



Enhancing Human-in-the-Loop Ontology Curation Results through Task Design

STEFANI TSANEVA and MARTA SABOU, Vienna University of Economics and Business, Austria and TU Wien, Austria

The success of artificial intelligence (AI) applications is heavily dependent on the quality of data they rely on. Thus, data curation, dealing with cleaning, organising, and managing data, has become a significant research area to be addressed. Increasingly, semantic data structures such as ontologies and knowledge graphs empower the new generation of AI systems. In this article, we focus on ontologies as a special type of data. Ontologies are conceptual data structures representing a domain of interest and are often used as a backbone to knowledge-based intelligent systems or as an additional input for machine learning algorithms. Low-quality ontologies, containing incorrectly represented information or controversial concepts modelled from a single viewpoint, can lead to invalid application outputs and biased systems. Thus, we focus on the curation of ontologies as a crucial factor for ensuring trust in the enabled AI systems. While some ontology quality aspects can be automatically evaluated, others require a human-in-the-loop evaluation. Yet, despite the importance of the field, several ontology quality aspects have not yet been addressed and there is a lack of guidelines for optimal design of human computation tasks to perform such evaluations. In this article, we advance the state-of-the-art by making two novel contributions. First, we propose a human computation (HC)-based approach for the *verification of ontology restrictions*—an ontology evaluation aspect that has not yet been addressed with HC techniques. Second, by performing two controlled experiments with a junior expert crowd, we empirically derive task design guidelines for achieving high-quality evaluation results related to (i) the *formalism for representing ontology axioms* and (ii) *crowd qualification testing*. We find that the representation format of the ontology does not significantly influence the campaign results. Nevertheless, contributors expressed a preference in working with a graphical ontology representation. Additionally, we show that an objective qualification test is better fitted at assessing contributors' prior knowledge rather than a subjective self-assessment and that prior modelling knowledge of the contributors had a positive effect on their judgements. We make all artefacts designed and used in the experimental campaign publicly available.

CCS Concepts: • **Information systems** → *Inconsistent data*; **Data cleaning**; **Ontologies**; **Crowdsourcing**; • **Human-centered computing** → *User studies*; *User interface design*; • **Computing methodologies** → *Ontology engineering*; • **Theory of computation** → *Incomplete, inconsistent, and uncertain databases*; • **General and reference** → *Verification*;

Additional Key Words and Phrases: Ontology evaluation, human-in-the-loop, human computation

This work was supported by the FWF HOnEst project (V 754-N).

Authors' address: S. Tsaneva and M. Sabou, Vienna University of Economics and Business, Welthandelsplatz 1, Vienna, Austria, 1020 and TU Wien, Favoritenstraße 9-11, Vienna, 1040, Austria; e-mails: stefani.tsaneva@wu.ac.at, marta.sabou@wu.ac.at.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

1936-1955/2024/03-ART4 \$15.00

<https://doi.org/10.1145/3626960>

ACM Reference format:

Stefani Tsaneva and Marta Sabou. 2024. Enhancing Human-in-the-Loop Ontology Curation Results through Task Design. *ACM J. Data Inform. Quality* 16, 1, Article 4 (March 2024), 25 pages.
<https://doi.org/10.1145/3626960>

1 INTRODUCTION

The widespread adoption of artificial intelligence (AI)-based systems to support and improve nearly all aspects of human society triggered broadly raised concerns about the potential negative impacts of these systems should they fail to provide trustworthy and ethically acceptable outputs and behaviour [8]. Erroneous, biased, and unfair AI system output is often a consequence of the quality issues (incorrect facts, biased views) present in the underlying knowledge base that propagates the system output. For instance, the social-robot use case presented in [22] showcases the potential consequences of a low-quality knowledge corpus used by an AI system. In the scenario, a child communicates personal experiences and issues to a social robot that generates supportive responses relying on a knowledge base. However, if the knowledge base includes errors, the robot may provide unsuitable or inappropriate replies and thus cause emotional harm rather than support. A crucial part of avoiding such cases and creating trustworthy systems is data curation. Data curation deals with data quality issues and involves activities such as the identification and cleaning of missing and erroneous values [4].

In this article, we consider the curation of symbolic data structures that make up the knowledge corpus of AI systems. Specifically, we focus on evaluating conceptual domain knowledge structures such as ontologies, taxonomies, and knowledge graphs [7, 10]. In recent years, these resources, traditionally used to support knowledge-based systems, have been increasingly consumed as input by machine learning algorithms as well as aiming at improvement and higher interpretability of the produced outputs [2, 13, 25]. Since the quality of the input data is a crucial factor for the success of machine learning, the thorough evaluation of the utilised symbolic data structure is vital. As ontologies constitute a basis for knowledge graphs and are more complex versions of taxonomies, they are the focus of this article.

Low-quality ontologies that include incorrectly represented information or controversial concepts modelled only from a single viewpoint can lead to invalid or biased system outputs, thus negatively impacting the trustworthiness of the enabled AI system. To avoid such cases, intense work has been performed in the last decades in the area of *ontology evaluation* leading to a variety of automatic techniques (e.g., for the detection of syntax errors, hierarchy cycles, logical inconsistencies) as well as the realisation that several quality aspects (e.g., unintended use of modelling elements, incorrect domain knowledge, viewpoints) can only be tested by involving a human-in-the-loop (HiL) [26].

An example ontology evaluation aspect that requires a human contributor is the *verification of ontology restrictions* defined with universal (\forall) and existential (\exists) quantifiers. The use of these quantifiers, exemplified in Figure 1, is not trivial and often leads to ontology defects [19, 26, 27]. For instance, consider an ontology that defines *ProteinLoversPizza* through a restriction modelled with the universal quantifier as “any Pizza that has only Meat toppings”. With such modelling, a concrete pizza can be considered belonging to the class of *ProteinLoversPizzas* if (a) it has one or more *Meat* toppings and no other toppings; or (b) it has no toppings at all. Often case (b), known as the trivial satisfaction of the universal restriction, is not intended by the (junior) ontology engineer [19] and can lead to undesired system outputs (e.g., recommending a white pizza to clients actually opting for meat-filled pizzas) thus lowering users’ trust in the system. Such erroneous system outputs cannot be automatically detected; therefore, human involvement is critical for quality control of ontologies.

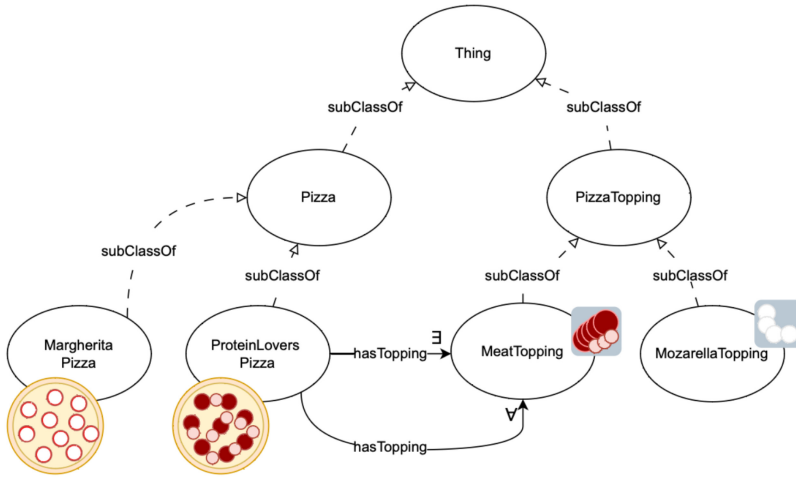


Fig. 1. A graphical representation of a simple ontology in the Pizza domain.

To reduce the cost of involving experienced ontology engineers and domain experts in ontology curation activities, human computation & crowdsourcing (HC&C) techniques offer a cost-effective alternative and have already been applied successfully for the evaluation of several ontology quality aspects [20]. For instance, HC&C approaches were employed to validate large biomedical ontologies [16], assess the quality of linked data by combining the efforts of experts and crowd workers [1], and investigate perception on viewpoints and controversial facts modelled in ontologies [3].

Yet, despite the importance of human-centric ontology evaluation as a key instrument for ensuring trustworthy AI systems, several issues remain. First, a number of ontology quality aspects, including the *verification of ontology restrictions*, has not yet been addressed with concrete HC-based applications (problem P1). Second, there are limited empirically gained insights into how to improve the results from HiL evaluation campaigns through HC task design (P2).

1.1 Contributions

We previously proposed the HERO methodology and tool support [24], which assist ontology engineers in conducting well-organised, consistent, and efficient HiL ontology evaluation campaigns and reduce time-intensive manual work. While we have already discussed the activities that are part of the proposed methodology, we extend our prior work in this article by formalising the evaluation process using a standardised notation, Business Process Model and Notation (BPMN), detailing the execution of each step of the process by describing a particular evaluation campaign and providing concrete artefacts to support a number of process steps. This leads to two main **contributions** (C) of this article:

- (C1) *Human-in-the-loop ontology restriction verification*. We exemplify the design of an HC&C-based semi-experts-sourcing application for ontology restriction verification. We chose this particular human-centric ontology evaluation aspect since it has not yet been solved with an HC-based approach. To the best of our knowledge, our HC&C approach is the first to address ontology modelling issues and distinguishes our work from prior research that primarily focused on domain factual correctness and relevance, thus addressing P1. We show that the designed HC&C application is suitable (100% accuracy achieved with a student semi-expert crowd) for leveraging human processing power as part of the ontology evalua-

tion process. Moreover, the approach acted as a tool for teaching novice ontology engineers to identify common defects and understand best practices, thus contributing to their skill to build high-quality data structures in the future.

- (C2) *Empirically gained human-in-the-loop task design guidelines*. To address P2, we conducted large-scale experiments with junior ontology engineers using the HC task from C1 to uncover task design aspects that can positively influence results from HiL ontology evaluation campaigns. In particular, we investigated whether the *representation of the ontology axioms* impacts the quality of the judgements in terms of the achieved accuracy and speed (RQ1) and whether *prior modelling knowledge* of the human contributors positively affects the campaign results (RQ2). Both task design aspects were previously lacking systematic exploration and uniform results. Thus, our investigations offer new insights for practitioners to refine and optimise their HiL ontology evaluation strategies. In terms of axiom representation (RQ1), we found no significant differences in the achieved evaluation results. However, the majority of the participants preferred a visual representation format. To assess participants' prior knowledge (RQ2), we designed a self-assessment and a complementing qualification test on ontology modelling. Our findings indicate that prior knowledge of ontology restriction models, as assessed by the qualification test, positively impacts the results of the evaluation campaign. The required level of expertise varies depending on the desired accuracy of the verifications.

1.2 Methodology

To establish the two information artefacts (C1&C2), we followed a design science methodology for information systems artefacts [6]. Figure 2 visualises how the relevance, design, and rigor cycles were realised.

- *Relevance cycle*. The relevance cycle connects new developments with the environment, including challenges and opportunities, and ensures that stakeholders' needs are considered. We have identified a number of problems (P1&P2 in Figure 2) relevant for the intersection of the semantic web and human computation domains, which we address with artefacts C1 and C2.
- *Rigor cycle*. The rigor ensures that new developments are grounded on existing theories. To that end, we adhere to the principles of experimental investigations in software engineering from Wohlin et al. [29] for the conducted experiments and follow the HERO [24] methodology for describing the evaluation process.
- *Design cycle*. The design cycle is the main cycle in design science projects, since it addresses the development and evaluation of the proposed artefacts. We conduct two experimental investigations for the evaluation of the designed HiL ontology verification approach and the investigated task design aspects.

The rest of the article is structured as follows. Section 2 gives an overview of related work in the area of human-in-the-loop ontology evaluation and Section 3 provides a summary of our previously designed HERO methodology. In Sections 4 to 6, we describe the HC-based ontology restriction verification (C1) and the conducted experimental investigations (C2) in the context of the HERO methodology. We conclude with a summary and closing remarks in Section 7.

2 RELATED WORK

Ontology quality issues and curation methods have been abundantly addressed in the literature for more than 20 years. McDaniel and Storey [14] reflect on the work in the ontology evaluation

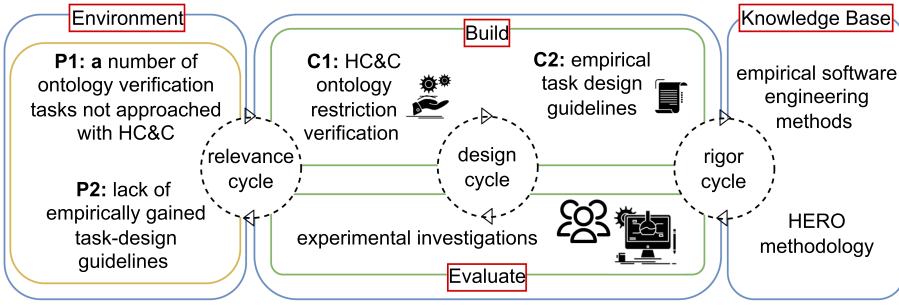


Fig. 2. Design science-based methodological approach.

domain from papers published in the last two decades and identify that semantic mistakes cannot always be automatically detected. While automatic methods might be fast and scalable, they also have their limitations that need to be addressed by human involvement.

2.1 Human-in-the-Loop Ontology Evaluation

Several papers have already identified ontology evaluation tasks that require human input [5, 18, 26, 28] and HC&C have been successfully applied for solving a portion of these identified tasks. For instance, Mortensen et al. presented a crowd-based verification of taxonomic relationships from a medical ontology [16]. The authors determine that the crowd performed almost as well as a single expert in identifying errors in ontology relations and can thus be used when the budget is limited or an expert is not available. In [3], the authors explore subjectivity modelled in ontologies. They argue that experts tend to build ontologies based on their personal beliefs and experience and thus apply crowdsourcing to investigate perception on viewpoints and controversial facts modelled in ontologies. Ontology enhancement achieved by crowdsourcing is investigated in [11], which also examines how non-experts can improve and complete ontology taxonomic knowledge. In [9], a framework for the syntactic and semantic quality evaluation of enriched ontologies is presented. The semantic evaluation relies on a crowdsourced approach in which crowd workers can agree (or disagree) with each proposed ontology enrichment. Despite the intense work of using human computation and crowdsourcing for various ontology evaluation tasks, as discussed above, there is a lack of a guidelines for creating optimal human-in-the-loop solutions for the evaluation of ontologies. Moreover, numerous tasks have not yet been addressed with HC-based applications, especially concerning modelling mistakes, such as the verification of ontology restrictions. Contribution C1 of this article, the application for human-in-the-loop ontology restriction verification, addresses this gap in the literature.

2.2 Worker Qualification for Human-in-the-Loop Ontology Evaluation

In the crowdsourcing domain, the qualification of workers and its influence on crowdsourced results has been abundantly discussed. However, the required qualification for semantic web verification tasks still requires further investigation. In [15], the authors performed a number of experiments for verifying hierarchical ontology relations from the medical domain. They showed that workers who pass a domain-specific (biology) qualification test perform best. In a follow-up study [17], they compared the effects of three qualification tests: two domain-specific tests (biology, medicine) and one test on the ontology domain. In contrast to their previous investigations, qualified workers did not perform better than others since they may have relied on intuition rather than the presented context.

In [27], the authors investigated the comprehension of ontology axioms. The qualification of the participants was subjectively measured based on a self-assessment test covering several knowledge categories (e.g., formal logics) and participants had to rate their expertise into the levels *no/little/some/expert* knowledge. The results indicate that knowledge in some areas (e.g., descriptive logics) could reduce the work time and increase the accuracy of the results. However, it is not clear what knowledge each expertise level (*no/little/some/expert*) covers and how an objective qualification test can be created adhering to the same principles.

While the qualification of crowd workers has been previously addressed in a number of studies, conclusions are indeterminate for ontology-related tasks. Additionally, details on how qualification tests are designed are omitted, the full tests are not made publicly available, and it is unclear whether they are reused for follow-up experiments. We address these limitations through contribution C2 – the empirically derived task design guidelines – by (i) designing a self-assessment test with justifications of what the knowledge levels (*no/little/some/expert*) entail; (ii) designing a qualification test with several expertise sections allowing us to categorise contributors objectively into the same knowledge levels as above; (iii) performing two experiments to test whether background modelling knowledge is beneficial for the acquired ontology verifications and whether a self-assessment or an objective qualification test is better fitted; (iv) providing both the subjective and objective assessments online for fellow researchers to reuse.

2.3 Ontology Representation

Ontologies are often represented in OWL,¹ RDF,² or other languages based on descriptive logics. Nevertheless, it is a challenge for novice ontology engineers to fully comprehend the meaning of axioms in such knowledge representation languages [19, 27]. Thus, ontology axioms are typically translated into natural language for the crowdsourced verification tasks, making them understandable also for lay users and semi-experts. In [19], the authors proposed the rephrasing of OWL ontology axioms into text, including keywords such as *amongst other things* to represent the open world assumption, and *some* and *only* to represent the existential and universal ontology restrictions (see Figure 14(a), Appendix A). The authors argued that, based on their experience, this phrasing has been shown to be effective in teaching novice ontology engineers (students). Nevertheless, we are not aware of an experimental investigation showing such effects.

In [27], Warren et al. perform experiments on the comprehension of axioms written in knowledge representation languages. They argued that natural language can be ambiguous and interpreted in multiple ways. In the performed experiments, the effect of used keywords in the representations of axioms is investigated. They showed that replacing *some* and *only* with *including* and *none or only*, respectively, to represent ontology restrictions can improve accuracy for simple restrictions. Warren et al. concluded that there are more investigations of the used keywords needed. For instance, they suggest for the representation of restrictions that the comprehension can be improved when substituting the original keywords by *at least one* and *none other than*, motivated by the explanation of description logics in [12]. Therefore, we compare this additional alternative phrasing from [27] against the representation from [19], and a graphical representation as part of contribution C2 – the empirically gained human-in-the-loop task design guidelines.

To summarise, in this article, we advance the state-of-the-art in terms of (C1) a HC-based application for the verification of ontology restrictions; and (C2) an experiment of HC task design

¹Web Ontology Language <https://www.w3.org/OWL>

²Resource Description Framework, <https://www.w3.org/RDF>

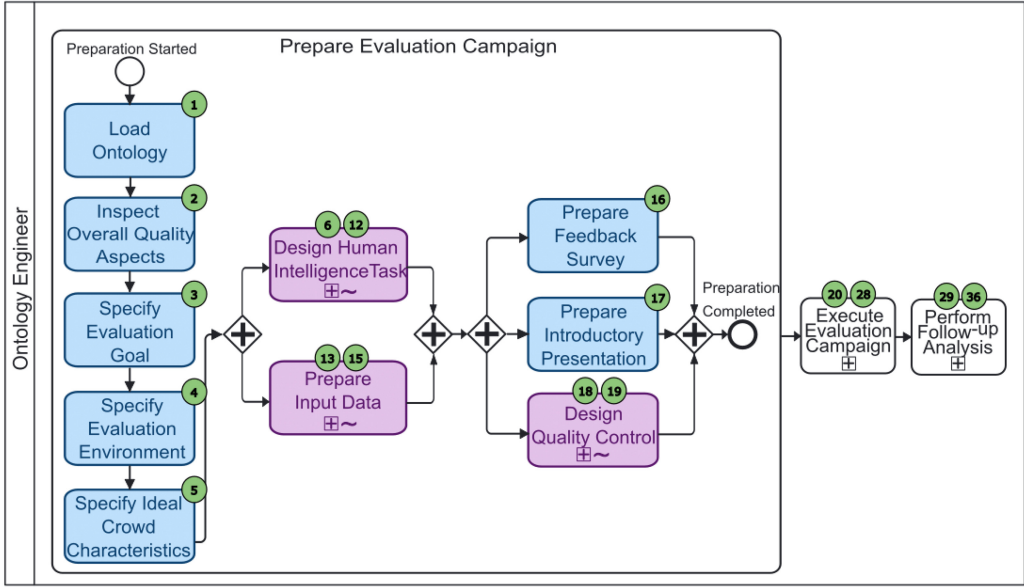


Fig. 3. Typical preparation stage of a HiL ontology evaluation process.

aspects for establishing guidelines on ontology representation and crowd qualification as part of the design phase of trustworthy AI systems.

3 BACKGROUND: THE HERO METHODOLOGY

With the aim of reducing the efforts needed to prepare, execute, and analyse human-centric ontology evaluation studies, we previously introduced HERO [24]. HERO is a process that guides ontology engineers through the relevant steps to be taken in HiL ontology evaluation campaigns. Thus, it can reduce the likelihood of crowdsourced errors or biases and improve the reliability of the results at the design phase of the campaign. The activities, part of the HERO methodology, can be divided into three main stages: preparation, execution, and follow-up analysis of the evaluation campaign. We formalised the process³ using Business Process Model and Notation (BPMN). In this section, we summarise each stage and the activities it includes.

It should be noted that HERO is designed with the goal of achieving broad applicability. Therefore, it encompasses activities⁴ that may not be relevant to every human-centric ontology curation campaign. The methodology covers a number of micro task-style approaches, defined through a variation of (a) the used software platform and (b) the type of human contributors involved. As such, HERO can be followed both when conducting evaluation campaigns utilising crowdsourcing platforms and lay users as well as when an expert evaluation is to be performed.

3.1 Preparation of the Evaluation Campaign

Figure 3 illustrates the activities within the preparation stage of HERO. The process begins with the selection of an ontology to be evaluated, its assessment in terms of overall quality aspects, and a specification of the aim of the evaluation (1–3 in Figure 3). To collect high-quality campaign results, the evaluation environment (4; e.g., crowdsourcing platform, games with a purpose, custom

³The resource is available at <https://doi.org/10.5281/zenodo.7643357>

⁴Depending on the evaluation campaign, some of these activities can be merged and completed simultaneously.

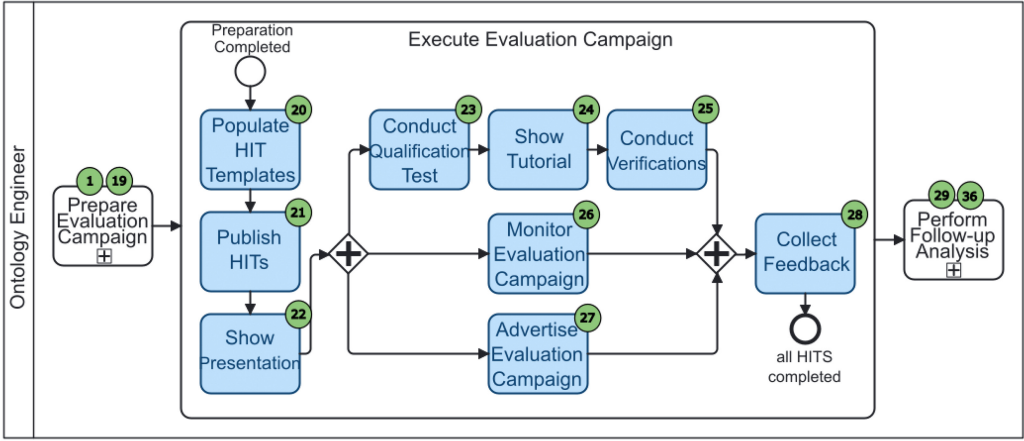


Fig. 4. Typical execution stage of a HiL ontology evaluation process.

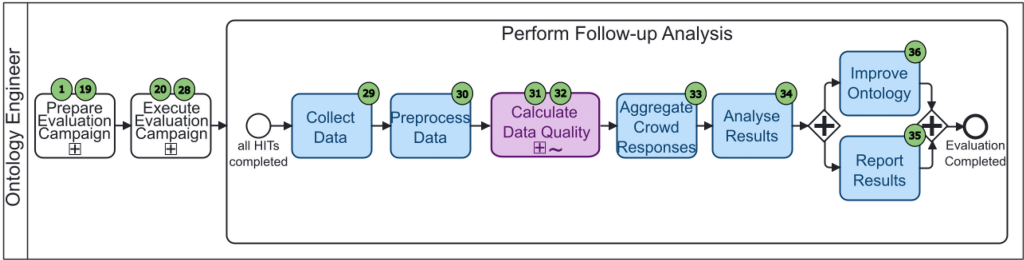


Fig. 5. Typical follow-up stage of a HiL ontology evaluation process.

interface, etc.) and crowd (5; e.g., lay users, skilled crowd workers, internal expert crowd, etc.) should be selected based on the evaluation aspect and size of the ontology. Once the aims and requirements of the evaluation campaign are set, the Human Intelligence Tasks (HITs) should be designed (6–12) and in parallel the input data should be prepared (13–15). For details of the separate activities that each of these iterative sub-processes includes, please refer to [24].

In some evaluation campaigns, it can be meaningful to prepare a feedback survey (16) for the crowd to fill in based on which task design aspects can be improved. Additionally, depending on the selected crowd, an introductory presentation can be prepared (17). To ensure high-quality results, quality control can be designed, i.e., creating training questions and seeding in control questions (18 and 19).

3.2 Execution of the Evaluation Campaign

Once the evaluation preparation is completed, the campaign can be conducted. First, the designed HITs are populated with the input data (20) and published (21 in Figure 4) and if a presentation was prepared it is shown to the crowd (22). Next, the actual campaign begins, i.e., the qualification testing is performed (23), a tutorial follows, i.e., the completion of the prepared training questions (24), and the verification tasks are completed by the crowd (25). In parallel, the campaign is continuously monitored (26) and advertised further if needed (27). Once all HITs are completed, feedback from the crowd can be collected (28).

3.3 Performing Follow-up Analysis

The next stage of the HERO methodology is the processing of the gathered results. The crowd responses are collected (29 in Figure 5) and preprocessed (30). Data quality statistics can be calculated next, such as trustworthiness of individual evaluators and inter-rater agreement (31 and 32). Next, workers' results should be aggregated (33) and the results analysed to obtain the final evaluation (34). Lastly, the evaluation results should be reported (35). The report can be used to improve the initial ontology (36).

In Sections 4 to 6, we follow HERO to describe the ontology restrictions verification campaign we conducted to systematically explain how it was carried out and to provide details on how the HERO activities can be addressed.

4 HERO STAGE 1: PREPARING AN HC-BASED ONTOLOGY RESTRICTION VERIFICATION CAMPAIGN

Following the HERO methodology, we describe the HC-based approach for verifying ontology restrictions (contribution C1) and the empirical investigation of two task design aspects and their influence on the quality of HiL evaluation campaign results: (1) the representational format of the ontology and (2) crowd qualification testing (contribution C2). In this section, we focus on the preparation of the evaluation campaign; Section 5 describes the execution of the study. Lastly, in Section 6, we discuss the follow-up analysis and the achieved campaign results.

4.1 Ontology (1 in Figure 3)

For our study, we used the Pizza Ontology,⁵ which is a known, high-quality educational ontology that contains many structures with the universal and existential quantifiers. In [19], the authors argue that this is also the most successful ontology in teaching Western audiences about ontology restrictions and common good practices. A simplified subset of the ontology is visualised in Figure 1.

4.2 Evaluation Goal (3 in Figure 3)

With the HiL curation campaign, we aim at investigating a human-computation-based approach for verifying ontology restrictions and the results that can be achieved with the proposed solution. Additionally, we look into several aspects of the task design, which are needed for future experiment developments. The focus lies on whether the representation of ontology axioms and prior modelling experience of the human contributors affect the quality of the collected results. Concretely, the influence of those two factors on the accuracy of the provided judgements as well as the speed of the verifications was observed. The following research questions are formulated:

RQ1: How does the representation of the ontology axiom impact the quality of the judgements in terms of the achieved accuracy and speed? It is important to investigate which formalism is best fitted for a HiL task design to guide researchers on how to represent ontologies to achieve high-quality results in future HiL evaluations.

RQ2: How does prior modelling knowledge of the human contributors impact the quality of their judgements in terms of accuracy and speed? A positive effect of prior modelling experience, measured with a self-assessment, has already been shown in [27]. We aim at gathering additional experimental data to validate those findings. It is essential to investigate this task

⁵<https://protege.stanford.edu/ontologies/pizza/pizza.owl>

design aspect with the intention of providing clear guidance on the needed skills of the contributors for future experiments. RQ2 investigates which background knowledge areas positively influence the verification results and how an objective qualification test can be designed.

To investigate and answer RQ1 and RQ2, we formulate the alternative hypotheses:

H1: The formalism in which ontology axioms are represented has a statistically significant influence on the performance and speed of the contributors.

H2: Prior modelling knowledge has a statistically significant positive influence on the performance and speed of the contributors.

4.3 Evaluation Environment (4 in Figure 3)

As a platform for performing the verification tasks, the sandbox of Amazon Mechanical Turk⁶ (mTurk) was used—a crowdsourcing platform that offers the possibility to harness the wisdom of the crowd. Requesters implement their outsourced work assignments as jobs and each job contains several HITs that are simple and independent pieces of work that can be solved by a global workforce—the human contributors, also called *workers*.

While we chose to use an internal semi-expert crowd for the experiment (see Section 4.4), we still utilised a crowdsourcing platform for the evaluation campaign for several reasons: (1) mTurk randomises the order of HITs within a job and thus deals with possible sequence bias; (2) the platform offers the possibility to skip a HIT and return to it later, thus contributing to the collection of better-quality results; and (3) by using a readily available platform, we reduce the time needed for the preparation of the campaign by avoiding the implementation of a custom interface.

4.4 Crowd Characteristics (5 in Figure 3)

The conducted experiment relied on a student crowd rather than a layman crowd for several reasons. First, to investigate RQ1 and RQ2, it was required that a portion of the participants have background knowledge in modelling and some understanding of graphical representations. Such a crowd can be difficult to collect in a crowdsourcing platform with layman contributors. Second, working with a student crowd allows for a more controlled environment. We prepared a self-assessment and a qualification test, which each participant completed, allowing for a better understanding of the participants' prior knowledge. Third, the experiment offers novice ontology engineers an environment to test their knowledge and understand common ontology engineering mistakes. In total, 88 masters' students taking an introductory course to semantic systems participated in the experiment.

4.4.1 Self-assessment. The study participants first needed to complete a self-assessment, in which they rate their knowledge in different knowledge areas (English skills, formal logic, ontology modelling, model engineering, web-based representational languages) into the categories *no/little/some/expert knowledge*. The same knowledge categorisation was used in the self-assessment test from [27]. However, we additionally specify for each category what the knowledge levels entail. An example question and the defined expertise levels for the ontology modelling domain are shown in Figure 6. The complete self-assessment is published as a Zenodo resource.⁷

4.4.2 Qualification Test. With the purpose of evaluating the knowledge of the participants objectively, a qualification test was designed that only targets the ontology modelling knowledge of

⁶<https://www.mturk.com>

⁷The resource is available at <https://doi.org/10.5281/zenodo.7643357>

Ontology Modeling Skills

For the questions below, please consider the following levels:

1 - no knowledge = I have no knowledge in the area.

2 - little knowledge = I am aware of the basic components of ontologies and can recognise them in graphical and textual representations.

3 - some knowledge = I have an understanding of the implications of ontology axioms and restrictions.

4 - expert knowledge = I can perform reasoning with ontology models, as well as compare and relate them to each other.

Q5: How would you rate your knowledge in ontology modeling? *

	1	2	3	4	
no knowledge	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	expert knowledge

Fig. 6. Example of a question from the self-assessment test and the provided knowledge expertise definitions.

the students with a focus on the understanding of universal and existential quantifiers. The qualification test complements the self-assessment and is needed because participants have a subjective view of their own competencies. The test includes 9 questions of different difficulty levels, and has a maximum score of 11 points. Based on the acquired score, the participants can be objectively sorted into the same categories as above: *no/little/some/expert knowledge*. For instance, one of the *little-knowledge*-difficulty questions required participants to be able to recognise classes and relations given an axiom as shown in Figure 7. To be sorted into the *some knowledge* category, the participants had to understand the meaning of a simple axiom. An example question from this category is visualised in Figure 8. For the expert category, the ability to compare two axioms was required (see Figure 9). The sorting of the participants into the expertise categories had the following criteria:

(1) *no knowledge (novice)*: scored at most 3/11 points overall.

(2) *little knowledge (beginner)*: scored 3/4 points on the little-knowledge section or 4/11 points overall.

(3) *some knowledge (intermediate)*: scored 2/3 points on the some-knowledge-section and at least 5/10 overall, or 7/11 points overall.

(4) *expert knowledge (expert)*: scored 3/4 points on the expert-knowledge-section and 10/11 points overall.

By sorting the participants into different categories instead of simply labelling them as qualified/not-qualified, we are able to investigate the verification accuracy each expertise group can achieve.

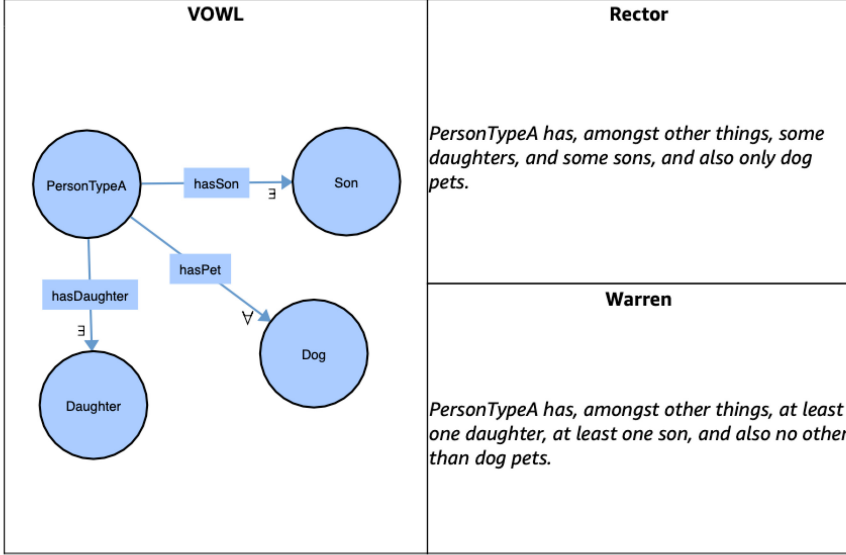
As seen in Figures 7 to 9, each question includes three ontology axiom representations, which we investigated in the experiment, aiming to avoid the introduction of bias in regards to the representation of the restrictions. The complete qualification test is available as a Zenodo resource.⁸

4.5 Input Data Preparation (13–15 in Figure 3)

4.5.1 Ontological Elements (13, 11). From the ontology, all existential and universal restrictions are extracted automatically and then grouped together on the same relation, forming ontology restriction axioms (ORAs). Each such axiom represents a small ontology that fully describes a

⁸The resource is available at <https://doi.org/10.5281/zenodo.7643357>

Consider the model, represented in 3 equivalent formalisms (VOWL | Rector | Warren) and answer questions 1 & 2 below.



1. Identify the main model components from the model

How many named classes can you identify from the model?


How many relations can you identify from the model?

Fig. 7. Example of a question from the little-knowledge section of the qualification test on ontology restriction modelling.

specific relation and can be evaluated independently from the rest of the axioms (see Figures 14 and 15, Appendix A, for examples of such an axiom represented using different formalisms). To allow for defect detection, the ontology was manually seeded with various defects, resulting in 15 correct and 15 incorrect ORAs used in the experiment.

4.5.2 Context (14). Since ontologies are often reused and extended during their lifetime, there is usually no specification document to compare them against. Instead, human domain knowledge is required to verify whether the model elements correctly represent the real-world domain entities they describe. However, previous research [15, 17] has shown that providing the evaluators with enough domain information has positive effects on the acquired verification accuracy. Therefore, to support the human participants in the ontology verification task, each ORA to be evaluated is matched to a real-world context entity EV_{ORA} , which acts as evidence and guides the human workers in their decision on whether the ORA is valid. Furthermore, the context entity is (i) representative of the ORA and (ii) small enough to be a part of an HC task, while (iii) providing enough context to the contributors, as suggested in [21]. In the performed campaign, an image is chosen for presenting the EV_{ORA} (see Figure 10, area 1).

Consider the model, represented in 3 equivalent formalisms (VOWL | Rector | Warren) and answer question 3 below.



<p>VOWL</p> 	<p>Rector</p> <p><i>PetLoverTypeA has, amongst other things, only Dog pets.</i></p> <hr/> <p>Warren</p> <p><i>PetLoverTypeA has, amongst other things, no other than Dog pets.</i></p>
--	--

3. Select the statement that describes instances of PetLoverTypeA correctly.

- ☐ Instances of PetLoverTypeA must have a Dog pet and cannot have other types of pets.
- ☐ Instances of PetLoverTypeA might not have a Dog pet and cannot have other types of pets.
- ☐ Instances of PetLoverTypeA must have a Dog pet and can also have other types of pets.
- ☐ Instances of PetLoverTypeA might not have a Dog pet and can also have other types of pets.

Fig. 8. Example of a question from the some-knowledge section of the qualification test on ontology restriction modelling.

Consider models A and B describing PetLoverTypeG and PerLoverTypeF, each represented in 3 equivalent formalisms (VOWL | Rector | Warren) and answer question 8 below.

Model A: PetLoverTypeG	Model B: PerLoverTypeF
<p>VOWL</p> 	<p>VOWL</p> 
<p>Rector</p> <p><i>PetLoverTypeG has, amongst other things, only pets that are not Dogs.</i></p>	<p>Rector</p> <p><i>PetLoverTypeF has, amongst other things, only Dog pets.</i></p>
<p>Warren</p> <p><i>PetLoverTypeG has, amongst other things, pets that are no other than not Dogs.</i></p>	<p>Warren</p> <p><i>PetLoverTypeF has, amongst other things, no other than Dog pets.</i></p>

8. Is it true that PetLoverTypeG is disjoint to PetLoverTypeF? That is, there can be no instance that is at the same time of type PetLoverTypeG and PetLoverTypeF.

- ☐ Yes
- ☐ No

Fig. 9. Example of a question from the expert-knowledge section of the qualification test on ontology restriction modelling.

4.6 Human Intelligence Task Design (6–12 in Figure 3)


In the field of ontology evaluation, a number of human intelligence tasks have been established, primarily for assessing factual correctness or domain relevance. Nevertheless, no HIT design has been tailored for the verification of ontology modelling decisions.

See Instructions

5

instructions on the correct usage of ontology restrictions

Please make sure you are familiar with the rules and examples provided in the **Instructions** before answering the question.

Pizza Menu	Model
 <p>ROSA V P</p> <p>Gorgonzola, Mozzarella, Tomato</p>	<p>Rosa pizzas have, amongst other things, some Tomato topping, and some Mozzarella topping, and some Gorgonzola topping, and also only Gorgonzola, Mozzarella, and/or Tomato toppings.</p>

1

context entity (E_{VORA}) in a representational format of choice

2

ontology restriction axiom (ORA) in a representational format of choice

Does the model represent the pizza menu item correctly ?

☐ The model correctly represents the menu item.
☐ For the model to correctly represent the menu item, one or more existential (some) restrictions need to be added.
☐ For the model to correctly represent the menu item, one or more universal (only) restrictions need to be added.
☐ For the model to correctly represent the menu item, one or more universal (only) restrictions need to be replaced by existential (some) restrictions.
☐ For the model to correctly represent the menu item, one or more existential (some) restrictions need to be replaced by universal restrictions (only).

Comment (optional)
In case you have any remarks please add them here

4

area for comments and other remarks

3

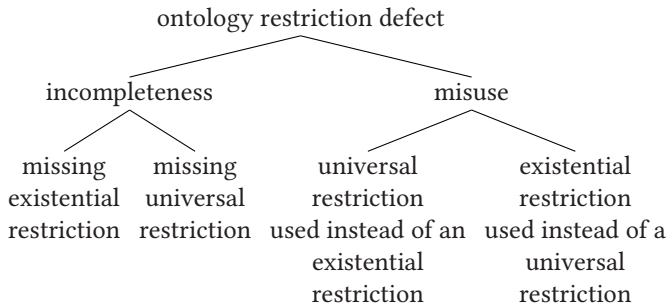
verification options corresponding to a defect taxonomy

Submit

Fig. 10. Example of a Human Intelligence Task (HIT) for the verification of ontology restrictions.

In the following, we introduce the novel design of a human intelligence task for verifying the correct usage of ontology restrictions. Some of the HIT design aspects are grounded on principles established by prior research on ontology verification tasks (i.e., context and instructions inclusion, e.g., [17]). Further design decisions are inspired by HC tasks in the Software Engineering domain, such as providing answer options based on a predefined modelling defect set [21]. In addition, the HIT design allows for the investigation of new task design aspects by supporting various formats of the included elements (e.g., presentation modality, context modality).

4.6.1 HIT Question Format and Answer Options (7,8). For the evaluation campaign, we decided to use close-ended questions. For the definition of possible answer options, we identify types of defects typical to the usage of the existential and universal quantifiers and organise them into a defect taxonomy. The taxonomy informs the HIT design so that the evaluators are guided through their tasks and supported in determining the most plausible defects. Below, the identified defects for the particular use case are shown. The four possible defects for this kind of verification can be also seen in the corresponding HIT (Figure 10, area 3).



4.6.2 Presentation Modality (10). There are different possibilities of how ontologies can be formalised. While ontologies are most commonly represented in OWL and RDF, HC&C approaches usually use a description of the ontological elements in natural language. While natural language description could be easier to understand for lay users, graphical representations might be helpful for those who have previous experience with model engineering. A well-known visual representation of ontologies is VOWL.⁹ For the experimental investigation of RQ1, we consider two textual representations, the first proposed by Rector et al. [19] and its suggested alternative phrasing by Warren et al. [27], as well as the graphical representation VOWL. Figures 14 and 15 (Appendix A) show how a Margherita Pizza defined in OWL can be paraphrased into the Rector and Warren formalisms or visualised as a VOWL graph.

4.6.3 HIT User Interface (6). For each ORA and a context entity EV_{ORA} , a HIT allows for the detection of various defect types. The task design follows the micro-tasking approach of splitting the complex problem of evaluating the quality of an ontology into smaller verification tasks focused on a single ontology axiom at a time. In Figure 10, we see an example of a HIT, in which the EV_{ORA} (1 in Figure 10) is shown as an image of a pizza menu item. On the right side of the HIT we have the ORA (2), which is represented in a formalism of choice. Here, a textual formalism proposed in [19] is used. The evaluator needs to decide whether the ontology axiom correctly represents the real-world entity and select one of the provided verification options (3). Each answer option presents a possible scenario of model changes and also corresponds to a defect from the defined defect taxonomy to allow easy aggregation and evaluation of the results. Evaluators can also leave free-text comments (4), allowing for the possibility that new defects are identified or ambiguities in the question design or axiom representation are established.

4.6.4 HIT Instructions (9). Since it is important to ensure that the contributors have enough context to make a correct decision, the modelling theory behind ontology quantifiers is provided in an instruction panel (5 in Figure 10), available throughout all verification tasks. The instructions contain definitions and descriptions adopted for the selected formalism in which the ORAs are presented, and also offer examples of correct and incorrect modelling choices with justifications. The designed tasks used in the evaluation campaign are published as a Zenodo resource.¹⁰

4.6.5 Follow-up Scripts (12). Before finalising the task design, we performed a pilot study within our research group to (a) identify aspects that can be improved and (b) ensure that the task design allows for the collection of all data required at the analysis stage. For this purpose, we prepared initial analysis scripts.

4.7 Data Quality Control (18 and 19 in Figure 3)

Training questions (18) were prepared in the form of a tutorial job in order to allow the participants to get to know the question format as well as mTurk better before working on the verification jobs. This was an important part of the campaign, since we relied on a student crowd rather than mTurk crowd workers. The tutorial had the same structure as the verification jobs; however, the data was from a different domain (Wine Ontology). Including tutorial questions ensures that all participants acquire the basic knowledge needed to use the crowdsourcing system prior to the actual experiment.

⁹Visual Notation for OWL Ontologies, <http://vowl.visualdataweb.org/v2/>

¹⁰The resource is available at <https://doi.org/10.5281/zenodo.7643357>

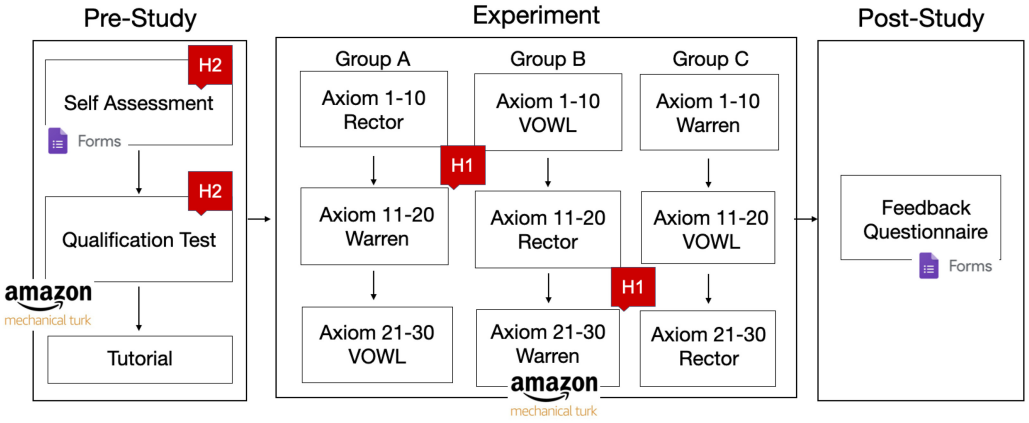


Fig. 11. Overview of the experiment workflow and its relation to the research hypothesis.

4.8 Introductory Presentation and Feedback Survey (16 and 17 in Figure 3)

We prepared a *feedback questionnaire* (16) to be filled in by each study participant, aiming to determine whether the experiment was designed well and how it was perceived by the participants. We believe that including the workers' opinions when defining guidelines for the optimal design of human-centric (ontology verification) tasks is crucial and can impact the success of future HC campaigns.

Our crowd included students, who participated on a voluntary basis and could receive extra credit for their class. Therefore, we prepared an *introductory presentation* (17) explaining what the campaign's goal is and additional organisational aspects.

5 HERO STAGE 2: EXECUTION OF THE ONTOLOGY EVALUATION CAMPAIGN

After *populating* the prepared HIT templates, *publishing* all jobs on mTurk, and *presenting* the experiment to the students (20–22 in Figure 4), we proceeded with the main part of the experiment. The flow of the experiment and all included stages are shown in Figure 11. The participants were given 2 hours to complete all parts of the experiment (pre-study, experiment, and post-study). Each student performed the tasks at home at a specified time. During the experiment, a Zoom meeting was active to *monitor the campaign* and solve organisational aspects and technical issues with the platform (26 in Figure 4).

As part of a pre-study, the participants completed the prepared *self-assessment*, the *qualification test*, and the prepared *tutorial* (23 and 24 in Figure 4). To conduct the *verification experiment* (25), three use cases — A, B, and C — were defined. Each student was assigned to one of the scenarios, forming three equally sized groups of students. Each use case involves 3 verification jobs, containing 10 HITs each and using a different formalism to represent the ORAs. The need for investigating three scenarios comes from the investigation of hypothesis H1. To be fair and unbiased, each group sees the same task sections in the same order; however, the ORAs are shown in a different formalism. Tasks from the same modelling formalism are grouped together in jobs to lower the cognitive overhead of switching between different representations for the workers. As shown in Figure 11, Group A started working on the first job seeing the ORAs in the Rector formalism, continued working on the second section in a formalism proposed by Warren, and finished with the VOWL formalism in the last job. Group B started with the same job, but saw the ORAs in the VOWL formalism and so on. The HITs within each job are automatically randomised by mTurk

for each participant to make sure that some questions are not overlooked and the sequence bias is avoided. Lastly, the *feedback questionnaire* was filled in by each study participant (28 in Figure 4).

6 HERO PHASE 3: FOLLOW-UP ANALYSIS OF THE EVALUATION CAMPAIGN

Once all HITs were collected, the mTurk responses were collected and preprocessed to allow for the analysis of the achieved results. Each submitted HIT results in an individual judgement of the ORA's correctness. By aggregating all collected judgements on a particular ORA, the final defect type can be identified and an ontology defects report can be created (29, 30, and 33–35 in Figure 5). In this section, we discuss the obtained results from the evaluation campaign. In Section 6.1, we give an overview of the gathered results. In Section 6.2, we look in detail into the two defined hypotheses. Lastly, in Section 6.3, we discuss the results from a replication study.

6.1 Overall Results

In total, 2629 student responses were gathered and each verification task received approximately 28 to 30 responses. Overall, 92.58% of those responses were correct and a single judgement took on average about a minute (56.79 seconds). When aggregating the results with a relative majority voting strategy, a 100% accuracy is reached, leading to the correct verification of each ORA. These results show that the proposed HC-based method can be applied to gather high-performance results on the task of verifying ontology restrictions.

6.1.1 Population Qualification. In order to understand the results of the experiment and how they were achieved, it is important to look into the qualification of the evaluator population prior to the study and their distribution into the three groups. Figure 12 shows the participants' knowledge in each of the groups that the students were separated into based on their self-assessments in different areas as well as their scores on the qualification test. From the graphics, we see that beginners and more experienced students were relatively equally distributed into the three student groups (A, B, and C). The experiment workflow did not allow for this split to be intentionally made; however, this distribution is favourable for the interpretation of the results.

6.1.2 Formalism-Based Results.

Important for *RQ1* are the achieved results based on the formalism, in which the ORAs were presented. Table 1 shows the results from the evaluation tasks based on the representation of the axioms. For each formalism, the average percentage of correct responses per HIT and the average verification time per HIT are provided.

It can be seen that while the results

Table 1. Results of the Initial Experiment in Terms of Accuracy and Speed Based on Ontology Representational Formalism and Feedback from the Participants

formalism	avg correctness per HIT	avg time per HIT	student preference
Rector	92.28%	55.69 s	10%
VOWL	93.76%	53.88	74%
Warren	91.74%	57.79 s	16%

are very similar, the verifications performed in the VOWL formalism have slightly higher accuracy and the average time needed for evaluating an axiom is lower than in the textual representations. Another important factor to consider is what formalism was easiest to understand from the perspective of the evaluators. Based on the feedback provided during the post-study, the majority (74%) of the participants preferred the graphical VOWL representation to the textual paraphrasing of the axioms. Moreover, based on the insights gathered from the comment option in the HIT, it becomes clear that some wording of the textual formalisms was hard to understand for students (e.g., the textual representations of the union shown as *and/or* and the meaning of *amongst other*

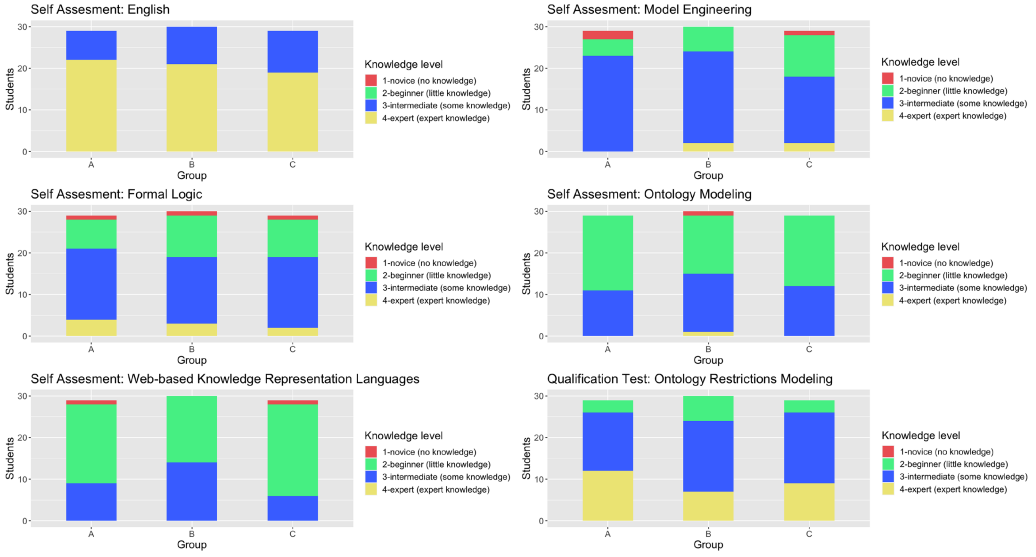


Fig. 12. Background knowledge of the experiment participants in different areas based on the self-assessment and qualification test.

things in the context of the ORAs). These findings reinforce the claim from [27] that natural language is ambiguous and keywords used to describe restrictions should be further investigated.

6.1.3 Qualification-Based Results. To explore *RQ2* we looked at the point-biserial correlation between the verification accuracy and speed and the background knowledge in each of the assessed areas and found that the prior knowledge levels detected with the qualification test has the highest influence on the accuracy of the judgements of the contributors with a correlation value $r(86) = 0.35$. We also found that prior knowledge in model engineering can slightly decrease ($r(86) = -0.22$, $p\text{-value} = 0.04$) the time needed for performing the verification.

Additionally, we look at the correctness of verifications from each expertise group to assist researchers in determining an appropriate qualification threshold based on the target accuracy of the verification campaign (Table 2). Our qualification test classified none of the students as novice (no knowledge), 12 as beginners (little knowledge), 48 as intermediate (some knowledge) and 28 as experts (expert knowledge). Beginners' responses had an overall correctness of 86.59%, intermediates submitted 91.92% correct verification, while expert judgements reached correctness of 96.18%. When aggregating the responses using a relative majority vote, counting verifications without a clear plurality as incorrect, beginners performed with 93.33% accuracy, intermediates 100%, and experts 98.89%. The average time needed to complete the judgements also differs between the three expertise groups. Beginners took 60.6 s for a single response on average, intermediates 55.85 s, and experts 55.03 s. Surprisingly, when responses were aggregated, intermediate raters' votes outperformed expert-level judgements, even though individual expert judgements initially exhibited the highest quality. These findings suggest that multiple less-experienced raters can effectively replace an expert crowd for certain tasks once their responses are combined.

In Table 2, we also report the (defect type-)weighted averages of the precision, recall and F1-scores. Since the aggregated votes of intermediates and experts reached very high accuracy scores, the precision, recall, and F1 metrics also yield (almost) perfect scores. However, beginners' scores indicate a slightly higher precision than recall, meaning that they sometimes overlooked defects in the ORAs.

Table 2. ORA Verification Scores Achieved by Each Expertise Group

Crowd Characteristics		Scores over all collected judgements		Majority Vote Metrics			
Qualification	N	Correctness	Avg Time per Judgement	Accuracy	Weighted Avg		
					Precision	Recall	F1
all participants	88	92.53%	56.79s	100%	100%	100%	100%
beginner	12	86.59%	60.6s	93.33%	94.63%	92.22%	93.11%
intermediate	48	91.92%	55.85s	100%	100%	100%	100%
expert	28	96.18%	55.03s	98.89%	98.91%	98.89%	98.87%

6.1.4 Inter-rater Agreement. To calculate the inter-rater agreement among the participants, we calculate a Krippendorff’s alpha coefficient for each study group (A, B, and C). The alpha scores vary slightly between the groups, with Group A having $\alpha = 0.802$, Group B 0.759, and Group C 0.857. Overall, these scores indicate a substantial to high level of agreement among the raters and a good reliability of the gathered results.

6.1.5 Performance Changes over Time. An additional interesting aspect to be explored based on the experiment results is whether the participants learned the “patterns” of the ontology axioms and included defects over time and whether this changed their performance in terms of time needed to perform a single verification and the accuracy of their results.

We found that there is no significant improvement of verification quality over time (Pearson correlation coefficient $r(2618) = 0.074$, p -value = 0.71). Since the verification accuracy was at a high level from the beginning, further improvements can be challenging to achieve, and small changes are not statistically significant. Nevertheless, we observed that the time students needed for the verification of each successive task decreased slightly ($r(2618) = -0.25$, $p = 2.2e-16$) as they got exposed to more HITs.

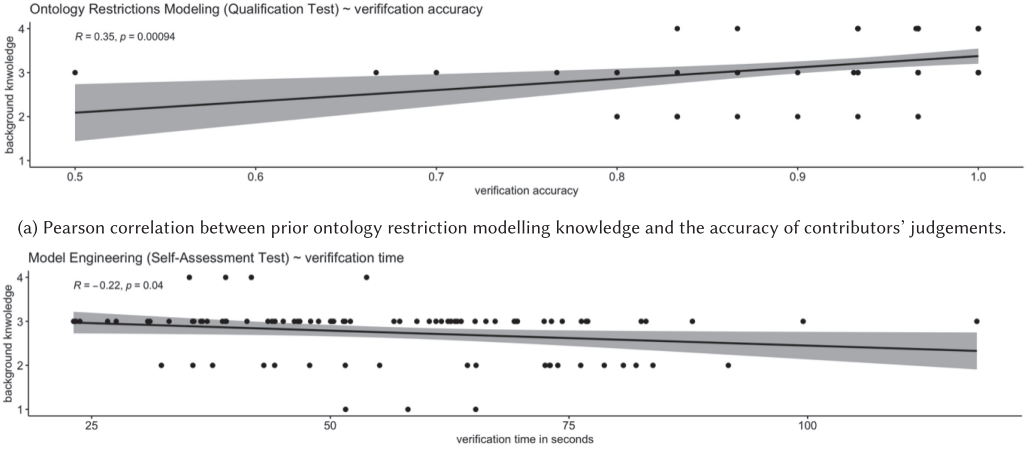
These results suggest that while the accuracy of verification remained consistently high, students became more efficient at completing the tasks as they gained experience.

6.1.6 Participants’ Feedback. The proposed HC-based verification of ontology restriction was well received by the students performing the verification jobs and proved to be very useful as part of distance learning. Some of the received comments pointed out that the experiment helped improve the learning process of ontology engineering in a fun way using pizzas, while others reported having improved their understanding of the use of ontology quantifiers in general as a basis for modelling high-quality and correct ontologies in the future.

6.2 Hypothesis Testing

To investigate the formulated hypotheses, we explore the impact that the representational formalism and the prior qualification of the workers have on the acquired results. The first independent variable —the representational formalism— can be either the Rector, Warren, or VOWL formalism. The prior modelling knowledge (subjective and objective) is the second set of independent variables, each of which takes values from 1 (no knowledge) to 4 (expert knowledge). The dependent variables influenced by the above factors are the *accuracy* of the achieved results as well as the *time* needed for performing the verification tasks. To investigate the significance of the presented results, we performed hypothesis testing using parametric tests.

To explore *Hypothesis H1*, we compare the effect of the three treatments (ORA formalism representations) on the performance in terms of accuracy and speed. For this, we used a one-way ANOVA test, which is suited when we have categorical independent variables and an interval dependent variable.



(b) Pearson correlation between between prior knowledge in Model Engineering and the time needed for completing the judgements.

Fig. 13. Influence of prior modelling knowledge of the workers on their performance in terms of accuracy (a) and speed (b).

The performed test reveals no statistically significant difference in the performance results between the group means. For the accuracy of verifications, $F(2, 87) = 0.41$ and the p -value equals 0.664, while for the time needed to perform the verifications $F(2, 87) = 0.26$ and $p = 0.773$. While there are slight differences in the accuracy and speed of the collected judgements depending on the used ORA representation, those dissimilarities are not statistically significant. Therefore, we fail to reject the Null Hypothesis of $H1$.

There could be several causes as to why the representation formalism did not significantly affect the accuracy of the ORA verifications. Unfortunately, the collected feedback and comments from the submitted HITs did not provide further insights into a concrete reason. One possible explanation is that the experiment design included an initial training phase for the participants through the qualification test and tutorial, in which participants became familiar with all three representations. Since they became comfortable with each of the representations prior to the verification jobs, this could have led to their consistently high results throughout all tasks.

For investigating *Hypothesis H2*, we measure the significance of the relationship between prior modelling knowledge and the verification accuracy and response times by calculating a Pearson product-moment correlation. At a 95% confidence interval, we report the following statistically significant effects:

- Prior knowledge of ontology restriction modelling positively influenced the verification results ($r(86) = 0.35$, $p = 0.001$). A visual representation of the effects can be seen in Figure 13(a).
- Surprisingly, students who rated their knowledge as higher in web-based knowledge representational languages performed slightly worse on the evaluations ($r(86) = -0.21$, $p = 0.045$). However, the self-assessment is subjective; therefore, further experimental investigations are needed where these areas are evaluated objectively as well.
- Prior modelling knowledge reduced the time needed for performing the verifications ($r(86) = -0.22$, $p = 0.04$) as illustrated in Figure 13(b).

Based on the observed results, we reject the Null Hypothesis of $H2$. As previously explored in [27], we also observed a positive effect of the evaluators' previous knowledge on the verification results. The scores of the qualification test, which objectively evaluated the knowledge of

contributors on the usage of the ontology quantifiers, were shown to have the strongest correlation to the achieved verification results. Meanwhile, a subjective judgement of the contributors on their prior knowledge in model engineering correlates to the speed of their verifications.

6.3 Replication Study

To test the credibility of the observed results, we conducted a replication study. Following the same setup of the original experiment, we asked 78 additional masters' students, taking the same semantic systems introductory course in the consecutive semester, to participate in the experiment. Below, we report on the results gathered during the replication study.

Table 3 shows the achieved results based on the representation in which the ontology axioms were presented (*Hypothesis H1*). We can clearly see that the accuracy is highest when the graphical formalism VOWL is used whereas the time needed for each judgement is the lowest, as with the previous experiment. Nevertheless, those results are again not statistically significant (accuracy: $F(2,87) = 2.08$, $p = 0.131$, and speed: $F(2,87) = 0.65$, $p = 0.523$) for an alpha level of 5%. As with the first conducted experiment, we gathered students' feedback in regards to their preference of used formalism. Again, VOWL was the favourite for the majority (62.8%) of the students.

While we were unable to see any statistically significant differences in the achieved accuracy and speed based on the ORA representation, it is clear that students preferred the graphical representation. Thus, we plan to use VOWL for future experiments in which further design aspects will be investigated.

Next, we calculate Pearson product-moment correlation to investigate the effects of prior modelling knowledge on the performance of the contributors (*Hypothesis H2*). At a 95% confidence interval, we report that prior ontology restriction modelling knowledge positively influenced the performance of the evaluators ($r(74) = 0.32$, $p = 0.004$). Nevertheless, we do not see any statistically significant effects ($r(74) = -0.05$, $p = 0.638$) of prior modelling knowledge of the participants on the speed of their verifications.

With these results, we show how important a qualification test is for ensuring better accuracy of the gathered judgements. While a self-assessment can offer a broader understanding of various background knowledge areas of the participants, it only allows for capturing subjective assessments. Such measurements can be less reliable than an objective qualification and can affect the reproducibility of the results.

7 CONCLUSION AND OUTLOOK

Ontologies are symbolic data structures widely adopted in various research fields and application areas making background knowledge accessible to knowledge-based applications or machine learning algorithms. Human-in-the-loop ontology evaluation is often applied to ensure ontology quality aspects that cannot be automatically assessed and thus favours the establishment of trustworthy AI systems. However, there is currently a lack of investigation of several ontology quality aspects (e.g., the verification of ontology restriction) using human computation and crowdsourcing techniques. Moreover, there are no empirically gained guidelines on how to best design the human-centric ontology evaluation tasks to achieve high-quality results. In this article, we present work that addresses these gaps and contributes towards the realisation of trustworthy AI systems.

Table 3. Results of a Replication Study in Terms of Accuracy and Speed Based on Ontology Representational Formalism and Feedback from the Participants

formalism	avg correctness per HIT	avg time per HIT	student preference
Rector	85.86%	46.14s	12.8%
VOWL	90.12%	41.89s	62.8%
Warren	84.83%	45.49s	24.4%

- (1) First, we propose a human intelligence task, designed for ontology restriction verification. This is a novel task in the ontology evaluation area, specifically for the assessment of ontology modelling aspects, and can be applied to verify other ontology modelling issues from various ontologies as well beyond the pizza ontology used as part of the experiments. We perform two large-scale experiments with semi-expert crowds and show that 100% accuracy of assessments can be achieved when a majority vote aggregation is applied.
- (2) Second, we empirically derive guidelines on aspects of optimal task design (i.e., ontology representation and crowd qualification testing), which can support other researchers focusing on HC-based evaluation of ontologies or even other conceptual structures. Our results imply that while the representation of the knowledge axioms does not significantly impact the quality of the gathered judgements, contributors preferred a visual representation. Additionally, we show that objective qualification testing can be applied to achieve higher quality of evaluation results. While expert and intermediate ontology engineers provide judgements with 98.89% to 100% correctness, beginners can also produce high-quality verifications (approximately 93%). Thus, depending on the goal of the evaluation campaign and available budget, a different crowd might be selected.
- (3) Lastly, we provide a BPMN formalisation of our previously proposed human-in-the-loop ontology evaluation methodology (HERO [24]) and exemplify its usage based on the conducted campaigns. All resources created for the experimental investigations are publicly available and can be reused in future ontology evaluation campaigns.

7.1 Broader Implications for (Ontology) Curation Campaigns

Based on the results gathered through the performed experimental investigations, we summarise key implications and recommendations for performing ontology curation campaigns:

- *Qualification test necessity.* The obtained experiment results indicate the importance of an objective qualification test as apposed to a self-assessment. An objective test can be adopted to assess particular knowledge skills that will be needed in the curation campaign and can support the reproducibility of the gathered results. Results obtained through a self-assessment can be misleading since each contributor has one's own interpretation of the questions and one's own skill set. Providing a structured knowledge level scale for contributors to identify their own strengths still failed to produce accurate assessments. These findings can be of interest in various communities working on human-in-the-loop solutions for which contributors should have particular skills.
- *Qualification test design.* We believe that the design of the applied qualification test could have had a high impact on the high accuracy of gathered results. The approach relied on assigning contributors to a qualification level according to their skills, which they showed by working on increasingly difficult problems. The test questions were organised and assessed such that they correspond to particular ontology modelling skill needed for the verification of ontology restrictions. The contents of the qualification test can be adopted following the same structural approach to support further aspects of ontology curation campaigns and similar domains.
- *Acceptance of ontology representation.* While the representation of the ontology did not have a direct impact on the gathered campaign results, contributors expressed their preference towards a graphical representation. It should be noted that the acceptance of the representation format could be dependant on the crowd characteristics. We therefore recommend the usage of the VOWL representation for future ontology curation campaign, where a crowd with prior modelling knowledge is selected.

- *Task design.* We exemplified the design of a HC-based task for one particular ontology modelling quality aspect- ontology restrictions verification. The designed HITs rely on several important elements to be included also for further (ontology) modelling evaluation tasks: (1) instructions on modelling choices; (2) context representing relevant domain information; (3) judgement options based on a predefined set of modelling issues.

7.2 Limitations and Outlook

In spite of the valuable insights gained through this study, it is essential to recognise several limitations and areas for which future work is still needed.

- *Evaluated ontology.* In the conducted experiments, we utilised a small ontology (the Pizza Ontology), thus giving only partial insights into the benefits of the designed artefacts. While the Pizza Ontology includes a variety of axioms showcasing the usage of the ontology restrictions, further investigations are needed to ensure external validity. Therefore, we plan a number of follow-up studies with larger and real-world ontologies as future work.
- *Crowd setting.* A further aspect to be considered for future research is the examination of whether similar results can be obtained when conducting the experiment with a layman crowd. While representing axioms in VOWL might be suitable for novice ontology engineers, participants without any background knowledge in modelling could find an alternative representation considerably easier to comprehend. Thus, we are currently planning a differentiated replication of the experiment presented in this article in which we utilise layman workers from Amazon Mechanical Turk to assess the generalizability of the findings.
- *Further task design aspects.* In this article, we investigate two important task design aspects: ontology axiom representation and crowd qualification testing. While we provide insight into the design of these aspects, other factors such as the context provided and the HIT answer options have not been thoroughly investigated yet. Furthermore, the approach can be extended to support multiple ontology quality aspects. At the moment, there are still several human-centric ontology errors that have not yet been approached using human computation techniques. It is essential to explore and address these quality aspects to develop a versatile and generic approach for tackling various ontology evaluation challenges.
- *Hybrid human-AI verification.* Human-in-the-loop campaigns are time-intensive and face scalability issues when the goal is to assess the quality of larger knowledge structures. There is a need to explore the integration of state-of-the-art ontology evaluation algorithms with human-in-the-loop approaches with the aim of reducing human efforts and minimising costs. We believe that our contributions can be used as a basis for a hybrid human-machine approach towards designing and conducting ontology curation campaigns and have already proposed an approach for establishing such hybrid solutions based on multi-agent theories in [23].

With the envisioned future work outlined above, this study has the potential to make a significant contribution towards the field of human-in-the-loop curation of semantic data structures, thus preventing biased and unfair applications relying on an ontological component.

APPENDIX

A ONTOLOGY AXIOM REPRESENTATION FORMALISMS

Figures in this appendix depict how an OWL axiom can be represented as a textual axiom following guidelines from [19] and [27] as well as a VOWL graph.

<p>OWL: Class(MargheritaPizza complete Pizza restriction (hasTopping someValuesFrom Tomato) restriction (hasTopping someValuesFrom Mozzarella) restriction (hasTopping allValuesFrom (Tomato or Mozzarella))) Paraphrase: A margherita pizza is any pizza which, amongst other things, has some tomato topping and also some mozzarella toppings and also has only mozzarella and/or tomato toppings.</p>	<p>OWL: Class(MargheritaPizza complete Pizza restriction (hasTopping someValuesFrom Tomato) restriction (hasTopping someValuesFrom Mozzarella) restriction (hasTopping allValuesFrom (Tomato or Mozzarella))) Paraphrase: A margarita pizza is any pizza which, amongst other things, has at least one tomato topping and also at least one mozzarella topping and also has no other than mozzarella and/or tomato toppings.</p>
(a) Rector formalism, taken from [19]	(b) Warren formalism.

Fig. 14. Example of an OWL axiom describing a Margherita Pizza, paraphrased into natural language following (a) a formalism proposed by Rector et al. and (b) its alternative suggested by Warren et al.

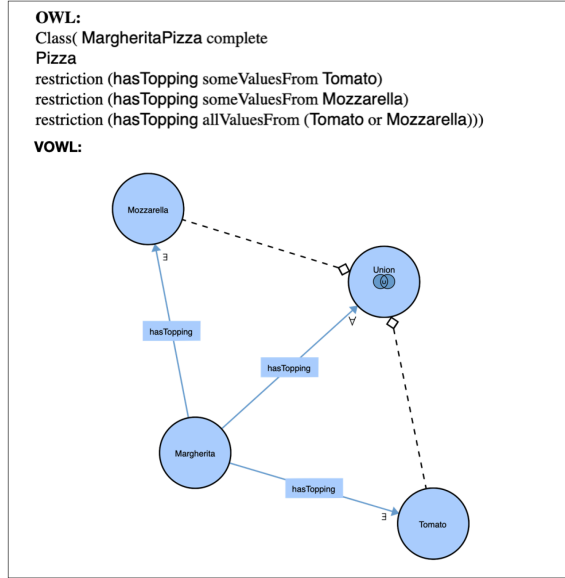


Fig. 15. Example of an OWL axiom describing a Margherita Pizza, represented in the VOWL formalism.

ACKNOWLEDGMENTS

We thank all contributors from the user studies for their involvement in the experiment.

REFERENCES

- [1] Maribel Acosta, Amrapali Zaveri, Elena Simperl, Dimitris Kontokostas, Sören Auer, and Jens Lehmann. 2013. Crowdsourcing linked data quality assessment. In *The Semantic Web – ISWC 2013*, Harith Alani, Lalana Kagal, Achille Fokoue, Paul Groth, Chris Biemann, Josiane Xavier Parreira, Lora Aroyo, Natasha Noy, Chris Welty, and Krzysztof Janowicz (Eds.). Springer, Berlin, 260–276.
- [2] Anna Breit, Laura Waltersdorfer, Fajar J. Ekaputra, Marta Sabou, Andreas Ekelhart, Andreea Iana, Heiko Paulheim, Jan Portisch, Artem Revenko, Anneten ten Teije, and Frank van Harmelen. 2023. Combining machine learning and semantic web: A systematic mapping study. *ACM Comput. Surv.* (Mar. 2023).
- [3] Eden S. Erez, Maayan Zhitomirsky-Geffet, and Judit Bar-Ilan. 2015. Subjective vs. objective evaluation of ontological statements with crowdsourcing. *Proceedings of the Association for Information Science and Technology* 52, 1 (2015), 1–4.
- [4] André Freitas and Edward Curry. 2016. *Big Data Curation*. Springer International Publishing, Cham, 87–118.
- [5] Florian Hanika, Gerhard Wöhlgenannt, and Marta Sabou. 2014. The uComp protege plugin for crowdsourcing ontology validation. In *International Semantic Web Conference (Posters & Demos)*. 253–256.
- [6] Alan R. Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. 2004. Design science in information systems research. *Management Information Systems Quarterly* 28 (03 2004), 75–105.

- [7] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutiérrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan F. Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. *Knowledge Graphs*. Number 22 in Synthesis Lectures on Data, Semantics, and Knowledge. Morgan & Claypool.
- [8] Eric Horvitz. 2017. AI, people, and society. *Science* 357, 6346 (2017), 7–7.
- [9] Vivek Iyer, Lalit Mohan Sanagavarapu, and Y. Raghu Reddy. 2022. A framework for syntactic and semantic quality evaluation of ontologies. In *Secure Knowledge Management in the Artificial Intelligence Era*, Ram Krishnan, H. Raghav Rao, Sanjay K. Sahay, Sagar Samtani, and Ziming Zhao (Eds.). Springer International Publishing, Cham, 73–93.
- [10] Diana Kalibatiene. 2021. *Ontology-Based Information Systems Establishment and Recent Development*. Special Issues in *Journal for Applied Sciences*. Section “Computing and Artificial Intelligence”.
- [11] Chepkoech C. Kiptoo. 2020. Ontology enhancement using crowdsourcing: A conceptual architecture. *International Journal of Crowd Science* 4, 3 (2020), 231–243.
- [12] Markus Krötzsch, František Simančík, and Ian Horrocks. 2012. A description logic primer. *arXiv* (2012).
- [13] Maxat Kulmanov, Fatima Zohra Smaili, Xin Gao, and Robert Hoehndorf. 2020. Semantic similarity and machine learning with ontologies. *Briefings in Bioinformatics* 22, 4 (10 2020).
- [14] Melinda McDaniel and Veda C. Storey. 2019. Evaluating domain ontologies: Clarification, classification, and challenges. *ACM Computing Surveys (CSUR)* 52, 4 (2019), 1–44.
- [15] Jonathan M. Mortensen. 2013. Crowdsourcing ontology verification. In *The Semantic Web – ISWC 2013*. Springer, Berlin, 448–455.
- [16] Jonathan M. Mortensen, Evan P. Minty, Michael Januszzyk, Timothy E. Sweeney, Alan L. Rector, Natalya F. Noy, and Mark A. Musen. 2015. Using the wisdom of the crowds to find critical errors in biomedical ontologies: A study of SNOMED CT. *Journal of the American Medical Informatics Association* 22, 3 (2015), 640–648.
- [17] Jonathan M. Mortensen, Mark A. Musen, and Natalya F. Noy. 2013. Crowdsourcing the verification of relationships in biomedical ontologies. In *AMIA Annual Symposium Proceedings*, Vol. 2013. American Medical Informatics Association, 1020–1029.
- [18] María Poveda-Villalón, Asunción Gómez-Pérez, and Mari Carmen Suárez-Figueroa. 2014. Oops!(ontology pitfall scanner!): An on-line tool for ontology evaluation. *International Journal on Semantic Web and Information Systems (IJSWIS)* 10, 2 (2014), 7–34.
- [19] Alan Rector, Nick Drummond, Matthew Horridge, Jeremy Rogers, Holger Knublauch, Robert Stevens, Hai Wang, and Chris Wroe. 2004. OWL pizzas: Practical experience of teaching OWL-DL: Common errors & common patterns. In *International Conference on Knowledge Engineering and Knowledge Management*. Springer, Berlin, 63–81.
- [20] Marta Sabou, Lora Aroyo, Kalina Bontcheva, Alessandro Bozzon, and Rehab K. Qarout. 2018. Semantic web and human computation: The status of an emerging field. *Semantic Web* 9, 3 (2018), 291–302.
- [21] Marta Sabou, Dietmar Winkler, Peter Penzerstadler, and Stefan Biffl. 2018. Verifying conceptual domain models with human computation: A case study in software engineering. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 6. 164–173.
- [22] Isabella Saccardi, Duygu Sezen Islakoglu, Anouk Neerincx, and Federica Lucia Vinella. 2022. Symbiotic child emotional support with social robots and temporal knowledge graphs. In *Workshop on Human-Centered Design of Symbiotic Hybrid Intelligence, Collocated with the 1st International Conference on Hybrid Human Artificial Intelligence (HHAI2022)*. arXiv: <https://doi.org/10.48550/arXiv.2205.13229>
- [23] Stefani Tsaneva. 2023. Evaluating knowledge graphs with hybrid intelligence. In *The Semantic Web: ESWC 2023 Satellite Events* (Hersonissos, Greece). Springer International Publishing.
- [24] Stefani Tsaneva, Klemens Käsznar, and Marta Sabou. 2022. Human-centric ontology evaluation: Process and tool support. In *Knowledge Engineering and Knowledge Management: 23rd International Conference, EKAW 2022, Bolzano, Italy, September 26–29, 2022, Proceedings*. Springer, 182–197.
- [25] Frank van Harmelen and Annette ten Teije. 2019. A boxology of design patterns for hybrid learning and reasoning systems. *Journal of Web Engineering* 18, 1 (2019), 97–124.
- [26] María Poveda Villalón and Asunción Gómez Pérez. 2016. Ontology evaluation: A pitfall-based approach to ontology diagnosis. *PhD Tesis, Universidad Politécnica de Madrid, Escuela Técnica Superior de Ingenieros Informáticos* (2016).
- [27] Paul Warren, Paul Mulholland, Trevor Collins, and Enrico Motta. 2019. Improving comprehension of knowledge representation languages: A case study with description logics. *International Journal of Human-Computer Studies* 122 (2019), 145–167.
- [28] Gerhard Wohlgenannt, Marta Sabou, and Florian Hanika. 2016. Crowd-based ontology engineering with the uComp protégé plugin. *Semantic Web* 7, 4 (2016), 379–398.
- [29] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Experimentation in Software Engineering*. Springer Science & Business Media.

Received 14 March 2023; revised 29 August 2023; accepted 29 September 2023