Briefings in Bioinformatics, 21(2), 2020, 473–485

doi: 10.1093/bib/bbz009
Advance Access Publication Date: 4 February 2019
Review article

Evaluation of ontology structural metrics based on public repository data

Manuel Franco, Juana María Vivo, Manuel Quesada-Martínez, Astrid Duque-Ramos and Jesualdo Tomás Fernández-Breis

Corresponding author: Jesualdo Tomás Fernández-Breis, Departamento de Informática y Sistemas, Universidad de Murcia, IMIB-Arrixaca, 30100, Murcia, Spain. Tel.: +34-868884613; E-mail: jfernand@um.es

Abstract

The development and application of biological ontologies have increased significantly in recent years. These ontologies can be retrieved from different repositories, which do not provide much information about quality aspects of the ontologies. In the past years, some ontology structural metrics have been proposed, but their validity as measurement instrument has not been sufficiently studied to date. In this work, we evaluate a set of reproducible and objective ontology structural metrics. Given the lack of standard methods for this purpose, we have applied an evaluation method based on the stability and goodness of the classifications of ontologies produced by each metric on an ontology corpus. The evaluation has been done using ontology repositories as corpora. More concretely, we have used 119 ontologies from the OBO Foundry repository and 78 ontologies from AgroPortal. First, we study the correlations between the metrics. Second, we study whether the clusters for a given metric are stable and have a good structure. The results show that the existing correlations are not biasing the evaluation, there are no metrics generating unstable clusterings and all the metrics evaluated provide at least reasonable clustering structure. Furthermore, our work permits to review and suggest the most reliable ontology structural metrics in terms of stability and goodness of their classifications.
Availability: http://sele.inf.um.es/ontology-metrics

Key words: biological ontologies; quantitative metrics; metrics comparison; data analysis

Introduction

The development and application of biological ontologies have increased significantly in recent years [1–3]. Their success lies in the combination of four main features present in almost all ontologies: standard identifiers for classes and relations that represent the phenomena within a domain, a vocabulary for a domain, metadata that describes the intended meaning of the classes and relations and machine-readable axioms and definitions that enable computational access to some aspects of the meaning of classes and relations [4]. The availability of hundreds of ontologies has provoked the need for repository-based initiatives to find and share their knowledge easily. Examples of such repositories are the OBO Foundry [5], AgroPortal [6], OntoBee [7], the Ontology Lookup Service (OLS) [8], AberOWL [9] or NCBO BioPortal [10]. According to the OBO Foundry website (http://www.obofoundry.org/), 'OBO Foundry participants voluntarily adhere to and contribute to the development of an evolving set of principles including open use, collaborative development, non-overlapping and strictly-scoped content, and common syntax

Manuel Franco is a professor in the Department of Statistics and Operations Research at the University of Murcia.
Juana María Vivo is a professor in the Department of Statistics and Operations Research at the University of Murcia.
Manuel Quesada-Martínez is an assistant professor in the Department of Statistics, Maths and Informatics at the Miguel Hernández University.
Astrid Duque-Ramos is a professor in the Department of Systems Engineering at the University of Antioquia.
Jesualdo Tomás Fernández-Breis is a full professor in the Department of Informatics and Systems at the University of Murcia and a member of the IMIB-Arrixaca Bio-Health Research Institute.
Submitted: 22 October 2018; **Received (in revised form):** 20 December 2018

and relations, based on ontology models that work well, such as the Gene Ontology (GO)'. Assuring the quality of ontologies has traditionally been a hard, tedious, manual task. The development and maintenance of a number of ontologies are now fortunately supported by automated tools, which permit to ensure some quality properties of the ontologies. However, even the development teams of those ontologies would benefit of the availability of automatic methods to provide information about the ontologies, which could help them to analyse those development and maintenance processes.

In general, the quality of an ontology is measured by analysing the degree in which the ontology meets its design requirements. The use of metrics is a good practice for evaluation processes, which have to be objective and reproducible. The community has recognised the necessity of reference methods to measure the quality of ontologies [2, 11], but there has been no community agreement so far [12]. However, the ontology engineering community has proposed both qualitative [13, 14] and quantitative approaches [15–19]. Gangemi *et al.* [13] propose a diagnostic task based on ontology descriptions, using three categories of criteria (structural, functional and usability profiling). Rogers [14] applies four qualitative criteria (philosophical rigour, ontological commitment, content correctness and fitness for a purpose). Yao *et al.* [19] and Tartir and Arpinar [18] define a series of metrics for evaluating structural properties in the ontology. Works like [15, 16, 20] evaluate the ontology from a realism-based perspective that demands manual judgement of users. In addition, works like [18, 19, 21–23] use metrics to measure quality-related properties of the ontologies. Those works have contributed to propose a set of metrics, mostly dealing with structural aspects of ontologies. Unfortunately, the evaluation of the methods and the metrics is very limited despite having demonstrated their usefulness in particular scenarios. The validity of those metrics as measurement instrument has not been sufficiently studied by the ontology engineering community.

In this work we aim at increasing the knowledge about ontology structural metrics. We study the validity of a set of structural metrics for assessing relevant features of ontologies based on the use of corpora of ontologies. For this purpose, we propose a method for evaluating metrics based on the information available in public ontology corpora. This will allow us to analyse the structural metrics on each ontology repository. In our approach, the values of each metric are clustered in five groups by analysing the distribution of its values. Each cluster is assigned a quality score in the range $\{1, \ldots, 5\}$, analogously to the standard Likert scale [24]. Since the method is corpus based, the clusters may vary for different corpora.

In this framework, the evaluation of structural metrics will be illustrated by using the OBO Foundry and AgroPortal repositories. The OBO Foundry repository has been selected because it is a general repository of biological and biomedical ontologies, which are supposed to share certain building principles. AgroPortal contains vocabularies and ontologies for agronomy, food, plant sciences and biodiversity [6], so it allows an analysis not specific of a unique corpus and domain.

The main contributions of this method are (1) the analysis of the correlations between structural metrics, (2) the validation of structural metrics by analysing the stability and goodness of the clusters and (3) the identification of the most reliable metrics for classifying ontologies in terms of stability and goodness of their classifications. We believe that this work allows to generate new insights in the field of ontology engineering and to shed light on ontology evaluation methods.

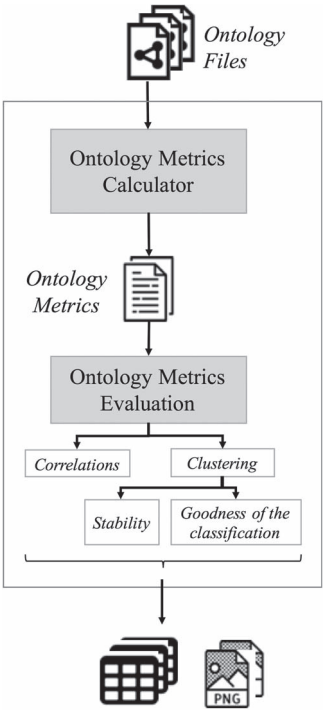

Figure 1. General overview of the process for the evaluation of the structural ontology metrics on a given corpus. First, the Ontology Metrics Calculator obtains the values of the *m* metrics for the *n* ontologies in the corpus. Then, the metrics are evaluated. The evaluation starts by finding correlations between the metrics. After that, each metric is independently evaluated. For this purpose, a clustering is performed on its *n* values. The result of the clustering is used for analysing the stability of the metric and the goodness of the classification generated by such a metric. The application of the method to different corpora requires different, independent executions of the process.

Methods

A general overview of the method applied in this work can be seen in Figure 1. First, a series of metrics (Metrics and scaling function) are calculated for each ontology included in the repository. The evaluation of a metric is then performed through different studies: analysis of correlations between the metrics (Correlation between the set of metrics) and assessment of its stability and the goodness of its classifications (Validation of the clusters obtained using the dynamic scale function). The description of our experimental study is reported in Experimental setup.

Metrics and scaling function

In this work, we focus on 19 ontology structural metrics (Table 1) that measure a series of facets of the ontology such as cohesion, the existence of multiple inheritance in the ontology or the richness of the ontology in terms of properties or comments.

The metrics have a function $f(x)$ associated, whose domain is an ontology and whose ranges are the raw values of the metrics that have different units of measurement. The evaluation of the ontology as a whole has to consider the values from all the metrics. A scaling function is used to bridge the different ranges of the metrics, $n\left(f(x)\right)$ being a function that generates an ordered factor of $k = 5$ categories in a dynamic scale, which is based on experimental data used as reference, i.e. $n\left(f(x)\right)$ partitions the range of $f(x)$ in five non-prefixed continuous intervals that

Table 1. Definition of the 19 metrics evaluated: column 1 shows the acronym of the metric, column 2 describes the ontology facet measured by the metric, column 3 describes how the metric is calculated and column 4 includes the references in which the metrics have been proposed or adapted to ontologies

Name of the metric	Facet of the ontology measured	Description	Reference
CBOnto	Coupling	Number of direct ancestor of classes divided by the number of classes minus subclasses of thing	[17, 25]
DITOnto	Depth of the hierarchy	Length of the longest path from thing to leaf classes	
NOCOnto	Descendants	Number of the direct subclasses divided by the number of classes minus the number of leaf classes	
RFCOnto	Properties usage	Number of usages of object and data properties and superclasses divided by the number of classes	
WMCOnto	Complexity	Mean length of the path from thing to a leaf classes	
NOMOnto	Properties	Mean number of object and data property usages per class	[26]
NACOnto	Ancestors of leaf classes	Mean number of superclasses per leaf classes	
LCOMOnto	Cohesion	Mean length of all paths from leaf classes to thing	[27]
ANOnto	Annotations	Mean number of annotations properties per classes	[18]
CROnto	Individuals	Mean number of individuals per classes	
AROnto	Attribute richness	Number of restrictions of the ontology per classes	
INROnto	Descendants	Mean number of subclasses per classes	
PROnto	Property richness	Number of subclass of relationships divided by the number of subclass of relationships and properties	
RROnto	Properties usage	Number of usages of object and data properties and super classes divided by the number of classes	
TMOnto	Multiple inheritance	Mean number of classes with more than one ancestor	[17]
POnto	Ancestors	Mean number of direct ancestor per class	
CBOnto2	Coupling	Mean number of direct ancestor per classes	
TMOnto2	Multiple inheritance	Mean number of direct ancestor of classes with more than one direct ancestor	
WMCOnto2	Complexity	Mean number of path from thing to leaf classes	

contain all the observed samples in the experimental data. It should be noted that we call values to the measurements of the metrics and scores to the scaled values. The clustering algorithm needs to know which values produced by $f(x)$ correspond to the highest categories of the factor associated. Thereby, analogously to the standard Likert scale, five predefined scores {1, ..., 5} are used, where 1 is associated with the lowest category of the factor and 5 with the highest one. In this context, 5 is not necessarily associated with the highest values of a particular metric.

An ontology set $\theta = \{\theta_1, \ldots, \theta_n\}$ is received as input and generates a vector of raw values $R_{M_i} = \{R_{\theta_1}, \ldots, R_{\theta_n}\}$ for each metric in $M = \{M_1, \ldots, M_m\}$. The application of a scaling function transforms R_{M_i} vectors into a scaled vector $N_{M_i} = \{N_{\theta_1}, \ldots, N_{\theta_n}\}$. This dynamic scale has been used to analyse the evolution of ontologies, using as experimental data those obtained by processing different versions of the same ontology [28, 29].

From the information of a given experimental dataset, the dynamic scale uses the *k*-means algorithm *m* times, one for each metric in *M*, in order to find a partitioning of the ontologies into five non-empty and non-overlapping categories. By maximising the compactness of the ontologies within categories (minimising the intra-cluster variance) and maximising the separability between the categories (maximising the inter-cluster variance) in each iteration, the new centroids are recalculated from the previous partitioning, and then the new cluster assignment is generated by reallocating each R_{θ_j} to the nearest centroid.

Figure 2 graphically shows the application of the dynamic scaling function using a corpus of ontologies, θ, for each metric in *M*. Specifically, (1) shows the graphical representation of the raw values of R_{M_i} for all the ontologies; (2) depicts the scores of N_{M_i}, i.e. the results of the dynamic scaling function for all the ontologies; and (3) displays the five categories of ontologies for the M_i metric that are determined by the N_{M_i} scores.

Correlation between the set of metrics

The correlations between the set of metrics will be studied using the data obtained for all the ontologies. For this purpose, we will calculate the Pearson correlation coefficient between all the pairs of metrics using as input the raw data obtained for all the ontologies θ of a corpus, measuring the strength and direction of the linear relationship between each pair. This analysis will allow us to determine whether certain pairs of metrics are representing the same ontology quality facet and to incorporate new methods that will be useful for validating metrics.

Validation of the clusters obtained using the dynamic scale function

The robustness of the dynamic scale is analysed by using validation procedures of non-hierarchical clustering. For this purpose, two important characteristics of the cluster validation will be performed on the clusterings generated by the dynamic scale

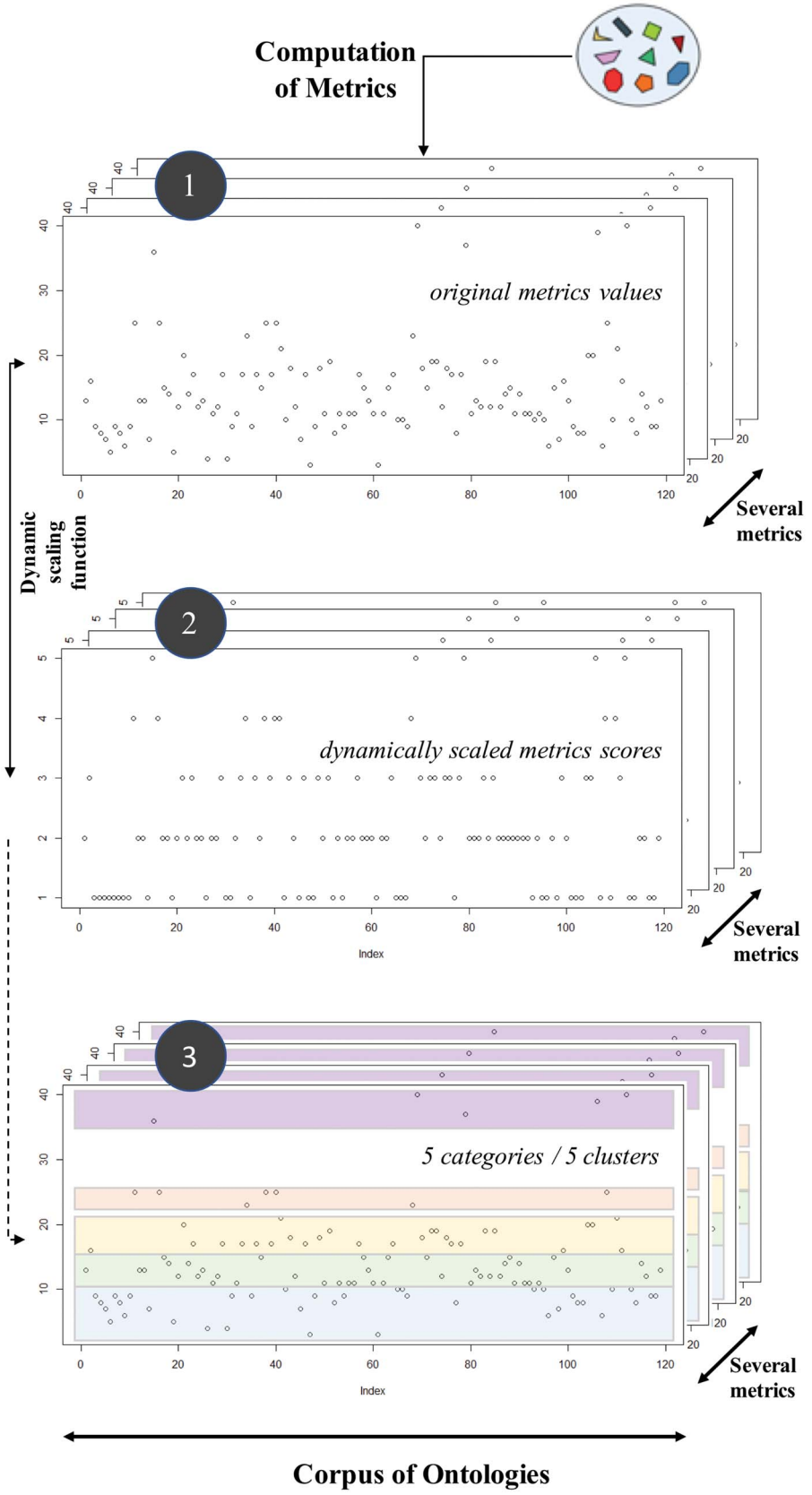

Figure 2. Graphical representation of the process for the application of the dynamic scaling function to an ontology corpus, which consists of three steps: (**1**) computation of the *m* metrics for the *n* ontologies in the corpus, so obtaining *m* values for each ontology; (**2**) a clustering is performed for each metric applying the *k*-means algorithm to its *n* values; and (**3**) each cluster is assigned a score in the range 1 to 5.

function: stability of the clusters and validity of the clusters. We describe next the methods used for both studies.

Stability of the clusters

The stability of the clusters generated by a partitioning algorithm means that the clustering is not meaningfully affected by small variations in the data, and thus stability may be measured by taking into account changes in the clusters $\left(C_1, \ldots, C_5\right)$ when the sample varies [30]. We can apply a bootstrap resampling method to assess the stability of each category of the dynamic scale clustering, $S_{M_i}\left(C_j\right)$ for $j = 1, \ldots, 5$, for each metric M_i, based on a similarity measure between sets, called Jaccard coefficient [31], as described by Hennig [32]. In detail, for each category C_j, the Jaccard coefficient is the proportion of concordant ontologies between C_j and the most similar cluster in one bootstrapped clustering of R_{M_i}. Thereby, $S_{M_i}\left(C_j\right)$ is the mean of the Jaccard coefficient values of the b bootstrap replicates. The number b of bootstrap replications is usually chosen according to the computational complexity of the estimators in order to achieve more relative reliable and accurate results. Thus, for each metric M_i in M, we have computed the category stabilities $S_{M_i}\left(C_j\right)$ for $j = 1, \ldots, 5$, by setting $b = 50, 100, 500$ and 1000, respectively. For interpretation purpose, we use the $S_{M_i}\left(C_j\right)$ scores to classify the categories as follows:

- **Unstable:** the category should not be trusted when $S_{M_i}\left(C_j\right) \in \left[0, 0.60\right[$.
- **Doubtful:** a pattern is recognised in the data, but there is uncertainty about which ontologies exactly should belong to the category when $S_{M_i}\left(C_j\right) \in \left[0.60, 0.75\right]$.
- **Stable:** the category should be trusted when $S_{M_i}\left(C_j\right) \in \left]0.75, 0.85\right]$.
- **Highly stable:** there is high certainty about which ontologies belong to the category when $S_{M_i}\left(C_j\right) \in \left]0.85, 1\right]$.

Furthermore, the corresponding category stability scores can be aggregated to form a single stability criterion for each metric that can be used to compare the different metrics. Therefore, assuming the same relative importance of the categories, the most straightforward aggregation is to compute and use the stability mean as global stability index for each metric, $S\left(M_i\right)$ for $i = 1, \ldots, m$. For example, using 1000 replicates, the stability of DITOnto categories is (0.84, 0.58, 0.55, 0.66, 0.69) on the OBO Foundry repository, and it is (0.94, 0.84, 0.78, 0.73, 0.68) on AgroPortal. Hence, the global stability index of DITOnto, $S(DITOnto)$, is 0.66 and 0.79, respectively.

Validity of the clusters

The validity of the clusters assesses the goodness of the clustering. There are several validity indexes available, such as Silhouette width (*sil*) [33], Calinski–Harabasz (*ch*) [34], Dunn (*dunn*) [35] and Davies–Boudin (*db*) [36] measurements, which can be used to analyse the quality of the classification obtained by using the dynamic scale function. They take into consideration the compactness of the ontologies into the same category and the separability between categories [37], which are two internal characteristics for the cluster validation. We focus our attention on the *sil* index to compute and compare the quality of the clustering outputs found by the different metrics because it enables to measure the goodness of the classification for both ontologies and metrics. More precisely, this measurement provides an assessment of how similar an ontology is to other ontologies from the same cluster and dissimilar to all the other clusters. The average on all the ontologies quantifies how appropriately the ontologies are clustered.

Firstly, the *sil* coefficient for each metric of a particular ontology θ_l represents the degree of confidence in the clustering, and it is given by

$$sil_{M_i}\left(\theta_l\right) = \frac{b_l - a_l}{\max\left(a_l, b_l\right)}, \text{for } l = 1, \ldots, n$$

where a_l is the mean distance between the ontology θ_l and all other ones in the same category and b_l is the mean distance between the ontology θ_l and the ones of the 'nearest neighbouring category'. Its value ranges from −1 to 1. Thus, for each ontology θ_l, $sil_{M_i}\left(\theta_l\right)$ measures how well it has been classified, which can be interpreted as in [33]. A large value close to 1 indicates that the ontology tends to be 'well-classified'. A value close to zero means that the ontology lies equally far away from the category assigned and the nearest neighbouring one. A negative value close to −1 shows that the ontology is 'misclassified'.

Secondly, the overall goodness of the clustering for a metric M_i is evaluated by the global Silhouette coefficient, which is defined by the mean of the *sil* scores, $\overline{sil}\left(M_i\right) = \sum_{l=1}^{n} sil_{M_i}\left(\theta_l\right)/n$ for $i = 1, \ldots, m$. Kaufman and Rousseeuw [38] suggested the interpretation of the global Silhouette width score as the effectiveness of the clustering structure, in terms of the metrics:

- There is no substantial clustering structure when $\overline{sil}\left(M_i\right) \in \left[-1, 0.25\right]$.
- The clustering structure is weak and could be artificial when $\overline{sil}\left(M_i\right) \in \left]0.25, 0.50\right]$.
- There is a reasonable clustering structure when $\overline{sil}\left(M_i\right) \in \left]0.50, 0.70\right]$.
- A strong clustering structure has been found when $\overline{sil}\left(M_i\right) \in \left]0.70, 1\right]$.

Analogously, *ch*, *dunn* and *db* indexes might be also applied to provide assessments of the global goodness of the clustering for each metric as the global Silhouette width index. However, unlike $\overline{sil}$ index, there is no consensual threshold for these validity indexes in order to interpret a clustering as 'misclassified' or 'well-classified'.

Experimental setup

In this work, we have focused on two corpora of ontologies: the OBO Foundry and the AgroPortal. We applied the OQuaRE platform (http://sele.inf.um.es/oquare) for the calculation of metrics. This platform uses the OWL API [39] and Neo4j (https://neo4j.com/). We actually used a web service to execute the metrics over the ontologies of our corpus in its server and to obtain an XML file with all the results. The platform offers the possibility of using two reasoners, ELK [40] and Hermit [41]; for this experiment we selected the ELK reasoner, which works with the OWL 2 EL profile (https://www.w3.org/TR/owl2-profiles/#OWL_2_EL_2).

The ontologies were processed using their Unified Resource Identifier (URI) for retrieving the corresponding file. Only ontologies in OBO or OWL formats were considered in this study. We were able to process 119 ontologies from the OBO Foundry and 78 from AgroPortal. The whole description of the two corpora can

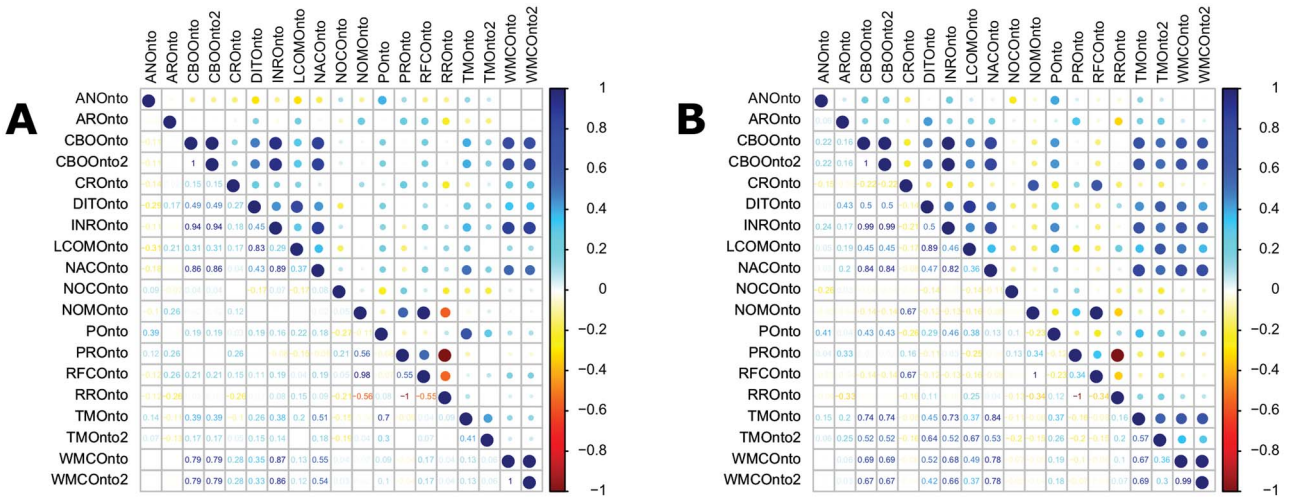

Figure 3. Pearson correlation coefficient between metrics. Graphical representations display the correlation matrix between pairs of the metrics on OBO Foundry (**A**) and AgroPortal (**B**) corpora, respectively. Scale on the right side reflects the corresponding intensity color to Pearson's correlation values, varying from −1 (red) to 1 (blue). Above the main diagonal, color intensity and size of the circles are proportional to the correlation values shown below such a diagonal.

be found in Supplementary File 1, which includes the name, the URI and the version of each ontology used in our study.

We processed the XML file, extracted the metrics raw scores and used R [42] for performing the statistical analysis. In particular, we used the following R packages for the statistical analysis: *corrplot* for correlations [43], *fcp* for stability analysis [44] and *cluster* for Silhouette graphics and the analysis of the goodness of the classifications [45].

Finally, our method can be applied to any corpus of ontologies through our website: http://sele.inf.um.es/ontology-metrics.

Results

Correlations between metrics

Figure 3 displays the correlations between pairs of metrics, using the raw values obtained for the OBO Foundry (Figure 3A) and AgroPortal (Figure 3B) ontologies processed. Most of the pairs of metrics have a correlation in absolute value under 0.80. In both repositories, we have obtained two pairs of metrics with a perfect correlation: <CBOnto, CBOnto2> and <PROnto, RROnto>:

- CBOnto and CBOnto2 are very similar, but CBOnto2 has an additional factor that includes in the computation the top level nodes of the ontologies. The calculation of CBOnto2 using ELK reasoner makes this additional factor to be 0, so both metrics have the same values on both corpora. This would not happen using an OWL 2 Description Logics (DL) reasoner such as Hermit.
- Both PROnto and RROnto account for relations. OWL relations can be classified in taxonomic and non-taxonomic ones. Each one of these two metrics measures the proportion of one of such types, which justifies this perfect negative correlation.

The next highest correlated pair is <WMCOnto, WMCOnto2> with a correlation close to 1 (0.9996 in OBO Foundry and 0.9881 in AgroPortal). In this case, they measure structural facets related to paths from leaf nodes to the root node of an ontology. While WMCOnto takes into account the length of the paths, WMCOnto2 takes into account the number of them.

Note that the pair <RFCOnto, NOMOnto> also achieves a correlation close to 1 (0.9801 in OBO Foundry and 0.9999 in AgroPortal). Both metrics are related to the use of properties. NOMOnto measures the mean number of properties use per class, whereas RFCOnto additionally uses the mean number of superclasses per class.

Figure 4 includes the pairs of metrics with correlations higher than 0.8 in absolute value for both repositories. The correlation between <CBOnto2, INROnto> is due to the fact that both deal with hierarchical relations. On the contrary, the correlations <INROnto, NACOnto> and <DITOnto, LCOMOnto> are not due to shared facets.

Stability of the clusters of the metrics

Table 2 shows the category stability scores $S_{M_i}(C_j)$, $j = 1, \ldots, 5$, and their global stability values $S(M_i)$ for different number b of bootstrap replications for the metrics ANOnto and AROnto from OBO Foundry and AgroPortal corpora. From both repositories, the global stability scores for each metric and for different bootstrap replicates are displayed in Figure 5. The convergence of the stability indexes can be observed when 500 replicates are used. The detailed results for the rest of the metrics on OBO Foundry and AgroPortal corpora can be found in Supplementary File 2.

According to Figure 5, the global stability of each metric tends to increase smoothly and converge when raising b. In fact, 17 out of 19 metrics remain in the same stability degree regardless of the value of b for OBO Foundry and 16 out of 19 metrics for AgroPortal. Moreover, the global stability scores obtained a range from 0.66 to 0.86 for OBO Foundry (0.61 to 0.88 for AgroPortal), so there are no 'Unstable' clusterings of the metrics, and specifically, 12 (10) of them achieved $S(M_i) > 0.75$, indicating that the 63.16% (52.63%) of all metrics provided 'Stable' or 'Highly stable' clusterings. In detail, 36.84% (47.37%) metrics are classified as 'Doubtful', 57.89% (47.37%) are 'Stable' and 5.26% (5.26%) are 'Highly stable' (see Table 3). Conceptually, having stable metrics means that the inclusion of new ontologies in the corpus would not have a meaningful impact on the current dynamic scaling of the metrics.

All these results support the clusters performed by the dynamic scale function with five categories, although a detailed analysis on the category stability scores shows that there is

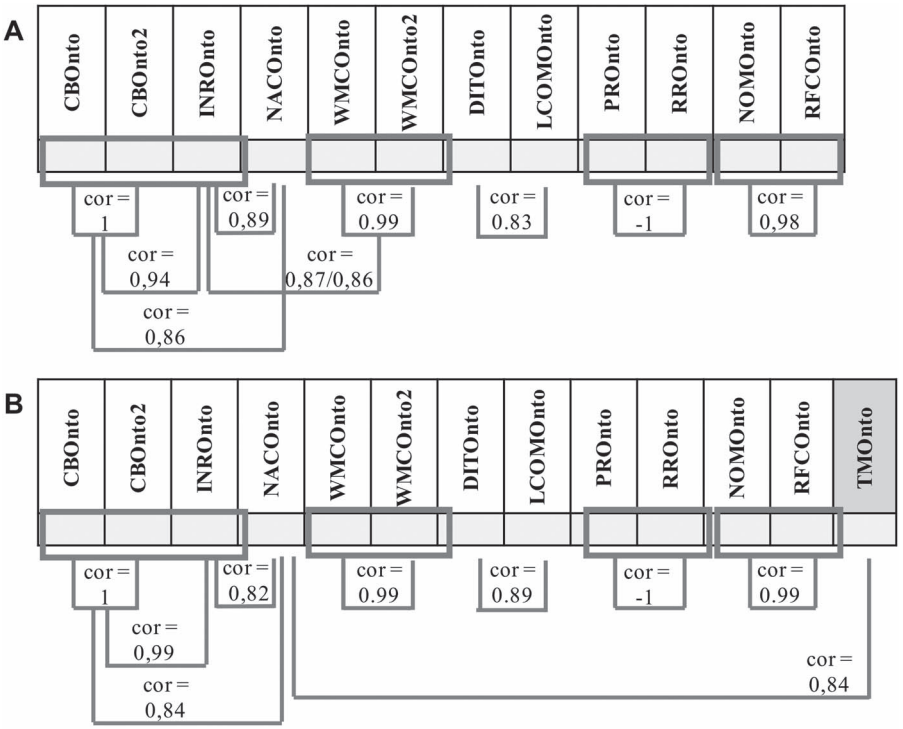

Figure 4. Summary of pairs of metrics with Pearson's correlations higher than 0.8 in absolute value for each ontology repository: (**A**) OBO Foundry and (**B**) AgroPortal.

Table 2. Stability of the clusters for the ANOnto and AROnto metrics on OBO Foundry and AgroPortal repositories. The stability scores for each category were computed for b = 50, 100, 500 and 1000 bootstrap replications. The last column presents the global stability scores for these metrics according to b on each corpus

		b	Category 1	Category 2	Category 3	Category 4	Category 5	Global stability score
OBO Foundry	**ANOnto**	**50**	0.88	0.65	0.89	0.73	0.73	0.77
		100	0.90	0.71	0.90	0.73	0.73	0.79
		500	0.91	0.77	0.91	0.74	0.75	0.82
		1000	0.92	0.78	0.92	0.75	0.76	0.82
	AROnto	**50**	0.94	0.38	0.79	0.67	0.74	0.71
		100	0.94	0.41	0.80	0.67	0.62	0.69
		500	0.94	0.40	0.77	0.68	0.63	0.69
		1000	0.95	0.42	0.78	0.68	0.64	0.69
AgroPortal	**ANOnto**	**50**	0.96	0.61	0.57	0.78	0.81	0.75
		100	0.96	0.58	0.56	0.77	0.76	0.73
		500	0.96	0.56	0.57	0.78	0.75	0.72
		1000	0.95	0.57	0.58	0.79	0.74	0.73
	AROnto	**50**	0.98	0.71	0.73	0.78	0.87	0.81
		100	0.98	0.68	0.76	0.79	0.87	0.81
		500	0.98	0.67	0.78	0.82	0.89	0.83
		1000	0.98	0.68	0.77	0.82	0.89	0.83

certain margin of improvement yet because if at least one single cluster has $S_{M_i}(C_j)<0.6$, then the clustering should be repeated with fewer categories. For example, the global stability score of AROnto is 0.69 on OBO Foundry repository, but the category 2 is 'Unstable' because of its score 0.42 (see Table 2).

Validity of clusters of the metrics

We analyse now the validity of the clusterings of the dynamic scale function. For each metric, the Silhouette width index provides validity measurements of the ontologies with respect to their classification by the scaling function and of the entire clustering. Moreover, this measure can also supply complementary information about the validity of those categories of the clustering by using the mean value of the ontologies belonging to each category.

Figure 6 shows the partial representation of the Silhouette widths of the CROnto, RROnto and WMCOnto metrics. The results of all the metrics can be found in Supplementary File 3 and 4. The Silhouette plot displays a measure of how close each ontology in one category is with respect to ontologies in the neighbouring categories and thus provides a way to visually

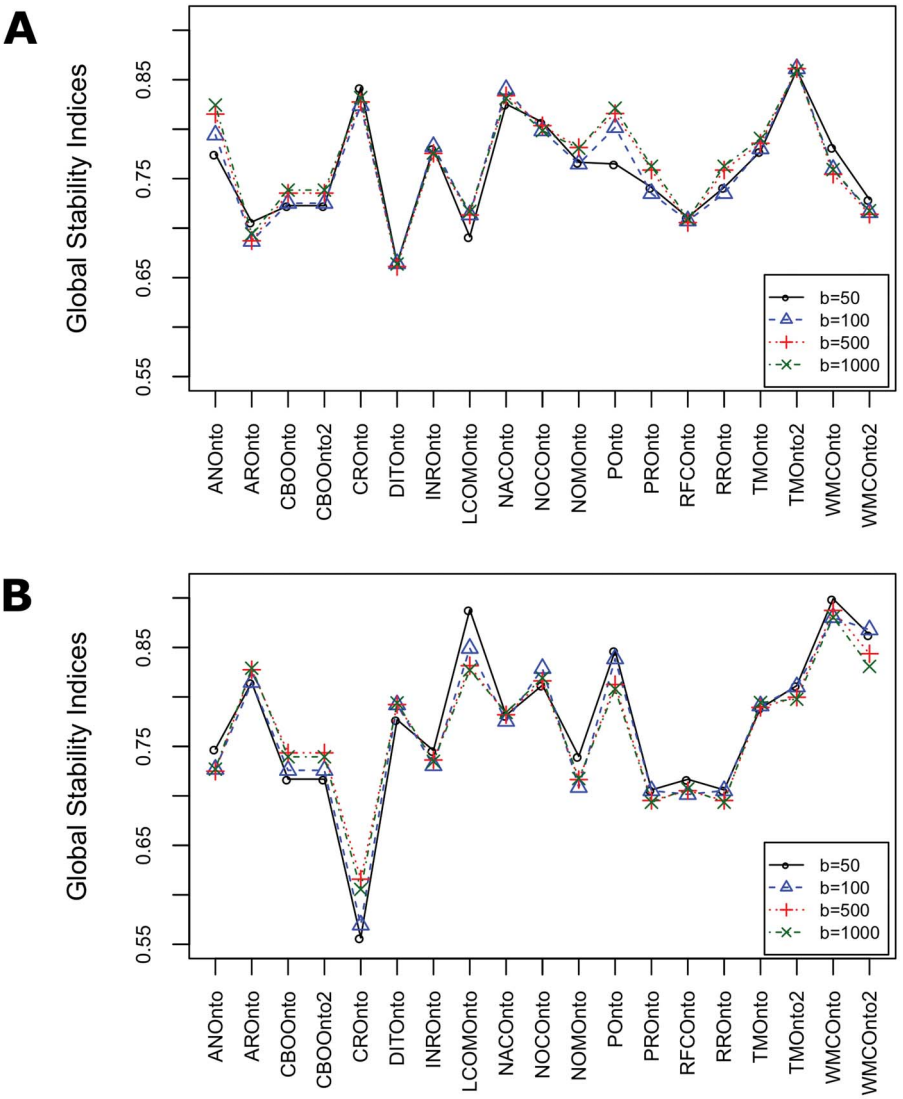

Figure 5. Global stability scores of the metrics. Both graphics display of scores of the global stability index obtained for each metric according to different bootstrap replicates, b = 50 (square symbol and solid lines), 100 (triangle symbol and dashed lines), 500 (plus symbol and dotted lines) and 1000 (star symbol and dotted and dashed lines), on each ontology repository: (**A**) OBO Foundry and (**B**) AgroPortal.

Table 3. Classification of the category stabilities and of the metric stabilities on the OBO Foundry and the AgroPortal corpora, respectively. The percentages represent the rates of each classification for all the 19 structural metrics from their stability score by using b = 1000 bootstrap replications

		Category 1	Category 2	Category 3	Category 4	Category 5	Global stability score
OBO Foundry	**Unstable**	0.00%	10.53%	10.53%	5.26%	0.00%	0.00%
	Doubtful	36.84%	26.32%	36.84%	63.16%	21.05%	36.84%
	Stable	31.58%	57.89%	42.11%	21.05%	26.32%	57.89%
	Highly stable	31.58%	5.26%	10.53%	10.53%	52.63%	5.26%
AgroPortal	**Unstable**	0.00%	10.53%	15.79%	15.79%	5.26%	0.00%
	Doubtful	31.58%	52.63%	26.32%	31.58%	31.58%	47.37%
	Stable	21.05%	21.05%	42.11%	26.32%	21.05%	47.37%
	Highly stable	47.37%	15.79%	15.79%	26.32%	42.11%	5.26%

assess the validity of ontology clusterings and categories for each metric. In this case, the global Silhouette width ranges from 0.51 to 0.86 in OBO Foundry and 0.57 to 0.95 in AgroPortal (see Table 4), so there are no metrics obtaining unstructured clustering neither weakly structured. More concretely, 31.58% (42.11%) of the metrics supplies categories with 'Strong structure' and 68.42% (57.89%) of them provide categories with 'Reasonable structure' on OBO Foundry (AgroPortal).

Moreover, we can try to identify metric clusters that could be improved by analysing the Silhouette width scores of the

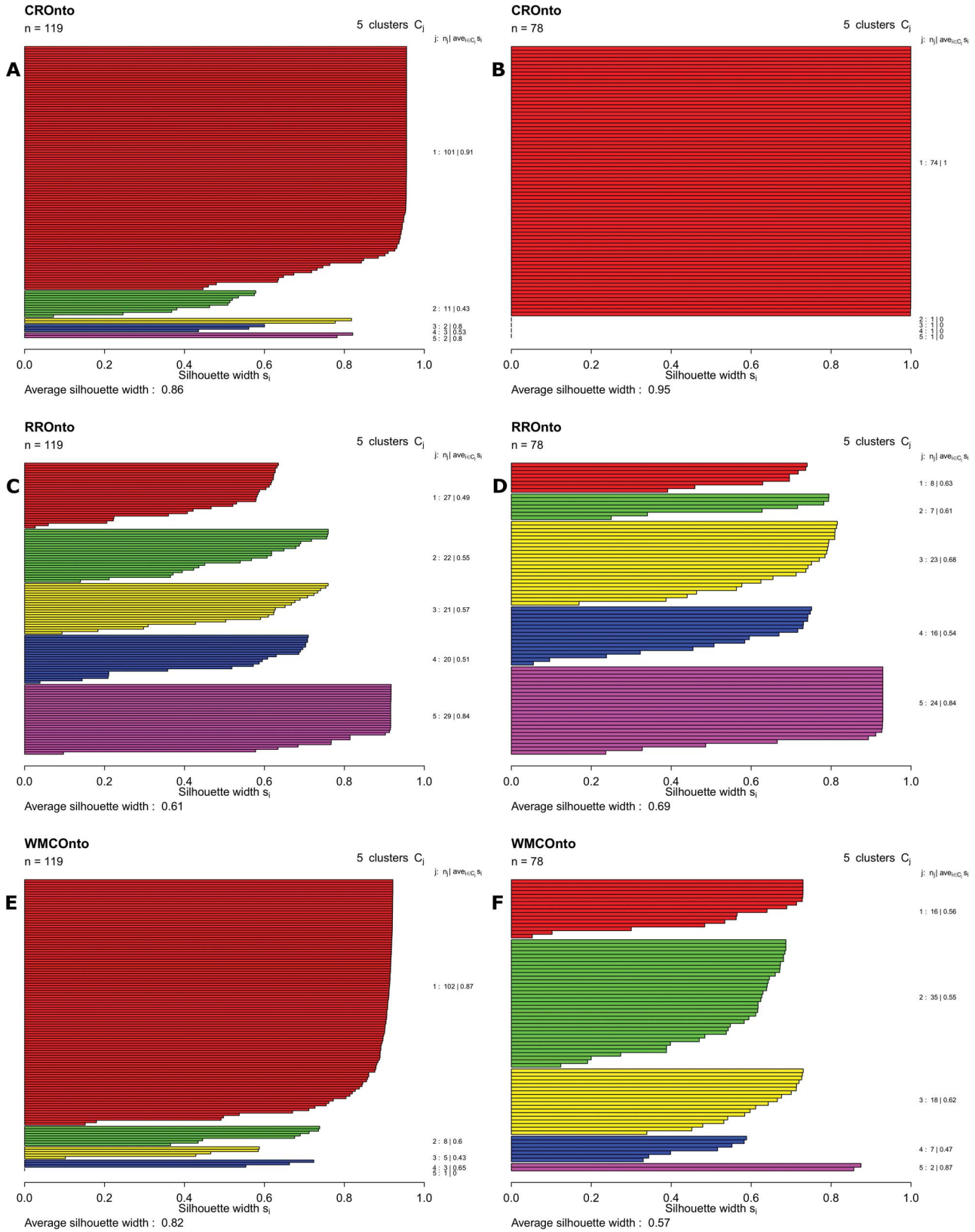

Figure 6. Clustering of the ontologies are shown by the Silhouette plots for the CROnto, RROnto and WMCOnto metrics from OBO Foundry (**A**, **C**, **E**) and AgroPortal (**B**, **D**, **F**) corpora. Ontologies within each cluster are depicted in decreasing order of their *sil* scores. For each cluster, the number of ontologies along with its mean *sil* score is noted. The global Silhouette width index is also reported for each metric.

Table 4. Validity of the clusters of the metrics on OBO Foundry and AgroPortal repositories. The clustering structure of each metric was classified from its global Silhouette width score

	OBO Foundry		AgroPortal	
	Global Silhouette width	Classification	Global Silhouette width	Classification
ANOnto	0.71	Strong	0.68	Reasonable
AROnto	0.80	Strong	0.86	Strong
CBOOnto	0.67	Reasonable	0.65	Reasonable
CBOOnto2	0.67	Reasonable	0.65	Reasonable
CROnto	0.86	Strong	0.95	Strong
DITOnto	0.51	Reasonable	0.63	Reasonable
INROnto	0.64	Reasonable	0.61	Reasonable
LCOMOnto	0.53	Reasonable	0.57	Reasonable
NACOnto	0.62	Reasonable	0.78	Strong
NOCOnto	0.61	Reasonable	0.61	Reasonable
NOMOnto	0.64	Reasonable	0.77	Strong
POnto	0.58	Reasonable	0.72	Strong
PROnto	0.61	Reasonable	0.69	Reasonable
RFCOnto	0.57	Reasonable	0.76	Strong
RROnto	0.61	Reasonable	0.69	Reasonable
TMOnto	0.62	Reasonable	0.79	Strong
TMOnto2	0.72	Strong	0.86	Strong
WMCOnto	0.82	Strong	0.57	Reasonable
WMCOnto2	0.78	Strong	0.57	Reasonable

ontologies. For example, the CROnto clustering has a strong structure, $\overline{sil}(CROnto)$ is 0.86 in OBO Foundry and 0.95 in AgroPortal. Although Silhouette widths of ontologies are positive in OBO Foundry, the mean in Category 2 (11 ontologies) is 0.43, but 1 out of 11 ontologies is close to 0 (see Figure 6A). Ontologies with Silhouette widths close to 0 are considered to be in the middle of two categories, and then it is not well-classified by the metric. In AgroPortal, each one of Categories 2 to 5 of CROnto only has one ontology with Silhouette score 0, so they are not well-classified by this metric. In the case of WMCOnto, the clustering structure is different in both repositories, $\overline{sil}(WMCOnto)$ is 0.82 in OBO Foundry and 0.57 in AgroPortal, strong and reasonable structures, respectively. Here, Category 5 is also formed by an ontology with Silhouette score 0 in OBO Foundry, and then it is not well-classified by WMCOnto. However, it is formed by two ontologies with Silhouette score 0.87 in AgroPortal, and thus both ontologies are well-classified by WMCOnto. Finally, approaches like these can be included to point out the most stable metrics for both repositories and to rank the metrics by validity or goodness of the clustering according to their silhouette widths, as it is shown in Table 4.

Discussion and perspectives

The method and the results

The increasing interest in ontologies makes necessary to develop effective quantitative methods for ontology evaluation. Reaching a community consensus about which properties are desirable in ontologies is hard, and it is even harder to agree on the quality-oriented classifications of the values associated with the quantitative measurements that describe the quality of an ontology. Besides, it is still a challenge to provide insights about whether the evaluation and classification of ontologies using structural quality metrics is a valid measuring instrument. In this work, we have analysed whether a set of selected metrics provides stable categories, structured clusterings and well-classified ontologies. In order to improve the usefulness of such a set of metrics, we have also discussed the correlations between them using experimental data obtained from two repositories of ontologies.

The analysis of correlations between metrics may help to optimise the set of metrics to use and to prevent biased evaluations when the metrics are perfectly correlated, and they are measuring similar ontology facets. We have found low correlations between the majority of the metrics, which is a good indicator, and we can say that these correlations are not biasing the evaluation. Nevertheless, those low correlations do not depend on the corpus of ontologies used since we obtain similar results for the two corpora analysed here. Thus, we can conclude that these metrics are not *ad hoc* to a particular corpus, but they can be reused in several ones. Moreover, in our study, the analysis of correlations has permitted to identify relationships between metrics; for instance, CBOnto and CBOnto2 provide the same clustering, and PROnto and RROnto provide completely opposite clusterings for both ontology repositories. These correlations can be used to normalise metrics (e.g. CBOnto and CBOnto2) or predict the behaviour of others (e.g. WCOnto and WCOnto2). Providing the same or opposite clusterings does not necessarily imply that the metrics are redundant. That would depend on the correlation value. In this case, CBOnto and CBOnto2 have perfect positive correlation in the two repositories used in this study, so one of them could be discarded to analyse those repositories. However, such redundancy might not be detected in other repositories. Furthermore, as explained in the Results, such redundancy would not happen in case of using a DL reasoner. PROnto and RROnto have perfect negative correlation, so they could be used to predict each other. We would not consider them redundant because they are not measuring the same property and their correlation might not be perfect in other corpora.

The normalisation of metrics would avoid computing unnecessary metrics, which would contribute to the performance of the execution, especially in corpora including a large set of ontologies. We recommend providing users with mechanisms to select the more explanatory metrics rather than removing metrics. This would enable different profiles of evaluation, which

could be supported by a pre-analysis of the ontologies considered representative of certain domains.

The stability analysis of the clusterings generated by the metrics on both ontology repositories has pointed out that the dynamic scale function using the standard Likert scale levels provides clusterings that are not 'unstable' for all the metrics (see Table 3). Furthermore, according to the results shown in Table 4, the global validity scores of the Silhouette width indicate that the clusterings obtained for all metrics have strong or reasonable structure. Therefore, the evaluation of these ontology structural metrics seems to indicate that their clusterings are not only stable but also ontologies are well-classified and categories are well-structured. Moreover, the classifications shown can be used to select the most stable metrics and the strongest structured metrics for classifying each repository. For example, Table 4 shows that 6 out of 19 metrics are classified as 'Strong' on OBO Foundry and 8 out of 19 on AgroPortal. Of 19 metrics, 11 have the same classification in both repositories. Also, the information from both repositories can be combined to conclude that:

- All the metrics have at least reasonable structure. AROnto, CROnto and TMOnto2 have the strongest structure. These three metrics are related to the ratio of attributes, individuals and direct ancestors.
- NACOnto, NOCOnto, POnto, TMOnto, TMOnto2 and WMCOnto provide stable or highly stable clusters. These metrics are related to the ratio of ancestors, descendants, multiple inheritance and number of paths, that is, facets related to the taxonomic component, which is fundamental in bio-ontologies.

In this work, we have used 19 ontology structural metrics. Other popular structural metrics such as consistency or formal correctness have not been included in this study because they are usually implemented as boolean functions, and both of them should actually be considered requirements for ontologies. Structural metrics can contribute to provide a useful view on the quality of the ontologies, but the whole picture requires to take into account non-structural metrics related to aspects such as content coverage, maintainability or performance. Our method can also be applied to non-structural metrics as long as they are quantitative, and the number of different values of the metric permits to create five clusters. It should be noted that some relevant non-structural metrics such as completeness or domain coverage are hardly automatically calculated, but the method could be applied if quantitative values are provided for them.

Application to ontology repositories

The method has been applied in this work to two ontology repositories, namely, the OBO Foundry and the AgroPortal, although it could have been applied to other repositories such as OLS or NCBO BioPortal. We considered that the OBO Foundry, OLS and NCBO BioPortal are general repositories in the area of biomedical and biological ontologies, so we selected one of them as representative of this category. AgroPortal is more oriented to agriculture; it potentially has a more different profile, which made it interesting for our study.

There is an overlap between the content of these repositories, since all of them contain the most relevant bio-ontologies such as the GO. The overlap between the two repositories used is around 25% of the size of AgroPortal, but the method applied to evaluate each metric in each corpus is not affected by such an overlap. Besides, the versions of such common ontologies in each repository were different in our experimental dataset, and their metrics were different. Hence, they are considered different ontologies in our study.

The results obtained in both repositories are similar. However, there are some differences due to the content of each repository. Consequently, some metrics could be appropriate for certain repositories and not for other ones. Our further work includes the analysis of the metrics of ontologies present in other repositories.

Impact and usefulness for the ontology community

The analysis of ontology metrics serves different purposes, since each metric provides certain information about a feature of the ontology. Hence, the results of this study should be useful for different types of users, among which we especially mention ontology developers, ontology users, ontology repository managers and developers of ontology metrics and ontology evaluation methods.

Ontology developers can use the metrics with the aim of comparing the properties of the ontologies against similar ones or to analyse the impact of their modelling decisions in such properties. For example, in [28, 29], ontology metrics are used to analyse the evolution of ontologies, and some findings identified that the drop in the values of some metrics were due to a major change in the modelling approach. That kind of support would not be possible without the metrics, and that is why we need to use reliable metrics. Reliability has to be interpreted in this work as metrics that provide stable and good classifications of ontologies.

Ontology users could use the values of reliable metrics to support an informed decision about which ontology to reuse. If we accept that users could make their decision based on the metrics of the existing ontologies, that would be another reason for ontology developers to care about.

Ontology repository managers should provide the information of the metrics to the users. These repositories contain many ontologies, so they provide a good context for a comparative evaluation of ontologies based on reliable metrics. Ontology repository managers could use our results to select which metrics are provided to the users in their repository and which ones could be the most interesting for analysing the repository content. In summary, reliable metrics help to make informed decisions. Finally, our method provides a way for analysing the reliability of a metric for a given repository. Our method could provide a benchmark for the developers of ontology metrics and the developers of ontology evaluation methods.

Limitations and further work

The major limitation of our approach is the lack of a gold standard. The ontology community has not developed a corpus of good ontologies that could be used for performing an external validation.

As we have already mentioned, our method can be applied to any repository of OBO/OWL ontologies. However, we recommend its usage on repositories with a process to control the addition of ontologies. The lack of such process would be a risk for the effective application of our method, thus probably getting biased results and misinterpretations.

The method presented in this paper should help anyone interested in applying a quantitative metric to evaluate ontologies. The method permits to study the properties of such metric using the content of ontology repositories. Hence, the reliability of the metric can be tested against different repositories. The method can be applied to other quantitative metrics if the distribution of values of the metric permits to create five clusters. It should be noted that we apply bootstrap resampling to calculate the degree of stability of a metric. This imposes the requirement of being able to obtain five clusters in each iteration, i.e. at least five different measurements of a metric are needed.

Currently, our method allows to achieve stable and well-structured categories, but the global stability could be improved by using the optimal number of categories for each metric. A detailed exploration of the Silhouette graphics shows that there exist some ontologies doubtfully classified in some clusterings (ontologies with low or negative Silhouette widths). We would expect this situation to be mitigated when using the optimal number of categories for each metric. Adjusting the method to work with the optimal number of categories will also permit to include more metrics in further studies.

Our method could be applied to metrics for resources that are not ontologies. An example could be the evaluation of the metrics that are currently being investigated and developed by the FAIR community for assessing the fairness of datasets [27].

Repositories usually store different types of ontologies (e.g. top-level versus domain ontologies or domain ontologies classified by subdomains). For example, CROnto, which has a strong structure according to our results, deals with individuals, which are not expected in some types of ontologies. The ontologies of a certain type could share some properties, so their optimal classification could be type dependent. Future work will include these aspects by the comparative analysis of the results for different number of categories for each metric and a comparative study of different repositories and types of ontologies.

We will also study corpora composed of versions of the same ontologies to study the reliability of the metrics for ontology evolution purposes.

Key Points

- We have evaluated relevant properties of the metrics for the evaluation of ontologies by using two corpora of ontologies, OBO Foundry and AgroPortal.
- The existing correlations between the metrics analysed would not bias the assessment of the quality of the ontologies.
- The clusterings generated by the dynamic scale are stable and are well-structured, which reinforce the usefulness of these metrics.
- This study is novel in the field of evaluation and classification of ontological structural metrics, and similar approaches might be used for other metrics.
- This kind of approach may well help users to understand the properties of the corpus under analysis, which can generate new insights in the properties of the ontologies of a repository.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

This work has been partially funded by the Spanish Ministry of Economy, Industry and Competitiveness, the European Regional Development Fund (ERDF) Programme and the Fundación Séneca through grants TIN2014-53749-C2-2-R, TIN2017-85949-C2-1-R and 19371/PI/14.

References

1. Legaz-García MC, Martínez-Costa C, Menárguez-Tortosa M, *et al*. A semantic web based framework for the interoperability and exploitation of clinical models and EHR data. *Knowl Based Syst* 2016;**105**:175–89.
2. Rubin DL, Shah NH, Noy NF. Biomedical ontologies: a functional perspective. *Brief Bioinformatics* 2007;**9**(1):75–90.
3. Viale P, Bora JJ, Benegui M, *et al*. Human endocrine system modeling based on ontologies. *Knowl Based Syst* 2016;**111**: 113–32.
4. Hoehndorf R, Schofield PN, Gkoutos GV. The role of ontologies in biological and biomedical research: a functional perspective. *Brief Bioinformatics* 2015;**16**(6):1069–80.
5. Smith B, Ashburner M, Rosse C, *et al*. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;**25**(11):1251–5.
6. Jonquet C, Toulet A, Arnaud E, *et al*. AgroPortal: a vocabulary and ontology repository for agronomy. *Comput Electron Agri* 2018;**144**:126–43.
7. Ong E, Xiang Z, Zhao B, *et al*. Ontobee: a linked ontology data server to support ontology term dereferencing, linkage, query and integration. *Nucleic Acids Res* 2017;**45**(D1):D347–52.
8. Côté R, Reisinger F, Martens L, *et al*. The ontology lookup service: bigger and better. *Nucleic Acids Res* 2010;**38**(suppl 2): W155–60.
9. Hoehndorf R, Slater L, Schofield PN, *et al*. Aber-owl: a framework for ontology-based data access in biology. *BMC Bioinformatics* 2015;**16**(1):26.
10. Whetzel PL, Noy NF, Shah NH, *et al*. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res* 2011;**39**(suppl 2): W541–5.
11. Neuhaus F, Vizedom A, Baclawski K, *et al*. Towards ontology evaluation across the life cycle. *Appl Ontol* 2013;**8**(3): 179–94.
12. Hoehndorf R, Dumontier M, Gkoutos GV. Evaluation of research in biomedical ontologies. *Brief Bioinformatics* 2012;**14**(6):696–712.
13. Gangemi A, Catenacci C, Ciaramita M, *et al*. Modelling ontology evaluation and validation. *The Semantic Web: Research and Applications. 3rd European Semantic Web Conference, ESWC 2006 Budva, Montenegro, June 11–14, 2006 Proceedings, 2006*. pp. 140–54. Springer Berlin Heidelberg, Berlin.
14. Rogers J. Quality assurance of medical ontologies. *Methods Inf Med* 2006;**45**(3):267–74.
15. Ceusters W. Applying evolutionary terminology auditing to the gene ontology. *J Biomed Inf* 2009;**42**(3):518–29.
16. Ceusters, W. Applying evolutionary terminology auditing to SNOMED CT. In: *AMIA Annual Symposium Proceedings, 2010*. Vol. 2010, p.96. American Medical Informatics Association, United States.
17. Duque-Ramos A, Fernández-Breis JT, Stevens R, *et al*. OQuaRE: a SQuaRE-based approach for evaluating the quality of ontologies. *J Res Prac Inf Tech* 2011;**43**(2):159–76.

18. Tartir S, Arpinar IB. Ontology evaluation and ranking using OntoQA. In: *ICSC '07: Proceedings of the International Conference on Semantic Computing, 2007*. pp. 185–92. IEEE Computer Society, Washington, DC, USA.
19. Yao H, Orme A, Etzkorn L. Cohesion metrics for ontology design and application. *J Comp Sci* 2005;**1**(1):107–13.
20. Ceusters W, Smith B A realism-based approach to the evolution of biomedical ontologies. In: *AMIA Annual Symposium Proceedings, 2006*. Vol. 2006, p.121. American Medical Informatics Association, United States.
21. Ashraf J, Chang E, Hussain OK, *et al*. Ontology usage analysis in the ontology lifecycle: a state-of-the-art review. *Knowl Based Syst* 2015;**80**:34–47.
22. McDaniel M, Storey VC, Sugumaran V. The role of community acceptance in assessing ontology quality. In: *International Conference on Applications of Natural Language to Information Systems*, 2016. pp. 24–36. Springer, Switzerland.
23. Poveda-Villalón M, Gómez-Pérez A, Suárez-Figueroa MC. Oops!(Ontology Pitfall Scanner!): an on-line tool for ontology evaluation. *Int J Semant Web Inf Syst* 2014;**10**(2):7–34.
24. Likert R. A technique for the measurement of attitudes. *Arch Psychol* 1932;**22**(140):1–55.
25. Chidamber SR, Kemerer CF. A metrics suite for object oriented design. *IEEE Trans Softw Eng* 1994;**20**(6): 476–93.
26. Li W. Another metric suite for object-oriented programming. *J Syst Softw* 1998;**44**(2):155–62.
27. Wilkinson MD, Sansone SA, Schultes E, *et al*. A design framework and exemplar metrics for FAIRness. *Sci Data* 2018;**5**: 180118.
28. Duque-Ramos A, Quesada-Martínez M, Iniesta-Moreno M, *et al*. Supporting the analysis of ontology evolution processes through the combination of static and dynamic scaling functions in oquare. *J Biomed Semant* 2016;**7**(1): 63–83.
29. Quesada-Martínez M, Duque-Ramos A, Iniesta-Moreno M, *et al*. Preliminary analysis of the OBO Foundry ontologies and their evolution using OQuaRE. *Stud Health Technol Inform* 2017;**235**:426–30.
30. Cheng R, Milligan GW. Measuring the influence of individual data points in a cluster analysis. *J Classification* 1996;**13**: 315–35.
31. Jaccard C. Distribution de la flore alpine dans le Basin de Dranses et dans quelques regions voisines. *Bulletin de la Societe Vaudoise des Sciences Naturelles* 1901;**37**:241–72.
32. Hennig C. Cluster-wise assessment of cluster stability. *Comput Stat Data Anal* 2007;**52**:258–71.
33. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987;**20**:53–65.
34. Calinski T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat* 1974;**3**(1):1–27.
35. Dunn JC. Well-separated clusters and optimal fuzzy partitions. *J Cybern* 1974;**4**(1):95–104.
36. Davies DL, Bouldin DW. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell* 1979;**PAMI-1**(2):224–7.
37. Lord E, Willems M, Lapointe F-J, *et al*. Using the stability of objects to determine the number of clusters in datasets. *Inf Sci* 2017;**393**:29–46.
38. Kaufman L, Rousseeuw P. *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, New Jersey, Canada: Wiley, 1990.
39. Horridge M, Bechhofer S. The OWL API: a Java API for OWL ontologies. *Semantic Web* 2011;**2**(1):11–21.
40. Kazakov Y, Krötzsch M, Simančík F. Elk reasoner: architecture and evaluation. In: *Proceedings of the 1st International Workshop on OWL Reasoner Evaluation, 2012*. CEUR-WS.org.
41. Shearer R, Motik B, Horrocks I. HermiT: a highly- efficient OWL reasoner. In: *OWLED*, Vol. 432, 2008, p. 91.
42. Team RC. *R language definition*. Vienna, Austria: R Foundation for Statistical Computing, 2000.
43. Wei T, Simko V. corrplot: Visualization of a Correlation Matrix. R package version 0.77, 2016.
44. Hennig, C. fpc: Flexible Procedures for Clustering. R package version 2.1–10, 2015.
45. Maechler M, Rousseeuw P, Struyf A, *et al*. cluster: Cluster Analysis Basics and Extensions. R package version 2.0.6, 2017.