A Survey of Identifiers and Labels in OWL Ontologies

Nor Azlinayati Abdul Manaf, Sean Bechhofer, and Robert Stevens

School of Computer Science
The University of Manchester, UK
norazlinayati.abdulmanaf@postgrad.manchester.ac.uk,
{sean.bechhofer, robert.stevens}@manchester.ac.uk

Abstract. We present a survey of the usage and style of identifiers and labels of named entities in a corpus of OWL ontologies. We investigated the frequency of use of both labels and meaningful or meaningless identifiers in those ontologies. We also surveyed common practices of lexical encoding styles for identifiers. We found that most ontologies do not use labels for named entities. When they do use labels, those labels are mostly meaningful and most ontologies also used meaningful identifiers. CamelCase style appears to be the most widely used style of lexical encoding for identifiers. We observed, however, that the majority of the ontologies use a mixture of two or more lexical encoding styles. The result of this survey is useful when considering strategies, for example, natural language generation from ontologies or converting artefacts, such as OWL ontologies, into languages like the Simple Knowledge Representation System (SKOS), where the notion of label is important. Given that labels are optional in OWL ontologies, what is the best way to handle the label selection when converting them into SKOS? Merging multiple entities may require selection from labels or identifiers assigned to these entities for `skos:prefLabel` and `skos:altLabel`.
Keywords: survey, identifiers, labels.

1 Introduction

In this paper we present a survey of how identifiers and labels are used within Web Ontology Language (OWL)[1] ontologies. We are interested in transforming such ontologies in to other forms—such as natural language and in to other Semantic Web representations such as the Simple Knowledge Organisation System (SKOS)[2]. In these transformations it is important to be able to deal with both identifiers and labels in OWL ontologies. In natural language generation, for example, a human understandable form of the entity needs to be available to place within a natural language setting [1]. In SKOS, a concept has an alternate and preferred label—from where do these labels arise (identifier or label) and how is a choice made between preferred and alternate labels? [2–4]. As OWL does not

[1] `http://www.w3.org/2004/OWL/`
[2] `http://www.w3.org/2004/02/skos/`

mandate use of labels, but an Internationalized Resource Identifier (IRI) alone can be used to 'identify' an entity for both machine and human, how does any transformation programme deal with labels and identifiers? [5]. What labelling and identifier situations should a designer of such a system expect to encounter? This survey was thus motivated by a need to understand the degree and style of use of labels and identifiers in OWL ontologies. Once this is known, strategies to deal with various situations can be made with an understanding of the cost and benefit of realising those strategies.

Ontologies are used to capture knowledge about some domain of interest. An ontology describes the concepts in the domain and also the relationships that hold among these concepts. These concepts can be represented by classes or individuals, and the relationships are represented using properties. In OWL, the concepts and properties can be referred to as entities. A named entity refers to a named class, a named individual or a named property. Each named entity must have a unique identifier, called an IRI. An IRI refers to an object that can act as a reference to something that has identity.

Identifiers are not only used by computers, but also by humans. Humans prefer using meaningful identifiers—the name encapsulates the nature of the entity that it names. An identifier is meaningful if there is a direct relationship between the natural language term used and the characteristics about the entity being identified. For example, if the entity is used to represent a concept "dog", then using `dog` as an identifier for this concept helps to make the identifier meaningful (to an English speaker).

It is also possible, however, to have a meaningless identifier, or also called a "semantic-free" identifier. For example, `ABC_20020` is a semantic-free identifier. An identifier is meaningless or semantic free if there is no direct relationship between the natural language term used and the characteristics about the entity being identified. In OWL it is possible to separate the IRI for the entity and the label for that entity (usually provided through `rdfs:label`). When identifiers are meaningless in human terms, the entity needs a label that is a natural language term for that entity. This can have several desirable effects, including: ability for having different language renderings; being able to change the label without having to change the identity of the entity (which is useful when the ontology is being used to encode data); and so on.

As identifiers in OWL can contain no spaces, meaningful identifiers that would normally contain spaces have to be encoded in a way that excludes spaces, but retains the meaningful nature of the identifier for human readers. An identifier can be encoded in various lexical encoding styles. Using internal upper case letters within an identifier to denote word boundaries–camel case style (eg. PetOwner or petOwner); underscore style (_) (eg. pet_owner); and hyphen style (-), are among the styles that are used in meaningful IRIs. This means that a meaningful identifier needs some processing to have it in a conventional form for human reading—that is, containing spaces between words.

We used a survey to determine the current use of labels and identifiers in ontologies including the naming convention of identifiers. The goal was to allow us to answer the following questions:

1. Given that labels are optional, what is the frequency of label use in an ontology?
2. Given that a label should be meaningful, and if labels are used in an ontology, what is the frequency of meaningful labels used?
3. Given that an entity could have multiple labels, what is the frequency of an entity having multiple labels?
4. What is the frequency of meaningful and meaningless identifiers used in an ontology?
5. What is the frequency of the following combinations between identifiers and labels used in an ontology?
   (a) an entity with a meaningful identifier and has meaningful label(s)?
   (b) an entity with a meaningful identifier and has meaningless label(s)?
   (c) an entity with a meaningful identifier and has no label?
   (d) an entity with a meaningless identifier and has meaningful label(s)?
   (e) an entity with a meaningless identifier and has meaningless label(s)?
   (f) an entity with a meaningless identifier and has no label?
6. What is the frequency of camel-case, underscore style and hyphen styles encoding used in an ontology to encode identifiers?

2 Materials and Methods

An overview of the methods used to answer these questions is:

1. Corpus preparation;
2. Isolation of identifiers and labels;
3. Determination of whether the identifiers and labels are meaningful or meaningless;
4. Result recording;
5. Data analysis.

2.1 Corpus Preparation

In this survey we used ontologies in the TONES repository[3]. We also searched on Google using `filetype:owl` for more OWL ontologies to be added to our corpus. From the search result, we looked through each ontology for OWL constructs such as `owl:Class, owl:Individual, owl:ObjectProperty` or `owl:DataProperty` to be considered as a "valid" OWL ontology for this survey. All collected ontologies from both sources were compared to eliminate duplication. We utilized the OWL API[4] for loading and managing the OWL ontologies. All ontologies were locally stored for future reference.

[3] `http://owl.cs.manchester.ac.uk/repository/`
[4] `http://owlapi.sourceforge.net/`

2.2 Isolation of identifiers and labels

For each ontology, we isolated the identifiers and labels for each named entity. The identifier was extracted from the IRI for each named entity. If the IRI contained a fragment identifier, then the identifier for this entity is the fragment identifier (the fragment after the (#) character). For example, for the named class `http://owl.cs.manchester.ac.uk/2010/people#person`, we extracted `person` as the identifier. Otherwise, we took the last portion of the path component as an identifier (the fragment after the last (/) character[5]. For example, for the named class `http://owl.cs.manchester.ac.uk/2010/pizza/pizzaTopping`, we extracted `pizzaTopping` as the identifier.

Entity labels were identified through the annotation property `rdfs:label` in the ontology. An entity is considered to have a label if there exist one or more `rdfs:label` associated with the entity. We also considered labels made through sub-properties of `rdfs:label`.

2.3 Determination of whether the identifiers and labels are meaningful or meaningless

Our aim is to test a label or identifier against the Web to see if it is meaningful. Gaining many pages or 'hits' to a query based on a label or identifier would suggest that it is meaningful—based on the assumption that use of the string on Web pages suggests its use in natural language. This is a two-stage process.

Normalise the lexical encoding style of identifiers Prior to this test, however, an identifier must be put in to a form suitable for querying as identifiers are formed with no spaces. As described in Section 1, identifiers are normally encoded in various lexical encoding styles. In order to determine the meaningfulness of an identifier, the human brain will apply some cognitive manipulation on the encoded identifier into a form (space-separated form) that could be more readily interpreted. We called this process 'normalisation'. To check the meaningfulness of an identifier, we normalised the string used by transforming it in to a space separated form. In order to do the transformation, we first needed to identify the style of lexical encoding used to encode the identifiers. We have identified the following commonly used lexical encoding styles, and limit our categorisation to these, placing any identifiers not using these styles in an "Other style" category.

1. CamelCaseStyle.
2. Underscore_style.
3. Hyphen-style.
4. HybridCamelCase_underscore_style.
5. HybridCamelCase-hyphen-style.
6. Hybrid-hyphen_underscore_style.
7. Single word

[5] `http://www.ietf.org/rfc/rfc2396.txt`

8. Other style—any identifiers that are encoded using other than the styles mentioned above are grouped under this category.

All single word identifiers are grouped in the "single word" category. This category can be considered as a "wild card" as it is compatible with all other categories. Therefore, we used to following rules to decide the lexical encoding style used in the ontologies.

1. If all identifiers in an ontology are encoded using single word, then classify the ontology as having only single word identifiers.
2. If some identifiers are encoded using the single word style and only one other lexical encoding style is used for the rest of the identifiers in an ontology, the single word category can be made compatible with the one lexical encoding style, and the ontology is classify to have that one lexical encoding style. For example, the rest of the identifiers in an ontology were encoded with camelCase style, then, the single word category can be made compatible with camelCase style and the ontology can be classified as encoded using camelCase style.
3. If some identifiers are encoded using single words and more than one lexical encoding style is used for the rest of the identifiers in an ontology, then the single word category cannot be made compatible with any of these styles and the ontology is classified as having a mixture of lexical encoding style.

Once the lexical encoding style has been identified, we then normalise the identifier into a space-separated form to be used in the meaningfulness checking.

Check for meaningfulness For our meaningfulness check we used a Web search query using the Bing API [6]. For each label and normalised identifier, we sent a Web search query to the World Wide Web (WWW) to search for the number of websites with the words in the labels and normalised identifiers. We are interested in the number of results returned from the search query. There are three options for sending strings to the Web search query. First, using the quotation (" ") around the search string. For example, string `hello world` is search as `"hello world"`. Since this query searched for exactly the same occurrence of string in the Web it returned limited number of search results due to the reason that not all words in an identifier occurred together in natural language presented in the Web. Second, search for the string without the quotation. For example, using string as `hello world` as the search string. This type of query searched for the words in an identifier that occur anywhere in a Web page, and not necessarily in the same order. The result return by this type of query is moderate and acceptable. Third, if a string consists of multiple words, search for the words as separate string with and/or without quotation [6]. For example, searching for `hello world` as separate query `hello` and `world`. Since this query searched for the string separately, the result returned could be two different number and further processing is needed to determine which one should

[6] `http://www.bing.com/developers`

be chosen. For this survey, we chose the second option to search the terms together without quotation. Additionally, we used the order of words as how it appears in the identifiers and labels. For example, if the normalised identifier is "hello world", then we use a string `hello world` with the same order for the Web search query. We set a threshold value of 100 hits which is used to determine the meaningfulness of the searched term. A hit result that is below 100 is considered not meaningful. The choice of 100 as the threshold is a heuristic based on running a few ontologies from various domains and simply judging a reasonable threshold. We found that ontologies with medical terminologies, get fewer hits for meaningful identifiers.

2.4 Result recording

We recorded the results of this survey at various stages. All extracted identifiers and labels were recorded in XML format for future reference and analysis. We also recorded the hit results for each of the identifiers and labels from the Web search query in CSV format for future reference and analysis.

2.5 Data analysis

Based on the recorded results, we calculated and recorded for each ontology and for each named entity, the following:

1. frequency of labels used;
2. frequency of meaningful and meaningless labels;
3. frequency of identifiers with one label and more than one labels;
4. frequency of meaningful and meaningless identifiers;
5. frequency of the combination of identifiers and labels;
6. frequency of lexical encoding styles.

For each entity type in each ontology, we also calculated the proportion of these frequency with respect to its total entity for each criterion listed above in the form of a percentage. Finally, we calculated the mode, mean and median of these percentages for each of the criteria.

3 Results

We used 219 valid ontologies from the TONES repository[7], after discarding any URIs that no longer existed or were too big to be loaded in our machine. There were 354 hits returned from the search query[8] after all the duplicate results were omitted. After looking at each URLs, only 264 URLs represented valid OWL ontologies, the rest were URLs linked to pages that no longer existed. We also compared the list of URLs with the ontologies from the TONES repository

[7] As at 22 February 2010
[8] As at 30 March 2010

to avoid duplication. Out of the remaining 241 ontologies, we have randomly selected 87 ontologies to be added to our corpus—making a total of 306 ontologies.

Out of 306 ontologies, 5 ontologies contained none of the named entities—leaving 301 ontologies in the corpus. There are 296 ontologies containing named classes; 105 ontologies with named individual; 264 ontologies with named object properties and 138 ontologies with named data properties. The rest of the analyses were performed on ontologies that contained named entities.

Table 1 shows the result summary with the number of ontologies for each criteria surveyed[9]. The mean shown represents the mean of the proportion of the measured criteria.

Type	Classes		Individuals		Object Properties		Data Properties	
	Count	Mean	Count	Mean	Count	Mean	Count	Mean
Total ontologies	**296**		**105**		**264**		**138**	
Ontologies with Labels	122	32.9%	32	20%	82	27.8%	23	14.4%
Meaningful labels	122	89.4%	32	94.4%	82	95.6%	23	91.7%
Meaningless labels	70	6.9%	12	5.6%	13	4.4%	6	8.3%
Ontologies with single label	121	93.9%	31	88.6%	81	97.5%	22	94.5%
Ontologies with multi labels	21	6.1%	8	11.5%	4	2.5%	2	5.5%
Meaningful identifiers	286	85.2%	103	90.6%	263	97.8%	137	97.2%
Meaningless identifiers	135	14.9%	49	9.4%	37	2.3%	23	3.5%
Meaningful identifiers								
with meaningful labels	107	18.8%	31	16.3%	81	25.2%	21	11.9%
with meaningless labels	46	1.4%	9	1.3%	11	1.2%	3	0.7%
with no label	242	65%	90	70%	200	71.3%	125	83.9%
Meaningless identifiers								
with meaningful labels	66	10.1%	9	3.2%	5	1.1%	3	1.5%
with meaningless labels	53	2.6%	6	0.2%	7	0.1%	5	0.4%
with no label	76	2.1%	42	7.1%	28	1.1%	16	1.7%
Lexical Encoding Style of Identifiers								
CamelCase style	116	64.9%	35	47.4%	133	52.5%	97	62.4%
Underscore_style	0	1%	4	7.3%	63	24.4%	4	5.9%
Hyphen-style	0	0.8%	0	0.6%	5	2.1%	1	0.9%
CamelCase_underscore style	40	27%	5	23.4%	35	4.9%	4	7%
CamelCase-hyphen style	1	1.6%	0	0.9%	0	0.4%	1	1.8%
Hyphen-underscore_style	0	0.1%	1	1%	0	0.1%	0	0%
Single word	4	4.5%	52	20.4%	0	14.9%	7	22.2%
Others	0	0.1%	0	0.01%	0	0%	0	0%
Mixture	135		52		58		24	

Table 1. Number of ontologies for different criteria surveyed. (The mean was calculated over the total ontologies for each entity type)

[9] The complete analysis of the result for this survey is made available at `http://www.myexperiment.org/packs/110`

4 Discussion

Table 1 provides basic answers to the questions raised in Section 1 in numerical terms. Here we present here some observations based on an initial analysis of those results along with closer examination of some of the ontologies.

First, we appreciate that the technique used to determine meaningfulness of both labels and identifiers – using a search with a fixed cutoff threshold – is rather basic. The threshold value was selected based on some preliminary experiments, but it is likely that the use of a single static threshold value is not appropriate for all domains – see the discussion below. However, for the purpose of this survey, the technique is enough to show some interesting results. We are currently extending and exploring possible mechansisms for the selection of variable threshold values based on the content of each ontology rather than having a single static cut-off value for all ontologies.

Labels are not widely used in all named entity types. However, when labels are used in an ontology, those labels are usually meaningful. In terms of the number of labels per entity, we observed that, for all named entity types, almost all ontologies contained single labels. Where an ontology does contain more than one label per entity, closer investigation revealed that the multiple labels were used to represent labels in different languages. Single labels usually represent labels in one language only.

Almost all of the ontologies used meaningful identifiers for named entities, with object property and data property entities showing the highest use of meaningful identifiers. Further analysis, shows that those identifiers for object and data properties that are classified as meaningless are actually meaningful, but the meaningfulness test gave a hit below the threshold (as discussed above). As for meaningless identifiers, even though the result shows that quite a number of ontologies used meaningless identifiers, their percentage of usage (in terms of the proportion of entities in the ontologies) is quite small.

For all named entity types, most of the ontologies contained meaningful identifiers with no label. This observation supports our findings that labels are not widely used in the ontologies and most ontologies do have meaningful identifiers. Interestingly, we observed that there are also a few ontologies that use meaningless identifiers with meaningless labels or no labels. However, their mean percentage of use is rather small. We suspect again that our approach in identifying the meaningfulness of terms is a factor in this abnormality. Having a meaningless identifier and no label makes little sense; it is reasonable to suspect that specialised language will appear meaningless in the face of the simple threshold approach used.

As for lexical encoding styles, the result show that camel case style is the most used lexical encoding style for all named entity type. There are also a significant number of ontologies that use a mixture of lexical encoding styles. A small number of ontologies used unidentified lexical encoding style under the others category. Further analysis of this category showed the identifiers classified into this category used other punctuation symbols such as dot (.) to encode the identifiers. Some example of identifiers in this category are as follows:

– E1.CRM_Entity (combination of dot (.) and underscore)
– E71.Man-Made_Thing (combination of dot (.), hyphen and underscore)
– erbB-2_Genes (combination of camel case style (erbB), hyphen and underscore)

The small numbers obtained for the "others" category suggests that the categories identified in Section 2.3 are indeed sufficient to characterize the bulk of ontologies in the corpus.

5 Related Work

There are several surveys that analyse Semantic Web documents especially OWL ontologies to help understanding of the nature of OWL ontologies. Bechhofer and Volz [7] surveyed a sample of 227 OWL ontologies to answer the question of "how much OWL DL is there on the Web?" and found that is "not much". A majority of them are OWL Full, which in many cases were caused by syntactic errors such as missing type triples. However, they presented a patching technique for these errors and increase this "a little bit". In [8], Wang et al. extended the work in [7] to a much larger samples size. They were interested in evaluating those ontologies to determine trends in modeling practices, OWL construct usages and OWL species utilization. They surveyed a sample of 1 300 ontological documents, not only OWL ontologies, but also RDFS documents. The survey reported in our paper adds to these surveys and takes a finer grained look at identifiers and labels within ontologies. The information gained is important, as discussed in the introduction, for deciding upon strategies for handling the 'names of entities' within software where some human orientated presentation is required.

6 Conclusion

We found that most ontologies do not use labels for named entities. When they do use labels, these labels are mostly meaningful. Only a few ontologies have more than one label per named entity. We also found that most of the ontologies do use meaningful identifiers and if they do use meaningless identifiers, these identifiers only occupied a small portion of the ontologies. Most ontologies that have meaningful identifiers do not have labels. Interestingly, there are also a few ontologies that used meaningless identifiers with meaningless labels or no label; though this may well be an artefact of our test for meaningfulness.

Camel case style appears to be the most widely used lexical encoding style for identifiers. However, most ontologies are inconsistent in their identifier encoding style, as more than one style is used to encode the identifiers within an ontology.

We hope to extend this survey on a larger corpus of ontologies. For example, collecting for more ontologies from various other sources like the Swoogle [10] and Watson [11]. It also might be interesting if we could extend the survey to not only

[10] `http://swoogle.umbc.edu/`
[11] `http://kmi-web05.open.ac.uk/WatsonWUI/`

investigate the use of labels and identifiers, but also other OWL constructs such as property restrictions, to have a better understanding of the common practice of use of these constructs in the existing OWL ontologies.

We can raise further questions about the effect of the domain for which an ontology was built on its style of identifier and label use. For example, the Open Biomedical Ontologies consortium [9] have a policy of semantic free identifiers and use of labels. In addition, the question of whether the ontology is one that is 'in service' with a community—that is, it is actually being used to do a job of work—rather than being one developed for research purposes makes a difference to identifier and lable use would be a useful one to answer.

When transforming OWL ontologies into other forms – such as natural language or to other Semantic Web representations such as SKOS, an understanding of the use of labels and identifiers within the ontologies is beneficial. If nothing else, it allows developers to make judgements about situations for which strategies should be developed.

Acknowledgements: This work was funded in part by the SWAT project EP/G032459/1. The authors would like to thank Majlis Amanah Rakyat (MARA), an agency under the Malaysian Government, for funding the student. Many thanks to the reviewers who gave insightful comments and suggestion to improve this paper.

References

1. Smart, P.R.: Controlled natural languages and the semantic web. Technical report ITA/P12/SemWebCNL, School of Electronics and Computer Science, University of Southampton (2008)
2. Jupp, S., Stevens, R., Bechhofer, S., Kostkova, P., Yesilada, Y.: Document navigation: Ontology or knowledge organisation system? In: Network Tools and Applications in Biology (NETTAB'2007) - A Semantic Web for Bioinformatics: Goals, Tools, Systems, Applications. (2007)
3. Jupp, S., Stevens, R., Bechhofer, S., Yesilada, Y., Kostkova, P.: Knowledge representation for web navigation. In: Semantic Web Applications and Tools for the Life Sciences (SWAT4LS 2008) Workshop. (2008)
4. Abdul Manaf, N.A., Bechhofer, S., Stevens, R.: Exploring the relationships between OWL and SKOS. ISWC 2009 Doctoral Consortium (2009)
5. Cimino, J.J.: Desiderata for controlled medical vocabularies in the twenty-first century. In: Methods of Information in Medicine. (1998) 394–403
6. Alani, H., Brewster, C.: Ontology ranking based on the analysis of concept structures. In: Proceedings of the Third International Conference on Knowledge Capture(K-CAP 05), Banff, Canada, ACM (2005)
7. Bechhofer, S., Volz, R.: Patching syntax in owl ontologies. In: Proceedings of the 3rd International International Semantic Web Conference. (2004)
8. Wang, T.D., Parsia, B., Hendler, J.: A survey of the web ontology landscape. In: In Proc. of the International Semantic Web Conference, ISWC. (2006)
9. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Eilbeck, L.J.G.K.: The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology **25 (11)** (2007) 1251–1255