Submitted to Briefings in Bioinformatics, YYYY, 0–0
doi: 10.1093/bib/xxxxx
Advance Access Publication Date: DD Month YYYY

Evaluation of ontology structural metrics based on public repository data

Manuel Franco[1], Juana-María Vivo[1], Manuel Quesada-Martínez[2], Astrid Duque-Ramos[3] and Jesualdo Tomás Fernández-Breis[4,*]

[1]Departamento de Estadística e Investigación Operativa, Universidad de Murcia, 30100, Murcia, Spain

[2]Center of Operations Research (CIO), Miguel Hernández University of Elche, 03202, Elche, Spain

[3]Departamento de Sistemas, Facultad de Ingenierías, Universidad de Antioquia, Medellín, 050010,Colombia

[4]Departamento de Informática y Sistemas, Universidad de Murcia, IMIB-Arrixaca, 30100, Murcia, Spain

* **Corresponding author**: Jesualdo Tomás Fernández-Breis, Departamento de Informática y Sistemas, Universidad de Murcia, IMIB-Arrixaca, 30100, Murcia, Spain, email:jfernand@um.es, phone: +34868884613

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

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 last 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.

Keywords: Biological ontologies, quantitative metrics, metrics comparison, data analysis
Supplementary information: Supplementary data are available at *Briefings in Bioinformatics* online.

1 Introduction

The development and application of biological ontologies have increased significantly in recent years [24, 36, 41]. 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 [18]. 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 [38], AgroPortal [21], OntoBee [31], the Ontology Lookup Service (OLS) [8], AberOWL [17], or BioPortal [43]. The OBO Foundry [38] is likely to be the largest initiative that pursues the collaborative development of biomedical ontologies by applying shared modelling principles such as open use, collaborative development, non-

overlapping, strictly-scoped content or common syntax and relations[1]. In this case, the quality of an ontology is checked by hand and related to the adherence and application of their set of design principles, which is a hard and tedious task. This example illustrates the benefit of the availability of automatic methods to provide information about the quality of the ontologies.

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 [30, 36], but there has been no community agreement so far [16]. However, the ontology engineering community has proposed both qualitative [13, 34] and quantitative approaches [3, 4, 11, 39, 44]. Gangemi et al. [13] propose a diagnostic task based on ontology descriptions, using three categories of criteria (structural, functional and usability profiling). Rogers [34] applies four qualitative criteria (philosophical rigour, ontological commitment, content correctness, and fitness for a purpose). Yao et al. [44] and Tartir and Arpinar [39] define a series of metrics for evaluating structural properties in the ontology. Works like [3–5] evaluate the ontology from a realism-based perspective that demands manual judgement of users. In addition, works like [1, 29, 32, 39, 44] 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 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, ..., 5}, analogously to the standard Likert scale [26]. 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, which allow to analyse corpora formed by 119 and 78 ontologies, respectively. The OBO Foundry repository has been selected because their ontologies are supposed to share certain building principles, which makes us think that it constitutes a repository of homogeneous ontologies. AgroPortal contains vocabularies and ontologies for agronomy, food, plant sciences and biodiversity [21], 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 stable metrics for classifying ontologies. We believe that this work allows to generate new insights in the field of ontology engineering and to shed light on ontology evaluation methods.

[1] http://www.obofoundry.org/principles/fp-000-summary.html

2 Methods

2.1 Metrics and scaling function

In this work we focus on 19 ontology structural metrics (Table 1) which 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 which 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, being a function $n(f(x))$ 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(f(x)) partitions the range of $f(x)$ in 5 non-prefixed continuous intervals that 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 to associate. 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, which is not necessarily associated with the highest values of a particular metric.

An ontology set $\theta = \{\theta_1, \dots, \theta_n\}$ is received as input and generates a vector of raw values $R_{M_i} = \{R_{\theta_1}, \dots, R_{\theta_n}\}$ for each metric in $M = \{M_1, \dots, M_m\}$. The application of a scaling function transforms R_{M_i} vectors into a scaled vector $N_{M_i} = \{N_{\theta_1}, \dots, N_{\theta_n}\}$. This dynamic scale has been used to analyse the evolution of ontologies, using as experimental data those obtained processing different versions of the same ontology [12, 33].

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 5 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 1 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 5 categories of ontologies for the M_i metric which are determined by the N_{M_i} scores.

2.2 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 which will be useful for validating metrics.

2.3 Validation of the clusters obtained us-ing 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 function: stability of the clusters, and validity of the clusters. We describe next the methods used for both studies.

2.3.1 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 $(C_1, \dots, C_5)$ when the sample varies [6]. We can apply a bootstrap resampling method to assess the stability of each category of the dynamic scale clustering, $S_{M_i}(C_j)$ for $j = 1, \dots, 5$, for each metric M_i, based on a similarity measure between sets, called Jaccard coefficient [20], as described by Hennig [14]. 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, $S_{M_i}(C_j)$ 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 $S_{M_i}(C_j)$ for $j = 1, \dots, 5$, by setting $b = 50, 100, 500$ and 1000, respectively. For interpretation purpose, we use the $S_{M_i}(C_j)$ scores to classify the categories as follows:

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

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, $S(M_i)$ for $i = 1, \dots, m$. 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.

2.3.2 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*) [35], Calinski-Harabasz (*ch*) [2], Dunn (*dunn*) [10], and Davies-Boudin (*db*) [9] 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 [27], 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.

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

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

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, $sil_{M_i}(\theta_l)$ measures how well it has been classified, which can be interpreted as in [35]. 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, $\overline{sil}(M_i) = \sum_{l=1}^{n} sil_{M_i}(\theta_l)/n$, for $i = 1, \dots, m$. Kaufman and Rousseeuw [22] 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 $\overline{sil}(M_i) \in [-1, 0.25]$.
- The clustering structure is weak and could be artificial when $\overline{sil}(M_i) \in (0.25, 0.50]$.
- There is a reasonable clustering structure when $\overline{sil}(M_i) \in (0.50, 0.70]$.
- A strong clustering structure has been found when $\overline{sil}(M_i) \in (0.70, 1.00]$.

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".

2.4 Experimental setup

In this work we have focused in two corpora of ontologies: the OBO Foundry ($n = 119$) and the AgroPortal ($n = 78$). For each ontology, we searched for its latest version in each repository. The whole description of the corpora can be found in Supplementary File 1 and Supplementary File 2.

We applied the OQuaRE platform[2] for the calculation of metrics. The n sizes of both corpora are just those ontologies correctly processed by this platform. This platform uses the OWL API [19] and Neo4j[3]. 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 [23] and Hermit [37]; for this experiment we selected the ELK reasoner, which works with the OWL 2 EL profile[4]. We processed the XML file, extracted the metrics raw scores and used R [40] for performing the statistical analysis. In particular, we used the following R packages for the statistical analysis: *corrplot* for correlations [42], *fcp* for stability analysis [15] and *cluster* for Silhouette graphics and validity analysis [28].

[2] http://sele.inf.um.es/oquare
[3] https://neo4j.com/
[4] https://www.w3.org/TR/owl2\discretionary- profiles/ #OWL_2_EL

3 Results

3.1 Correlations between metrics

Figure 2 displays the correlations between pairs of metrics, using the raw values obtained for the whole ontology set of OBO Foundry (Figure 2(a)) and AgroPortal (Figure 2(b)) repositories. The 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 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 with 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 3 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.

3.2 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 4. The convergence of the stability indexes can be observed when 500 replicates are used. The detailed results for the rest of metrics on OBO Foundry and AgroPortal corpora can be found in Supplementary File 3.

According to Figure 4, 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 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 5 categories, although a detailed analysis on the category stability scores shows that there is 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).

3.3 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 5 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 4 and Supplementary File 5. 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 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 provides 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 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 5(a)). 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-classifed by WMCOnto. However, it is formed by two ontologies with Silhouette score 0.87 in AgroPortal and thus both ontologies are well-classifed 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.

4 Discussion and Perspectives

The increasing interest in ontologies makes necessary to develop effective quantitative methods for ontology evaluation. Reaching a communi-

ty 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, the correlations do not depend on the corpus of ontologies used since we obtain similar results for the two corpora analysed here, so 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). The normalisation of metrics would avoid computing unnecessary metrics, which would contribute to the performance of the execution, specially in corpora including a large set of ontologies. However, we do not recommend to remove metrics, but to provide users with mechanisms to select the more explanatory 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 which 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 well-classified ontologies and well-structured categories. 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. 13 out of 19 metrics have the same classification in both repositories. Also, the information from both repositories can be combined to select the three strongest structured metrics (AROnto, CROnto and TMOnto2). These metrics are related to the ratio of attributes, individuals and direct ancestors, which are relevant ontology features.

As it has been mentioned, 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. There is a number of ontologies common to the OBO Foundry and AgroPortal repositories. 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 of this study should be useful for different types of users, among which we especially mention ontology repository managers and ontology users. 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. Ontology users could drive their attention to the metrics that provide a better classification when, for instance, evaluating or selecting ontologies for reuse.

Currently, our method allows to achieve stable and good 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). Moreover, ontology repositories usually store different types of ontologies (e.g. top-level vs domain ontologies or domain ontologies classified by subdomains). For example, one of the strongest metric is CROnto, which deals with individuals, which are not expected in some types of ontologies. We suspect that the ontologies of a certain type could share different properties, so their optimal classification could be different as well. 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.

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.

Funding

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

- **Manuel Franco** is a Professor in the Department of Statistics and Operational Research at the University of Murcia.
- **Juana-María Vivo** is a Professor in the Department of Statistics and Operational 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 Project Manager 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 member of the IMIB-Arrixaca Bio-Health Research Institute.

Conflict of Interest: none declared.

References

1. Ashraf, J., Chang, E., Hussain, O. K., and Hussain, F. K. (2015). Ontology usage analysis in the ontology lifecycle: A state-of-the-art review. Knowledge-Based Systems, 80, 34–47.
2. Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics, 3(1), 1–27.
3. Ceusters, W. (2009). Applying evolutionary terminology auditing to the gene ontology. Journal of Biomedical Informatics, 42(3), 518 – 529.
4. Ceusters, W. (2010). Applying evolutionary terminology auditing to SNOMED CT. In AMIA annual symposium proceedings, volume 2010, page 96. American Medical Informatics Association.
5. Ceusters, W. and Smith, B. (2006). A realism-based approach to the evolution of biomedical ontologies. In AMIA Annual Symposium Proceedings, volume 2006, page 121. American Medical Informatics Association.
6. Cheng, R. and Milligan, G. W. (1996). Measuring the influence of individual data points in a cluster analysis. Journal of Classification, 13, 315–335.
7. Chidamber, S. R. and Kemerer, C. F. (1994). A metrics suite for object oriented design. IEEE Transactions on Software Engineering, 20(6), 476–493.
8. Côté, R., Reisinger, F., Martens, L., Barsnes, H., Vizcaino, J. A., and Hermjakob, H. (2010). The ontology lookup service: bigger and better. Nucleic Acids Research, 38(suppl 2), W155–W160.
9. Davies, D. L. and Bouldin, D. W. (1979). A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2), 224–227.
10. Dunn, J. C. (1974). Well-Separated Clusters and Optimal Fuzzy Partitions. Journal of Cybernetics, 4(1), 95–104.
11. Duque-Ramos, A., Fernández-Breis, J. T., Stevens, R., and Aussenac- Gilles, N. (2011). OQuaRE: A SQuaRE-based approach for evaluating the quality of ontologies. Journal of Research and Practice in Information Technology, 43(2), 159–176.
12. Duque-Ramos, A., Quesada-Martínez, M., Iniesta-Moreno, M., Fernández-Breis, J. T., and Stevens, R. (2016). Supporting the analysis of ontology evolution processes through the combination of static and dynamic scaling functions in oquare. Journal of Biomedical Semantics, 7(1), 63–83.
13. Gangemi, A., Catenacci, C., Ciaramita, M., and Lehmann, J. (2006). The Semantic Web: Research and Applications: 3rd European Semantic Web Conference, ESWC 2006 Budva, Montenegro, June 11-14, 2006 Proceedings, chapter Modelling ontology evaluation and validation pages 140–154. Springer Berlin Heidelberg, Berlin.
14. Hennig, C. (2007). Cluster-wise assessment of cluster stability. Computational Statistics & Data Analysis, 52, 258–271.
15. Hennig, C. (2015). fpc: Flexible Procedures for Clustering. R package version 2.1-10.
16. Hoehndorf, R., Dumontier, M., and Gkoutos, G. V. (2012). Evaluation of research in biomedical ontologies. Briefings in Bioinformatics, 14(6), 696–712.
17. Hoehndorf, R., Slater, L., Schofield, P. N., and Gkoutos, G. V. (2015). Aber-owl: a framework for ontology-based data access in biology. BMC Bioinformatics, 16(1), 26.
18. Hoehndorf, Robert and Schofield, Paul N. and Gkoutos, Georgios V. (2015). The role of ontologies in biological and biomedical research: a functional perspective. Briefings in Bioinformatics, 16(6), 1069–1080.
19. Horridge, M. and Bechhofer, S. (2011). The OWL API: A Java API for OWL Ontologies. Semantic Web, 2(1), 11–21.
20. Jaccard, C. (1901). Distribution de la flore alpine dans le Basin de Dranses et dans quelques regions voisines. Bulletin de la Societe Vaudoise des Sciences Naturelles, 37, 241–272.
21. Jonquet, C., Toulet, A., Arnaud, E., Aubin, S., Yeumo, E. D., Emonet, V., Graybeal, J., Laporte, M.-A., Musen, M. A., Pesce, V., and Larmande, P. (2018). AgroPortal: A vocabulary and ontology repository for agronomy. Computers and Electronics in Agriculture, 144, 126–143.
22. Kaufman, L. and Rousseeuw, P. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley.
23. Kazakov, Y., Krötzsch, M., and Simanc˘ík, F. (2012). Elk reasoner: Architecture and evaluation. In Proceedings of the 1st International Workshop on OWL Reasoner Evaluation, pages –.
24. Legaz-García, M. C., Martínez-Costa, C., Menárguez-Tortosa, M., and Fernández-Breis, J. T. (2016). A semantic web based framework for the interoperability and exploitation of clinical models and EHR data. Knowledge-Based Systems, 105, 175–189.
25. Li, W. (1998). Another metric suite for object-oriented programming. 44, 155–162.
26. Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology, 22(140), 1–55.
27. Lord, E., Willems, M., Lapointe, F.-J., and Makarenkov, V. (2017). Using the stability of objects to determine the number of clusters in datasets. Information Sciences, 393, 29–46.
28. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2017). cluster: Cluster Analysis Basics and Extensions. R package version 2.0.6.
29. McDaniel, M., Storey, V. C., and Sugumaran, V. (2016). The Role of Community Acceptance in Assessing Ontology Quality. In International Conference on Applications of Natural Language to Information Systems, pages 24–36. Springer.
30. Neuhaus, F., Vizedom, A., Baclawski, K., Bennett, M., Dean, M., Denny, M., Grüninger, M., Hashemi, A., Longstreth, T., Obrst, L., et al. (2013). Towards ontology evaluation across the life cycle. Applied Ontology, 8(3), 179–194.
31. Ong, E., Xiang, Z., Zhao, B., Liu, Y., Lin, Y., Zheng, J., Mungall, C., Courtot, M., Ruttenberg, A., and He, Y. (2017). Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration. Nucleic Acids Research, 45(D1), D347–D352.
32. Poveda-Villalón, M., Gómez-Pérez, A., and Suárez-Figueroa, M. C. (2014). Oops!(ontology pitfall scanner!): An on-line tool for ontology evaluation. International Journal on Semantic Web and Information Systems (IJSWIS), 10(2), 7–34.
33. Quesada-Martínez, M., Duque-Ramos, A., Iniesta-Moreno, M., and Fernández-Breis, J. T. (2017). Preliminary Analysis of the OBO Foundry Ontologies and Their Evolution Using OQuaRE. Informatics for Health: Connected Citizen-Led Wellness and Population Health, 235, 426–430.
34. Rogers, J. (2006). Quality assurance of medical ontologies. Methods of Information in Medicine, 45(3), 267–274.
35. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65.
36. Rubin, D. L., Shah, N. H., and Noy, N. F. (2007). Biomedical ontologies: a functional perspective. Briefings in Bioinformatics, 9(1), 75–90.
37. Shearer, R., Motik, B., and Horrocks, I. (2008). HermiT: A Highly- Efficient OWL Reasoner. In OWLED, volume 432, page 91.
38. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology, 25(11), 1251–1255.
39. Tartir, S. and Arpinar, I. B. (2007). Ontology evaluation and ranking using OntoQA. In ICSC '07: Proceedings of the International Conference on Semantic Computing, pages 185–192, Washington, DC, USA. IEEE Computer Society.
40. Team, R. C. (2000). R language definition. Vienna, Austria: R foundation for statistical computing.
41. Viale, P., Bora, J. J., Benegui, M., and Basualdo, M. (2016). Human endocrine system modeling based on ontologies. Knowledge-Based Systems, 111, 113–132.
42. Wei, T. and Simko, V. (2016). corrplot: Visualization of a Correlation Matrix. R package version 0.77.
43. Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., and Musen, M. A. (2011). BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Research, 39 (suppl 2), W541–W545.
44. Yao, H., Orme, A., and Etzkorn, L. (2005). Cohesion metrics for ontology design and application. Journal of Computer Science, 1(1), 107–113.

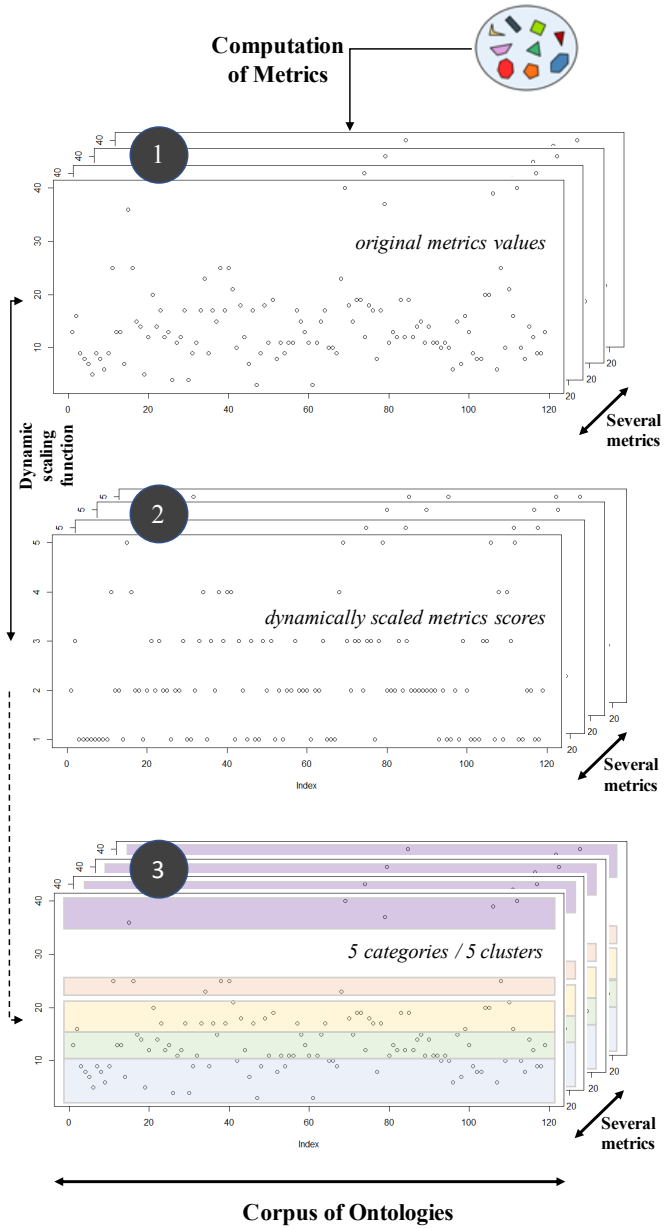

Figure 1. Graphical representation of the application of the dynamic scaling function using a corpus.

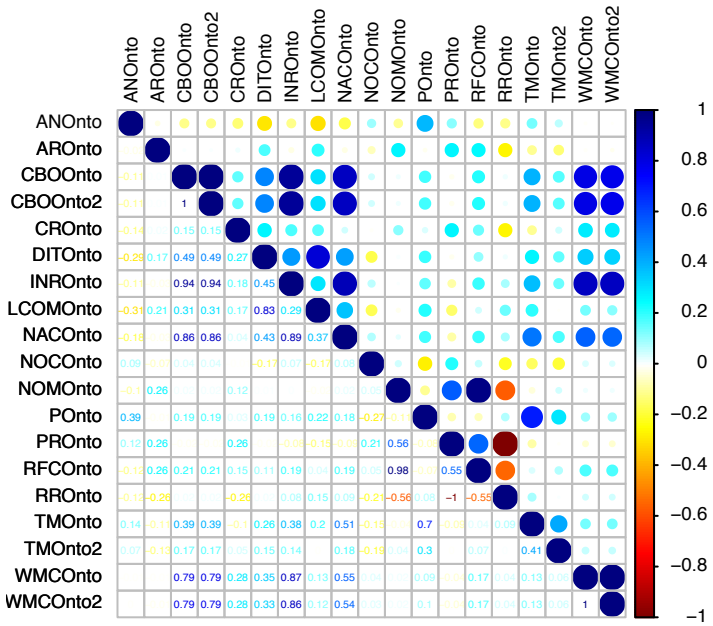

Figure 2 a). Pearson's correlation coefficient between metrics: OBO Foundry.

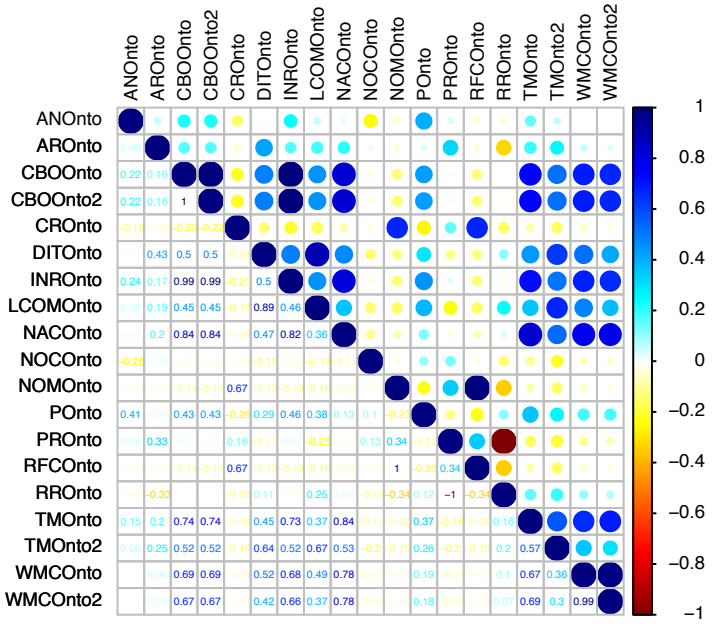

Figure 2 b). Pearson's correlation coefficient between metrics: AgroPortal.

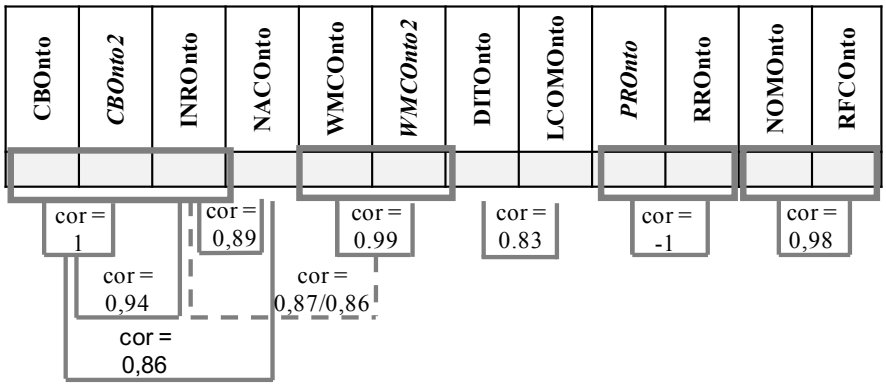

Figure 3 a). Pairs of metrics with correlations higher than 0.8 in absolute value: OBO Foundry.

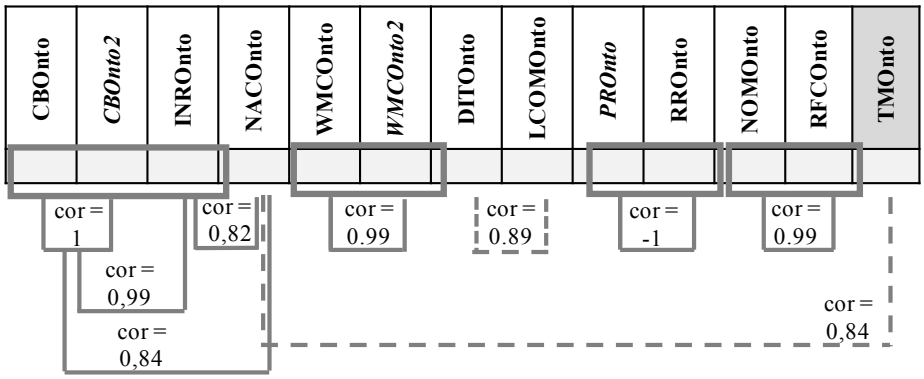

Figure 3 b). Pairs of metrics with correlations higher than 0.8 in absolute value: AgroPortal.

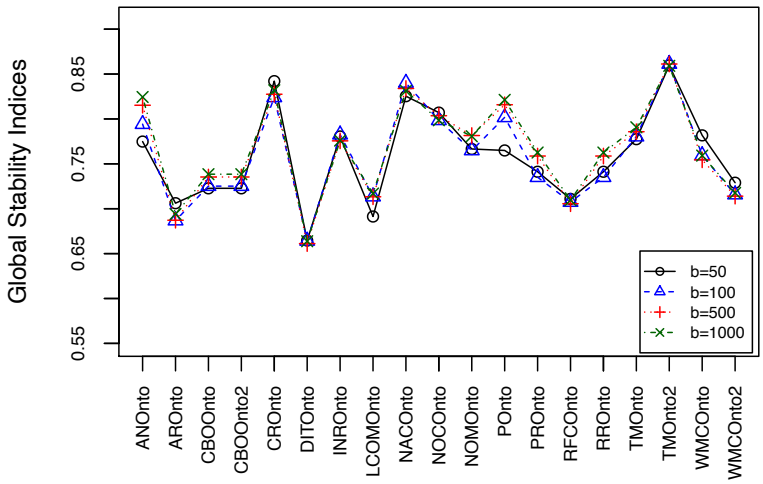

Figure 4 a). Category and global stability scores of ANOnto and AROnto metrics for b=50,100,500,1000: OBO Foundry.

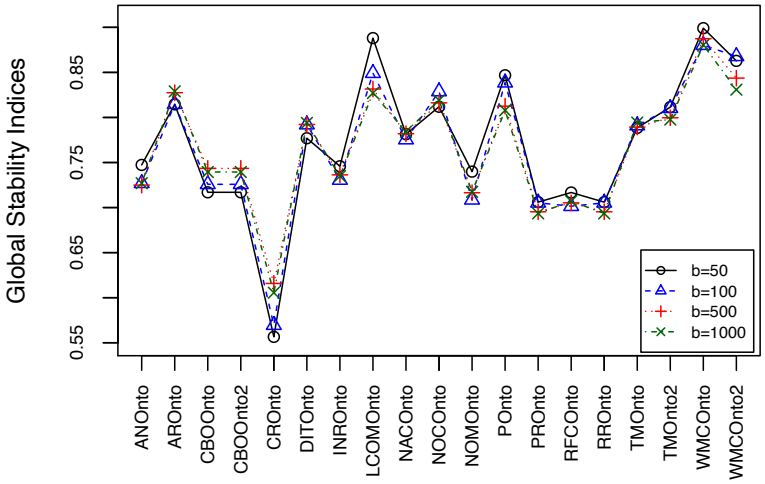

Figure 4 b). Category and global stability scores of ANOnto and AROnto metrics for b=50,100,500,1000: AgroPortal.

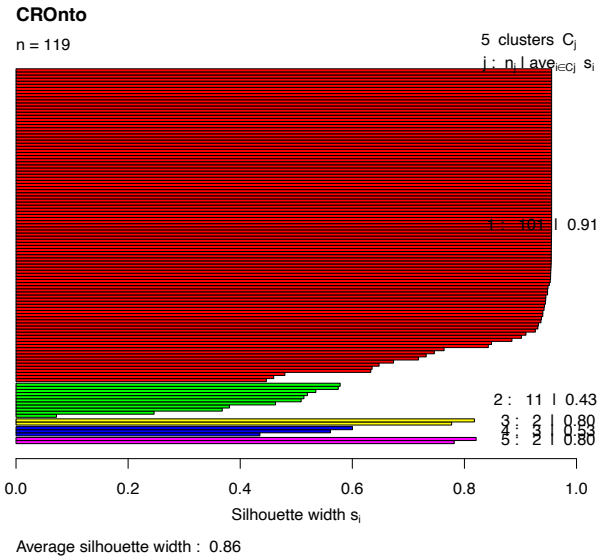

Figure 5 a). Silhouette graphics of three selected metrics representing different behaviours: OBO Foundry – CROnto.

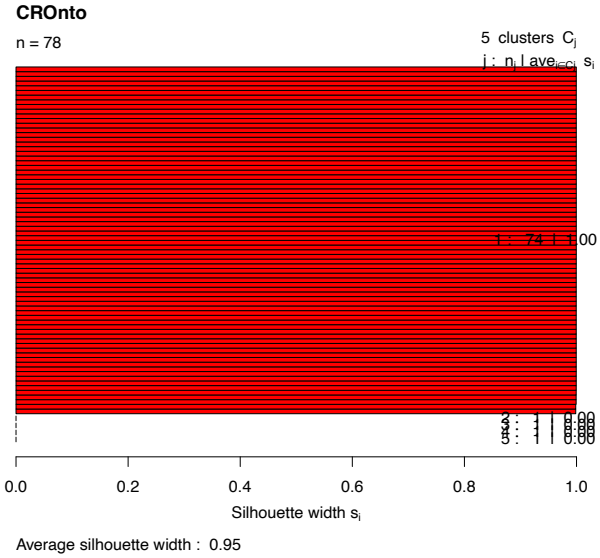

Figure 5 b). Silhouette graphics of three selected metrics representing different behaviours: AgroPortal – CROnto.

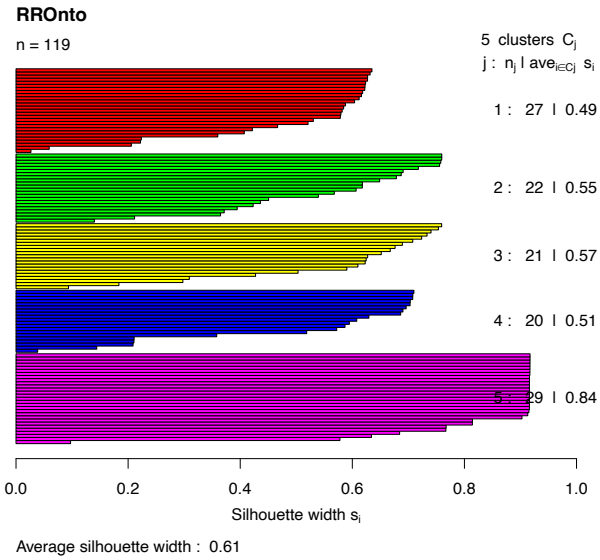

Figure 5 c). Silhouette graphics of three selected metrics representing different behaviours: OBO Foundry – RROnto.

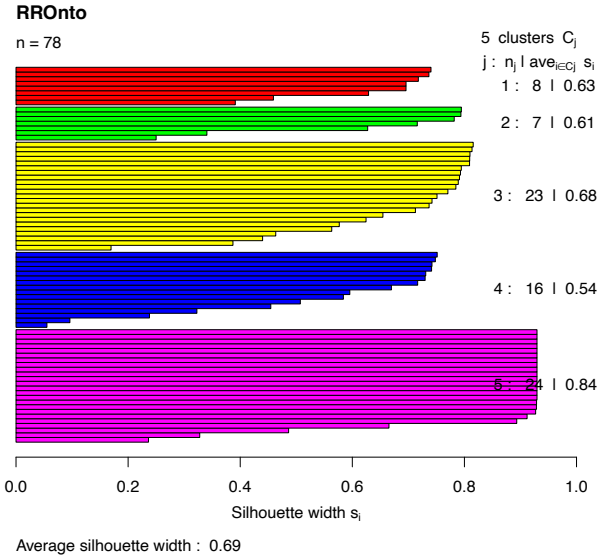

Figure 5 d). Silhouette graphics of three selected metrics representing different behaviours: AgroPortal – RROnto.

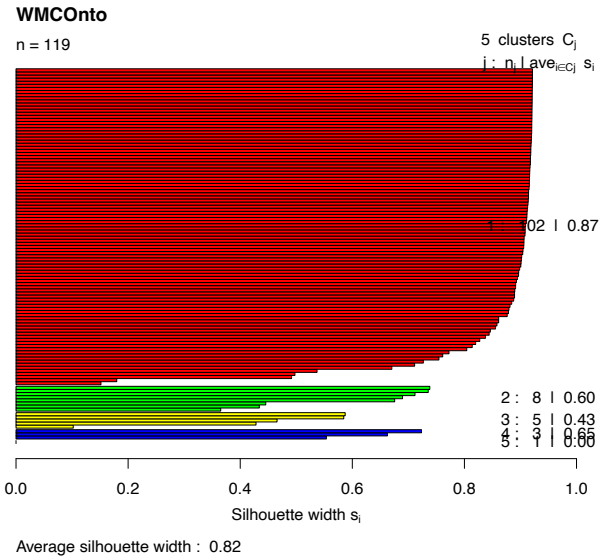

Figure 5 e). Silhouette graphics of three selected metrics representing different behaviours: OBO Foundry – WMCOnto.

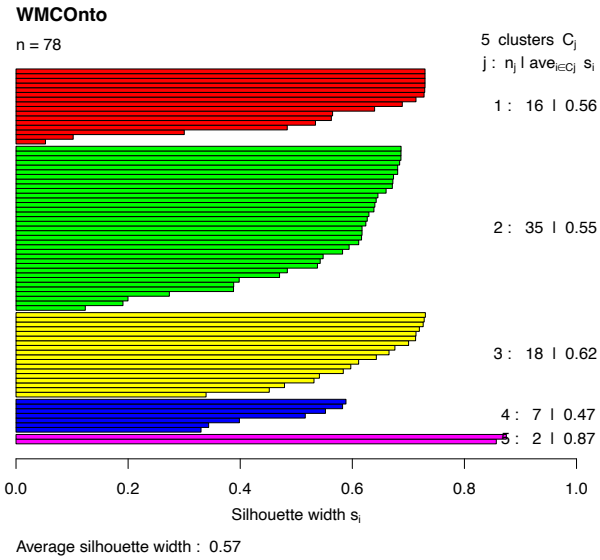

Figure 5 f). Silhouette graphics of three selected metrics representing different behaviours: AgroPortal – WMCOnto.