# A Snapshot of Ontology Evaluation Criteria and Strategies

Auriol Degbelo
Institute for Geoinformatics, University of Muenster
Heisenbergstrasse 2, 48149
Muenster, Germany
degbelo@uni-muenster.de

## ABSTRACT

Ontologies are key to information retrieval, semantic integration of datasets, and semantic similarity analyses. Evaluating ontologies (especially defining what constitutes a "good" or "better" ontology) is therefore of central importance for the Semantic Web community. Various criteria have been introduced in the literature to evaluate ontologies, and this article classifies them according to their relevance to the design or the implementation phase of ontology development. In addition, the article compiles strategies for ontology evaluation based on ontologies published until 2017 in two outlets: the Semantic Web Journal, and the Journal of Web Semantics. Gaps and opportunities for future research on ontology evaluation are exposed towards the end of the paper.

## CCS CONCEPTS

• **Computing methodologies** → **Ontology engineering**; • **Information systems** → *World Wide Web*;

## KEYWORDS

ACM proceedings, LaTeX, text tagging

**ACM Reference format:**
Auriol Degbelo. 2017. A Snapshot of Ontology Evaluation Criteria and Strategies. In *Proceedings of Semantics2017, Amsterdam, Netherlands, September 11–14, 2017,* 8 pages.
https://doi.org/10.1145/3132218.3132219

## 1 INTRODUCTION

Ontologies are one of the basic components enabling the Semantic Web vision. They have both theoretical and practical relevance. On the one hand, they are theories (the grail of researchers), a point that is reflected in Guarino's definition: "[a]n ontology is a logical theory accounting for the intended meaning of a formal vocabulary" [16]. On the other hand, they are implementable, that is, they can (when encoded in one or more ontology implementation languages) be processed by machines, and be used in a number of applications such as knowledge management systems, semantic portals and recommender systems, to mention but a few.

Since ontologies are of great relevance to the Semantic Web, the importance of ontology evaluation cannot be overstated. Ontology evaluation assesses *how good* an ontology performs with respect to specific criteria, and is necessary for tasks such as ontology ranking or ontology re-use. Despite various proposals for ontology evaluation, there is still a need for discussions and reflections about *what to consider* while assessing ontology (i.e., evaluation criteria), and *how best to carry it out* (i.e., evaluation strategies and best practices). As Ferrario and Grüninger [8] recently pointed out: "We have not yet reached a consensus about ontology evaluation". This work aims at informing ongoing discussions on ontology evaluation, through a look into the practice of ontology evaluation in the Semantic Web community. Investigating the practice of ontology evaluation is necessary (and useful) for an evidence-based discussion of the topic. The article looks at the following four research questions:

- RQ1: what are possible criteria to evaluate the design and the implementation of ontologies?
- RQ2: what are possible strategies to evaluate the design and the implementation of ontologies?
- RQ3: what are current links between theory and practice of ontology evaluation?
- RQ4: does the design vs. implementation distinction matter in practice?

The first research question (RQ1) is answered through a review of criteria for ontology evaluation suggested in previous work. Strategies for ontology evaluation extracted from ontologies published until 2017 in the Semantic Web Journal, and the Journal of Web Semantics are useful to provide a preliminary answer to the second research question (RQ2). Putting findings from the first two research questions in relation sheds some light on the relationship between theory and practice of ontology evaluation (RQ3), and the relevance of the design vs. implementation distinction in practice (RQ4). Section 2 motivates the use of the design vs. implementation distinction in this paper. Section 3 compiles criteria for ontology evaluation from the literature, and organizes them according to their relevance to the design or the implementation phase of ontology development. Section 4 presents ontology evaluation strategies obtained by examining articles of the two outlets mentioned above. Section 5 discusses possible answers to RQ3 and RQ4. Related work is introduced in Section 6 before Section 7 concludes the paper.

## 2 BACKGROUND

A discussion of ontology evaluation necessitates an explicit commitment to a *method* for ontology development. There is, as [30] recently indicated, no agreed upon method for ontology development. Despite the variety of these methods (for examples, see [6, 21, 30, 34]), it can be argued that two recurrent tasks during ontology building are of primary importance for the Semantic Web:

*Semantics2017, September 11–14, 2017, Amsterdam, Netherlands*

ACM ISBN 978-1-4503-5296-3/17/09.
https://doi.org/10.1145/3132218.3132219

modelling semantics, and encoding it. The former is a design task, and the latter an implementation task (see [27]).

Calls for the distinction between design tasks and implementation tasks in the context of ontology building were also implicit in [1, 20]. Bittner and Donnelly [1] argued for the need to understand a computational ontology as consisting of two complementary components: (i) an expressive ontology specified in first-order logic, and (ii) an ontology developed in description logics, which is computationally efficient, and thus useful for computer implementations. Likewise, Guizzardi [20] called for two classes of ontology languages for the discipline of ontology engineering: 'well-founded ontology representation languages' (with a focus on representation adequacy regardless of the consequent computational costs), and 'lightweight representation languages' (with a focus on adequate computational properties for the ontologies encoded in these languages). Guizzardi also indicated that the name 'ontology representation languages' when applied to the so-called Semantic Web languages (e.g., OWL) is a misnomer, since these languages are motivated by epistemological and computational concerns, not ontological ones.

Table 1 presents the peculiarities of each of the tasks of ontology design and ontology implementation. The table is based essentially on discussions provided in [20, 27]. Some comments on the table are in order at this point. First, that some languages are classified as implementation languages in the table implies by no means that they cannot be (or have not been) used for design, but such a use might come at the expense of greater expressiveness and understanding. Reasons why OWL may not be expressive enough for modelling were discussed in [27]. Second, it must be admitted that ontology implementation languages need also expressiveness to a certain degree. Nonetheless, the key feature of ontology implementation languages is not expressiveness per se, rather it is *the compromise between expressiveness and efficient reasoning support*. This has been the reason for not explicitly mentioning 'expressiveness' in the column IMPLEMENTATION of the table. The characteristics of the design, and the implementation stages of ontology development are briefly introduced next.

## 2.1 Design (or modelling) stage

The steps of the design stage mentioned here are partially modified from the seminal work of Grüninger and Fox [15]. The design stage is an iterative process consisting of three steps: (i) identification of a motivating scenario, (ii) identification of the terms of the ontology, and (iii) formal specification of the terms of the ontology.

*Identification of a motivating scenario.* A motivating scenario is a story problem or example which is not adequately addressed by existing ontologies (see [35]). A motivating scenario provides a set of intuitively possible solutions to the scenario problem, and helps to understand the motivation for the proposed ontology in terms of its applications. Given the motivating scenario, a set of questions to be answered by the ontology (i.e., competency questions) is extracted. Competency questions specify the (expressiveness) requirements for an ontology, and are a means to give an informal justification of the necessity of the ontology to be developed (see [15, 35]). Overall, the specification of a motivating scenario, and of competency questions helps to delineate the scope of the ontology.

*Identification of the terms of the ontology.* At this stage, terms from existing ontologies are considered for re-use and new terms introduced (if necessary), to cover the needs arising from the motivating scenario. The terms of the ontology are the *objects*, *relations* and *attributes* that are required to answer the competency questions. Once the terms have been identified, they can be aligned to a foundational ontology. Brodaric and Probst [5] point out that an alignment is realized by establishing an *is-a* relation between an ontology element and a foundational ontology element. Some benefits of ontology alignment presented in [29] are: (i) conceptual disambiguation, (ii) increased axiomatization, (iii) improved design, and (iv) possible comparison of several aligned ontologies.

*Formal specification of the terms of the ontology.* At this step of the design stage, axioms are provided, and the terms of the ontology are specified using an ontology design language. An axiom "contains formulas which are considered to be always true (and therefore *sharable* among multiple agents), independently of particular states of affairs" [17]. Axioms are useful to specify the definitions of terms in the ontology, and constraints on their interpretation (see [15, 35]). The whole obtained by putting together terms and axioms is a *logical theory*, and this theory represents the outcome of the design stage.

## 2.2 Implementation (or encoding) stage

As Bittner *et al.* [2] pointed out: "[o]nce one has developed a highly expressive theory, less expressive logics with better computational properties can be used to implement certain portions of the full theory for specific purposes". The goal of the implementation stage is to isolate portions of the theory that have some desired computational properties (e.g., tractability). The terms of the ontology identified during the design stage (all or some of them) are re-used during the implementation stage. A *subset of the axioms* from the design stage is isolated and implemented in an ontology implementation language. A distinguishing criterion between ontology implementation language and ontology design language is that the former *must be machine-readable*, whereas the latter *needs not be*. The outcome of the implementation stage is a computational artifact that can be used in practical tasks such as query disambiguation, query term expansion, relevance ranking and web resource annotation. The design and implementation phases, as described above, are used in the remainder of this article to frame the discussion on ontology evaluation.

# 3 ONTOLOGY EVALUATION CRITERIA

Ontology evaluation is defined in this work after Gómez-Pérez *et al.* [13] as a technical judgment of the ontology with respect to a frame of reference. The frame of reference, as Gómez-Pérez *et al.* point out, can be requirements specification, competency questions, and the real world. Ontology evaluation can be carried out for two major goals mentioned in [39]: *tracking progress in ontology development*, and *ontology selection*. The distinction mentioned in [11] between *technical evaluation*, and *user evaluation* fits with these two goals. Technical evaluation is carried out by ontology developers and aims at tracking progress during ontology development; user evaluation is done by end-users of the ontology, and aims at selecting an ontology for a given purpose.

**Table 1: Key features of the design and implementation stages of an ontology building process (from [7])**

| | DESIGN | IMPLEMENTATION |
|---|---|---|
| Goal | Support human understanding | Support automated reasoning |
| End consumer | Humans (in tasks such as communication & domain analysis) | Machines (in tasks such as inference & reasoning) |
| Requirements for supporting languages | Conceptual clarity, Expressiveness | Efficient automated reasoning, Decidability, Scalability |
| Examples of supporting languages | First-order logic, Haskell, UML, Isabelle/HOL | OIL, DAML, DAML+OIL, RDFs, OWL, LINGO |

There has been work covering technical evaluation only (e.g., OntoClean [19]) or focusing solely on user evaluation (e.g., ONTOMETRIC [28]). Examples of work covering both technical and user evaluation are [10, 25]. Kehagias *et al.* [25] proposed a set of criteria to ensure ontology validation. The authors distinguish between internal measures concerned with the ontologies themselves (e.g., density, cognitive adequacy), and external measures concerned with their take-up and use within user communities (e.g., availability, ease and effectiveness of access). Gangemi *et al.* [10] identified three types of measures for ontology evaluation: structural measures, functional measures, and usability-related measures. Structural and functional measures are pertinent to both *technical* and *user evaluation*, usability measures are relevant to *user evaluation.* Reviews of approaches for ontology evaluation can be found in e.g., [3, 23, 32].

The works on ontology evaluation aforementioned have one point in common, namely that they have not discussed ontology evaluation in relation to the design-implementation distinction. To fill this gap, the following repartition of previous criteria is suggested. The criteria listed below were extracted from the existing literature on ontology evaluation. They can be used for both technical and user evaluation. Since the design stages yields a theory (see Section 2.1), criteria suitable for this stage are those which assess theoretical properties of the ontology. On the contrary, the implementation stage results in a computational artifact (see Section 2.2). That is, criteria useful for the evaluation of the implementation stage assess computational properties of the ontology. Knowing how the criteria are related to specific phases is vital for the development of tools which support automatic evaluation of either the design, or the implementation (or both phases) of ontology building.

**Design evaluation**: criteria possibly relevant for the evaluation of the design stage of an ontology development process include:

- *accuracy* [36] (i.e., correct representation of aspects of the real world)
- *adaptability* [14, 36] (i.e., ease of performing changes)
- *clarity* [14] (i.e., effective communication of the intended meaning of defined terms)
- *cognitive adequacy* [25] (i.e., match between formal and cognitive semantics)
- *completeness* [12, 36] (i.e., appropriate coverage of the domain of interest)
- *conciseness* [12, 36] (i.e., absence of unnecessary or useless definitions or axioms)
- *consistency* [14, 36] (i.e., incapacity of getting contradictory conclusions from valid input data)
- *expressiveness* [15, 31] (i.e., number of competency questions that the ontology can answer)
- *grounding* [14, 26] (i.e., number of assumptions done by the ontology's underlying philosophical theory about reality)

**Implementation evaluation**: criteria possibly useful for the evaluation of the implementation stage of an ontology development process include:

- *computational efficiency* [36] (i.e., ease and speed of processing by reasoners)
- *congruency* [25] (i.e., fitness between ontology and corpus terms)
- *practical usefulness* [31] (i.e., number of practical problems to which the ontology can be applied)
- *precision* [25] (i.e., fraction of retrieved instances by the ontology that are relevant)
- *recall* [25] (i.e., fraction of relevant instances that are retrieved by the ontology)

The criteria listed above provide an answer to RQ1.

## 4 ONTOLOGY EVALUATION STRATEGIES

Section 3 has discussed *what aspects* could be evaluated during the design or implementation of ontologies. The focus of this section is on *how* evaluation can be done (i.e., approaches and strategies[1]). Previous work has discussed various approaches for the evaluation of ontologies. For example, Hammar and Sandkuhl [22] differentiate *validation by example* from *empirical validation.* Validation by example happens when one or more examples are presented in natural language to illustrate the relevance of the concepts of the ontology in a theoretical manner. Empirical validation occurs when some sort of experimental procedure or case study is performed to assess the ontology. Hoehndorf *et al.* [23] distinguish between *direct evaluation*, *application-based evaluation*, and *analysis-based evaluation* of ontologies. A direct evaluation assesses intrinsic properties of the ontology (e.g., consistency, expressivity) via peer-review; an application-based evaluation assesses the ontology, based on an application that makes use of it; and an analysis-based evaluation performs a scientific data analysis that relies on an ontology, and evaluates the success of the analysis using criteria established in a scientific domain. According to Hoehndorf *et al.*, these three evaluation approaches play distinct roles and are complementary:

[1] A note on the terminology: 'approaches' is used in this context to refer to general ways of dealing with ontology evaluation, whereas 'strategies' denote specific actions.

peer review can assure the ontology's adherence to scientific reporting standards; an application-based evaluation ensures that ontologies can be used efficiently; and the evaluation using a scientific analysis ensures that ontologies lead to verifiable novel insights in science. Obrst *et al.* [32] listed five approaches for the evaluation of ontologies in the life sciences: (i) evaluation with respect to the use of an ontology in an application, (ii) evaluation with respect to domain data sources, (iii) assessment by humans against a set of criteria, (iv) evaluation of ontologies in terms of their impact on natural language processing tasks, and (v) the use of 'reality itself' as a benchmark. Brank *et al.* [4] mentioned four approaches: gold-standard-based (i.e., comparing the ontology to a gold standard which may itself be an ontology); application-based (i.e., using the ontology in an application and evaluating the results); data-driven (i.e., comparison with a source data about the domain that is to be covered by the ontology); and human-based (i.e., humans try to assess how well an ontology meets a set of predefined criteria). There are some similarities between the two lists of [4] and [32]: gold-standard-based approaches are related to (v) above; application-based approaches are equivalent to (i); data-driven ones are similar to (ii); and human-based approaches correspond to (iii).

Despite much theoretical and conceptual work on ontology evaluation, empirical assessments of ontology evaluation within the Semantic Web community have not been often undertaken. To get a glimpse of how ontology evaluation has been practiced so far, this article offers an assessment of articles published in the Journal of Web Semantics (JWS) and the Semantic Web Journal (SWJ). These two outlets were chosen because (i) they focus on Semantic Web research and development, and (ii) they are currently the two most visible and influential journals within the community[2]. Given their good standings, it seems reasonable to assume that (i) their published ontologies have undergone some rigorous evaluation, and (ii) the evaluation criteria used are meaningful for the community. The value of insights from an empirical assessment is that they highlight areas, if any, where theoretical and applied work on ontology evaluation can learn from each other.

**Data collection**: Two inclusion criteria were defined for the selected papers as follows: (i) the paper should be about ontology, and (i) the paper should describe the process of ontology development. Articles which (i) described other aspects of Semantic Web research (e.g., ontology ranking, ontology tools, ontology matching, ontology-based systems), (ii) focused solely on ontology acceptance by the community, or (iii) only described an ontology without the process of obtaining it, were excluded from the analysis. Ontology papers from the JWS were retrieved using the query http://www.websemanticsjournal.org/index.php/ps/search/search?type=Ontology+Paper. Ontology papers from the SWJ were collected by applying keyword search (i.e., "ontolog" & "Ontolog") over http://www.semantic-web-journal.net/issues. "ontolog" & "Ontolog" were used as keywords to take into account variants of the word such as 'ontology', 'ontologies', 'Ontology' and 'Ontologies'. The searches returned in total 74 articles (N=26, JSW; N=48, SWJ). The data was collected in March 2017.

**Data analysis**: After application of the inclusion and exclusion criteria mentioned above, 26 papers were left (N=11, JWS; N=15, SWJ) which are summarized in Table 2. The papers cover the timeframe 2003-2017. Though ontology evaluation criteria are the main focus of the work, they are strongly tied to the method of ontology development as mentioned in Section 2. For this reason, the 26 papers were annotated taking into account: presence (or absence) of a design/implementation phase, ontology language(s) used for the design/implementation phase (if any), and criteria used for evaluation of each of the phases (if any). Since there is currently no consensus regarding what ontology evaluation is (let alone how to document it), the 'principle of charity' from [37] was adopted when annotating the 26 articles from the sample. That is, articles which followed the steps of ontology design/implementation (as described in Section 2), and used the criteria described in Section 3 were labelled as having a design/implementation stage, and having used the criteria (even if the terminology used in these articles was different from the terminology introduced in this paper).

**Results**: Table 2 shows that *expressiveness* stands out from the criteria for ontology design evaluation suggested in the literature. The works examined used different strategies: J1, J2, J9, S1, S3, S6, S14 used *diagrams* to show how the ontology can be used to model some illustrative examples; S2 described in *natural language* how the ontology concepts can be used to model relevant examples; J3, S5, S11, S12, and S13 presented *Turtle excerpts* to illustrate the use of the ontology to describe resources; S7 used both diagrams and RDF serializations; J7 explicitly listed *competency questions*, and discussed how the ontology can be used to answer them; S8 translated all the competency questions of the ontology into SPARQL and offered an endpoint against which they could be tested; and S10 provided examples of queries, both in natural language and SPARQL which the ontology can help to answer. With respect to ontology implementation, *practical usefulness* is by and large the criterion often used. Most ontology papers mentioned one (or several) applications which use the ontology. Furthermore, J7 compared the performance of a baseline application with an ontology-based application. In addition to practical usefulness, S2 took *computational efficiency* into account by providing some data about the time needed for reasoning when applying the rules of the ontology. A criterion (which is not on the list from Section 3), but was discussed by J7, S10, S11 is *adoption of the ontology* by the community. Ontology adoption is a useful evaluation criterion which is not specific to design or implementation, but applies to both taken together. The strategies articulated in this paragraph offer a beginning of an answer to RQ2.

## 5 DISCUSSION

The analysis of the sample ontology papers leads to some observations which are discussed in this section.

**Theory vs. practice of ontology evaluation**: there is a recognizable gap between the theory and practice of ontology evaluation. Apart from expressiveness and practical usefulness, very few of the

[2] The statement is based on data compiled in June 2016 by Google Scholar. See https://scholar.google.com/citations?hl=en&view_op=search_venues&vq=semantic+web (last accessed: April 18, 2017).

**Table 2: Ontology articles analyzed and their characteristics.**

| | Articles in alphabetical order; [Year] | D. Phase? | I. Phase? | D. Language | I. Language | D. Evaluation | I. Evaluation |
|---|---|---|---|---|---|---|---|
| | The Journal of Web Semantics (https://www.journals.elsevier.com/journal-of-web-semantics/) | | | | | | |
| J1 | Design and use of the Simple Event Model (SEM); [2011] | Yes | Yes | RDFs | RDF | Expressiveness | Practical usefulness |
| J2 | DOLCE ergo SUMO: On foundational and domain models in the SmartWeb Integrated Ontology (SWIntO); [2007] | Yes | Yes | UML, RDFs | RDFs | Expressiveness | Practical usefulness |
| J3 | FaBiO and CiTO: ontologies for describing bibliographic resources and citations; [2012] | Yes | Yes | OWL 2 DL | RDF | Expressiveness | Practical usefulness |
| J4 | Key choices in the design of Simple Knowledge Organization System (SKOS); [2013] | Yes | Yes | OWL 1 | RDF/OWL | - | Practical usefulness |
| J5 | Ontologies for ecoinformatics; [2006] | Yes | Yes | OWL-DL | OWL-DL | - | Practical usefulness |
| J6 | SwetoDblp ontology of computer science publications; [2007] | Yes | Yes | - | RDF | - | Practical usefulness |
| J7 | The Data Mining OPtimization ontology; [2015] | Yes | Yes | OWL 2 | OWL 2 | Expressiveness | Practical usefulness |
| J8 | The National Cancer Institute's thesaurus and ontology; [2003] | - | Yes | - | OWL-Lite | - | - |
| J9 | The SSN ontology of the W3C Semantic Sensor Network incubator group; [2012] | Yes | Yes | OWL 2 | OWL 2 | Expressiveness | Practical usefulness |
| J10 | Translating the foundational model of anatomy into OWL; [2008] | Yes | - | OWL DL/Full | - | - | - |
| J11 | Web authoring for accessibility (WAfA); [2007] | Yes | Yes | OWL DL | OWL DL | - | Practical usefulness |
| | The Semantic Web Journal (http://www.semantic-web-journal.net/) | | | | | | |
| S1 | An ontology design pattern and its use case for modeling material transformation; [2017] | Yes | Yes | first-order-logic/DL | OWL | Expressiveness | - |
| S2 | An OWL ontology library representing judicial interpretations; [2016] | Yes | Yes | OWL 2 | OWL 2 | Expressiveness | Computational efficiency, practical usefulness |
| S3 | LOTED2: an ontology of European public procurement notices; [2016] | Yes | Yes | OWL 2 DL | OWL 2 DL | Expressiveness | - |
| S4 | OLiA - Ontologies of linguistic annotation; [2015] | Yes | Yes | OWL 2 DL | OWL 2 DL | - | Practical usefulness |
| S5 | Ontology for observations and sampling features, with alignments to existing models; [2017] | Yes | - | OWL 2 DL | - | Expressiveness | - |
| S6 | Ontology of units of measure and related concepts; [2013] | Yes | Yes | OWL 2 | OWL 2 | Expressiveness | Practical usefulness |
| S7 | Overview of the MPEG-21 media contract ontology; [2016] | Yes | Yes | OWL 2 | OWL 2 | Expressiveness | Practical usefulness |
| S8 | PPROC, an ontology for transparency in public procurement; [2016] | Yes | Yes | OWL | OWL | Expressiveness | Practical usefulness |
| S9 | The Bowlogna ontology: Fostering open curricula and agile knowledge bases for Europe's higher education landscape; [2013] | Yes | Yes | OWL | OWL | - | Practical usefulness |
| S10 | The collections ontology: creating and handling collections in OWL 2 DL frameworks; [2014] | Yes | Yes | OWL 2 DL | OWL 2 DL | Expressiveness | Practical usefulness |
| S11 | The document components ontology (DoCO); [2016] | Yes | Yes | OWL | OWL | Expressiveness | Practical usefulness |
| S12 | The publishing workflow ontology (PWO); [2017] | Yes | Yes | OWL 2 DL | OWL 2 DL | Expressiveness | - |
| S13 | Time ontology extended for non-Gregorian calendar applications; [2016] | Yes | Yes | OWL | OWL | Expressiveness | - |
| S14 | Using an ontology for representing the knowledge on literary texts: The Dante Alighieri case study; [2017] | Yes | Yes | RDFs | RDFs | Expressiveness | Practical usefulness |
| S15 | Using the relation ontology Metarel for modelling Linked Data as multi-digraphs; [2014] | Yes | Yes | - | RDF | - | Practical usefulness |

criteria suggested by previous work have been often used. There is therefore a clear evidence that practice still needs to make more use of the criteria suggested by theory (i.e., the literature). This poses also the question of the actual relevance of criteria other than expressiveness and practical usefulness for ontology evaluation in the Semantic Web community.

Regarding the approaches for ontology evaluation listed in Section 4, Hammar and Sandkuhl's distinction between validation by example and empirical validation seems to be the best fit for the current practice of ontology evaluation in the Semantic Web. As observed in the previous section, design is often validated by examples and implementation by some experimental procedure (i.e., application development). Hoehndorf *et al.*'s distinction between direct evaluation and application-based evaluation is also appropriate for ontology evaluation practice. Direct evaluation touches on the intrinsic properties of the ontology and happens during the design stage; application-based evaluation is relevant for the implementation stage. However, (and contrary to Hoehndorf *et al.*'s suggestion) direct evaluation *needs not* happen through peer-review. J2 and S4 provided examples of evaluation of ontologies in terms of their impact on natural language processing tasks. J7 used a baseline application as gold standard. There is no documentation of analysis-based evaluation, evaluation with respect to domain data sources, assessment by humans against a set of criteria, and 'reality itself' used as benchmark, in the sample analyzed. In sum (and apropos RQ3), there seems to be a gap between theory and practice, with only few of the criteria suggested in the literature being actually used while evaluating existing Semantic Web ontologies.

**Design vs. implementation of ontology in practice**: In practice, design and implementation of ontologies seem to be intertwined: that is, design is almost always followed by implementation for Semantic Web ontologies. This could be because OWL is a language that can be used for both purposes. It is also striking to observe that languages such as Haskell, first-order logic or Isabelle which are used for ontology design in other communities (cf. [1, 2, 9, 23, 33]) are barely present in Table 2. One way to identify the reasons for this, is to look at why some of the papers analyzed went beyond OWL. S1 (from Table 2) used first order logic to model some statements which could not be expressed within OWL 2 DL. This suggests that more expressive languages than OWL are needed *in some cases*. Nevertheless, the apparent rarity of these cases is a signal that additional empirical investigations are needed to find out the extent to which they actually matter for the Semantic Web. Finally, Table 2 suggests no apparent relationship between language for ontology design/implementation, and criteria used. In summary (and with respect to RQ4), the design and implementation stages of ontology building seem interwoven during Semantic Web ontology development, with OWL often being used for both stages.

**Limitations**: despite the large timeframe (2003-2017) covered and the prominent outlets selected, the relatively small sample of papers analyzed suggests that caution should be made generalizing the observations of this article. Future work could replicate this work using a larger sample including articles from the International Semantic Web Conference (ISWC), and/or the Extended Semantic Web Conference (ESWC). Focusing only on ontologies published in the JWS, and the SWJ has also put some constraints on the number of papers selected. Therefore, the sample can also be enlarged through the inclusion of *vocabularies* from the JWS, the SWJ, the ISWC, and the ESWC (provided documentation about the creation process of, and evaluation of these vocabularies is available). Furthermore, ontologies may be distinguished into three types, after [18]: lightweight ontologies (which mainly focus on machine interoperability); reference ontologies (which carefully aim to avoid misunderstandings among humans); and foundational ontologies (which provide the basic common vocabulary for conceptual modeling and ontological analysis). Traditionally, lightweight ontologies have attracted more attention in the Semantic Web community. Outlets such as the Formal Ontology in Information Systems, or the Applied Ontology journal have a traditional focus on reference ontologies, and foundational ontologies. Analyzing articles from these outlets may have produced another picture of ontology evaluation criteria and strategies. At last, it has been partly subjective to classify the papers analyzed, because they were described at different levels of granularity. Involving the authors of the surveyed papers in the classification process may have refined the outcomes of the analysis.

## 6 RELATED WORK

A recent ontology summit on ontology evaluation [30] has suggested a model of the ontology life cycle with eight phases. Design & Implementation which were discussed here correspond to one phase of [30], namely the "Ontology Development Phase". By focusing solely on this phase, and discussing it in greater detail, this work has aimed to provide a greater understanding of how ontology development is currently applied in the Semantic Web community. Moreover, Neuhaus *et al.* [30] suggested questions that an assessment of the "Ontology Development Phase" could answer. Examples of these questions include: Does the formalization capture the intent of the competency question appropriately? Does the ontology support operational requirements (e.g., performance, precision, recall)? The criteria listed in Section 3, and the strategies identified in Section 4 provide an indication of how the Semantic Web community has approached these questions.

As discussed in Section 3, several criteria for the evaluation of ontology were proposed in existing work (e.g., [10, 25]). A discussion of how these criteria relate to design and implementation activities was still missing. This article introduced a proposal to fill this gap. In practice, the transition between design activities and implementation activities may not be clear-cut (see [30], and also Section 5). However (and in line with [30]), the conceptual distinction between design evaluation criteria, and implementation evaluation criteria remains useful because both activities have different goals, and lead to different outcomes. From a researcher's point of view (and as indicated in [7]), ontology design is useful to explore possible coherent ways of approaching an issue; ontology implementation is helpful to expose areas for future investigations regarding available technologies. In addition, Janowicz and Hitzler posed the question "where is the sweet spot for ontologies that go beyond surface semantics?" [24] after reflecting on the relation between linked data, semantic annotations and ontologies. They added that the question has no simple answer. The design vs. implementation distinction may help make some guesses. According to Section 2, Design & Implementation yield two outcomes: logical theories of a domain, and computational artifacts. Since implementation concerns (i.e., computational artifacts) are not the main focus of ontologies going beyond surface semantics (i.e., foundational ontologies or reference ontologies), their attractiveness might reside in the logical theories they produce. Theories are vehicles for thought in a community (see [38]). The call to arms[3] in Frank van Harmelen's keynote address at the 10th International Semantic Web Conference recently

[3] See http://videolectures.net/iswc2011_van_harmelen_universal/; last accessed: May 19, 2017).

reminded that theories of the information universe are still needed. As a result, the sweet spot for ontologies that go beyond surface semantics may be their supply of a *greater human understanding of the information universe and/or a scientific domain.*

Finally, syntheses of ontology evaluation practices for other communities exist. Hoehndorf *et al.* [23] scoped out biomedical ontologies, while Obrst *et al.* [32] focused on the life sciences. By looking at articles from the JWS and the SWJ, this article has offered a starting point for a better understanding of the practice of ontology evaluation in the Semantic Web community.

## 7 CONCLUSION

Despite much theoretical and conceptual work on ontology evaluation, consensus is yet to be reached regarding ontology evaluation. Through a clear commitment to an ontology development method, and an empirical assessment of ontology evaluation within the Semantic Web community, this article has made the following contributions to the ongoing discussion on ontology evaluation:

- *Organization of existing evaluation criteria according to their relevance for the design/implementation activities during ontology development*: ontology evaluation is necessarily tied to an ontology development method (and its corresponding phases). Since design and implementation are recurrent activities in existing methods, they could offer a useful frame for further discussions on the road towards a consensus on ontology evaluation criteria and strategies;
- *Elicitation of strategies for ontology evaluation based on practice*: expressiveness has often been demonstrated in practice through the use of diagrams and/or turtle excerpts, while practical usefulness is often assessed through one or more applications which use the ontology;
- *Identification of links between theory and practice of ontology evaluation*: among the several possible ways of characterizing ontology evaluation, the distinctions between validation by example vs. empirical validation, and direct evaluation vs. application-based evaluation seem the most promising to describe the practice of ontology evaluation in the Semantic Web community. In addition, expressiveness and practical usefulness seem to be often used in practice, whereas several other criteria (e.g., conciseness, grounding, precision, recall) seem to have had lower adoption by the community so far;
- *Opportunities for future research*: these include the need to develop strategies to broaden aspects of ontologies considered during evaluation; investigating when criteria other than expressiveness and practical usefulness are sensible for ontology evaluation in practice; and increasing the understanding of ontology evaluation by observing its practice in additional prominent outlets of the Semantic Web.

## ACKNOWLEDGMENTS

The author gratefully acknowledges funding from the European Union through the GEO-C project (H2020-MSCA-ITN- 2014, Grant Agreement Number 642332, http://www.geo-c.eu/).

## REFERENCES

[1] T Bittner and M Donnelly. 2007. Logical properties of foundational relations in bio-ontologies. *Artificial Intelligence in Medicine* 39, 3 (2007), 197–216.

[2] T Bittner, M Donnelly, and B Smith. 2009. A spatio-temporal ontology for geographic information integration. *International Journal of Geographical Information Science* 23, 6 (2009), 765–798.

[3] J Brank, M Grobelnik, and D Mladenić. 2005. A survey of ontology evaluation techniques. In *Information Society 2005 - 8th International Multiconference*, O Markič, M Gams, U Kordež, M Heričko, D Mladenić, M Grobelnik, I Rozman, V Rajkovič, T Urbančič, M Bernik, and M Bohanec (Eds.). Ljubljana, Slovenia, 166–169.

[4] J Brank, M Grobelnik, and D Mladenić. 2007. Automatic evaluation of ontologies. In *Natural Language Processing and Text Mining*, A Kao and S R Poteet (Eds.). Springer-Verlag London Limited, 193–219.

[5] B Brodaric and F Probst. 2008. DOLCE ROCKS: integrating geoscience ontologies with DOLCE. In *AAAI Spring Symposium: Semantic Scientific Knowledge Integration*, D McGuinness, P Fox, and B Brodaric (Eds.). AAAI, Stanford, California, USA, 3–8.

[6] O Corcho, M Fernández-López, and A Gómez-Pérez. 2003. Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data & Knowledge Engineering* 46, 1 (2003), 41–64.

[7] A Degbelo. 2015. *Spatial and temporal resolution of sensor observations.* Dissertations in Geographic Information Science, IOS Press. 206 pages.

[8] Roberta Ferrario and Michael Grüninger. 2017. Applied ontology: A foreword by the new editors-in-Chief. *Applied Ontology* 12, 1 (mar 2017), 1–4.

[9] A Frank. 2003. Ontology for spatio-temporal databases. In *Spatio-Temporal Databases: The CHOROCHRONOS Approach*, T Sellis, M Koubarakis, A U Frank, S Grumbach, R H Güting, C S Jensen, N Lorentzos, Y Manolopoulos, E Nardelli, B Pernici, B Theodoulidis, N Tryfona, H Schek, and M Scholl (Eds.). Springer-Verlag Berlin Heidelberg, Chapter 2, 9–77.

[10] A Gangemi, C Catenacci, M Ciaramita, and J Lehmann. 2006. Modelling ontology evaluation and validation. In *The Semantic Web: Research and Applications, 3rd European Semantic Web Conference (ESWC 2006)*, Y Sure and J Domingue (Eds.). Springer, Budva, Montenegro, 140–154.

[11] A Gómez-Pérez. 2001. Evaluation of ontologies. *International Journal of Intelligent Systems* 16, 3 (2001), 391–409.

[12] A Gómez-Pérez. 2003. Ontology evaluation. In *Handbook on ontologies* (1st ed.), S Staab and R Studer (Eds.). Springer Berlin Heidelberg, 251–273.

[13] A Gómez-Pérez, N Juristo, and J Pazos. 1995. Evaluation and assessment of the knowledge sharing technology. In *Towards Very Large Knowledge Bases*, N J I Mars (Ed.). IOS Press, Enschede, The Netherlands, 289–296.

[14] T Gruber. 1995. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies* 43, 5-6 (1995), 907–928.

[15] M Grüninger and M S Fox. 1995. Methodology for the design and evaluation of ontologies. In *Proceedings of the IJCAI Workshop on Basic Ontological Issues in Knowledge Sharing*. Montreal, Quebec, Canada.

[16] N Guarino. 1998. Formal ontology and information systems. In *Proceedings of the 1st International Conference on Formal Ontology in Information Systems (FOIS'98)*, N Guarino (Ed.). IOS Press Amsterdam, Trento, Italy, 3–15.

[17] N Guarino and P Giaretta. 1995. Ontologies and knowledge bases towards a terminological clarification. In *Towards very large knowledge bases*, N J I Mars (Ed.). IOS Press, Enschede, The Netherlands, 25–32.

[18] Nicola Guarino and Mark Musen. 2015. Applied ontology: The next decade begins. *Applied Ontology* 10, 1 (may 2015), 1–4.

[19] N Guarino and C Welty. 2002. Evaluating ontological decisions with OntoClean. *Commun. ACM* 45, 2 (2002), 61–65.

[20] G Guizzardi. 2007. On Ontology, ontologies, conceptualizations, modeling languages, and (meta)models. In *Proceedings of the 2007 conference on Databases and Information Systems IV: Selected Papers from the Seventh International Baltic Conference DB&IS'2006*, O Vasilecas, J Eder, and A Caplinskas (Eds.). IOS Press, Vilnius, Lithuania, 18–39.

[21] Giancarlo Guizzardi. 2010. Theoretical foundations and engineering tools for building ontologies as reference conceptual models. *Semantic Web* 1, 1 (2010), 3–10.

[22] K Hammar and K Sandkuhl. 2010. The state of ontology pattern research: a systematic review of ISWC, ESWC and ASWC 2005-2009. In *Proceedings of the 2nd International Workshop on Ontology Patterns - WOP2010*, E Blomqvist, V Chaudhri, O Corcho, V Presutti, and K Sandkuhl (Eds.). CEUR-WS.org, Shanghai, China.

[23] Robert Hoehndorf, Michel Dumontier, and Georgios V. Gkoutos. 2013. Evaluation of research in biomedical ontologies. *Briefings in Bioinformatics* 14, 6 (2013), 696–712.

[24] Krzysztof Janowicz and Pascal Hitzler. 2013. Thoughts on the complex relation between linked data, semantic annotations, and ontologies. In *Proceedings of the Sixth International Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR '13)*, Paul. N. Bennett, Evgeniy Gabrilovich, Jaap Kamps, and Jussi Karlgren (Eds.). ACM, San Francisco, California, USA, 41–44.

[25] D D Kehagias, I Papadimitriou, J Hois, D Tzovaras, and J Bateman. 2008. A methodological approach for ontology evaluation and refinement. In *The 2nd International Conference of ASK-IT*. Nuremberg, Germany.

[26] W Kuhn. 2009. Semantic engineering. In *Research Trends in Geographic Information Science*, G Navratil (Ed.). Springer Berlin Heidelberg, 63–76.

[27] W Kuhn. 2010. Modeling vs encoding for the Semantic Web. *Semantic Web* 1, 1 (2010), 11–15.

[28] A Lozano-Tello and A Gómez-Pérez. 2004. Ontometric: a method to choose the appropriate ontology. *Journal of Database Management* 2, 15 (2004), 1–18.

[29] P Mika, D Oberle, A Gangemi, and M Sabou. 2004. Foundations for service ontologies: aligning OWL-S to DOLCE. In *Proceedings of the 13th international conference on World Wide Web*, S I Feldman, M Uretsky, M Najork, and C E Wills (Eds.). ACM, New York, New York, USA, 563–572.

[30] Fabian Neuhaus, Amanda Vizedom, Ken Baclawski, Mike Bennett, Mike Dean, Michael Denny, Michael Grüninger, Ali Hashemi, Terry Longstreth, Leo Obrst, Steve Ray, Ram Sriram, Todd Schneider, Marcela Vegetti, Matthew West, and Peter Yim. 2013. Towards ontology evaluation across the life cycle: The Ontology Summit 2013. *Applied Ontology* 8, 3 (2013), 179–194.

[31] N F Noy and C D Hafner. 1997. The state of the art in ontology design: a survey and comparative review. *AI Magazine* 18, 3 (1997), 53.

[32] L Obrst, W Ceusters, I Mani, S Ray, and B Smith. 2007. The evaluation of ontologies. In *Semantic Web - Revolutionizing Knowledge Discovery in the Life Sciences*, C Baker and K Cheung (Eds.). Springer, 139–158.

[33] C Stasch, S Scheider, E Pebesma, and W Kuhn. 2014. Meaningful spatial prediction and aggregation. *Environmental Modelling & Software* 51 (2014), 149–165.

[34] Y Sure, S Staab, and R Studer. 2009. Ontology engineering methodology. In *Handbook on ontologies* (2nd ed.), S Staab and R Studer (Eds.). Springer Berlin Heidelberg, 135–152.

[35] M Uschold and M Grüninger. 1996. Ontologies: principles, methods and applications. *The Knowledge Engineering Review* 11, 2 (1996), 93–136.

[36] D Vrandečić. 2009. Ontology evaluation. In *Handbook on ontologies* (2nd ed.), S Staab and R Studer (Eds.). Springer Berlin Heidelberg, 293–313.

[37] N L Wilson. 1959. Substances without substrata. *The Review of Metaphysics* 12, 4 (1959), 521–539.

[38] Jacob O. Wobbrock and Julie A. Kientz. 2016. Research contributions in human-computer interaction. *Interactions* 23, 3 (2016), 38–44.

[39] J Yu, J A Thom, and A Tam. 2009. Requirements-oriented methodology for evaluating ontologies. *Information Systems* 34, 8 (2009), 766–791.