Qualitative Analysis of Vocabulary Evolution on the Linked Open Data Cloud

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

[1] Christian-Albrechts University, Kiel, Germany
stu120798@mail.uni-kiel.de, asc@informatik.uni-kiel.de
[2] ZBW – Leibniz Information Centre for Economics, Kiel, Germany
a.scherp@zbw.eu

Abstract. We analyse the evolution of vocabularies on the Linked Open Data cloud. Based on the recent statistics of the LOD cloud, we have selected the twelve most dominant vocabularies in terms of their use in different pay-level domains. The number of versions we found for these vocabularies range between 2 to 11. While some ontologies exist for more than 10 years (e.g., FOAF) others are only online since a few years (like DCAT). Our analysis shows that many changes occurred on annotation properties. This reflects a need for more clarification of the terms, especially at early versions of the vocabularies. The majority of changes in the vocabularies are due to changes in other, imported vocabularies. Thus, there is a co-evolution of different vocabularies. This insight has practical impacts to ontology engineers. They not only need to consider the evolution of the vocabularies they directly use, but also those they import and indirectly depend on.

1 Introduction

The Semantic Web main objective is to give data in the web a well-defined meaning. Those meanings can be represented using ontologies, which can be defined as a descriptive form for the concepts and items in a specific field or domain and provides specifications for those items and its relations to other concepts [1]. The ontologies are subject to change (evolve) over time for many reasons, such as changes in the ontology's domain, resolving errors appeared in the current or previous versions of the ontology, changes in its external vocabularies that are used to establish those ontologies, or any other reasons for updating the ontology.

Creating a new version of an ontology requires processes to handle and manage multiple versions of that ontology [5]. Many research focused on analysing the evolution of some ontologies, and does not focus on the core of establishing those ontologies; the vocabularies. We mean by vocabulary, a collection of basic terms (types and properties) that have a broad meaning. Those vocabularies can be general (suitable for all domains) or specific (some domains or a single domain).

In this paper, we focus on analysing the changes that occurred on a selected set of vocabularies. We analyse them from different perspectives and observe how vocabularies are influenced by changes made in their dependency vocabularies. Furthermore, we analyse the types of changes that occurred in vocabularies terms (classes and properties). Those changes can be additions, deletions, modifications, or renaming of vocabulary terms. Other changes such as splits/merges can be considered as additions/ deletions processes.

We clarified the percentage of changes occurred on the examined vocabularies that caused by the external vocabularies they depend on establishing their vocabularies, and which vocabularies are depend on their terms on establishing and evolving the vocabulary. All those analyses are useful for ontology engineers when they are establishing a new ontology or updating an existing one by having a clear idea about dependencies and relations between vocabularies to choose vocabularies terms that meet their needs.

The remainder of the paper is organized as follows: In Section 2, we present and discuss related work. We present our methodology for analysing vocabularies in Section 3. In Section 4 we give an overview for the examined vocabularies. In Section 5 we make a discussion of vocabularies evolution. Conclusion and future work in presented in Section 6.

2 Related Work

Current research focused on measuring what are the changes in the LOD cloud, but not how they are changed. For example, Dividino et al. [2] proposed a framework to measure the evolution of the data in a dataset over time. They applied their dynamics functions on 84 weekly snapshots from DyLDO dataset, which results a number that can be used to represent how data in the dataset are evolved. Furthermore, Dividino et al. [3] analysed the usage of vocabularies on the LOD cloud over time and observed how they are changed according to their usage. They analysed the combination of classes and properties that describe a resource and applied their analysis on a dataset by taking 53 weekly snapshots from DyLDO dataset. Over six months, Kfer et al. [4] observed the documents retrieved from DyLDO dataset they created. They analysed those documents using different factors, their lifespan, the availability of those documents and their change rate. Also, they analysed the RDF content that are frequently changed (added or removed). Finally they observed how links between documents are evolved overtime (either increased or decreased). To keep track what are the changes happened when publishing a new version of ontology, Neubert [6] compared the SKOS vocabulary versions files and found the differences between them and then stored those differences into two separated named graphs; insertions and deletions. Then he used the version history graph to link the insertion and deletion graphs with vocabulary versions files. Walk et al. [8] studied and analysed the user behaviour during editing ontologies to support the ontologies' editing tools. They derived nine hypotheses to describe the users' change behaviour, and then applied those hypotheses on four real-world ontology projects.

They found that the hierarchical structure hypothesis had the highest influence on the editing behaviourr. Furthermore, Walk et al. [9] analysed the collection of actions in the change-logs files that made by users in the collaborative ontology engineering environment, to increase the quality of ontologies they design. They applied Markov chains into the International Classification of Diseases (Revision 11) dataset. Zablith et al. [10] published a survey presenting an ontology evolution cycle, trying to gather researchers' work in ontology evolution community. Furthermore, they analyse the different approaches of each stage in the ontology evolution process. They suggest to integrate the tools used for ontology evolution, and share the research in this field using Web portals, beside sharing some common use cases that needs to evolve.

3 Methodology

Schmachtenberg et al.[7] published a report that provides a detailed statistics about the LOD cloud. They analysed a subset of the Linked Data Web. The subset is based on crawling seed URIs from the datahub.io[1] dataset, BTC 2012 dataset[2], and public-lod@w3.org[3] mailing list.

Based on their report, we selected the top used vocabularies in their crawled subset of the LOD Cloud and have different available versions to download. Our methodology can be expressed in the following steps:

- **Selection criteria:** We chose the vocabularies that have been used in more than five datasets (0.49% of 1014 datasets used in the statistics [7]).
- **Exclusion:** We excluded the vocabularies that have been used in less than five datasets because they are rarely used as shown in the previous statistics. Furthermore, we exclude some of the upper level ontologies like RDF, RDFS and OWL. For RDF and RDFS there are only one downloadable version for each of them. For OWL vocabulary, we excluded it because all our selected vocabularies and in all their versions use the same entities from it, and they are five Annotation properties (backwardCompatibleWith, deprecated, incompatibleWith, priorVersion, and versionInfo) beside "Thing" class. Therefore, we decided not to include them in this paper.
- **Selection result:** Based on our criteria in selecting the vocabularies, we examined 62 vocabularies that have been used in more than five datasets. We found twelve vocabularies from the 62 vocabularies had more than one version and can be downloaded. We tried to collect vocabularies that cover all the topical domains in the LOD Cloud.
- **Downloading:** We downloaded the available versions for the selected vocabularies using Linked Open Vocabularies (LOV) observatory[4] and the official

[1] http://datahub.io/dataset?tags=lod
[2] http://km.aifb.kit.edu/projects/btc-2012/
[3] http://lists.w3.org/Archives/Public/public-lod/
[4] http://lov.okfn.org/dataset/lov

sites of some vocabularies. By using Protg 4.3.0[1], we extracted the differences between each version to capture the evolution of those vocabularies. Table 1 shows the number of downloaded versions for each vocabulary, the period from the first to latest version for those vocabularies, the evolution duration in years/months, and the average number of changes per year.

– **Analysing:** We analysed the changes that occurred in different versions of the vocabularies (creating, deleting, modifying, or renaming). Those changes can be on classes, properties (with different types), datatypes, or individuals. We analysed the changes using different classifications. First, we count the number of changes for each type of entities, i.e., classes and properties. Subsequently, we observed the percentage of internal changes versus external changes. Internal changes are changes that occurred on the entities (classes and properties) that originally introduced and developed by ontology engineers of the examined vocabulary. On the contrary, external changes are changes on the vocabularies' entities from other vocabularies that are used to establish the examined vocabulary.

Table 1: Number of downloaded versions for the examined vocabularies and their evolution period, i.e., first appearance to latest version, sorted by the number of datasets they were used in based on the State of the LOD Cloud report 2014. In addition, the table shows to the evolution period in years/months and the average number of changes per year

Vocabulary	**#Versions**	**Evolution period**	**Duration in years/months**	**#Changes per year**
FOAF	10	03.04.2005-14.01.2014	8 years & 10 months	30
DCterm	3	14.01.2008-14.06.2012	4 years & 5 months	26
SKOS	8	26.03.2004-18.08.2009	5 years & 4 months	44
Cube	4	27.11-2010-31.07.2014	3 years & 8 months	10
bibo	2	03.06.2008-04.11.2009	1 year & 5 months	13
DCAT	6	24.04.2012-31.05.2014	2 years & 1 month	47
GN	11	05.10.2006-29.10.2012	6 years & 1 month	34
SWC	3	27.02.2009- 11.05.2009	3 months	184
PROV	3	03.05.2012-11.01.2015	2 years & 8 months	106
AIISO	3	14.05.2008-25.09.2008	4 months	79
ORG	10	06.06.2010-12.04.2014	3 years & 10 months	70
Cal	2	07.04.2004-11.01.2015	10 years & 9 months	6

4 Overview of the Vocabularies from LOD Cloud

In the following figures and tables, we summarize the statistics of the selected vocabularies and their different versions. According to the 1014 datasets crawled

[1] http://protege.stanford.edu

in the State of the LOD Cloud report (Version 0.4, 08/30/2014) [7], the examined vocabularies in our study were used in multiple topical domains.

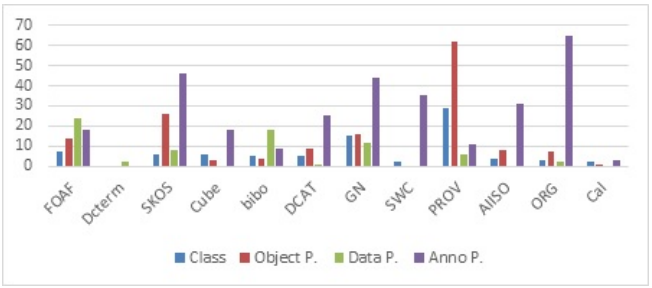
Table 2 shows the topical domains for our selected vocabularies. In addition, we show the percentage of the datasets crawled in [7] that use these vocabularies. The table shows twelve vocabularies. Please remind that we selected those vocabularies based on the availability of their versions (if they have) to download. Any vocabulary that does not have versions, or their previous versions are not available to download, are excluded from our analysis. We found that 65% of the 62 examined vocabularies just have one version, and 15% of those 62 vocabularies have more than one version but they are not available to download. Therefore, we excluded them from our study. In Table 2, we can see that two vocabularies (FOAF and DCterm) are used in more than 50% of datasets crawled in [7]. In addition, half of the selected vocabularies are classified as cross-domain vocabularies.

Table 2: Vocabularies according to their domains and dataset's percentage based on the State of the LOD Cloud report 2014 [7].

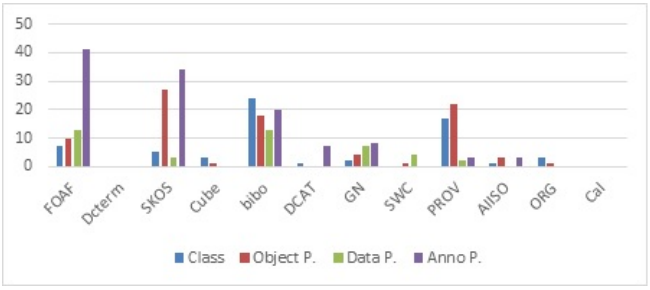
Vocabulary	**Domain**	**Number of datasets**
FOAF	Cross-domain	701 (69.13%)
DCterm	Cross-domain	568 (56.02%)
SKOS	Publications/ Cross-domain/ Geographic	143 (14.10%)
Cube	Government/ Geographic	114 (11.24%)
bibo	Cross-domain/ Social web/ Media/ Publications/Life Sciences	62 (6.11%)
DCAT	User-generated content/ Government/ Cross-domain	59 (5.82%)
GN	Geographic/ Life Sciences/ Media/ Social web	27 (2.66%)
SWC	Social web	27 (2.66%)
PROV	Government/ Cross-domain	21 (2.07%)
AIISO	Publications/ Life Sciences	17 (1.68%)
ORG	Social web	14 (1.38%)
Cal	Social web	9 (0.89%)

Fig 1 represent the total number of each change type, i.e. created, deleted, modified, or renamed classes, object properties, data properties, annotation properties, for all vocabularies we included in our analysis. From figure 1c, we can observe that most of the changes are related to the modification changes type, especially in object properties and classes. Also, the second most changes are the creating changes type, mostly in the annotation properties (figure 1a).

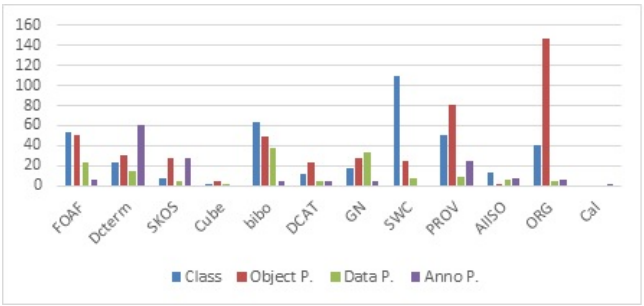
Table 3 shows the external vocabularies that are used in establishing each vocabulary in our study. From the external vocabularies listed in this table, we can note two things: First, there are three vocabularies (FOAF, DCterm and SWC) stuck on their external vocabularies list that used in establishing their first edition until the recent one. Second, we can note that there are two vocabularies



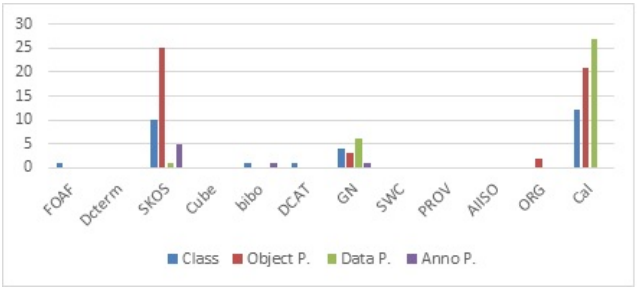
(a) Total number of created entities



(b) Total number of deleted entities



(c) Total number of modified entities



(d) Total number of renamed entities

Fig. 1: Total number for each change type (Create, delete, modify, and rename) for all versions of the examined vocabularies

(GN and ORG) have many changes on their external vocabularies list during their evolution period. GN partially changed its external vocabularies five times during six years, and ORG four times during approximately four years.

The percentage of internal changes compared to external changes in the vocabularies versions are shown in Table 4. We calculate the total percentage for all internal changes occurred through the vocabulary evolution overall its versions; from first to latest version. We can see that three vocabularies with more than 90% of internal changes (DCterm, GN, and PROV). Furthermore, there is just one vocabulary (Cube) with a percentage of internal changes less than 50%.

5 Discussion of Vocabularies Evolution

We observe that most of the vocabularies if they evolves, most of the changes occurred on annotation properties for more explanations (metadata) to clarify classes, properties or individuals. This can be expressed as a need for more clearance for terms, especially between the early versions. Another observation is that vocabularies are highly static w.r.t. the number of external vocabularies used to establish and develop them. We can conclude that the topical domains are fully covered with terms in the existing vocabularies, and if there is a need of change, ontology engineers can modify the existing vocabularies.

In most of the examined vocabularies, the terms that changed were internal terms, i.e. the terms that were created for the examined ontology by ontology engineers, not the terms that are imported from external vocabularies during establishing or evolving the vocabulary. On the other hand, some vocabularies such as Cube, bibo, DCAT, and ORG are changed because some changes in their external vocabularies occurred. For example, in the Cube vocabulary, the percentage of the internal change is 43%, and the remaining percentage of change is caused by other external terms over its four published versions of this vocabulary. The first version from the Cube vocabulary was published in 27.11.2010, and some of its external vocabularies are DCterm and FOAF. DCterm published its latest version in 14.06.2012, before the next version of Cube (was in 02.03.2013) was published.

Another example is the bibo ontology. Analysing ten versions from that vocabulary, we conclude that 35% of changes are caused by external vocabularies. bibo uses many external vocabularies such as DCterm and SKOS, and both of them had versions between the two published versions of bibo (first version was in 03.06.2008 and the latest version was in 04.11.2009).

Vocabularies such as DCterm, GN, and PROV are dependent on their own terms, and most of the changes are made on those terms. For example, in the GN ontology, we analysed eleven versions, and the percentage of internal changes are 97%, which means that when the ontology engineers needs to change, they change their own terms. Another observation in DCterm and PROV vocabularies is that their external vocabularies as shown in Table 4 are small, and they are organized as upper level ontologies.

Table 3: External vocabularies usage during examined vocabularies' evolution

Vocab.	External vocabularies	Notes
FOAF	dc/owl/rdf/rdfs/vs/wot/xml/xsd	For all versions
DCterm	dcam/owl/rdf/rdfs/skos/xml/xsd	For all versions
SKOS	dc/dcterm/owl/rdf/rdfs	In first version
	dc/dcterm/owl/rdf/rdfs/foaf/vs	V. from 31.03.2005 to 29.08.2008
	dcterm/owl/rdf/rdfs	V. from 17.03.2009 to 18.08.2009
Cube	dcterm/foaf/owl/rdf/rdfs/ scovo/skos/void/xml/xsd	For all versions
bibo	Address/dc/dcterm/event/foaf/owl/rdf/rdfs/ skos/time/vann/vs/ wgs84_pos/xml/xsd	V. 03.06.2008
	Event/foaf/ns/owl prism/rdf/rdfs/schema/ skos/dcterm/vann/xml/xsd	V. 04.11.2009
DCAT	dcterm/owl/rdf/rdfs/xml/xsd	In first version
	dc/dcterm/dctype/foaf/owl/rdf/rdfs/ schema/skos/vcard/vann/voaf/xml/xsd	In other remaining versions
GN	skos/owl/rdf/rdfs/xml/xsd	V. from 05.10.2006 to 10.05.2010
	cc/dcterm/foaf/owl/rdf/rdfs/skos/ wg84_pos/xml/xsd	V. 22.09.2010
	dcterm/foaf/owl/rdf/rdfs/skos/xml/xsd	V. 05.10.2010
	cc/dcterm/foaf/owl/rdf/rdfs/skos/vann/ voaf/xml/xsd	V. 14.02.2012
	adms/cc/dcterm/foaf/mrel/owl/rdf/rdfs/skos/ vann/xml/xsd	V. 29.10.2012
SWC	bibtex/dc/dcterm/doap/foaf/geo/cal/misc/owl/rdf/ rdfs/sioc/swrc_ext/vcard/vs/wordnet/xml/xsd	For all versions
PROV	owl/rdf/rdfs/skos/xml/xsd	In first version
	owl/rdf/rdfs/xml/xsd	In other remaining versions
AIISO	cc/dc/dcterm/dctype/owl/rdf/rdfs/skos/vann/ xml/xsd	In first two versions
	cc/dc/dcterm/dctype/foaf/owl/rdf/rdfs/skos/vann/ vs/xml/xsd	In latest version
ORG	dc/foaf/gr/opmv/owl/time/rdf/rdfs/skos/vcard/ xml/xsd	In first version
	dcterm/foaf/gr/opmv/owl/time/rdf/rdfs/skos/ vcard/xml/xsd	V. from 08.10.2010 to 30.09.2012
	dcterm/foaf/gr/opmv/owl/time/prov/rdf/rdfs/ skos/vcard/xml/xsd	V. 06.10.2012
	dcterm/foaf/gr/owl/time/prov/rdf/rdfs/skos/ vcard/xml/xsd	V. from 15.02.2012 to 12.04.2014
Cal	dt/owl/rdf/rdfs/xml/xsd	In first version
	dc/dt/xhtml/owl/rdf/rdfs/xml/xsd	In latest version

Table 4: Internal changes percentage

Vocabulary	Internal changes percentage
FOAF	79%
DCterm	99%
SKOS	89%
Cube	43%
bibo	65%
DCAT	71%
GN	97%
SWC	86%
PROV	98%
AIISO	89%
ORG	73%
Cal	77%

The vocabularies in our study are different in their evolution period and the number of versions published so far. In our study, we are trying to analyse the evolution behaviour for those vocabularies. Some vocabularies such as FOAF, SKOS, GN, and ORG have many versions; 10, 8, 11, and 10, respectively. GN has eleven versions in the period from 05.10.2006 to 29.10.2012, and ORG have ten versions in the period from 06.06.2010 to 12.04.2014. We think this large number of versions is caused by their topical domain they are used in, especially in social web ontologies.

The last observation is that the vocabularies used in publications, geographic, social web, and government topical domains have the largest number of changes, this is obvious from Fig 1 using SKOS, GN, SWC, FOAF, and PROV vocabularies number of changes. For example, PROV added 29 classes and 62 object properties, and modified 51 classes and 80 object properties during its evolution period (Three versions published from 03.05.2012 to 11.01.2015). Another example is SWC, the ontology engineers modified 110 classes and 25 object properties through the three published versions.

6 Conclusion and Future Work

Analysing the change behaviour of vocabularies can help ontology engineers in establishing new ontologies and evolve existing ones. This study can give a clear view about ontology dependencies (External vocabularies), and the relation between change in ontologies and their external ones.

Changes are mostly made for internal terms if they compared by external terms percentage. Furthermore, most of the vocabularies have a static number of external ontologies to depend on during the evolution period, and if they add or remove some vocabularies, the number of those additions and deletions is small.

Topical domains such as publications, geographic, social web, and government have a high percentage for change if they compared with other domains.

As a future work, we will analyse the usage of vocabularies' in the Dynamic Linked Data Observatory (DyLDO) dataset. We will select the vocabularies that have a version before and after a specific snapshot of a DyLDO crawl (the first snapshot was in 06.05.2012, the latest is until today). Furthermore, we will include the vocabularies not considered so far in this study. Establishing a framework for ontologies' concepts history tracking system will be useful for ontology engineers.

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

References

1. Eder, J., & Koncilia, C. (2004, January). Modelling changes in ontologies. In On the Move to Meaningful Internet Systems 2004: OTM 2004 Workshops (pp. 662-673). Springer Berlin Heidelberg.
2. Dividino, R., Gottron, T., Scherp, A., & Grner, G. (2014). From Changes to Dynamics: Dynamics Analysis of Linked Open Data Sources. In PROFILES14: Proceedings of the Workshop on Dataset ProfiIling and Federated Search for Linked Data.
3. Dividino, R. Q., Scherp, A., Grner, G., & Grotton, T. (2013). Change-a-LOD: Does the Schema on the Linked Data Cloud Change or Not?. In COLD.
4. Kfer, T., Abdelrahman, A., Umbrich, J., OByrne, P., & Hogan, A. (2013). Observing linked data dynamics. In The Semantic Web: Semantics and Big Data (pp. 213-227). Spring-er Berlin Heidelberg.
5. Klein, M. C., & Fensel, D. (2001, July). Ontology versioning on the Semantic Web. In SWWS (pp. 75-91).
6. Neubert, J. (2015): Leveraging SKOS to trace the overhaul of the STW Thesaurus for Eco-nomics, In: Proceedings of the International Conference on Dublin Coreand Metadata Appli-cations 2015, So Paulo (Forthcoming)
7. Schmachtenberg, M., Bizer, C., & Paulheim, H. (2014). State of the LOD Cloud 2014. URL: http://linkeddatacatalog. dws. informatik. unimannheim. de/state.
8. Walk, S., Singer, P., Noboa, L. E., Tudorache, T., Musen, M. A., & Strohmaier, M. (2015). Understanding how users edit ontologies: comparing hypotheses about four real-world pro-jects. In The Semantic WebISWC 2015 (pp. 551-568). Springer International Publishing.
9. Walk, S., Singer, P., Strohmaier, M., Helic, D., Noy, N. F., & Musen, M. A. (2015). How to apply Markov chains for modeling sequential edit patterns in collaborative ontology-engineering projects. International Journal of Human-Computer Studies, 84, 51-66.
10. Zablith, F., Antoniou, G., d'Aquin, M., Flouris, G., Kondylakis, H., Motta, E., ... & Sabou, M. (2015). Ontology evolution: a process-centric survey. The Knowledge Engineering Review, 30(01), 45-75.