IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 11 July 2023, accepted 12 September 2023, date of publication 18 September 2023,
date of current version 21 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3316512

RESEARCH ARTICLE

HURON: A Quantitative Framework for Assessing Human Readability in Ontologies

FRANCISCO ABAD-NAVARRO, CATALINA MARTÍNEZ-COSTA, AND JESUALDO TOMÁS FERNÁNDEZ-BREIS, (Senior Member, IEEE)
Departamento de Informática y Sistemas, Universidad de Murcia, CEIR Campus Mare Nostrum, IMIB-Arrixaca, 30100 Murcia, Spain

Corresponding author: Jesualdo Tomás Fernández-Breis (jfernand@um.es)

This work was supported in part by MCIN/AEI/10.13039/501100011033 under Grant PID2020-113723RB-C22 and Grant RYC2020-030190-I, in part by the Horizon Europe HORIZON-HLTH-2021-TOOL-06-03 under Grant 101057603, and in part by Horizon-HLTH-2022-Tool-12-Two-Stage under Grant 101080875.

ABSTRACT The increasing use of ontologies requires their quality assurance. Ontology quality assurance consists of a set of activities that allow analyzing the ontology, identifying strengths and weaknesses, and proposing improvement actions. Human readability is a quality aspect that improves the use and reuse of ontologies. Human readable content refers to the natural language content consumed by humans and by the growing number of embedding methods applied to ontologies. The ontology community has proposed best practices for human readability, but there is no standardized framework for its evaluation. We aim to provide a framework for analyzing the human readability based on quantitative metrics to support ontology developers' decisions. We present the HURON framework, which consists of the specification of five quantitative metrics related to the human readability of ontology content and a software tool to implement them. The metrics take into account the number of names, descriptions, or synonyms, and also assess the application of systematic naming conventions and the 'lexically suggest, logically define' principle. Target values are provided for each metric to help to interpret them. HURON can also be used to assess compliance with best practices. We have applied HURON to a representative set of biomedical ontologies, the OBO Foundry repository. The results showed that, in general, the OBO Foundry ontologies comply with the expected number of descriptions and names in their classes, and both lexical and semantically formalized contents are aligned. However, most of the ontologies did not follow a systematic naming convention. In general, the ontologies in this repository show adherence to some of the best practices, although areas for improvement were identified. A number of recommendations are made for ontology developers and users.

INDEX TERMS Knowledge engineering, ontologies, quality assurance, readability metrics, semantic web.

I. INTRODUCTION

Ontologies play a key role in knowledge engineering by providing a common conceptualization of a domain. They have been successfully applied in various domains, but especially in biology and biomedicine, with different purposes [1], [2], [3], [4], [5].

At the time of writing, repositories such as BioPortal [6] had more than 1,000 ontologies and both the Open Biological and Biomedical Ontology (OBO) Foundry [7] and the Ontology Lookup Service (OLS) [8] had more than 250 ontologies. The number of ontologies in these repositories continues to grow, demonstrating their relevance and impact.

Unlike other artifacts used in data management systems, such as relational databases, which are developed specifically for particular applications, ontologies should be created in a standardized way to facilitate their reuse. The sharing and reuse orientation of ontologies has also made them fundamental for achieving Findable, Accessible, Interoperable, and Reusable (FAIR) datasets [9]. As a result, ensuring the quality of ontologies has become an important need.

Ontology quality assurance consists of a set of activities that allow analyzing the ontology, identifying strengths and

The associate editor coordinating the review of this manuscript and approving it for publication was Mansoor Ahmed.

weaknesses, and proposing and implementing improvement actions. It should be noted that the quality of the ontology has to be analyzed in terms of its requirements. In recent years, several approaches have been proposed to study the quality of ontologies. For example, Vrandečić [10] proposed to evaluate the accuracy, the completeness, the conciseness, the consistency, the computational efficiency, the adaptability, and the clarity. Later, our research group proposed the OQuaRE framework [11], [12], [13], which classifies these characteristics in aspects related to the ontology structure, its functional adequacy, compatibility, transferability, operability, and its quality in use. The evaluation of these aspects is a time-consuming task that usually needs of domain experts. Fortunately, the community is evolving towards the use of metrics for measuring the quality of ontologies [14], [15], [16]. This is in line with the increasing use of metrics for supporting the development of methods for the analysis of data repositories or software systems [17], [18], [19], [20].

It should be noted that the content of ontologies must be understandable by both humans and machines. In the context of software applications, the content of ontologies must be processed and used by the machine to perform automated tasks, but it may also be displayed through the user interface, so it must be understandable by humans. This means that ontology evaluation should consider both human-readable and machine-readable content. In Web Ontology Language (OWL) ontologies, annotation properties typically provide human-readable information (also known as lexical content) for describing entities. These annotation properties allow the definition of labels, synonyms, descriptions, etc., which are not used in Description Logic (DL) reasoning [21], but provide relevant information for using the ontology. These annotation properties were used in [22] to define metrics for assessing structural accuracy and readability in the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) [23].

Previous works (see, for example [24]) have shown that ontologies contain human-readable content that is not expressed in a logical form for the machines, this is the so-called 'hidden semantics', which is knowledge captured in the content for humans but not expressed in a machine-processable way. The consistency between both the lexical and the axiomatic content of the ontology can be checked by applying the principle 'lexically suggest, logically define' [25]. An example is the name of the concept *'left atrium endocardium'*, which can suggest a semantic relationship with the concepts *'left'* and *'endocardium'*. This principle has been used for ontology analysis and axiom discovery [26], [27].

Both lexical and semantic ontology content is used by algorithms derived from Natural Language Processing (NLP) techniques such as Word2Vec [28], OWL2Vec* [29] or RDF2vec [30], to represent ontology entities as numerical vectors. These vectors can be further used for measuring similarity between entities [31], [32], or for predictive tasks [33], [34], [35]. In this sense, the human-readable content of the ontologies is a key factor to improve the performance of these techniques, since most of them use this content to create a corpus of text to be used as input for pure NLP-based techniques, obtaining the vectors from this text.

In the literature, best practices related to the human-readable content of ontologies have been proposed (see e.g. [25], [36]). However, to the best of our knowledge, existing ontology evaluation frameworks have not proposed a systematic way to evaluate these best practices or the human readability of ontology content. Consequently, there is a lack of systematic methods for detecting ontologies and specific entities with readability problems. Making this information available would allow developers to improve their ontologies based on informed decisions. In addition, this information would also be helpful to users when deciding which ontology to reuse.

Our hypothesis is that we can define quantitative metrics that capture information that supports the assessment of human readability of ontologies and adherence to best practices. The availability of a set of standardized human readability metrics would contribute to a more comprehensive evaluation of biomedical ontologies. It would also provide the basis for developing a fundamental understanding of the human readability of ontologies, which could be relevant to related areas such as natural language processing. To this end, we propose a set of metrics to capture a set of best practices related to the human-readable content of ontologies.

The main research question (RQ1) is to what extent quantitative metrics are useful for assessing the human readability of ontologies and the adherence to related best practices. For this purpose, we will use the OBO Foundry repository as a use case. The OBO Foundry [7] is a bioontology community pursuing the development of an orthogonal, interoperable collection of bioontologies. The OBO Foundry has defined a set of principles to be followed by ontologies developed by the community. These principles are the closest thing we know to an evaluation of these best practices. The principles deal with aspects such as the openness of the ontologies, the format, the construction of the Uniform Resource Identifier (URI), the versioning of the ontologies, the use of formal relations, the need for maintenance and collaboration, or how to define definitions and names, the latter two being directly related to human readability. The second research question (RQ2) is what recommendations can be derived from our study to help ontology developers improve the human readability of ontologies. The third research question (RQ3) is to what extent the OBO Foundry ontologies show compliance with the best practices. Regarding RQ3, the OBO Foundry has recently made an effort to encode the principles as operational rules that can be tested by validation checks [37], which is a step forward in ontology quality assurance, and will serve to compare the information about the ontologies generated by both approaches.

Therefore, our work will make the following contributions: (1) we will provide a framework consisting of a set of metrics for systematically evaluating the human-readable content of biomedical ontologies; (2) we will provide a tool for systematically evaluating the human-readable content of biomedical ontologies; (3) we will generate knowledge about the human-readable content of biomedical ontologies in a well-known repository; and (4) we will provide some recommendations on how ontology developers could improve the human-readability of ontologies. Finally, although the framework has been applied to biomedical ontologies in this work, it could be useful for other domains and also for other types of semantic resources.

II. BEST PRACTICES ON HUMAN READABLE CONTENT

In this section we describe some best practices related to human-readable content of ontologies proposed by the community. In this study, we assume OWL2 ontologies.[1] OWL2 ontologies consist of a set of axioms that describe entities (e.g., classes, individuals) and relationships between entities (via properties). Classes are described by three types of properties: data type properties, object properties, and annotation properties. Annotation properties are the ones used to provide human-readable content, so they are the relevant ones for this work. Annotation properties are not used to reason with OWL2 content. OWL2 provides some built-in annotation properties, such as *rdfs:label* for assigning names to classes, *rdfs:comment* for providing descriptions, but none specific to synonyms. The content of the annotation properties can have a language associated with it, allowing multilingual ontologies to be defined. There is no cardinality limit on the number of instances of the same annotation property associated with the same class, allowing the definition of multiple labels or comments associated with the same language. In addition, OWL2 ontology developers can create their own annotation properties.

A. CLASS NAMES

In this section we address three best practices related to the naming of classes:

- *Number of names in classes*: each class must have one name.
- *Naming style*: class names must define the concept represented as well as possible, using a systematic nomenclature [36].
- *Lexically suggest, logically define*: the information that can be inferred from class names must be reflected by ontology axioms [25].

Each ontology class is expected to have a standardized canonical name consisting of a comprehensive textual representation that is easily understood by users from different backgrounds. This is usually provided by adding annotation properties such as *rdfs:label*, *skos:prefLabel*, or *foaf:name* to the class. In some ontologies, classes may have more than one name due to multilingualism or the use of multiple properties derived from design choices. For example, the SNOMED CT OWL version [38] uses the annotation property *skos:prefLabel* to represent the class name, e.g. 'genetic disease', and the property *rdfs:label* to concatenate the semantic type to the class name, e.g. 'genetic disease (disorder)'.

This canonical name must follow a systematic naming convention that defines the concept as well as possible. For example, the class name 'juvenile osteochondrosis of the foot' may be used to indicate that this is a type of osteochondrosis of the foot that affects young people. However, this class is also known as 'Kohler's disease', from which it can only be inferred that it is a disease discovered by Kohler, which is less meaningful to non-experts in the field.

The OBO Foundry Principle 12 (Naming Conventions) [36] provides a set of rules and tips for good class naming. These include:

- Use explicit and concise names: Keep names short and memorable, but precise enough to capture the intended meaning. For example, the aforementioned class name 'juvenile osteochondrosis of the foot' captures the meaning of the concept better than the name 'Kohler's disease'.
- Use names that are self-explanatory and understandable when viewed outside the immediate context of the ontology. An example of this can be seen in the Disease Ontology excerpt in Figure 1, where classes are shown as blue circles, human-readable content as green squares, and the annotation properties used to link them as arrows. Here, the class labeled *disease by infectious agent* satisfies the rule because its name is understandable by itself without knowing the context of the class. Conversely, if the class had been named 'by infectious disease', the context provided by the parent class (*disease*) would be required to know that the class is related to diseases, which is against the rule.
- Recycle strings: Word compositions should be constructed in a consistent manner, rather than using parasynonymous strings interchangeably. In the previous example, both the *disease by infectious agent* and *disease* classes contain the string 'disease' in their names, which follows the rule. Using synonymous words like 'disease' and 'disorder' interchangeably would be bad practice.
- Use genus-differentia style names: Class names should reflect the differentiation that distinguishes the class from its parent class. This rule suggests that the name of a class should follow the name of its parent classes, lexically indicating that the subclasses are specializations of the parent class. The class *disease by infectious agent* is a subclass of *disease*, so its name indicates that it is a special type of disease caused by

[1] https://www.w3.org/TR/owl2-overview/

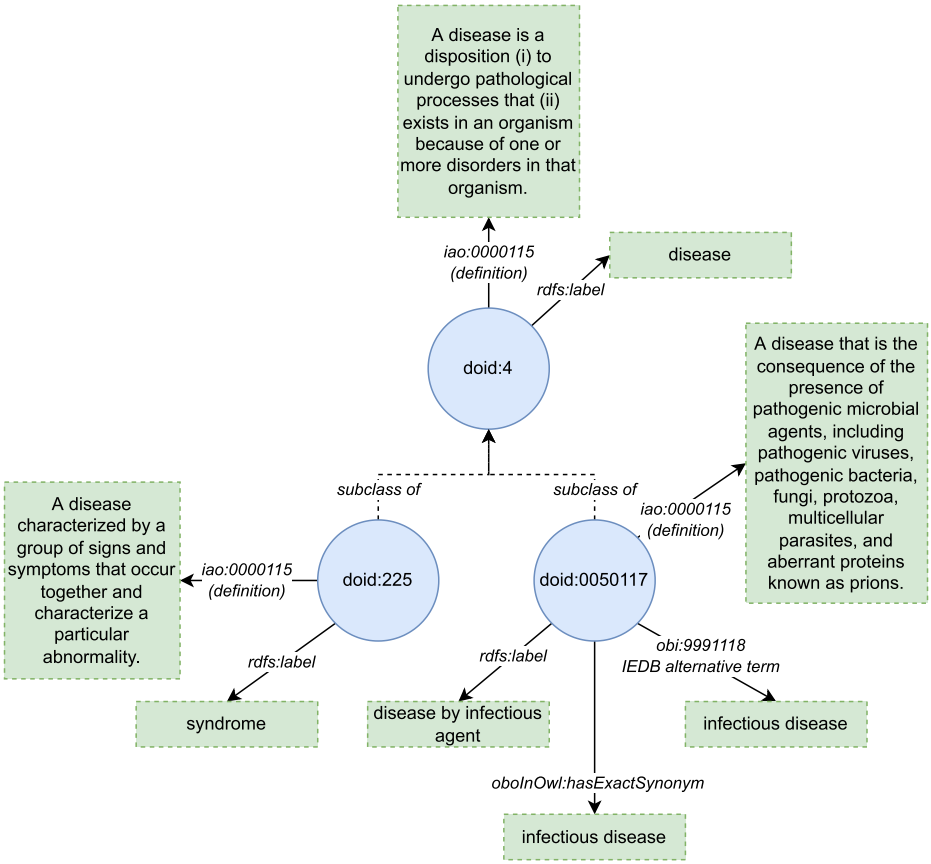

FIGURE 1. Excerpt of the disease ontology.

infectious agents, thus satisfying the rule. The class *syndrome* is also a subclass of *disease*, but it does not contain the word 'disease' in its name, which violates the rule.

In addition to the previous recommendations, in [25] the principle '*lexically suggest, logically define*' suggests that the name of the class should be aligned with the logical axioms associated with the class. In the example of the class name *juvenile osteochondrosis of foot*, this means that this class must be semantically related to the ontology classes *osteochondrosis* and *foot*, if they exist.

B. CLASS DESCRIPTIONS

Human readable descriptions facilitate the understanding and reuse of the ontology. The OBO Foundry Principle 6[2] states that '*The ontology has textual definitions for most of its classes and especially for top-level terms*'. These descriptions are more informative than a simple name, since a name may lead to misunderstandings due to polysemy. Descriptions are incorporated into ontology classes using annotation properties such as *skos:definition*, *rdfs:comment*, or *dc:description*. Ideally, each class should have a description. For multilingual ontologies, the number of descriptions in classes should be equal to the number of languages supported by the ontology.

[2]https://obofoundry.org/principles/fp-006-textual-definitions.html

C. CLASS SYNONYMS

Ontology classes need to be enriched with synonyms whenever possible, which would facilitate the understanding of the meaning of the class by human users from different backgrounds and improve the use of the ontology for NLP processes, as suggested in [39]. Synonyms are alternative names to the canonical name that are widely accepted by the community and used in certain domains. As previously mentioned in Section II-A, the canonical name 'juvenile osteochondrosis of foot' is a good name because it defines the concept with a high level of detail. Nonetheless, in the medical field, this concept is also widely known as 'Kohler disease'. Therefore, 'Kohler disease' can be included as a synonym. Synonyms are included in OWL ontologies by using annotation properties such as *skos:altLabel* or *oboInOwl:hasExactSynonym*.

III. METHODS

In this section, we define the HURON framework, which consists of a set of quantitative metrics and the software tool that implements them. The framework also defines associations between the metrics and best practices.

A. THE HURON METRICS FOR HUMAN READABLE CONTENT

We propose a set of quantitative metrics that provide useful information for assessing the degree of implementation of the previous best practices. We will use the annotation properties resulting from our previous analysis of the most commonly used annotation properties in the BioPortal repository [22], plus some additional ones.

1) NAMES PER CLASS

This metric accounts for the number of names associated with classes, and uses the list of annotation properties from the ontology community for names (see Table 1). Then, the metric *names per class* is calculated as the number of names associated with ontology classes divided by the total number of classes in the ontology.

The value range of this metric is the set of real positive numbers. Values less than one indicate that there are classes without names in the ontology. Conversely, a value greater than 1 indicates that there are classes with multiple names, possibly caused by the inclusion of multilingual names or by some design decision, such as the one commented on previously for SNOMED CT in Section II-A.

To illustrate how this metric is calculated, we will use the Disease Ontology excerpt shown in Figure 1. The figure contains three classes with several annotation properties describing them, the name being provided by the *rdfs:label* property. Since there are 3 classes and 3 *rdfs:label* properties associated with classes, the value of the metric *names per class* is $\frac{3}{3} = 1$.

2) DESCRIPTIONS PER CLASS

This metric takes into account the number of descriptions associated with classes that can also be provided using different annotation properties. Specifically, Table 2 provides the identified list of annotation properties used by the community to encode descriptions. The metric *descriptions per class* is calculated as the total number of descriptions associated with ontology classes divided by the total number of classes in the ontology.

As an example for the calculation of the metric, the disease ontology excerpt (see Figure 1) contains three classes, and each of them has a description provided by the property *iao:0000115* ('definition', from the Information Artifact Ontology). So the metric would give a result of 3 descriptions over 3 classes, i.e. 1 *description per class*, which is in line with the recommendation.

3) SYNONYMS PER CLASS

Similarly, the metric *synonyms per class* takes into account the number of synonyms associated with classes, which can also be provided using different annotation properties. Table 3 contains the identified list of annotation properties used by the community to add synonyms. The metric *synonyms per class* is calculated as the number of synonyms associated with classes divided by the total number of classes in the ontology.

In the example from Figure 1, only one of the three classes is annotated with synonyms. Specifically, the class *doid:0050117* is annotated with two synonyms using the properties *obi:9991118* (alternative term used by the IEDB) and *oboInOwl:hasExactSynonym*. Thus, the ontology contains 2 synonyms and 3 classes, giving $\frac{2}{3} = 0.67$ *synonyms per class*. Note that the metric does not check the textual content of the synonym, which in this case is the same.

4) METRICS BASED ON LEXICAL REGULARITIES

In this section, we describe metrics that exploit the lexical structure and content of the names of the ontology classes. These metrics use the concept of lexical regularity (LR) [27], [40], which is a consecutive list of words that appear recurrently in a set of class names.

When a lexical regularity exhibited by a class is equal to its full canonical name, the class is called *lexical regularity class* (LR class). For example, the class *process* from the *Basic Formal Ontology* (BFO) (see Figure 2) is an *LR class* because the text 'process' is a lexical regularity that is also exhibited by the classes *process profile* and *process boundary*.

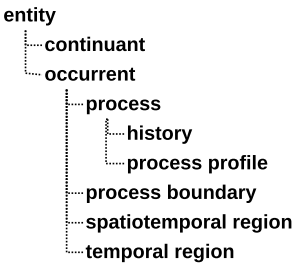

FIGURE 2. Extract from the class hierarchy of the BFO ontology, version dated 2019-08-26.

In this work, we use the following two metrics based on lexical regularities:

- *Systematic naming*: This is related to the ontology design principle that classes in the same taxonomy should share part of their name, since subclasses are specializations of the parent class. In other words, class names should follow a *genus-differentia* style. This metric is calculated as the ratio of subclasses of an *LR class* that share the lexical regularity of the parent class. This requires calculating how many subclasses of a given *LR class* have the lexical regularity in their name (positive cases) and how many do not (negative cases). The value of the metric is calculated by dividing the positive cases by the total number of cases and is done for each *LR class*. An example of the metric focusing on the *process* taxonomy from the BFO ontology (see Figure 2) results in a value of 0.5. In this case, the *LR class process* has two subclasses: *history*, which is a negative case because it does not have the lexical regularity of the parent, and *process profile*, which is a positive case.

TABLE 1. Annotation properties used to encode names.

Annotation property URI	Description
http://www.w3.org/2004/02/skos/core#prefLabel	Preferred label from the SKOS vocabulary
http://www.w3.org/2000/01/rdf-schema#label	Label from the RDF vocabulary
http://schema.org/name	Name from the schema.org vocabulary
http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#P108	Preferred name from the NCI Thesaurus
http://purl.obolibrary.org/obo/IAO_0000589	OBO Foundry unique label from the Information Artifact Ontology
http://xmlns.com/foaf/0.1/name	Name from the FOAF vocabulary

TABLE 2. Annotation properties used to encode descriptions.

Annotation property URI	Description
http://purl.obolibrary.org/obo/IAO_0000115	Definition from the Information Artifact Ontology
http://www.w3.org/2004/02/skos/core#definition	Definition from the SKOS vocabulary
http://www.w3.org/2000/01/rdf-schema#comment	Comment from the RDF vocabulary
http://purl.org/dc/elements/1.1/description	Description from the DC terms vocabulary
http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#P97	Definition from the NCI Thesaurus

TABLE 3. Annotation properties used to encode synonyms.

Annotation property URI	Description
http://www.w3.org/2004/02/skos/core#altLabel	Alternative label from the SKOS vocabulary
http://www.geneontology.org/formats/oboInOwl#hasExactSynonym	Exact synonym from the oboInOwl vocabulary
http://www.geneontology.org/formats/oboInOwl#hasRelatedSynonym	Related synonym from the oboInOwl vocabulary
http://www.geneontology.org/formats/oboInOwl#hasBroadSynonym	Broad synonym from the oboInOwl vocabulary
http://www.geneontology.org/formats/oboInOwl#hasNarrowSynonym	Narrow synonym from the oboInOwl vocabulary
http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#P90	Fully qualified synonym from the NCI Thesaurus
http://purl.obolibrary.org/obo/IAO_0000118	Alternative term from the Information Artifact Ontology
http://purl.obolibrary.org/obo/OBI_9991119	Alternative term used by the Functional Genomics Data (FGED) Society.
http://purl.obolibrary.org/obo/OBI_9991118	Alternative term used by the IEDB.
http://purl.obolibrary.org/obo/OBI_0001847	Alternative term used by the ISA tools project (http://isa-tools.org).
http://purl.obolibrary.org/obo/OBI_0001886	Alternative term used by the National Institute of Allergy and Infectious Diseases (NIAID), the Genomic Sequencing Centers for Infectious Diseases (GSCID), and the Bioinformatics Resource Centers (BRC).

- *Lexically suggest, logically define (LSLD)*: It is related to the design principle of the same name [25], which we interpret as follows: what is expressed in natural language for humans should be expressed as logical axioms for the machine. It is calculated as the ratio in which an *LR class* is semantically related to other classes that exhibit its lexical regularity. Here, two classes are semantically related if there exists a path of arbitrary length between them through the axioms in the ontology (i.e., subclass of, equivalence, or domain/range property). Disjoint axioms are not considered for this ontology traversal. For computational reasons, we limited the length of the path between classes to 5; thus, classes that are semantically related by longer paths are considered to be not semantically related. In this case, positive cases are classes that exhibit a lexical regularity and are semantically related to the corresponding *LR class*. Negative cases are classes that exhibit lexical regularity and are not semantically related to the corresponding *LR class*. The value of the metric is calculated by dividing the positive cases by the total number of cases. This metric is calculated for each *LR class*. In our BFO example (Figure 2), the *LSLD* metric would check if the classes with the lexical regularity 'process' are semantically related to the class *process*. In this case, *process profile* is a positive case because it is a subclass of *process*, but *process boundary* is a negative case because it is not semantically related to *process*. Consequently, the value of the metric for the *LR class process* is also 0.5.

Finally, the value of the *systematic naming* and *LSLD* metrics for an ontology is calculated as the sum of all positive cases divided by the sum of all positive and negative cases obtained by each *LR class* in the ontology. The values of both metrics are in the range [0, 1], where the highest values represent the best values for the metric. In this work, the implementation of both metrics assumes the annotation property *rdfs:label* as the canonical name, since it is the most frequently used by the analyzed ontologies.

B. ASSOCIATION BETWEEN BEST PRACTICES AND METRICS

Table 4 summarizes the associations between best practices and metrics, showing which metrics are used to evaluate each best practice. We also provide the target values for the metrics, which represent the threshold for estimating the ontology's compliance with the best practice associated with each metric:

- The target value for the metrics *names per class* and *descriptions per class* is 1 for monolingual ontologies. For multilingual ontologies, the optimal value for these metrics is equal to the number of languages supported by the ontology.
- The target number of *synonyms per class* depends on the concept represented by that class, since there are classes with a higher number of synonyms than others. Nevertheless, we have set a target value of 1 synonym per class based on the study of the OBO Foundry repository, which is shown in Section IV-E.
- The target value for the *systematic naming* and the *LSLD* metrics is their maximum possible value, 1.

C. CLUSTERING-BASED ANALYSIS

We performed a metric-by-metric clustering analysis to obtain information about how each metric partitions the corpus of ontologies. We used Evaluome [41], [42] to perform a k-means based clustering of ontologies, using each metric separately as a unique feature. Evaluome performs clusterings for a range of values of k and returns the value of two statistical properties of the clusterings, namely stability and goodness. The stability measures the effect of small variations on the data, and its values are in the range [0, 1], whose interpretation is described in Table 5. Goodness measures how closely the instances in a category are related and how well a category is separated from the rest of the categories, and returns values in the range $[-1, 1]$ (see Table 6 for the interpretation of the values).

In this work, we have used the interval [2,11] as the range for the values of k. Given a value of k, the method requires at least k different values for the metric used for the clustering to be able to provide a result. This method suggests the optimal number of ontology clusters (k) for each metric by analyzing the values of stability and goodness. The clustering using the optimal value of clustering k will be described, although we have also manually inspected the results of other values of k.

D. THE HURON SOFTWARE TOOL

We have developed a software tool that provides a command line interface to compute the metrics of any set of ontologies (see https://github.com/fanavarro/huron). This tool was implemented in Java and uses the OWL API [43] for ontology parsing, the ELK reasoner [44] for axiom discovery, and the OntoEnrich framework for lexical regularity extraction [26]. For ontologies that could not be processed by ELK, the structural reasoner provided by the OWL API was used.

A web version of the tool is available at https://semantics.inf.um.es/huron. In this web version, the user can enter a set of ontologies by specifying their IRI, which should redirect to an OWL file, and select the set of metrics to be calculated. Optionally, the user can choose to perform the analysis of the metrics based on Evaluome [41], [42] described in this work. This analysis generates several plots showing (1) the global distribution of each metric, shown as violin plots; (2) the correlation between the metrics, shown as a heatmap; and (3) a clustering analysis, whose output is a plot per metric showing the distribution and the range of values of each group of ontologies found, together with the stability and the quality of the resulting clustering. The results of the request are sent by email in a zip file containing a CSV file with the metric values for each ontology and the plots resulting from the analysis, if requested by the user.

IV. RESULTS

In this section, we describe the results of applying our metrics to a corpus of biomedical ontologies extracted from the OBO Foundry repository [7]. The full results are available in our GitHub repository.[3] The OBO Foundry aims to provide a set of orthogonal biomedical ontologies and is therefore a good candidate for providing a wide variety of ontologies in the biomedical domain. The content of this repository is expected to follow the OBO principles. First, we describe how the content of the OBO Foundry repository was processed. We then present the results in terms of the distribution of metric values and the metrics for names, descriptions and synonyms.

A. DATA PROCESSING

At the time of our work (March 2022), the repository contained 182 active ontologies considered for this study, 10 of which were classified as OBO Foundry member ontologies, while the remaining 172 were candidates. They are available at https://doi.org/10.5281/zenodo.4701571. In addition, we provide a CSV file containing the acronyms and full names of the ontologies considered in this work in our GitHub repository[3]. The 182 OBO Foundry ontologies were downloaded using the scripts available in our GitHub repository[3]. The metrics were then calculated using HURON release v0.0.2, presented in Section III-D.

A timeout of 12 hours was applied to the OBO Foundry candidate ontologies. This restriction was not applied to the OBO Foundry member ontologies due to their relevance and smaller number of ontologies. The final corpus contained 142 ontologies for the following reasons

- A total of 38 ontologies could not be parsed due to incompatibilities with the OWLAPI version, including the member ontologies *Human Disease Ontology* (DOID) and the *Plant Ontology* (PO).
- The 12 hours timeout expired for the NCBI Taxon ontology.
- The *Mathematical Modeling Ontology* (MAMO) did not use the *rdfs:label* property for the ontology class names. This was required to calculate the *systematic naming* and the *LSLD* metrics.

B. DISTRIBUTION OF VALUES OF THE METRICS

The values of the metrics considered in this work (i.e., *names, synonyms, and descriptions per class*, and *LSLD* and *systematic naming*) did not fit a normal distribution in this corpus. Figure 3 contains violin and box plots summarizing the

[3]https://github.com/fanavarro/lexical-analysis-obo-foundry

TABLE 4. Association between the best practices and the metrics that measure them.

Characteristic	Best practice	Metric	Target value
Class name	Number of names in classes	Names per class	1
	Naming style	Systematic naming	1
	Lexically suggest, logically define	LSLD	1
Class description	Number of descriptions in classes	Descriptions per class	1
Class synonym	Number of synonyms in classes	Synonyms per class	1

TABLE 5. Interpretation of cluster stability values.

Stability	Interpretation
[0, 0.60)	Unstable
[0.60, 0.75]	Doubtful
(0.75, 0.85]	Stable
(0.85, 1]	Highly stable

TABLE 6. Interpretation of cluster goodness values.

Goodness	Interpretation
[-1, 0.25)	There is no substantial clustering structure
[0.25, 0.50]	The clustering structure is weak and could be artificial
(0.50, 0.70]	There is a reasonable clustering structure
(0.70, 1]	Strong clustering structure

distribution of each metric and its values obtained from the ontology corpus, including the p-values resulting from the Shapiro-Wilk normality test [45]. The Shapiro-Wilk test tests the null hypothesis that a sample comes from a normally distributed population.

C. NAMES

1) NUMBER OF NAMES

The metric *names per class* has a median of 1, a mean of 1.015, and a standard deviation of 0.077. This fact is consistent with the best practice described in Section II-A, which states that each ontology class must have a unique canonical name. Thus, the evaluated ontologies were generally compliant with this recommendation.

There were some outlier ontologies with values slightly higher than 1, such as the *Prescription of Drugs Ontology* (PDRO), the *Biological Imaging Methods Ontology* (FBBI), and the *Compositional Dietary Nutrition Ontology* (CDNO), which had 1.4 *names per class*. Upon closer inspection, we found that PDRO and CDNO use the annotation property *rdfs:label* with different language tags to include multilingual names in some classes. Also, FBBI included an OBO Foundry unique label in several ontology classes through the *IAO:0000589* property from the *Information Artifact Ontology* (IAO), in addition to the *rdfs:label*.

Conversely, there were 15 ontologies, such as the *Units of Measurement Ontology* (UO) and the *Food Interactions with Drugs Evidence Ontology* (FIDEO), that did not reach 1 *names per class*, although this value was close to 1 in all cases. This was mainly caused by reusing classes from other ontologies without including the corresponding annotations. For example, FIDEO reuses several *Basic Formal Ontology* (BFO) classes that do not contain annotations. This could be mitigated by including these annotations manually, or by importing the external ontology using an 'import' statement.

Using Evaluome, the optimal number of clusters obtained for the *names per class* metric was 2, reaching a stability of 0.922 and a goodness of 0.972, indicating highly stable clusters with a strong structure. Cluster 1 was formed by only 4 ontologies, with values ranging from 1.436 to 1.472 *names per class*, while cluster 2 was formed by the remaining 129 ontologies, with values ranging from 0.921 to 1.152 (see Figure 4A).

Here, cluster 1 represents the ontologies with more than 1 *names per class* in general, due to multilingual names or the use of more than one property to contain names, as commented before. On the other hand, cluster 2 includes the ontologies whose values for the metric name per class were around 1. The optimal clustering did not suggest a cluster to represent ontologies with values significantly lower than 1 for this metric, which would be contrary to the best practice commented in Section II-A. In summary, the results support that the ontologies in this corpus follow this best practice.

Finally, we explored different k partitions for this metric in addition to the optimal one, and found $k = 4$ to be the most informative. These results are shown in Figure 5. Here we found cluster 1 to be the most representative, with 123 ontologies having almost 1 name per class. Cluster 2 was formed by 2 ontologies (not shown in the figure due to their low density) with a *names per class* slightly higher than 1; while this value was slightly lower than 1 for cluster 4, since 4 ontologies belong to this cluster. Finally, cluster 3 was formed by 4 ontologies with more than one name per class.

2) SYSTEMATIC NAMING

The value obtained for the *systematic naming* metric is often low, with a mean of 0.2, a median of 0.14, and a standard deviation of 0.173. In general, this means that ontology classes do not follow a *systematic naming* along the hierarchies, so the *genus-differentia* naming style is not used.

However, some ontologies received values close to one for the metric, which is the highest possible value. This is the case of the *Teleost Taxonomy Ontology* (TTO), which received a value of 0.998. This ontology represents a taxonomy of organisms that uses scientific notation for naming them, thus presenting a *genus-differentia* style. For example, the fact that 'Eudontomyzon danfordi', 'Eudontomyzon hellenicus', and

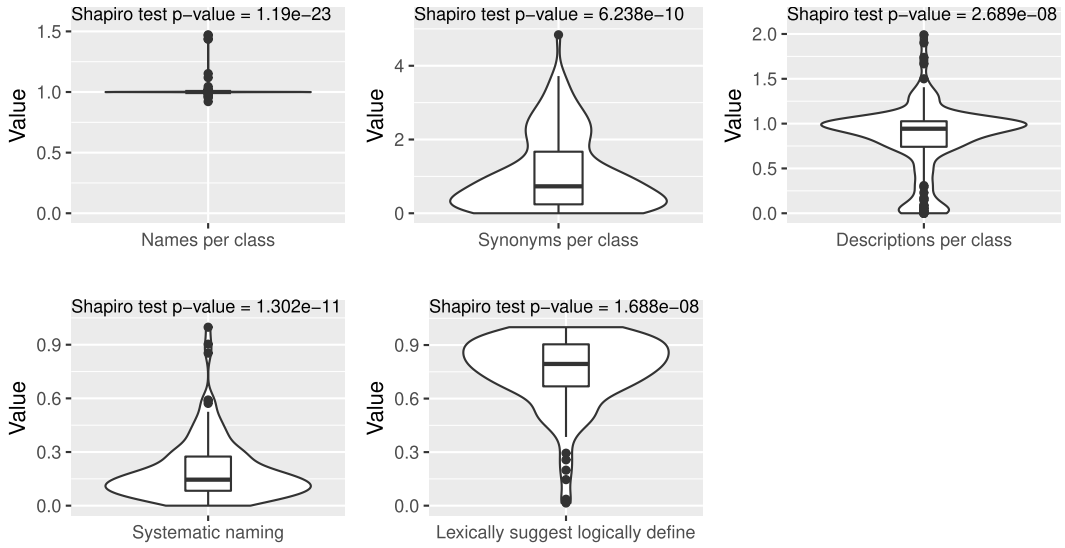

FIGURE 3. Violin plots showing the distribution of each metric along the ontology corpus. The p-values of the Shapiro normality test are shown.

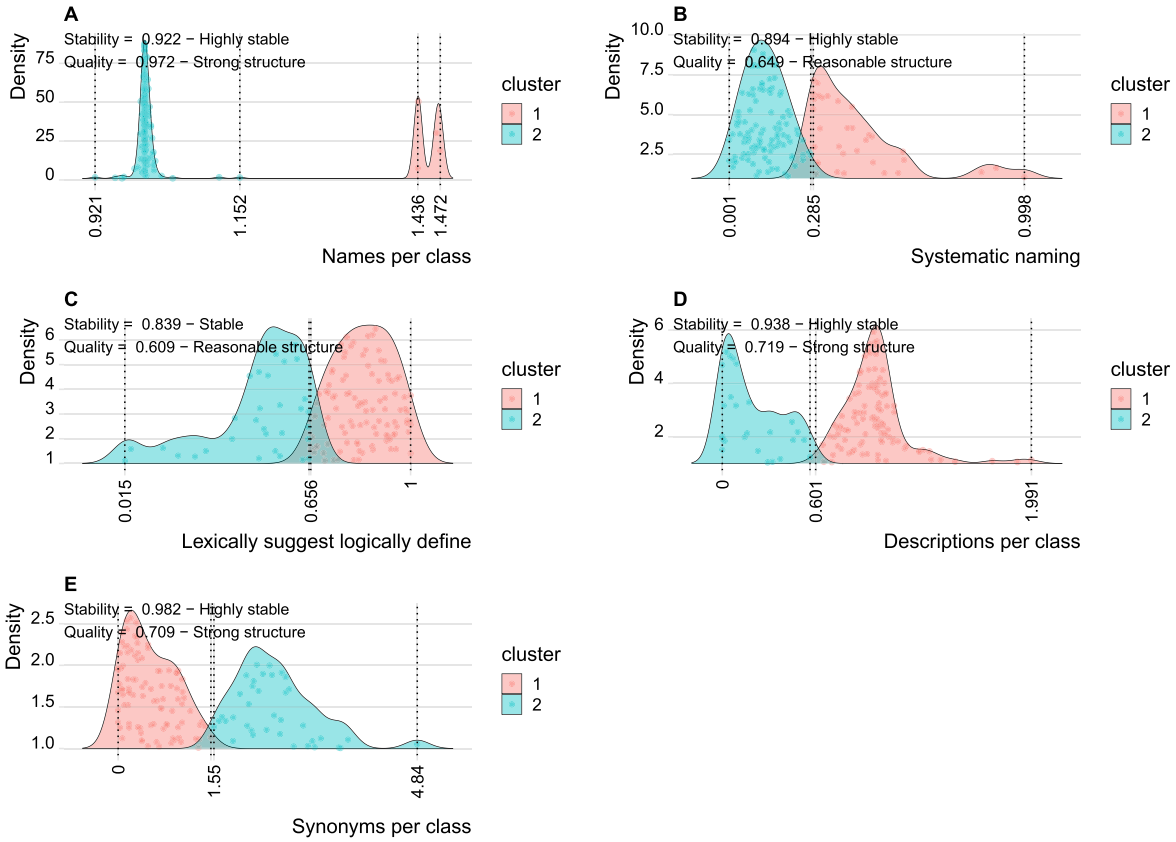

FIGURE 4. Distribution of the metrics, separated by each cluster of ontologies found by Evaluome.

'Eudontomyzon lanceolata' are specializations of 'Eudontomyzon' is represented semantically by defining them as subclasses of *Eudontomyzon*, and also lexically by including the word 'Eudontomyzon' in their names.

Evaluome found that the optimal number of clusters for the *systematic naming* metric was 2. This cluster configuration achieved a stability of 0.8937970 and a goodness of 0.6487532, indicating a highly stable clustering with a

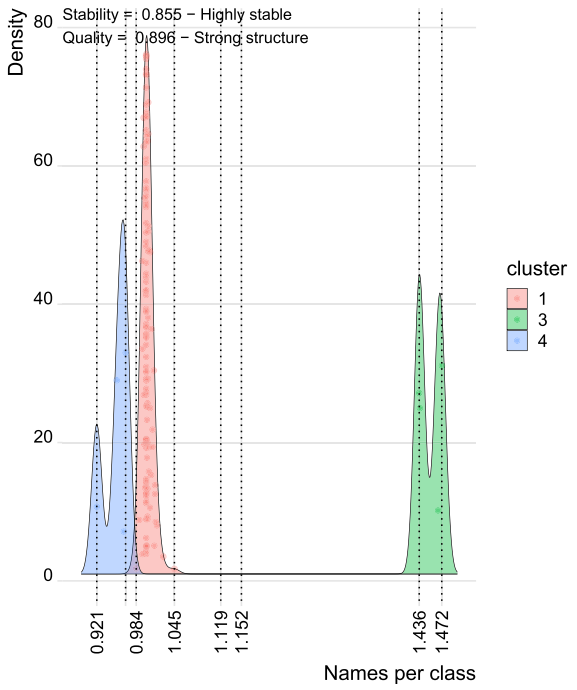

FIGURE 5. Distribution of the *names per class* metric, separated by each cluster of ontologies for that metric by using $k = 4$.

reasonable structure. Cluster 1 contained 33 ontologies with *systematic naming* values between 0.285 and 0.998, while cluster 2 contained 100 ontologies with values between 0.001 and 0.276 for the metric (see Figure 4B). Here, cluster 1 represents the ontologies with a higher value for the *systematic naming* metric, which includes only the 24.8% of the considered ontologies and has even a wider range of values than cluster 2. For its part, cluster 2 grouped the ontologies with lower values, which was the majority of them (75.2%). This fact supports that, in general, the ontologies in the corpus do not follow a stable *systematic naming*.

3) LEXICALLY SUGGEST LOGICALLY DEFINE

The value obtained for the *LSLD* metric is generally high, with a median of 0.794 and a mean of 0.756 with a standard deviation of 0.198. This indicates that *LR classes* are usually semantically related by ontology axioms with classes showing their lexical regularity in the name, as recommended by the 'lexically suggest, logically define' principle.

The *Confidence Information Ontology* (CIO), the *Glycan Naming and Subsumption Ontology* (GNO), the *Human Developmental Stages Ontology* (HSAPDV), and the *Mouse Developmental Stages Ontology* (MMUSDV) obtained the maximum *LSLD* metric value. However, these ontologies have a small number of *LR classes* (see Table 7), so the metric was calculated using a small fraction of the ontology. This may indicate that the human-readable content of these ontologies does not provide much information about the semantics, although the information provided is consistent with the semantic definitions. For example, GNO contained 112, 377 classes, of which only *GNO:00000001* (glycan) was an *LR class*. The lexical regularity 'glycan' is exhibited by 13, 729 classes in the ontology, and all of them were semantically related to the class 'glycan', thus having a *LSLD* metric of 1. Note that this value was calculated from a single LR class in the ontology, so it is not very informative.

TABLE 7. Top-5 ontologies according to the *LSLD* metric, along with the number of LR classes detected in each one.

Ontology	LSLD value	Number of LR classes
CIO	1	12
GNO	1	1
HSAPDV	1	2
MMUSDV	1	1
OMO	1	5

A representative example of the ontology corpus for the *LSLD* metric was the *Informed Consent Ontology* (ICO), which received the median of the repository (0.794). A number of 184 *LR classes* were detected in this ontology, and most of the classes that included the lexical regularity in their names were semantically related to the corresponding LR class. For example, 77 classes included the lexical regularity 'specimen' in their names, and 76 of them were semantically related to the LR class 'specimen'.

Evaluome found 2 ontology clusters as the best clustering configuration, achieving a stability of 0.8387104 and a quality of 0.6092190, indicating stable clusters with a reasonable cluster structure. Cluster 1 grouped 102 ontologies with the highest values for the metric, ranging from 0.656 to 1. On the other hand, cluster 2 contained 31 ontologies with lower values for this metric, ranging from 0.015 to 0.650 (see Figure 4C). In this case, most of the ontologies (77%) belonged to the cluster with the highest values for the metric. This indicates that the lexical information contained in class names is generally consistent with the semantic definitions declared in the corresponding ontology, thus following the principle of 'lexically suggest, logically define'.

4) JOINT ANALYSIS OF SYSTEMATIC NAMING AND LSLD

Although both *systematic naming* and *LSLD* were evaluated independently, we found interesting to perform a cluster analysis by using both of them as features for ontology classification. Here we used the silhouette method [46] to identify the optimal number of clusters resulting from the application of the k-means algorithm and the Euclidean distance, since Evaluome only works with single variables. The optimal number of clusters identified by the silhouette method was 3 (see Figure 6). The resulting clusters are shown in Figure 7, where cluster 1 was the largest, formed by 82 ontologies with high values for the *LSLD* metric (mean of 0.83) and low values for the *systematic naming* metric (mean of 0.15). Cluster 2 was formed by 31 ontologies with low values for both metrics (*LSLD* mean = 0.47; *systematic*

naming mean = 0.11). Cluster 3 was formed by 20 ontologies with high values for both metrics (*LSLD* mean = 0.88; *systematic naming* mean = 0.52). Thus, cluster 3 represents the ontologies that follow the class naming recommendations to a higher degree, i.e. the *LR classes* follow a systematic naming, and most of the classes showing lexical regularities are linked to the corresponding *LR classes*. Cluster 2, on the other hand, represents the ontologies that should improve this aspect. Here, the *LR classes* do not follow a systematic naming and the classes exhibiting lexical regularities are related to the corresponding LR class in a lower degree. On a middle ground, cluster 1, which is the most representative of the OBO Foundry repository due to its size, was formed by ontologies that follow the principle of 'lexically suggest, logically define', but do not follow a systematic naming. More precisely, the *LR classes* of these ontologies follow a systematic naming to a low degree, but they are linked to those classes that show their lexical regularities to a high degree.

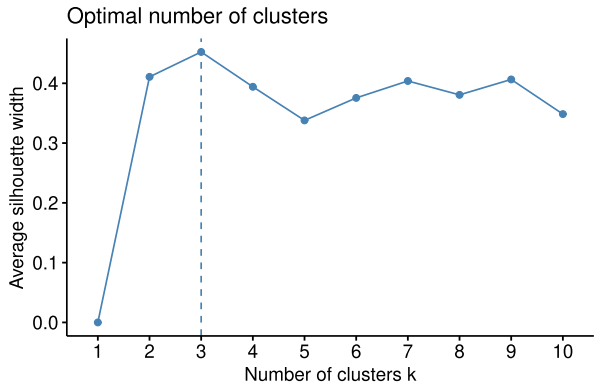

FIGURE 6. Silhouette plot to identify the optimal number of ontology clusters by using the *LSLD* and the *systematic naming* metric together as features.

In addition, we found a positive correlation between the values of the *systematic naming* and the *LSLD* metrics. In other words, ontologies with high values for the *systematic naming* metric are likely to have high values for the *LSLD* metric as well. Specifically, the Spearman correlation test yielded $\rho = 0.37; p < 0.05$. This makes sense because any subclass of an LR class that exhibits the corresponding lexical regularity is counted as a positive case for both metrics.

D. DESCRIPTIONS

As noted in Section II-B, all classes would contain at least one description in the best case scenario. The distribution of values for this metric has a median of 0.943, a mean of 0.845, and a standard deviation of 0.39. This is close to the minimum acceptable value for the metric. In particular, 52 ontologies met this rule, and 90 had less than 1 description per class. In addition, there were 48 ontologies with more than one description per class on average, such as the Gene Ontology or the Protein Ontology.

The *Systems Biology Ontology* (SBO) obtained a value of 0.999 *descriptions per class*, which is very close to the target value. Like SBO, many ontologies obtained a value for this metric very close to 1, such as the *Zebrafish Phenotype Ontology* (ZP) (0.998 *descriptions per class*, 1, 114 of 55, 186 classes without description), the *Ontology of Arthropod Circulatory Systems* (OARCS) (0. 997 *descriptions per class*, 1 of 308 classes without description), or the *C. elegans developmental ontology* (WBLS) (0.997 *descriptions per class*, 12 of 794 classes without description). These ontologies lack descriptions for a small fraction of their classes, so they can be improved with relatively little effort. Conversely, some ontologies obtained a lower number of *descriptions per class*, such as MCO (0.528 *descriptions per class*, 1, 607 out of 3, 383 classes without description) or FIDEO (0.565 *descriptions per class*, 180 out of 402 classes without description). In addition, 3 ontologies, namely the *Mouse adult gross anatomy ontology* (MA), the *Teleost taxonomy ontology* (TTO), and the *Zebrafish developmental stages ontology* (ZFS), did not contain any description in classes, demonstrating that this aspect is not considered in their development.

Finally, Evaluome proposed 2 clusters according to the description per class metric, reaching a stability of 0.9375347 and a goodness of 0.7186178, indicating a highly stable clustering with a strong structure. On the one hand, cluster 1 contained 108 ontologies whose values for the metric ranged from 0.601 to 1.991, representing the ontologies with a higher number of *descriptions per class*. On the other hand, cluster 2 had only 25 ontologies with a lower value for the metric, specifically between 0 and 0.565 *descriptions per class* (see Figure 4D).

E. SYNONYMS

The value obtained for the *synonyms per class* metric ranged from 0 to 4.84, with a median of 0.73 and a mean of 1, and a standard deviation of 1. Table 8 shows the top 5 ontologies with the highest number of *synonyms per class*, including the *Microbial Conditions Ontology* (MCO), the *National Cancer Institute Thesaurus* (NCIT), the *Protein Modification Ontology* (MOD), the *Fission Yeast Phenotype Ontology* (FYPO), and the *Compositional Dietary Nutrition Ontology* (CDNO). On the other hand, no synonyms were found in the classes of the following ontologies (i.e. a value of 0 *synonyms per class*): *The Core Ontology for Biology and Biomedicine* (COB), *The Disease Drivers Ontology* (DISDRIV), *The Nomenclatural Ontology for Biological Names* (NOMEN), *The Systems Biology Ontology* (SBO), *The Units of Measurement Ontology* (UO), and *The Zebrafish Developmental Stages Ontology* (ZFS). As shown in Figure 8C, most synonyms were included by using *oboInOwl:hasExactSynonym* and *oboInOwl:hasRelatedSynonym*; however, seven other annotation properties with marginal use for including synonyms were identified in the repository.

Evaluome grouped the ontologies into 2 clusters according to this metric, achieving a stability of 0.9819980 and a

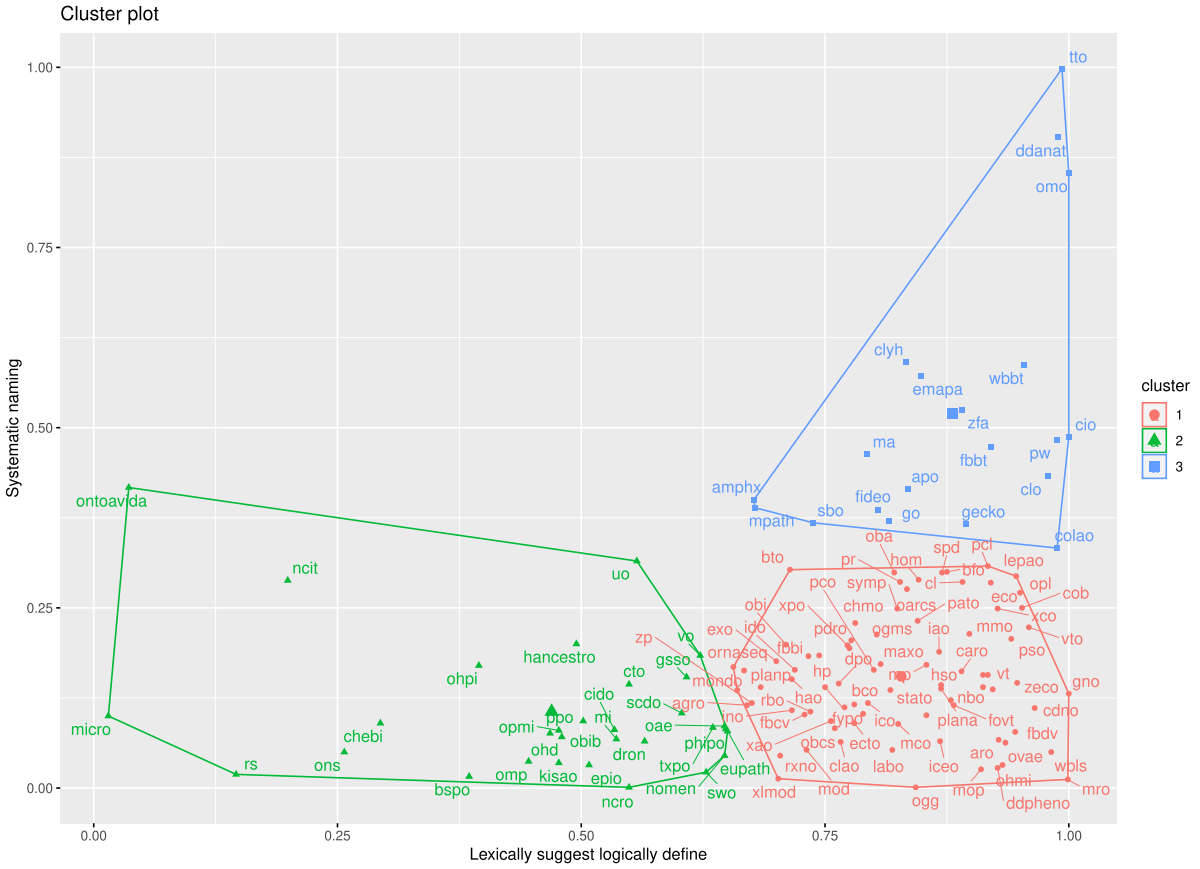

FIGURE 7. Clusters of ontologies found by using the *LSLD* and the *systematic naming* metrics together as features and $k = 3$.

TABLE 8. Top-5 ontologies according to the *synonyms per class* metric.

Ontology	Synonyms per class
MCO	4.840
NCIT	3.720
MOD	3.598
FYPO	3.571
CDNO	3.278

goodness of 0.7089322, indicating a highly stable cluster with a strong structure. Cluster 1 contained 96 ontologies with a low number of *synonyms per class*, ranging from 0 to 1.5, while cluster 2 contained 37 ontologies with a higher number of *synonyms per class*, ranging from 1.55 to 4.84 (see Figure 4E).

V. DISCUSSION

The HURON framework has made it possible to obtain knowledge about the adherence to best practices related to the human readability of ontologies (RQ1). In this work, we did not use a predetermined threshold for the values of the metrics to determine compliance with a best practice. However, we did use a target value for their interpretation. In addition, we used clustering to classify the ontologies according to their calculated metric values. This served to detect different levels of readability of the ontologies in the OBO Foundry repository. Next, we provide a general discussion of the work done.

A. THE METRICS

In this work, we have proposed a target value per metric to serve as a reference for the analysis of the ontologies. However, different target values for the same metric could apply to different ontologies. For example, a value higher than 1 *name per class* may be justified by the design decision of including multilingual names or additional annotations for different name types. In this case, the target value for *names per class* metric should depend on the number of languages supported by the ontology. Accordingly, the fact that three ontologies have a value of 1.4 *names per class* could indicate that there are classes without labels for all languages used in the ontology.

Regarding the *systematic naming* metric, we observed the existence of *LR classes* with short names representing very general concepts, usually located at high levels of the ontology, such as 'role', 'object' or 'entity'. These classes can be considered as 'domain-independent' classes and are usually present in top-level ontologies such as BFO, which are imported to be used as a semantic framework

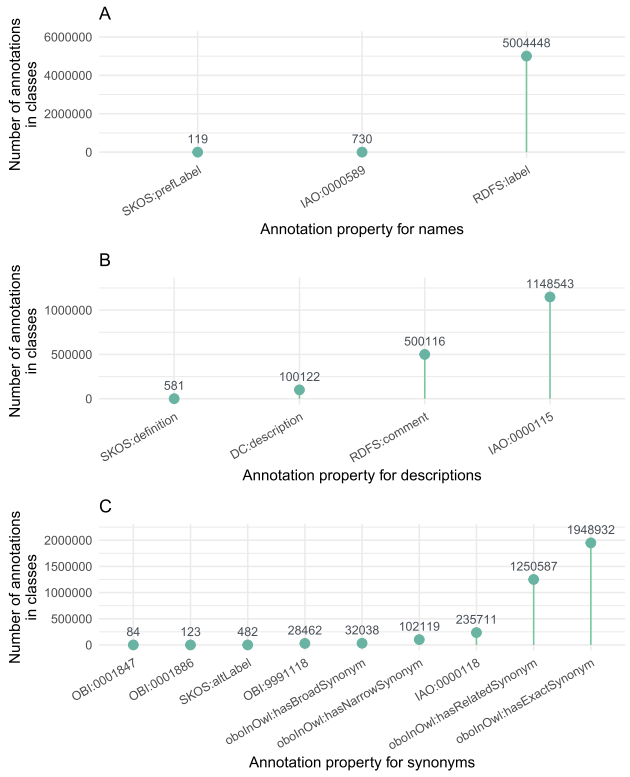

FIGURE 8. Usage, in terms of the number of annotations in classes, of each annotation property identified in the OBO Foundry repository that encodes (A) names, (B) descriptions, and (C) synonyms.

for the development of new ontologies. Maintaining these words as part of the names along the subclass hierarchy is difficult due to the large number of subclasses by transitivity. For example, in the *ontology of core ecological entities* (ECOCORE), the LR class 'entity' has 5, 439 subclasses by transitivity, but only 43 of them include the word 'entity' as part of the name. In general, the classes exhibiting lexical regularities with a broad meaning in their names are not subclasses of the corresponding *LR class*, resulting in low values for the metric. Conversely, *LR classes* that refer to more concrete domain-dependent concepts are more likely to have higher *systematic naming* scores. This fact is mainly due to a lower number of transitive subclasses to evaluate for the corresponding *LR class*, which makes it easier to maintain the lexical regularity in these subclasses. As an example taken from ECOCORE, the *LR class* 'aorta' is a domain-dependent class that has only three subclasses that exhibit the corresponding lexical regularity: 'dorsal aorta', 'left dorsal aorta' and 'right dorsal aorta', thus having a *systematic naming* value of 1. To quantify this fact, we performed a Spearman correlation test between the number of transitive subclasses of each LR class found in the entire repository, and the value obtained for that LR class for the *systematic naming* metric. The Spearman correlation showed, as depicted in Figure 9, a negative correlation between the number of transitive subclasses and the value for the *systematic naming* metric ($\rho = -0.229$; $p = 2.19 \cdot 10^{-239}$). In addition, we provide a supplementary CSV file available on our GitHub[3], where this correlation analysis was performed for each ontology in the repository individually, showing that 92 out of 126 ontologies resulted in a significant negative correlation, whereas only 7 showed a significant positive correlation ($p < 0.05$).

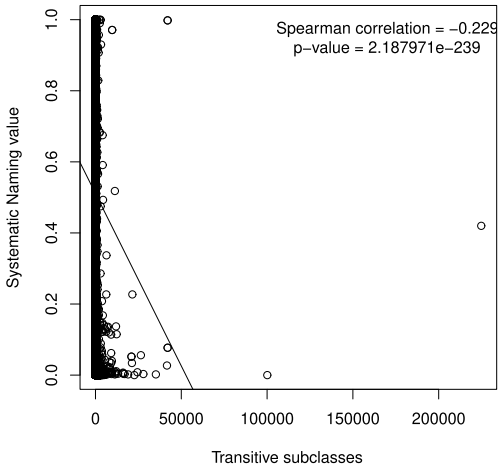

FIGURE 9. Correlation between the *systematic naming* score and the number of transitive subclasses found for all the LR classes in the repository.

The *LSLD* metric can help to identify possible ontology errors by looking at *LR classes* that negatively affect the metric. For example, looking at the LR class 'informed consent form' from ICO, 36 classes contained this lexical regularity in their names (e.g., 'signing an informed consent form'), but 18 of them were not semantically related to the LR class 'informed consent form'. In this case, the *LSLD* metric helps to detect classes whose semantics may not be fully formalized.

The value of the metric *synonyms per class* seems to be related to the nature of the domain. Ontologies whose classes can be named using different naming systems are more likely to have a higher number of synonyms. In addition, domain ontologies that contain very specific classes are also likely to have more synonyms. This is the case in the chemical domain, where compounds defined as domain classes can be named using different nomenclatures (e.g., traditional, systematic, inventory). For example, the highest number of *synonyms per class* (4, 839) was achieved by the Microbial Conditions Ontology (MCO) (see Table 8), which has a strong chemical background. This ontology reuses a large number of terms from the CHEBI ontology by adding them directly to the ontology, including all annotations of the class from CHEBI. For example, the class *chebi:29222* (hypochlorite) includes the synonyms 'oxidochlorate(1-)', '[ClO](-)' or 'hypochlorite'. However, the terms in CHEBI contain annotations such as *chebi:charge*, *chebi:mass*, or *chebi:monoisotopicmass*, which were erroneously included as synonyms in MCO, possibly due to an automatic

import process, resulting in a higher number of synonyms compared to CHEBI. Ontologies with a lower number of *synonyms per class* are usually ontologies with very general concepts that are less likely to have synonyms. This is the case of the previously mentioned ontologies with 0 *synonyms per class*. They include concepts like 'information representation' or 'geographic location' from COB, or 'chemical driver' or 'nutrient deficiency' from DISDRIV. However, ontologies such as ZFS or SBO contained low-level classes such as 'gastrula' or 'enzymatic rate law for irreversible non-modulated non-interacting bireacting enzymes', which were not enriched with synonyms, possibly due to a lack of synonyms in the corresponding domains. In addition, we found that UO included synonyms by using both *oboInOwl:hasExactSynonym* and *oboInOwl:hasRelatedSynonym*; however, these properties were included by using an undefined prefix in the ontology, thus preventing the proper recognition of synonyms. In addition, NOMEN also included synonyms by using its own annotation properties (*http://purl.obolibrary.org/obo/NOMEN:0000017* and *http://purl.obolibrary.org/obo/NOMEN_0000018*), which were not considered by the *synonyms per class* metric.

The clusterings performed showed that the metrics provide useful information for analyzing the ontologies. The method automatically selects the optimal number of groups based on statistical features of a set of clusterings. Each cluster contains ontologies with similar values for a given metric, which facilitates the analysis of a repository based on the metrics. Since there is no community proposal or agreement on the existence and characterization of readability levels, we cannot assign a specific readability level to each cluster at this time.

It should be noted that the target value of a metric for an ontology may depend on design decisions (e.g., multilingual ontologies should have multiple names per class) that need to be taken into account to determine the readability level. Identifying such design decisions can be complex in most cases. Consequently, the metrics may be analyzed differently when trying to characterize a repository and when analyzing a single ontology. In the first case, the values of the metrics and the structure of the corresponding clusterings will show which ontologies exhibit similar behavior with respect to these metrics. In the second case, the values of the metrics need to be tested against the design decisions of the ontologies.

B. THE REPOSITORY

Regarding RQ3, our metrics suggest that the ontologies in the OBO Foundry repository show adherence to best practices regarding the number of classes and descriptions, and the application of the 'lexically suggest, logically define' principle. Figure 8A shows that the most used annotation property for class names is *rdfs:label* with a marginal use of *skos:prefLabel* and *iao:0000589*, which is in line with the recommendations of the OBO Foundry Principle 12 on naming conventions [36]. Furthermore, the ontologies in this repository use the properties *rdfs:comment* and *IAO:0000115* (definition) to include descriptions, in accordance with OBO Foundry Principle 6 about textual definitions[2], which recommends using these two properties to include descriptive information in classes. Figure 8B shows that both annotation properties are widely used in the repository, compared to other properties that show minimal use, such as *dc:description* or *skos:definition*. The results of the *names per class* metric and the clustering obtained, with only 4 ontologies in cluster 1 (values from 1.436 to 1.472), means that this community is not developing multilingual ontologies and that most names are provided in English only.

Conversely, they do not show compliance with systematic naming or the number of synonyms. It should be noted that the analysis of the results obtained for these two best practices shows a high dependency on the nature of the domain modeled. This means that there are some ontologies that would never obtain a good value for the metrics due to the nature of the domain. It should be remembered that from a quality perspective, the metrics provide information that must be interpreted in terms of the design requirements of the ontology. Therefore, not showing compliance with a particular best practice may not be a sign of lower quality, as this best practice may not be applicable to the ontology. Therefore, in this article, we do not evaluate the quality of the ontologies, but we analyze the adherence to the principles.

The OBO Foundry uses automatic methods implemented in ROBOT to evaluate the OBO Foundry Principles. According to the results of the evaluation, they rate the ontology for this principle as pass, warning or fail, which we believe is an important, positive step forward in quality assurance for this community. It should be noted, however, that we are interested in generating information for ontology developers that could help them improve their ontologies, not in rating them, and our efforts are independent of any particular ontology community. Next, we describe some differences between our work and the evaluations performed by the OBO Foundry.

On the one hand, OBO Foundry automatically checks its Principle 12 (naming conventions) by applying the following criteria (https://obofoundry.org/principles/checks/fp_012):

1) Each label must be unique. If this requirement is not met, update at least one label to distinguish between the two terms. Add the original label to a *oboInOwl:hasExactSynonym* (alternatively, narrow, related, or broad) or *IAO:0000118* (alternative term) annotation.
2) Each entity must not have more than one label. If this criteria is not met, determine which label most accurately describes the term. Change the other label(s) to *oboInOwl:hasExactSynonym* (alternatively, narrow, related, or broad) or *IAO:0000118* (alternative term).
3) Each entity should have a label using *rdfs:label*. If this criteria is not met, add an *rdfs:label* annotation to each term that is missing a label.

This check only considers the number of names per class and the annotation property used to include the name, and thus does not cover all the recommendations of Principle 12. The requirement to have only one label per class is contrary to multilingual ontologies, which could have multiple labels in multiple languages, and no recommendation is given on how to include names in multilingual ontologies. However, it is fine to have one if the OBO Foundry ontologies are intended to be in English only, although to the best of our knowledge this is not claimed as a principle. In addition, our *systematic naming* and *LSLD* metrics are in line with most of the recommendations of Principle 12 that are not already covered by the OBO Foundry automatic check, so our results could be helpful for their evaluation work.

On the other hand, the application of the OBO Foundry Principle 6 (textual definitions)[2] is also automatically evaluated by the following criteria (https://obofoundry.org/principles/checks/fp_006):

1) Each definition must be unique. If this criterion is not met, update any duplicate definition to include some detail that distinguishes one term from another.
2) Each entity may have no more than one textual definition. If a term has more than one definition, combine the definitions. Alternatively, change a definition to a *rdfs:comment* if it only contains further details.
3) Each entity should have a textual definition using *IAO:0000115* (definition). If this criterion is not met, add a *IAO:0000115* (definition) annotation to each term that lacks a definition.

In this context, the metric *description per class* measures this aspect in a more general way, taking into account not only the annotation properties proposed by OBO Foundry, *IAO:0000115* (definition) and *rdfs:comment*, but also others such as *skos:definition* or *dcterms:description*. In addition, the criteria used by OBO Foundry for checking class definitions would be contrary to multilingual ontologies, which would not pass the check due to having one definition per supported language.

It should also be noted that the OBO Foundry has traditionally distinguished between member and candidate ontologies. The member ontologies have been manually checked and are expected to comply with the OBO Foundry principles. Table 9 shows the mean value obtained by each group and the p-value indicating the statistical significance using the Wilcoxon Rank Sum Test, also known as the Mann-Whitney test. In general, the member ontologies had a higher number of *synonyms per class* (1.43 vs. 1.02) and *descriptions per class* (0.94 vs. 0.84). In addition, the member ontologies obtained higher values for the *systematic naming metric* (0.26 vs 0.19). However, the Wilcoxon Rank Sum Test did not show a statistically significant difference between member and candidate ontologies.

For most metrics, two levels of readability have been identified in the repository by the clusters. These clusters describe the content of the repository for that metric.

TABLE 9. Comparison between the member and the candidate OBO Foundry ontologies. For each metric, the mean values for the member and the candidate set are shown, together with the p-value resulting from the Wilcoxon test.

Metric	Mean of members	Mean of candidates	p-value
Names per class	0.996625	1.0086642	0.36938832
Synonyms per class	1.429750	1.0170746	0.23752194
Descriptions per class	0.939250	0.8357090	0.49848390
Systematic naming	0.262000	0.1927857	0.08917916
Lexically suggest logically define	0.751625	0.7564173	0.91095520

The clusters separated the ontologies with higher values for each metric from those with lower values. The range of values associated with each metric could be used to determine quality-related thresholds, but that is beyond the scope of this article.

C. RECOMMENDATIONS

Next, we make some recommendations (RQ2) to the ontology developers and users inspired by the results of our research.

1) RECOMMENDATION 1: REUSE OF HUMAN-READABLE CONTENT

The minimum information to be reused to ensure human readability of ontologies should include names, descriptions and synonyms. MIREOT [47] recommends reusing labels and some textual information in addition to reusing URIs. Most of the ontologies in our corpus show adherence to the best practice of defining human-readable names for ontology classes. The exceptions are mostly due to the reuse of URIs, but not to the reuse of their annotations. Reusing only URIs makes sense from an ontology maintenance perspective, since the human-readable content of the ontology is more likely to change. However, this has limitations for an ontology that is intended to be human-readable. Ontology developers should not only reuse the minimum information suggested by MIREOT, but also follow its recommendations regarding labels and textual information.

2) RECOMMENDATION 2: DEVELOPMENT BASED ON CONTINUOUS INTEGRATION

Ontology developers should use a continuous integration process. Recommendation 1 may create additional overhead to keep reused content up to date. The development of some ontologies, such as the Gene Ontology, follows a continuous integration process. Such an approach would allow for the inclusion of a step to obtain the updated content from the reused ontologies. Interestingly, the OBO Foundry provides the Ontology Development Kit [48], which provides workflows for managing ontologies with continuous integration. This feature is not common in most standard ontology editors.

3) RECOMMENDATION 3: HUMAN READABILITY OF MULTILINGUAL ONTOLOGIES

Ontology developers must take care of readability in all languages associated with the ontology. Considerable research has been done on the multilingualism of ontologies, one of the challenges of the multilingual Web of data [49]. Our results show that the developers of the ontologies in this corpus have not paid enough attention to this aspect. The effort and cost of developing and maintaining a multilingual resource is high, we need only think of a large resource like SNOMED CT and its versions in several languages, but this has the advantage of facilitating the use of the ontology as part of the content shown to non-scientists, citizens for example, who may not speak English. Most of the ontologies in the corpus studied are not that large, so some improvements in this direction are possible. In fact, there are automatic translation tools that should be explored, always taking into account the specificity of each domain, which could require a very rigorous, supervised process.

4) RECOMMENDATION 4: STANDARDIZATION

Ontology developers need to follow community standards for readability. Our results showed that the OBO Foundry repository has a high degree of standardization in terms of the annotation properties used to set names and descriptions in classes. Figure 8 shows this homogeneity in annotation property usage. This is mainly due to the rules and recommendations provided by the OBO Foundry through its principles. Nevertheless, previous studies have shown that a variety of annotation properties are used to provide human-readable information in other repositories such as BioPortal, which is a sign of lack of standardization in the field (see Table 1 of [22] as an example). This can make it difficult to find human-readable content. Recently, the BioLink model [50] has been proposed as a universal schema for biomedical knowledge graphs. It proposes its own annotation properties for human-readable information, such as for description (*biolink:description*), but also defines the mappings to other existing properties. Thus, it can be said to provide an explicit catalog of properties that can be used in the knowledge graph. Our recommendation to the community is to move in this direction of having a clear catalog of properties that can be used for human-readable information. Regarding synonyms, our results show the use of custom properties to provide synonyms. Our recommendation is to use known annotation properties (see Table 3) for this purpose.

5) RECOMMENDATION 5: NAMING CONVENTIONS

Ontology developers should consider the benefits of systematic naming conventions to improve readability. Most OBO Foundry ontologies have a weak naming style according to the values of the *systematic naming* metric. This metric is high when the genus-differentia style of naming is used. This type of naming allows for a better understanding of ontology classes, since class hierarchies can be inferred from class names alone. In this context, the Protégé ontology editor [51] facilitates the use of the genus-differentia style by providing the ability to include common prefixes or suffixes when creating a set of classes. For example, the ontology developer can specify the suffix 'DorsalAorta' when creating the subclasses of DorsalAorta, and then the URIs of the created subclasses would start with 'DorsalAorta'. However, this functionality is only available at the URI level and does not apply to annotation properties. To mitigate this, Protégé allows the creation of *rdfs:label* annotations from the URIs, but this would not be useful for opaque URIs. Finally, it should be noted that a perfect score for the *systematic naming* metric is unlikely to be achievable in real-world scenarios; therefore, we recommend keeping the value of this metric as close to 1 as possible, while avoiding the inclusion of artificial suffixes or prefixes in class names that have a negative impact on the readability of the ontology.

6) RECOMMENDATION 6: DESCRIPTIONS ARE AS IMPORTANT AS NAMES FOR HUMAN READABILITY

Ontology developers should explain the meaning of classes by providing descriptions. The values for the metric *descriptions per class* are lower and have more variance than for *names per class* in the OBO Foundry repository, as shown in Figure 10. This clearly indicates that ontology developers do not pay as much attention to descriptions as they do to names when developing ontologies. Nevertheless, we believe that the description of a class is a key factor in making an ontology shareable, as it adds additional information that could clarify how to use that class, thus avoiding misunderstandings due to possible polysemous names. Therefore, we recommend to consider the descriptions with the same level of importance as the names when developing an ontology.

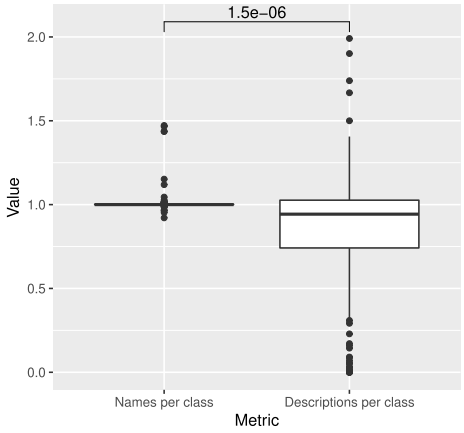

FIGURE 10. Comparison between the *names per class* and the *descriptions per class* between the OBO Foundry ontologies. The p-value returned by the Wilcoxon test is shown.

D. LIMITATIONS AND FUTURE WORK

We have presented several metrics that provide information for evaluating the human readable content in ontologies.

These metrics and related best practices can help to analyze and improve the human readability of ontologies, but they are not sufficient to perform a complete quality assurance analysis. To achieve this, they should be combined with other metrics to cover other relevant aspects of quality assurance, such as coverage in a particular domain or conciseness. For this reason, we plan to develop new metrics that cover these other aspects and combine them with metrics included in frameworks such as OQuaRE [12].

Some ontologies received low values for some of the metrics; for example, *The Core Ontology for Biology and Biomedicine* (COB), or *The Zebrafish developmental stages ontology* (ZFS), which presented the minimum value for *synonyms per class*. This can be justified by the type of knowledge included. Therefore, further research studying the values of the metrics by type of ontology could provide insights on the need to consider the type as a factor in the process of determining thresholds. Moreover, we identified cases where the reuse of top-level ontologies, such as BFO, negatively affects the *systematic naming* metric due to the inclusion of domain-independent classes that are also *LR classes*, with a high number of subclasses that do not exhibit the lexical regularity of the parent class. As future work, we plan to study the impact of classes reused from top-level ontologies on the *systematic naming* metric.

In addition, the *LSLD* and *systematic naming* metrics are calculated over the *LR classes*; however, we found ontologies, such as GNO, that had a low number of LR classes. In these cases, the values obtained for the metrics could not be representative for the ontology, since they represent a small part of the ontology. On the one hand, we decided to use only *LR classes* to calculate these metrics to avoid a high penalty on the score obtained. The fact that a class is an LR class is an indicator of the semantic relationship with the classes that have this regularity class, so we focused on these classes. In other words, the metrics would have returned much lower values if all ontology classes had been considered for their calculation. On the other hand, we are aware that this is a limitation when evaluating ontologies with a small number of lexical regularities. To correct this behavior, we plan to generate further information based on the metrics presented in this article, and to develop the metrics to take into account aspects such as the number of lexical regularities of an ontology as an additional metric, and to use this value as a weight for the *LSLD* and *systematic naming* metrics.

We have mentioned that reuse by URI creates some human readability problems due to the lack of reuse of names, descriptions or synonyms. It would also be interesting to study the real impact of this by checking whether this human-oriented content is available in the source ontology, and to study in detail the human readability of highly reused ontologies.

An area of interest for further research is the determination of thresholds for the different metrics, which would allow a more comprehensive assessment of the human readability of ontologies. This would require further experiments with other repositories.

Finally, although the framework has been applied to a repository of biomedical ontologies, the presented metrics and related best practices are domain independent, so they can be applied to evaluate the readability of any ontology from any domain. For this purpose, it would be interesting to evaluate other domains where the use of ontologies is increasing, such as the Internet of Things [52] or agronomy [1].

VI. CONCLUSION

In this work, we have proposed the HURON framework, which processes the natural language content of ontologies to compute a set of quantitative metrics related to the human readability of the ontology. The metrics have been mapped to existing best practices to support their interpretation, and implemented in a software tool that can be freely used by ontology developers and users. We applied our method to a representative set of biomedical ontologies, the OBO Foundry repository. In general, our metrics suggest that the ontologies in the repository adhere to best practices with respect to the number of classes and descriptions, and the application of the 'lexically suggest, logically define' principle. Conversely, they do not show adherence to systematic naming or the number of synonyms, although these two best practices may be highly dependent on the nature of the domain being modeled. These results have served to propose a set of recommendations to ontology developers, since the metrics are useful for generating information related to human readability. The data generated in this study could be helpful in detecting ontologies and specific entities with potential deficiencies, thus allowing their improvement based on informed decisions.

REFERENCES

[1] C. Jonquet, A. Toulet, E. Arnaud, S. Aubin, E. Dzalé Yeumo, V. Emonet, J. Graybeal, M.-A. Laporte, M. A. Musen, V. Pesce, and P. Larmande, "AgroPortal: A vocabulary and ontology repository for agronomy," *Comput. Electron. Agricult.*, vol. 144, pp. 126–143, Jan. 2018.

[2] T. Tudorache, "Ontology engineering: Current state, challenges, and future directions," *Semantic Web*, vol. 11, no. 1, pp. 125–138, Jan. 2020.

[3] C. Gaudet-Blavignac, V. Foufi, M. Bjelogrlic, and C. Lovis, "Use of the systematized nomenclature of medicine clinical terms (SNOMED CT) for processing free text in health care: Systematic scoping review," *J. Med. Internet Res.*, vol. 23, no. 1, Jan. 2021, Art. no. e24594.

[4] R. Vita et al., "Standardization of assay representation in the ontology for biomedical investigations," *Database*, vol. 2021, Jul. 2021, Art. no. baab040.

[5] G. K. Mazandu, J. Hotchkiss, V. Nembaware, A. Wonkam, and N. Mulder, "The sickle cell disease ontology: Recent development and expansion of the universal sickle cell knowledge representation," *Database*, vol. 2022, Apr. 2022, Art. no. baac014.

[6] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen, "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.*, vol. 39, pp. W541–W545, Jul. 2011.

[7] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis, "The OBO foundry: Coordinated evolution of ontologies to support biomedical data integration," *Nature Biotechnol.*, vol. 25, no. 11, pp. 1251–1255, Nov. 2007.

[8] S. Jupp, T. Burdett, C. Leroy, and H. E. Parkinson, "A new ontology lookup service at EMBL-EBI," in *Proc. SWAT4LS*, 2015, pp. 118–119.

[9] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. D. S. Santos, P. E. Bourne, and J. Bouwman, "The FAIR guiding principles for scientific data management and stewardship," *Sci. Data*, vol. 3, no. 1, pp. 1–9, Mar. 2016.

[10] D. Vrandečić, "Ontology evaluation," in *Handbook on Ontologies*. Cham, Switzerland: Springer, 2009, pp. 293–313.

[11] A. Duque-Ramos, J. Fernandez-Breis, R. Stevens, and N. Aussenac-Gilles, "OQuaRE: A square-based approach for evaluating the quality of ontologies," *J. Res. Pract. Inf. Technol.*, vol. 43, pp. 159–176, May 2011.

[12] A. Duque-Ramos, M. Boeker, L. Jansen, S. Schulz, M. Iniesta, and J. T. Fernández-Breis, "Evaluating the good ontology design guideline (GoodOD) with the ontology quality requirements and evaluation method and metrics (OQuaRE)," *PLoS One*, vol. 9, no. 8, Aug. 2014, Art. no. e104463.

[13] A. Duque-Ramos, M. Quesada-Martínez, M. Iniesta-Moreno, J. T. Fernández-Breis, and R. Stevens, "Supporting the analysis of ontology evolution processes through the combination of static and dynamic scaling functions in OQuaRE," *J. Biomed. Semantics*, vol. 7, no. 1, pp. 1–20, Dec. 2016.

[14] M. McDaniel, V. C. Storey, and V. Sugumaran, "Assessing the quality of domain ontologies: Metrics and an automated ranking system," *Data Knowl. Eng.*, vol. 115, pp. 32–47, May 2018.

[15] M. Amith, Z. He, J. Bian, J. A. Lossio-Ventura, and C. Tao, "Assessing the practice of biomedical ontology evaluation: Gaps and opportunities," *J. Biomed. Informat.*, vol. 80, pp. 1–13, Apr. 2018.

[16] A. Fernández-Izquierdo, M. Poveda-Villalón, A. Gómez-Pérez, and R. García-Castro, "Towards metrics-driven ontology engineering," *Knowl. Inf. Syst.*, vol. 63, no. 4, pp. 867–903, Apr. 2021.

[17] N. Medeiros, N. Ivaki, P. Costa, and M. Vieira, "Vulnerable code detection using software metrics and machine learning," *IEEE Access*, vol. 8, pp. 219174–219198, 2020.

[18] J. A. Bernabé-Díaz, M. Franco-Nicolás, J. M. Vivo-Molina, M. Quesada-Martínez, A. Duque-Ramos, and J. T. Fernández-breis, "An automated process for the repository-based analysis of ontology structural metrics," *IEEE Access*, vol. 8, pp. 148722–148743, 2020.

[19] L. Šikic, P. Afric, A. S. Kurdija, and M. ŠIlic, "Improving software defect prediction by aggregated change metrics," *IEEE Access*, vol. 9, pp. 19391–19411, 2021.

[20] K. Phung, E. Ogunshile, and M. Aydin, "Error-type—A novel set of software metrics for software fault prediction," *IEEE Access*, vol. 11, pp. 30562–30574, 2023.

[21] F. M. Donini, M. Lenzerini, D. Nardi, and A. Schaerf, "Reasoning in description logics," *Princ. Knowl. Represent.*, vol. 1, pp. 191–236, Feb. 1996.

[22] F. Abad-Navarro, M. Quesada-Martínez, A. Duque-Ramos, and J. T. Fernández-Breis, "Analysis of readability and structural accuracy in SNOMED CT," *BMC Med. Informat. Decis. Making*, vol. 20, no. S10, pp. 1–21, Dec. 2020.

[23] K. Donnelly, "SNOMED-CT: The advanced terminology and coding system for eHealth," *Stud. Health Technol. Informat.*, vol. 121, no. 121, p. 279, 2006.

[24] A. Third, "Hidden semantics: What can we learn from the names in an ontology?" in *Proc. 7th Int. Natural Lang. Gener. Conf.*, vol. 5, 2012, pp. 67–75.

[25] A. Rector and L. Iannone, "Lexically suggest, logically define: Quality assurance of the use of qualifiers and expected results of post-coordination in SNOMED CT," *J. Biomed. Informat.*, vol. 45, no. 2, pp. 199–209, Apr. 2012.

[26] M. Quesada-Martínez, J. T. Fernández-Breis, R. Stevens, and N. Aussenac-Gilles, "OntoEnrich: A platform for the lexical analysis of ontologies," in *Proc. Int. Conf. Knowl. Eng. Knowl. Manage.* Cham, Switzerland: Springer, 2014, pp. 172–176.

[27] P. van Damme, M. Quesada-Martínez, R. Cornet, and J. T. Fernández-Breis, "From lexical regularities to axiomatic patterns for the quality assurance of biomedical terminologies and ontologies," *J. Biomed. Informat.*, vol. 84, pp. 59–74, Aug. 2018.

[28] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[29] J. Chen, P. Hu, E. Jimenez-Ruiz, O. M. Holter, D. Antonyrajah, and I. Horrocks, "OWL2Vec*: Embedding of OWL ontologies," *Mach. Learn.*, vol. 110, no. 7, pp. 1813–1845, Jun. 2021.

[30] P. Ristoski and H. Paulheim, "RDF2Vec: RDF graph embeddings for data mining," in *Proc. Int. Semantic Web Conf.* Cham, Switzerland: Springer, 2016, pp. 498–514.

[31] A. Ritchie, J. Chen, L. J. Castro, D. Rebholz-Schuhmann, and E. Jimenez-Ruiz, "Ontology clustering with OWL2Vec," in *Proc. CEUR Workshop*, vol. 2918, 2021, pp. 54–61.

[32] T. Racharak, "On approximation of concept similarity measure in description logic ELH with pre-trained word embedding," *IEEE Access*, vol. 9, pp. 61429–61443, 2021.

[33] İ. Pembeci, "Using word embeddings for ontology enrichment," *Int. J. Intell. Syst. Appl. Eng.*, vol. 4, no. 3, pp. 49–56, 2016.

[34] Ş. Kafkas, S. Althubaiti, G. V. Gkoutos, R. Hoehndorf, and P. N. Schofield, "Linking common human diseases to their phenotypes; development of a resource for human phenomics," *J. Biomed. Semantics*, vol. 12, no. 1, pp. 1–15, Dec. 2021.

[35] J. Chen, Y. He, Y. Geng, E. Jimenez-Ruiz, H. Dong, and I. Horrocks, "Contextual semantic embeddings for ontology subsumption prediction," 2022, *arXiv:2202.09791*.

[36] D. Schober, B. Smith, S. E. Lewis, W. Kusnierczyk, J. Lomax, C. Mungall, C. F. Taylor, P. Rocca-Serra, and S.-A. Sansone, "Survey-based naming conventions for use in OBO foundry ontology development," *BMC Bioinf.*, vol. 10, no. 1, pp. 1–9, Dec. 2009.

[37] R. Jackson et al., "OBO foundry in 2021: Operationalizing open data principles to evaluate ontologies," *Database*, vol. 2021, Oct. 2021, Art. no. baab069, doi: 10.1093/database/baab069.

[38] SNOMED International. *SNOMED CT OWL Toolkit*. Accessed: Jul. 2023. [Online]. Available: https://github.com/IHTSDO/snomed-owl-toolkit

[39] J. L. Fink, P. Fernicola, R. Chandran, S. Parastatidis, A. Wade, O. Naim, G. B. Quinn, and P. E. Bourne, "Word add-in for ontology recognition: Semantic enrichment of scientific literature," *BMC Bioinf.*, vol. 11, no. 1, pp. 1–8, Dec. 2010.

[40] J. T. Fernandez-Breis, L. Iannone, I. Palmisano, A. L. Rector, and R. Stevens, "Enriching the gene ontology via the dissection of labels using the ontology pre-processor language," in *Proc. Int. Conf. Knowl. Eng. Knowl. Manage.* Cham, Switzerland: Springer, 2010, pp. 59–73.

[41] J. A. Bernabé-Díaz, M. Franco, J.-M. Vivo, M. Quesada-Martínez, A. Duque-Ramos, and J. T. Fernández-Breis. (2022). *evaluomeR: Evaluation of Bioinformatics Metrics*. R Package Version 1.7.8. [Online]. Available: https://github.com/neobernad/evaluomeR

[42] J. A. Bernabé-Díaz, M. Franco, J.-M. Vivo, M. Quesada-Martínez, and J. T. Fernández-Breis, "An automated process for supporting decisions in clustering-based data analysis," *Comput. Methods Programs Biomed.*, vol. 219, Jun. 2022, Art. no. 106765.

[43] M. Horridge and S. Bechhofer, "The OWL API: A Java API for OWL ontologies," *Semantic Web*, vol. 2, no. 1, pp. 11–21, 2011.

[44] Y. Kazakov, M. Krötzsch, and F. SimanČík, "The incredible ELK," *J. Automated Reasoning*, vol. 53, no. 1, pp. 1–61, Jun. 2014.

[45] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, nos. 3–4, pp. 591–611, Dec. 1965.

[46] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0377042787901257

[47] M. Courtot, F. Gibson, A. L. Lister, J. Malone, D. Schober, R. R. Brinkman, and A. Ruttenberg, "MIREOT: The minimum information to reference an external ontology term," *Appl. Ontology*, vol. 6, no. 1, pp. 23–33, 2011.

[48] N. Matentzoglu et al., "Ontology development kit: A toolkit for building, maintaining and standardizing biomedical ontologies," *Database*, vol. 2022, Oct. 2022, Art. no. baac087.

[49] J. Gracia, E. Montiel-Ponsoda, P. Cimiano, A. Gómez-Pérez, P. Buitelaar, and J. McCrae, "Challenges for the multilingual web of data," *J. Web Semantics*, vol. 11, pp. 63–71, Mar. 2012.

[50] D. R. Unni et al., "Biolink model: A universal schema for knowledge graphs in clinical, biomedical, and translational science," *Clin. Transl. Sci.*, vol. 15, no. 8, pp. 1848–1855, Aug. 2022.
[51] M. A. Musen, "The protégé project: A look back and a look forward," *AI Matters*, vol. 1, no. 4, pp. 4–12, Jun. 2015.
[52] A. Pliatsios, K. Kotis, and C. Goumopoulos, "A systematic review on semantic interoperability in the IoE-enabled smart cities," *Internet Things*, vol. 22, Jul. 2023, Art. no. 100754, doi: 10.1016/j.iot.2023.100754. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S254266052300077X

FRANCISCO ABAD-NAVARRO was born in Murcia, Spain, in 1989. He received the degree in computer engineering, the master's degree in bioinformatics, and the master's degree in new computer science technologies from the University of Murcia, Spain, in 2012, 2016, and 2018, respectively, where he is currently pursuing the Ph.D. degree in computer science.

Since 2016, he has been a Research Assistant with the Department of Computing and Systems, University of Murcia. He is currently a Co-Founder of the spin-off LongSeq Applications. His research interests include semantic web technologies and its applications, such as ontology evaluation or information retrieval and bioinformatics, concretely long read sequencing technologies.

CATALINA MARTÍNEZ-COSTA received the B.Sc. degree in computer engineering, the M.Sc. degree in informatics and mathematics applied to science and engineering, and the Ph.D. degree in computer science from the University of Murcia, in 2007, 2008, and 2011, respectively. She was a Visiting Researcher with UCL, London, U.K., in 2010, and a Postdoctoral Researcher with the Institute of Medical Informatics, Medical University of Graz, Austria (2012–2019). She is also a member of the IMIB-Arrixaca Bio-Health Research Institute. In 2019, she was awarded with a grant for young investigators from the Spanish Ministry of Science and Innovation (JIN program). She is also a Ramón y Cajal Postdoctoral Researcher with the Faculty of Computer Science, University of Murcia. She has been leading the design and implementation of semantic data harmonization and integration approaches in the context of EU projects, since 2012. Her current research interest includes developed within the field of biomedical informatics. More specifically, the application of semantic technologies on biomedical data, i.e., knowledge representation and management.

JESUALDO TOMÁS FERNÁNDEZ-BREIS (Senior Member, IEEE) received the degree in computer engineering and the Ph.D. degree in computer science from the University of Murcia, Spain, in 1999 and 2003, respectively. He is currently a Full Professor with the Faculty of Computer Science, University of Murcia. He is also a member of the IMIB-Arrixaca Bio-Health Research Institute. He has been leading research projects related to semantic web technologies, since 2004. His current research interests include the application of semantic technologies for the development of learning health systems and the development of quality assurance methods for ontologies and terminologies. He is also a Co-Founder of the spin-off LongSeq Applications.