A semantic approach for ontology evaluation

Montserrat Batet and David Sánchez
UNESCO Chair in Data Privacy
Department of Computer Engineering and Mathematics
Universitat Rovira i Virgili
Av. Països Catalans, 26. 43007, Tarragona, Spain
{montserrat.batet, david.sanchez}@urv.cat

***Abstract*—In recent years, ontologies have experienced an enormous development due to their importance in knowledge-based systems. Because of the discrepancies that may appear during the modeling of ontologies, and due to the availability of ontologies covering overlapping domains, ontology evaluation is crucial in order to select the most appropriate ontology for a specific application. Many ontology evaluation mechanisms available in the literature assess the quality of ontologies according to their structural features. Even though, most of them propose ad-hoc scores aggregating different features, which lacks semantic and mathematical coherence. In this paper, we present an intuitive measure for ontology evaluation that quantifies the *semantic dispersion* of the ontology, which is both mathematically and semantically coherent. Our proposal is inspired in the standard notion of numerical dispersion of a sample and on a recent empirical study showing which ontological features can better predict the accuracy of ontologies. Our experiments, performed over a set of widely used ontologies, suggest that our measure positively correlates with such features, while offering a more coherent ontology evaluation score.**

Knowledge representation; Ontology evaluation; semantics; dipsersion

I. INTRODUCTION

The proliferation of ontologies during the last years has been motivated by the success of the WWW and the aim of global initiatives like the Semantic Web [1] of incorporating semantic content to web resources in order to improve the machine understanding of their contents.

Ontologies, which can be defined as a formal representation of a shared conceptualization [2], describe classes (concepts), attributes, individuals (instances), and relations between classes in a structured and machine readable way. Ontologies have been extensively used in many applications that require managing and interpreting textual data from a semantic perspective such as information retrieval [3] and extraction [4, 5], document classification [6] or data privacy [7-9].

Nowadays, thousands of ontologies covering a variety of overlapping topics and domains are available through the Web [10]. Ontologies come from a vast variety of different sources, spanning institutions and persons aiming for different goals and quality criteria. However, due to the complexity, and potential discrepancies of knowledge representation, ontologies can model domains of knowledge in significantly different (better or worse) ways. Moreover, the lack of consensus in ontology development [11, 12] makes difficult the assessment of the suitability of an ontology for a specific task. Thus, ontology evaluation, whose aim is to assess the quality of ontologies, is essential for a wide adoption of ontologies because it provides feedback to ontology developers and knowledge engineers and gives insights on the adequacy of ontologies to their users [13, 14].

From the different dimensions that can be evaluated, the semantic accuracy of the ontology, which states the semantic coherency and suitability of the components of the ontology, is the most important one, because it directly influences the accuracy of ontology-based systems. As it will be discussed in the next section, most of the proposed works on ontology accuracy evaluation are based on the analysis of the structural features (e.g., number of classes or taxonomic depth) of ontologies. In these works, the different structural features are aggregated using ad-hoc measures and empirically tuned weights, which lack semantic and mathematical coherence.

To tackle these limitations, in this paper we present a measure to quantify the semantic accuracy of ontologies. In contrast to most of the existent ontology evaluation approaches, our proposal aims at providing a semantically-coherent, numerically normalized and integrated quantification of the semantic accuracy of a given ontology. To ensure its applicability, our approach solely relies on taxonomic knowledge that, contrary to other non-taxonomic knowledge (e.g. part-of, concept features, axioms, etc.) is available in all ontologies [10].

Inspired by a recent study [15], which identified how the variance of the taxonomic depth and breadth can be used to reasonably predict the semantic accuracy of ontologies, we propose a measure to semantically quantify the dispersion of the concepts modeled in the taxonomic structure of the ontology, in a similar way as standard numerical dispersion measures do for a sample of values. In the experiments reported in this paper, we will also show that, for several well-known ontologies, our proposed *semantic dispersion* measure positively correlates with the aforementioned structural characteristics that reasonably predict the accuracy of ontologies, while offering a more coherent and intuitive ontology evaluation score.

The rest of the paper is organized as follows. Section 2, discusses related works on ontology evaluation. Sections 3

and 4 introduce the proposed measure and illustrate, by means of several examples, its suitability as a measure of *semantic dispersion* and its relationship with the structural features that better predict the ontological accuracy. Section 5 details the empirical experiments performed with several well-known ontologies. The final section depicts the conclusions of the work.

II. RELATED WORK

Ontology evaluation approaches consider different quality dimensions such as the semantic accuracy of the ontology, adaptability to different contexts, clarity for human readers, completeness, conciseness and logical consistency [16]. The most important dimension from the perspective of automated knowledge-based systems is the *semantic accuracy* of the ontology because the semantic coherency and adequacy of the modeling of the ontological components (i.e., classes, relationships, etc.) directly influences the precision of knowledge-based semantic assessments or inferences [15].

Most of the approaches that evaluate the ontological accuracy rely on the analysis of structural features (i.e., number of classes, relationships, taxonomic depth, branching factor, etc.) of ontologies [15-17]. These approaches can be easily and efficiently calculated from the graph defined by the semantic links modeled in the ontology. Moreover, they provide an objective and numerical assessment of the ontology quality without depending on biased or domain-dependent external resources, or on the subjective criterion of a human expert. However, because each individual structural feature provides a partial picture of the ontology quality, the different features are usually aggregated using ad-hoc measures and weights, a solution that lacks theoretical basis and, thus, generality. This issue is closely related with the fact that, traditionally, there were not evidences of which features are the most suitable ones to quantify the semantic accuracy of ontologies. Finally, many methods rely on non-taxonomic structural features (e.g., non-taxonomic relationships, attributes, etc.) which are hardly found in available ontologies [10], thus hampering their applicability.

A recent study [15] have identified how certain structural features can be used to predict the semantic accuracy of ontologies. In that study the semantic correctness of the relationships modeled in a set of ontologies was manually evaluated. Then, the authors studied different structural features and identified which ones could be used to identify ontologies with a correct knowledge modeling. In particular, the authors mainly focused on taxonomic features. They conclude that the structural characteristics that, individually, seem to better predict the ontological accuracy are:

- The *maximum depth*: length of the longest taxonomic branch in the ontology, measured as the number of concepts from the root node to the leaves of the taxonomy.
- The *average depth:* average length of all taxonomic branches.
- The *variance of the taxonomic depth*: dispersion of depths of the longest branches with respect to the average depth, which is computed as the standard mathematical variance.
- The *maximum breadth*: width of the taxonomic level of the ontology with the largest number of concepts.
- The *variance of the taxonomic breadth*: dispersion of breadths of all taxonomic levels with respect to the average breadth.

More specifically, the authors argue that *depth* and *breadth variances* are the best predictors of the ontology accuracy. They also note that structural features commonly used in the ontology evaluation literature, such as the total number of concepts or the number of properties of such concepts, do not show a significant correlation with the ontological accuracy.

The main limitation of the previous work is that the values individually quantifying such features may provide contradictory predictions when considered as a whole. Thus, an ad-hoc aggregation or selection of features based on empirical hypothesis or experiments is still needed. In fact, in [15] these features are used as thresholds for binary classifiers and not as a continuous quality score, which is the tool that enables to rank a set of ontologies according to their accuracy. Moreover, the values of these features are greatly influenced by the cardinality (i.e., number of concepts) of the ontology, thus hampering the comparison of ontologies with significantly different sizes. Finally, an approach based solely on structural features omits the semantics (i.e., the relationships defined between concepts) explicitly modeled in the ontology.

III. ONTOLOGY ACCURACY EVALUATION MEASURE

Recalling the main conclusions reported in [15], we have that the variances of the taxonomic depth and breadth of an ontology seem to be positively correlated with the semantic accuracy of the knowledge modeled in that ontology. Thus, given that the variance is a measure of dispersion, we can assume that a largely dispersed taxonomic structure, which is reflected by a high depth and breadth variance, will suggest an accurate knowledge modeling. By inversing the argument, this also suggests that a semantically accurate knowledge representation of a domain rarely produces a perfectly balanced and homogenous structure.

These assumptions are coherent with the inherent nature of knowledge representation, which is an a posteriori formalization of the de facto consensus regarding the semantics of concepts as they are used and understood by human societies [18]. Since such semantics unpredictably and informally evolve through time as societies develop their knowledge, the inherently empirical knowledge representation process will unlikely produce perfectly homogenous and balanced knowledge structures.

Moreover, in a very homogenous structure, concepts are evenly distributed through the taxonomy, thus becoming less distinguishable from each other when analyzing their structural characteristics. Because many reasoners rely on such structure to infer concept semantics [19], a very homogenous structure is, in general, not desirable; in fact,

this homogeneity goes against the main purpose of knowledge modeling, which is to make concepts well-differentiated from each other in order to minimize the ambiguity of the semantic inferences relying on the modeled knowledge [19].

Given the above arguments, we propose an ontology accuracy evaluation score that relies on measuring the degree of *semantic dispersion* of concepts in a given ontology. Our goal is to obtain an integrated, generic, non-parameterized, intuitive and mathematically coherent score to quantify the dispersion of concepts modeled in an ontology by considering the intrinsic semantics of such concepts in the ontology.

From a numerical perspective, dispersion describes the spread of a sample, or its variation around a central value. Thus, the *semantic dispersion* of concepts in a given ontology should be quantified with regard to the center of the knowledge structure of that ontology. From a taxonomic perspective, the "center" or *centroid* of an ontology is the root node of the taxonomic tree, since it is the concept that subsumes the meaning of all the other concepts, which represent its specializations.

It is important to note that, even if the ontology incorporates several disjoint taxonomic trees, all of them can be naturally joined by a virtual root that generalizes all of them. In fact, this common root note is usually included in most knowledge repositories (e.g., *entity* in WordNet, *Concept* in SNOMED-CT, *Top* in the taxonomy of categories in the Open Directory Service, *thing* in OWL ontologies, etc.).

From a structural perspective, in a perfectly balanced taxonomy, in which all branches have the same depth and the branching factor is constant for all concepts, the root node of the taxonomy matches with the geometric center of the graph defined by such taxonomy; thus, coherently with its numerical definition, this central node or *centroid* is the one that minimizes the distances with respect to all the concepts in the ontology [20].

On the contrary, when taxonomic branches of the ontology present different depths (i.e., *depth variance*, which is limited by the *maximum depth*) and branching factors (that influence the *breadth variance*, and which is limited by the *maximum breadth*), the unbalancing of the ontology with respect to the *centroid* tends to increase. In such case, the degree of unbalancing quantifies how much concepts are dispersed with respect to the reference *centroid*. Thus, by measuring this degree of ontological unbalancing or dispersion, we are capturing the structural characteristics that are related to the accuracy of the ontology (i.e., depth and breadth variances).

In order to compute a *semantic dispersion* and thus, capture the semantics modeled in the ontology, we should be able to estimate the *semantic distance* between each concept and the reference *centroid* (i.e., the root node) of the taxonomy. A suitable *semantic distance* to evaluate and compare the accuracy of ontologies should not only capture the semantic differences between concepts, but it should not be influenced by the cardinality of the ontology; in this manner, the accuracy of ontologies with different cardinalities could be coherently computed and compared. Finally, given that many ontology only incorporate taxonomic relationships [10], the measure should only rely on taxonomic knowledge.

Within the literature of computational linguistics, many different mechanisms have been proposed to measure the semantic distance between concepts by capturing the semantics modeled in the ontology [21]. The simplest methods consider ontologies as graphs in which the distance between two concepts can be measured as the minimum number of semantic links (typically taxonomic) separating them [22]. Because absolute distance values are provided, distances tend to be larger as the ontology size increases. This goes against our goal of coherently compare ontologies with different cardinalities. Another drawback of these measures is their relatively low accuracy, which is caused by the fact that they only consider the shortest path connecting the two concepts to be compared. This causes that, in ontologies in which concepts are connected with several paths (i.e., multiple taxonomic inheritance exist), a lot of knowledge is omitted; as show in empirical studies, this negatively influences the accuracy of the similarity assessments [23].

To solve the drawbacks of these simple methods, feature-based distance measures have been proposed. They rely on comparing concepts according to the amount of semantic evidences that they have and do not have in common in the ontology. Based on this principle, in [23, 24] a state-of-the-art feature-based measure is proposed that measures the semantic distance $d(c_1,c_2)$ between two concepts c_1 and c_2 as a function of their number of non-common taxonomic ancestors divided (for normalization) by their total number of ancestors:

$$d(c_1,c_2)=\log_2\left(1+\frac{\left|T(c_1)\cup T(c_2)\right|-\left|T(c_1)\cap T(c_2)\right|}{\left|T(c_1)\cup T(c_2)\right|}\right) \qquad (1)$$

where $T(c_i)$ is the set of taxonomic ancestors of concept c_i in the ontology, including itself.

From a semantic perspective, this measure captures more knowledge than the methods based on counting links, since it implicitly considers *all* the paths connecting the two concepts by evaluating all of their ancestors. Moreover, thanks to the normalizing denominator, the distance can differentiate concept pairs with the same number of shared ancestors. Finally, the non-linearity of the calculus better aggregates the semantic evidences gathered from the ontology (i.e., the number of common and disjoint ancestors), because the relationship between the amount of such evidences and the semantic distance has also proven to be non-linear [25]. As a result, according to a standardized set of empirical experiments, (1) approximates human judgments of similarity better than other ontology-based measures [23, 24].

Numerically, (1) yields positive normalized values in the [0,1] range, thus, making it independent on the ontology size and suitable to coherently compare the accuracy of ontologies with different cardinalities. Finally, as

demonstrated in [24] and [26], (1) satisfies non-negativity, reflexivity, symmetry and triangular inequality, thereby being a distance measure in the mathematical sense. This is relevant in order to compute the *semantic dispersion* in a mathematically-coherent way.

By relying on the above semantic distance, our measure of *semantic dispersion* of an ontology is inspired in the numerical notion of dispersion of a sample quantified by its *standard deviation*. Contrary to other dispersion measures of a sample like the variance, a useful property of the standard deviation is that it is expressed in the same units as the data. So a *semantic dispersion* measure based on the standard deviation will provide values expressed in the same units as the semantic distance, thus enabling to define coherent ranges or intervals of semantic distances and to group or classify semantically similar concepts according to such intervals.

Formally, likewise the standard deviation of a numerical sample, we define the *semantic dispersion* of an ontology *O*, which taxonomically models a set of concepts *C*, as the square root of the average squared *semantic distance* between each concept c_i in *C* and the reference *centroid* (i.e., the root node) of *O*.

$$Dispersion(O) = \sqrt{\frac{\sum_{c_i \in C} d(c_i, Root(O))^2}{|C|}} \qquad (2)$$

where $|C|$ states the number of concepts in the ontology without counting the root, because we assume the root node is unique or it has been virtually added as a subsumer of all the other concepts.

It is important to note that the use of the squared semantic distances in the dispersion calculus makes differences between concepts positive and increases the contribution of the most scattered values, that is, those that contribute more to the unbalancing of the taxonomy. Moreover, because of the use of a semantically grounded distance, the resulting dispersion assessment captures the intrinsic semantics of the evaluated concepts, as they are modeled in the ontology.

Numerically, when the *semantic dispersion* is equal to zero, it indicates that all values are identical, whereas a high *dispersion* indicates that concepts are very spread out from the reference center and from each other, thus being well differentiated.

Finally, if we are solely interested in the *semantic dispersion* of a particular taxonomic branch of an ontology (e.g., a particular type of disease in an ontology modeling medical concepts), it is possible to restrict the dispersion calculation to just this branch. This can be done by using the common generalization of that branch (e.g., the concept representing a general disease type) as the reference node and computing the respective semantic distances of all its taxonomic specializations.

IV. EXAMPLE

This section illustrates how the *semantic dispersion* measure proposed in this paper behaves with regard to the taxonomic structure of an ontology. According to eq. (2), we have that the minimum *dispersion* is obtained with a perfectly balanced taxonomic structure, in which the root node is the exact geometric center, and in which all concepts in the taxonomy are structurally indistinguishable from each other (i.e., they are direct specializations of the root node with no additional differentiating features).

In Fig. 1 we show an example of such structure in which six concepts (six carnivore mammals) are modeled as direct specializations of the root node. From a knowledge representation perspective, this taxonomy is equivalent to a flat list of concepts. Semantically, this gives no insights about the commonalities or differences between concepts, thus going against the purpose of a proper knowledge modeling.

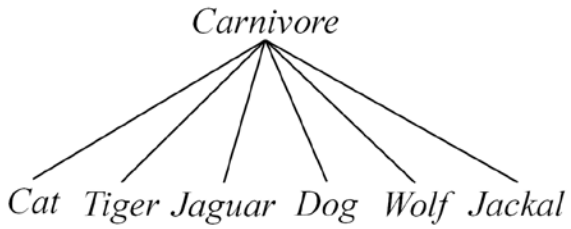


Figure 1. Sample ontology O_1.

By measuring the structural features of this taxonomy, we have that the *average* and *maximum depth* of the taxonomic branches is 1, the *depth variance* is 0, the *maximum breadth* is 6 and the *breadth variance* is 0 (without counting the root node, which is assumed to be unique or virtually added in all ontologies).

The low values of these features suggest an also low ontological accuracy, as discussed in the empirical study in [15]. Indeed, a taxonomic structure as simple as that of Fig. 1 makes concepts completely indistinguishable, since all of them have the same level of specificity and their semantic distances are identical.

The *semantic dispersion* of this ontology is:

$$Dispersion(O_1) = \sqrt{\begin{pmatrix} d(cat,carn)^2 + d(tiger,carn)^2 + d(jaguar,carn)^2 + \\ + d(dog,carn)^2 + d(wolf,carn)^2 + d(jackal,carn)^2 \end{pmatrix} / 6} =$$

$$= \sqrt{\begin{pmatrix} \left(\log_2\left(1+\frac{1}{2}\right)\right)^2 + \left(\log_2\left(1+\frac{1}{2}\right)\right)^2 + \left(\log_2\left(1+\frac{1}{2}\right)\right)^2 + \\ \left(\log_2\left(1+\frac{1}{2}\right)\right)^2 + \left(\log_2\left(1+\frac{1}{2}\right)\right)^2 + \left(\log_2\left(1+\frac{1}{2}\right)\right)^2 \end{pmatrix} / 6} = 0.585$$

In order to better represent the semantic differences of concepts, in Figure 2, we have alternatively modeled the same concepts by adding a new inner taxonomic level (*Feline* and *Canine*) that separates the set *Dog*, *Wolf* and *Jackal* from the set *Cat*, *Tiger* and *Jaguar*.

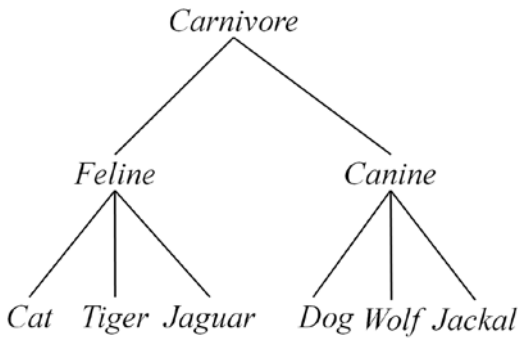


Figure 2. Sample ontology O_2.

The values of the structural features in this case are: the *average* and *maximum depth* of the taxonomic branches is 2, the *depth variance* is 0, the *maximum breadth* is 6 (the average breadth is (6+2)/2=4) and the *breadth variance* is $((2\text{-}4)^2+(6\text{-}4)^2)/2=4$.

Some of these values are still low and the root node remains as the exact geometric center of the taxonomy, but both the *maximum depth* and the *breadth variance* have increased as a result of the more accurate representation and better differentiation of concepts.

In this case, the *semantic dispersion* also increases accordingly:

$$Dispersion(O_2)=\sqrt{\begin{pmatrix} d(cat,carn)^2+d(tiger,carn)^2+d(jaguar,carn)^2+ \\ +d(dog,carn)^2+d(wolf,carn)^2+d(jackal,carn)^2+ \\ +d(feline,carn)^2+d(canine,carn)^2 \end{pmatrix}/8}=$$

$$=\sqrt{\frac{6\times\left(\log_2\left(1+\frac{2}{3}\right)\right)^2+2\times\left(\log_2\left(1+\frac{1}{2}\right)\right)^2}{8}}=0.702$$

In the final structure shown in Fig. 3, we represent a subset of the taxonomic structure that models this concepts in WordNet [27], in which *Tiger* and *Jaguar* has been better differentiated from *Cat* by adding a new inner node (*Phantera*, a *big cat* in WordNet).

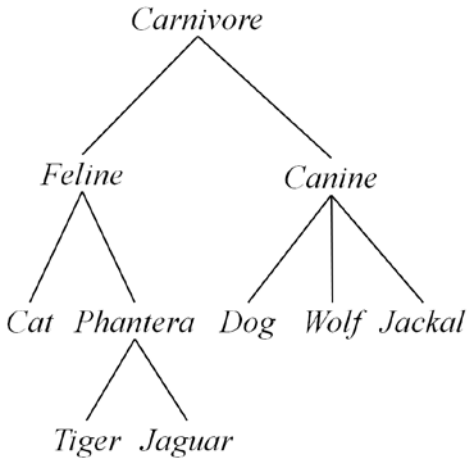


Figure 3. Sample ontology O_3.

In this case, the pair of concepts *Tiger* and *Jaguar* define a larger taxonomic branch, which makes that the *maximum depth* is now 3, the *average depth* is (2+3+3+2+2+2)/6=2.33, the *depth variance* is $(2\cdot(3\text{-}2{,}33)^2+4\cdot(2\text{-}2{,}33)^2)/6=0.22$, the *maximum breadth* is 5 (breadth average is now (2+5+2)/3=3) and the *breadth variance* is $((2\text{-}3)^2+(5\text{-}3)^2+(2\text{-}3)^2)/3=2$.

Even though *breadth* related features are now slightly lower, the increase in the *depth* related features suggests a less balanced structure. This is reflected by the fact that the root node is not the perfect geometric center of the tree anymore. In this case, however, an independent analysis of each structural feature does not show a clear improvement over the structure in Fig. 2: *breadth* related features are lower but *depth* related features are higher. Thus, it is difficult to conclude which structure seems to be more accurate.

In comparison, the *semantic dispersion*, which captures a coherent aggregation of all the taxonomic features, properly reflects the better differentiation between concepts of the structure in Fig. 3, resulting in a slight increase over the previous example:

$$Dispersion(O_3)=\sqrt{\begin{pmatrix} d(tiger,carn)^2+d(jaguar,carn)^2+d(cat,carn)^2+ \\ +d(panthera,carn)^2+d(dog,carn)^2+d(wolf,carn)^2+ \\ +d(jackal,carn)^2+d(feline,carn)^2+d(canine,carn)^2 \end{pmatrix}/9}=$$

$$=\sqrt{\frac{2\times\left(\log_2\left(1+\frac{3}{4}\right)\right)^2+5\times\left(\log_2\left(1+\frac{2}{3}\right)\right)^2+2\times\left(\log_2\left(1+\frac{1}{2}\right)\right)^2}{9}}=0.723$$

V. EXPERIMENTS

In this section we report the results of an empirical experiment in which our proposed measure of *semantic dispersion* is compared against the structural features of several widely-used ontologies and taxonomic trees. The goal is twofold: i) to show how our measure positively correlates with the structural features that, as discussed above, help to identify accurate ontologies and, ii) to illustrate how our measure is able to coherently aggregate such features and, thus, provide an intuitive and quantitative score for ontology evaluation.

We employed three widely-used ontologies [7, 8, 19, 24, 28] to perform the experiment:

- WordNet [27]: a lexical database that describes and structures more than 80,000 general concepts, which are semantically structured in an ontological way. The taxonomic structure in WordNet corresponding to nouns is subsumed by the abstract node *entity*, and represents most of the semantic relationships modeled in the ontology [29]. We used WordNet version 2.1.
- SNOMED-CT (The Systematized Nomenclature of Medicine, Clinical Terms) [30]: one of the largest sources included in the Unified Medical Language System (UMLS) of the US National Library of Medicine. It contains around 300,000 concepts organized into 18 overlapping hierarchies subsumed

by the abstract *SNOMED-CT concept*. We used the July 2013 release of SNOMED-CT.

- MeSH (The Medical Subject Headings) [31]: a hierarchy of medical and biological terms defined by the U.S National Library of Medicine to catalogue books and other library materials. It consists of a controlled vocabulary and a hierarchical tree with 16 categories containing more than 26,000 concepts. MeSH does not explicitly include a common root to all the 16 categories. Thus, in coherency with the premises of our work, we manually added the abstract *MeSH concept* as virtual root subsuming all of them. We used the 2011 release of MeSH.

Table 1 shows, for the three ontologies, the values of structural features considered in this study and of the *semantic dispersion* obtained with eq. (2). We can see that, for all the structural features, there is a perfect rank order positive correlation with regard to the *semantic dispersion*. Even though, quantitatively, the relationship between individual structural features and the *dispersion* is not linear. Indeed, structural features such as the maximum and average *depth* or *breath* closely depend on the size of the ontology. This makes difficult to coherently compare ontologies with significantly different cardinalities. In comparison, the *semantic dispersion*, provides values normalized in the range [0..1] that, as discussed, above, do not depend on the cardinality of the ontology, but just on the structure of the taxonomy. Thus, it provides a clearer assessment of ontological accuracy.

TABLE I. STRUCTURAL FEATURES AND *SEMANTIC DISPERSION* OF WORDNET, SNOMED AND MESH

Measure	**Ontology**		
	WordNet	***SNOMED-CT***	***MeSH***
Number of concepts	81,426	297,799	26,157
Maximum depth	17	19	12
Average depth	8.3	7	5.4
Depth Variance	3.38	3.89	1.17
Maximum breadth	17,837	60,778	7,775
Average breadth	4,790	15,674	2,012
Breadth Variance	32,474,163	375,721,551	6,226,389
Dispersion	**0.921**	**0.938**	**0.903**

In the previous experiment, results show that, as a whole, SNOMED-CT is the ontology with the largest dispersion followed by WordNet and MeSH; this ordering also matches with that defined by the cardinalities of the three ontologies. However, the different scopes of the three ontologies does not allow a fair comparison of their accuracy with regard to their dispersion: SNOMED-CT and MeSH are biomedical repositories with a high degree of overlap [32], whereas WordNet is a general-purpose repository that models many different domains.

In order to compare the dispersion and, thus, the assessed accuracy, of the three ontological structures within the scope of the same domain, we focused on the taxonomic trees of several domains that the three ontologies have in common: *drug*, *cell* and *virus*. These domains are necessarily related to biomedicine, because of the constrained scopes of SNOMED-CT and MeSH. Moreover, they are general enough to cover (i.e., subsume) a significant amount of concepts in each of the three ontologies and, thus, produce representative results. Following the procedure detailed at the end of the third section to measure the dispersion of a taxonomic tree, Tables 2, 3 and 4 show the values of the structural features and of the *semantic dispersion* of the taxonomic structures of the three domains in the three ontologies.

TABLE II. STRUCTURAL FEATURES AND *SEMANTIC DISPERSION* OF THE "DRUG" TAXONOMY IN WORDNET, SNOMED AND MESH

Measure	**Ontology ("Drug" taxonomy)**		
	WordNet	***SNOMED-CT***	***MeSH***
Number of concepts	908	6,798	8,975
Maximum depth	7	10	11
Average depth	3.73	4.74	5
Depth Variance	0.95	1.51	1.19
Maximum breadth	287	1,728	2,709
Average breadth	130	680	816
Breadth Variance	11,082	408,767	892,278
Dispersion	**0.851**	**0.932**	**0.898**

TABLE III. STRUCTURAL FEATURES AND *SEMANTIC DISPERSION* OF THE "CELL" TAXONOMY IN WORDNET, SNOMED AND MESH

Measure	**Ontology ("Cell" taxonomy)**		
	WordNet	***SNOMED-CT***	***MeSH***
Number of concepts	129	1,136	501
Maximum depth	6	6	9
Average depth	3.09	3.88	3.88
Depth Variance	1.28	1.88	2.18
Maximum breadth	37	274	125
Average breadth	21	189	55
Breadth Variance	122	6,362	1,718
Dispersion	**0.782**	**0.867**	**0.881**

TABLE IV. STRUCTURAL FEATURES AND *SEMANTIC DISPERSION* OF THE "VIRUS" TAXONOMY IN WORDNET, SNOMED AND MESH

Measure	**Ontology ("Virus" taxonomy)**		
	WordNet	***SNOMED-CT***	***MeSH***
Number of concepts	72	2,005	573
Maximum depth	4	10	7
Average depth	2.76	5.93	3.94
Depth Variance	0.43	1.69	0.89
Maximum breadth	31	649	146
Average breadth	18	200	77
Breadth Variance	103	46,083	3200
Dispersion	**0.761**	**0.881**	**0.862**

Within the scope of these domains, the pairwise relationship between each structural feature and the *semantic dispersion* does not perfectly correlates in all cases. For the *virus* domain, we still get the same perfect correlation that we observed when considering the whole ontologies, even though, in this case, MeSH had a larger *semantic dispersion*

than WordNet, which is coherent with the strictly biomedical scope of the former.

However, for the *cell* and *drug* domains, the independent analysis of individual structural features provides contradictory conclusions. This shows the limitations of using these structural features as individual ontology evaluation scores. For the *cell* domain, we observe that *breadth* related features would suggest that SNOMED-CT offers the most accurate modeling, whereas *depth* related features would indicate that MeSH is preferable. In this case, the *breadth* related features, which are absolute values quantifying the number of concepts at each taxonomic level, are very influenced by the larger cardinality of the *cell* taxonomy in SNOMED-CT (i.e., 1,136 concepts in SNOMED-CT vs. 501 in MeSH). As discussed previously, this biases the results towards the largest ontology, with independence on the actual dispersion. The *semantic dispersion*, on the contrary, provides a coherent aggregation of such features, which in this case favors MeSH over SNOMED-CT due to the larger differences between depth-related features.

A similar situation is observed for the *drug* domain: *breadth variance* is maximal for MeSH, whereas *depth variance* is maximal for SNOMED-CT. As above, the larger cardinality of the MeSH taxonomy of *drugs* in this case distorts the conclusions extracted from the *breadth* analysis, which is greatly influenced by such cardinality. Contrary, the *semantic dispersion*, which is not influenced by the taxonomy size thanks to its normalized calculations, shows a preference towards SNOMED-CT.

The conclusions of this empirical analysis, which are coherent with the theoretical assumptions discussed in previous sections, are:

- Individual structural features may provide contradictory conclusions due to the partial and non-normalized scopes.
- The *breadth* is more influenced by the cardinality of the ontology than the *depth* and, thus, it tends to provide higher values for the larger knowledge structures.
- The *semantic dispersion* is not influenced at all by the cardinality of the taxonomy and it is a direct function of the unbalancing of the taxonomic structure, which is a function of both the *depth* and the *breadth*; thus, it better serves as an integrated, intuitive and normalized score for ontological accuracy evaluation.

VI. Conclusion

In this paper we presented an intuitive measure to quantify the *semantic dispersion* of the taxonomic structure of a given ontology. Given that, according to previous empirical studies, this dispersion seems to be a good predictor of the ontology accuracy, the proposed measure can be used as an automatic and integrated ontology evaluation score.

Unlike other works on ontology evaluation, which propose ad-hoc and weighted aggregations of heterogeneous features, our measure is a semantically and mathematically coherent one, in that it builds on the standard notion of numerical dispersion (i.e., the usual standard deviation of a sample) and on the well-established foundations of semantic similarity/distance assessment to capture the semantics inherent to the knowledge model led in the ontology.

Given that our measure provides normalized values, it enables a coherent comparison of ontologies with significantly different cardinalities. Moreover, its applicability is ensured by the fact that it focuses only on the taxonomic features that, contrary to other non-taxonomic features, are available in all ontologies.

Acknowledgment

The authors are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization.

This work was partly supported by the European Commission under the FP7 project “Inter-Trust”, by the Spanish Government (through projects ICWT TIN2012-32757, ARES-CONSOLIDER INGENIO 2010 CSD2007-00004 and BallotNext IPT-2012-0603-430000) and by the Government of Catalonia through grant 2014 SGR 537.

This work was also made possible through the support of a grant from Templeton World Charity Foundation. The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of Templeton World Charity Foundation.

References

[1] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities," *Scientific American*, vol. 284, pp. 34-43, 2001.

[2] N. Guarino, "Formal Ontology in Information Systems," in *1st International Conference on Formal Ontology in Information Systems, FOIS 1998*, Trento, Italy, 1998, pp. 3-15.

[3] P. Castells, M. Fernández, and D. Vallet, "An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval " *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, pp. 261-272, 2007.

[4] D. C. Wimalasuriya and D. Dou, "Ontology-based information extraction: An introduction and a survey of current approaches," *Journal of Information Science*, vol. 36, pp. 306-323, 2010.

[5] D. Sánchez, D. Isern, and M. Millán, "Content Annotation for the Semantic Web: an Automatic Web-based Approach," *Knowledge and Information Systems*, vol. 27, pp. 393-418, 2011.

[6] M. Batet, "Ontology based semantic clustering," *AI Communications*, vol. 24, pp. 291-292, 2011.

[7] S. Martínez, D. Sánchez, and A. Valls, "A semantic framework to protect the privacy of electronic health records with non-numerical attributes," *Journal of Biomedical Informatics*, vol. 46, pp. 294-303, 2013.

[8] D. Sánchez, M. Batet, and A. Viejo, "Automatic general-purpose sanitization of textual documents," *IEEE Transactions on Information Forensics and Security*, vol. 8, pp. 853-862, 2013.

[9] D. Sánchez, M. Batet, and A. Viejo, "Utility-preserving sanitization of semantically correlated terms in textual documents," *Information Sciences*, vol. 279, pp. 77-93, 2014.

[10] L. Ding, *et al.*, "Swoogle: A Search and Metadata Engine for the Semantic Web," in *thirteenth ACM international conference on Information and knowledge management, CIKM 2004*, Washington, D.C., USA, 2004, pp. 652-659.

[11] A. Zouag and R. Nkambou, "A survey of domain ontology engineering: Methods and tools," in *Advances in Intelligent Tutoring Systems*, ed, 2010, pp. 103–119.

[12] E. Simperl, M. Mochol, T. Bürger, and I. Popov, "Achieving maturity: the state of practice in ontology engineering in 2009," in *On the Move to Meaningful Internet Systems: OTM 2009*, 2009, pp. 983–991.

[13] D. Vrandecic, S. Pinto, C. Tempich, and Y. Sure, "The diligent knowledge processes," *Journal of Knowledge Management*, vol. 9, pp. 85–96, 2005.

[14] O. Corcho, M. Fernández-López, A. Gómez-Pérez, and A. López-Cima, "Building legal ontologies with methontology and webode," in *Law and the semantic web*, Springer-Verlag, Ed., ed, 2005, pp. 142–157.

[15] M. Fernández, C. Overbeeke, M. Sabou, and E. motta, "What Makes a Good Ontology? A Case-Study in Fine-Grained Knowledge Reuse," in *4th Asian Conference on The Semantic Web*, 2009, pp. 61-75.

[16] D. Vrandecic, "Ontology Evaluation," in *Handbook on Ontologies*, ed: Springer Berlin Heidelberg, 2009, pp. 293-313.

[17] S. Tartir, B. Arpinar, and A. P. Sheth, "Ontological Evaluation and Validation," in *Theory and Applicaitons of Ontology*, ed: Springer, 2010, pp. 115-130.

[18] A. Gómez-Pérez, M. Fernández-López, and O. Corcho, *Ontological Engineering*, 2nd ed.: Springer-Verlag, 2004.

[19] G. Pirró, "A semantic similarity metric combining features and intrinsic information content," *Data & Knowledge Engineering*, vol. 68, pp. 1289-1308, 2009.

[20] S. Martínez, A. Valls, and D. Sánchez, "Semantically-grounded construction of centroids for datasets with textual attributes," *Knowledge-Based Systems*, vol. 35, pp. 160-172, 2012.

[21] S. Harispe, D. Sánchez, S. Ranwez, S. Janaqi, and J. Montmain, "A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain," *Journal of Biomedical Informatics*, vol. 49, pp. 38-53, 2014.

[22] R. Rada, H. Mili, E. Bichnell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, pp. 17-30, 1989.

[23] M. Batet, D. Sánchez, and A. Valls, "An ontology-based measure to compute semantic similarity in biomedicine," *Journal of Biomedical Informatics*, vol. 44, pp. 118-125, 2011.

[24] D. Sánchez, M. Batet, D. Isern, and A. Valls, "Ontology-based semantic similarity: A new feature-based approach," *Expert Systems with Applications*, vol. 39, pp. 7718-7728, 2012.

[25] B. Lemaire and G. Denhière, "Effects of High-Order Co-occurrences on Word Semantic Similarities," *Current Psychology Letters - Behaviour, Brain and Cognition*, vol. 18, p. 1, 2006.

[26] M. Batet, A. Valls, and K. Gibert, "A distance function to assess the similarity of words using ontologies," in *Proceeding of the XV congreso español sobre tecnologías y lógica fuzzy, ESTYLF 2010*, Huelva, Spain, 2010, pp. 561-566.

[27] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Cambridge, Massachusetts: MIT Press, 1998.

[28] M. Batet, D. Sanchez, A. Valls, and K. Gibert, "Semantic similarity estimation from multiple ontologies," *Applied Intelligence*, vol. 38, pp. 29-44, 2013.

[29] A. Devitt and C. Vogel, "The topology of WordNet: Some Metrics," in *2nd Global Wordnet Conference, GWC 2004*, Brno, Czech Republic, 2004, pp. 106-11.

[30] K. Spackman, "SNOMED CT milestones: endorsements are added to already-impressive standards credentials," *Healthcare Informatics*, vol. 21, pp. 54-56, 2004.

[31] S. J. Nelson, D. Johnston, and B. L. Humphreys, "Relationships in Medical Subject Headings," in *Relationships in the Organization of Knowledge*, ed: K.A. Publishers, 2001, pp. 171-184.

[32] D. Sánchez and M. Batet, "A semantic similarity method based on information content exploiting multiple ontologies," *Expert Systems with Applications*, vol. 40, pp. 1393-1399, 2013.