Detecting Good Practices and Pitfalls when Publishing Vocabularies on the Web

María Poveda-Villalón[1], Bernard Vatant[2], Mari Carmen Suárez-Figueroa[1], Asunción Gómez-Pérez[1]

[1]Ontology Engineering Group. Universidad Politécnica de Madrid. Spain.
[2]Mondeca, Paris, France.

mpoveda@fi.upm.es, bernard.vatant@mondeca.com, {mcsuarez, asun}@fi.upm.es

Abstract. The uptake of Linked Data (LD) has promoted the proliferation of datasets and their associated ontologies bringing their semantic to the data being published. These ontologies should be evaluated at different stages, both during their development and their publication. As important as correctly modelling the intended part of the world to be captured in an ontology, is publishing, sharing and facilitating the (re)use of the obtained model. In this paper, 11 evaluation characteristics, with respect to publish, share and facilitate the reuse, are proposed. In particular, 6 good practices and 5 pitfalls are presented, together with their associated detection methods. In addition, a grid-based rating system is generated showing the results of analysing the vocabularies gathered in LOV repository. Both contributions, the set of evaluation characteristics and the grid system, could be useful for ontologists in order to reuse existing LD vocabularies or to check the one being built.

Keywords: ontology, vocabulary, linked data, ontology publication, ontology evaluation, pitfalls, good practices

1 Introduction

Vocabularies or Ontologies[1] bring their semantics to Linked Data (LD)[2] [3], by formally defining shared sets of classes and properties, using semantic standards such as RDFS or OWL. When a vocabulary element is used in a RDF dataset through its URI, nothing more is generally declared in this dataset about this element, and that is a good practice since datasets have not to re-define URIs already defined in external vocabularies. In order to understand the meaning of such an URI, both humans and applications should be able to de-reference it, discover the context in which it has

[1] At this moment, *there is no clear division between what is referred to as "vocabularies" and "ontologies"* (http://www.w3.org/standards/semanticweb/ontology). For this reason, we will use both terms indistinctly in this paper.

[2] http://www.w3.org/standards/semanticweb/ontology

been formally defined. This context is typically a RDFS or OWL file and the matching HTML documentation. Both files are, in the best of cases, available from the ontology URI through proper content negotiation implementation over HTTP protocol. Both human-readable and machine-consumable information should provide, not only the semantics of the elements defined in its namespace, but also a reasonable amount of metadata about the vocabulary (dates, creator, publisher, versions, etc.).

The Linked Open Vocabularies project (LOV)[3] is intended to gather and describe those vocabularies used or potentially usable by LD and to provide indicators of their relevancy. Each vocabulary in LOV is described by metadata gathered either from its formal publication, or from the vocabulary documentation or communication with the publishers, or from the vocabulary content itself. Two years after its launch, LOV has been widely acknowledged and embraced by the LD community.

A fundamental feature for the scope of our research is that each entry in LOV is uniquely identified by a vocabulary URI, and is generally associated with a unique namespace URI. Given the variety of interpretations and so many different implementation practices we have discovered in LOV, either OWL standards have underspecified the definition and relationship between those two URIs, or the specification has been largely either ignored or misunderstood. The simplest configuration is to have these two URIs being the same. But many other configurations are possible and are indeed observed. Moreover, content negotiation on the namespace does not necessarily leads to vocabulary URI.

An application dedicated to consume vocabularies is not likely to be prepared to such a variety of configurations. It is likely to identify the vocabulary by either or both the namespace URI or the vocabulary URI. Vocabulary publishing practices can be classified as "good" or "bad" insofar as they ease or impede such applications.

In this paper, we have conducted a detailed analysis of more than 350 vocabularies gathered in the LOV registry. Our aim is to automatize the detection of good practices and common pitfalls when publishing vocabularies in order to ease the work of applications willing to access and consume LOV vocabularies with no more initial information than the vocabulary URI, its namespace and the prefix assigned in LOV. The results of such scan is: (1) a non exhaustive list of best practices and common pitfalls about publishing LD vocabularies, (2) specific methods for detecting such good practices and common pitfalls, (3) some metadata about ontology quality (regarding the appearance or lack of good practices and pitfalls) that could be added to the vocabulary metadata stored in LOV, and (4) the inclusion of pitfalls in services such as OOPS![4] to help eager vocabulary managers to check the quality of their vocabularies prior to their publication.

The structure of the paper is the following. Section 2 introduces and describes the framework with the evaluation characteristics to be used in the evaluation of LOV vocabularies. Section 3 presents the detection methods implemented for checking the characteristics presented in Section 2. The results of executing such detection methods over 355 vocabularies registered in LOV and an analysis of the obtained results are shown in Section 4. Finally, Section 5 exposes related research efforts and Section 6 presents some concluding remarks and future lines of work.

[3] http://lov.okfn.org

[4] http://www.oeg-upm.net/oops

2 Good practices and pitfalls for publishing vocabularies

Main guidelines for publishing data over the web are the extremely well-known Linked Data principles and the Linked Open Data 5 Star rating system defined by Tim Bernes-Lee[5]. More precisely, the rating system defines the following levels (taken literally from the source):

- ***LOD1.*** *Available on the web (whatever format) but with an open licence, to be Open Data*
- ***LOD2.*** *Available as machine-readable structured data (e.g. excel instead of image scan of a table)*
- ***LOD3.*** *As (2) plus non-proprietary format (e.g. CSV instead of excel)*
- ***LOD4.*** *All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff*
- ***LOD5.*** *All the above plus Link your data to other people's data to provide context*

More specific recommendations about publishing ontologies on the web have been proposed inspired by the above-mentioned 5-star linked data scale. We will refer to it along this paper as the "Linked data vocabulary 5-start rating system"[6] that defines the following recommendations (taken literally from the source):

- ***LDV1.*** *Publish your vocabulary on the Web at a stable URI*
- ***LDV2.*** *Provide human-readable documentation and basic metadata such as creator, publisher, date of creation, last modification, version number*
- ***LDV3.*** *Provide labels and descriptions, if possible in several languages, to make your vocabulary usable in multiple linguistic scopes*
- ***LDV4.*** *Make your vocabulary available via its namespace URI, both as a formal file and human-readable documentation, using content negotiation*
- ***LDV5.*** *Link to other vocabularies by re-using elements rather than re-inventing*

Along the rest of the paper we will refer to the points stated in these two rating systems as LOD or LDV plus its ordinal numeration according to the lists above. We will use some of these points or recommendations to support the good practices and pitfalls proposed in this paper. We will also point to the 10 rules [1] for designing persistent URI, since some points are also applicable.

In the following, we describe the 11 characteristics we have identified when publishing ontologies on the Web. It should be noted that in the remaining the term "characteristics" will be used for referring to the set of both good practices and pitfalls. That is, there are 11 characteristics described here, 6 of them represent good practices and 5 of them represent pitfalls. Each characteristic has an identifier, a description and one example of an ontology holding that characteristic. The identifiers are on the form of GPX for good practices where the X is a numerical identifier, in this case starting in 1. For pitfalls, the identifiers are on the form of PY where Y is a numerical identifier. In this case, as the pitfalls here defined will be included in OOPS! catalogue[7], the numeration follows to the one given in the catalogue to avoid confusion and help the reader to find each pitfall both along this paper and within the

[5] http://www.w3.org/DesignIssues/LinkedData.html
[6] http://bvatant.blogspot.fr/2012/02/is-your-linked-data-vocabulary-5-star_9588.html
[7] http://www.oeg-upm.net/oops/catalogue.jsp

catalogue by the same identifier**.** For the examples, we refer to the vocabularies registered in LOV.

2.1 Good practices proposal

The following six characteristics represent our proposal of good practices in ontologies regarding publishing issues and metadata in an online ontology.

GP1. Provide RDF description: In order to make an ontology more reusable one should publish it on an stable URI (LDV1) providing machine-readable formats using open standards from W3C to identify things (LOD4).

- **Example:** the "Configuration ontology (cold)" ontology with URI http://purl.org/configurationontology provides a turtle serialization when looking up its URI.

GP2. Provide HTML documentation: It is important to provide human-readable documentation (LDV2) so that third parties (data publishers, ontology developers, etc.) can understand the ontology more easily, boosting, therefore, its use (e.g. describing data from) and reuse (e.g. within another ontology).

- **Example:** "Accommodation Ontology Language Reference (acco)", which URI is http://purl.org/acco/ns, provides HTML documentation by redirecting to http://ontologies.sti-innsbruck.at/acco/ns.html.

GP3. Content negotiation for RDF: According to (a) LDV4, (b) the best recipes for publishing vocabularies[8] "*It is accepted as a principle of good practice that HTTP clients SHOULD include an 'Accept:' field in a request header, explicitly specifying those content types that may be handled*." and (c) the rule "*Implement 303 redirects for real-world objects*" proposed in [1], it is a good practice to provide RDF description of the vocabulary using content negotiation mechanisms to retrieve it when the Accept header indicates this format.

- **Example: "**ACM Classification Ontology (acm)" with URI http://www.rkbexplorer.com/ontologies/acm provides correct content negotiation mechanism when asking for RDF content.

GP4. Content negotiation for HTML: According to (a) LDV4, (b) the best recipes for publishing vocabularies "*It is accepted as a principle of good practice that HTTP clients SHOULD include an 'Accept:' field in a request header, explicitly specifying those content types that may be handled*." and (c) the rule "*Implement 303 redirects for real-world objects*" proposed in [1], it is a good practice to provide HTML description of the vocabulary using content negotiation mechanisms to retrieve it when the Accept header indicates this format.

- **Example:** "Agent Relationship Ontology (agrelon)" with URI http://d-nb.info/standards/elementset/agrelon.owl# implements correct content negotiation mechanism when requesting (X)HTML.

GP5. Provide *vann* metadata: As an ontology URI does not necessarily corresponds to the namespace where the ontology elements are defined it is a good practice to indicate by means of metadata the namespace used for defining

[8] http://www.w3.org/TR/swbp-vocab-pub/

them. In this sense, we also consider a good practice to indicate a preferred prefix used when referring to the given ontology. This good practice is related to LDV2 as it is related with the metadata provided within the ontology.

- **Example:** "The Lingvoj Ontology (lingvo)" with URI http://www.lingvoj.org/ontology it a good example of providing vann metadata to indicate the preferred namespace and prefix for the ontology.

GP6. Well-established prefix: Even though it is no crucial, it would be desirable that a prefix used for a given vocabulary is well-established and there is consensus about it across applications. For example, in the case of "foaf" there is no doubt to which vocabulary is this prefix referring to.

- **Example: "**Algorithms Ontology (algo)" with URI http://securitytoolbox.appspot.com/securityAlgorithms has a consistent prefix across systems, in this case, LOV and prefix.cc[9].

2.2 Pitfalls proposed

The following five characteristics represent our proposal for pitfalls in ontologies regarding publishing issues and metadata. These five characteristics represent undesirable situations to be found in an online ontology, or in other words, a publisher team would not like to see these characteristics in its ontologies.

P36. URI contains file extension: Guidelines in [1] suggest avoiding file extension in persistent URIs, particularly those related to the technology used, as for example ".php" or ".py". In our case we have adapted it to the ontology web languages used to formalized ontologies and their serializations. In this regard, we consider as pitfall including file extensions as ".owl", ".rdf", ".ttl", ".n3" and ".rdfxml" in an ontology URI.

- **Example:** "BioPAX Level 3 ontology (biopax)" ontology's URI (http://www.biopax.org/release/biopax-level3.owl) contains the extension ".owl" related to the technology used.

P37. Ontology not available: This bad practice is about not meeting LOD1 from Linked Data star system that stars "On the web" and LDV1 that says "Publish your vocabulary on the Web at a stable URI".

- **Example: "**Ontology Security (ontosec)" which URI is http://www.semanticweb.org/ontologies/2008/11/OntologySecurity.owl is not available online as RDF nor as HTML[10].

P38. No OWL ontology declaration: The *owl:Ontology* tag aims at gathering metadata about a given ontology as version information, creation date, etc. It is also used to declare the inclusion of other ontologies. Not declaring this tag is consider as a bad practice for owl ontologies as it is a symptom of not providing useful metadata as proposed in LDV2.

- **Example:** "Creative Commons Rights Expression Language (cc)" ontology with URI http://creativecommons.org/ns does not have any

[9] http://prefix.cc/

[10] By the time of carrying out this study at 19th of June of 2013.

owl:Ontology declaration in its RDF file even though there are other OWL elements used as, for example, *owl:equivalentProperty*.

P39. Ambiguous namespace: In the case of not having defined the ontology URI nor the *xml:base* namespace, the ontology namespace is matched to the file location. This situation is not desirable as the location of a file might change while the ontology should remain stable as proposed in LDV1.

- **Example:** "Basic Access Control ontology (acl)" with URI http://www.w3.org/ns/auth/acl has no *owl:Ontology* tag nor *xml:base* definition.

P40. Namespace hijacking: This bad practice refers to the situation when an ontology is reusing or referring to terms from other namespaces that are not defined in such namespace. This is an undesirable situation as no information could be retrieve when looking up those undefined terms, in addition, there would be no meaning or semantic behind them. In addition this practice is against Linked Data publishing guidelines provided in [3] "*Only define new terms in a namespace that you control*."

- **Example:** the **"**WSMO-Lite Ontology (wl)" which URI is http://www.wsmo.org/ns/wsmo-lite#, uses http://www.w3.org/2000/01/rdf-schema#Property" that is not defined in the rdf namespace (http://www.w3.org/2000/01/rdf-schema#) instead of using http://www.w3.org/1999/02/22-rdf-syntax-ns#Property**,** that is actually defined in the rdfs namespace (http://www.w3.org/1999/02/22-rdf-syntax-ns#).

2.3 Dependencies between good practices and pitfalls

It is obvious that some good practices and pitfalls appearance is conditional upon the appearance of another one. In this sense, some characteristics block the potential appearance of others, for example, if it not possible to retrieve the RDF description of an ontology it cannot be checked whether it has *vann* metadata defined in it. These connections are shown in Figure 1 by means of the relation "X depends on Y".

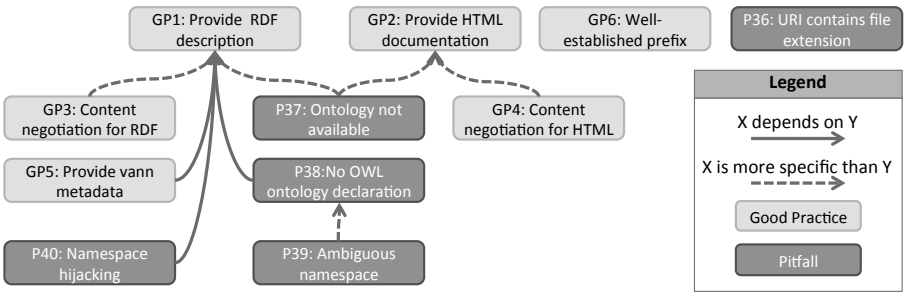


Figure 1. Dependencies between good practices and pitfalls.

Another dependency between characteristics is the case of a good practice or a pitfall being more specific than other. For example, providing HTML documentation

implementing correct content negotiation mechanisms is more specific than just serving HTML documentation. These associations are shown in Figure 1 by means of the relation "X is more specific than Y". For these cases we need the most general characteristic to be true in order to check a more specific one. The opposite is also possible, for example, for "P37. Not available" to be possible "GP1. Provide RDF description" and "GP2. Provide HTML description" have to be false. This information is important from the publisher point of view as it indicates which detections could be affected when correcting another issue.

3 Description of the methods used to identify good practices and pitfalls in ontologies

In this section, the detection methods used within this study for each good practice (Section 3.1) and pitfall (Section 3.2) are detailed. These methods have been coded and applied over the 355 vocabularies registered in LOV at the moment of carrying out this study. The results and analysis of such execution are shown in Section 4.

3.1 Good practices detection methods

Detection method for GP1. Provide RDF description: To check whether the ontology, given its URI it, can be loaded and processed by means of an RDF API, in our case we use JENA[11].

Detection method for GP2. Provide HTML documentation: To check whether, given an ontology URI, an HTML document is retrieved when requesting HTML in the accept header. This is checked by means of looking for HTML tags in the retrieved content. We do not use any HTML parser as they add the tag needed to make a valid HTML page from sources that do not really follow this syntax.

Detection method for GP3. Content negotiation for RDF: To check whether, given an ontology URI, it provides an rdf/xml serialization when asking for RDF in the accept header and it implements the redirections mechanism: 303-200. We use Vapour[12] for checking this point and adapted its behaviour for purl ontologies considering also the sequence 302-303-200.

Detection method for GP4. Content negotiation for HTML: To check whether, given an ontology URI, it provides an HTML document when asking for HTML in the accept header and it implements the redirections mechanism: 303-200. We use Vapour for checking this point and adapted its behaviour for purl ontologies considering also the sequence 302-303-200.

Detection method for GP5. Provide *vann* metadata: To check whether there is at least one result for the following SPARQL query executed over the ontology model loaded in JENA:

[11] http://jena.apache.org/

[12] http://validator.linkeddata.org/vapour

```
SELECT ?prefPrefix ?prefNS WHERE{
       OPTIONAL {?s vann:preferredNamespacePrefix ? prefPrefix.}
       OPTIONAL {?s vann:preferredNamespaceUri ?prefNS.}}
```

Detection method for GP6. Well-established prefix: To check that the prefix defined in LOV for a given ontology matches with the one defined in prefix.cc. The detection method first, checks if given the ontology namespace we obtain from prefix.cc the same prefix as declared in LOV. If no prefix is retrieved, the service is used the other way around, the namespace recorded in prefix.cc for the prefix given in LOV is requested. If the two prefixes (the one from LOV and the one obtained, if any, from prefix.cc) are equal we say that the ontology meets this characteristic, otherwise it does not.

3.2 Pitfalls detection methods

Detection method for P36. URI contains file extension: To check whether the ontology URI contains the string ".owl" or ".rdf" or ".n3" or ".ttl".

Detection method for P37. Ontology not available: To check whether neither GP1 nor GP2 hold, that is, if they both are false.

Detection method for P38. No OWL ontology declaration: To check whether there is an "owl:Ontology" tag defined in the ontology or not. It is worth mentioning that this check is done over the raw text containing the RDF code and applying the following seven regular expressions:

```
<owl:Ontology rdf:about="
<rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Ontology"/>
a(\\s+)owl:Ontology(\\s*);
rdf:type(\\s+)owl:Ontology(\\s*);
a(\\s+)owl:Ontology(\\s*),
rdf:type(\\s+)owl:Ontology(\\s*),
<owl:Ontology>
```

Detection method for P39. Ambiguous namespace: To check whether the RDF code of a given ontology matches at least one of the following cases:

a. There is no "owl:Ontology" tag declaration nor "xml:base" defined.
b. There is no "owl:Ontology" tag declaration and the "xml:base" is empty.
c. The "rdf:about" in the "owl:Ontology" tag declaration is empty and there is no "xml:base" defined.
d. The rdf:about in the "owl:Ontology" tag and the "xml:base" are empty.

Detection method for P40. Namespace hijacking: For detecting this pitfall we rely on Triple Checker[13]. It should be noted that we only consider as error the case of an ontology using undefined terms in a namespace even though Triple Checker also warns about other issues. For example, analysing "Appearances Ontology Specification"[14] we consider as P40 the case of the term

[13] http://graphite.ecs.soton.ac.uk/checker/

[14] A copy of the result given by Triple Checker for the "Appearances Ontology Specification" at 11th of June of 2013 is available at http://goo.gl/MD9FDo

"http://swrc.ontoware.org/ontology#date-added" however other warnings are not, for example, the one given for the term "http://rdf.muninn-project.org/ontologies/muninn#wikipedia_version".

4 Results and Analysis over LOV vocabularies

In this section, results and statistics for the LOV ecosystem status at 19th of June of 2013 are shown. Main motivations for choosing vocabularies in LOV as vocabulary registry for carrying out this study are the facts that (a) the ecosystem is updated and manually curated, (b) contains a reasonable number of vocabularies registered (more than 350) and (c) provides complete information and trustable values for the data needed (in our case we need: namespace, URI and prefix) for each vocabulary.

Figure 2 shows for each characteristic (good practice or pitfall) how many times it has been detected within the 355 analyzed vocabularies. For example, the first column shows that "GP1. Provide RDF description" appears in 308 ontologies (marked as 'Good Practice detected') and it does not appear in 47 (marked as 'Good Practice not detected'). For the pitfalls, in the seventh column we observe that "P36. URI contains file extension" appears in 39 ontologies (marked as 'Pitfall detected'), while it does not appear in 316 (marked as 'Pitfall not detected').

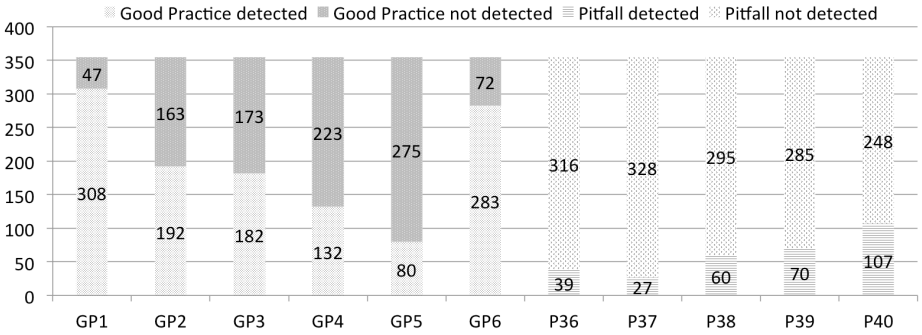
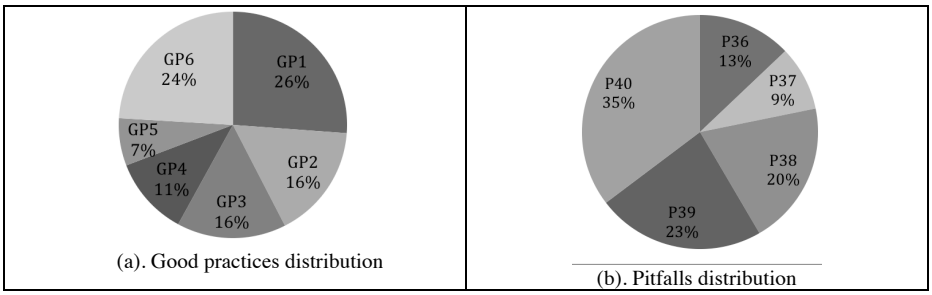


Figure 2. Good practices and pitfalls frequency.

The information shown in Figure 3 represents the distribution of good practices (a) and pitfalls (b) among the total number of appearance. That is, looking at the pie chart in the right, the slice for P40 means that among all the pitfalls appearances over the 355 ontologies (a total of 303: the sum of all the values for 'Pitfall detected' in Figure 2), 35% it has been a case of "P40. Namespace hijacking".

From Figure 2 and Figure 3 we see that most of the good practices are present in more than half of the ontologies analysed, being the most popular "GP1. Provide RDF description" and "GP6. Well-established prefix". Even though GP1 is the good practice appearing most it is still alarming that in more than 40 ontologies they could not have been processed programmatically. This is clearly a problem as it impedes the the ontology (re)usability and, in case some data is annotated with such an ontology, it semantics could not be retrieved, turning it into meaningless data. The high appearance of GP6 might be surprising as it is quite specific and requires and extra effort

from ontology managers. This high frequency is due to the efforts from LOV curators editing prefix.cc content to keep as many prefixes as possible equal in both systems. Regarding the pitfalls, we can observe that they are scarcely present apart from "P40. Namespace hijacking" that have a high frequency. Which is also alarming as defining terms in namespaces out of our control would lead to de-referenceability and lack of semantics issues, indeed it clearly goes against main guidelines for publishing LD [3].



(a). Good practices distribution

(b). Pitfalls distribution

Figure 3. Good practices and pitfalls appearance distribution.

Figure 4 shows the number of ontologies that have a given number of good practices and pitfalls. For example, the bubble in the top row and third column starting from the left means that there are 32 ontologies having 2 good practices and 0 pitfalls. In this grid we see that most of the ontologies have none, one or two pitfalls while most of the ontologies have between 2 and 5 good practices. Even though the general landscape is not bad, there is still work to do in order to achieve the ideal situation where all the vocabularies are placed in the right corner at the top, that is, having the maximum number of good practices implemented and none pitfalls. It should be noted that the information shown in Figure 4 has been condensed and that a detailed grid showing the name of the ontologies is available at http://goo.gl/zu9ZbW.

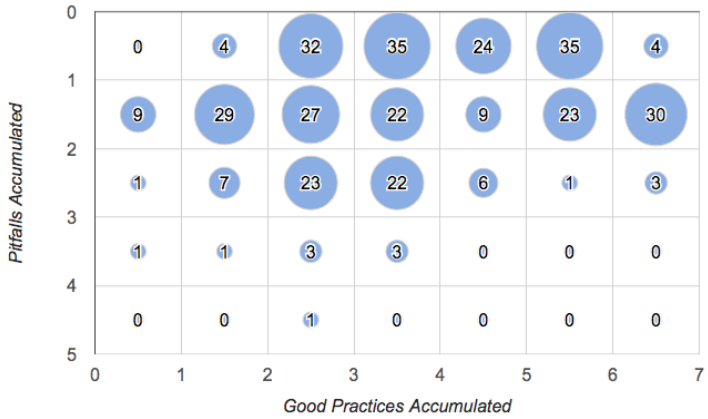


Figure 4. Number of vocabularies by good practices and pitfalls accumulated grid.

5 Related work

Ontology evaluation is a key process that should be performed at different stages of the ontology development and deployment. As important as correctly modelling the intended part of the world to be captured in an ontology, is publishing the model following good practices and avoiding bad practices.

However, apart from the aforementioned publishing recommendations (See Section 2), to the best of our knowledge, most of the evaluation approaches are focused on the ontology content quality or syntax checking and there is not too much research on approaches for validating the ontology publication process.

Regarding ontology content quality evaluation, and not directly related to LD features, it is important to mention plug-ins for desktop applications as XDTools plug-in[15] for NeOn Toolkit and OntoCheck plug-in[16] for Protégé; the wiki-based ontology editor MoKi [5] that incorporates ontology evaluation functionalities; and the online tool OOPS! [6] that detects potential pitfalls in ontologies.

In addition, validation services for RDF and LD have also been developed. One of the most popular tools is the W3C RDF Validation Service[17] that checks the syntax of RDF documents. In this regard, RDF:Alerts[18] also checks for syntax errors, undefined terms, among others. Regarding protocol issues, the online tools Vapour[19] [2] and Hyperthing[20] aim at validating the compliance of a resource according to LD publication rules. These tools check the de-referenceability of a given URI.

We can also mention evaluation works with respect to SKOS vocabularies where several tools have been proposed. Those that check characteristics related to LD, are qSKOS [4] that checks missing in and out links, broken links, undefined SKOS resources and HTTP URI scheme violation; and PoolParty[21] that also checks URI correctness.

6 Conclusions and future work

Along this paper 6 good practices and 5 pitfalls have been proposed and described. Detection methods for each of them have also been suggested and implemented[22]. With this contribution, ontology evaluation tools and quality features catalogues could be extended. In addition, an evaluation of the good practices and pitfalls detection has been carried out over 355 vocabularies registered in LOV.

A grid-based rating system has also been proposed. In this grid[23] the vocabularies are positioned according to the total number of good practices and pitfalls appearing.

---

[15] http://neon-toolkit.org/wiki/2.3.1/XDTools
[16] http://protegewiki.stanford.edu/wiki/OntoCheck
[17] http://www.w3.org/RDF/Validator
[18] http://swse.deri.org/RDFAlerts
[19] http://validator.linkeddata.org/vapour
[20] http://www.hyperthing.org/
[21] http://demo.semantic-web.at:8080/SkosServices/check
[22] Complete execution results are provided at http://goo.gl/zu9ZbW
[23] It refers to the detailed grid available at http://goo.gl/zu9ZbW instead of the one in section 4.

This grid could be used by (a) LOV curators in order to identify which vocabularies need to be reviewed and (b) vocabulary authors and publishers in order to detect possible improvements by meeting more good practices and avoiding pitfalls.

First conclusion we can draw is that vocabularies in LOV seem to be well maintained and likely to be high quality. It could be due to the fact that the LOV ecosystem is reviewed and conflictive vocabularies authors are contacted when a problem is encountered and, in the worst case, the vocabularies are deleted from the ecosystem. In this way, LOV administrators keep a high standard for the vocabularies registered. That is, it is a goodness of a semi-handcrafted registry against crawlers gathering vocabularies and ontologies over the web with little or no review and maintenance.

Second, it is worth mentioning that some practices that one would not expect to find in a stable and well-established ontology are surprisingly quite present within the analysed ontologies, e.g. making the RDF code of the ontology available online or not using terms from other namespaces that are not actually defined in such namespace.

Third, it is worth mentioning that it is difficult to define the division line between good practices and pitfalls as in some cases the absence of a good practice (e.g. "GP1. Provide RDF description") could be taken as a pitfall and the other way round. However, it does not hold for all of the good practices and pitfalls defined in this work. For example, the lack of some pitfalls (e.g. "P40. Namespace hijacking") does not really represent a good practice or a high quality point for the ontology.

Future lines of work will include to deal with the detection of (a) metadata about licences in order to check LDV1; (b) other kind of metadata apart from *vann* annotation, for example, creators, authors, dates, languages, etc. as proposed in LDV2; (c) linguistic information in order to check LDV3 and (d) reused terms within the analysed ontology in order to check LDV5. In addition different importance levels could be attached to each good practice or pitfall, as it is obvious that, for example, an ontology containing the file extension in its URI is not as critical as a case of namespace hijacking. This information would be useful to assess and rank ontologies weighting the evaluation results for the good practices and pitfalls observed.

As complement to this work, we propose, as future work, to provide guidelines to solve the problems when a good practice in not implemented or a pitfall is detected.

Finally, we propose to execute described methods over an ontology registry as LOV in regular basis in order to observe the evolution of the quality of the ecosystem as a whole and for each particular vocabularies in particular and draw trends and patterns when publishing vocabularies.

Acknowledgments. This work has been partially supported by the Spanish project *BabelData* (TIN2010-17550), the mobility and internationalization program by the *Consejo Social* of the *Universidad Politécnica de Madrid* and the French project Datalift (ANR-10-CORD-009).

References

1. Archer, P., Goedertier, S., and Loutas, N. *D7.1.3 – Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs and the EC*. Deliverable. December 17, 2012.

2. Berrueta, D., Fernández, S., and Frade, I. *Cooking HTTP content negotiation with Vapour*. ESWC2008 workshop on Scripting for the Semantic Web (SFSW2008), Tenerife, Spain. June 2, 2008.
3. Heath, T., Bizer, C.: *Linked data: Evolving the Web into a global data space* (1st edition). Morgan & Claypool (2011)
4. Mader, C., Haslhofer, B., & Isaac, A. *Finding quality issues in SKOS vocabularies*. In Theory and Practice of Digital Libraries. Springer Berlin Heidelberg. 2012.
5. Pammer, V. *PhD Thesis: Automatic Support for Ontology Evaluation Review of Entailed Statements and Assertional Effects for OWL Ontologies*. Engineering Sciences. Graz University of Technology.
6. Poveda-Villalón, M., Suárez-Figueroa, M.C., Gómez-Pérez, A. *Validating ontologies with OOPS!*. 18th International Conference on Knowledge Engineering and Knowledge Management. (EKAW2012) Galway, Ireland, 8 - 12 October 20