Metrics and methods for comparative ontology evaluation

Amanda Hicks
Ph.D, Philosophy, State University of New York at Buffalo
Assistant Professor, Health Outcomes & Policy Faculty,
Institute for Child Health Policy,
University of Florida, Gainesville, Florida, Estados Unidos.
https://com-hop.sites.medinfo.ufl.edu/files/2015/06/CV-September2016.pdf
E-mail: aehicks@ufl.edu

Submetido em: 10/07/2017. Aprovado em: 05/10/2017. Publicado em: 28/12/2017.

ABSTRACT

While progress has been made toward describing the need for ontology evaluation and offering proposals concerning what properties to measure and how, work remains to develop ontology evaluation as a rigorous discipline. Ontologies as information artifacts have a variety of aspects that can inform their evaluation, both in terms of what is evaluated and the metrics used. Ontology evaluation as a discipline requires (1) having a systematic account of the different aspects of ontologies and the properties relevant to those aspects, (2) critically developing methods for examining those properties, (3) developing comparative metrics that allow ontology engineers to compare the effects of various modeling choices and allow users to compare the merits of existing ontologies, and (4) charting possible pitfalls of evaluation. This paper considers various properties of ontologies that have been proposed and organizes these properties according to different aspects of ontologies. To begin bringing previous work together and to illustrate where pitfalls and potential solutions might enter into a rigorous evaluation, I offer a more in depth (though still partial) analysis of evaluating the correctness of ontologies. I conclude with a discussion of next steps in systematizing ontology evaluation.

Keywords: Ontologies. Ontologies Evaluation.

Métricas e métodos para a avaliação comparativa de ontologias

RESUMO

Progresso tem sido feito no sentido de descrever a necessidade de avaliação de ontologias, bem como nas propostas para mensurá-las e como mensurá-las, mas a literatura ainda carece de trabalhos sobre a avaliação de ontologias como uma disciplina rigorosa. Ontologias como artefatos de informação apresentam uma variedade de aspectos que podem fornecer subsídios para avaliação, tanto em termos do que avaliar como em termos das métricas adotadas. A avaliação de ontologias é uma disciplina que requer (1) consideração sistemática dos diferentes aspectos das ontologias e das propriedades que são relevantes para tais aspectos; (2) desenvolvimento crítico de métodos para examinar as propriedades mencionadas; (3) desenvolvimento de métricas que permitam aos engenheiros de ontologias comparar os efeitos das diversas decisões de modelagem possíveis, bem como a possibilidade do usuário comparar os méritos de ontologias existentes; (4) identificação das principais práticas que levam a erros de avaliação. O presente artigo considera as várias propriedades das ontologias que têm sido propostas e organiza tais propriedades de acordo com diferentes aspectos ontológicos. Inicia-se apresentando trabalhos relacionados anteriores, ilustrando as práticas que levam a erros, bem como soluções potenciais, para depois oferecer uma análise detalhada (ainda que parcial) da avaliação da correção de ontologias. Conclui-se com uma discussão sobre os próximos passos necessários para sistematizar a avaliação de ontologias.

***Palavras-chave**: Ontologias. Avaliação de ontologias.*

Métricas y métodos para la evaluación comparativa de ontologías

RESUMEN

Progreso ha sido hecho en el sentido de describir la necesidad de evaluación de ontologías, como también en las propuestas para mensurarlas y como mensurarlas, pero la literatura aún carece de trabajos sobre la evaluación de ontologías como una disciplina rigorosa. Ontologías como artefactos de información presentan una variedad de aspectos que pueden proveer subsidios para la evaluación, tanto en términos de que evaluar como en términos de las métricas adoptadas. La evaluación de ontologías es una disciplina que requiere (1) consideración sistemática de los diferentes aspectos de las ontologías y de las propiedades que son relevantes para tales aspectos; (2) desarrollo crítico de métodos para examinar las propiedades mencionadas; (3) desarrollo de métricas que permitan a los ingenieros de ontologías comparar los efectos de las diversas decisiones de modelaje posibles, como también la posibilidad de que el usuario compare los méritos de ontologías existentes; (4) identificación de las principales prácticas que llevan a errores de evaluación. El presente artículo considera las varias propiedades de las ontologías que han sido propuestas y organiza tales propiedades de acuerdo con diferentes aspectos ontológicos. Se inicia presentando trabajos relacionados anteriores, ilustrando las prácticas que llevan a errores, como también soluciones potenciales, para ofrecer después un análisis detallado (aún que parcial) de la evaluación de la corrección de ontologías. Se concluye con una discusión sobre los próximos pasos necesarios para sistematizar la evaluación de ontologías.

***Palabras-clave**: Ontologías. Evaluación de ontologías.*

INTRODUCTION

The call for a rigorous discipline of ontology evaluation is not new (GANGEMI et al., 2005; GANGEMI et al., 2006; GOMEZ-PEREZ, 2001; GUARINO, 2004; HOEHNDORF et al., 2007). While some progress has been made toward describing the need for ontology evaluation and offering proposals concerning what properties to measure and how, we still have work remains to develop ontology evaluation as a rigorous discipline.

Ontologies as information artifacts have a variety of aspects that can inform their evaluation, both in terms of what is evaluated and the metrics used. Ontology evaluation as a discipline requires (1) having a systematic account of the different aspects of ontologies and the properties relevant to those aspects, (2) critically developing methods for examining those properties, (3) developing comparative metrics that allow ontology engineers to compare the effects of various modeling choices and allow users to compare the merits of existing ontologies, and (4) charting possible pitfalls of evaluation.

The next section of this paper considers various properties of ontologies that have been proposed for evaluation – namely ontologies as representational artifacts, logical theories, mathematical objects, parts of information systems, and community resources for reuse – and organizes these properties according to different aspects of ontologies. To begin bringing previous work together and to illustrate where pitfalls and potential solutions might enter into a rigorous evaluation, I offer a more in depth (though still partial) analysis of evaluating the correctness of ontologies. I conclude with a discussion of next steps in systematizing ontology evaluation.

THE SCOPE OF ONTOLOGIES UNDER CONSIDERATION AND SCOPE OF EVALUATION

While there is disagreement about what should be properly considered an ontology, an assumption of this paper is that, to the extent that this disagreement is fueled by normative considerations of what an ontology *should* be, ontology evaluation as a discipline should take an ecumenical view of the kinds of artifacts it evaluates to further inform the normative debate. Nevertheless, some constraints and a working definition of ontologies need to be offered as a starting point.

The kinds of artifacts under consideration in this paper have a human readable component, a machine readable component, and aim to formally represent something (whether material reality, concepts, linguistic intuitions, or something else). Ontologies are often created to be shared and reused by a community of informatitions, so a discipline of ontology evaluation will need to consider ontology reuse as well. However, for the purpose of this paper we do not consider this to be a necessary condition for an ontology since many of the considerations of evaluating an ontology would apply even to idiosyncratic, proprietary artifacts meant for use by a small number of users.

In this paper, the term 'ontology' does not refer exclusively to realist ontologies. This is not an assertion about whether conceptualist ontologies are as good as realist ontologies, nor is it a statement about whether wordnets can *function* as ontologies. Instead it is an agnostic stance intended to avoid begging the question about what makes a good ontology. Ultimately the discipline of ontology evaluation ought be able to meaningfully compare different types of resources, so arguments for adopting one or the other will not based solely on appeals to common sense or conjecture, but rather will be grounded in some evaluative data to support the claim that one is preferable to another. We, therefore, adopt the definition offered in Neuhaus et al. (2013, p.):

> Ontologies are human-intelligible and machine-interpretable representations of some portions and aspects of a domain, where the domain can be portions of the physical world or ways in which human agents mentally represent the physical world. That being said, we do believe that one should always be aware of which kind of ontology one is working with; a representation of concepts should not be mistaken for a representation of material reality.

Hoehndorf et al. (2013) argue that ontologies should always be evaluated as a part of an information system rather than as an ontology alone. I agree that a robust discipline of ontology evaluation must include evaluating ontologies as parts of information systems.

To take this a step further, such evaluation ideally ought to compare the performance of two or more ontologies in the same information system for the same task. However, I do not believe that ontologies cannot be evaluated according to their intrinsic properties. On the contrary, a complete science of ontology will be able to describe how different intrinsic properties of an ontology, including philosophical assumptions, affect the performance of information systems for particular tasks. Such an achievement requires analyzing and evaluating properties of ontologies, measuring their performance, and then synthesizing the results of both steps.

In designing ontology evaluation studies, we ought to have clear answers to the following:

1. What aspect of the ontology is under consideration?
2. What properties of the ontology, whether intrinsic or extrinsic, are being investigated?
3. What method will accurately and reliably capture this property?
4. What metrics can quantify that property, either directly or by proxy?
5. How can the metric be designed to yield accurate and reliable comparisons of evaluation results?

The next section provides a brief overview of 1 and 2 together.

VARIOUS ASPECTS AND PROPERTIES OF ONTOLOGIES

Ontologies are complex artifacts that can be considered from a variety of view points and disciplines. Accordingly, various approaches and proposals for ontology evaluation focus on different aspects of ontologies including ontologies as representational artifacts, as logical theories, as mathematical graphs, as parts of information systems, and as community resources for reuse.

Each of these aspects are discussed and used to organize properties of ontologies that have been proposed for evaluation.

As representational artifacts an ontology can be evaluated for the truth and accuracy of their representations, whether the allowable formal interpretations include all and only the intended interpretations, whether domain experts approve of the content, and breadth and granularity of domain coverage.

The logical system used in the ontology can be evaluated for validity, soundness, completeness, decidability (OBRST et al., 2007), syntactic lawfulness and richness (AMITH; TAO, 2017). The ontology as an axiomatic theory can be evaluated for logical properties such as consistency.

Ontologies that are graphs are mathematical objects, so their topological properties can be measured. Some of the topological metrics can be used as proxies for desirable attributes of ontologies, such as whether there is a single root node, multiple inheritance, depth, and fanned-outness (WALOSZEK, 2012). Waloszek (2012) offers an in depth discussion of this.

As parts of information systems ontologies can be evaluated against the requirements of the information system (NEUHAUS et al., 2013), use cases, their data sources (OBRST et al., 2007), and competency questions (GRÜNINGER; FOX, 1995). Ultimately, ontologies ought to be compared in terms of how they affect the overall performance of the system in a manner that isolates the contributions of various ontologies in the information system. This requires developing information systems in a manner that allows swapping out ontologies to compare results on the same task.

Ontologies as a community resource, including as controlled terminologies, can be evaluated according to their suitability for use and reuse. As such their ranking within a community can be measured and evaluated, including how many other ontologies reuse the current one (AMITH; TAO, 2017), how frequently the ontology is used in information systems, and how successful it is for integrating data.

As resources for use and reuse by humans, ontologies can be evaluated for the intelligibility of terms and definitions in the ontology (AMITH; TAO, 2017; OBRST et al., 2007), including what Amith and Tao (2017) calls "clarity" – the ratio of terms in the ontology that are ambiguous with respect to some source lexicon such as WordNet (MILLER, 1995) – "interpretability" or the ratio of terms in the ontology that have at least one word sense in some source lexicon (AMITH; TAO, 2017), the number duplicate terms in the ontology itself (AMITH; TAO, 2017), and the degree to which there is community consensus that the ontology contains the relevant classes with correct definitions (OBRST et al., 2007). The O^2 framework proposes usability-measures that quantify the number of annotations on the ontology according to a typology of annotations that promote reuse, e.g., recognition annotations, which describe an ontology's structure and purpose (GANGEMI et al., 2005). This latter proposal is likely to play a large role in ensuring that ontologies conform to FAIR principles, which aim to make digital artifacts findable, accessible, interpretable, and reusable (WILKINSON et al., 2016).

In what follows, we focus on ontologies as representational artifacts and walk through considerations of measuring the correctness of the representation. We discuss methodological concerns, pitfalls, and solutions that ought to be addressed for a fully-fledge discipline of ontology evaluation.

THE CASE OF CORRECTNESS

When considering ontologies as representational artifacts, the natural question arises of whether they represent their subject matters well. A precise answer to this question requires a clearly defined notion of a good representation and some way of determining and measuring the quality of the representation.

Previous discussions of ontology evaluation have dealt with the quality of the representation in terms of fidelity (NEUHAUS et al., 2013), accuracy (AMITH; TAO, 2017), functional measures (GANGEMI et al., 2006), precision and recall of intended models (GUARINO, 2004), expert agreement (GANGEMI et al., 2005), and coverage (GANGEMI et al., 2005; ROSPOCHER et al., 2012; ZHU et al., 2009). What follows is a guided tour of some considerations that go into evaluating and comparatively measuring the correctness of ontologies.

Neuhaus et al. (2013) describe fidelity as a property of ontologies to be evaluated. Fidelity includes correctness of the statements in the ontology, both human readable and machine readable, but fidelity is also intended to capture whether the axioms and documentation are in agreement with each other. This latter criterion is not strictly about the quality of the representational aspect of an ontology, but rather about its usability and accessibility to humans. We certainly consider this important, but an orthogonal issue to the one described here, so in what follows the word 'correctness' will be used to describe the accuracy of the statements in the ontology and is distinguished from 'fidelity' which also describes agreement of the ontology with documentation.

The questions at hand for evaluating the correctness of the ontology include whether the human readable statements (e.g., definitions, examples, etc.) and axioms are correct. However, it is important to note that the standard of correctness is different for realist, conceptualist, and linguistic ontologies. For a realist ontology, this amounts to asking whether the statements are true statements about the world. For a conceptualist ontology, correctness means that representation corresponds to the conceptualization of a particular person or group, and for wordnets, correctness means whether the statements are readily agreed upon by native speakers of the language (MILLER; FELLBAUM, 1991). Evaluating the Descriptive Ontology for Cognitive and Linguistic Entities (MASOLO et al., 2002) according to a realist conception of truth would be as inappropriate as evaluating the Basic Formal Ontology (ARP et al., 2015) according to whether it conforms to a particular person's conception of the world. Evaluating the correctness of an ontology requires first understanding what the ontology is supposed to be a representation of.

SOME PREVIOUS SUGGESTIONS

How should the statements of an ontology be evaluated for correctness? And how can the result be quantified for comparative evaluation? It is widely acknowledged that evaluating ontologies for correctness requires domain experts (NEUHAUS et al., 2013; AMITH; TAO, 2017), which is often approached manually and is therefore both labor intensive and expensive. Neuhaus et al. (2013) suggests some automated methods that include checking for logical consistency, checking whether allowable formal models match intended models, and comparing the structure of the source ontology to some target ontology. Gordon et al. (2013) describe a semi-automatic method for constructing questions for expert review of the ontology discussed further below.

We consider checking for logical consistency is a task that evaluates ontologies as logical theories rather than as representations. For realist ontologies, logical consistency also evaluates the representation since it is a common assumption going back to Aristotle that reality is consistent, and so contradictions must be false. Since human agents often hold contradictory beliefs, a conceptualist ontology can be correct and logically inconsistent.

While allowable models can be automatically generated, it is not clear how these can be tested against intended models in an automated or even semi-automated way since a set of intended models needs to be constructed somehow and Neuhaus et al. (2013) do not describe how to construct that set. If we already had an ontology that was known to define all and only the intended models, the problems of evaluation and of creating a good ontology would already be solved.

Comparing the structure of two ontologies is an interesting proposal, but it is not clear what its value is. For the comparison to be meaningful, we must have reason to believe that the target ontology is of sufficient quality to function as a gold-standard. If this is the case, it is not clear why one would not simply adopt this ontology rather than create a new one. Also, comparing the structure of ontologies presumably relies on at least some ontology mapping, but this itself is an area that is in need of methods for evaluation. Finally, if the ontologies have different ontological commitments, their correct representations may not be sufficiently isomorphic to produce a meaningful comparison. Without an authoritative digital source of knowledge that can be compared to an ontology (which is to say, with an ontology already known to be correct), human experts are indispensable.

MANUAL EVALUATION

Studies for manual evaluation by domain experts[1] need to have explicit methods and criteria and, wherever possible, metrics developed. For example, when designing an evaluation study, the criteria for correctness need to be carefully considered and clearly explicated to the domain experts. Consider the sentence, "Timolol is an ingredient of Timolol opthalmic solution." If simply presented with this sentence and asked whether it is true, most cardiologists wouldassert that it is true since, by definition, all Timolol ophthalmic solution has Timolol as an ingredient. If an ontology aims to represent linguistic intuitions (or perhaps conceptualizations), this is fine. However, if this is intended as a natural language expression of a description logic axiom in a realist ontology, it should be read as "All Timolol is an ingredient of some Timolol opthalmic solution", which is false. Only some Timolol is an ingredient of Timolol opthalmic solution. This distinction would impact the evaluation results of RxNorm since it contains the triple (LIU et al., 2005):

Timolol ingredient_of 'Timolol opthalmic solution'.

The Bacterial Clinical Infectious Disease Ontology (BCIDO) is an example of a biomedical ontology that was evaluated for correctness by domain experts. Gordon et al. (2013) give a brief report of ontology evaluation during the development phase of the BICDO in which a knowledge elicitation technique laddering was used obtain information from infectious disease fellows and the answers were compared to statements in the ontology. For example, the domain experts were asked, "Can you tell me some bacteria that causes acute meningitis?" While the article is sparse on the details of the evaluation, considering possible scenarios in light of this question provides an opportunity to think through some important methodological details. Suppose a domain expert answers the question with "Streptococcus pneumonia." It is now the task of the ontology engineer to look for an axiom in the ontology that expresses the relation between acute meningitis and Streptococcus pneumonia elicited. Suppose the following class description were in the ontology:

'Streptococcus pneumoniae' causes 'acute meningitis'[2]

This would entail that every Streptococcus pneumoniae causes acute meningitis, which is false. Most people who are infected with Streptococcus pneumoniae do not develop acute meningitis. Not all Streptococcus pneumoniae infect an organism, and finally, a single bacterium is not sufficient to cause meningitis. So the domain expert's ontologically naive statement 'Streptococcus pneumoniae causes acute meningitis' needs to be appropriately translated to a more accurate statements (or sets of statements) in the ontology such as

'colony of Streptococcus' bearer_of 'infectious agent causing acute meningitis disposition'.

Ben Abacha et al. (2016) report a semi-automatic approach to evaluating the correctness of statements in an ontology by converting formal statements to natural language statements with templates.

[1] (OBRST et al., 2007) describe human evaluation as an approach.

[2] This axiom is not actually in BCIDO. Instead this class description is one disjunct in disjunctive class description that contains seventeen total disjunctions of the form (causes x). We use this example for simplicity; however, the same critique applies to the actual axioms in BCIDO.

For example, the template that corresponds to the is_a relation is as follows:

Is CLASS a type of CLASS?

Notice that this template can lead to many false positives. For example, the question "Is a cat a type of pet?" is likely to elicit the answer, yes. In which case, an is_a link between cat and pet in a realist ontology would not be detected as false. This approach, however, is suitable for a wordnet that is meant to represent linguistic intuitions. These errors can be avoided by presenting natural language statements that involve an explicit quantifier to the domain specialist whenever the semantics of the logical language of the ontology use such quantifiers.

From these considerations, we can abstract the beginning of an evaluation approach the laddering approach for correctness evaluation used in Gordon et al. (2013) with the following steps.

1. employ a knowledge elicitation technique with domain experts;
2. translate natural language sentences to rigorously formalized ontological statements and/or vice versa as necessary;
3. compare the knowledge elicited with the knowledge encoded in the ontology, taking care to.

Next, we discuss how to measure correctness. The results reported in Gordon et al. (2013) simply state that comparison of the results with the ontology "demonstrated agreement with BCIDO class hierarchies". The vagueness of this summary points to an outstanding question in ontology evaluation. How could we measure this agreement for a more rigorous evaluation? One approach is to compute an accuracy measure as proposed by Amith and Tao (2017) and Burton-Jones et al. (2005) that is the ratio of true statements to the number of statements in the ontology[3].

Since the result is a ratio rather than a count, the accuracy measures of two ontologies could be compared. However, to ensure a meaningful comparison, the ontologies ought to be normalized prior to computing conducting the evaluation (VRANDEČIĆ; SURE, 2007). A correct normalization will ensure that statements that are explicit in one ontology but inferred in another are both evaluated. Consider the following three equivalent sets of is_a statements:

(1)

lizard is_a mammal
cat is_a mammal
mammal is_a animal

(2)

lizard is_a mammal
lizard is_a animal
cat is_a mammal
mammal is_a animal

(3)

lizard is_a mammal
lizard is_a animal
cat is_a mammal
cat is_a animal
mammal is_a animal

Although each set of statements is logically equivalent to the others (assuming the standard interpretation of 'is_a' as transitive), (1) has a correctness score of .66, (2) has a correctness score of .75, and (3) has a correctness score of .80. Normalizing ontologies to generate an ontology where all implied statements are explicit and therefore included in the evaluation is necessary for a reliable and comparable fidelity score. Through a process of normalization like that proposed by (VRANDEČIĆ; SURE, 2007), each set of statements would be normalized to (3), which is therefore the set of propositions to evaluate for fidelity, so .80 is the correct fidelity score for (1), (2), and (3).

[3] Note that this requires an elicitation technique that also determines whether statements in the ontology are false.

FINAL REMARKS

This paper joins the general call for a rigorous discipline of ontology evaluation. It proposes that we systematize both the intrinsic and extrinsic properties of an ontology that can be evalauted according to the various aspects of an ontology. An initial review of the literature on ontology evaluation has revealed properties that emphasize ontologies as representational artifacts, as logical theories, as mathematical graphs, as parts of information systems, and as community resources for reuse. Further systematizes will likely reveal more categories and more fine grained distinctions. A discipline of ontology evaluation will also involve critically developing methods for examining those properties, developing comparative metrics that allow ontology engineers to compare the effects of various modeling choices and allow users to compare the merits of existing ontologies, and charting possible pitfalls of evaluation. As an example, we discussed methodological issues related to evaluating ontologies for correct representations.

The standard of correctness is different for realist, conceptualist, and linguistic ontologies, and this should be acknowledged in evaluation, and also considerations for applications. We also observed that elicitation techniques of domain knowledge from domain experts should be careful to present the domain knowledge in a manner that reflects the structure of the logical language the ontology is written in. For example, statements in OWL ontologies should be presented in natural language with an explicit quantifier to avoid ambiguity. Finally, metrics ought to be developed that can compare the correctness of ontologies and careful consideration ought to be given to the design of the evaluation study to ensure a comparable result. For example, the knowledge elicitation technique should detect both true and false statements in the ontology, and ontologies ought to be normalized to ensure meaningful comparison.

Future work for developing a systematic discipline of ontology evaluation includes a more complete survey of methods used to carry out ontology evaluations along with a detailed critique of their successes and pitfalls. Metrics ought to be developed that allow comparison and more work needs to be done determining what methods of ontology normalization are optimal for comparative metrics. While this paper has not discussed the many tools that exist for checking the quality of ontologies, undoubtedly the field will need more tools that can support comparing the performance of ontologies in information systems. Finally, data need to be generated that allow us to trace performance errors and successes in information systems to intrinsic properties of ontologies to help guide ontology development and selection.

ACKNOWLEDGEMENTS

I would like to thank Bill Hogan, Selja Seppälä, Andrew Spear, and Brian Stucky for helpful conversations during the writing of this paper. Work on this paper was supported in part by the NIH/NCATS Clinical and Translational Science Awards to the University of Florida UL1 TR000064. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

REFERENCES

AMITH, M.; TAO, C. Modulated evaluation metrics for drug-based ontologies. *Journal of Biomedical Semantics*, v. 8, n. 1, 2017.

ARP, R. et al. *Building ontologies with basic formal ontology*. New York: MIT Press, 2015.

BEN ABACHA, A. et al. Towards natural language question generation for the validation of ontologies and mappings. *Journal of Biomedical Semantics*, n. 7, 2016.

BURTON-JONES, A. et al. A semiotic metrics suite for assessing the quality of ontologies. *Data & Knowledge Engineering*, v. 55, n. 1, 2005.

GANGEMI, A. et al. A theoretical framework for ontology evaluation and validation. *SWAP*, v. 166, 2005.

_____. Modelling ontology evaluation and validation. *Semantic Web*, n. 4011, 2006.

GOMEZ-PEREZ, A. Evaluation of ontologies. *International Journal of Intelligent Systems*, v. 16, n. 3, 2001.

GORDON, C. L. et al. Design and evaluation of a bacterial clinical infectious diseases ontology. *AMIA Annual Symposium Proceedings*, 2013.

GRÜNINGER, M.; FOX, M. S. The role of competency questions in enterprise engineering. In: BENCHMARKING: theory and practice. Boston, MA: Springer US, 1995.

GUARINO, N. Toward a formal evaluation of ontology quality. *Ieee Intelligent Systems*, v. 19, n. 4, 2004.

HOEHNDORF, R. et al. Evaluation of research in biomedical ontologies. *Briefings in Bioinformatics*, v. 14, n. 6, 2013.

LIU, S. et al. Rxnorm: prescription for electronic drug information exchange. *IT professional*, v. 7, n. 5, 2005.

MASOLO, C. et al. *Wonderweb deliverable D17:* the wonderweb library of foundational ontologies and the DOLCE ontology. [S.l.: s.n.], 2002.

MILLER, G. A. Wordnet: a lexical database for English. *Communications of the ACM*, v. 38, n. 11, 1995.

_____. FELLBAUM, C. Semantic networks of English. *Cognition*, v. 41, n. 1, 1991.

NEUHAUS, F. et al. Towards ontology evaluation across the life cycle. *Applied Ontology*, v. 8, n. 3, 2013.

OBRST, L. et al. The evaluation of ontologies: toward improved semantic interoperability. In: SEMANTIC web. Cheung, K.-H. Boston, MA: Springer US, 2007.

ROSPOCHER, M. et al. Corpus-based terminological evaluation of ontologies. *Applied Ontology*, v. 7, n. 4, 2012.

VRANDEČIĆ, D.; SURE, Y. How to design better ontology metrics. *The Semantic Web*, 2007.

WALOSZEK, W. Measures for evaluation of structure and semantics of ontologies. *Metrology and Measurement Systems*, v. 19, n. 2, 2012.

WILKINSON, M.D. et al. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, n. 3, 2016.

ZHU, X. et al. A review of auditing methods applied to the content of controlled biomedical terminologies. *Journal of Biomedical Informatics*, v. 42, n. 3, 2009.