Linked Open Vocabulary Ranking and Terms Discovery

Ioannis Stavrakantonakis, Anna Fensel, Dieter Fensel
University of Innsbruck, STI Innsbruck
Technikerstr. 21a
6020 Innsbruck, Austria
firstname.lastname@sti2.at

ABSTRACT

Searching among the existing 500 and more vocabularies was never easier than today with the Linked Open Vocabularies (LOV) curated directory list. The LOV search provides one central point to explore the vocabulary terms space. However, it can be still cumbersome for non-experts or semantic annotation experts to discover the appropriate terms for the description of given website content. In this direction, the proposed approach is the cornerstone part of a methodology that aims to facilitate the selection of the highest ranked terms from the abundance of the registered vocabularies based on a keyword search. Moreover, it introduces for the first time the role of the contributors' background, which is retrieved from the LOV repository, in the ranking of the vocabularies. With this addition, we aim to address the issue of very low scores for the newly published vocabularies. The paper underlines the difficulty of selecting vocabulary terms through a survey and describes the approach that enables the ranking of vocabularies within the above mentioned methodology.

CCS Concepts

•**Information systems** → **Web data description languages;**

Keywords

vocabulary term; ontology search; semantic annotation

1. INTRODUCTION

The past few years many formats and vocabularies have been developed to support the implementation of semantic annotations that enable a website to be leveraged to a resource of machine understandable content. Thus, there is a matrix of combinations that can be used in order to add meaning to the content of a website.

According to Common Web Crawler[1] only ca. 17% of domains were found with triples in 2014, ca. 13.8% in 2013 and ca. 5.6% in 2012. In addition, the corpus analysis by [11] and [3] depicts a significant growth rate year over year, showing the realisation by the website owners of the importance of exposing structured data on the Web. On the other hand, the vocabulary discovery and implementation ease still remains an obstacle for the developers. On an earlier study, that we conducted, aiming to explore the uptake of semantics in the tourism business domain, which relies to a great extent on the online presence, we realised that local hotel businesses are far behind the structured data paradigm with triples existence ratio to be significantly low. As presented in [18], we discovered only 5% of hotel websites with any type of semantic annotations within the borders of Austria in 2014 and in a sample of ca. 2100 URLs.

Exploring further the vocabulary space, we observe an increase of the registered vocabularies, which is illustrated by the Linked Open Vocabulary directory creators in [19]. Therefore, it is important to bridge the gap between the two sides in order to have structured machine understandable entities and websites in the Web of Data in the future. Not facilitating the discovery of vocabularies, while they grow in number makes it more and more difficult for the consumers of vocabularies to produce added value by using them on the Web. Furthermore, in a recent survey by Meusel et al. in [10], the authors concluded that *"the adoption of new classes and properties in general is happening slowly, and that there are certain parts of the schema which are barely used at all."*, referring to the schema.org classes and properties, which is the most famous vocabulary as it is defined by the major search engine providers.

The topic of selecting the most semantically relevant terms for a given piece of information has been under research for a couple of publications lately; mostly because it is believed to be a barrier to the uptake of the vocabularies and ontologies in the modelling of real world scenarios as well as the inclusion of semantic annotations in the webpages. Systematic examination of existing Linked Data, like in [9], proves the complexity in the generation of semantic annotations as over half of the examined sites confuse Object Properties with Datatype Properties.

To address the difficulty in the annotations creation process, we designed an approach on top of the provided LOV functionality that aims to facilitate the selection of vocabulary terms. In this respect, the presented work focuses a) on the low scoring of newly published vocabularies within ranking formulas that work on top of the LOV scoring, like our LOVR framework [17] and b) on the recommendation of vo-

[1] http://www.webdatacommons.org/structureddata/

SEMANTiCS 2016, September 12-15, 2016, Leipzig, Germany

 ISBN 978-1-4503-4752-5/16/09…$15.00

DOI: http://dx.doi.org/10.1145/2993318.2993338

cabulary terms for static parts of the webpage content, like the multimedia HTML tags; and c) proposes a new vocabulary, the *vSearch*, that can be used to describe the results of a keyword based search.

The remainder of the paper describes related approaches that aim to address similar obstacles in the process of creating semantic annotations in Section 2; discusses a survey that we conducted to measure the ease of selecting vocabulary terms in Section 3; presents the proposed ranking approach as part of the LOVR framework in Section 4; evaluates the approach in Section 5; and closes with our conclusions in Section 6.

2. RELATED WORK

Bridging the gap between the need of search engines to make sense of unstructured crawled data and website developers to provide explicit meaning to website content via semantic annotations is crucial for realising the vision of the Semantic Web and the autonomous agents. In this respect, the Linked Open Vocabulary initiative [20] has a fundamental role by providing a curated directory of vocabularies. Each vocabulary in LOV is represented with a profile page[2] that provides useful metadata about the vocabulary itself, e.g. the namespace, the number of classes and properties of the latest version, the number of incoming and outgoing links, the versions history, the authors and the raw vocabulary schema in N3 notation. The curation follows a few specific rules related to the description of the vocabulary, as listed in [20]: a) be written in RDF and be dereferenceable; b) be parsable without error; c) terms should have an rdfs:label; d) reuse relevant existing vocabularies and e) provide metadata about itself. Therefore, the aforementioned rules do not guarantee the trust degree or effectiveness of a vocabulary, albeit the verification of the description and definition completeness.

In addition, a number of discovery interfaces have been implemented including a search for the vocabularies and the terms in order to find the most relevant resources for a given keyword; a SPARQL endpoint to query the data repository; and a JSON based REST API that facilitates the integration of the vocabularies and terms search functionality with external applications, like the approach that we present in this paper.

The LOV platform served 1.4 million queries, for a six months window in 2015, in total across the various search types, i.e. terms, vocabularies and agents. The agent type was introduced in 2015 and refers to the vocabulary contributors. The breakdown of the number of queries depicts that 92% of queries with a keyword were made for terms, while only 39% of the total number of term searches are using keywords as published in [19]. Another interesting aspect of the search figures is the percentage breakdown, which shows that 74% of the total searches refer to agent searches. That reflects a new dimension that we should at least experiment with, i.e. the authors of the vocabularies, as some of the enlisted persons are considered key people of the vocabulary community and it seems that could be the starting point for a user search in the process of finding vocabularies.

Introducing the related work with the LOV service reflects our consideration of the presented approach being strongly connected with the contribution of LOV. Research work relevant to vocabulary recommendation, ranking of vocabularies and usage of the LOV data in general are the directions that are considered related to our contribution.

Atemezing and Troncy in [1] examine the problem of vocabularies recommendation based on a ranking metric that has been developed by introducing the concept of Information Content (IC) to the context of LOV. The IC approach aims to rank the vocabularies, by evaluating the terms occurrence in comparison to the maximum term occurrence in the set of vocabularies and then leveraging the term rankings with a sum and a weight depending on the centrality of the vocabulary in the set.

On the ontology ranking topic, Butt et al. in [5] and [4] propose the DWRank algorithm. DWRank consists of two main scores, i.e. the Hub score and the Authority score, which measure the centrality of the concepts within the ontology and the ontology authoritativeness (i.e. the importance of the ontology in the ontologies space), respectively. DWRank ranks the concepts defined within the ontologies in order to find the best match to a keyword search based on the graph of vocabularies and without any Linked Open Data (LOD) usage parameter.

Discovering vocabularies can be assisted via many different directions apart from the ranking of vocabularies and vocabulary terms. Schaible et al. in [13] aim to support the ontology engineer by providing a methodology that guides the creation of Linked Data through the best practices for modelling new entities [8]. The approach is mainly based on Swoogle[3] and the SchemEX index and consists of an iterative process, where each iteration cycle finishes with the definition of one or several mappings of data to vocabulary terms. Furthermore, the vocabulary reuse is studied in [14] by presenting various approaches and one of the many extracted insights is the fact that using popular terms from popular vocabularies is preferred over using mainly one domain specific vocabulary that covers the needs of the given data. This insight is taken in consideration in our approach and reflected in the formulas presented later.

TermPicker presented in [15] experiments towards the direction of suggesting types and properties from vocabularies that other LOD providers have combined together with the one the engineer has used to model the given part. To achieve that, the authors introduce the schema-level patterns (SLPs), that represent the connection between two sets of RDF types (vocabulary classes) via a set of properties.

Ellefi et al. in [6] propose an approach to recommend datasets to a given non-linked dataset. The aim of the recommendation framework is to provide the user with an ordered list of datasets that are potential candidates for interlinking with the given input dataset. They base the interlinking on building a profile graph that provides information about the relationship between a document and a topic by following a topic modelling process.

In the next section, we provide the results of a survey we conducted in the context of the presented research. The survey is used in a twofold way within this publication. Firstly, it proves the difficulty of manually discovering and choosing vocabulary terms and secondly the results are compared against the output of the proposed algorithm and approach.

[2]LOV schema.org profile:
http://lov.okfn.org/dataset/lov/vocabs/schema

[3]http://swoogle.umbc.edu/

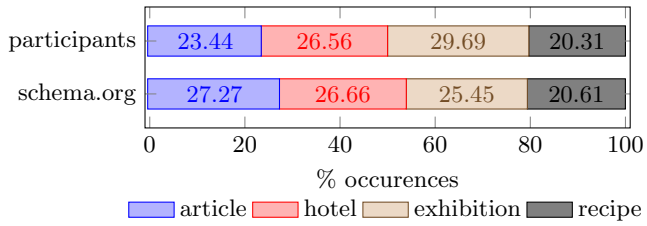

Figure 1: Distribution comparison of the number of participants per use case with the schema.org terms usage per use case.

3. SURVEY FOR MANUALLY DISCOVERING VOCABULARY TERMS

In the scope of the approach design we ran a survey to identify and measure the difficulties in the process of annotating a webpage in a way that would allow us to generalise and draw insights up to a significant extent and accuracy. The survey task was about discovering the appropriate vocabulary terms for a given webpage. The participants were asked to provide a justification on their decision of the proposed terms over other candidate vocabulary terms and also to specify their main difficulties throughout the discovery. In addition, they were asked to measure the time needed to complete the task. However, they were given a timeframe of one week to complete the task in order not to be artificially stressed. The assignment asked to use solely the LOV search for the discovery of terms. Therefore, the participants browsed through the webpage content to identify important keywords and then used them in the LOV search to retrieve vocabulary terms and pick the one that better refers to the input keyword according to the participant.

Also, they were asked to provide vocabulary terms for any keyword that they decided to be relevant, and additionally to provide alternative vocabulary terms that played the role of candidate terms during their selection of the result term for a given keyword. There was no explicit guideline whether they should include annotations for the multimedia content or only the text content. The participants were familiar to Computer Science topics, but without prior experience in the semantic annotations topic.

In order to introduce variety in the examined content types, we chose the four following tasks and randomly assigned one of them to each participant:

- *Article* from the NASA news[4] online feed,
- *Hotel* room webpage[5] from Austria,
- *Museum exhibition*[6] webpage of Louvre,
- *Recipe* webpage[7] for a pizza.

[4]http://www.nasa.gov/feature/jpl/nasas-curiosity-rover-team-confirms-ancient-lakes-on-mars
[5]http://www.mohr-life-resort.at/zimmer-und-preise/detail.html?rid=12
[6]http://www.louvre.fr/en/expositions/winged-victory-samothracerediscovering-masterpiece
[7]http://www.cookingchanneltv.com/recipes/debi-mazar-and-gabriele-corcos/margherita-pizza.print.html

Each one of the use cases requires a different set of vocabulary terms and describes a totally different concept. However, a few common terms can be used in order to describe basic elements of a webpage, like an image, a title of an entity or a hyperlink to another resource. The first use case is a news article about the discovery of evidence of water on planet Mars including pictures, address of the author and publication date among other details. The second one is a webpage about a hotel room offer mainly populated with prices, images and description of the amenities. The third use case refers to a museum exhibition and informs the visitors about the visiting period, the title of it, the admission fee, etc. Finally, the last use case is one of the most common semantic annotation examples, i.e. a recipe and describes all the steps, the ingredients, serving details and nutritional data. The distribution of the participants per use case is almost uniform with an average of 16 per use case and a total of 64 participants with valid submissions; detailed percentages are shown in Figure 1.

As it is shown in Figure 2, the time needed to complete the task varies depending on the use case difficulty to define the entities that could be annotated and to discover the appropriate terms. Putting all the distribution histograms together in the form of box plots for the four use cases shows that the article case holds the highest median value while the recipe case has the lowest measured time median together with the exhibition. We use the box plots diagrams in order to allow the comparison of all the distribution diagrams at the same time. The boxes represent the range in which the 50% of the data points fall in, while the two horizontal lines (whiskers) above and below the main box part refer to the maximum and minimum values respectively. The dots depict the outliers in the dataset and the horizontal line in the box depicts the median.

Also, it is interesting to realise that the article and the hotel cases are those with the most spread distributions in the experiment, which gives an indication that the various participants of this specific case interpreted differently the search results and searched for a different number of keywords; maybe due to uncertainty of which parts of the content are eligible for annotation.

Analysing the term URIs that the participants proposed gives indicators about the pitfalls that are hidden in the transformation of a webpage to an annotated data node. Furthermore, taking into consideration the reasons of their decisions we can realise that a few basic requirements should be met in order to make a vocabulary an option and potential solution for the vocabulary needs of the webpage development process.

The total number of the proposed term URIs is 499, while the used terms are 300. The difference is due to the fact that the participants were asked to include in their selection alternative candidate terms accompanying their main choice. As shown in Figure 3, the median of the number of selected terms for all the use cases is between 9 and 12. There are no outliers, albeit the fact that the maximum values are roughly twice as large compared to the 3^{rd} quartile (upper edge of the box). The hotel use case has attracted the highest median and maximum number of terms, which can be justified by the fact that the content of the respective webpage domain is simpler with easily recognisable entities (e.g. room, price). For the one third (33%) of the proposed term URIs, the namespace is the schema.org, while from the

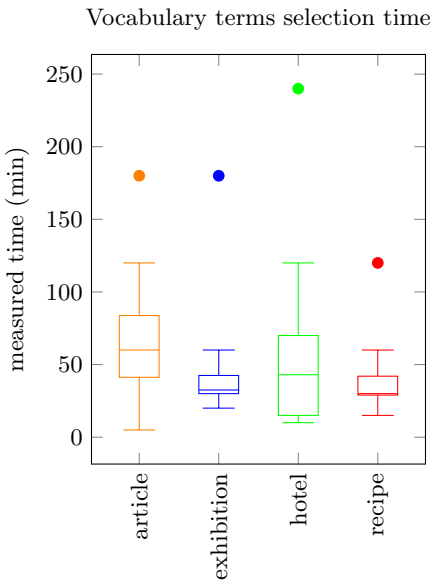

Figure 2: Distribution of the time needed by the participants to select vocabulary terms for each use case.

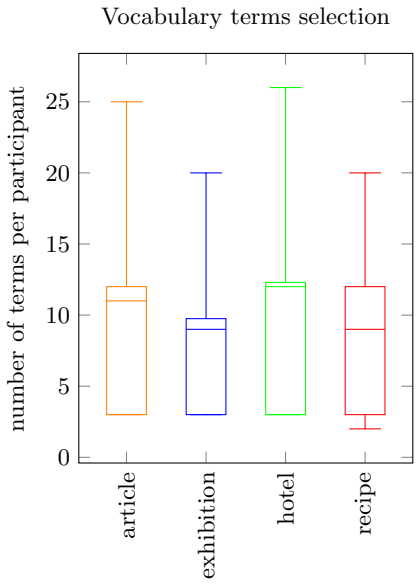

Figure 3: Distribution of the number of selected terms per participant for each use case.

used ones the same metric climbs to the 47%. This result shows a trend of schema.org terms being in favour over the rest of the candidate terms selected. Among the justification comments about their decision, the participants mention the good documentation of the vocabulary, the lack of documentation of terms in the rest of vocabularies or even vocabulary websites not accessible due to technical issues, i.e. 404 error pages. Grouping the results by use case, we see that the schema.org terms are following roughly the same distribution like the one of participants number per use case as shown in Figure 1.

Another interesting observation is the limited annotation of the multimedia content of the use cases. Only 10% (7 out of the 64) of the participants proposed to annotate the images of the webpage, while all the cases included images. This shows the difficulty of comprehensively addressing the topic of transforming a webpage to a machine understandable node.

Finally, we realised that some answers proposed completely wrong terms, like the *schema.org/HousePainter*, which refers to a house painting service according to the vocabulary documentation. Obviously, this term is irrelevant to the given context of all the use cases. Building on top of this survey, we describe in the next section (Section 4) the core processes and algorithms of our approach that aims to address all the above described difficulties that hinder the generation of high quality semantic annotations. Afterwards, at the evaluation section, we compare the results of the survey with the results from our approach against the same use cases that were given to the participants.

4. EXTENDING THE LOVR RANKING DIMENSIONS

Taking the aforementioned survey in mind, we designed an approach that is capable to facilitate the recommendation of vocabulary terms for a given webpage. The first version of the LOVR approach (stands for Linked Open Vocabulary Recommendation) with details about the framework infrastructure is presented in [17]. In the current paper, we mostly focus on the updated part regarding the ranking of vocabularies that are registered under LOV; the ranking of the vocabulary terms and the description of the discovered terms as part of a new vocabulary.

The ranking score across the LOV space facilitates the exploration of the most used vocabularies, reuse of those that are widely accepted by the community and the most relevant terms to the target webpage. In the presented contribution we simplify the LOVR approach by focusing on the part of discovering terms and skipping the part of extracting concepts from the webpage to use them as keywords. We assume the set of keywords K is already defined.

4.1 Ranking Vocabularies

Leveraging the findings presented in the related work section about the importance of the vocabulary authors, we introduce a dimension that reflects exactly that insight by giving a special weight to the authority of the vocabulary authors.

The idea behind the new metric is based on the assumption that a vocabulary created by authors that have already contributed to the vocabulary space has higher probability of being accepted by the community and broadly used by

the vocabulary engineers and web development stakeholders. Therefore, a newly introduced vocabulary will be preferred over another one if and only if the authors of it had created in the past vocabularies with a good ranking score. In this way, we aim to address an issue with new vocabularies that we have observed in all the existing approaches, to the best of our knowledge. Also, in the previous version of LOVR we had underlined a weakness of the formulas, which were agnostic about the existence period of the vocabulary leading to implicitly penalising those vocabularies that are newer in the LOV space, and probably less used in the LOD cloud and less reused by other vocabularies.

Thus, introducing the author as the common ground between two or more vocabularies assists our approach to promote newly created vocabularies by authors that have provided vocabularies in the past with a desired quality level. Definition 1 reflects the above described approach and allows to overcome the cold start issue of a newborn vocabulary in the LOV space by giving a score equal or higher than it would be assigned if the author metadata was not considered in the formulas.

Definition 1. If V_a represents the set of vocabularies that author a has a role in, A_v represents the set of authors of vocabulary v, $S_{v,i,k}$ refers to the score of vocabulary k for author i, and $V_{a,i}$ is set of vocabularies related to author i, then let S_v be the score of vocabulary v as defined by:

$$V_a = V_1, V_2, ...V_m, m \in N$$

$$A_v = a_1, a_2, ...a_n, n \in N$$

$$S_v = \frac{1}{|A_v|} \sum_{i=1}^{n} \frac{\sum_{k=1}^{m} S_{v,i,k}}{|V_{a,i}|}, n = |A_v|, m = |V_{a,i}|$$

In brief, the formula in Definition 1 calculates the average of the scores of the various vocabularies for each of the authors of vocabulary v and then it produces the score for v as the average of the sum of all the cumulative scores per author. Therefore, the new ranking metric for a vocabulary v has changed and follows the equation of Definition 2.

Definition 2. If B_v represents the number of vocabularies that are incoming for v, i.e. linking to it, $|LOV|$ is the total number of vocabularies, and S_v is the score based on the authors for v, then let $VSR_{LOV,V}$ be the score of vocabulary v as defined by:

$$VSR_{LOV,V} = \frac{B_v}{|LOV|} + S_v$$

To evaluate the effectiveness of the defined formula, we randomly selected vocabularies from the LOV dataset and compared the score S_v based on this formula with the score they would get without this dimension in the calculations. Table 1 depicts the comparison results, by providing the vocabulary URI at the leftmost column, the default score at the next column and the score taking into consideration the authors at the rightmost column.

As we can see from the data of Table 1, there are vocabularies that did not improve in the ranking score with the new aspect. For example the *og* vocabulary is still ranked very low as the authors of it do not appear in any other vocabulary. However, the score of *event*, *dcterms*, *foaf*, *schema*, *vcard* has improved and especially in the case of *event* we consider it to be a significant difference that could help the terms of it to appear higher in the ranking of a result set.

Vocabulary V	$VR_{LOV,V}$	$VSR_{LOV,V}$
dbpedia-owl:	0,021	0,021
dcterms:	0,806	0,948
event:	0,065	0,071
foaf:	0,599	0,781
gr:	0,071	0,071
og:	0,000	0,000
schema:	0,091	0,131
sioc:	0,038	0,038
skos:	0,371	0,371
vcard:	0,025	0,032

Table 1: LOV Vocabulary ranking examples of the old and the new ranking scores in comparison.

```
SELECT DISTINCT ?p ?vocab {
GRAPH <http://lov.okfn.org/dataset/lov>{
 ?vocab a voaf:Vocabulary.
 ?vocab2 a voaf:Vocabulary.
 {?vocab dc:contributor ?p}
 UNION
 {?vocab dc:creator ?p}.
 {?vocab2 dc:contributor ?p}
 UNION
 {?vocab2 dc:creator ?p}.
 FILTER(?vocab2 != ?vocab).
}} ORDER BY desc (?p)
```

Listing 1: SPARQL to retrieve the vocabularies of an author in the LOV graph.

Using the SPARQL endpoint of LOV[8] to run the query shown in Listing 1, we were able to extract the vocabularies that share creators or contributors. Studying those vocabularies we can find a lot of examples of vocabularies that would be ranked low due to the number of incoming links and the usage in datasets, which would be both zero.

4.2 Discovering Vocabulary Terms

In all the search processes, ranking plays a crucial role as it is the filter that removes entries from the final result set but also is responsible for bringing at the top positions the most relevant results. In this respect the discovery of vocabulary terms for a given set of keywords can be accomplished by our proposed approach by combining together the vocabulary ranking metrics with the vocabulary term ranking metrics. The ranking of terms follows the same approach as the vocabularies, but also takes in consideration the various figures from vocab.cc [16] and LODStats [2]. However, in the scope of this paper, we assume for simplicity that the ranking of terms is only affected by the vocabulary ranking and the relevance of a term to the given keyword. The relevance score is provider by the LOV search endpoint.

Apart from the vocabulary terms that are considered for inclusion due to the text content of the webpage, there is one

[8]http://lov.okfn.org/dataset/lov/sparql

HTML tag	Term t	Term t range
<img>	schema:image	schema:ImageObject
<video>	schema:video	schema:VideoObject
<audio>	schema:audio	schema:AudioObject

Table 2: Mappings between static elements and vocabulary terms used by the recommendation algorithm at the second recommendation stage. The *schema:* namespace stands for the URI http://schema.org/.

more category of terms equally important for the online visibility of the webpages and the completeness of an approach like the one proposed in this paper. We refer to this category as static recommendations, because of the mappings that are used. Thus, a predefined set of mappings is leveraged in order to provide suggestions for annotating parts of HTML markup related to the various multimedia types, e.g. images and videos. An overview of the terms and the mappings that we are considering for this version of LOVR are shown in Table 2. For the static elements, *schema.org* is preferred over other vocabularies as the terms can have as domain the most generic type of item, i.e. *schema:Thing*.

4.3 Describing the Generated Vocabulary

The generated set of terms is a new vocabulary that is provided to the user with all the needed metadata about the terms ranking and mappings to the webpage content. In this respect, we designed a new vocabulary, that is mostly a technical vocabulary that facilitates the output of the discovery results to the user. The vocabulary itself combines existing vocabularies and introduces a new namespace (i.e. vSearch) for the properties and classes that weave them together with the existing vocabulary properties and classes. The relationships among the vSearch properties and classes is depicted in Figure 4. The purpose of the vSearch vocabulary is to provide the appropriate properties for the description of a search query with the related results and the accompanied ranking. For the ranking properties, we reuse the vRank vocabulary[9] described in [12]. The vSearch vocabulary is hosted under http://vocab.sti2.at/vsearch.

At the core of the vocabulary, the main entity i.e. *Query*, could have one or more keywords (using the *hasResultTerm* object property) and one or more results (using the *hasRank* object property), which are instances of the *ResultTerm* type. Each result is connected with a *Rank* instance from the vRank vocabulary accompanied with the vrank:rankValue property. Additionally the *ResultTerm* is mapped with a 1:1 relationship with a URI that identifies the term, via the *termURI* datatype property.

Finally, it is important to mention that the vSearch vocabulary has been designed taking into consideration the need of having a vocabulary that can be used to describe a search activity. Thus, the vocabulary can be used beyond the scope of the proposed approach in order to describe a search together with the products of it and their ranking score.

5. EVALUATION

The plain keyword search discovery of vocabulary terms presented in the survey of Section 3 highlighted a few issues that the the proposed approach aims to address.

Firstly, the duration of the vocabulary terms selection process across all the use cases lasted for not less than one hour in average, while in a lot of the participants submissions we observed, as shown in Figure 2, very high values that go beyond the two hours. From this perspective, the proposed discovery of vocabulary terms proves to be much more efficient.

Furthermore, the participants mostly used the ranking of the terms as that is provided by the LOV search of terms, which ignores the usage of the authors as that was introduced in our approach in Section 4 and aims to improve the score of vocabularies with potential. Improving the score of new vocabularies relies on the connections that vocabularies share through their authors and contributors. Comparing the use cases to results computed with the approach we realised some differences in the selection of terms. For example, the recipe use case was completely dominated in the proposed approach by the schema.org terms, while in the manual discovery in the survey we observed more vocabularies to be proposed, but still the presence of schema.org was strong.

Running the SPARQL query of Listing 2 against the LOV repository, we were able to evaluate the effectiveness of the approach on new vocabularies.

```
SELECT DISTINCT ?date ?title ?vocab {
 GRAPH <http://lov.okfn.org/dataset/lov>{
 ?vocab a voaf:Vocabulary.
 ?vocab dcterms:title ?title.
 ?vocab dcterms:issued ?published.
 FILTER (STR(?published) > "2015")
 BIND (STR(?published) as ?date)
}} ORDER BY DESC(?date) ?title
```

Listing 2: SPARQL to retrieve vocabularies that were published in LOV after 2015.

As we can see in Table 3, the introduced approach cannot be considered a methodology that will completely eliminate the cold start problem in the ranking of a new vocabulary in the vocabulary space. However, as it is reflected by the score of *wfdesc*, we can expect a significant impact on the scoring of vocabularies that enclose some expertise in the field of ontology engineering. In the case of this specific vocabulary, the creators and contributors are five authors, with few of them being involved in very important vocabularies, like *skos*. The rest of the examples in Table 3 have scores down to zero due to the lack of any incoming links (reuse), while the new aspect of the authors experience does not help to improve the score, due to the fact that the contributors of them have not been involved in vocabularies with good scores. The same pattern is met in Table 1, where we observe a significant difference between the $VR_{LOV,V}$ and the $VSR_{LOV,V}$ for the *schema* vocabulary due to a contributor of it that is involved in very popular vocabularies like *foaf*. In this case, there is no doubt about the importance of the *schema* vocabulary, which is aligned to the increased probability of reuse that our metric would have foreseen in case we were not aware of the success of it.

Regarding the inclusion of static parts to the generated dataset we can easily observe the improved comprehensiveness of the proposed result terms in comparison to the pre-

[9]http://lov.okfn.org/dataset/lov/vocabs/vrank

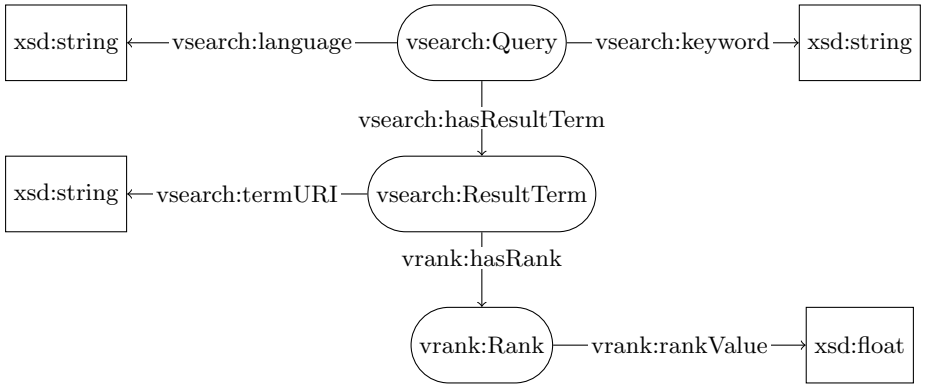

Figure 4: The vSearch vocabulary.

Vocabulary V	$VR_{LOV,V}$	$VSR_{LOV,V}$
kees:	0,000	0,000
jup:	0,000	0,000
cwork:	0,000	0,000
provoc:	0,000	0,000
wfdesc:	0,001	0,024

Table 3: LOV Vocabulary ranking examples of the old and the new ranking scores in comparison, for vocabularies that were introduced within and after 2015 (sample).

Term	Survey occurences
schema:Recipe	15%
schema:ingredients	15%
schema:totalTime	7%
schema:prepTime	7%
schema:recipeYield	7%

Table 4: Comparison of the proposed approach results with the survey participants' input.

vious version of the approach and to other approaches that ignore those information bits. It could be argued that the recommendation of the static parts is not as important as the rest of the generated vocabulary, however, we find it extremely useful for new vocabulary engineers or developers to be explicitly informed about the importance of annotating the multimedia content of the webpages. An indication of the impact of those terms in the LOVR approach is the results of the survey that we presented in Section 3, where none of the participants provided vocabulary terms and suggestions for annotating the multimedia content, which existed in all the use cases. Furthermore, underlying the importance of explicitly specifying the multimedia content on a webpage, we would like to stress out that the search engine providers not only crawl text from the Web sphere, but also images and videos and on top of the crawled data they provide the corresponding search functionality, which benefits and becomes more accurate when the multimedia content is explicitly annotated.

Finally, comparing the output result set of terms with each one of the individual collections compiled by the survey participants presented in Section 3, we realise a significant difference between the selected terms by the participants and our approach. Table 4 refers to the recipe use case and shows that only a few participants selected the same terms with the approach algorithm. Some of the participants chose a different one from the top three candidates, e.g. for the *Recipe* class, while some other terms were wrong (e.g. the schema:CookAction) as such information cannot derive from the webpage. From a qualitative perspective, the result of the approach gave an uncomplete result set with a recall of 71% but precision of 100%, by missing the cooking time and cooking method, which were provided by the participants combined all sets together. Therefore, it can be considered as a good starting point and assistant for the selection of vocabulary terms.

6. CONCLUSION

Summarising the contributions presented in this paper, they cover three main directions that support each other but can be separately exploited as well. In Section 3, we discussed the outcomes of a survey that we ran about the creation of semantic annotations, while it was later leveraged to support the evaluation of Section 5. Second contribution is the new dimension of the ranking algorithm presented in Section 4, which plays a significant role at the backbone of our LOVR approach. Finally the third contribution is related to the vocabulary introduced in the Section 4.3, which can be used to describe any search activity and the corresponding results together with the ranking of the latter.

Regarding the improvements of the already published approach, we underline two main parts, i.e. a) the inclusion of static elements in the recommendations, which enables the enrichment of the webpages with annotations that increase the visibility of the pages in more than one search types, e.g. image search; and b) the inclusion of the author in the ranking formulas in order to overcome the difficulty of assigning a score to vocabularies that are not used in datasets and are not reused by other vocabularies yet. The outcome of the presented research work shows evidence that it is feasible to rely on the vocabulary contributors background in the ranking of the vocabularies to improve the ranking performance of vocabularies that are new. Therefore, vocabularies created by authors, which are also contributors of vocabularies that were successful in terms of reuse and realisation in datasets, will receive a better ranking score in order to appear higher in the search term results. In this way we aim

to increase the probabilities of vocabularies being used in datasets and reused by new vocabularies instead of creating yet another ontology for the same topic.

In a recent publication about the evolution of the structured data on the Web [7], the authors highlight a few lessons learned since the beginning of Web-scale structured data exchange efforts. First and foremost they stretch the importance of making it easy for publishers and developers to produce structured data. Thus, providing them with examples/recipes of annotating data rather than lengthy specifications is also suggested as second lesson. Our approach in this publication aims at the same direction; facilitating the generation of semantic annotations based on existing vocabularies and the webpage content by providing to the vocabulary engineers and web developers with a set of candidate vocabulary terms that functions as a dynamic vocabulary covering the needs of the targeted webpage. Therefore, the user of the proposed tool saves all the time and effort needed to go through the existing vocabularies and the LOV search results in order to discover relevant terms and combine them together.

Finally, the outlook from the presented work includes the finalisation of the framework by incorporating the feedback that is being gathered from the usage of the implemented methodology as a Web service. The implemented framework is open sourced under the MIT License and the source code is available on GitHub[10], which allows future researchers to extend the implemented methodology, replace parts with other external libraries, or even extract the core ranking module and reuse it in other frameworks.

7. ACKNOWLEDGMENTS

This work has been partially supported by the Internet Foundation Austria in the scope of Netidee 2015 and the EU projects BYTE, ENTROPY, EUTravel, the OeAD project LDCT, as well as the FFG project TourPack.

8. REFERENCES

[1] G. A. Atemezing and R. Troncy. Information content based ranking metric for Linked Open Vocabularies. In *Proceedings of the 10th International Conference on Semantic Systems*, pages 53–56. ACM, 2014.

[2] S. Auer, J. Demter, M. Martin, and J. Lehmann. Lodstats–an extensible framework for high-performance dataset analytics. In *Knowledge Engineering and Knowledge Management*, pages 353–362. Springer, 2012.

[3] C. Bizer, K. Eckert, R. Meusel, H. Mühleisen, M. Schuhmacher, and J. Völker. Deployment of RDFa, microdata, and microformats on the Web – a quantitative analysis. In *The Semantic Web–ISWC 2013*, pages 17–32. Springer, 2013.

[4] A. S. Butt. Ontology search: Finding the right ontologies on the Web. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 487–491. International World Wide Web Conferences Steering Committee, 2015.

[5] A. S. Butt, A. Haller, and L. Xie. Relationship-based top-k concept retrieval for ontology search. In *Knowledge Engineering and Knowledge Management*, pages 485–502. Springer, 2014.

[10] https://github.com/istavrak/vocab-recommender

[6] M. B. Ellefi, Z. Bellahsene, S. Dietze, and K. Todorov. Beyond established knowledge graphs-recommending Web datasets for data linking. In *European Conference on Web Engineering*, pages 262–279. Springer, 2016.

[7] R. Guha, D. Brickley, and S. MacBeth. Schema.org: Evolution of structured data on the Web. *Queue*, 13(9):10, 2015.

[8] T. Heath and C. Bizer. Linked Data: Evolving the Web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136, 2011.

[9] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the pedantic Web. 2010.

[10] R. Meusel, C. Bizer, and H. Paulheim. A Web-scale study of the adoption and evolution of the schema.org vocabulary over time. In *Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics*, page 15. ACM, 2015.

[11] R. Meusel, P. Petrovski, and C. Bizer. The WebDataCommons Microdata, RDFa and microformat dataset series. In *The Semantic Web–ISWC 2014*, pages 277–292. Springer, 2014.

[12] A. Roa-Valverde, A. Thalhammer, I. Toma, and M.-A. Sicilia. Towards a formal model for sharing and reusing ranking computations. In *Proc. of the 6th Intl. Workshop on Ranking in Databases In conjunction with VLDB*, volume 2012, 2012.

[13] J. Schaible, T. Gottron, S. Scheglmann, and A. Scherp. Lover: support for modeling data using linked open vocabularies. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, pages 89–92. ACM, 2013.

[14] J. Schaible, T. Gottron, and A. Scherp. Survey on common strategies of vocabulary reuse in linked open data modeling. In *The Semantic Web: Trends and Challenges*, pages 457–472. Springer, 2014.

[15] J. Schaible, T. Gottron, and A. Scherp. TermPicker: Enabling the reuse of vocabulary terms by exploiting data from the Linked Open Data cloud. In *European Semantic Web Conference*, pages 101–117. Springer, 2016.

[16] S. Stadtmüller, A. Harth, and M. Grobelnik. Accessing information about linked data vocabularies with vocab.cc. In *Semantic Web and Web Science*, pages 391–396. Springer, 2013.

[17] I. Stavrakantonakis, A. Fensel, and D. Fensel. Linked Open Vocabulary recommendation based on ranking and Linked Open Data. In *Proceedings of the 5th Joint International Semantic Technology Conference*. Springer, 2015.

[18] I. Stavrakantonakis, I. Toma, A. Fensel, and D. Fensel. Hotel websites, Web 2.0, Web 3.0 and online direct marketing: The case of Austria. In *Information and Communication Technologies in Tourism 2014*, pages 665–677. Springer International Publishing, 2014.

[19] P.-Y. Vandenbussche, G. A. Atemezing, M. Poveda-Villalón, and B. Vatant. Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the web. *Semantic Web*, (Preprint):1–16, 2015.

[20] P.-Y. Vandenbussche and B. Vatant. Linked Open Vocabularies. *ERCIM news*, 96:21–22, 2014.