolution of vocabulary terms in KGs. We address three research questions: (1) *When are the newly created terms adopted in KGs?* (2) *What is the use rate of terms for a set of vocabularies in each dataset?* (3) *Are the deprecated terms still used in KGs?*

To address these questions, we analyzed various vocabularies to better understand by whom and how they are used, and how these changes are adopted in evolving KGs. Formally, we understand a vocabulary V as a set of terms T. A term T is either a class C or a property P. A set of terms relates to a vocabulary as $T(V) = C(V) \bigcup P(V)$. Changes in a vocabulary V are changes on its terms, i. e., classes and properties. Data that uses terms of a changed vocabulary should be updated accordingly. In a previous work [2], we manually conducted a qualitative analysis of vocabulary evolution on the LOD cloud. We analyzed the changes for a set of vocabularies by clarifying which terms changed, the type of change, and if those changes were done on terms defined in the vocabularies or on the classes and properties that were imported from other vocabularies. In this paper, we consider the two basic types of changes: addition and deletion. Any other change, e. g., a modification, can be expressed by these two basic changes. We use three well-known dataset: Dynamic Linked Data Observatory (DyLDO) [8], Billion Triples Challenge (BTC)[1], and Wikidata[2].

Our experiments showed that even if the frequency of vocabularies terms changes is rather low, they have a large impact on the real data. Most of the newly coined terms are adopted in less than one week after their publishing date.

Our work may help ontology engineers to select classes and properties that fit their needs when creating or updating ontologies. For example, we believe it can make ontology engineers more aware of who uses their terms and how. Furthermore, it may foster a better understanding what is the impact of their changes and how they are adopted, to possibly learn from previous experience what change is effective and what not. This study may also assists data publishers in updating their models by providing information about vocabulary changes.

The remainder is structured as follows. In Section 2, we review related work. We present our methodology in Section 3, and describe the datasets considered in Section 4. We outline our results in Section 5, we discuss them in Section 6 and conclude in Section 7.

2 Related Work

In terms of analyzing the use of structured data on the web, some works focused on *schema.org*. Meusel et al. [9] analyzed its evolution and adoption. They made a comparison of the use of *schema.org* terms over four years by extracting the structured data from the web pages that use this vocabulary from *WebDataCommons* Microdata datasets[3]. They discovered that not all terms in *schema.org* are

[1] http://challenge.semanticweb.org/, last accessed: 29/11/2017
[2] https://www.wikidata.org, last accessed: 29/11/2017
[3] http://webdatacommons.org/, last accessed: 10/10/2017

used and deprecated terms are still used, as it is also illustrated in this work. Furthermore, they found that publishing new types and properties is preferred over using *schema.org*'s extension mechanism. Guha et al. [6] investigated the use of the *schema.org* in the structured data of a set of web pages. They analyzed a sample of 10 billion web pages crawled from Google index and *WebDataCommons* and found that about 31 % of those pages had some *schema.org* elements and estimated that around 12 million websites are using *schema.org* terms. In contrast to this work, they did not consider the changes in vocabulary terms. Additionally, we are not limited to one vocabulary only.

Mihindukulasooriya et al. [10] conducted a quantitative analysis for studying the evolution of DBpedia, *schema.org*, PROV-O, and FOAF ontologies. They proposed recommendations such as the need of dividing large ontologies into modules to avoid duplicates when adding new terms and adding provenance information beside the generic metadata when the change occurred. Papavassiliou et al. [11] proposed a framework that automatically identifies changes for both schema and data. They provide a formal language for identifying ontology changes and a change detection algorithm. Roussakis et al. [12] introduced a framework for analyzing the evolution of LOD datasets. Their framework allows users to identify changes in datasets versions and make a complex analysis on the evolved data.

Other works exploited DyLDO to study the use of vocabularies. Dividino et al. [4] analyzed how the use of RDF terms on the LOD cloud changed over time. They studied the combination of terms that describe a resource but did not investigate whether a vocabulary and its terms have changed. The authors applied their analysis on a dataset of 53 weekly snapshots from the DyLDO dataset, as it is also investigated in this work. Over six months, Käfer et al. [7] observed the documents retrieved from DyLDO. They analyzed those documents using different factors, their lifespan, the availability of those documents and their change rate. Also, they analyzed the RDF content that is frequently changed (triple added or removed). Additionally, they observed how links between documents are evolved over time. While their study is important for various areas such as smart caching, link maintenance, and versioning, it does not include information about adopting new and deprecated terms.

Gottron et al. [5] provided an analysis of the LOD schema information by analyzing the BTC 2012 dataset in three different levels. The first level concerns unique subject URIs by studying the dependency relations between the classes and their properties. They found a redundancy between classes and the attached properties. The second level addresses Pay-Level Domains (PLDs) by dividing the BTC 2012 dataset into individual PLDs. They found that 20 % of the PLDs can be ignored without losing the graph explanation. The third level focuses on the vocabularies by analyzing how important a vocabulary is for describing the data. They stated that data publishers either made a strong schematic design, or apply a combination between a set of vocabularies to model their data.

Finally, some studies analyzed the use of vocabularies with other sources. Vandenbussche et al. [15] published a report that describes Linked Open Vo-

cabularies (LOV). It provides statistics about LOV and its capabilities such as the total number of terms and the top-10 searched terms, but does not include information about adopting new terms and which PLD uses which vocabulary. Rathachari et al. [3] proposed a model that facilitates the understanding of organisms. Their model presents the changes in taxonomic knowledge in RDF form. The proposed model acts as a history tracking system for changing terms but gives no information about how and when the terms are used, and which PLDs adopted the changed terms. Schaible et al. [13] published a survey of the most preferred strategies for reusing vocabulary terms. The participants, 79 Linked Data experts and practitioners, were asked to rank several LOD modeling strategies. The survey concluded that terms widely used are considered as a better approach. Furthermore, the use rate of vocabularies is a more important argument for reuse than the frequency of a single vocabulary term. Their survey can help to understand why there are some terms frequently used and why some of them are not used at all.

3 Analysis Method

Our analysis method consists of two steps. First, we determined vocabularies that have more than one published version on the web. Second, we investigated how the changed terms of vocabularies are adopted and used in the evolving KGs. For the first step, we relied on Schmachtenberg et al. [14] who published a report with detailed statistics about a large-scale snapshot of the LOD cloud. The snapshot comprises seed URIs from the datahub.io dataset[4], the BTC 2012 dataset[5], and the public-lod@w3.org mailing list[6]. We selected a set of vocabularies that satisfy the following set of conditions and characteristics. (1) The vocabulary have at least two versions published on the web to make a comparison between them. (2) These two versions are covered by the dataset that we investigate. For example, for the DyLDO dataset, there is to be one version of the vocabularies that have been published after May 6th, 2012. This is needed since at this date the first snapshot of the DyLDO dataset has been crawled. (3) The vocabulary terms are directly used for modeling some data, i. e., the vocabulary terms occur in at least one triple in the published dataset. In contrast, vocabularies could also be just linked from a data publisher, where changes of external vocabularies may not have any impact on the published data.

On the basis of these criteria, we examined 134 of the most used vocabularies listed in the state of the LOD cloud 2014 report by Schmachtenberg et al. [14]. We found 18 vocabularies that have more than one version. From them, 13 vocabularies have changes (additions or deprecations) on terms created by the ontology engineers of those vocabularies in the timeframe of the considered datasets. We downloaded the different versions using the Linked Open Vocabu-

[4] http://datahub.io/group/lodcloud, last accessed: 10/10/2017
[5] http://km.aifb.kit.edu/projects/btc-2012/, last accessed: 10/10/2017
[6] http://lists.w3.org/Archives/Public/public-lod/, last accessed: 10/10/2017

laries (LOV) observatory[7]. Due to the low number of changes in the vocabulary, we did not use data mining techniques. Instead, we exploited the PromptDiff Protégé 4.3.0[8] plugin to identify the vocabulary changes. This plugin identifies simple as well as complex changes, and shows the difference between two versions. The vocabularies selected are listed in Table 1, which also provides the number of versions considered for each vocabulary and the total number of changes (additions and deletions) occurred. Considering all the vocabularies and all their versions the total number of terms studied is 936.

Table 1: Overview of the vocabularies and their changes.

Vocabulary	Versions	Changes
Asset Description Metadata Schema (ADMS)[9]	2	18
Citation Typing Ontology (CiTO)[10]	3	218
The data cube vocabulary (Cube)[11]	2	6
Data Catalog Vocabulary (DCAT)[12]	2	13
A vocabulary for jobs (emp)[13]	2	1
Ontology for geometry (geom)[14]	2	2
The Geonames ontology (GN)[15]	7	31
The music ontology (mo)[16]	2	46
Open Annotation Data Model (oa)[17]	2	31
Core organization ontology (org)[18]	2	8
W3C PROVenance Interchange (Prov)[19]	5	168
Vocabulary of a Friend (voaf)[20]	4	8
An extension of SKOS for representation of nomenclatures (xkos)[21]	2	1

Subsequently, we investigated how the vocabulary terms changed are used in the evolving KGs. We extracted all PLDs from the subject of the crawled triples that use any of the terms from the 13 vocabularies above. Specifically, we relied

[7] http://lov.okfn.org/dataset/lov, last accessed: 10/10/2017
[8] http://protege.stanford.edu, last accessed: 10/10/2017
[9] https://www.w3.org/TR/vocab-adms/, last accessed: 10/11/2017
[10] https://sparontologies.github.io/cito/current/cito.html, last accessed: 10/11/2017
[11] http://www.w3.org/TR/vocab-data-cube/, last accessed: 10/11/2017
[12] https://www.w3.org/TR/vocab-dcat/, last accessed: 10/11/2017
[13] http://lov.okfn.org/dataset/lov/vocabs/emp, last accessed: 10/11/2017
[14] http://data.ign.fr/def/geometrie/20160628.htm, last accessed: 10/11/2017
[15] http://www.geonames.org/ontology/documentation.html, last accessed: 10/11/2017
[16] http://www.geonames.org/ontology/documentation.html, last accessed: 10/11/2017
[17] http://www.openannotation.org/spec/core/, last accessed: 10/11/2017
[18] https://www.w3.org/TR/vocab-org/, last accessed: 10/11/2017
[19] https://www.w3.org/TR/prov-o/, last accessed: 10/11/2017
[20] http://lov.okfn.org/vocommons/voaf/v2.3/, last accessed: 10/11/2017
[21] http://rdf-vocabulary.ddialliance.org/xkos.html, last accessed: 10/11/2017

on the Guava library[22], which returns the PLD from any given URL. Besides the date of the first appearance of a vocabulary term, we also recorded the number of triples which contain that term. This information is then used to compute the adoption time of terms over the dataset snapshots.

4 Datasets

We applied our analysis approach on three large-scale KGs. The first two are DyLDO and BTC, which are obtained from the Linked Open Data cloud, and the third is Wikidata. We analyzed the use, changes and adoption of vocabularies within each individual dataset. We did not compare any results across the datasets because the results cannot be meaningfully compared. Below, we briefly characterize the datasets.

DyLDO is a repository to store weekly snapshots from a subset of web data documents [8]. For our study, we parse 242 snapshots (from May 2012 until March 2017). BTC[23] is yearly crawled from the LOD cloud from 2009 to 2012, as well as in 2014. We used all available versions to analyze the adoption of the extracted vocabularies in our study. Wikidata[24] is a knowledge base to collaboratively store and edit structured data. To analyze the Wikidata vocabulary, we first extracted the terms introduced by this vocabulary. Specifically, through the RDF Exports from Wikidata page[25], we parsed the terms and properties from the RDF dump files that were generated using the Wikidata toolkit[26]. We assumed that the first snapshot of those files is the first version of the Wikidata vocabulary, and based on this assumption we parsed the next dump files to extract the changes to the first version, and so on. Relying on the 25 RDF dump files (from April 2014 until August 2016), we extracted the terms that are added or deprecated. Subsequently, we parsed those files to extract the adoption of terms to analyze the adoption behavior for the Wikidata vocabulary's terms.

5 Results

In this section, we summarize our findings based on the experiments conducted. Section 5.1 presents the results of vocabulary changes, use, and adoption in the LOD Cloud, while Section 5.2 outline the same findings for Wikidata.

5.1 The LOD Cloud

Changes in LOD Vocabularies We studied the changes of terms in the vocabularies, focusing on creation and deprecation. Overall we observed 35 % of

[22] https://github.com/google/guava/, release 23.1
[23] http://challenge.semanticweb.org/, last accessed: 29/11/2017
[24] https://www.wikidata.org/, last accessed: 29/11/2017
[25] https://tools.wmflabs.org/wikidata-exports/rdf/, last accessed: 29/11/2017
[26] https://github.com/Wikidata/Wikidata-Toolkit, last accessed: 29/11/2017

newly created terms and 11 % of deprecated ones. 85 % of the vocabularies in this study have an increased number of terms. Two exceptions are *ADMS* and *CiTO*: the number of terms decreased for the former, while the latter vastly dropped in the number of classes.

During our analysis, we noticed that some of the deprecated properties were reintroduced later. These reintroduced terms belongs to the *CiTO* and *GN* vocabularies. The former deprecated 18 properties in May 2014 (introduced in March 2010), which reappeared in the version that was published in March 2015, i. e. after around ten months. The latter reintroduced three deprecated properties: `alternateName` (creation: October 2006, deprecation: September 2010, recreation: February 2012), `name` (creation: October 2006, deprecation: September 2010, recreation: October 2010), and `shortName` (creation: September 2010, deprecation: May 2010, recreation: February 2012). *GN* reintroduced two out of three deprecated terms after about 17 months and one shortly after (13 days).

Use of LOD Vocabularies We analyzed the use of the selected vocabularies by considering triples which contains one of their terms in the predicate and/or the object position and a PLD in the subject. *Geonames.org* is the PLD that uses most terms of the selected vocabularies in the BTC 2009 and 2010 datasets (Table 2). In BTC 2011 and 2012, *zitgist.com* and *rdfize.com* are the most frequent PLDs, while in BTC 2014 and DyLDO, *dbtune.org* accounts for the highest use. However, the number of triples in BTC 2009, 2011, and 2012 is significantly lower than for the other datasets. The PLDs with the highest use of certain terms vary over time. For example, *geonames.org*'s triples did not disappear in BTC 2011. It still accounts for around 500,000 triples, i. e. much less than BTC 2009 and BTC 2010, but Table 2 only lists the PLD that has the highest amount of triples for each dataset (*zitgist.com* for BTC 2011). Please note that there are different crawling strategies for each BTC dataset and this may contribute to the variations in the number of triples.

Table 2: PLDs with the highest use of terms from the selected vocabularies for each of the datasets.

Dataset	PLD	Triples
BTC 2009	geonames.org	81M
BTC 2010	geonames.org	7M
BTC 2011	zitgist.com	2.6M
BTC 2012	rdfize.com	3.8M
BTC 2014	dbtune.org	81.5M
DyLDO	dbtune.org	160M

In DyLDO, the use of most vocabularies is steady. Figure 1 shows the vocabularies with a varying use. Notably, *mo* shows increasing and declining intervals, *Prov* is increasing in popularity despite some slight negative picks, while *ADMS*

had a significant drop in 2015 after an initial increasing use, although it seems again slightly increasing from 2015 to 2017. Furthermore, *Cube* had a pick towards the end of 2015 to then come back to its initial use rate, while *emp* seems no more used from 2015.

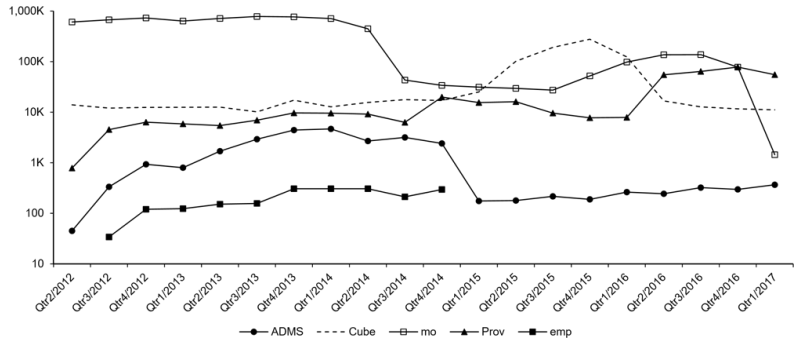

Fig. 1: The mean number of triples that use terms for a subset of the vocabularies considered by DyLDO snapshots aggregated in quarters.

The great majority of the deprecated terms (87 %) are still used after deprecation. We found that *geonames.org* is the PLD that most frequently uses deprecated terms in the BTC and DyLDO datasets. For instance, Figure 2 shows the use of the `gn:Country` class in DyLDO, which was deprecated in September 2010. Despite various fluctuations, its use increased until August 2015, then declined and increased again to reach a peak in August 2016.

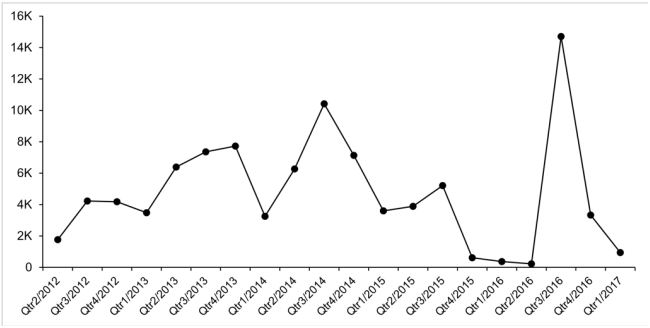

Fig. 2: The use of `gn:Country` class in the DyLDO dataset.

Adoption of LOD Vocabualry Changes The majority of the newly created terms were adopted in less than 10 days, as showed in Table 3. The triples column

represents the total number of triples in DyLDO that contains the adopted terms, while μ and σ are the average number of days before adoption and the standard deviation, respectively. Additionally, adopting *geom* and *GN* terms took long time.

Table 3: The adoption of newly created terms for each of the vocabularies.

Vocabulary	New terms	Adopted terms	Triples	μ	σ
ADMS	6	100 %	31K	7	0
CiTO	80	100 %	281K	7	0
Cube	5	100 %	15K	7	0
DCAT	5	100 %	104K	8.4	3.13
emp	1	100 %	4K	7	0
geom	2	100 %	16K	420	0
GN	21	100 %	160M	127.76	255.33
mo	44	100 %	45M	8.75	9.68
oa	21	0 %	-	-	-
org	8	100 %	173K	7	0
Prov	106	85 %	121M	30.15	37.49
voaf	10	100 %	75K	43.33	68.58
xkos	1	0 %	-	-	-

After being adopted, 50 % of the newly created terms decreased in use during the considered period, 47 % showed a steady use, while 3 % increased. For example, during its evolution, the *voaf* vocabulary created 10 new terms. All but one of those have a decline in the use. Figure 3 shows only six terms as the remaining are exploited in much fewer triples. In general, a similar trend holds for all the vocabularies. More details about the adoption time of other vocabularies are available in an extended technical report [1].

Not all terms are adopted. For example, the percentage of adoption for half of the vocabularies is less than 50 % of terms in the BTC dataset (in total, 50 % of all terms were not used). While in DyLDO, the percentage of unused terms of all vocabularies was 23 %, and only one vocabulary (*CiTO*) adopted 60 % of the terms, while the remaining vocabularies less than 40 % (Table 4). Notably, the 21 new terms of the *oa* vocabulary and the only *xkos* term are never adopted.

5.2 Wikidata

After parsing the terms and properties from the RDF dump files for the period from April 2014 until August 2016, we have extracted the added and deprecated terms of the Wikidata vocabulary. Figure 4 presents the total number of classes and properties in each Wikidata snapshot, which grows to reach 11 classes and 27 properties in August 2017. Ontology engineers added 3 classes and 9 properties during the analyzed period. Notably, there are no terms that are deprecated during the ontology evolution.

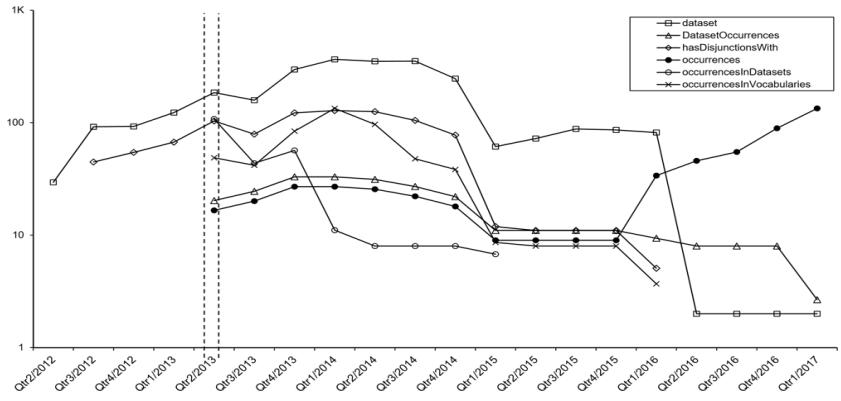

Fig. 3: The use (amount of triples in which a term occurs) of the *voaf's* newly created terms by quarters of DyLDO snapshots. The vertical dashed lines represent the publishing time of new versions of the vocabulary. Please note that two versions of *voaf* have been published before the first snapshot of DyLDO (i. e. `dataset` and `hasDisjunctionsWith` are newly created in versions released before the second quarter of 2012).

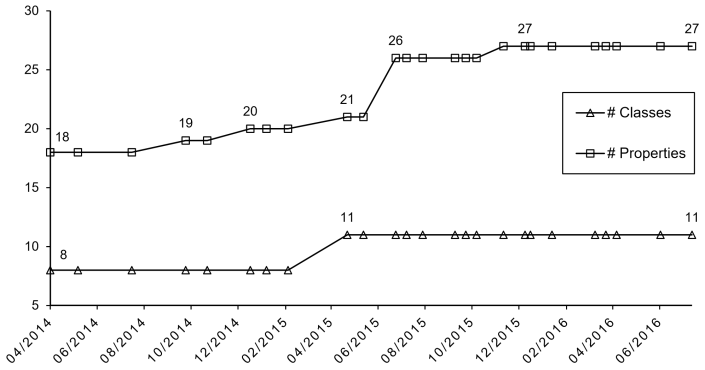

Fig. 4: Total number of terms of the Wikidata vocabulary per RDF dump file.

Figure 5 illustrates the use of newly created terms in Wikidata. Only 5 out of 12 terms are adopted. `NormalRank` and `rank` are much more used than the other new terms. Furthermore, the actually adopted terms among all the newly created ones are adopted directly after their creation date (i. e., on the same day). One possible reason is that Wikidata is a more controlled and centralized environment than a distributed KG, such as the LOD cloud, as discussed in Section 6.2.

Table 4: The percentage of unused terms in the BTC and DyLDO datasets.

Vocabulary	Total terms	BTC	DyLDO
ADMS	31	68 %	3 %
CiTO	220	72 %	60 %
Cube	37	35 %	0 %
DCAT	23	48 %	9 %
emp	31	87 %	6 %
geom	34	100 %	3 %
GN	43	26 %	9 %
mo	208	36 %	2 %
oa	63	83 %	35 %
org	44	20 %	11 %
Prov	143	22 %	24 %
voaf	24	33 %	8 %
xkos	35	63 %	14 %

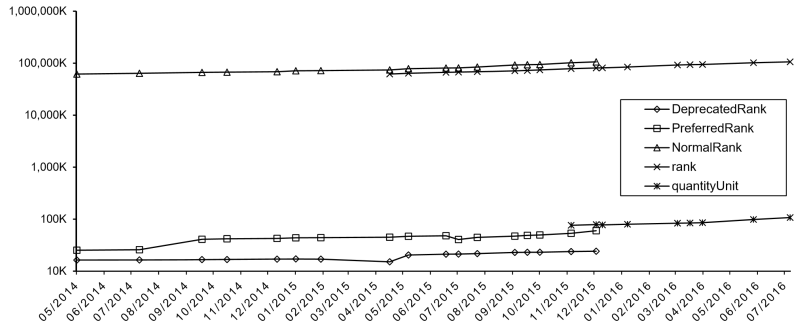

Fig. 5: The amount of triples that the adopted newly created classes and properties of the Wikidata vocabulary after parsing Wikidata RDF dump files.

6 Discussion

We found that not all vocabulary changes are reflected in the data in KGs, and there is a need for a service to track vocabulary changes. Such service helps ontology engineers and data publishers in updating their ontologies and models. In Section 6.1, we discuss the results related to the LOD Cloud, and in Section 6.2 we discuss the results of changes and adoption of the Wikidata terms.

6.1 The LOD Cloud

Changes in LOD Vocabularies The number of terms' changes is small. This is in line with existing studies [2, 6, 9]. However, those changes may have a large impact on the data of KGs. For example, the new version of the *oa* vocabulary caused a significant increased of its use: the triples containing its terms almost triplicates (from roughly 400 hundred to over 1100). In general, the changes

impact on the use either in an increasing or decreasing way (6 and 5 out of 13 vocabularies, respectively), although with varying time. For *DCAT* there is an increase so delayed in time (3 years) which is probably not due to the new version. More details are available in our extended technical report [1].

Most of the vocabularies increased in the total number of terms. This suggests that more knowledge is represented in the LOD cloud, requiring new terms. One exception is *CiTO*, which consisted of 94 classes and 36 properties when initially published. The second version counted only one class and 50 properties. Specifically, all the 94 classes were replaced with the new class *CitationAct* and most of the 36 properties of the first version were substituted. The third version provided 91 properties, although 18 of the new properties were reintroduced from the first version (deprecated in the second and reintroduced in the third). In practice, almost a new ontology was built. This is particularly important since *CiTO* has grown much in popularity (BTC 2014 contained over 300 thousand triples compared to 40 thousand in BTC 2011).

New versions of vocabularies, together with the great variety of vocabularies already existing, and the new vocabularies may overwhelm ontology engineers, which need to choice among a vast amount of alternative terms when building or updating their ontologies. Similar issues may occur to data publishers when deciding which vocabularies to exploit in their datasets. Missing some changes and consequently not updating an ontology or a dataset is likely (see the following discussion on the use of terms), notably in a distributed environment as the LOD cloud. This holds particularly for deprecation. There is a lack of tools to notify ontology engineers and data publishers when there are changes in the vocabularies. Such tools may help ontology engineers to select classes and properties that they want to use by knowing the latest updates of terms, and help them in updating their vocabularies. They could also assist data publishers in updating their models by providing a history of changes for the terms they use. While these systems can ease the maintenance of vocabularies and datasets, more advanced one could also recommend terms and vocabularies according to the specific needs of their users. The insights provided in this study can be beneficial to build such tools.

Use of LOD Vocabularies Cross-domain (*Prov*, *voaf*, and *ADMS*) and Geographic (*Cube* and *GN*) vocabularies were the most popular among data publishers. Some of them are exploited by few PLDs. For instance, *w3.org* widely used *ADMS* terms at the beginning of the investigated time-frame, while later *deri.de* accounted for the highest use of this vocabulary. On the other hand, some vocabularies have been used by various PLDs. For example, *Cube* has been employed by *ontologyCenter.com, esd.org.uk, linked-statistics.org*, and *linkedu.en*. This may suggest that some vocabularies are applicable in multiple domains, while others are more application-specific, but it should be further investigated.

Overall, *geonames.org* and *dbtune.org* are the most frequent PLDs. In the BTC 2009 and BTC 2010 datasets, *geonames.org* was the PLD that uses most of the terms. This is caused by the wide use of the *GN* vocabulary in those years.

Later, *dbtune.org* accounted for the highest number of triples in the BTC 2014 and DyLDO snapshots from 2012 to 2014.

Although some terms are deprecated, 87 % of them were still exploited. This is in line with [9]. *Geonames.org* is the PLD with the highest number of deprecated terms. For example, in the BTC 2011 dataset, *geonames.org* used six deprecated terms in about 522 thousand triples. That number declined to three terms and roughly 181 thousand triples in BTC 2012, but increased again to 49 terms in BTC 2014 (5.5 thousand triples). It seems is that data publishers did not update their data models. A possible reason of this is that they are not aware of changes in the vocabularies exploited. Thus, as previously discussed, they could benefit from tools to notify these changes.

In order to provide information about the status of a term, the Vocabulary Status ontology[27] can be used. This ontology consists of three properties: vs:term_status, vs:moreinfo, and vs:userdocs. Unfortunately, this ontology is not widely used. Only 7 out of the 134 vocabularies investigated in our paper rely on it.

Adoption of LOD Vocabualry Changes Most of the newly coined terms are adopted rather quickly (in less than one week). Surprisingly, we even found some terms adopted before their official publishing date. We believe that some of the new versions of vocabularies are already online and can be used before their official announcement. In some cases, it may take time to finish the procedures to publish the new version of the vocabulary. Thus, data publishers can access the new terms before their formal release, simply because they are available online.

Although most of the terms have fast adoption time, some vocabularies, such as *GN*, took more than 120 days, in average, to adopt new terms. However, this average does not reflect the actual adoption behavior: the new version of *GN* provides 21 new terms, 17 terms are adopted within 7 days, while the remaining 4 terms are adopted in over 600 days. Therefore, the average result was affected by those few terms that have a vast adoption time.

Another interesting point is that some newly created terms are never adopted. For example, ontology engineers published a new version of the *oa* vocabulary in June 2016, with 21 new classes and properties. None of those terms have been adopted (at least until April 2017, the last DyLDO snapshot considered), while the first version of *oa* was published in February 2013 with 42 terms and all but one were adopted in less than 3 months. However, the reasons why those terms are unused likely depends on the specific application scenario. For instance, not all terms need to be currently in use: some could be designed for future applications. Furthermore, although some terms are not used in the LOD cloud, they may be exploited in other forms. We do not mean that every term has to be adopted: we aim to raise awareness to ontology engineers that there are some of their terms that are never adopted. We also believe that raising awareness to data publishers about the existence of other terms in an ontology in use may further stimulate the reuse of ontologies.

[27] https://www.w3.org/2003/06/sw-vocab-status/note.html, last accessed: 27/02/2018

6.2 Wikidata

We found that the Wikidata vocabulary showed no deprecated terms, although some were never adopted during the investigated time-frame (e. g., the `Article` class). Likewise most of the LOD vocabularies, the Wikidata vocabulary counts a small number of additions (3 classes and 9 properties) and no deprecation.

Three classes (`DeprecatedRank`, `NormalRank`, and `PreferredRank`) suddenly disappeared from Wikidata statements after the snapshot in December 2015, after about 8 months (they were created in May 2015). There is a huge difference in the number of triples in which the terms occur. For instance, the `NormalRank` and `Statement` classes have been used in about 106 and 81 million triples, respectively. The other classes (except `Item`) are used in less than 2.4 million triples. The same observation can be made for the properties: all but `rank` appeared in less than 2.7 million triples, while `rank` accounted for approximately 62 million triples when introduced in May 2015, then reached about 106 million triples in August 2016. The wide exploitation of these terms suggests a pressing necessity for adding them to the vocabulary.

Only 5 out of 12 of the newly created terms are adopted and their adoption occurs directly after their creation date. This was expected in Wikidata which is a more controlled and centralized environment than a distributed KG as the LOD cloud. Surprisingly, the majority of new terms (2 classes and 9 properties) seems not adopted in any statements of Wikidata. However a deeper analysis showed that these are used to define properties and their types, except the `Article` class, which needs further investigation.

7 Conclusion and Future Work

Even small changes of vocabulary terms can have a deep impact on the real data that use those terms. Most of newly coined terms are adopted immediately afterwards, while 50 %, and 23 % of the terms studied are never adopted in the BTC and DyLDO datasets, respectively. Unexpectedly, some deprecated terms have been recreated after some time by their deprecation. Deprecation is a critical operation, notably in a distributed KG as the LOD cloud. We are not surprised that most of the deprecated terms are still used, because data publishers may not be aware of the changes to the exploited vocabularies. We think that this study can help ontology engineers and data publishers in updating their ontologies and datasets. Providing a service to notify changes on ontologies can simplify the update of vocabulary and datasets, as well as foster the adoption of new terms. In order to reproduce our research and extend on them, we provide all results and datasets to the public[28]. As future work, we plan to study the impact of vocabulary changes on the ontology network and provide a service for tracking changes on vocabulary which incorporate the insights of this study. Furthermore, we plan to investigate the impact of similarity measures on the reuse of vocabulary terms on the LOD cloud.

[28] https://figshare.com/s/d5487f88a2bdfab4c2ee

Acknowledgment This work was supported by the EU's Horizon 2020 programme under grant agreement H2020-693092 MOVING.

References

1. Abdel-Qader, M., Scherp, A.: Towards understanding the evolution of vocabulary terms in knowledge graphs. ArXiv e-prints (Sep 2017), https://arxiv.org/pdf/1710.00232.pdf
2. Abdel-Qader, M., Scherp, A.: Qualitative analysis of vocabulary evolution on the linked open data cloud. In: PROFILES Workshop co-located with ESWC, Volume 1598. CEUR-WS. org (2016)
3. Chawuthai, R., Takeda, H., Wuwongse, V., Jinbo, U.: Presenting and preserving the change in taxonomic knowledge for linked data. Semantic Web 7(6), 589–616 (2016)
4. Dividino, R., Scherp, A., Gröner, G., Grotton, T.: Change-a-LOD: does the schema on the linked data cloud change or not? In: Consuming Linked Data Workshop co-located with ISWC, Volume 1034. pp. 87–98. CEUR-WS. org (2013)
5. 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 and Parallel Databases 33(4), 515–553 (2015)
6. Guha, R.V., Brickley, D., Macbeth, S.: Schema.org: Evolution of structured data on the web. Communications of the ACM 59(2), 44–51 (2016)
7. Käfer, T., Abdelrahman, A., Umbrich, J., O'Byrne, P., Hogan, A.: Observing linked data dynamics. In: ESWC. pp. 213–227. Springer (2013)
8. Käfer, T., Umbrich, J., Hogan, A., Polleres, A.: Towards a dynamic linked data observatory. LDOW co-located with WWW (2012)
9. Meusel, R., Bizer, C., Paulheim, H.: A web-scale study of the adoption and evolution of the schema.org vocabulary over time. In: International Conference on Web Intelligence, Mining and Semantics. p. 15. ACM (2015)
10. Mihindukulasooriya, N., Poveda-Villalón, M., García-Castro, R., Gómez-Pérez, A.: Collaborative ontology evolution and data quality-an empirical analysis. In: International Experiences and Directions Workshop on OWL. pp. 95–114. Springer (2016)
11. Papavassiliou, V., Flouris, G., Fundulaki, I., Kotzinos, D., Christophides, V.: On detecting high-level changes in RDF/S KBs. In: ISWC. pp. 473–488. Springer (2009)
12. Roussakis, Y., Chrysakis, I., Stefanidis, K., Flouris, G., Stavrakas, Y.: A flexible framework for understanding the dynamics of evolving RDF datasets. In: ISWC. pp. 495–512. Springer (2015)
13. Schaible, J., Gottron, T., Scherp, A.: Survey on common strategies of vocabulary reuse in linked open data modeling. In: ESWC. pp. 457–472. Springer (2014)
14. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the Linked Data best practices in different topical domains. In: ISWC. pp. 245–260. Springer (2014)
15. Vandenbussche, P.Y., Atemezing, G.A., Poveda-Villalón, M., Vatant, B.: Linked open vocabularies (LOV): a gateway to reusable semantic vocabularies on the web. Semantic Web 8(3), 437–452 (2017)