Analysis of Advanced Complexity Metrics of Biomedical Ontologies in the Bioportal Repository

Yannick Kazela Kazadi[1]*, Jean Vincent Fonou-Dombeu[2]

[1] Departement of ICT, Vaal University of Technology, Private Bag X021, Andries Potgieter Blvd Vanderbijlpark, South Africa
[2] Department of Software Studies, Vaal University of Technology, Private Bag X021, Andries Potgieter Blvd.

* Corresponding author. Tel.: +27 (0)71 073 6311; email: yan.kazela@gmail.com
Manuscript submitted May 25, 2016; accepted October 26, 2016.
doi: 10.17706/ijbbb.2017.7.1.20-32

Abstract: There is an increase in the number of biomedical ontologies on the semantic web. Therefore, it is important to evaluate their complexity to promote their sharing and reuse in the biomedical domain. This study analyses and discusses the advanced complexity features of the biomedical ontologies stored in the BioPortal repository. A set of 100 biomedical ontologies from the BioPortal repository was collected. Thereafter, the collected ontologies are assigned to the analysis process to compute their advanced complexity metrics including the: size of the vocabulary, entropy of ontology graphs, the average number of paths per class, the tree impurity, class richness, percentage of part-of relations in the total number of relations, and many more. The results show that the biomedical ontologies studied are highly complex; this finding is evidenced by the analysis of their size of the vocabulary, average number of paths and entropy of ontology graph. However, it was interesting to learn that the structure of these ontologies favour their easy reuse and maintenance; these findings were reached through the analysis of the tree impurity, class and relationship richness of these ontologies.

Key words: Biomedical ontology, BioPortal, complexity of ontology, primitive ontology metrics, advanced complexity metrics.

1. Introduction

Ontology is a formal, explicit specification of a shared conceptualization of a domain of knowledge [1]. It represents knowledge as a set of concepts within a domain and the relations between them. An ontology defines the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms [2]. The emergence of the semantic web has resulted in the need for the use and development of ontologies. Therefore, as the ontologies of a given domain (medicine, geology, biomedical, e-science, etc.) grow in size and number, it is important to evaluate their complexity to help semantic web developers better understand, reuse, integrate and maintain them [3].

The evaluation of complexity of existing ontologies would reveal their underlying characteristics and provide relevant information for improving their quality for better reuse as well as estimating the cost and time for their future maintenance [4]. It is argued that[3] a quantitative measurement of the complexity of ontology improves the understanding of its structure and enables a better evaluation of its design as well as the control of its development process. Nowadays, one of the active areas of ontology development is the biomedical domain where a large number of ontologies are being developed to study classes of entities such

as the substances, qualities and processes in realities which are of biomedical significance. These classes include substances such as the mitral valve and glucose, qualities such as the diameter of the left ventricle and the catalytic function of enzymes, and processes such as blood circulation and hormone secretion [5].

This paper determines and analyses the advanced complexity features of biomedical ontologies stored in the BioPortal repository. These advanced complexity features are determined using the basic semantic features of the ontologies and appropriate programming constructs and algorithms. The BioPortal repository includes 491 biomedical ontologies and provides tools and services for browsing the ontologies. Developed during the early 2000, BioPortal is a community-based ontology repository for biomedical ontologies where users can publish, submit new versions, browse, and access the ontologies and their components through a set of REST services and SPARQL [6]. The Web interface of BioPortal allows users to browse the list of ontologies, search and comment on the terms in the ontologies, annotate text with ontology terms, and search an ontology-based index of biomedical resources [7]. Ontologies in the BioPortal are grouped into 18 categories such as: anatomy, chemical, health, human, immunology, molecule, protein, taxonomic classification, and so on [8]. However, if a new ontology falls in a category that does not exist, the administrator of the ontology can register a new category [6].

Although many studies have been conducted on the assessment of the BioPortal content [6], [7], to our knowledge, no study has focused on determining the advanced complexity features of BioPortal ontologies. In fact, the ontology metrics such as the number of classes, properties, instances, root and leaf classes, and the maximum number of children provided in BioPortal are the basic characteristics of these ontologies and constitute the primitive metrics for expressing the complexity of ontology [3], [4], [9], [10]. Therefore, advanced metrics such as the size of the vocabulary, entropy of ontology graphs, the average number of paths per class, the tree impurity, class richness, percentage of part-of relations in the total number of relations, etc., need to be determined to discuss the advanced complexity features of the BioPortal ontologies and their impact on the sharing, reuse and maintenance of these ontologies [3], [4], [9], [10].

The rest of the paper is structured as follows. Section 2 discusses related works. The materials and methods used in the study are presented in section 3. Section 4 presents experimental results and discussions. A conclusion is drawn in Section 5.

2. Related Works

Research studies have demonstrated that the level of complexity of a software artefact determines its quality [11]-[13]. In the Object-Oriented field, this has led to the proposal of several metrics for software quality including the cyclomatic complexity [11], coupling [12], Chidamber and Kemerer (CK) Object-Oriented design [13] and Metrics for Object Oriented Design (MOOD) metrics [14]. Similarly, many researchers have proposed metrics that can be used to determine the complexity of ontology. In [15] eight metrics for measuring the structural complexity of ontology represented in the form of UML class diagram are proposed. Additionally, three specific metrics for measuring the size of ontology are also proposed in [15]. Inspired by the usage of UML for representing classes and the relations between them, authors in [10] proposed a method that consists in weighting class dependence graphs to represent ontology class diagram; they further presented a structured complexity measure of the ontology based on entropy distance. This consists in assigning a value to each of the ontology classes and relations through a simple algorithm and then applying these values to the Shannon's entropy function.

In [16] three metrics, namely, number of root classes, number of leaf classes, and average depth of inheritance tree to measure the cohesiveness of an ontology are presented. This study was inspired from the principle of cohesiveness in an Object-Oriented class diagram presented by [15]. Another study based on the concept of software metrics is from [3]. This study proposed a suite of ontology metrics to measure

the design complexity of ontologies. In [4] a suite of metrics for the measurement of the complexity of an ontology are presented. These metrics examine the quantity, ratio and correlativity of concepts and relationships of ontologies.

With regard to the biomedical domain, [17] proposed a tool that enables users to select suitable biomedical ontologies for use when building applications that integrate clinical and biological data. Although some ontology metrics such as the scope of ontology, granular density and ontology integration are tackled in the study, the focus was not on the analysis and discussion of the advanced complexity features of biomedical ontologies.

Several studies have used the BioPortal as a dataset. Vescovo et al. [18] presented a large-scale investigation into the decomposability and modular aspects of 181 BioPortal ontologies and demonstrated that most of them can be split into small logically coherent parts from which modules can be efficiently assembled before reasoning. A semantic query engine that provides semantic reasoning and query processing is presented in [19]; the queries are further translated and executed on BioPortal ontologies.

3. Material and Methods

3.1. Primitive Metrics of Ontology

The primitive metrics that determine the basic characteristics of ontology include the:

- Number of classes (|C|) - Total number of classes or concepts of an ontology [9].
- Number of properties (|P|) - Total number of properties of an ontology [3], [4].
- Number of instances (|I|) - Total number of instances or individuals of an ontology [9], [10].
- Maximum depth (Λ) - Longest depth of inheritance of concepts of the ontology. The depth of inheritance of a given concept is the longest path from this concept to the root concept in the inheritance hierarchy of the ontology [9].
- Maximum number of children (NOC_{max}) - Number of subclasses of the upper class in the inheritance tree in the ontology [3].
- Average number of children (IR_s) - Average number of subclass relations per class in the ontology [3], [9].

Among the abovementioned primitive ontology metrics, the number of classes, properties and instances, and the maximum depth are used as the basis for defining the advanced complexity metrics of ontology as shown in the next subsection.

3.2. Advanced Complexity Metrics of Ontology

The advanced complexity metrics of ontology include:

- Size of vocabulary (SOV) – this metric defines the total number of named classes and instances, and properties in the ontology; it is defined as in Equation 1:

$$SOV = |C_n| + |P| + |I_n| \quad (1)$$

where $|P|$ represents the number of properties of the ontology, and $|C_n|$ and $|I_n|$, the number of named classes and instances, which are classes and instances with URIs (Uniform Resource Identifiers), respectively [20]. This is in contrast with the anonymous classes and instances which are classes and instances without URIs. A higher SOV implies that the ontology is big in size and would require a lot of time and effort to build it [3].

- The average number of paths per concept (ρ) - indicates the average connectivity degree of a concept to the root concept in the ontology inheritance hierarchy [4]. A higher ρ indicates the existence of a

high number of inheritance relationships in the ontology; it also shows that there is a high number of interconnections between classes in the ontology. This metric is defined as in Equation 2:

$$\rho = \frac{\sum_{i=1}^{m} p_i}{|C|} \tag{2}$$

where p_i is the number of paths of a given concept. The value ρ for any ontology must be greater or equal to 1; a $\rho = 1$ indicates that an ontology inheritance hierarchy is a tree.

- Tree Impurity (TIP) - this metric is used to measure how far an ontology inheritance hierarchy deviates from a tree; the TIP is defined as in Equation 3:

$$TIP = |R'| - |C'| + 1 \tag{3}$$

where R' and C', represent the sets of relations and concepts in the inheritance hierarchy, respectively. The rational of the TIP is that a well-structured ontology is composed of classes organized through inheritance relationships. A TIP=0 means that the inheritance hierarchy is a tree. The greater the TIP, the more the ontology inheritance hierarchy deviates from the tree and the greater its complexity is.

- The longest path length of a concept (λ_i) - this metric indicates the location of a concept in the ontology inheritance tree; a higher λ_i shows that the class C_i resides deeper in the inheritance hierarchy and reuses more information from its ancestors; it also indicates that the class is more difficult to maintain as it is likely to be affected by changes in any of its ancestors [3], [4]. It is defined as in Equation 4:

$$\lambda_i = \max(pl_{i,k}), 1 \le k \le p_i \tag{4}$$

where, $pl_{i,k}$ represents the length of the k-th path for the i-th concept for which λ_i is being calculated and p_i is the number of paths of that concept.

- The average path length of a concept ($\overline{\lambda_i}$) - this metric defines the average number of ancestors of a concept C_i in each of its path. It is calculated as in Equation 5:

$$\overline{\lambda_i} = \frac{\sum_{k=1}^{p_i} pl_{i,k}}{p_i} \tag{5}$$

A higher $\overline{\lambda_i}$ indicates that ~~a~~ the class C_i inherits the characteristics of many other classes in the ontology; any changes to the inherited classes will require more effort to maintain the class C_i [10].

- The average path length of an ontology ($\overline{\Lambda}$) - this metric indicates the average number of concepts in a path in the ontology. An ontology with a bigger $\overline{\Lambda}$ indicates that there too many inheritance relationships in the ontology; as a consequence, the management and manipulation of concepts in such ontology could be a complex task [4]. It is defined as in Equation 6:

$$\overline{\Lambda} = \frac{\sum_{i=1}^{m} \sum_{k=1}^{p_i} pl_{i,k}}{\sum_{i=1}^{m} p_i} \tag{6}$$

This metric is obtained from the ratio of the sum of the path lengths $(pl_{i,k})$ of each of the m concepts in the ontology over the sum of the number of paths (P_i) of concepts.

- Entropy of ontology graph (EOG) - this metric is the application of the logarithm function to a probability distribution over the ontology graph in order to provide a numerical value that can be used as an indicator of the graph complexity [3]. It is defined as in Equation 7:

$$EOG = -\sum_{i=1}^{n} p(i)\log_2(p(i)) \tag{7}$$

where $p(i)$ is the probability for a concept to have i relations. The minimum value of EOG corresponds to EOG=0, it is obtained when concepts have the same distribution of relations in the ontology, that is, all the nodes of the ontology sub-graphs have the same number of edges. Therefore an ontology with a smaller EOG can be considered as less complex in terms of relations distribution [3].

- Relationship Richness (RR) – it explains the distribution of relations in an ontology. It is the ratio of the total number of relations over the sum of the number of subclass relations and the number of relations in the ontology [9]. It is defined in Equation 8:

$$RR = \frac{|R|}{|SC| + |R|} \tag{8}$$

where, $|R|$ and $|SC|$ represent the number of relations between classes and the number of subclass relations in the ontology, respectively. A RR value close to one indicates that most of the relations between concepts in the ontology are not subClassOf relations, while a RR close to zero specifies that the subClassOf relations are predominant amongst the concepts of the ontology [21].

- Class Richness (CR) - the value of this metric explains the distribution of individuals or instances in the ontology [9]. It is the ratio of the total number of classes having at least one instance (|C'|) over the total number of classes (|C|) in the ontology. Its definition is provided in Equation 9.

$$CR = \frac{|C'|}{|C|} \tag{9}$$

According to Tartir *et al*. [9], a CR close to one indicates that most of the ontology classes have instances.

3.3. Calculation of Advanced Complexity Metrics

For each ontology in the dataset, the advanced complexity metrics in Equations 1 to 9 are computed in Java Jena Application Programming Interface (API) [22]. A Jena Model is built for each ontology. Thereafter, the Jena Model is processed to compute the relevant primitive ontology metrics which are then used to calculate the advanced complexity metrics. The processing of the Jena Model requires the design and use of various Java constructs including Arrays, Queues, Lists and Iterators as well as the design and implementation of appropriate algorithms.

4. Experiments

4.1. Dataset

The dataset is constituted of 100 ontologies downloaded from the BioPortal Repository. These ontologies are listed in Tables 1 and 2 and are the semantic modelling of different branches of the biomedical domain. They include:

- Ontologies of different kinds of diseases and their impact on human and animal bodies – Examples are

the Alzheimer disease ontology (O_2 in Table 1), HIV ontology (O_{73} in Table 2) and Dengue Fever ontology (O_7 in Table 1).

Table 1. List of Biomedical Ontologies in the Dataset — Part I

Index	Ontology Name	Index	Ontology Name
O_1	Information Consent Ontology	O_{26}	Non-codingRNA
O_2	Alzheimer's Disease Ontology	O_{27}	Semantic Science Ontology
O_3	Bone dysplasia Ontology	O_{28}	Statistic Ontology
O_4	Cigarette Smoke Exposure Ontology	O_{29}	Neural Electromagnetic Ontology
O_5	Ontology of vaccine advert events	O_{30}	New Born Ontology
O_6	Dermatology Lexicon	O_{31}	Parkinson Disease Ontology
O_7	Dengue Fever Ontology	O_{32}	Animal trait ontology
O_8	Galen Ontology	O_{33}	Ontology of Pneumology
O_9	Human Dermatological Ontology Disease	O_{34}	Metagenome and Microbiology Ontology
O_{10}	Human Interaction Network Ontology	O_{35}	Human Physiology simulation ontology
O_{11}	Natural Products Ontology	O_{36}	Sleep Domain Ontology
O_{12}	NCI Thesaurus	O_{37}	The Drug-Drug Interaction Ontology
O_{13}	Ontology of Adverse Events	O_{38}	Hymenoptera Anatomy Ontology
O_{14}	Ontology of drug neuropathy adverse event	O_{39}	Congenital Health Defects
O_{15}	Orphanet Rare Disease Ontology	O_{40}	Environment ontology for livestock
O_{16}	Uber Anatomy Ontology	O_{41}	Phenotype Quality Ontology
O_{17}	Vaccine Ontology	O_{42}	Human dermatological disease Ontology
O_{18}	Experimental Factor Ontology	O_{43}	Cognitive Atlas Ontology
O_{19}	Human Disease Ontology	O_{44}	Cell type ontology
O_{20}	Cell Ontology	O_{45}	Ontology of physics for biology
O_{21}	Human Phenotype Ontology	O_{46}	Ontology of MicroRNA Target
O_{22}	Chemical Entities of Biological Interest	O_{47}	Mass Spectrometry
O_{23}	Diabetes Ontology	O_{48}	Adult mouse brain
O_{24}	Nano particle Ontology	O_{49}	Ontology of biological and clinical statistic
O_{25}	Pathogenic diseases	O_{50}	Radio oncology ontology

Table 2. List of Biomedical Ontologies in the Dataset — Part II

Index	Ontology Name	Index	Ontology Name
O_{51}	Vertebrate Skeletal Ontology	O_{76}	Eagle resource research
O_{52}	BioAssay Ontology	O_{77}	Plant experimental assay Ontology
O_{53}	Emotion Ontology	O_{78}	Ontology of Drug Neuropathy adverse events
O_{54}	Neuroscience Ontology	O_{79}	Neural-Immune Gene Ontology
O_{55}	Neuroscience Information Ontology	O_{80}	Kinetic simulation algorithm ontology
O_{56}	Ontology of genetic interval	O_{81}	Chemical Information Ontology
O_{57}	Population and Community Ontology	O_{82}	Sequence phenotype ontology
O_{58}	Beta Cell Genomics Ontology	O_{83}	Disease core rare disease Ontology
O_{59}	Enano Mapper Ontology	O_{84}	Drug Interaction Knowledge Base Ontology
O_{60}	Experimental Factor Ontology	O_{85}	Cell line Ontology
O_{61}	Immuno-genetics Ontology	O_{86}	Breast Cancer Ontology
O_{62}	NanoParticle Ontology	O_{87}	Multiple Sclerosis Ontology
O_{63}	Brain Region Ontology	O_{88}	Autism spectrum ontology
O_{64}	Mental Functioning ontology	O_{89}	Infectious Disease Ontology
O_{65}	Clinical Measurement Ontology	O_{90}	Translational medicine ontology
O_{66}	Fission Yeast Phenotype Ontology	O_{91}	Ecosystem ontology
O_{67}	Adult Brain Ontology	O_{92}	Ontology of alternative medicine
O_{68}	Clinical Trials Ontology	O_{93}	Family Health History Ontology
O_{69}	Fanconi Anemia Ontology	O_{94}	Symptom Ontology
O_{70}	Medical image simulation	O_{95}	Cancer Management and research ontology
O_{71}	Anatomical entity Ontology	O_{96}	Biomedical Resource Ontology
O_{72}	Single-Nucleotide Polymorphism Ontology	O_{97}	Growth medium ontology
O_{73}	HIV Ontology	O_{98}	Epidemiology Ontology
O_{74}	Cardiac Electrophysiology Ontology	O_{99}	Ontology of clinical research
O_{75}	Flora phenotype	O_{100}	Mental State Assessment

- Ontologies of human and animal anatomy — These ontologies encompass the vertebrate skeletal

ontology (O_{51} in Table 2) and anatomical entity ontology (O_{71} in Table 2).

- Ontologies of treatment products and their effects on the human body — Examples of these ontologies include the vaccine ontology (O_{17} in Table 1), the ontology of adverse events (O_{13} in Table 1) and the Natural products ontology (O_{11} in Table 1).
- Ontologies of organization of molecules and proteins and their different processes in the human and animal bodies – Examples are: cell ontology, sequence phenotype ontology and the Non-coding RNA (O_{26} in Table 1).
- Ontologies of cancer and treatment methods - Examples of these include the Breast cancer ontology (O_{86} in Table 2), cancer management and research ontology (O_{95} in Table 2) and the radio oncology ontology (O_{50} in Table 1).

4.2. Computer Environments

The experiments were carried out on a computer with the following characteristics: 64-bit Genuine Intel (R) Celeron (R) CPU 847, Windows 8 release preview, 2 GB RAM and 300 GB hard drive. The algorithms for computing and analyzing the complexity metrics were implemented in Java Jena API [22] configured in Eclipse Integrated Development Environment (IDE) Version 4.2.

4.3. Experimental Results

4.3.1. Calculation of primitives metrics of ontology

In order to compute the advanced complexity metrics for all the ontologies in the dataset, it was necessary to determine the basic semantic characteristics of these ontologies such as the number of classes, properties and instances. To this end, appropriate data structures and programs were designed and implemented in Java Jena API.

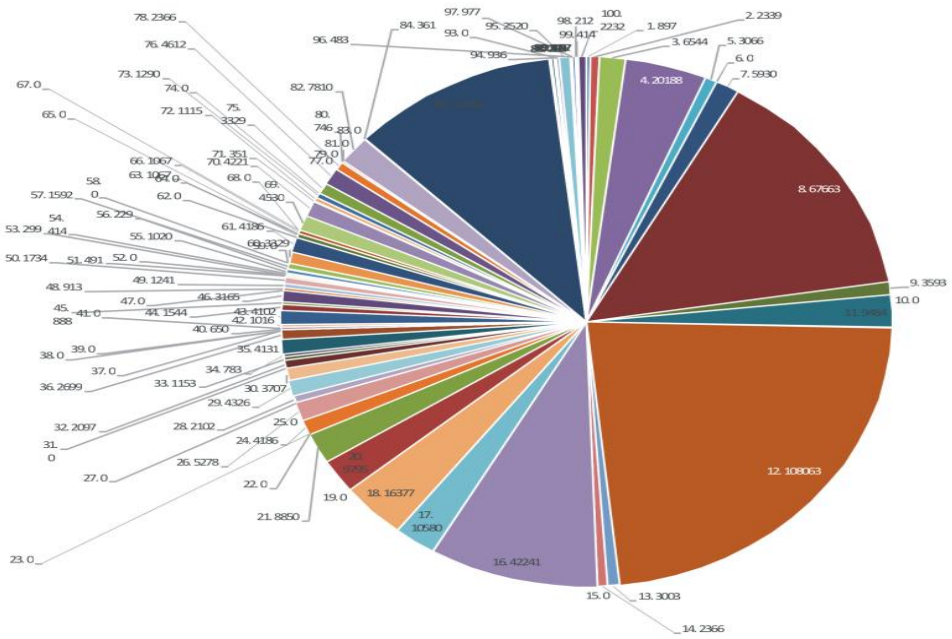

Fig. 1. Chart of the number of concepts in the biomedical ontologies in the dataset.

Fig. 1, 2 and 3 depict the charts of the basic semantic characteristics of the biomedical ontologies in the dataset including the number of classes (Fig. 1), properties (Fig. 2) and instances (Fig. 3). These characteristics appear in Fig. 1, 2 and 3 as pairs of values in the form x.y. The value x represents the index of the ontology in Table 1 or Table 2 and the value y, either the number of concepts, properties or instances of

the ontology O_x. For instance, the largest bar in the chart in Fig. 1 corresponds to the pair 12.108063; this means the ontology O_{12} in Table 1, that is, the National Cancer Institute (NCI) Thesaurus has 108063 classes. Similarly, the pair 78.2366 at the top left of Fig. 1 means that the ontology O_{78} in Table 2, that is, the Ontology of Drug Neuropathy adverse events has 2366 classes.

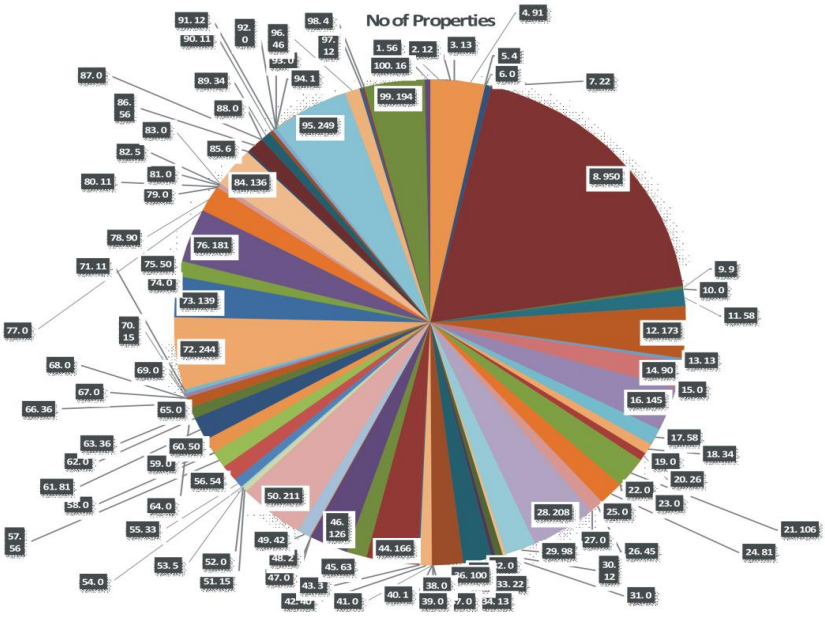

Fig. 2. Chart of the number of properties of biomedical ontologies in the dataset.

Fig. 3 depicts the chart of the number of instances in the biomedical ontologies in the dataset. The ontology with the most instances is O_{11} in Table 1, that is, the Natural Products Ontology with 22012 instances, followed by the NCI Thesaurus (O_{12} in Table 1) with 4141 instances. Overall, Fig. 3 shows that the majority of selected ontology in the BioPortal as dataset for this study had a lower number of instances.

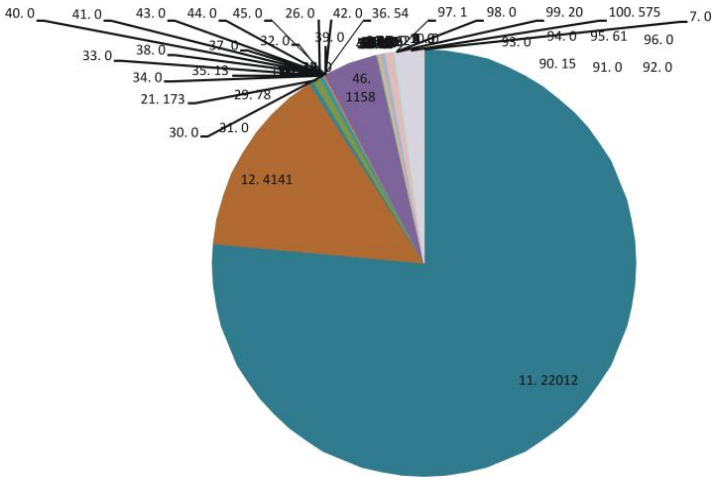

Fig. 3. Chart of the number of the instances of biomedical ontologies in the dataset.

4.3.2. Analysis and discussion of the advanced complexity metrics of biomedical ontologies

1) Size of the Vocabulary (SOV) – Fig. 4 presents the results of the measurement of the SOV for all

ontologies in the dataset. These results are grouped into 8 ranges from the range of ontologies with a SOV less than 1k (i.e. 1000) to the range of the ones with SOV >100k (i.e. 100000).

The majority (56%) of the ontologies in the dataset have a SOV between 1000 and 15000, followed by those with a SOV less than 1000 (31%); 5% of ontologies in the dataset have a SOV between 15000 to 30000 and 2% a SOV of more 100000. These results indicate that the large majority of ontologies in the dataset are constituted of thousands or tens of thousands of components. Then, it would be beneficial for semantic web developers in the biomedical domain to consider the reuse of these larger ontologies (Uber Anatomy ontology (O_{16}, SOV=42386), Vaccine Ontology (O_{17}, SOV=10706)) rather than trying to build new related ontologies *de novo*. The SOV of these ontologies also suggests that they would require a larger amount of time and effort to build [3].

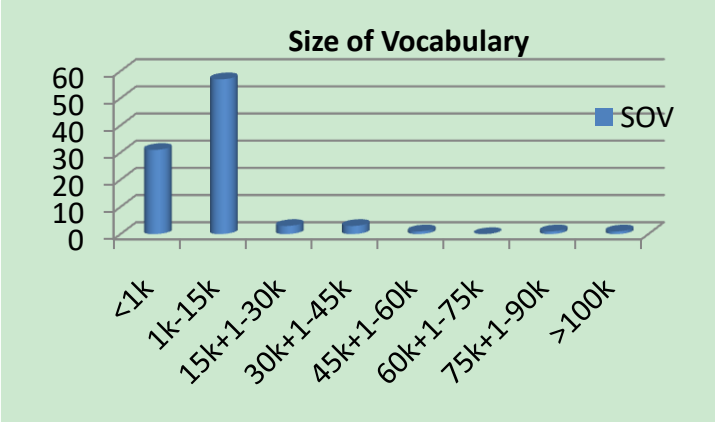

Fig. 4. The size of vocabulary.

2) Average path length of the ontology and Average Number of Paths per Concept – Fig. 5 presents a joint analysis of the average path length of the ontology and the average number of paths per concept or class (ρ). The values of these 2 metrics for all the ontologies in the dataset are grouped into 11 ranges as in Fig. 5. In Fig. 5 it is shown that a considerable proportion of the ontologies in the dataset (36%) have a ρ value less than 5; a larger number of these ontologies (37%) have a ρ between 6 and 15; 6% of ontologies in the dataset have a ρ between 36 and 45. A smaller number of ontologies (4%) have ρ in one of the following ranges 16-25, 26-35, 46-55 and 66-75.

From the analysis of the values of the ρ for all the ontologies in the dataset, one can conclude that the majority of the ontologies in the dataset have multiple paths from the root class to given classes; this indicates that in most of these ontologies the inheritance relationships among the classes are intense and constitute a sign of higher complexity of these ontologies. Once more, building similar ones from scratch would require a lot of time and effort [4]. Fig. 5 also portraits that the majority of ontologies in the dataset (94%) have smaller $\overline{\Lambda}$ values (less than 5). This indicates that changes in a class in these ontologies would have a less impact on its sub-classes [4].

3) Entropy of the Ontology Graph or Inheritance Hierarchy – Fig. 6 presents the chart of EOG for the ontologies in the dataset. The bars in the chart in Fig. 6 represent the percentage of ontologies with EOG in the corresponding range of EOG values. Fig. 6 depicts that many of the ontologies in the dataset have EOG between 2 and 2.499 (41%); followed by those with EOG in the range of 1.5 to 1.999. A significant group of these ontologies have EOG between 1 and 1.499 (14%). A smaller number of the ontologies in the dataset have EOG close to zero. This indicates that the structures of the majority of ontologies in the dataset are less regular, which is a sign of higher complexity of these ontologies [3].

4) Tree Impurity (TIP) – Fig. 7 presents results of the calculation of TIP for all the ontologies in the dataset. These results are classified into 5 groups in Fig. 7 based on the TIP values.

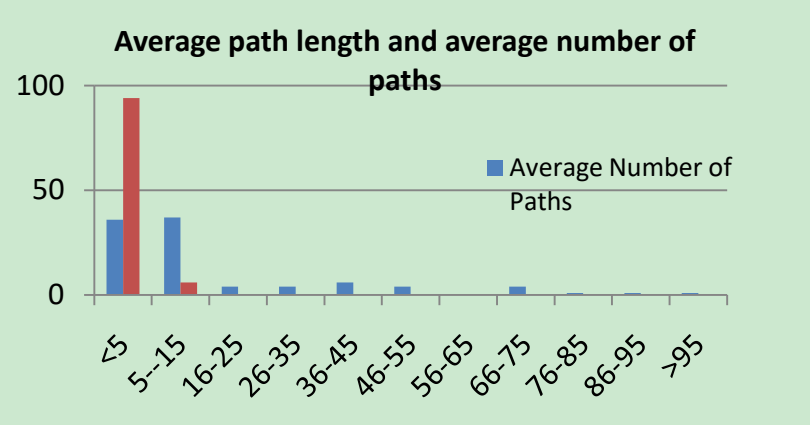

Fig. 5. Average path length and average number of paths.

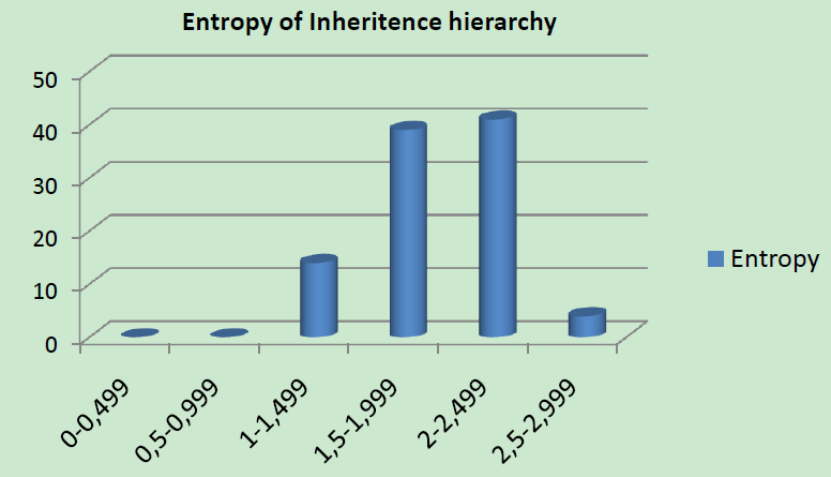

Fig. 6. Entropy of inheritance hierarchy.

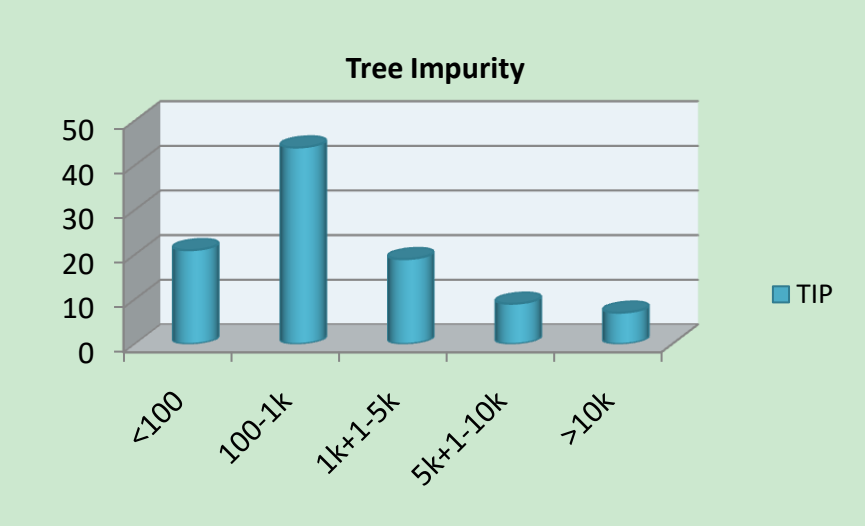

Fig. 7. Tree impurity.

It is shown in Fig. 7 that an important number of ontologies in the dataset (44%) have TIP between 100 and 1000 (k); followed by those with TIP below 100 (21%). The remaining groups of ontologies have TIP in

the ranges (1k+1) to 5k (1001 to 5000), (5k+1) to10k (5001 to 10000) and >10k (10000). These results suggest that the average number of subclass relations per class is low in these ontologies; this indicates that they can be easily reused and maintained [3].

5) Relationship Richness – Fig. 8 presents the results of a joint analysis of the relationship and class richness metrics. Fig. 8 shows that 99% of the ontologies in the dataset have a RR between 0.5 and 0.74999 and all of them have CR values less than 0.25. This indicates that there is a balance between the number of SubClassOf and non-SubClassOf relationships and that most of the classes of these ontologies do not have instances.

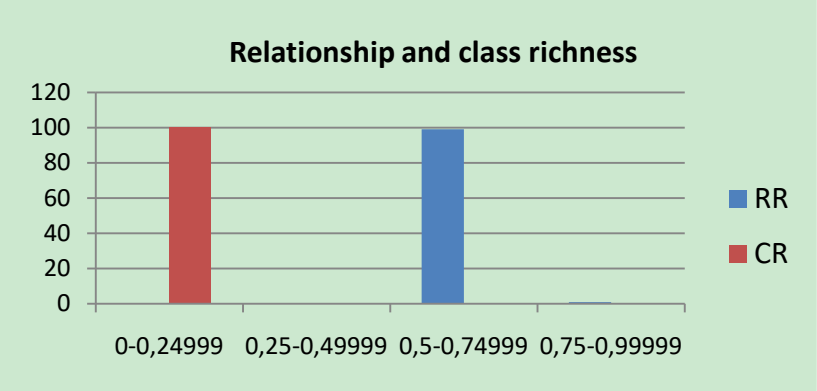

Fig. 8. Relationship and class richness.

5. Conclusion and Future Work

This paper presented an analysis of the advanced complexity features of 100 biomedical ontologies from the BioPortal repository. These features include the average number of paths per class, the size of vocabulary, the entropy of an ontology graph, the class and relationships richness, the tree impurity, the average path length of the ontology and the average number of paths per concept. The analysis of these advanced complexity metrics of biomedical ontologies in the dataset portrays that the majority of these ontologies have large size of vocabulary (SOV), and bigger average path length (ρ) and entropy of ontology graph (EOG). These findings indicate that the biomedical ontologies in the dataset are highly complex [3], [4]. It would therefore be advised to consider the reuse and sharing of these ontologies in the biomedical domain rather than trying to build similar ontologies *de novo*; the reuse may consist in using (1) parts of existing biomedical ontologies to build new ones or (2) the full ontologies in new applications [23]. In fact, ontology reuse (1) reduces human efforts required to formalized new ontologies from scratch, (2) increases the quality of the resulting ontologies because the reused ontologies have already been tested, (3) simplifies the mapping between ontologies built using shared components of existing ontologies, and (4) improves the efficiency of ontology maintenance [23]. Furthermore, the analysis of the tree impurity (TIP), relationship richness (RR) and class richness (CR) metrics revealed that the biomedical ontologies in the dataset can be easily reused and maintained [3], [9], [21]. These findings are supported by the fact that the biomedical ontologies concerned are available for download free of charge on the BioPortal repository and many researches [6], [7], including this study, provide metadata that may be useful in understanding, reusing, sharing and maintaining these ontologies in the biomedical domain.

In future, we intend to develop a framework for classifying the ontologies in the dataset based on their level of complexity.

References

[1] Gruber, T. R. (1993). Toward principles for the design of ontologies used for knowledge sharing. *International Journal Human-Computer Studies*, *43(1)*, 907-928.

[2] Neches, R., Fikes, R. E., Finin, T., Gruber, T. R., Senator, T., & Swartout, W. R. (1990). Enabling technology for knowledge sharing. *AI Magazine*, *12(3)*, 36-56.

[3] Zhang, H., Li, Y. F., & Tan, H. B. K. (2010). Measuring design complexity of semantic web ontologies. *Systems and Software*, *83(3)*, 803-814.

[4] Yang, Z., Zhang, D., & Ye, C. (2006). Evaluation metrics for ontology complexity and evolution analysis. *Proceedings of the IEEE International Conference e-Business Engineering* (pp. 162–170).

[5] Bodenreider, O., Peters, L., Kapusnik-Uner, E., & Nguyen, T. (2011). An approximate matching method for clinical drug names. *Proceedings AMIA Annual Symposium* (pp. 1117-1126).

[6] Salvadores, M., Alexander, P. R., Musen, M. A., & Noy, N. F. (2013). BioPortal as a dataset of linked biomedical ontologies and terminologies in RDF. *Semant Web*, *4(3)*, 277-284.

[7] Whetzel, P. L, & Team, N. (2013). NCBO technology: Powering semantically aware applications. *Journal of Biomedical Semantics, 4(1)*, 88-95.

[8] Wbio. Welcome to the NCBO Bioportal. Retrieved from the website: http://bioportal.bioontology.org

[9] Tartir, S., Arpinar, B., Moore, M., Sheth, A., & Aleman-Meza, B. (2005) OntoQA: Metric-based ontology quality analysis. *Proceedings of IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources* (pp. 45 -53).

[10] Kang, D., Xu, B., Lu, J., & Chu, W. C. (2004). A complexity measure for ontology based on UML. *Proceedings 10th IEEE International Workshop on Future Trends of Distributed Computing Systems* (pp. 222-228).

[11] Mccabe, T. J. (1976). A complexity measure. *IEEE Transactions on Software Engineering, 2(4)*, 308–320.

[12] Fenton N., & Melton A. (1990). Deriving structurally based software measures. *Journal of Systems and Software*, *12(2)*, 177-187.

[13] Chidamber S, & Kemerer C. (1994). A metrics suite for object oriented design. *IEEE Transactions on Software Engineering*, *20(6)*, 476-493.

[14] Brito, E., Abreu, F., & Melo, W. (1996). Evaluating the impact of object-oriented design on software quality. *Proceedings of 3rd International Metric Symposium* (pp. 90-99).

[15] Manso, M, Genero, M, & Piattini, M. Non-redundant metrics for UML class diagram structural complexity. *Proceedings 15th International Conference Advanced Information Systems Engineering* (pp. 50-65).

[16] Yao. H., Orme. A., & Etzkorn. L. (2005). Cohesion metrics for ontology design and application. *Journal of Computer Science*, *1(1),* 107-113.

[17] Maiga, G., & Ddembe, W. (2009). Flexible biomedical ontology selection tool. *International Journal of Computing and ICT Research, Special Issue, 3(1)*, 53-66.

[18] Vescovo, C., Parsia, B., Sattler, U., & Schneider, T. (2011). The modular structure of an ontology: Atomic decomposition. *Proceedings International Joint Conference Artificial Intelligence* (pp. 2232–2237).

[19] Zheng, S., Wang, F., & Lu, J. (2014). Enabling ontology based semantic queries in biomedical database systems. *International journal of Semantic Computing*, *8(1)*, 67–83.

[20] Mallea, A., Arenas, M., Hogan, A., & Polleres, A. (2011). On blank nodes. *Proceedings 10th International Semantic Web Conference* (pp. 421-437).

[21] Sugumaran, V., & Gula, J. T. (2012). *Applied Semantic Web Technologies*. Tailor & Francis Group.

[22] Mcbride, B. (2001), Jena: Implementing the RDF model and syntax specification. *Proceedings 2nd International Workshop on the Semantic Web - SemWeb'2001* (pp. 308-320).

[23] Ding, Y., Lonsdale, D., Embley, D. W., Hepp, M., & Xu, L. (2007). Generating ontologies via language

components and ontology reuse. *Proceedings of the 12th International Conference on Applications on Natural Language to Information Systems [NLDB'07]* (pp. 131-142).

Yannick Kazela Kazadi is currently a master student in information and communication technology at the Vaal University of Technology. He holds a baccalareus of technology in information technology from the Vaal University of Technology (South Africa) and a bachelor of science in mathematics and computer sciences from the Mouloud Mammeri University of Tizi-Ouzou (Algeria). Mr Kazadi's research interests are in: semantic web, ontology engineering, and machine learning.

Jean Vincent Fonou-Dombeu is a senior lecturer of software studies at the Vaal University of Technology. He holds a PhD in computer science from the North-West University, South Africa, an MSc in computer science from the University of KwaZulu-Natal, South Africa, and BSc (Hons) and BSc in computer science from the University of Yaoundé I, Cameroon. Dr Fonou-Dombeu's research is in ontology engineering, semantic web and application of semantic technologies in e-government, e-business and e-science. His research has appeared in peer reviewed journals such as Lecture Notes in Computer Science (LNCS), International Journal of Web and Semantic Technology (IJWesT), Journal of Emerging Technologies in Web Intelligence (JETWI) and African Journal of Information Systems. Dr Fonou-Dombeu has also presented papers at international conferences in South Africa, France, Slovenia, Italy, Germany and Spain.