Semantic variance: an intuitive measure for ontology accuracy evaluation

David Sánchez[1], Montserrat Batet, Sergio Martínez, Josep Domingo-Ferrer

UNESCO Chair in Data Privacy, Department of Computer Science and Mathematics, Universitat Rovira i Virgili, Av. Països Catalans, 26, 43007 Tarragona, Catalonia (Spain)

Abstract

Ontology evaluation is a relevant issue in the field of knowledge representation. It aims at quantifying the quality of ontologies, so that potential users can have an idea of their accuracy and thereby select the most appropriate ontology for a specific application. Many of the ontology evaluation methods and frameworks available in the literature assess the quality of ontologies according to their structural features, even though most of these methods propose *ad hoc* aggregations of such features that lack a theoretical basis. Inspired by recent empirical studies showing that some structural features are better suited to predict the semantic accuracy of ontologies, we present in this paper the notion of *semantic variance* of an ontology. Semantic variance is an intuitive and inherently semantic measure to evaluate the accuracy of ontologies. Unlike *ad hoc* methods, our proposal is a mathematically coherent extension of the standard numerical variance to measure the semantic dispersion of the taxonomic structure of ontologies. In our experiments performed over a set of widely used ontologies, the proposed *semantic variance* positively correlated with the structural features of ontologies that best predicted their accuracy in previous studies. Moreover, our measure also provided a good prediction of the ontological accuracy in one of the most essential knowledge-based tasks: assessing the semantic similarity between concepts. These results suggest that the *semantic variance* can be used as a generic, quantitative and theoretically coherent score to evaluate the accuracy of ontologies.

Keywords: Ontologies, Ontology evaluation, Semantic similarity.

[1] Corresponding author. Address: Department of Computer Engineering and Mathematics. Universitat Rovira i Virgili. Avda. Països Catalans, 26. 43007. Tarragona, Catalonia (Spain)
Tel.: +34 977 559657; Fax: +34 977 559710;
E-mail: david.sanchez@urv.cat.

1. Introduction

In the last decade, ontologies have experienced an enormous development motivated by the growing interest in exploiting the contents of the World Wide Web, which is driven by global initiatives like the Semantic Web [1]. Ontologies provide a formal representation of a shared conceptualization by means of classes, instances, properties and semantic relationships [2]; that is, they offer a structured and machine-readable representation of the semantics related to one or several domains of knowledge. Ontologies have in fact become the cornerstone of many knowledge-based engineering applications that require managing and interpreting data (usually text) from a semantic perspective. Just to name a few, ontologies are extensively used to improve information retrieval [3] and extraction [4-6], resource classification [7, 8], to achieve interoperability between heterogeneous systems [9] or to interpret textual data semantics in areas such as automatic reasoning [10, 11] or data privacy [12-14]. Moreover, ontologies are the key to bring semantic content to electronic textual resources under the umbrella of the Semantic Web [1].

Due to their inherent interest, thousands of ontologies have been developed and made available through the Web that cover a variety of overlapping topics and domains [15]. However, the lack of consensus in ontology development [16, 17] raises the following question: which ontology, among those covering a certain domain of knowledge or modeling the same concepts, is the best one for a specific task? Ontology evaluation and, more specifically, our work here try to answer the previous question.

1.1. Background on ontology evaluation

Ontology evaluation specifically aims at measuring the quality of ontologies, either to provide feedback to ontology developers and knowledge engineers or to give insights on the adequacy of ontologies to their users [18, 19]. The dimensions considered in ontology evaluation are, however, quite diverse. In [20] a characterization of the quality criteria usually employed in the ontology evaluation literature is provided in terms of *accuracy* of the semantics represented by the ontology, *adaptability* to different contexts, *clarity* for human readers, *completeness*, *conciseness* and logical *consistency*. From the perspective of automated knowledge-based systems, *semantic accuracy*, which captures the semantic coherency and suitability of the definition and description of ontological components (i.e., classes, relationships, etc.), is the most important dimension because it directly influences the precision of the semantic assessments or inferences [21]. More specifically, since semantics are inherently human features, semantic accuracy is seen as the ability to properly represent semantics as understood

by society, because this will enable knowledge-based systems to better mimic human reasoning, which is usually the main goal of such systems.

Ontology evaluation can be performed i) manually, which requires a human expert to measure the quality of the ontology [22], ii) automatically, in which case the ontology quality is measured according to a specific criterion (i.e., a Golden Standard [23]), or iii) oriented to a specific task, in which case the ontology quality is a function of the accuracy of the results provided by the application in a specific task [24]. However, on the one hand, manual approaches are hardly scalable given the large amount of ontologies that are currently available, whereas a task-oriented evaluation may only provide a biased assessment of ontology quality if the task is not generic enough. Automatic methods, on the other hand, are more generic and scalable [25, 22], but they require defining appropriate evaluation criteria.

If we focus on the evaluation of the ontological accuracy, which is the aim of this paper, most approaches rely on the analysis of structural features of ontologies [21, 26, 20]. Indeed, since the same knowledge (i.e., a set of specific concepts belonging to a domain) can be modeled in many different (more or less accurate) ways, and ontologies model concepts as semantic graphs, the resulting graph structures (i.e., number of classes, relationships, taxonomic depth, branching factor, etc.) can vary highly for ontologies describing the *same* knowledge [20]. Evaluation approaches based on these features assume that, given a set of manually constructed or validated ontologies (i.e., with a minimum structural coherency), their structure may give insights on the accuracy of the modeled knowledge. The advantages of this kind of analysis are: i) it provides a numerical, unambiguous and objective characterization of the ontology, ii) it can be easily and efficiently calculated from the graph defined by the semantic links modeled in the ontology, iii) its assessment does not depend on potentially subjective, biased or domain-dependent external resources such as other knowledge sources or experts' criteria. However, the analysis of the current literature [27, 28, 21, 29-31] also reveals the following drawbacks: i) individual structural features (e.g., maximum taxonomic depth, number of classes, etc.) only provide a partial picture of the ontology quality, so different features are usually aggregated using *ad hoc* measures and weights, a solution that lacks a sound theoretical basis and, hence, generality; ii) many methods rely on non-taxonomic structural features (e.g., non-taxonomic relationships, attributes, logical restrictions, etc.), which are hardly found in available ontologies (most of them only model taxonomic relationships [15]); and iii) usually, it is not clear which features are the most suitable ones to quantify the semantic accuracy of ontologies. A recent study [21] has shed some light on the latter issue by identifying how certain structural features can be used to reasonably predict the semantic accuracy of ontologies. Specifically, the authors manually evaluated the semantic correctness of the relationships modeled in a set of ontologies obtained

by means of ontology alignment techniques and manually validated by knowledge experts; as a result, the set of ontologies were classified as reliable or non-reliable. Then, the authors studied which structural characteristics could be used to identify the reliable ontologies. The authors mainly focused on taxonomic features (e.g., number of classes, taxonomic depth, breadth, etc.), which are the kind of features that are available in most ontologies. They concluded that the maximum, average and variance of the taxonomic depth, and the maximum and variance of the taxonomic breadth (i.e., number of concepts at each taxonomic level) seem the best predictors of the semantic accuracy of ontologies: the larger these features are, the higher the probability that the ontology is a reliable one. However, since there is no unique characteristic that can be used to assess the accuracy of an ontology and individual feature values may be contradictory to each other, an *ad hoc* aggregation or selection of features based on empirical hypotheses is still needed. 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, such structural characteristics are used as thresholds for binary classifiers (i.e., reliable or non-reliable ontologies) and not as a continuous quality score. Yet the latter score is the most suitable tool for ontology evaluation, as it enables selecting the *best* ontology from a set of candidates modeling the same knowledge.

1.2. Contributions and plan of this paper

In this paper we present and formalize the notion of *semantic variance* of an ontology as an evaluation measure to quantify the ontology's semantic accuracy. In contrast to the *ad hoc* nature of most of the available ontology evaluation approaches, our proposal aims at measuring the semantic dispersion of an ontology by means of a mathematically coherent interpretation of the classical numerical variance of a sample, but applied to the knowledge structure of the ontology. Moreover, it solely relies on taxonomic knowledge, which is what all available ontologies have in common [15]. By implicitly capturing the unbalance of the taxonomic structure of the ontology, our measure aims at providing a semantically coherent and numerically normalized quantification of the ontological dispersion, which does not depend on the cardinality of the ontology. We will also show that such dispersion correlates with the structural features that, as discussed in previous studies [21], seem to be related to the accuracy of the ontology. Thus, the *semantic variance* measure could be used as a unique, generic, non-parameterized and quantitative score to evaluate the accuracy of ontologies modeling the same domain. Last but not least, this score is intuitive, theoretically coherent and easy to implement.

The theoretical hypotheses of this work are also empirically validated by showing the positive correlation between the relevant structural characteristics of several well-known ontologies and

the proposed notion of *semantic variance*. Moreover, the suitability of the *semantic variance* as a tool to select the most accurate ontology from a set of ontologies covering the same domain of knowledge is also evaluated by computing the performance of these ontologies in one of the most basic and relevant ontology-based tasks: assessing the semantic similarity between concepts [32].

The rest of this paper is organized as follows. Section 2 introduces and formalizes the notion of *semantic variance* and illustrates its suitability as a measure of semantic dispersion and its relationship with some relevant structural features through several examples. Section 3 details the empirical experiments performed with several well-known ontologies and shows how our measure can be of use as an intuitive score for ontology accuracy evaluation. Section 4 describes other applications of the *semantic variance*, such as measuring economic diversity, measuring biodiversity and also quantifying the protection of confidential information in database anonymization. The final section gathers some conclusions.

2. Semantic variance

As mentioned above, the empirical results reported in [21] suggest that the structural features that, individually, can be used to predict the ontologies with the best reliability are:

- *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), *average depth* (average length of all taxonomic branches) and *depth variance* (dispersion with respect to the average depth, computed as the standard mathematical variance).
- *Maximum breadth* (width of the taxonomic level of the ontology with the largest number of concepts) and *breadth variance* (dispersion with respect to the average breadth).

The authors of that study argue that, among the above features, the depth and breadth variances are the best predictors (i.e., the larger they are, the higher the probability of the ontology being reliable). In contrast, other structural features commonly used in the ontology evaluation literature, such as the total number of concepts or the number of properties, do not show a significant relationship with the ontological accuracy.

These results suggest that, in general, the dispersion and unbalance of the taxonomic structure of an ontology, which are mainly reflected by its variance in depth and breadth and are limited by the maximum depth/breadth of the taxonomy, are positively correlated with the semantic

accuracy of the knowledge modeled in the ontology. This also suggests that a semantically accurate representation of the knowledge of a domain is rarely achieved by a homogeneous taxonomic structure. This is coherent with the inherent nature of knowledge representation, which is an *ex post* formalization of the *de facto* semantic consensus of human society [33]. Since such semantics unpredictably and informally evolve as societies develop their knowledge, the posterior formalization is unlikely to produce homogenous and balanced knowledge structures. Moreover, a very homogeneous knowledge structure indicates that concepts are evenly distributed through the taxonomy; thus, due to this homogeneity, they become less distinguishable from each other according to that structure. This goes against the main goal of knowledge modeling, which is making concepts well differentiated from each other in order to minimize the ambiguity of semantic inferences obtained from the modeled knowledge [34]. Paradoxically, a problem of the above-detailed features is that they are highly dependent on the size of the ontology (especially for the maximum and average depth and breadth). Hence, large ontologies will tend to systematically provide larger values than smaller ones. The question is, how can we measure in an integrated and semantically and mathematically coherent way the degree of unbalancing or *semantic dispersion* of an ontology so that it can be used as a quantitative evaluation measure of ontological accuracy? To answer this question, we propose the notion of *semantic variance* of an ontology, which is inspired by the usual notion of numerical sample variance.

Within a numerical domain, the usual *variance* is used to quantify the dispersion of a sample of values with regard to the center (mean) of that sample. It is computed as the average squared difference (or distance) between each element s_i of a sample S and the mean of the sample [35] :

$$Var(S) = \frac{\sum_{s_i \in S} (s_i - \overline{S})^2}{|S|} \quad (1)$$

The square makes differences positive and increases the contribution of the most scattered values. In the extreme cases, a zero variance indicates that all values are identical, whereas a high variance indicates that values are very spread out from the center and from each other, hence being clearly distinguishable.

In essence, this numerical variance is what we aim to mimic in the context of ontologies: we want to capture the degree of dispersion of concepts modeled in a given ontology with respect to the center of the knowledge structure of that ontology. Ideally, this should not depend on the size of the ontology, just like the numerical variance does not depend on the size of the sample.

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 generalizes the meaning of all the other concepts, which are its specializations. Note that, even if the ontology incorporates several disjoint taxonomic trees, all of them can be joined by a virtual root that subsumes all of them, as it is done in most knowledge repositories (e.g., *entity* in WordNet, *Concept* in SNOMED-CT, *Top* in ODP, *thing* in OWL ontologies, etc.). In a perfectly balanced taxonomy, in which all branches have the same depth and the branching factor is constant, this root node coincides with the geometric center of the graph defined by the taxonomy; thus, this central node is the one that minimizes the distances with respect to all the concepts in the ontology [36]. In contrast, when taxonomic branches present different depths (i.e., higher *depth variance*, which is limited by the *maximum depth*) and branching factors (i.e., higher *breadth variance*, which is limited by the *maximum breadth*), the unbalance of the ontology with respect to the root node (i.e., the center in a perfectly balanced taxonomy) increases. Thus, by measuring the degree of unbalance, we are quantifying the dispersion of concepts in the ontology that, as discussed earlier, is a direct function of the structural characteristics that are related to the ontological accuracy. We name this notion of semantic dispersion *semantic variance* of an ontology. Consistently with the numerical case, we define it as follows:

Definition (Semantic variance)*: Given an ontology O, which models in a taxonomic way a set of concepts C, the* semantic variance *of O is computed as the average of the squared semantic distance d(·,·) between each concept* $c_i \in C$ *in O and the taxonomic Root node of O. If we denote by |C| the cardinality of C excluding Root(O), the mathematical expression of the semantic variance of O is:*

$$Semantic\_Variance(O) = \frac{\sum_{c_i \in C} d(c_i, Root(O))^2}{|C|} \quad (2)$$

Note that we include all the concepts in the ontology in the variance calculation because, even though *depth*-related features only refer to the leaves of the taxonomy, *breadth*-related features also consider inner nodes.

By means of the proposed *semantic variance* and given a set of ontologies modeling the same domain, the *a priori* most accurate ontology can be selected as the one with the maximum variance. The proposed measure can also be applied only to particular taxonomic branches of a set of ontologies. This is relevant because the different scopes and goals for which ontologies are designed may not allow comparing ontologies as a whole, but only those taxonomic branches modeling the same domains. With our method, this partial comparison can be done by

using the common generalization of that branch as the root, and by computing the distances towards all of its taxonomic specializations. In this manner, we are not only able to evaluate and select entire ontologies, but also to measure if a specific taxonomic branch of an ontology provides a better differentiation of concepts in that branch than its counterpart branch in another ontology.

2.1. Computing the semantic distance

The key element to measuring the *semantic variance* as defined in Expression (2) is the calculation of the *semantic distance* $d(\cdot,\cdot)$ between each concept in the ontology and the root node of the taxonomy. On the one hand, a suitable distance measure ought to accurately capture the semantic differences between concepts. On the other hand, in order to coherently compare the variances of ontologies with different number of concepts, the distance values should not depend on the cardinality of the ontology.

Within the literature of computational linguistics, a plethora of approaches have been proposed to measure the semantic distance by exploiting the knowledge modeled in the ontology [37]. The simplest methods evaluate ontologies as directed graphs in which the distance between two concepts is measured as the number of edges of the shortest path between them [38]. Since there is no normalization, distance values tend to be larger as the ontology size increases. Thus, if we use this approach to compute the *semantic variance*, large ontologies will systematically yield larger variance values. Another drawback of edge-counting measures is their relatively low accuracy, which is motivated by the fact that they only consider the shortest path connecting the two concepts; in ontologies in which concept pairs are connected by several paths, a lot of explicit knowledge is omitted, which negatively influences the similarity assessment accuracy [39].

To solve the drawbacks of edge-counting measures, feature-based approaches are proposed. They compare concepts according to the amount of semantic evidences that they have and do not have in common. In [39, 32], 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:

$$d(c_1,c_2)=\log_2\left(1+\frac{\left|T(c_1)\cup T(c_2)\right|-\left|T(c_1)\cap T(c_2)\right|}{\left|T(c_1)\cup T(c_2)\right|}\right) \tag{3}$$

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

From a semantic perspective, the measure of Expression (3) captures more taxonomic knowledge than edge-counting methods, since it implicitly considers *all* the paths connecting the two concepts. 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 calculation better aggregates the semantic evidences gathered from the ontology (i.e., 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 [40]. As a result, Expression (3) approximates human judgments of similarity better than other ontology-based measures, as demonstrated for several standard evaluation benchmarks [39, 32].

Further, in contrast with the absolute distance values provided by edge-counting methods, Expression (3) yields positive normalized values in the [0,1] range. Thus, the measure does not depend on the ontology size and is therefore suitable to measure and coherently compare *semantic variance*s of ontologies with different cardinalities. Finally, as demonstrated in [32] and [41], Expression (3) satisfies non-negativity, reflexivity, symmetry and the triangular inequality, thereby being a distance measure in the mathematical sense. This is relevant in order to apply the *semantic variance* as a mathematically coherent replacement of the standard numerical variance in algorithms or methods dealing with semantic values [42, 43].

2.2. Examples and discussion

When using Eqs. (2) and (3), we have that the minimum *semantic variance* is obtained with a perfectly balanced taxonomic structure, in which the root node matches the geometric center of the graph and in which all concepts in the taxonomy are structurally indistinguishable from each other (i.e., they are direct specializations of the root node). Figure 1 shows an example of such a structure in which four concepts extracted from WordNet [44] (*Orange, Clementine, Strawberry* and *Blackberry*) are modeled as direct specializations of the root node (*Fruit*). In practice, such taxonomy is equivalent to a flat list of concepts, which gives no insight about their semantic commonalities or differences, thus failing to offer proper knowledge modeling.

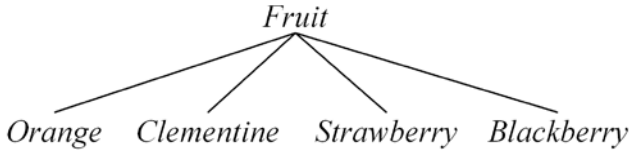

Figure 1. Sample ontology O_1

Structurally, in this extreme case, the *average* and *maximum depth* of the taxonomic branches is 1 and the *depth variance* is 0. Regarding the breadth, the *maximum breadth* is 4 and the *breadth variance* is 0, without considering the root node that is assumed to be unique (or virtually added) in all ontologies. The low values of these structural features suggest that, according to the empirical study in [21], the ontological accuracy is also low. This is coherent, since in a taxonomic structure as simple as the one shown in Figure 1, all concepts have the same level of specificity and their semantic distances are identical, which makes them structurally indistinguishable. In this case, the *semantic variance* of the ontology is:

$$Semantic\_Variance(O_1) = \frac{d(Orange, Fruit)^2 + d(Clem., Fruit)^2 + d(Straw., Fruit)^2 + d(Black., Fruit)^2}{4} =$$

$$= \frac{\left(\log_2\left(1+\frac{1}{2}\right)\right)^2 + \left(\log_2\left(1+\frac{1}{2}\right)\right)^2 + \left(\log_2\left(1+\frac{1}{2}\right)\right)^2 + \left(\log_2\left(1+\frac{1}{2}\right)\right)^2}{4} = 0.342$$

Next, let us consider the ontology represented in Figure 2, which shows an alternative representation of the same four concepts. In this case, an inner taxonomic level (*Citrus* and *Berry*) has been added to differentiate the pair of concepts *Orange* and *Clementine* from the pair *Strawberry* and *Blackberry*.

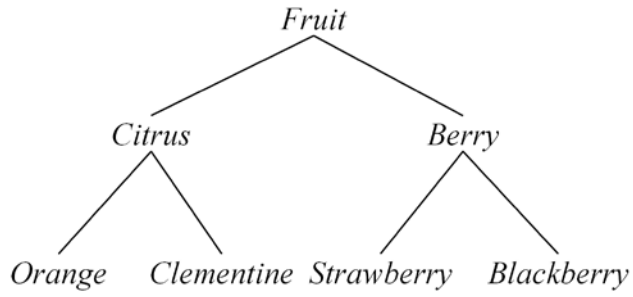

Figure 2. Sample ontology O_2

In this case we have that the *average* and *maximum depth* of the taxonomic branches is 2 and the *depth variance* is 0. Moreover, by considering all the nodes in the ontology except the root node, the *maximum breadth* is 4 (the average breadth is 3) and the *breadth variance* is $((2-3)^2+(4-3)^2)/2=1$. The structural measures of taxonomic dispersion are still low and the root is

still the perfect geometric center of the taxonomy, but both the *maximum depth* and the *breadth variance* have increased as a result of the better differentiation between concepts. The *semantic variance* of this ontology also increases accordingly:

$$Semantic\_Variance(O_2) = \left(\begin{array}{l} d(Orange, Fruit)^2 + d(Clem., Fruit)^2 + d(Straw., Fruit)^2 + \\ + d(Black., Fruit)^2 + d(Cit., Fruit)^2 + d(Ber., Fruit)^2 \end{array} \right) / 6 =$$

$$= \left(\begin{array}{l} \left(\log_2 \left(1 + \frac{2}{3} \right) \right)^2 + \left(\log_2 \left(1 + \frac{2}{3} \right) \right)^2 + \left(\log_2 \left(1 + \frac{2}{3} \right) \right)^2 + \\ + \left(\log_2 \left(1 + \frac{2}{3} \right) \right)^2 + \left(\log_2 \left(1 + \frac{1}{2} \right) \right)^2 + \left(\log_2 \left(1 + \frac{1}{2} \right) \right)^2 \end{array} \right) / 6 = 0.476$$

Finally, let us consider a third sample ontology shown in Figure 3, in which *Clementine* has been more differentiated from *Orange* by adding a new inner node (*Mandarin*). This is precisely the way in which these concepts are modeled in WordNet.

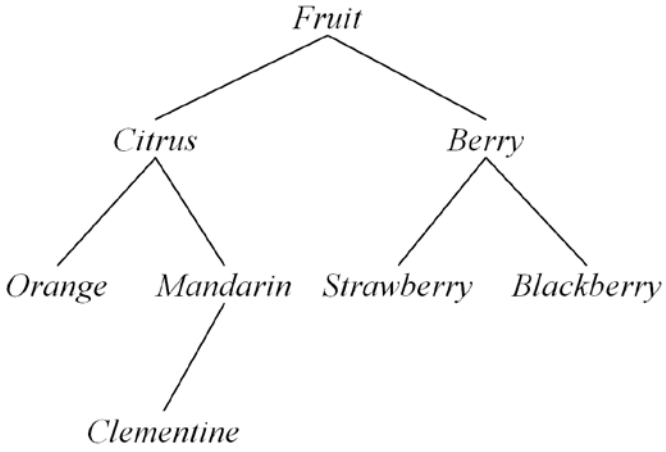

Figure 3. Sample ontology O_3

Due to the longer taxonomic branch defined by the concept *Clementine*, the *maximum depth* is now 3, the *average depth* is (2+3+2+2)/4=2.25 and the *depth variance* is $((2-2,25)^2+(3-2,25)^2+(2-2,25)^2+(2-2,25)^2)/4=0.8125$. On the other hand, the *maximum breadth* is still 4 (breadth average is now 2,33) and the *breadth variance* is $((2-2,33)^2+(4-2,33)^2+(1-2,33)^2)/3=1.55$. The increase in the *depth* and *breadth variances* suggests a less balanced structure, which is reflected by the fact that the root node is not the perfect geometric center of the tree anymore. The *semantic variance* also reflects this dispersion increase:

$$Semantic\_Variance(O_3) = \left(\begin{array}{l} d(Oran., Fr.)^2 + d(Clem., Fr.)^2 + d(Straw., Fr.)^2 + \\ + d(Black., Fr.)^2 + d(Cit., Fr.)^2 + d(Ber., Fr.)^2 + d(Mand., Fr.)^2 \end{array} \right) / 7 =$$

$$= \left(\begin{array}{l} \left(\log_2 \left(1 + \frac{2}{3} \right) \right)^2 + \left(\log_2 \left(1 + \frac{3}{4} \right) \right)^2 + \left(\log_2 \left(1 + \frac{2}{3} \right) \right)^2 + \\ + \left(\log_2 \left(1 + \frac{2}{3} \right) \right)^2 + \left(\log_2 \left(1 + \frac{1}{2} \right) \right)^2 + \left(\log_2 \left(1 + \frac{1}{2} \right) \right)^2 + \left(\log_2 \left(1 + \frac{2}{3} \right) \right)^2 \end{array} \right) / 7 = 0.501$$

The behavior of the *semantic variance* is driven by the semantic distance defined in Eq. (3) that, given a set of concepts/domain to be modeled, it increases when:

- The *maximum* and *average depth* and the *depth variance* increase: when taxonomic branches are longer, the number of disjoint ancestors between concept pairs also tends to increase, thereby capturing more evidences about their differences and resulting in larger distances and larger *semantic variance*. For a fixed number of inner taxonomic nodes, a greater depth variance also increases the *semantic variance*, since the most distant concepts to the root node (i.e., those in the longer branches) dominate the calculation over the closest ones, due to the squared distances. Moreover, a large depth variance also allows more degrees of freedom for the distance calculation and, thus, distances become more fine-grained and less homogeneous.
- The *maximum breadth* and the *breadth variance* increase: again, such features depend directly on the degree of unbalance between the sets of taxonomic ancestors (i.e., inner nodes of the taxonomy), which are evaluated by the semantic distance and, thus, reflected in the *semantic variance*. In fact, a larger breadth variance will make concepts less evenly distributed, thereby increasing the influence in the calculation of the most scattered ones due to the squared distances.

More specifically, given a set of concepts to be modeled, and starting from the trivial structure shown in Figure 1, any possible change of the taxonomic structure in order to better differentiate concepts will either i) maintain the geometric balance (i.e., root node as the center) but add more inner nodes that contribute to increasing the *maximum* and *average depth* and the *breadth variance* (like in Figure 2)), or ii) introduce a degree of unbalance, which will increase the *depth variance* and/or the *breadth variance* (like in Figure 3). In both cases, the *semantic variance* will increase. Hence, the *semantic variance* quantifies the degree of deviation from the trivial structure in which concepts are structurally indistinguishable. Furthermore, it is proportional to the degree of semantic dispersion and, thus, of semantic distinguishability, of concepts modeled in the ontology, which is desirable from the perspective of knowledge engineering [34]. Other interesting characteristics of the *semantic variance* are: i) due to the squared distance calculation of the standard variance on which it relies, it increases as the

unbalance of the taxonomy increases, because the increasing distance of the farthest concepts to the root node will have a greater influence in the calculation; ii) the semantic distance measure $d(\cdot,\cdot)$ aggregates features in a logarithmic way, which better correlates with the non-linear nature of semantic evidences [40]; iii) the *semantic variance* does not depend on the cardinality of the ontology, because both the aggregation of distances and the distance itself are normalized with respect to the number of concepts in the ontology; and iv) it yields values bounded in the range [0..1] (which is the range of the distance $d(\cdot,\cdot)$), thus facilitating the comparison of variance values computed from different ontologies.

3. Empirical results

In this section we detail the results of a set of experiments aimed to illustrate:

1) How the *semantic variance* aggregates and positively correlates with the structural features that, as discussed above, help to identify semantically accurate ontologies.
2) How the *semantic variance* can be used as a quantitative ontology evaluation score to select the most accurate ontology from a set of ontologies modeling the same domain, and how this predicted accuracy is reflected, in practice, in the results achieved by an essential knowledge-based task.

In the experiments, we used three well-known ontologies: WordNet, as a domain-independent repository, and SNOMED-CT and MeSH as domain-specific ontologies modeling biomedical concepts. All three ontologies have been widely used to evaluate semantic measures [45, 34, 32] and to guide knowledge-based systems [12, 13].

WordNet [44] is a freely available lexical database that describes more than 80,000 general concepts, which are semantically structured in an ontological way. The taxonomic structure in WordNet corresponding to nouns is very comprehensive and represents most of the semantic relationships modeled in the ontology [46]. The root node of the taxonomy, which is used in the calculations of the *semantic variance* is *entity*. We used WordNet version 2.1 in all the experiments.

The Systematized Nomenclature of Medicine, Clinical Terms (SNOMED-CT) [47] is one of the largest sources included in the Unified Medical Language System (UMLS) of the U.S. National Library of Medicine. It contains around 300,000 concepts organized into 18 overlapping hierarchies. Concepts in SNOMED-CT typically present a high degree of multiple inheritance (i.e., they may have multiple ancestors at the same taxonomic level) and are linked with

approximately 1.36 million relationships. The root node that subsumes all the hierarchies and which is used in our calculations is *SNOMED-CT Concept*. We used the July 2013 release of SNOMED-CT in our experiments.

The Medical Subject Headings (MeSH) [48] is a hierarchy of medical and biological terms defined by the U.S National Library of Medicine to catalogue books and other library materials, and to index articles for inclusion in health-related databases including MEDLINE. 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, consistently with the premises of our work, we added the abstract *MeSH concept* as virtual root subsuming all of categories to be used in our calculations. The 2011 release of MeSH was used in our experiments.

3.1. Relationship between ontological features and semantic variance

Table 1 shows an overview of the structural features of the three ontologies with respect to their *depth* and *breadth* features, and the *semantic variance* obtained with Eq. (2). We can see that, for all structural features, there is a positive correlation with the *semantic variance*. In fact, the Spearman rank order correlation coefficient is 1 in all cases, since the relative ordering according to each feature and the *semantic variance* match perfectly. Quantitatively, the relationship between individual structural features and the *semantic variance* is not linear (i.e., differences in the *semantic variances* of the different ontologies are much lower than for the structural features). Indeed, the structural features closely depend on the cardinality of the ontology, thus making it difficult to compare ontologies with significantly different sizes. Moreover, in [21] structural values were not considered as final ontology evaluation scores but just evidences to distinguish reliable ontologies. In contrast, the *semantic variance* yields values normalized in the range [0..1] that do not depend on the cardinality of the ontology, but just on the dispersion of the taxonomic structure. In the second part of this experimental study, we will analyze the behavior of the *semantic variance* as a quantitative ontology evaluation score.

Table 1. Structural features and *semantic variances* of WordNet, SNOMED-CT and MeSH.

Ontology	*Max. depth*	*Avg. depth*	*Depth variance*	*Max. breadth*	*Avg. breadth*	*Breadth variance*	***Semantic Variance***
WordNet	17	8.3	3.38	17,837	4,790	32,474,163	**0.849**
SNOMED-CT	19	7	3.89	60,778	15,674	375,721,551	**0.880**
MeSH	12	5.4	1.17	7,775	2,012	6,226,389	**0.815**

The comparison of the three ontologies shows that SNOMED-CT has the largest *semantic variance*, followed by WordNet and, finally, MeSH. However, the scopes of the three ontologies do not allow a fair comparison of the semantic accuracy because they model different domains: SNOMED-CT and MeSH are both biomedical knowledge repositories with a high degree of overlap [49], whereas WordNet is a general-purpose repository that models many different domains. Thus, one would expect SNOMED-CT and MeSH to be more accurate than WordNet with regard to the modeling of medical concepts. In order to fairly compare the dispersion of three structures within the same domain, as introduced in Section 2 we computed the *semantic variance* of taxonomic branches of the three ontologies whose scopes match. To do so, we extracted the taxonomic branch that corresponds to a physical *disease* from the three ontologies. In SNOMED-CT, this corresponds to the set of specializations of the *Disease (disorder)* concept, which acts as the root node for the semantic variance calculation; for MeSH it corresponds to the third main taxonomy, *C-Disease*, and for WordNet it corresponds to the tree under the *ill health* concept. Table 2 shows the values of the structural features and the *semantic variance* for these taxonomic structures.

Table 2: Structural features and *semantic variances* of the *disease* taxonomic structure of WordNet, SNOMED-CT and MeSH.

Ontology (disease)	*Max. depth*	*Avg. depth*	*Depth variance*	*Max. Breadth*	*Avg. Breadth*	*Breadth variance*	***Semantic Variance***
WordNet	8	4.35	1.45	353	142	13,007	**0.722**
SNOMED-CT	11	5.12	1.45	17,657	6,026	40,916,395	**0.905**
MeSH	8	4.07	0.71	1,650	562	373,333	**0.785**

The pairwise relationship between *breadth* and *semantic variance* shows again a perfect Spearman correlation. SNOMED-CT's *disease* taxonomy has the largest *semantic variance* but it is now followed by MeSH and WordNet. If we use the semantic variance as a measure of accuracy, it turns out that, with regard to the modeling of medical entities, the two biomedical ontologies (MeSH and SNOMED-CT) are more accurate than the general-purpose WordNet, even though WordNet as a whole presents a larger dispersion than MeSH.

On the other hand, the correlation between *depth* and *semantic variance* (and thus, between *depth* and *breadth*) is not positive in this case, even though the differences between depth-related values for the three taxonomic structures are relatively small. This mismatch shows the limitations of these structural features as individual ontology evaluation scores, which may yield contradictory assessments. In contrast, the *semantic variance* provides a semantically coherently aggregated score of taxonomic dispersion that is a function of both the *depth* and the *breadth*. In

this case, the larger differences in the *breadth variance* between the three structures dominate the *semantic variance*.

3.2. Task-oriented ontology evaluation

As discussed in the introduction, one of the goals of ontology evaluation is to facilitate the selection of the most suitable ontology for a particular task in those cases in which several ontologies modeling the same domain are available. On the other hand, task-oriented ontology evaluation consists in measuring the quality of an ontology according to the accuracy of the results that it yields for a specific ontology-based task [24]. In this section, we combine both lines in order to evaluate the suitability of the *semantic variance* as an ontology evaluation score that can be used to select the most suitable ontology (from a set of ontologies modeling the same domain) in one of the most essential ontology-based tasks: assessing the semantic similarity (or distance) between concepts.

Semantic similarity aims at mimicking the human reasoning when assessing the similarity or the distance between concepts mentioned in a context (e.g., a sentence, a document, a database, etc.). Thus, it constitutes a key tool for understanding textual resources. Most of the semantic similarity/distance measures available in the literature (such as those introduced in Section 3.1), exploit the knowledge modeled in one or several ontologies to obtain a numerical score for a given pair of concepts [45, 37, 49]. Thus, semantic similarity assessment is one of the most general ontology-based tasks and it is the cornerstone of many applications such as document classification [50], semantic disambiguation [51] or information retrieval [52]. The evaluation of the accuracy of semantic similarity assessment is usually tackled by comparing human judgments of similarity against the computerized scores for a set of term pairs [53]. The correlation between both assessments objectively quantifies the accuracy of a given semantic measure.

Different benchmarks exist in the literature to evaluate semantic measures. They consist of a set of term pairs with similarity ratings provided by a set of human experts. Given the scope of the ontologies considered in this empirical study, we focused on the Pedersen et al.'s benchmark [53], which has become the *de facto* evaluation standard within the biomedical domain [37, 51, 54]. It consists of 30 pairs of medical terms, whose similarity has been assessed by a group of experts of the Mayo Clinic in the range [1..4]. Table 3 lists the set of term pairs and the averaged experts' similarity scores, and indicates which pairs are included in any of the three ontologies considered in this study. Also, we marked in boldface those pairs modeled as *diseases* in the three ontologies, as in the second part of the previous experiment (Table 2).

Table 3: Set of 30 medical term pairs with averaged experts' similarity scores from the Pedersen et al. benchmark [53]. The last three columns state whether the term pair is included in WordNet, SNOMED-CT or MeSH, respectively. In boldface we represent those that specifically correspond to a physical disease and belong to each ontology.

Term 1	*Term 2*	*Similarity*	*WordNet*	*SNOMED-CT*	*MeSH*
Renal failure	Kidney failure	4.0	**Yes**	**Yes**	**Yes**
Heart	Myocardium	3.3	Yes	Yes	Yes
Stroke	Infarct	3.0	Yes	Yes	**Yes**
Abortion	Miscarriage	3.0	Yes	**Yes**	**Yes**
Delusion	Schizophrenia	3.0	Yes	**Yes**	Yes
Congestive heart failure	Pulmonary edema	3.0	No	**Yes**	**Yes**
Metastasis	Adenocarcinoma	2.7	Yes	**Yes**	**Yes**
Calcification	Stenosis	2.7	Yes	Yes	**Yes**
Diarrhea	Stomach cramps	2.3	No	Yes	No
Mitral stenosis	Atrial fibrillation	2.3	**Yes**	**Yes**	**Yes**
Chronic obstructive pulmonary disease	Lung infiltrates	2.3	No	No	No
Rheumatoid arthritis	Lupus	2.0	**Yes**	**Yes**	**Yes**
Brain tumor	Intracranial hemorrhage	2.0	**Yes**	**Yes**	**Yes**
Carpal tunnel syndrome	Osteoarthritis	2.0	**Yes**	**Yes**	**Yes**
Diabetes mellitus	Hypertension	2.0	**Yes**	**Yes**	**Yes**
Acne	Syringe	2.0	Yes	Yes	Yes
Antibiotic	Allergy	1.7	Yes	Yes	Yes
Cortisone	Total knee replacement	1.7	No	Yes	Yes
Pulmonary embolus	Myocardial infarction	1.7	No	**Yes**	**Yes**
Pulmonary fibrosis	Lung cancer	1.7	No	**Yes**	**Yes**
Cholangiocarcinoma	Colonoscopy	1.3	No	Yes	Yes
Lymphoid hyperplasia	Laryngeal cancer	1.3	No	**Yes**	**Yes**
Multiple sclerosis	Psychosis	1.0	Yes	**Yes**	Yes
Appendicitis	Osteoporosis	1.0	Yes	**Yes**	**Yes**
Rectal polyp	Aorta	1.0	No	Yes	No
Xerostomia	Alcoholic cirrhosis	1.0	Yes	**Yes**	**Yes**
Peptic ulcer disease	Myopia	1.0	**Yes**	**Yes**	**Yes**
Depression	Cellulitis	1.0	Yes	**Yes**	Yes
Varicose vein	Entire knee meniscus	1.0	No	Yes	Yes
Hyperlipidemia	Metastasis	1.0	Yes	**Yes**	**Yes**

Using this benchmark we configured the following evaluation scenario. First, we defined several data sets with the subsets of term pairs that can be found in the ontologies as a whole and those that were modeled as *diseases*. According to Table 3, we created the following data sets:

- *Dataset 1* contains the 27 medical term pairs of the benchmark that can be found both in SNOMED-CT and in MeSH, which correspond to those marked with "yes" in the fifth and sixth column of Table 3.
- *Dataset 2* contains the 16 medical term pairs that are physical diseases in SNOMED-CT and MeSH, which correspond to those marked with a boldface "yes" in the fifth and sixth column of Table 3.
- *Dataset 3* contains the 7 medical term pairs that are physical diseases in SNOMED-CT, MeSH and WordNet, which are marked with a boldface "yes" in the fourth, fifth and sixth column of Table 3.

After that, we computed the Pearson correlation between the set of human similarity ratings (reported in the third column of Table 3), and the set of semantic distance values obtained with the measure introduced in Expression (3) for the term pairs of the three data sets detailed above. In order to measure the distances, we used the different medical knowledge bases considered in this study: SNOMED-CT and MeSH as a whole, and the taxonomic branches corresponding to diseases of SNOMED-CT, MeSH and WordNet, as detailed in the previous section. Note that, since Expression (3) measures distance whereas human ratings in the benchmark quantify similarity (i.e., the opposite of distance), we needed to invert distances by changing their sign. As stated above, the correlation between both sets of values quantifies the accuracy of the semantic similarity assessment and, since we use the same similarity measure in all tests, it thus measures the quality of the associated ontology/taxonomy for this particular task. As a proof of the statistical significance of the correlation values obtained for each dataset and ontology/taxonomy, we also measured the *p*-value of the Pearson correlation; according to [55], a *p*-value below 0.001 is a proof of statistical significance under the strictest standards. Finally, we compared the correlation values with the *semantic variances* reported in the previous section for each ontology and taxonomic tree, so that we could assess whether the *semantic variance* was a good score for evaluating the quality of ontologies and, thus, whether it could be used to guide the selection of the most appropriate ontology for this particular task. The results of this experiment are shown in Table 4.

Table 4: Pearson correlation coefficients (and *p*-values in parentheses) for each data set and ontology/taxonomy between the experts' similarity scores in the Pedersen et al. benchmark [53] and the similarity assessed by Eq. (3). The last column shows the *semantic variance* of each ontology and taxonomy.

Ontology	*Correl. Dataset 1 (p-value)*	*Correl. Dataset 2 (p-value)*	*Correl. Dataset 3 (p-value)*	***Semantic variance***
SNOMED-CT	0.69 (*p*: 0.000068)	-	-	**0.880**
MeSH	0.65 (*p*: 0.00024)	-	-	**0.815**
WordNet disease	-	-	0.79 (*p*: 0.0346)	**0.722**
SNOMED-CT disease	-	0.83 (*p*: 0.000068)	0.9 (*p*: 0.0058)	**0.905**
MeSH disease	-	0.77 (*p*: 0.00048)	0.9 (*p*: 0.0058)	**0.785**

It is important to note that we refrained from including WordNet in the comparison of the *semantic variances* of the complete ontologies. As stated above, because WordNet is a general-purpose ontology whereas both SNOMED-CT and MeSH focus on biomedical terms, this comparison is not fair for WordNet, because its *semantic variance* reflects the dispersion of concepts belonging to many different domains and not only biomedical terms. On the other hand, the three ontologies were considered when evaluating only the taxonomic trees and term pairs corresponding to *diseases* because the scopes of the three structures are comparable.

We can see that in most cases there is a positive correlation between the accuracy of the semantic assessment and the *semantic variance* of the ontology: the greater the *semantic variance*, the more accurate is the assessment. In fact, the Spearman rank order correlation coefficient is 1 in all cases. Only with *Dataset 3* (i.e., the smallest one), SNOMED-CT and MeSH gave the same correlation (0.9) but different variances (0.905 and 0.785, respectively); however, this result is not as statistically significant (i.e., *p*-values above 0.001) as those obtained for the same ontologies with the larger *Dataset 2*, which also considered physical diseases and which showed a positive correlation between the accuracy of similarity assessments and the *semantic variance*. These results illustrate how the semantic similarity/distance assessment benefits from a knowledge structure that, due to its taxonomic unbalance (reflected by a larger *semantic variance*), differentiates concepts better. Certainly, large *depth* and *breadth variances*, which increase the *semantic variance* as shown in the previous experiments, provide more degrees of freedom to the semantic assessment because

similarity/distance values become less homogeneous and more fine-grained. This observation is also coherent with the empirical results reported in the literature on semantic similarity, in which the most accurate measures are usually those that best exploit the differences between concepts explicitly modeled in the ontological structures [32]. This also suggests that, given a certain semantic measure, the most appropriate ontology would also be the one that differentiates concepts best.

Furthermore, we can see that the numerical scales of the *semantic variance* and the Pearson correlation values are quite linearly proportional for each data set; this contrasts with the numerical ranges of the structural features analyzed in the previous section, which were much broader, because they depended on the cardinality of the ontologies. For example, as shown in Table 1, even though the values of the structural features of SNOMED-CT are much larger than those of MeSH (because SNOMED-CT has around 300,000 concepts whereas MeSH only incorporates around 22,000), their *semantic variances* are not that different (0.880 vs. 0.815) nor are the semantic assessment accuracies they achieve (0.69 vs. 0.66). As stated in Section 3, this behavior is the result of the normalized results provided by Eq. (3) and the non-linear integration of semantic features. In fact, as shown in Table 4, the *disease* taxonomy of SNOMED-CT yields a *semantic variance* greater than the whole taxonomy (0.905 vs. 0.880), which shows how, regardless the size of the evaluated structure, the *disease* taxonomy of SNOMED-CT is more spread out than the whole ontology. Moreover, as shown in Tables 2 and 4, while individual structural features (i.e., *depth* and *breadth*) led to contradictory conclusions on the accuracy of MeSH and WordNet when evaluating the *disease* taxonomy, the *semantic variance* positively correlates with the semantic similarity accuracy. These results suggest that the *semantic variance* can be used as a better predictor of the accuracy of an ontology in this task, thus allowing the selection of the most appropriate ontology from a set of overlapping ones. Lastly, for a more accurate prediction in specific domains or applications, the *semantic variance* calculation can be applied not only to complete ontologies but also to the specific taxonomic tree(s) in the ontology that model(s) the domain of interest, as illustrated in the tests related to the *disease* trees.

4. Other applications of the semantic variance

Even though this work focuses on the suitability of the *semantic variance* as a measure for ontology evaluation, it can be also applied to other contexts. Thanks to the mathematical coherency of its formulation (inspired in the usual numerical variance) and of the distance calculation (which fulfills the basic metric properties), our measure can be applied to quantify

the dispersion of a sample of semantic values. In this case, the "reference" root node should be replaced in Eq. (2) by the concept that acts as the actual centroid (mean) of the sample of textual values, which should be mapped to ontological concepts [36]. The *semantic variance* can thus act as a replacement of the standard numerical variance in algorithms and methods dealing with textual data. In this role, it is an alternative to other taxonomic variances proposed in the literature, like [56, 42, 57], with the advantage that the *semantic variance* is more similar to the numerical variance, it is better grounded in the notion of semantic similarity and it yields values normalized within the range [0..1]. Therefore, the semantic variance can also be used for the applications of the marginality-based variance described in [57] and [42], which we review in the next subsections.

4.1. Measuring economic diversity

As explained in [57], in official statistics companies are associated an attribute "Economic activity", which takes values in some hierarchical classification. Then, given a representative sample of companies in a country, let us consider the ontology "induced" by the sample, which is obtained as follows: a) prune those nodes in the hierarchical classification that do not lie in the path from the root to any leaf corresponding to a value in the sample; b) in case a value is repeated in the sample, consider the corresponding leaf as many times as the number of repetitions. The semantic variance of such an induced ontology is a measure of the country's economic diversity. For example, in Europe the NACE hierarchical classification of economic activities (standardized by the European Commission [58]) is used for economic activity. NACE is a hierarchy with up to four levels: from higher to lower, "Section", "Division", "Group" and "Class". Specifically, Section A stands for "Agriculture, hunting and forestry", Section B for "Fishing", Section C for "Mining and quarrying", Section D for "Manufacturing", etc. Clearly, a country focusing mostly on agriculture is less economically diverse than a country striking a good balance among the various activities. This idea is captured and quantified by the semantic variance: if a representative sample of companies from the former country is taken, the sampled companies will mostly fall in NACE Section A, whereas a representative sample from the latter country will include a good balance of companies in all sections; hence, the semantic variance of the ontology induced by the sample of the latter country will be higher than the semantic variance of the ontology induced by the sample of the former country.

4.2. Measuring biodiversity

Given a representative sample of plants and/or animals of a certain ecosystem, the semantic variance of the sample based on the taxonomy of plants/animals is a measure of the ecosystem's biodiversity. The details are analogous to the previous example on economic diversity.

4.3. Data anonymization

In database anonymization, the attributes in a database are classified as *identifiers* (to be suppressed before any release), *quasi-identifiers* and *confidential attributes*. Quasi-identifier attributes are those that, in combination, can be linked with external information to re-identify (some of) the respondents to whom (some of) the records in the database refer; this is called *re-identification disclosure*. Example quasi-identifier attributes are age, job, zip code, etc. Confidential attributes are those conveying sensitive information on the respondent (e.g. disease, political opinion, etc.). The *k*-anonymity privacy model [59] focuses on thwarting re-identification of respondents by modifying the quasi-identifier attributes before release so that any combination of their values is shared by at least *k* records.

Yet, if a group of *k* records sharing the same quasi-identifier values also have very similar or even the same values for a confidential attribute (e.g. they all suffer from AIDS), then an intruder does not need to re-identify his target within the group of *k* records: he knows that his target suffers from AIDS. This is called *attribute disclosure*.

To remedy the insufficient protection offered by *k*-anonymity against attribute disclosure, several extensions have been proposed, with *l*-diversity [60] being among the most popular. *l*-diversity requires that, for each confidential attribute in the released data, there exist at least *l* "well-represented" values in a group of records sharing a combination of quasi-identifier values. The simplest meaning of "well-represented" means just "different", but then the *l* values might not be different enough (e.g. imagine a confidential attribute "Disease" for which all values in the group are sexually transmitted diseases). The authors of *l*-diversity proposed other diversity measures, like entropy, etc., but none of them adequately captures the semantics of the confidential attribute values. The semantic variance is clearly useful here: *protection against attribute disclosure is sufficient only if the semantics of the values of each confidential attribute within each group of records sharing quasi-identifier values are different enough, or equivalently if their semantic variance is above a certain pre-selected threshold* (the threshold is a privacy parameter analogous to *l* in *l*-diversity).

Also, in order to protect sensitive data, semantic noise can be used to distort a sample of textual data (i.e., replacing original values by other similar ones taken from their semantic neighborhood in an ontology) in such a way that the noise is adapted to the sample dispersion rather than being fixed by the ontology; specifically, given an original value to be anonymized, the semantic variance could be used to determine the range of ontology concepts around the original value among which the anonymized value is to be randomly drawn, analogously to what it is done for numerical data [61].

It is important to note that, in the above described applications, the *semantic variance* of a sample will depend both on the dispersion of the values in the sample and also on the inherent dispersion of the ontology to which those values are mapped. In order to minimize the influence of the ontology structure in the calculation, the *semantic variance* of the ontology can be used as a normalizing factor for the *semantic variance* of the sample. In this manner, the variances of samples mapped to ontologies with different taxonomic dispersions will become more comparable.

5. Conclusions and future work

In this paper, we have presented the *semantic variance*, an intuitive measure to quantify the degree of semantic dispersion of the taxonomic structure of an ontology (or of a specific taxonomic tree within a larger ontology). Since according to previous empirical studies [21] this dispersion seems to be a good predictor of the ontological accuracy, the proposed *semantic variance* can be used as an automatic ontology evaluation score. Thus, if several ontologies with similar scopes are available, by evaluating them with the proposed measure we are able to select the (*a priori*) most accurate ontology.

Given that our measure is based on the analysis of structural features, it offers a numerical, unambiguous and objective characterization of the ontology, which is easy to implement, efficient to compute and does not depend on the judgment of potentially subjective experts [21, 26, 20]. Moreover, unlike other works based on structural features [27, 28, 21, 29-31], which mainly 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 variance and on the well-established foundations of semantic similarity/distance assessment [39, 32]. Unlike approaches based on individually analyzing structural features [21] and other variance measures [56, 42, 57], which provide absolute values that are greatly influenced by the cardinality of the ontology, the results provided by our measure are

normalized to the ontology size, and also constrained in the [0..1] range. This enables an intuitive and coherent comparison of ontologies with significantly different sizes. Finally, our measure can be applied to any ontology because it solely focuses on taxonomic relationships, which are available in all ontologies [15].

The empirical results obtained on a set of widely used ontologies support our theoretical hypotheses. On the one hand, the *semantic variance* positively correlates with the structural features that suggest a good ontological accuracy. On the other hand, our measure acts as an accurate predictor of the ontology quality in one of the most essential ontology-based tasks: the assessment of the semantic similarity between concepts.

As future work we plan to study other semantic evidences that could be incorporated into the assessment of the ontology accuracy. For example the coherency of the informativeness of the concepts as modeled in the ontology with respect to their actual usage could be also used as an indication of ontological accuracy. To do so, we can compare the informativeness of concepts in the ontology, which can be computed as a function of their specificity in the taxonomy [62, 63], with the informativeness of the same concepts computed from their distribution in corpora [64].

Acknowledgements and disclaimer

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 FP7 projects “DwB”, “Inter-Trust” and H2020 “CLARUS”), by the Spanish Government (through projects ICWT TIN2012-32757, CO-PRIVACY TIN2011-27076-C03-01 and BallotNext IPT-2012-0603-430000) and by the Government of Catalonia (under grant 2014 SGR 537). The last author is partially supported as an ICREA-Acadèmia researcher by the Government of Catalonia and by a Google Faculty Research Award. 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, 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 284 (2001) 34-43.

[2] N. Guarino, Formal ontology in information systems, In: N. Guarino (ed). Proc. of 1st International Conference on Formal Ontology in Information Systems, FOIS 1998, Trento, Italy, 1998, pp. 3-15.
[3] P. Castells, M. Fernández, D. Vallet, An adaptation of the vector-space model for ontology-based information retrieval IEEE Transactions on Knowledge and Data Engineering 19 (2007) 261-272.
[4] D. Sánchez, D. Isern, M. Millán, Content annotation for the Semantic Web: an automatic Web-based approach, Knowledge and Information Systems 27 (2011) 393-418.
[5] C. Vicient, D. Sánchez, A. Moreno, An automatic approach for ontology-based feature extraction from heterogeneous textual resources, Engineering Applications of Artificial Intelligence 26 (2013) 1092-1106.
[6] D.C. Wimalasuriya, D. Dou, Ontology-based information extraction: An introduction and a survey of current approaches, Journal of Information Science 36 (2010) 306-323.
[7] G. Allampalli-Nagaraj, I. Bichindaritz, Automatic semantic indexing of medical images using a web ontology language for case-based image retrieval, Engineering Applications of Artificial Intelligence 22 (2009) 18-25.
[8] M. Batet, Ontology based semantic clustering, AI Communications 24 (2011) 291-292.
[9] A. Valls, K. Gibert, D. Sánchez, M. Batet, Using ontologies for structuring organizational knowledge in home care assistance, International Journal of Medical Informatics 79 (2010) 370-387.
[10] L.W.C. Chan, Y. Liu, C.R. Shyu, I.F.F. Benzie, A SNOMED supported ontological vector model for subclinical disorder detection using EHR similarity, Engineering Applications of Artificial Intelligence 24 (2011) 1398–1409.
[11] P. Wriggers, M. Siplivaya, I. Joukova, R. Slivin, Intelligent support of engineering analysis using ontology and case-based reasoning, Engineering Applications of Artificial Intelligence 20 (2007) 709-720
[12] S. Martínez, D. Sánchez, A. Valls, A semantic framework to protect the privacy of electronic health records with non-numerical attributes, Journal of Biomedical Informatics 46 (2013) 294-303.
[13] D. Sánchez, M. Batet, A. Viejo, Automatic general-purpose sanitization of textual documents, IEEE Transactions on Information Forensics and Security 8 (2013) 853-862.
[14] D. Sánchez, M. Batet, A. Viejo, Utility-preserving sanitization of semantically correlated terms in textual documents, Information Sciences 279 (2014) 77-93.
[15] L. Ding, T. Finin, A. Joshi, R. Pan, R.S. Cost, Y. Peng, P. Reddivari, V. Doshi, J. Sachs, Swoogle: a search and metadata engine for the Semantic Web, In: Proc. of thirteenth ACM international conference on Information and knowledge management, CIKM 2004, Washington, D.C., USA, 2004, pp. 652-659.
[16] E. Simperl, M. Mochol, T. Bürger, I. Popov, Achieving maturity: the state of practice in ontology engineering in 2009, In: Proc. of On the Move to Meaningful Internet Systems: OTM 20092009, pp. 983–991.
[17] A. Zouag, R. Nkambou, A survey of domain ontology engineering: methods and tools, In: Advances in Intelligent Tutoring Systems, 2010, pp. 103–119.
[18] O. Corcho, M. Fernández-López, A. Gómez-Pérez, A. López-Cima, Building legal ontologies with METHONTOLOGY and WebODE, In: V.R.B.e. al. (ed). Law and the Semantic Web, Springer-Verlag, 2005, pp. 142–157.

[19] D. Vrandecic, S. Pinto, C. Tempich, Y. Sure, The diligent knowledge processes, Journal of Knowledge Management 9 (2005) 85–96.

[20] D. Vrandecic, Ontology evaluation, In: Handbook on Ontologies, Springer, 2009, pp. 293-313.

[21] M. Fernández, C. Overbeeke, M. Sabou, E. motta, What makes a good ontology? a case-study in fine-grained knowledge reuse, In: Proc. of 4th Asian Conference on The Semantic Web2009, pp. 61-75.

[22] A. Lozano-Tello, A. Gómez-Pérez, ONTOMETRIC: A method to choose the appropriate ontology, Journal of Database Management 15 (2004) 1-18.

[23] A. Maedche, S. Staab, Measuring similarity between ontologies, In: Proc. of 13th International Conference on Knowledge Engineering and Knowledge Management2002, pp. 251-263.

[24] M. Sabou, J. Garcia, S. Angeletou, M. d'Aquin, E. Motta, Evaluating the Semantic web: a task-based approach, In: Proc. of 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference2007, pp. 423-437.

[25] A. Burton-Jones, V. Storey, V. Sugumaran, P. Ahluwalia, A semiotic metrics suite for assessing the quality of ontologies, Data and Knoweldge Engineering 44 (2005) 84-102.

[26] S. Tartir, B. Arpinar, A.P. Sheth, Ontological evaluation and validation, In: Theory and Applicaitons of Ontology, Springer, 2010, pp. 115-130.

[27] H. Alani, C. Brewster, N. Shadbolt, Ranking ontologies with AKTiveRank, In: Proc. of 5th International Semantic Web Conference2006, pp. 1-15.

[28] P. Buitelaar, T. Eigner, T. Declerck, OntoSelect: a dynamic ontology library with wupport for ontology selection, In: Proc. of 3rd International Semantic Web Conference2004.

[29] A. Gangemi, C. Catenacci, M. Ciaramita, J. Lehmann, Modelling ontology evaluation and validation, In: Proc. of 3rd European Semantic Web Conference2006, pp. 140-154.

[30] N. Guarino, C. Welty, An overview of OntoClean, In: S.a.S. Staab, R. (ed). Handbook on Ontologies, Springer-Verlag, 2009, pp. 201-220.

[31] S. Tartir, I. Arpinar, M. Moore, A. Sheth, B. Aleman-Meza. OntoQA: Metric-based ontology quality analysis. In: IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogenous Data and KNoweldge Sources: IEEE Computer Society; 2005. p. 45-53.

[32] D. Sánchez, M. Batet, D. Isern, A. Valls, Ontology-based semantic similarity: a new feature-based approach, Expert Systems with Applications 39 (2012) 7718-7728.

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

[34] G. Pirró, A semantic similarity metric combining features and intrinsic information content, Data & Knowledge Engineering 68 (2009) 1289-1308.

[35] E. Parzen, Modern Probability Theory and its Applications, Wiley, 1960.

[36] S. Martínez, A. Valls, D. Sánchez, Semantically-grounded construction of centroids for datasets with textual attributes, Knowledge-Based Systems 35 (2012) 160-172.

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

[38] R. Rada, H. Mili, E. Bichnell, M. Blettner, Development and application of a metric on semantic nets, IEEE Transactions on Systems, Man, and Cybernetics 9 (1989) 17-30.

[39] M. Batet, D. Sánchez, A. Valls, An ontology-based measure to compute semantic similarity in biomedicine, Journal of Biomedical Informatics 44 (2011) 118-125.

[40] B. Lemaire, G. Denhière, Effects of high-order co-occurrences on word semantic similarities, Current Psychology Letters - Behaviour, Brain and Cognition 18 (2006) 1.

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

[42] J. Domingo-Ferrer, D. Sánchez, G. Rufian-Torrell, Anonymization of nominal data based on semantic marginality, Information Sciences 242 (2013) 35-48.

[43] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, S. Martínez, Enhancing data utility in differential privacy via microaggregation-based k-anonymity, VLDB Journal 23 (2014) 771-794.

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

[45] M. Batet, D. Sanchez, A. Valls, K. Gibert, Semantic similarity estimation from multiple ontologies, Applied Intelligence 38 (2013) 29-44.

[46] A. Devitt, C. Vogel, The topology of WordNet: some metrics, In: P. Sojka, K. Pala, P. Smrz, C. Fellbaum, P. Vossen (eds), Proc. of 2nd Global Wordnet Conference, GWC 2004, Brno, Czech Republic, 2004, pp. 106-11.

[47] K. Spackman, SNOMED CT milestones: endorsements are added to already-impressive standards credentials, Healthcare Informatics 21 (2004) 54-56.

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

[49] D. Sánchez, M. Batet, A semantic similarity method based on information content exploiting multiple ontologies, Expert Systems with Applications 40 (2013) 1393-1399.

[50] R.L. Cilibrasi, P.M.B. Vitányi, The Google Similarity Distance, IEEE Transactions on Knowledge and Data Engineering 19 (2006) 370-383.

[51] B.T. McInnes, T. Pedersen, Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text, Journal of Biomedical Informatics 46 (2013) 1116-1124.

[52] A. Budanitsky, G. Hirst, Evaluating wordnet-based measures of semantic distance, Computational Linguistics 32 (2006) 13-47.

[53] T. Pedersen, S. Pakhomov, S. Patwardhan, C. Chute, Measures of semantic similarity and relatedness in the biomedical domain, Journal of Biomedical Informatics 40 (2007) 288-299.

[54] D. Sánchez, M. Batet, Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective Journal of Biomedical Informatics 44 (2011) 749-759.

[55] V.E. Johnson, Revised standards for statistical evidence, In: Proc. of Proceedings of the National Academy of Sciences of the United States of America2013.

[56] J. Domingo-Ferrer, Marginality: a numerical mapping for enhanced exploitation of taxonomic attributes, In: Proc. of Modeling Decisions for Artifical Intelligence2012, pp. 367-381.

[57] J. Domingo-Ferrer, A. Solanas, A measure of variance for nominal hierarchical attributes, Information Sciences 178 (2008) 4644-4655. Erratum in 179 (2009) 3732.

[58] Eurostat. NACE Rev. 2: Statistical Classification of Economic Activities in the European Community. 2008

[59] P. Samarati, L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. 1998

[60] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, L-diversity: Privacy beyond k-anonymity, ACM Transactions on Knowledge Discovery from Data 1 (2007) 3.

[61] J. Domingo-Ferrer, F. Sebé, J. Castellà-Roca, On the security of noise addition for privacy in statistical databases, In: Privacy in Statistical Databases, Springer, 2004, pp. 149-161.

[62] D. Sánchez, M. Batet, A new model to compute the information content of concepts from taxonomic knowledge, International Journal on Semantic Web and Information Systems 8 (2012) 34-50.

[63] D. Sánchez, M. Batet, D. Isern, Ontology-based information content computation, Knowledge-based Systems 24 (2011) 297-303.

[64] P. Resnik, Using information content to evalutate semantic similarity in a taxonomy, In: C.S. Mellish (ed). Proc. of 14th International Joint Conference on Artificial Intelligence, IJCAI 1995, Montreal, Quebec, Canada, 1995, pp. 448-453.