The Current State of SKOS Vocabularies on the Web

Nor Azlinayati Abdul Manaf[1,2], Sean Bechhofer[1], and Robert Stevens[1]

[1] School of Computer Science, The University of Manchester, United Kingdom
{abdulman,seanb,robert.stevens}@cs.man.ac.uk
[2] MIMOS Berhad, Technology Park Malaysia, 57000 Kuala Lumpur, Malaysia

Abstract. We present a survey of the current state of Simple Knowledge Organization System (SKOS) vocabularies on the Web. Candidate vocabularies were gathered through collections and web crawling, with 478 identified as complying to a given definition of a SKOS vocabulary. Analyses were then conducted that included investigation of the use of SKOS constructs; the use of SKOS semantic relations and lexical labels; and the structure of vocabularies in terms of the hierarchical and associative relations, branching factors and the depth of the vocabularies. Even though SKOS concepts are considered to be the core of SKOS vocabularies, our findings were that not all SKOS vocabularies published explicitly declared SKOS concepts in the vocabularies. Almost one-third of the SKOS vocabularies collected fall into the category of *term lists*, with no use of any SKOS semantic relations. As concept labelling is core to SKOS vocabularies, a surprising find is that not all SKOS vocabularies use SKOS lexical labels, whether `skos:prefLabel` or `skos:altLabel`, for their concepts. The branching factors and maximum depth of the vocabularies have no direct relationship to the size of the vocabularies. We also observed some common modelling slips found in SKOS vocabularies. The survey is useful when considering, for example, converting artefacts such as OWL ontologies into SKOS, where a definition of typicality of SKOS vocabularies could be used to guide the conversion. Moreover, the survey results can serve to provide a better understanding of the modelling styles of the SKOS vocabularies published on the Web, especially when considering the creation of applications that utilize these vocabularies.

1 Introduction

We present a survey of Simple Knowledge Organization System (SKOS) vocabularies on the Web. The aim of this survey is to understand what a typical SKOS vocabulary looks like, especially in terms of the shape, size and depth of the vocabulary structure. We are also interested in determining the extent of usage of SKOS constructs. The results of this survey will equip us with a better understanding of the modelling styles used in the SKOS vocabularies published on the Web. This may be important when considering the creation of an application that utilizes these vocabularies, for example, when converting artefacts such as

E. Simperl et al. (Eds.): ESWC 2012, LNCS 7295, pp. 270–284, 2012.

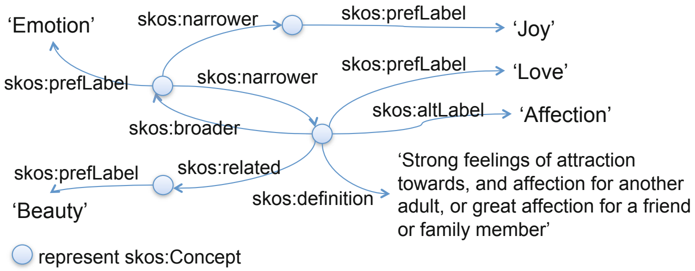


Fig. 1. An example of SKOS constructs usage

OWL ontologies into SKOS, where some typicality of SKOS vocabularies may be useful to guide the conversion.

SKOS[1], accepted as a W3C Recommendation in August 2009, is one of a number of Semantic Web knowledge representation languages. Other such languages include the Resource Description Framework (RDF)[2], the RDF Schema (RDFS)[3], and the Web Ontology Language (OWL)[4]. SKOS is a language designed to represent traditional knowledge organization systems whose representation has weak semantics that are used for simple retrieval and navigation. Such representation includes thesauri, subject headings, classification schemes, taxonomies, glossaries and other structured controlled vocabularies.

The basic element in SKOS is the *concept* which can be viewed as a 'unit of thought'; ideas, meanings or objects, that is subjective and independent of the term used to label them [1]. These concepts can be semantically linked through hierarchical and associative relations.

Figure 1 shows an example of the usage of SKOS constructs. There are four `skos:Concept` in the figure, representing the concept of *Emotion, Love, Joy* and *Beauty*. The `skos:broader` and `skos:narrower` properties are used to show that the concepts are hierarchically arranged, while the `skos:related` property is used to show the associative relations between the concepts. SKOS provides three properties for associating lexical labels to conceptual resources; `skos:prefLabel`, `skos:altLabel` and `skos:hiddenLabel`. SKOS documentation properties such as `skos:definition` are used to provide additional textual information regarding the concept.

One of SKOS' aims is to provide a bridge between different communities of practice within the Library and Information Sciences and the Semantic Web communities. This is accomplished by transferring existing models of knowledge organization systems to the Semantic Web technology context [1]. Knowledge organization system (KOS) is a general term, referring to the tools that present the organized interpretation of knowledge structures [2]. This includes a variety

[1] http://www.w3.org/2004/02/skos/
[2] http://www.w3.org/RDF/
[3] http://www.w3.org/2001/sw/wiki/RDFS
[4] http://www.w3.org/2001/sw/wiki/OWL

of schemes that are used to organize, manage and retrieve information. There are several types of KOS. Hodge [3] groups them into three general categories:

Term lists (flat vocabularies): emphasize lists of terms, often with definitions; e.g., authority files, glossaries, dictionaries, gazetteers, code lists;

Classifications and categories (multi-level vocabularies): emphasize the creation of subject sets; e.g., taxonomies, subject headings, classification schemes; and

Relationship lists (relational vocabularies): emphasize the connections between terms and concepts; e.g., thesauri, semantic networks, ontologies.

We wish to find how many SKOS vocabularies are publicly available for use on the Web and how many fall into one of the listed categories. Additionally, we are interested in learning what the SKOS vocabularies look like in terms of size, shape and depth of the vocabulary structure. We are interested in understanding which of the SKOS constructs listed in the SKOS Reference document [1] are actually being used in the SKOS vocabularies and how often these constructs are used.

Related Work. While research has attempted to characterize Semantic Web documents such as OWL ontologies and RDF(S) documents on the Web [4,5,6], to the authors knowledge there is no attempt at characterizing SKOS vocabularies. Our approach is similar to the work produced by Wang et. al. [4], which focused on OWL ontologies and RDFS documents.

2 Materials and Methods

The steps carried out in this survey are:

1. Preparing a candidate SKOS vocabulary corpus.
2. Identifying SKOS vocabularies.
3. Collecting survey data.
4. Filtering out multiple copies of the same SKOS vocabularies.
5. Analysing the corpus of vocabularies.

Apparatus. All experiments were performed on a 2.4GHz Intel Core 2 Duo MacBook running Mac OS X 10.5.8 with a maximum of 3 GB of memory allocated to the Java virtual machine. Two reasoners were used: JFaCT, which is a Java version of the FaCT++ reasoner, and Pellet. We used the OWL API version 3.2.4[5] for handling and manipulating the vocabularies.

Preparing a candidate SKOS vocabulary corpus. We employed several methods to gather the candidate SKOS vocabularies to be included in our corpus. First, we gathered vocabularies from two dedicated collections, which are the SKOS

[5] `http://owlapi.sourceforge.net/`

Implementation Report[6] and the SKOS/Datasets[7]. We chose the collections as a primary source because the vocabularies listed in these collections were compiled by the SKOS Working Group through its community call. We manually downloaded the vocabularies listed in these collections and stored them locally.

In the second method we utilised Semantic Web search engines such as Swoogle[8] and Watson[9]. Collecting vocabularies from these sources may enable us to gain some insights into the use of SKOS vocabularies in the community. We made use of the API provided by both search engines to programmatically gather the results from the relevant search. For both search engines, we used `thesaurus`, `skos` and `concept` as search terms. At this point, we collected the URIs of the vocabularies as our analysis tools will retrieve the documents from the Web given the URIs. We also used the Google search engine, using search keywords "skos" and relevant filetypes, which are ".skos", ".owl", ".rdf", ".ttl", ".nt", ".n3" and ".xml". We considered these terms as keywords in identifying SKOS vocabularies because some of them are the terms used as construct names in the SKOS data model.

Our third method used a Web crawler. We used an off-the-shelf web crawler called Blue Crab web crawler[10], which could be configured to crawl based on user specific settings.

Identifying SKOS vocabularies. For the purposes of this survey, we used the following definition of SKOS vocabulary. A SKOS vocabulary is a vocabulary that at the very least contains SKOS concept(s) used directly, or SKOS constructs that indirectly infer the use of a SKOS concept, such as use of SKOS semantic relations.

Each candidate SKOS vocabulary was screened in the following way to identify it as a SKOS vocabulary:

1. Check for existence of direct instances of type `skos:Concept`; if Yes, then accept the vocabulary as a SKOS vocabulary.
2. Check for existence of implied instances of `skos:Concept` due to domain and range restrictions on SKOS relationships (for example the subject of a `skos:broader`, `skos:narrower` or `skos:related` relationship is necessarily a `skos:Concept`); if Yes, then accept the vocabulary as a SKOS vocabulary.
3. Otherwise, do not accept this vocabulary as a SKOS vocabulary.

Consider the following vocabulary snippets written in Manchester Syntax[11]. `Vocabulary 1` and `Vocabulary 2` are accepted as SKOS vocabularies based on tests in Step 1 and Step 2, respectively. Meanwhile, `Vocabulary 3` is not

[6] `http://www.w3.org/2006/07/SWD/SKOS/reference/20090315/implementation.html`

[7] `http://www.w3.org/2001/sw/wiki/SKOS/Datasets`

[8] `http://swoogle.umbc.edu/`

[9] `http://kmi-web05.open.ac.uk/WatsonWUI/`

[10] `http://www.limit-point.com/products/bluecrab/`

[11] `http://www.w3.org/TR/owl2-manchester-syntax/`

accepted as a SKOS vocabulary according to our definition, even though this vocabulary uses SKOS constructs such as `skos:prefLabel` and `skos:altLabel`.

```
Vocabulary 1:          Vocabulary 2:         Vocabulary 3:
Individual: Emotion    Individual: Love      Individual: Love
  Types:                 Types:                Types:
    Concept                Thing                 Thing
Individual: Love       Facts:                Facts:
  Types:                 broader Emotion       prefLabel "Love",
    Concept                                    altLabel "Affection"
Individual: Beauty     Individual: Emotion
  Types:                 Types:
    Concept                Thing
```

Collecting survey data. Since we are interested in both the asserted and inferred version of the SKOS vocabularies, we performed the data recording in two stages; with and without invocation of an automatic reasoner such as Pellet or JFaCT. The data collected without invocation of an automatic reasoner may suggest the actual usage of SKOS constructs in those vocabularies. As for the rest of the analysis, such as identifying the shape and structure of the vocabularies, we need to collect the data from the inferred version of the vocabularies, hence the use of an automatic reasoner.

In the first stage of data collection, no automatic reasoner is invoked, since we are interested in the asserted version of the SKOS vocabularies. For each candidate SKOS vocabulary, we count and record the number of instances for all SKOS constructs listed in the SKOS Reference [1]. We also record the IRI of all *Concept Scheme* present in each SKOS vocabulary.

In the second stage of data collection, we applied a reasoner, and collected and recorded the following for each SKOS vocabulary:

1. Number of SKOS concepts.
2. Depth of each SKOS concept and maximum depth of the concept hierarchy.
3. Total number of links for `skos:broader`, `skos:narrower` and `skos:related` properties.
4. Total number of loose singleton concepts (concepts that are not connected to any other concepts).
5. Total number of root concepts (concepts with only `skos:narrower` relation, but no `skos:broader` relation).
6. Maximum number of `skos:broader` property.

Filtering out multiple copies of the same SKOS vocabularies. We use the recorded information in the previous stage to filter structurally identical SKOS vocabularies. We compare the *Concept Scheme* IRI to search for duplicate vocabularies. For two or more vocabularies having the same *Concept Scheme* IRI, we compare the record for each SKOS construct count. We make a pairwise comparison between each SKOS construct count, taking two vocabularies at a time.

1/ If the two vocabularies have identical records, we then check the content of these vocabularies. This is done by making a pairwise comparison between the instances of `skos:Concept` in one vocabulary to the other. If the two vocabularies have the same instances of `skos:Concept`, then one copy of these vocabularies is kept and the duplicate vocabulary is removed. Otherwise, follow the next step.

2/ If the two vocabularies do not have identical records or identical instances of `skos:Concept`, we assume that one vocabulary is a newer version of the other. We check, if the two vocabularies belong to the same category (either Thesaurus, Taxonomy, or Glossary), then we keep the latest version of the vocabulary and remove the older version. Otherwise, both vocabularies are kept.

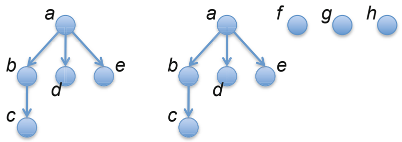
Analysing the survey results. In analysing the collected data, we found that some of the vocabularies that are known to be SKOS vocabularies, fail in Step 2 (to be identified as a SKOS vocabulary). We manually inspected these vocabularies. We found several patterns of irregularity in the vocabulary representation and considered them as *modelling slips* made by ontology engineers when authoring the vocabularies. For each type of modelling slip, we decide whether the *error* is intentional or unintentional, and if *fixing* the *error* would change the content of the vocabulary. If the *error* is unintentional and *fixing* the error does not change the content of the vocabulary, then we can apply *fixing* procedures to correct the modelling slips. All fixed vocabularies were included in the survey for further analysis.

We calculated the mode, mean, median and standard deviation for the occurrence of each construct collected from the previous process. The analysis focused on two major aspects of the vocabularies; the usage of SKOS constructs and the structure of the vocabularies. In terms of the usage of SKOS constructs, we analysed which constructs were most used in the SKOS vocabularies. As for the structural analysis of the SKOS vocabulary, we introduced a SKOS metric, $\mathcal{M}$, with eight tuples as follows:

$$\mathcal{M} =< \mathcal{S}, \mathcal{D}, \mathcal{L}, \mathcal{R}, \mathcal{MAX_B}, \mathcal{F_H}, \mathcal{B_H}, \mathcal{F_A} > \qquad (1)$$

where $\mathcal{S}$ is the size of vocabulary (represented by the number of SKOS *concepts*), $\mathcal{D}$ is the maximum depth of vocabulary structure, $\mathcal{L}$ is the number of *loose singleton concepts* in the vocabulary, $\mathcal{R}$ is the number of *root nodes* of the vocabulary structure, $\mathcal{MAX_B}$ is the maximum `skos:broader` relation for each *concept* in the vocabulary, $\mathcal{F_H}$ is the average hierarchical forward branching factor, $\mathcal{B_H}$ is the average hierarchical backward branching factor and $\mathcal{F_A}$ is the average associative forward branching factor.

According to [7], *branching factor* is the measure of the number of links coming in to or going out from a particular node. For a directed graph, there are two types of *branching factor*, namely ***forward branching factor (FBF)*** and ***backward branching factor (BBF)***. The *FBF* is the number of arcs or links going out from a node. The *BBF* is the number of arcs or links coming into a node.



(a) Example 1 (b) Example 2

Fig. 2. Example graphs for determining the structure of vocabulary

The *FBF* for hierarchical relations is calculated based on the number of `skos:narrower` relations of a particular concept. Whereas the *BBF* for hierarchical relations is calculated based on the number of `skos:broader` relations of a particular concept. As for associative relations, both *FBF* and *BBF* are calculated based on the number of `skos:related` relations of a particular concept. Since the `skos:related` relation is symmetric, both *FBF* and *BBF* for the associative relation of a particular vocabulary is the same.

We suspect that the *branching factor* values for each concept in a particular SKOS vocabulary are non-uniform. Therefore, we calculated the *average FBF* and *average BBF* for both hierarchical and associative relations. Note that we ignored the *loose singleton concepts* when calculating the *average branching factors*. The *average hierarchical FBF,* F_H, *average hierarchical BBF,* B_H and *average associative FBF,* F_A are given by the following equations:

$$\mathcal{F}_\mathcal{H} = \frac{n}{T_n}, \quad \mathcal{B}_\mathcal{H} = \frac{b}{T_b}, \quad \mathcal{F}_\mathcal{A} = \frac{r}{T_r} \qquad (2)$$

where n, b, and r are the total number of `skos:narrower`, `skos:broader` and `skos:related` relations, respectively, and T_n, T_n and T_n are the total number of concepts with `skos:narrower`, `skos:broader` and `skos:related` relations, respectively.

Figure 2 shows two graphs illustrating two different structures of example SKOS vocabulary. Each circle represents a SKOS concept and each directed link between two circles represents a `skos:narrower` relation. The SKOS metric for Figure 2(a) and 2(b) are given by:

$\mathcal{M}$	$\mathcal{S}$	$\mathcal{D}$	$\mathcal{L}$	$\mathcal{R}$	$\mathcal{MAX}_\mathcal{B}$	$\mathcal{F}_\mathcal{H}$	$\mathcal{B}_\mathcal{H}$	$\mathcal{F}_\mathcal{A}$
$\mathcal{M}_\mathcal{A}$	5	2	0	1	1	2	1	0
$\mathcal{M}_\mathcal{B}$	8	2	3	1	1	2	1	0

Note that even though the $\mathcal{F}_\mathcal{H}$ and $\mathcal{B}_\mathcal{H}$ for both examples are the same because each example has the same `skos:narrower` and `skos:broader` relations, the structure of both example is different. However, by looking at the $\mathcal{S}, \mathcal{L}, \mathcal{R}$, we may distinguish the structure of Example 1 from Example 2.

Following the categories of KOS discussed in Section 1, we also defined some rules using the SKOS metric, $\mathcal{M}$, to categorise the vocabularies in our corpus:

- If all $\mathcal{D}, \mathcal{F}_{\mathcal{H}}, \mathcal{B}_{\mathcal{H}}, \mathcal{F}_{\mathcal{A}} > 0$, then this vocabulary is categorised as a ***Thesaurus***.
- If all $\mathcal{D}, \mathcal{F}_{\mathcal{H}}, \mathcal{B}_{\mathcal{H}} > 0$ and $\mathcal{F}_{\mathcal{A}} = 0$, then this vocabulary is categorised as a ***Taxonomy***.
- If all $\mathcal{D}, \mathcal{F}_{\mathcal{H}}, \mathcal{B}_{\mathcal{H}}, \mathcal{F}_{\mathcal{A}} = 0$, then this vocabulary is categorised as a ***Glossary***.
- If the vocabulary does not belong to any of the above category, then this vocabulary is categorised as ***Others***. For example, the vocabulary uses only associative relation but not hierarchical relations.

3 Result and Observation

We collected 303 candidate SKOS vocabularies from the first method of corpus collection[12]. As for the second method, we collected 4220 URIs[13]. We collected 2296 URIs of candidate SKOS vocabularies from the third method[14]. This gives a total of 6819 candidate SKOS vocabularies.

After the SKOS vocabulary identification stage, 1068 vocabularies were identified as SKOS vocabularies according to our definition of SKOS vocabulary. The filtering of structurally identical SKOS vocabularies resulted in the exclusion of 603 identical SKOS vocabularies and 11 older versions of SKOS vocabularies, which gave us 454 SKOS vocabularies for further analysis.

Typology of modelling slips Based on our analysis of the collected result, we determined three types of modelling slips as follows:
Type 1: Undeclared property type. Each property used in the vocabulary is not explicitly typed as any of the property types such as `owl:objectProperty`, `owl:dataProperty`, `owl:annotationProperty`, etc. An example of this modelling slip is the use of `skos:broader` or `skos:narrower` properties in the vocabulary without explicit declarations as `owl:ObjectProperty`; such as through an `owl:import` of SKOS core vocabulary. This can cause tooling like the OWLAPI to treat the property as `owl:annotationProperty`. The *fixing* procedure for this type of modelling slip was to add the declarations for SKOS related properties. Applying the *fixing* procedure fixed 18 SKOS vocabularies.
Type 2: Inconsistent. We found 20 SKOS vocabularies were inconsistent. We classify the reasons for inconsistent vocabularies into:
- unintentionally typing an individual that is supposed to be an instance of `skos:ConceptScheme` class to also be an instance of `skos:Concept` class. The *fixing* procedure for this type of modelling slip was to remove the declaration of `skos:Concept` class from the individual. Applying the *fixing* procedure fixed 5 SKOS vocabularies.
- unintentionally using `skos:narrower` construct to relate between a member and its collection. We found that the `skos:member` property was declared in the

[12] This figure is valid as at 8 December 2010.

[13] as at 2 March 2011.

[14] Note that the web crawler runs for approximately three months, ended on 17 May 2011.

Table 1. Summary results

Stages	vocabs
Corpus preparation:	6819
- 1st method	303
- 2nd method	4220
- 3rd method	2296
SKOS vocabularies identification	1068
- 1st method	143
- 2nd method	432
- 3rd method	21
Structurally identical filtering	454
Fixing modelling slips:	24
- Type 1	18
- Type 2	6
Total SKOS vocabularies	478

Table 2. Exclusions

Exclusion type	vocabs
Plain HTML pages/blogs/forum	2986
Ontologies that are not SKOS vocabularies	1152
File not found	632
Connection refused	511
Network is unreachable	331
Connection timed out	284
Parsing error	266
Actual SKOS Core vocabulary	93
Failed to load import ontology	63
Modelling slips (Type 2 & Type 3)	23 (14 & 9)
Total exclusion	6341

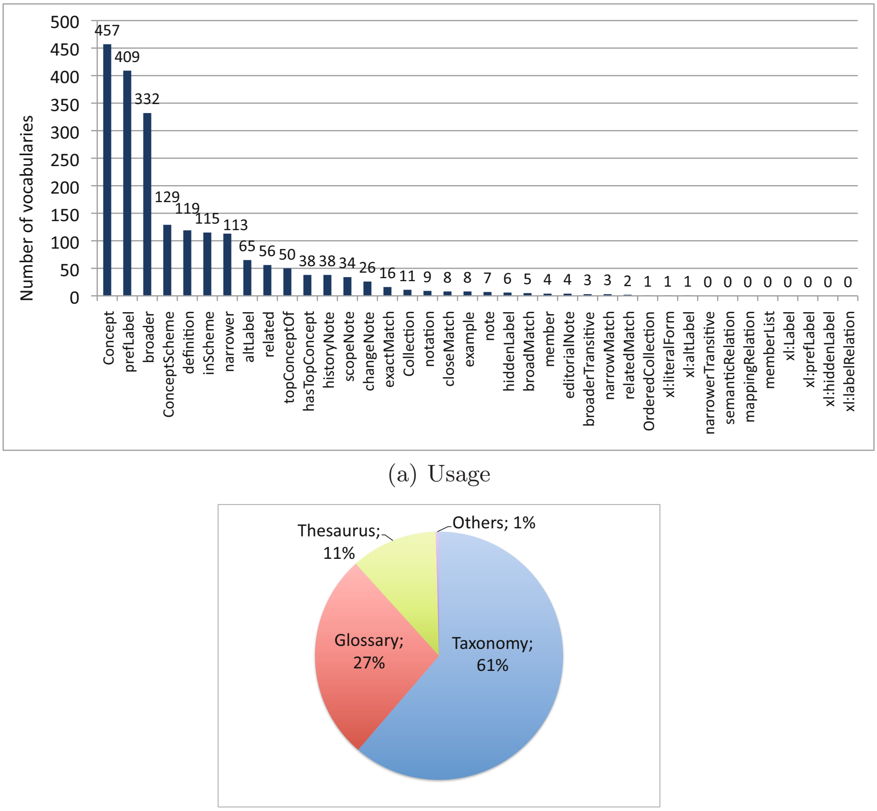
vocabulary but never used. The *fixing* procedure for this type of modelling slip was to replace the `skos:narrower` property with the `skos:member` property to show the relationship between a member and its collection. Applying the *fixing* procedure fixed 1 SKOS vocabulary.
- invalid datatype usage such as invalid dateTime datatype, invalid integer datatype, invalid string datatype, etc. This mistake was considered unfixable *errors*. There were 14 SKOS vocabularies excluded from the survey.

Type 3: Use of unsupported datatype. This was issued by the reasoner when encountering user-defined datatype. Note that this is not exactly a modelling slip instead a result from some limitations of the reasoner, thus, hindered us from getting the required data. To deal with this case, we first checked whether the user-defined datatype was actually in use to type the data in the vocabulary. If the datatype was not in use, we excluded the datatype from the datatype list and reclassified the vocabulary. There were 9 vocabularies excluded from the survey.

Fixing the modelling slips resulted in 24 additional SKOS vocabularies included in the corpus, which gave us the final number of 478 SKOS vocabularies. The summary figures and reasons for exclusion are presented in Tables 1 and 2. The full results and analysis can be found at `http://www.myexperiment.org/packs/237`.

Figure 3(a) shows the percentage of the SKOS construct usage. For each SKOS construct, a SKOS vocabulary was counted as using the construct if the construct is used at least once in the vocabulary. In this stage, we only counted the asserted axioms in each of the SKOS vocabularies.

Of all the SKOS constructs that are made available in the SKOS Recommendation[1] `skos:Concept`, `skos:prefLabel`, and `skos:broader` are the three most used constructs in the vocabularies, with 95.6%, 85.6% and 69.5%,



(a) Usage

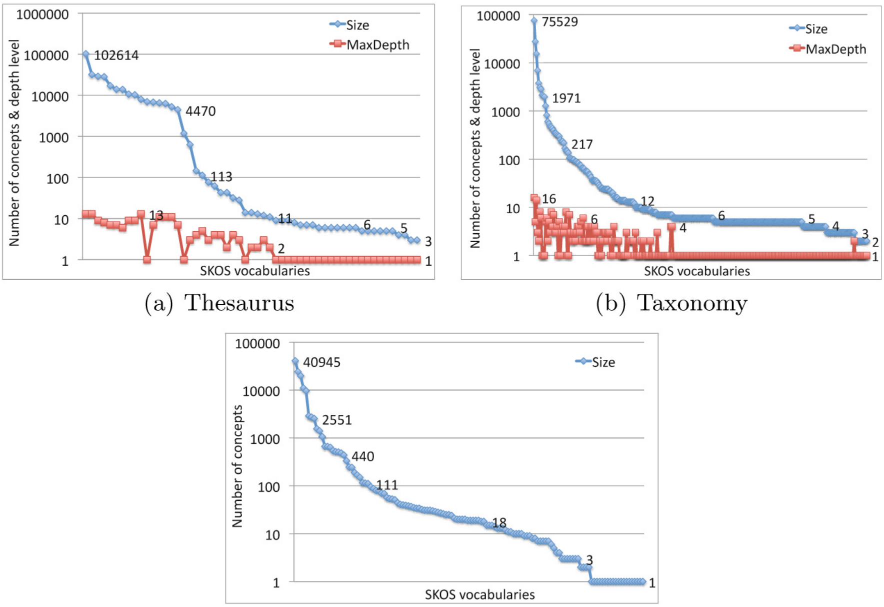
(b) Usage

Fig. 3. Construct Usage and Vocabulary Types

respectively. 28 out of 35 SKOS constructs were used in less than 10% of the vocabularies. There were eight SKOS constructs that were not used in any of the vocabularies.

The rules in Section 2 were used to categorise the vocabularies following the types of KOS as described in Section 1. Figure 3(b) shows a chart representing different types of SKOS vocabulary. As shown in this figure, 61% or 293 of the vocabularies are categorised as *Taxonomy*. The second largest type is *Glossary* with 27% or 129 vocabularies. 11% or 54 vocabularies fell into the *Thesaurus* category.

The remaining 1% or 2 vocabularies were categorised as *Others*. Further inspection revealed that the two vocabularies in the *Others* category are a *Glossary* with 4 `skos:related` properties on 2 pairs of the *concepts* (out of 333 *concepts*) and a snippet of a real SKOS vocabulary intended to show the use of `skos:related` property. We decided to reclassify these two vocabularies into the *Glossary* category for the first one and the *Thesaurus* category for the other.



(a) Thesaurus

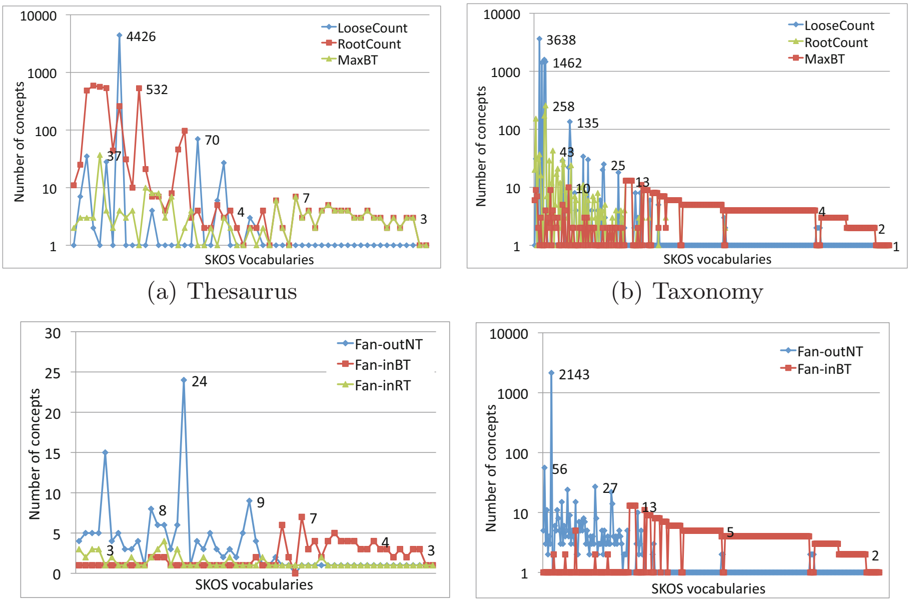
(b) Taxonomy

(c) Glossary

Fig. 4. Size of vocabulary and its maximum depth of vocabulary structure

Figure 4 plots the size of SKOS vocabularies and the maximum depth of the vocabulary structure. Each subgraph represents different categories of SKOS vocabulary; *Thesaurus*, *Taxonomy* and *Glossary*. Within each category, the vocabularies are sorted according to their size in descending order. The size of SKOS vocabulary was calculated based on the number of SKOS concepts in the vocabulary. The maximum depth of vocabulary was calculated based on hierarchical relations; `skos:broader` and `skos:narrower`; in the vocabulary. Figure 4(a) shows that the smallest size of vocabulary for the *Thesaurus* category is 3 concepts and the largest is 102614 concepts. The maximum depth of the vocabulary structure ranged from 1 to 13 levels. For the *Taxonomy* category as shown in Figure 4(b), the smallest size of vocabulary was 2 concepts and the largest was 75529 concepts. The maximum depth of the vocabulary structure ranged from 1 to 16 levels. Figure 4(c) plots size of vocabularies from the *Glossary* category. The smallest size of vocabulary is 1 concept and the largest is 40945 concepts. The maximum depth of the vocabulary structure for the *Taxonomy* category ranged from 1 to 16 levels. Note that there was no maximum depth for the *Glossary* category because there are no hierarchical relations were present in the vocabularies of this category.

Figure 5(a) and 5(b) show the number of loose concepts, root concepts and maximum `skos:broader` relations for each vocabulary structure. For the *Thesaurus* category, the loose concepts ranged from 0 concept (which means all concepts are connected to at least another concept) to 4426 concepts. The root concepts ranged from 1 root to 590 root concepts. The maximum `skos:broader`



(a) Thesaurus

(b) Taxonomy

(c) Thesaurus - branching factors

(d) Taxonomy - branching factors

Fig. 5. (a) Number of loose concepts, root concepts, maximum `skos:broader` relations, hierarchical branching factors, and associative branching factors of vocabulary structure

relations ranged from 1 relation to 37 relations. Only 11 out of 54 vocabularies had a single parent (mono-hierarchy) and the rest had at least 2 parent concepts.

As for the *Taxonomy* category, the loose concepts ranged from 0 concept to 3638 concepts. The root concepts ranged from 1 concept to 258 concepts. The maximum `skos:broader` relations ranged from 1 relation to 13 relations. 223 out of 293 vocabularies had more than 1 `skos:broader` relations, which meant that these were poly-hierarchy graphs.

Figure 5(c) and 5(d) show the number of hierarchical and associative branching factor for the *Thesaurus* and *Taxonomy* category. For the *Thesaurus* category, the smallest hierarchical FBF was 1 branch and the largest was 24 branches. The smallest hierarchical BBF was 1 branch and the largest was 7 branches. As for the associative FBF, the smallest was 1 branch and the largest was 4 branches. For the *Taxonomy* category, the smallest hierarchical FBF was 1 branch and the largest was 2143 branches. The smallest hierarchical BBF was 1 branch and the largest was 13 branches.

4 Discussion

As shown in Figure 3(a), of all the SKOS constructs that are available in the SKOS standards, only 3 out of 35 SKOS constructs are used in more than 60%

of the vocabularies. These constructs are `skos:Concept`, `skos:prefLabel` and `skos:broader`. Note that the `skos:Concept` construct usage is not 100%, due to not all vocabularies explicitly typing their individuals as `skos:Concept` in their vocabularies. However, these vocabularies use other SKOS constructs such as semantic relations that infer the individuals as `skos:Concept` due to its domain and range constraints. In some SKOS applications available on the Web like SKOS Reader[15], being able to recognise SKOS concepts is the key to display a SKOS vocabulary. Sometimes, these applications are not equipped with automatic reasoner that could infer the SKOS concepts. Therefore, not explicitly typing the resources as SKOS concepts in a SKOS vocabulary could prevent certain SKOS applications like SKOS Reader from behaving properly.

We found that `skos:broader` relations are used more frequently compared to `skos:narrower` relations, with 69.5% over 23.6% of vocabularies. Note that `skos:narrower` is an inverse relation of `skos:broader`. In those vocabularies where the ontology engineers only specify either one of these relations we think they may be taking advantage of its semantic property to infer the inverse relation. However, we found that 77 vocabularies specified both relations, `skos:broader` and `skos:narrower`, for any pair of concepts that had either relation. 7 of these vocabularies were originated from traditional KOS via some form of conversion process. Various works on converting from traditional KOS to SKOS can be found in [8,9,10,11].

Note also that the `skos:prefLabel` construct is not used in all of the vocabularies. Further analysis revealed that some vocabularies use `rdfs:label` for labelling their concepts. There are 25 out of 35 SKOS constructs used in less than 10% of the vocabularies. Some of these constructs are SKOS mapping properties, which are expected to be used in vocabularies that define alignment to other vocabularies. Other constructs that fell into this portion are mainly the SKOS documentation constructs and SKOS extended (SKOS-XL) constructs. The result also showed that more than 1000 vocabularies excluded from the survey used some of the SKOS constructs such as lexical labelling and documentation constructs.

From the results shown in Figure 3(b), the largest category of 61% of the vocabularies in our corpus are categorised as a *Taxonomy*. All the three categories, *Thesaurus*, *Taxonomy* and *Glossary*, are consistent with the KOS categories.

The results we found in this survey corresponded to some extent to the result of the Controlled Vocabulary Survey[16] conducted by the Semantic Web Company published in June 2011. They conducted an online survey, involving 158 participants which aimed to investigate and learn more about some aspects related to controlled vocabularies. Amongst the foci of interests are; i) preferred knowledge models, ii) main application areas, iii) importance of standards, and iv) trends in organization sizes. The result of their survey revealed that taxonomies and ontologies seem to be the preferred knowledge models. Semantic

[15] http://labs.mondeca.com/skosReader.html

[16] `http://poolparty.punkt.at/wp-content/uploads/2011/07/`
`Survey_Do_Controlled_Vocabularies_Matter_2011_June.pdf`

search, data integration and structure for content navigation are the main application areas for controlled vocabularies. Standards like SKOS have gained greater awareness amongst the participants, which shows that the web-paradigm has entered the world of controlled vocabularies.

The result shows that there is no direct relationship between size of vocabulary and its maximum depth of vocabulary structure. We can see from the graph that the maximum depth of small vocabularies is almost similar to the maximum depth of large vocabularies for both the *Thesaurus* and *Taxonomy* categories.

For the *Thesaurus* category, 43 out of 55 or 80% of the vocabularies have maximum `skos:broader` relations more than one. This means that at least one of the concepts in these vocabularies has more than one *broader concept*, which make them poly-hierarchy graphs. 93% of the vocabularies have more than one root concept, which means that these vocabularies have shape of multi-trees. As for the *Taxonomy* category, 76% of the vocabularies have maximum `skos:broader` relations more than one and 81% of the vocabularies have more than one *root concept*.

As for the hierarchical and associative branching factors result, we found one anomaly to the *Taxonomy* category where one of the vocabularies had depth, $\mathcal{D}$ of 1 and only one *root concept*. One particularly noteworthy value of the metric is that the hierarchical FBF, $\mathcal{F}_{\mathcal{H}}$ is 2143, which is one less than the total vocabulary size, $\mathcal{S}$, which is 2144. This vocabulary is an outlier, with a a single root node and a very broad, shallow hierarchy. If we were to exclude this vocabulary, the range of hierarchical FBF for the *Taxonomy* category is between 1 and 56. The hierarchical FBF alongside hierarchy depth are important in determining the indexing and search time for a particular query[12].

5 Conclusion

Our method for collecting and analysing SKOS vocabularies has enabled us to gain an understanding of the type and typicality of those vocabularies. We found out that all but two of the SKOS vocabularies that we collected from the Web fell into one of the categories listed by the traditional KOS; flat vocabularies (*Glossary*), multi-level vocabularies (*Taxonomy*) and relational vocabularies (*Thesaurus*). In the future, we plan to select several SKOS vocabularies from each category and study them in more detail in terms of the use and function of the vocabularies in applications.

Based on the results of this survey, a typical taxonomy looks like a polyhierarchy that is 2 levels deep, with a $\mathcal{FBF}_{\mathcal{H}}$ of 10 concepts and a $\mathcal{BFF}_{\mathcal{H}}$ of 3 concepts. A typical thesaurus also looks like a polyhierarchy that is 6 level deep, with a $\mathcal{FBF}_{\mathcal{H}}$ of 3 concepts and a $\mathcal{BFF}_{\mathcal{H}}$ of 2 concepts, additionally having associative relationships, $\mathcal{FBF}_{\mathcal{A}}$ of 1 concept.

In this survey, we collected 478 vocabularies that according to our definition are SKOS vocabularies. Three years after becoming a W3C Recommendation, the use of SKOS remains low. However, our total of SKOS vocabularies may be artificially low, with some being hidden from our collection method. The reasons for some of these vocabularies not being publicly accessible by an automated process could be due to confounding factors such as proprietary issue, vocabularies stored within SVN, etc. However, we are confident that we have done our best to deploy various methods in collecting SKOS vocabularies that are publicly available on the Web.

Acknowledgements. Nor Azlinayati Abdul Manaf is receiving the scholarship from Majlis Amanah Rakyat (MARA), an agency under the Malaysian Government, for her doctorate studies. Many thanks to the reviewers who gave insightful comments and suggestion to improve this paper.

References

1. Miles, A., Bechhofer, S.: SKOS simple knowledge organization system reference. W3C recommendation, W3C (2009)
2. Zeng, M.L., Chan, L.M.: Trends and issues in establishing interoperability among knowledge organization systems. JASIST 55(5), 377–395 (2004)
3. Hodge, G.: Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. Council on Library and Information Resources (2000)
4. Wang, T.D., Parsia, B., Hendler, J.: A Survey of the Web Ontology Landscape. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 682–694. Springer, Heidelberg (2006)
5. Ding, L., Finin, T.W.: Characterizing the Semantic Web on the Web. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 242–257. Springer, Heidelberg (2006)
6. Tempich, C., Volz, R.: Towards a benchmark for semantic web reasoners - an analysis of the DAML ontology library. In: EON (2003)
7. Poole, D.L., Mackworth, A.K.: Artificial Intelligence: Foundations of Computational Agents. Cambridge University Press, New York (2010)
8. van Assem, M., Malaisé, V., Miles, A., Schreiber, G.: A Method to Convert Thesauri to SKOS. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 95–109. Springer, Heidelberg (2006)
9. Panzer, M., Zeng, M.L.: Modeling classification systems in SKOS: some challenges and best-practice recommendations. In: Proceedings of the 2009 International Conference on Dublin Core and Metadata Applications, Dublin Core Metadata Initiative, pp. 3–14 (2009)
10. Summers, E., Isaac, A., Redding, C., Krech, D.: LCSH, SKOS and linked data. In: Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications (DC 2008), Dublin Core Metadata Initiative, pp. 25–33 (2008)
11. Binding, C.: Implementing Archaeological Time Periods Using CIDOC CRM and SKOS. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) ESWC 2010, Part I. LNCS, vol. 6088, pp. 273–287. Springer, Heidelberg (2010)
12. Roszkowski, M.: Using taxonomies for knowledge exploration in subject gateways. In: Proceedings of the 17th Conference on Professional Information Resources, INFORUM 2011 (2011)