Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Hlomani Hlomani [*], and Deborah Stacey
School of Computer Science, University of Guelph, 50 Stone Road East, Guelph, Ontario, N1G 2W1, Canada
E-mail: {hhlomani,dastacey}@uoguelph.ca

Abstract. Ontology evaluation is concerned with ascertain two important aspects of ontologies: quality and correctness. The distinction between the two is attempted in this survey as a way to better approach ontology evaluation. The role that ontologies play on the semantic web at large has been has been argued to have catalyzed the proliferation of ontologies in existence. This has also presented the challenge of deciding one the suitability of an given ontology to one's purposes as compared with another ontology in a similar domain. This survey intends to analyze the state of the art in ontology evaluate spanning such topics as the approaches to ontology evaluation, the metrics and measures used. Particular interest is given to Data-driven ontology evaluation with special emphasis on the notion of bias and it's relevance to evaluation results. Chief among the outputs of this survey is the gap analysis on the topic of ontology evaluation.

Keywords: Ontology, ontology evaluation, ontology evaluation methods, ontology evaluation metrics, ontology evaluation approach

1. Introduction

Ontology evaluation is a emerging field that has a number of frameworks and methodologies in existence [40]. The relevance and importance of ontology evaluation is evident in the role the play in the semantic web and ontology-enabled applications. They are the centrepiece of knowledge description in the semantic web allowing for the definition of a shared knowledge base that can be acted upon by agents acting on the behalf of humans. Given this role, they have since attracted a lot of interest from both academic and industrial domain leading to the proliferation of ontologies in existence. While this is attractive, it presents a challenge in deciding which ontology to reuse (as they are reusable knowledge artefacts) and hence the topic of ontology evaluation.

This paper discusses some of the frameworks and methodologies for ontology evaluation with the view of performing a gap analysis. It first gives a context to ontology evaluation by defining the notion of ontology evaluation (Section 2.1) and discussing ontology evaluation in the context of ontology reuse as an example scenario for the role of ontology evaluation (Section 2.2). This is then followed by an overview of related work that has be done on ontology evaluation by breaking it down into several categories (Section 4). Prior to the discussion on the different categories of ontology evaluation, this survey discusses measures/metrics that have been used in the different methods to decide the quality of ontologies (Section 3). Finally the paper concludes by discussing what still

[*] Corresponding author. E-mail: hhlomani@uoguelph.ca

0000-0000/14/$00.00 © 2014 – IOS Press and the authors. All rights reserved

needs to be done with respect to ontology evaluation (Section 6).

2. A context for ontology evaluation

An ontology has been previously defined as an explicit formal specification of a conceptualization where the conceptualization in this context refers to the abstraction of a domain of interest [18]. These abstractions are increasingly used in various fields such as information extraction, data integration, and the biggest of which is the semantic web. This apparent increase in the use of ontologies has lead to an increase in the number of ontologies in existence which in turn has heightened the need for evaluating the ontologies.

2.1. A definition

Generally ontology evaluation can be defined as the process of deciding on the quality of an ontology in respect to a particular criterion with the view of determining which in a collection of ontologies would best suit a particular purpose [2]. An interesting definition of ontology evaluation has be given by Gómez-Pérez et al. [17] and later echoed by Vrandecic et al. [40]. In these works, ontology evaluation is defined in the context of two interesting concepts; verification and validation. The definition is interesting because it also offers a way to categorize current ontology evaluation endeavours. Ontology verification is concerned with building an ontology correctly, while ontology validation on the other hand is concerned with building the correct ontology.

2.2. Ontology integration and merging

A considerable number of ontologies have been created. It follows, therefore, that the time has come for these ontologies to be reused. Ontology integration and merging offers the two most obvious scenarios for the uses of ontology evaluation under the frame of ontology reuse [31]. Their primary aim is to build an ontology from existing ontologies. There are subtle differences between the two concepts but both are a part of the ontology building process. Ontology integration is focused on building an ontology through specializing, extending and/or adapting existing ontologies that form a part of the resultant ontology. For example, given an algorithms ontology, one may introduce another class or type of algorithm (e.g greedy, simple recursive) within the same ontology thereby specializing the algorithms class. Ontology merging, on the other hand, is concerned with merging different ontologies into a single one. For example, if two ontologies model time: temporal ontology (O_1) and time ontology (O_2), then O_1 and O_2's concepts and relations are strategically merged together resulting in a new ontology ($O_1 + O_2 = O_\alpha$)

3. Metrics for ontology evaluation

3.1. Ontology evaluation metrics: State of the art

The relevance of ontologies as engineering artifacts has been justified in literature. Hence, like every engineering artifact, evaluation is a precursor to their (re)use. The challenge, however, has been on how to actually go about evaluating these engineering artifacts. In this regard, most questions have been directed at the applicability of software evaluation metrics to ontology evaluation. Arguments about this applicability stem from distinguishing ontologies from software processes, and rather seeing them as data models [20]. It has been said that “an ontology is a conceptualization of a domain" [18]. Guarino [20], however, reduces this definition to reflect that ontologies are rather *approximate* specifications of a domain and argues that ontology evaluation should reflect the degree of such approximation. In the same work, propositions of metrics relevant to information retrieval were suggested, more specifically the notions of precision and recall (equated with coverage). These propositions are echoed in Brewster et al. [4] and more recently in Ouyang et al. [25], albeit with a warning on naive interpretations of the meaning of precision, and recall or coverage.

Burton-Jones et al. [6]'s approach to ontology quality takes from the measurement tradition in software engineering. From this viewpoint, measures of a program's internal attributes (such as coupling and cohesion) are believed to influence external quality attributes (such as maintainability and performance), and hence their measurement is indicative of the program's quality at least in the context of the aforementioned quality attributes. In this regard, they proposed a theoretical framework based on a metric suite consisting of four metrics, and ten attributes, where the metrics are: (i) Syntactic quality, (ii) Semantic quality, (iii) Pragmatic quality, and (iv) Social quality [6].

Ontology quality metrics were described and partitioned into three main types in Gangemi et al. [14]:

(i) Structural metrics - those that are concerned with the syntax and semantics aspects, (ii) Functional metrics - those that are focused on the intended use of the ontology and its components (basically, their function in a given context), and (iii) Usability-profiling - those that are focused on the communication aspect of an ontology (are also reliant of the ontology annotations and are reminiscent of the pragmatic quality of Burton-Jones [6]).

3.2. Ontology evaluation measures: Perspective,criteria, metrics

After an analysis of the state of the art in ontology evaluation, this work perceives ontology evaluation to be done in the view of two complementary perspective: (i) Ontology Quality, and (ii) Ontology Correctness.

Ontology quality perspective Most research on ontology evaluation considers ontology evaluation from the ontology quality perspective. Burton-Jones et al. [6]'s approach discussed earlier epitomizes this quality-oriented view to ontology evaluation. They discuss a software engineering-influenced view to ontology quality where internal attributes (which are usually directly measurable) influence the external quality attributes (which are not usually directly measurable).

Ontology correctness perspective According to Guarino and Gómez-Pérez [20,17], ontology evaluation is defined in the view depicted in Figure 1. An ontology here is rather an *approximate* specification of a domain, whereas, ontology evaluation is concerned with the degree or rather the distance between this approximate conceptualization (the model) and the real world. This definition of ontology evaluation is consistent with the validation aspect of the ontology evaluation definition provided by Gómez-Pérez et al., and Vrandecic et al. [17,40] since the concern is in determining if the correct model of the domain was created. This has also been the current focus of recent research in data-driven ontology evaluation such as [4,25,22].

There exists different schools of thought on these perspectives. For example, does a good quality (referred here in terms of its standard) ontology necessarily mean it is correct? Conversely, does a correct ontology have any reflection on its quality? Take *maintainability* as decision criteria of the quality of an ontology for example. If it is measured through the relevant metrics such as coupling and get a high score, does this necessarily mean we have a correct ontology? Certainly, not. On the other hand, if we look at an ontology's accuracy as a measure of its correctness through measuring its coverage of the domain and get a high score, does it mean the ontology is of a high quality? These scenarios necessitate the need for separation of these concerns and advocates for separate determination of each.

3.3. A metric suite based on the ontology quality and correctness perspectives

With these perspectives in mind (quality and correctness), this paper proposes a four-layered metric suite for ontology evaluation. This metric suite is reminiscent of Button-Jones et al. [6]' metric suite for ontology auditing. The main difference is that in Button-Jones's case all metrics are cluttered under the umbrella of *ontology quality*. This paper distinguishs between quality and correctness foci as has already been discussed. Table 1 depicts this four-layered metric suite for ontology evaluation. The primary focus is to evaluate an ontology with the view of making a decision on whether to (re)use the ontology or not. Hence the first layer is the overall ontology evaluation. The second layer is based on the perspective from which the evaluation is conducted from, i.e. whether the evaluation is to determine the quality of an ontology or the correctness of the ontology. The third layer is concerned with the criterion used in deciding either the quality or correctness (or both depending on the purpose of the evaluation) of the ontology. The last layer specifies the quantitative measures which are indicative of the level of satisfaction of the criterion.

This metric suite has been largely based on Verendicic [41]'s evaluation criteria and Button-Jones [6]'s metric suite. The eight criteria by Verendicic is in turn derived from analysis of previous publications on ontology evaluation criteria, more precisely, the works of Gomez-Perez [17], Gruber [18], Gruninger and Fox [19], Obrst et al. [28], Gangemi et al. [13]. The metric suite is not by any means an exhaustive list of evaluation criterion nor does it exhaustively list all attributes that can be measured for each criterion but rather provides a frame from which the criterion or metric can be understood from. Table 2 provides the definitions of the terms in the metric suite.

4. Approaches to ontology evaluation

Now that the previous section has established what ontology evaluation is and given ontology reuse as an

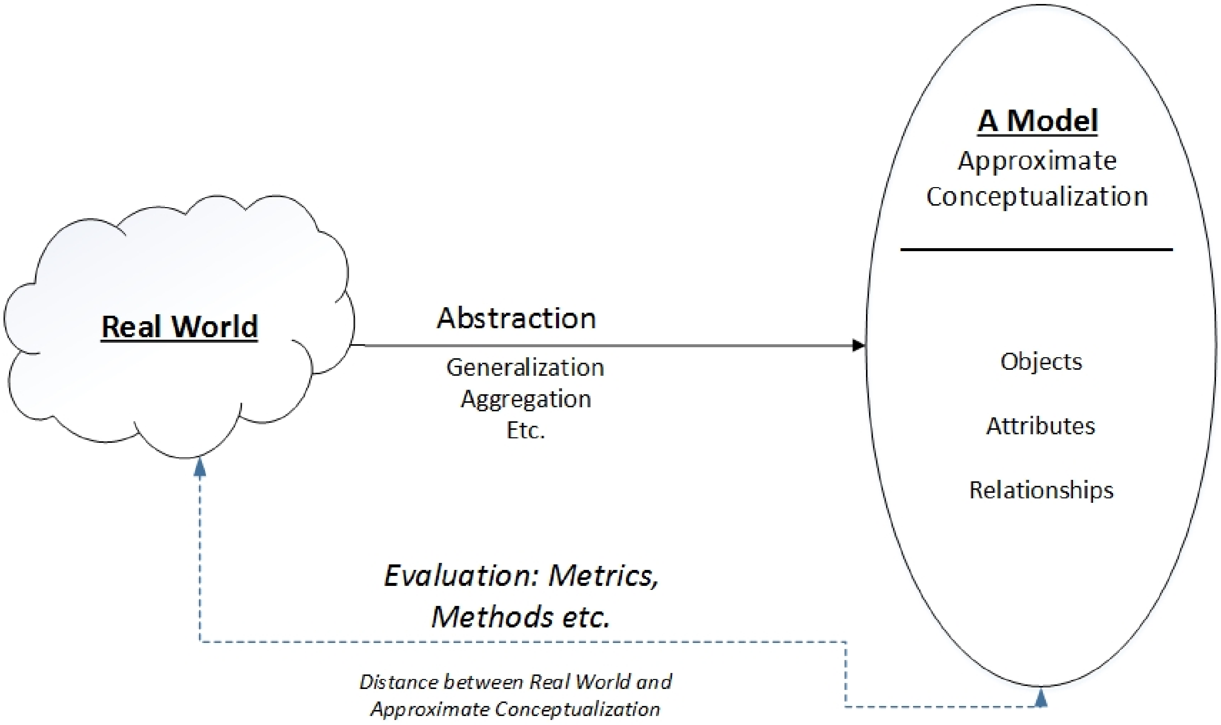
A Model
Approximate Conceptualization
Real World
Abstraction
Generalization
Aggregation
Etc.
Objects
Attributes
Relationships
Evaluation: Metrics, Methods etc.
Distance between Real World and Approximate Conceptualization

Fig. 1. Ontology evaluation in terms of the definition of the notion of ontology: What is an ontology?

Table 1
Categorization of Measures for Ontology Evaluation Adapted from [6,41]

Evaluation Perspective	Metric	Measure
Ontology Correctness	Accuracy	precision: total number correctly found over whole knowledge defined in ontology
		Recall: total correctly found over all knowledge that should be found
		Coverage
	Completeness	Coverage
	Conciseness	
	Consistency	Count: Number of terms with inconsistent meaning
Ontology Quality	Computational efficiency	Size
	Adaptability	Coupling: Number of external classes referenced
		Cohesion: Number of Root (NoR), Number of Leaf (NoL), Average Depth of Inheritance Tree of Leaf Nodes(ADIT-LN)
	Clarity	Number of word senses

exemplar of motivations to ontology evaluation, this section discusses ontology evaluation in the context of the categories within which each ontology evaluation framework can be classified.

Based on a survey of the literature, this work will discuss three ways in which ontology evaluation can be classified with possible interjections on the limitations and strengths of the evaluations. The work of [17] has already introduced a category based on the definition of ontology evaluation. In this view an evaluation framework or methodology can be classified as either being a verification or validation framework.

A second classification of ontology evaluation was introduced by [2], hereby referred to as the layered approach to ontology evaluation. In the layered approach, an ontology is considered to be a fairly complex struc-

Table 2

Definition of the metrics (terms)

Term	Definition
Accuracy	The criteria for determining is the asserted knowledge in the ontology agrees with the expert's knowledge about the domain. A higher accuracy will typically results from correct definitions and descriptions of classes, properties, and individuals. [41]
Adaptability	Measures the ease of use of an ontology in different contexts possibly by allowing it to be extend and specialized monotonically, i.e. without the need to remove axioms
Clarity	Measures how effectively the ontology communicates the intended meaning of the defined terms [41]
Cohesion	From an ontology point of view, cohesion refers to the relatedness of elements in ontologies. It is intended to measure modularity [13,43]. An ontology would have high cohesion if its its classes are strongly related therefore, high cohesion is a desirable property.
Completeness	Measures if the domain of interest is appropriately covered. All questions the ontology should be able to answer can be answered.
Computational efficiency	Relates to the speed at which tools can work with the ontology (e.g. reasoners)
Conciseness	Intended to reflect if the ontology defines irrelevant elements with regards to the domain to be covered or redundant representations of the semantics [41].
Consistency	Describes that the ontology does not include or allow for any contradictions.
Coupling	Reflects the number of classes from imported ontologies that are referenced in the the ontology.
Coverage	Reflects how well the ontology represents the domain it models.

ture and the argument is that it may be better to evaluate each level of the ontology separately than targeting the ontology as a whole.

A third classification referred to mostly by recent research in ontology evaluations [4,22,25] is based on (i) comparison against a gold standard, (ii) application or task-based evaluation, (iii) user-based evaluation, and (iv) data-driven evaluation. In the interest of completeness and brevity, the layers of an ontology along with their applicable approaches is depicted in Table 3. While the Table 3 shows the relation between the second and third types of classification, the first or rather the definition based classification can be found to span the different categories. This is to mean that, for example, within gold standard-based evaluations, there could exist methods and frameworks that are geared towards either verification or validation. Subsequent sections will discuss the different approaches to ontology evaluation.

4.1. Gold standard-based evaluation

This typically compares an ontology against a “gold-standard" which is suitably designed for the domain of discourse [2,10]. This may in fact be an ontology considered to be well-constructed to serve as a reference. Maedche and Staab [26]'s work epitomizes such an approach to ontology evaluation. Their work endeavours to propose measures that estimates the similarity of ontologies at the lexical and conceptual level through the evaluation of the extent to which one ontology is covered by the other. A similar approach is followed by Brank et al. [3]. While they share the same overall goal of comparing the two ontologies (the target and gold-standard), they differ from Maedche and Staab [26]'s work in that, Brank et al. focuses on the arrangement of the class instances and the hierarchical arrangement of the classes.

Within the gold-standard paradigm, Hlomani et al. [21] considered the evaluation of the adaptability of the context of an ontology from the point of view of a checklist. The checklist was specially designed for the evaluation of ontologies based on Gillespie et al. [16]'s knowledge identification framework geared towards the engineering of ontologies.

Ontology design best practices can also be considered to fall within the realm of gold-standard. These would include such work that evaluate ontologies for pitfalls [33,34], and ontology design patterns [12,15, 1]. Akin to reusable problem solving methods, ontology design patterns (ODPs) are considered to be encodings of best practices, that help in solving common

Table 3

An overview of approaches to ontology evaluation as related to the aspects (layers) of an ontology. Source: [2]

Level	**Approach to evaluation**			
	Gold Standard	Application-based	Data-driven	User-based
Lexical, vocabulary, concept, data	X	X	X	X
Hierarchy, taxonomy	X	X	X	X
Other semantic relations	X	X	X	X
Context, application		X		X
Syntactic	X			X
Structure, architecture, design				X

recurring problems and are generally employed in the ontology design lifecycle [1]. In the case of ontology pitfalls, Keet et al. [23] exemplifies such an evaluation and shows the prevalence of the pitfalls in ontologies created by both novice and expert engineers. Their results basically showed no statistical significance between ontologies created by novice engineers as compared to those created by experienced engineers (in the context of pitfalls).

Overall, the gold-standard does offer an avenue to evaluating ontologies. However, it has a major limitation that the gold standard itself needs to be evaluated. Thus far, it is difficult to establish the quality of the gold standard and hence, in the case of discrepancies in the results, it will be difficult to determine the source of the errors. It will be difficult to tell if the gold standard itself is incorrect or the results are in fact flawed.

4.2. Application or task-based evaluation

This would typically involve evaluating how effective an ontology is in the context of an application. Application here may be an actual software program or a use-case scenario. Porzel and Malaka [32] exemplify the software system instance of the task-based ontology evaluation. They define the task of tagging ontological relations, whereby, given a set of concept, the system has to tag concept pairs with appropriate relations. For such an experiment to work, they put forward that all the other components of the system should remain constant except for the ontology-dependent parts. This allows for the effects of the ontology on the performance of the system to be quantified. While this may be practical for the purposes of evaluating a single ontology, it may be challenging to evaluate a number of ontologies in an application area to determine which one is best fitted for the application especially in an automated fashion. Clarke et al. [8], on the other hand, exemplify a more use-case scenario type of task-based evaluation. Their work is specific to the gene ontology. In the gene-set enrichment analysis experiment, they purposed to assess the quality and utility of the gene ontology and its annotations.

There exists two main issues with the task-based approach to ontology evaluation. First, what is applicable in one application context may not be applicable in another. Therefore, it is hard to generalize the results of a task-based evaluation. Second, this is highly suitable to a small (local) set of ontologies and would certainly become unmanageable in an automated setting with a variable number of ontologies.

4.3. User-based evaluation

This typically involves evaluating the ontology through users' experiences. The drive for user-based evaluations is not in assessing semantic validity and consistency of the ontologies per say. It lies, however, in capturing the subjective information about the ontology which Supekar [39] argues is equally as important. This was proposed to be achieved through a metadata ontology that facilitates the capture of two type of information: (i) Source - metadata from the viewpoint of the ontology authors, and (ii) Third-party - metadata from the viewpoint of the users of the ontology, hence, the notion of peer reviews. On a different note, user subjectivity is expressed in Ouyang [25]'s work in the influences they impose on the different metrics applied to an ontology evaluation endeavour. They exert their influence on the results through the application of weights on each metric. For example, given the coverage, coupling and cohesion metrics defined in that work, each metric will be give a weighted value depend on how important the user deems the particular metric to be for their purposes.

The problem with this method is that it is difficult to establish objective standards pertaining to the criteria (metrics) for evaluation. In addition it is also hard to establish who the right users are.

4.4. Data-driven evaluation

This typically involves comparing the ontology(ies) against existing data about the domain the ontology models. This has been done from different perspectives. For example Patel et al. [30] considered it from the point of view of determining if an ontology refers to a particular topic(s). Simply put, an ontology is classified into an array of topics by first, extracting ontological elements (e.g. concept and a relation), and second, feeding these elements to a text classification model (supervised or unsupervised). Spyns et al. [38] attempted to analyze how appropriate an ontology covers a topic of the corpus through the measurement of the notions of precision and recall. Similarly, Brewster et al. [4] investigates how well a given ontology or set of of ontologies fit the domain knowledge. This is done by comparing ontology concepts and relations to text from documents about a specific domain and further refining the results by employing a probabilistic method to find the best ontology for the corpus. Ontology coverage of a domain was also investigated by Ouyang [25] where coverage is considered from the point of view of both the coverage of the concepts and the coverage of the relations. Following on Brewster et al. [4], Hlomani and Stacey [22] also investigated ontologies' coverage of a domain by particularly considering the workflow domain. They analyzed several ontologies' coverage of the workflow domain and further followed a statistical approach to find the best ontology for the workflow domain from a given set of ontologies.

The major limitation of current research within the realm of data-driven ontology evaluation is that domain knowledge is implicitly considered to be constant. This is inconsistent with reality and indeed it is inconsistent with the literature's assertions about the nature of domain knowledge. For example, Nonaka [27] asserts that domain knowledge is dynamic. Changes in ontologies has been partially attributed to changes in the domain knowledge. In some circles, ontological representation of the domain has been deemed to be biased towards their temporal, environmental, and spatial setting [2,4]. By extension, the postulation is that domain knowledge would vary (change) over these dimensions as well. Hence, a data-driven ontology evaluation should be directed to succinctly incorporate these salient dimensions of domain knowledge in an ontology evaluation effort with the view of proving their unexplored influence on evaluation measures.

5. Overall limitations of current approaches to ontology evaluation

Despite the efforts to address this need for ontology evaluation such as the approaches discussed in Sections 4.1 through 4.3, near ideal solutions to the ontology evaluation problem are yet to be introduced. While each of the approaches to ontology evaluation has its own limitations, this section discusses *subjectivity* as a common major limitation to current research in ontology evaluation. This survey demarcates this discussion into: (i) subjectivity in the selection of the criteria for evaluation and (ii) subjectivity in the thresholds for each criterion, and (iii) influences of subjectivity on the results of ontology evaluation.

5.1. Subjectivity in the criteria for evaluation

Ontology evaluation can be regarded over several different decision criteria. These criteria can be seen as the desiderata for the evaluation [41,6]. The first level of difficulty has been in deciding the relevant criteria for a given evaluation task. It has largely been the sole responsibility of the evaluator to determine the elements of quality to evaluate [41]. This brings about the issue of subjectivity in deciding which criteria makes the desiderata. This has largely been the issue with most of the approaches to ontology evaluation since with most, there has to be a criterion that decides the overall quality or correctness of the ontology.

To address this issue, two main approaches have been proposed in literature: (i) induction - empirical testing of ontologies to identify desirable properties of the ontologies in the context of an application, and (ii) deduction - deriving the most suitable properties of the ontologies based on some form of theory (e.g. based on software engineering as exemplified in Section 3.1). The advantages of these coincidentally seem to be the disadvantage of the other. For example, inductive approaches are guaranteed to be applicable for at least one context, but their results cannot be generalized to other contexts. Deductive approaches on the other hand, can be generalized to other contexts, but are not guaranteed to be applicable for any specific context. In addition, for deductive approaches, the first level of challenge is in determining the correct theory to base the deduction on. This then spirals back to the problem of subjectivity where the evaluator has to sift through a plethora of theories in order to justify selection.

5.1.1. Inductive approach to criteria selection

Preliminary work in the TOVE ontology by Fox et al. [11] epitomizes the inductive approach to ontology criteria selection. They proposed eight criteria for evaluating ontologies which includes: generality, competence, perspicuity, transformability, extensibility, granularity, scalability, and minimality. Of these eight, they considered competence to be the most important and explored their ontology in the context of answers to competency questions.

Recent work by Ning et al. [24] discusses ontology summarization. Their work is a perfect exemplar of the inductive approach to ontology feature selection for the purposes of evaluation. In this work, the importance of four ontology features is evaluated and ranked using Kendall's tau coefficient [36,35] . The ontology features under investigation were: density, name simplicity, popularity, and reference.

Žontar and Heričko [42] present a selection approach that might be seen to span the two categories (induction and deduction). Their work is inductive in that a rigorous analysis of possible candidate object-oriented software metrics is conducted in the form of a feasibility study that leads to a narrow list of relevant criteria. It can also be deductive in that, while, object-oriented software metrics is not a theory per say, it is however, well established and acknowledged as a measure of software quality and used here as a basis for the inclusion into the desiderata. A list of eighteen software metrics was then proposed.

The premise of Sicilia [37]'s work in that, ontologies are very heterogeneous in their structure and organization, guiding objectives and their level of formality and hence, suggest the need to perform exploratory work on ontology metrics to gain insight on the relevance of each metric. They empirically (following a statistical method) investigate eleven ontology metrics.

5.1.2. Deductive approaches to criteria selection

The deductive approach to ontology criteria selection is perhaps the most used of the two (induction and deduction). Chidamber and Kemerer [7]'s objected-oriented design (OO-D) metrics are such an example. Having been based on Bunge [5]'s formal ontology and measurement theory, they have been the common metrics used in ontology evaluation. Typical ontology evaluation works following the OO-D metrics include those that explore the notions of coupling and cohesion [10,25,29,43]. For example, in software engineering, cohesion metrics are mostly used to measure modularity. In the same light, metrics similar to software cohesion metrics have been defined to measure the relatedness of the classes (and/ or other elements) of an ontology [13]. These would include such measures as [13,43]: (i) **Number of Root Classes** (NoR) - Number of root classes explicitly defined in an ontology. This would mean a class with no explicit super class. (ii) **Number of Leaf Classes** (NoL) - Number of Leaf classes explicitly defined in an ontology. This would mean a class with no explicit subclass. (iii) **Average Depth of Inheritance Tree of Leaf Nodes** (ADIT-LN) - The sum of depths of all paths over the total number of paths. Depth in this case means the total number of nodes preceding the leaf node from the root. While the total number of paths is all distinct paths from each root node to the leaf node if there exists an inheritance path from the root to the leaf node. It is then believed that strong cohesion is desirable.

Semiotic theory has motivated the proposal of some metrics to assess the quality of ontologies. Examples of these includes the works of Burton-Jones et al., Gangemi et al., and Dividino [6,14,9]. These would include a metric suite to measure: syntactic quality, semantic quality, pragmatic quality, and social quality.

5.2. Subjectivity in thresholds

The issue of thresholds for ontology evaluation criteria (both quality and correctness) has been highlighted by Vrandecic [41]. He puts forward that the goal for ontology evaluation should not be to perform well for all criteria and also suggests that some criteria may even be contradictory. This then defaults to the evaluator to make a decision on the results of the evaluation over the score of each criterion. This leads to subjectivity in deciding the *optimal* thresholds for each criterion. For example, if a number of ontologies were to be evaluated for a specific application, it becomes the responsibility of the evaluator to answer questions like, *"Based on the evaluation criteria, when is Ontology A better than Ontology B?*.

5.3. Influences of subjectivity on the overall value of the measures/metrics

The default setting of good science is to exclude subjectivity from a scientific undertaking (e.g experiment) [27]. This has been typical of ontology evaluation. However, as has been discussed in Sections 5.1 and 5.2, humans are the objects (typically as actors) of research in most ontology evaluation experiments. The research itself can hence, not be free of subjectivity. This expresses bias from the point of view of the evaluator. There exists another form of bias, the kind that is inherent in the design of the ontologies. An ontology (a model of domain knowledge) represents the domain in the context of the time, place, and cultural environment in which it was created as well as the modellers perception of the domain [2,4].

The problem lies in the unexplored potential influence of this subjectivity in the evaluation results. If one takes a data-driven approach to ontology evaluation for example, it would be interesting to see how the evaluation results spreads over each dimension of the domain knowledge (i.e. temporal, categorical, etc.). This is based on equating subjectivity/bias to the different dimensions of domain knowledge. To give a concrete example, let us take Brewster et al. [4]'s results. These are expressed as a vector representation of the similarity score of each ontology showing how closely each ontology represents the domain corpus. This offers a somewhat one dimensional summarization of this score (coverage) where one ontology will be picked ahead of the others based on a high score. It, however, leaves unexplored how this score changes over the years (temporal) for example. This could reveal very important information such as the relevance of the ontology, meaning the ontology might be aging and needs to be updated as opposed to a rival ontology. The results of Ouyang et al. [25] are a perfect exemplar of this need. They reveal that the results of their coverage showed a correlation between the corpus used and the resultant coverage. This revelation is consistent with the notion of dynamic domain knowledge. In fact, a changing domain knowledge has been attributed to the reasons for changes to the ontologies themselves [27]. This offers an avenue to explore and accounts for bias and its influence on the evaluation results.

Thus far, to the best of our knowledge, no research in ontology evaluation has been undertaken to account for subjectivity. This has not been especially done to measure subjectivity in the context of a scale as opposed to binary (yes- it is subjective, or no - its not subjective). Hence, this provides a means to account for the influences of bias (subjectivity) on the individual metrics of evaluation that are being measured.

6. Conclusion

It is fitting to summarize this paper in terms of three important concepts: approach, methods and metrics. This summary details how these concepts fit together and narrow it down to the crust of this survey. Ontology evaluation was defined in the context of two important perspectives: ontology quality, and ontology correctness. Correctness is hereby defined in the context of the measure of the distance between the text about a domain (real world) and the model (formalized by ontologies).

The paper discusses methods, metrics, and approaches to ontology evaluation. Approaches offer general philosophies of ontology evaluation, while methods offer a means to measure the metrics that help in deciding both the quality and correctness of a given set of ontologies. Within an approach, several methods can be used.

This work has also decomposed literature on ontology evaluation measures into the two perspectives discussed earlier: ontology quality, and ontology correctness.

The survey also explores the notion of subjectivity. This is an important aspect of ontology evaluation that has not been considered in current research in ontology evaluation. Particular interest is on the influence of this subjectivity/ bias on the overall measures of ontology correctness which can be extended to other aspects of ontology evaluation (e.g. ontology quality).

References

[1] R. Alm, S. Kiehl, B. Lantow, and K. Sandkuhl. Applicability of quality metrics for ontologies on ontology design patterns. In *Proceedings of the 5th International Conference on Knowledge Engineering and Ontology Development*, Vilamoura, Portugal, 2013.

[2] J. Brank, M. Grobelnik, and D. Mladenić. A survey of ontology evaluation techniques. In *Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005)*, pages 166–170, 2005.

[3] J. Brank, D. Madenic, and M. Groblenik. Gold standard based ontology evaluation using instance assignment. In *Proceedings of the 4th Workshop on Evaluating Ontologies for the Web (EON2006)*, Edinburgh, Scotland, May 2006.

[4] C. Brewster, H. Alani, S. Dasmahapatra, and Y. Wilks. Data-driven ontology evaluation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 2004.

[5] M. Bunge. *Treatise on Basic Philosophy: Volume 3: Ontology I: The Furniture of the World*. Springer, 1 edition, June 1977.

[6] A. Burton-Jones, C. V. Storey, V. Sugumaran, and P. Ahluwalia. A semiotic metrics suite for assessing the quality of ontologies. *Data & Knowledge Engineering*, 55(1):84 – 102, 2005.

[7] S. R. Chidamber and C. F. Kemerer. A metrics suite for object oriented design. *IEEE Transactions on Software Engineering*, 20(6):476–493, June 1994.

[8] E. Clarke, S. Loguercio, B. Good, and A. Su. A task-based approach for gene ontology evaluation. *Journal of Biomedical Semantics*, 4(Suppl 1):S4, 2013.

[9] R. Dividino, M. Romanelli, and D. Sonntag. Semiotic-based Ontology Evaluation Tool (S-OntoEval). In E. L. R. A. (ELRA), editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008.

[10] F. Ensan and W. Du. A semantic metrics suite for evaluating modular ontologies. *Information Systems*, 38(5):745 – 770, 2013.

[11] M. S. Fox, M. Barbuceanu, M. Gruninger, and J. Lin. An organization ontology for enterprise modelling. In *Modeling, In: International Conference on Enterprise Integration Modelling Technology 97*. Springer, 1997.

[12] A. Gangemi. Ontology design patterns for semantic web content. In *Proceedings of the Fourth International Semantic Web Conference*, pages 262–276. Springer, 2005.

[13] A. Gangemi, C. Catenacci, M. Ciaramita, and J. Lehmann. Ontology evaluation and validation: an integrated formal model for the quality diagnostic task. Technical report, Laboratory of Applied Ontologies – CNR, Rome, Italy, 2005.

[14] A. Gangemi, C. Catenacci, M. Ciaramita, and J. Lehmann. Modelling ontology evaluation and validation. In *Proceedings of the 3rd European Semantic Web Conference (ESWC2006), number 4011 in LNCS, Budva*. Springer, 2006.

[15] A. Gangemi and V. Presutti. Ontology design patterns. In S. Staab and R. Studer, editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 221–243. Springer Berlin Heidelberg, 2009.

[16] M. G. Gillespie, H. Hlomani, D. Kotowski, and D. A. Stacey. A knowledge identification framework for the engineering of ontologies in system composition processes. In *IRI*, pages 77–82. IEEE Systems, Man, and Cybernetics Society, 2011.

[17] A. Gómez-Pérez. Ontology Evaluation. In S. Staab and R. Studer, editors, *Handbook on Ontologies*, chapter 13, pages 251–274. Springer Berlin Heidelberg, Berlin, Heidelberg, first edition, 2004.

[18] T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. In *International Journal of Human-Computer Studies*, pages 907–928. Kluwer Academic Publishers, 1993.

[19] M. Gruninger and M. S. Fox. Methodology for the design and evaluation of ontologies. In *International Joint Conference on Artificial Inteligence (IJCAI95), Workshop on Basic Ontological Issues in Knowledge Sharing*, 1995.

[20] N. Guarino. Toward a formal evaluation of ontology quality. *IEEE intelligent Systems*, 19(4):78–79, 2004.

[21] H. Hlomani, M. G. Gillespie, D. Kotowski, and D. A. Stacey. Utilizing a compositional knowledge framework for ontology evaluation: A case study on BioSTORM. In *Conference on Knowledge EngineeringÊand Ontology Development (KEOD*, Paris, France, 2011.

[22] H. Hlomani and A. D. Stacey. Contributing evidence to data-driven ontology evaluation: Workflow ontologies perspective. In *Proceedings of the 5th International Conference on Knowledge Engineering and Ontology Development*, Vilamoura, Portugal, 2013.

[23] M. C. Keet, C. M. Suńarez-Figueroa, and M. Poveda-Villalńon. The current landscape of pitfalls in ontologies. In *Proceedings of the 5th International Conference on Knowledge Engineering and Ontology Development*, Vilamoura, Portugal, 2013.

[24] N. Li, E. Motta, and M. d'Aquin. Ontology summarization: an analysis and an evaluation. In *The International Workshop on Evaluation of Semantic Technologies (IWEST 2010)*, Shanghai, China, 2010.

[25] O. Liubo, Z. Beiji, Q. Miaoxing, and Z. Chengming. A method of ontology evaluation based on coverage, cohesion and coupling. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on*, volume 4, pages 2451 –2455, july 2011.

[26] E. Maedche and S. Staab. Measuring similarity between ontologies. In *in Proceedings of the European Conference on Knowledge Acquisition and Management (EKAW*, pages 251–263. Springer, 2002.

[27] I. Nonaka and R. Toyama. The theory of the knowledge-creating firm: subjectivity, objectivity and synthesis. *Industrial and Corporate Change*, 14(3):419–436, June 2005.

[28] L. Obrst, W. Ceusters, I. Mani, S. Ray, and B. Smith. The evaluation of ontologies. In C. J. Baker and K.-H. Cheung, editors, *Revolutionizing Knowledge Discovery in the Life Sciences*, chapter 7, pages 139–158. Springer, 2007.

[29] S. Oh, H. Y. Yeom, and J. Ahn. Cohesion and coupling metrics for ontology modules. *Inf. Technol. and Management*, 12(2):81–96, June 2011.

[30] C. Patel, K. Supekar, Y. Lee, and E. K. Park. Ontokhoj: A semantic web portal for ontology searching, ranking and classification. In *In Proc. 5th ACM Int. Workshop on Web Information and Data Management*, pages 58–61, 2003.

[31] H. S. Pinto and J. P. Martins. Reusing ontologies. In *In AAAI 2000 Spring Symposium on Bringing Knowledge to Business Processes*, pages 77–84. AAAI Press, 2000.

[32] R. Porzel and R. Malaka. A task-based approach for ontology evaluation. In *Proc. of ECAI 2004 Workshop on Ontology Learning and Population*, Valencia, Spain, August 2004.

[33] M. Poveda, M. C. Suarez-Figueroa, and A. Gomez-Perez. Common pitfalls in ontology development. In C. . S. Papers., editor, *Current Topics in Artficial Intelligence, CAEPIA 2009 Selected Papers*. Springer-Verlag Berlin, 2010.

[34] M. Poveda-VillalŮn, M. del Carmen SuĞrez-Figueroa, and A. GŮmez-PŐrez. Validating ontologies with oops! In A. ten Teije, J. VŽlker, S. Handschuh, H. Stuckenschmidt, M. d'Aquin, A. Nikolov, N. Aussenac-Gilles, and N. Hernandez, editors, *EKAW*, volume 7603 of *Lecture Notes in Computer Science*, pages 267–281. Springer, 2012.

[35] D. Sheskin. *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, Fla. CRC Press, 1997.

[36] D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall, 2003.

[37] M. A. Sicilia, D. Rodríguez, E. García-Barriocanal, and S. Sánchez-Alonso. Empirical findings on ontology metrics. *Expert Systems with Applications*, 39(8):6706 – 6711, 2012.

[38] P. Spyns. EvaLexon: Assessing triples mined from texts. Technical Report 09, Star Lab, Brussels, Belgium, 2005.

[39] K. Supekar. A peer-review approach for ontology evaluation. In *8th Int. Protégé Conference, Madrid, Spain*, July 2005.

[40] D. Vrandecic. Ontology Evaluation. In S. Staab and R. Studer, editors, *Handbook on Ontologies*, pages 293–313. Springer Berlin Heidelberg, Berlin, Heidelberg, 2nd edition, 2009.

[41] D. Vrandecic. *Ontology Evaluation*. PhD thesis, Karlsruhe Institute of Technology, Karlsruhe, Germany, 2010.

[42] R. Žontar and M. Heričko. Adoption of object-oriented software metrics for ontology evaluation. In *Proceedings of the Fifth Balkan Conference in Informatics*, BCI '12, pages 298–301, New York, NY, USA, 2012. ACM.

[43] H. Yao, A. M. Orme, and L. Etzkorn. Cohesion Metrics for Ontology Design and Application. *Journal of Computer Science*, 1(1):107–113, 2005.