

# Human-Centric Ontology Evaluation

## A Human Computation approach for ontology restrictions verification

DIPLOMARBEIT

zur Erlangung des akademischen Grades

**Diplom-Ingenieurin**

im Rahmen des Studiums

**Wirtschaftsinformatik**

eingereicht von

**Stefani Stoynova Tsaneva, BSc**

Matrikelnummer 01527443

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Dipl.-Ing. Mag.rer.soc.oec. Dr.techn. Stefan Biffli

Mitwirkung: Projektass.(FWF) Reka Marta Sabou, MSc PhD

Wien, 2. März 2021

---

Stefani Stoynova Tsaneva

---

Stefan Biffli



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Human-Centric Ontology Evaluation

## A Human Computation approach for ontology restrictions verification

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

**Diplom-Ingenieurin**

in

**Business Informatics**

by

**Stefani Stoynova Tsaneva, BSc**

Registration Number 01527443

to the Faculty of Informatics

at the Vienna University of Technology

Advisor: Ao.Univ.Prof. Dipl.-Ing. Mag.rer.soc.oec. Dr.techn. Stefan Biffi

Assistance: Projektass.(FWF) Reka Marta Sabou, MSc PhD

Vienna, 2<sup>nd</sup> March, 2021

\_\_\_\_\_  
Stefani Stoynova Tsaneva

\_\_\_\_\_  
Stefan Biffi



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Erklärung zur Verfassung der Arbeit

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 2. März 2021

---

Stefani Stoynova Tsaneva



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Acknowledgements

I would like to thank Prof. Stefan Biffl for agreeing to supervise my master thesis as well as for providing insightful feedback and recommendations.

I wish to also express my gratitude to Dr. Marta Sabou for her dedicated support during each stage of the thesis. She provided guidance and encouragement when needed and her valuable remarks and suggestions helped me in enhancing the quality of my work.

Finally, I wish to thank all of the participants, who joined in the pilot and experimental evaluation, without whom the analysis would not have been possible.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.



# Abstract

Ontologies present a conceptual view of a domain of interest and are essential for systems requiring real-world knowledge. The correctness and quality of ontologies are of high importance, as incorrectly represented information or controversial concepts modeled from a single viewpoint can lead to invalid application outputs and biased systems. Several ontology quality issues can be detected automatically, such as the detection of syntax errors or hierarchy cycles, however, others require human involvement, e.g., identifying incorrectly modeled statements, or discovering concepts not compliant with how humans think. Such **human-centric ontology evaluation tasks** (HOETs), typically performed manually by domain experts or knowledge engineers, can be expensive, time-intensive, and have limited scalability. Human Computation (HC) techniques have been used as a promising approach to outsource HOETs to human contributors at a lower cost.

Despite the importance of human-centric ontology evaluation, a systematic understanding of the types of HOETs is still missing. Moreover, it is not clear which human-centric ontology evaluation has already been addressed with HC methods and how to use HC to realise those HOETs that were not yet investigated.

This thesis aims to address this research gap by following a *Design Science* methodology. First, *systematic literature review* methods are used to investigate human-centric ontology evaluation and a structured and unbiased review of HOETs, their characteristics and used solution approaches is provided. We also identify a list of HOETs for which a HC approach has still not been presented. Second, from this list, we select the task of ontology restriction verification and propose a corresponding *HC task design*. Third, an *experimental evaluation* of the proposed HC task design solution is performed using a student crowd in the context of *distance learning* approaches at Vienna University of Technology.

Based on the evaluation data we conclude that: (i) over 90% of the collected responses were correct; (ii) with the proposed evaluation method a 100% accuracy of the verifications can be reached using a majority vote aggregation; (iii) the knowledge representation formalism in which an ontology is presented to the contributors can influence the quality of their assessments; (iv) which formalism leads to the highest quality of verifications depends on the ontology axiom structure and the defect type; (v) prior modeling knowledge of the participants is a good predictor of their verification performance.

With the proposed HC method high-quality evaluations were achieved when the contributors are novice ontology engineers. In future, experimental investigations are needed where the solution is also explored with layman crowds. Several HOETs are identified that are still missing a HC approach, therefore, the proposed HC task can be further extended to support those and thus enable the verification of multiple ontology aspects.

# Contents

<b>Abstract</b>	<b>ix</b>
<b>Contents</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Research Problem . . . . .	1
1.2 Research Questions . . . . .	2
1.3 Methodology . . . . .	4
1.4 Thesis Structure . . . . .	6
<b>2 Background and Related Work</b>	<b>7</b>
2.1 Ontology Evaluation . . . . .	7
2.2 Human-Centric Ontology Evaluation Tasks . . . . .	9
2.3 Human Computation and Crowdsourcing . . . . .	10
2.4 Using Human Computation for Human-Centric Ontology Evaluation . . . . .	10
2.5 Verification of Ontology Restrictions . . . . .	11
2.6 Ontology Axiom Representations . . . . .	12
<b>3 Human-Centric Ontology Evaluation: Literature Review</b>	<b>17</b>
3.1 Literature Review Method . . . . .	17
3.2 Included Papers . . . . .	26
3.3 Evaluated Resource (SMS_RQ1) . . . . .	28
3.4 Evaluation Goal (SMS_RQ2) . . . . .	30
3.5 Evaluation Context (SMS_RQ3) . . . . .	36
3.6 Evaluation Population (SMS_RQ4) . . . . .	39
3.7 Evaluation Method (SMS_RQ5) . . . . .	44
3.8 Literature Review Summary . . . . .	49
<b>4 Ontology Restrictions Verification: a Human Computation Approach</b>	<b>51</b>
4.1 Misuse of the Universal and Existential Restrictions . . . . .	51
4.2 Ontology Restrictions Verification - a Human Computation Approach . . . . .	54
4.3 Human Computation Approach Summary . . . . .	60
	xi

<b>5</b>	<b>Setup of Evaluation Experiment</b>	<b>61</b>
5.1	Experiment Aims . . . . .	61
5.2	Experimental Setup . . . . .	62
5.3	Experiment Overview . . . . .	65
5.4	Evaluation Setup Summary . . . . .	68
<b>6</b>	<b>Evaluation</b>	<b>69</b>
6.1	Evaluation Methods . . . . .	69
6.2	Participants Background Knowledge . . . . .	70
6.3	Experiment Results . . . . .	72
6.4	Evaluation Summary . . . . .	83
<b>7</b>	<b>Conclusion and Future Work</b>	<b>87</b>
7.1	Answers to Research Questions and Discussion . . . . .	88
7.2	Limitations & Future Work . . . . .	89
	<b>List of Figures</b>	<b>91</b>
	<b>List of Tables</b>	<b>93</b>
	<b>Bibliography</b>	<b>95</b>
	<b>Appendices</b>	<b>99</b>
	Appendix A: Included Papers for HOETs Analysis . . . . .	99
	Appendix B: Experiment Starting Point for the Participants from Group A	105
	Appendix C: Self Assessment Test . . . . .	106
	Appendix D: Qualification Test . . . . .	110
	Appendix E: Feedback Questionnaire . . . . .	116

# Introduction

## 1.1 Research Problem

Since its development in 1989, the World Wide Web (WWW) has experienced major growth. In its early years, now referred to as "Read-Only Web" or Web 1.0, it relied on one-way communication. A small number of creators would publish static content and users with access to the Internet would be able to read it, without the possibility to add comments or provide some feedback. Later on, it evolved to Web 2.0 also known as the "Social Web", which extended the WWW further, allowing for two-way communication, social networks, a variety of content types, and web-based technologies.

With all the information being published, searching for relevant resources becomes challenging and the need to organize and categorize the published information significantly increases. The goal of Web 3.0, the "Semantic Web", is to make information machine-readable and allow for its further processing, sharing, and reuse [Berners-Lee et al., 2001]. Ontologies are fundamental for the Semantic Web as they provide a knowledge representation schema describing concepts of the domain in interest [Kehagias et al., 2008]. In computer science, they are defined as a "formal, explicit specification of a shared conceptualization" [Studer et al., 1998]. Ontologies thus represent a conceptual model consisting of information about concepts, relations, and instances. With the help of ontologies, the information published on the web is structured into categorical systems, allowing for knowledge to be made machine-readable.

Ontology Engineering, the process of creating and maintaining ontologies, is a time-consuming task, and therefore ontologies are often reused and extended during their lifetime. For some application domains, the ontology quality might seem insignificant. An example introduced in [Zaveri et al., 2016] is that while searching for entertainment topics such as which movie an author is related to, a missing movie would not be of such great importance. On the other hand, for a medical application, missing information could be crucial [Zaveri et al., 2016]. Low quality ontologies could lead to failed systems and

serious consequences- incorrectly represented information can lead to invalid application outputs and controversial concepts modeled from a single viewpoint could result in biased or discriminating systems. Therefore, the correctness and quality of ontologies are very important aspects, making ontology evaluation a crucial research area to be addressed [Brank et al., 2005].

Many quality issues of the ontology can be detected via automated methods. While those methods might be fast and scalable, they also have their limitations. Pitfalls such as hierarchy cycles or missing annotations can be easily detected automatically, however, there are some evaluation tasks, which require domain knowledge and can only be solved via human input [Villalón and Pérez, 2016]. An example of such a task taken from [Villalón and Pérez, 2016] is shown in Figure 1.1 and refers to the correct use of the existential and universal quantifiers in ontologies. In order to detect such errors, one would need background knowledge in the area of ontology modeling languages as well as human domain knowledge. A traditional approach to solve this task would be to involve ontology engineers or domain experts that would verify the correctness of the ontology.

An evaluation performed by experts might be more accurate than an automated process, however, it is a costly and time-intensive task [Villalón and Pérez, 2016]. Human Computation (HC) is a promising approach in which specific tasks of the system, which cannot be fully automated, are outsourced to human participants. This can be done for instance with the help of crowdsourcing methods. In Crowdsourcing the tasks are outsourced to an undefined group (crowd) of people (workers) using the Internet instead of a predefined study group. These paradigms have already been successfully integrated with domains such as software engineering [LaToza and Van Der Hoek, 2015] [Sabou et al., 2018b] and have been used for solving human-centric Semantic Web tasks [Sabou et al., 2018a].

There have been some usages of Human Computation and Crowdsourcing techniques in ontology evaluation research, however, a systematic understanding of the types of tasks where human involvement is required is still missing. Moreover, it is not clear which human-centric ontology evaluation has already been addressed with HC methods and how to use HC to realise those HOETs that were not yet investigated. Therefore, this master thesis aims to extend the available research by analysing known HOETs and proposing a HC task design for one such task not yet approached with Human Computation.

### 1.2 Research Questions

This master thesis aims to provide answers to the following research questions:

- **RQ1** Which ontology evaluation tasks cannot be yet reliably automated and need human involvement?

The first part of the thesis aims to identify and collect ontology evaluation tasks, which cannot be (yet reliably) fully automated and thus need a human-centric

<b>Title</b>	P14. Misusing "owl:allValuesFrom"	<b>Importance level</b>	Critical
<b>Aspects</b>	Modelling decisions	<b>Affects to</b>	Classes
<b>Description</b>			
This pitfall consists in using the universal restriction ( <code>owl:allValuesFrom</code> ) as the default qualifier instead of the existential restriction ( <code>owl:someValuesFrom</code> ). Additional information about this pitfall is provided in [1].			
<b>Examples</b>			
Graphical representation and/or OWL code		Natural language description	
		<p>In the graphical example a definition of the class "Book" is provided by means of an <code>owl:equivalentClass</code> axiom in the following way: <math>Book \equiv \exists producedBy.Writer \sqcap \forall uses.Paper</math>. While the <code>owl:someValuesFrom</code> axiom is properly used for stating that the book has to be produced by at least one writer, it is not correct to say that all the materials used in the book have to belong to the class "Paper", as for example, another material used during the production of a book might be "Ink", among others.</p>	
<b>How to solve it</b>			
<p>An universal restriction (<code>owl:allValuesFrom</code>) can be used to restrict the range of a relationship, i.e., to state that, in the context of an axiom, only individuals belonging to the specified class can act as objects of that relationship. Considering a class "ClassA" used as target of an universal restriction for a given relationship, one should:</p> <ol style="list-style-type: none"> <li>Answer the question "Is it possible to have individuals that do not belong to "ClassA" acting as object of such property?" If the answer is "yes", the universal restriction (<code>owl:allValuesFrom</code>) should be deleted. To analyse whether the universal restriction should be replaced by an existential one check the point (b).</li> <li>Check whether the intended meaning of the restriction is to state that at least one individual belonging to "ClassA" should appear as object in an instantiation of such property. In this case, an existential restriction (<code>owl:someValuesFrom</code>) should be used instead of the universal one (<code>owl:allValuesFrom</code>).</li> </ol>			
<b>References</b>	<p>[1] Rector, A., Drummond, N., Horridge, M., Rogers, J., Knublauch, H., Stevens, R., Wang, H., and Wroe, C. (2004). OWL pizzas: Practical experience of teaching OWL-DL: Common errors &amp; common patterns. In <i>Engineering Knowledge in the Age of the Semantic Web</i>, pages 63-81. Springer.</p>		

**Figure 1.1:** Example of a human-centric ontology evaluation task, reproduced from [Villalón and Pérez, 2016].

solution. This goes beyond the OOPS! (Ontology Pitfall Scanner!) collection [Poveda-Villalón et al., 2014] [Villalón and Pérez, 2016], which is described later in chapter 2 *Background and Related Work* - Concrete ontology evaluation tasks that were performed with human involvement were identified and extracted in a structured format.

- **RQ2** How can Human Computation techniques be used to evaluate ontologies regarding the correct usage of restrictions?

The second part of the thesis aims to identify how a HC approach of one human-

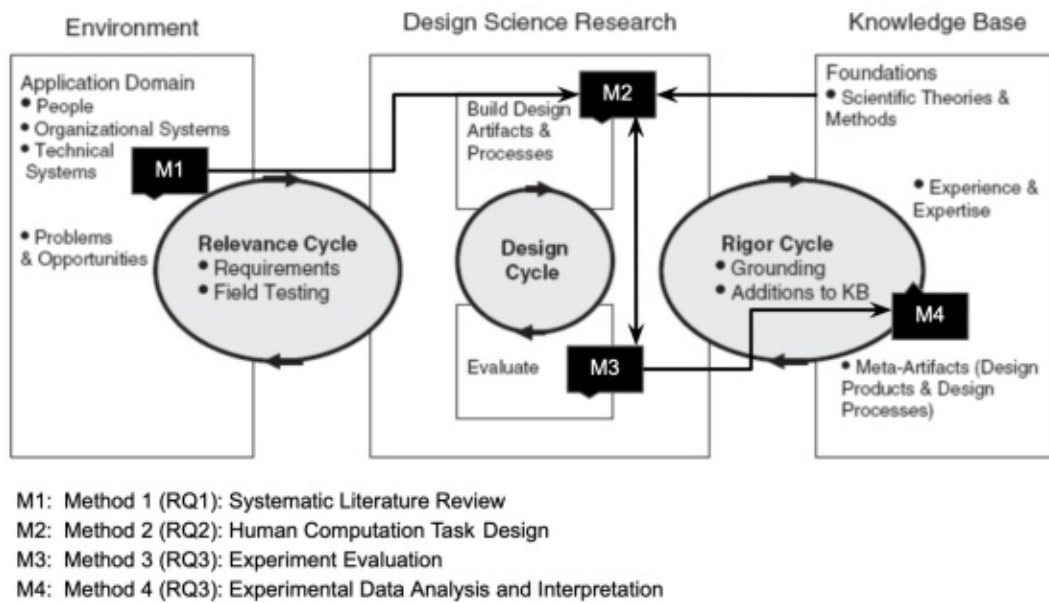
centric ontology verification task, not yet approached with HC techniques- the evaluation of ontology restrictions, can be designed and implemented in a crowd-sourcing platform.

- **RQ3** How suitable are Human Computation techniques for the evaluation task of verifying ontology restrictions?

The last part of the thesis deals with the evaluation of the proposed HC implementation. For this purpose an expert-sourcing (student) experiment is conducted, which aims to provide insights into different HC task design aspects for the ontology evaluation task of identifying the misuse of ontology restrictions.

### 1.3 Methodology

The thesis follows a Design Science approach. This research methodology aims to extend existing knowledge by building new innovative artefacts [Hevner et al., 2004]. Design Science does not strike to develop the best-optimized solution for a problem but instead provides the knowledge base for this [Dresch et al., 2014].



**Figure 1.2:** Methods applied in the master thesis

The methodological flow used for the thesis is visualised in Figure 1.2. The three cycles of the Design Science methodology are addressed as follows:

- **Relevance cycle** - The relevance factor plays an important role while identifying and understanding the problem to be solved.



The first part of the master thesis aims to answer RQ1 and focuses on researching ontology evaluation tasks that cannot be yet reliably fully automated as some human-input is required for their execution. In this step of the master thesis, characteristics of such tasks are gathered and possible solutions are identified. The aim is to identify **human-centric ontology evaluation tasks** (HOETs) and to extend current research by providing an unbiased structured review of these tasks.

This literature review follows a *Systematic Literature Review methodology* and is also part of a larger Systematic Mapping Study (SMS) [Kitchenham et al., 2011][Kitchenham and Charters, 2007] that goes beyond the scope of the thesis. The work in this thesis contributes to the

- (1) definition and piloting of the study protocol,
- (2) search and selection of research papers, and
- (3) data collection from a portion of the total papers identified for review.

The methodological approach of this step is further explained in section 3.1 *Literature Review Method*.

- **Rigor cycle** - The rigor connects the current research with existing knowledge and methodologies. The researchers should ground the new developments on existing methodologies and theories [Hevner, 2007].

This master thesis is grounded on existing concepts described in chapter 2 *Background and Related Work* and techniques and concepts identified during the literature research done in the relevance cycle.

- **Design cycle** - This cycle includes the iteration between building and evaluating the proposed artefact. This is the main cycle in Design Science projects and it enables the continuous feedback flow for the refinement of the artefact [Hevner, 2007].

For this part of this master thesis, in order to address RQ2, a *HC task design* of one of the identified HOETs not yet approached using Human Computation techniques—the evaluation of ontology restrictions, is proposed. During this phase, decisions on the design and implementation of the HC task are taken, such as whether to develop open-ended (crowd workers can provide their answers without constraints) or closed tasks (crowd workers need to select an answer from a predefined taxonomy of defects.). The implementation methodology is described in detail in section 4.2 *Ontology Restrictions Verification - a Human Computation Approach*.

Furthermore, to tackle RQ3 an *experimental evaluation* is performed in the form of a student experiment to evaluate the developed HC task design of the selected HOET. The setup and execution of the experiment are described further in chapter 5 *Setup of Evaluation Experiment*. The last part of the thesis deals with the *evaluation data analysis and interpretation* to determine how suitable Human Computation techniques were for solving the ontology evaluation task and to complete RQ3. What evaluation methods were used is addressed in detail in section 6.1 *Evaluation Methods*.

## 1.4 Thesis Structure

The main contributions of the thesis are structured in the following way:

- Chapter 2 *Background and Related Work* presents a discussion of related work in the field.
- Chapter 3 *Human-Centric Ontology Evaluation: Literature Review* provides insights into the current literature state about ontology evaluation tasks that require human involvement.
- Chapter 4 *Ontology Restrictions Verification: a Human Computation Approach* introduces a HC task design for the evaluation of ontology restrictions.
- Chapter 5 *Setup of Evaluation Experiment* describes the setup of the conducted experiment for the evaluation of the proposed approach in the previous chapter.
- Chapter 6 *Evaluation* shows the results of the conducted expert-sourcing (student) experiment and describes how suitable the proposed implementation was for solving the verification task of evaluating ontology restrictions.
- Chapter 7 *Conclusion and Future Work* concludes the findings of this thesis and suggests aspects to be considered in future work.

# Background and Related Work

This chapter addresses important for the master thesis research areas. First of all, in section 2.1 *Ontology Evaluation* background information on Ontology Evaluation is provided since this is the main research area to which the thesis contributes to. In section 2.2 *Human-Centric Ontology Evaluation Tasks* we look into known research, which introduces **human-centric ontology evaluation tasks** (HOETs) to gain some understanding of what such tasks could be and how they have already been approached. Furthermore, Human Computation (HC) and Crowdsourcing (C) techniques are explained in section 2.3 *Human Computation and Crowdsourcing* since those paradigms offer a solution to human-centric tasks and literature is presented, in which those techniques are applied successfully for solving evaluation tasks. Section 2.4 *Using Human Computation for Human-Centric Ontology Evaluation*, acts as an introduction to RQ1, and papers, which present approaches of HC&C methods used to solve HOETs are introduced. In section 2.5 *Verification of Ontology Restrictions*, for the second part of the thesis covering RQ2, which deals with the implementation of a HC approach for one specific HOET (the verification of ontology restrictions), papers that look into the misuse of ontology restrictions are presented. Lastly, in section 2.6 *Ontology Axiom Representations* different ontology representational formalisms, which are of importance for the evaluation performed for RQ3, are discussed.

## 2.1 Ontology Evaluation

Ontology evaluation has been abundantly addressed in the literature for more than 20 years. McDaniel and Storey reflect the work in the domain ontology assessment from papers published in the last two decades in [McDaniel and Storey, 2019]. The authors identify two distinct ways to approach ontology evaluation. The first, referred to as "glass box" or "component evaluation", looks at the characteristics of an ontology throughout its life cycle - its efficiency, accuracy, and appropriateness. A "black-box" or "task-based"

evaluation, on the other hand, considers the overall quality of the ontology when it is integrated into an application and measures the ontology's performance for a specific task [McDaniel and Storey, 2019]. Furthermore, in their survey, five main research areas of Ontology Evaluation are outlined based on their focus - Domain/Task Fit, Error Checking, Libraries, Metrics, and Modularization, as discussed next.

*Domain/Task Fit assessment* aims to evaluate how suitable an ontology is for a specific context based on its performance on a specific set of tasks. In order to perform this evaluation, the ontology engineers need to have predefined task criteria to compare the ontology against. Flaws in the ontology that can be identified in this type of assessment are for instance superfluous and missing concepts [McDaniel and Storey, 2019].

*Error Checking evaluation* focuses on the syntax, structure, and semantics of the ontology. Syntactical errors can easily be found as the ontology representational languages have a well-defined syntax that can be checked automatically. Semantic mistakes, however, can be more challenging to detect, as here contradictory meanings or incorrect interpretations need to be identified. Errors found during this type of assessment have the potential to be automated, however, not all defects are easy to locate. Nevertheless, identifying common ontology mistakes and correcting them improves the quality of an ontology and thus its usefulness. This evaluation approach does not provide a thorough assessment, especially on how suitable the ontology would be in a specific context, and should, therefore, be combined with other ontology assessments [McDaniel and Storey, 2019].

*The Libraries* research class focuses on creating repositories for the storage of ontologies and the maintenance of their quality. The purpose of those libraries is to ensure the quality of domain-specific ontologies so that they can be further reused [McDaniel and Storey, 2019].

*Metric Based assessments* aim to provide an objective evaluation of the ontology, based on specific attributes. The assessment metrics need to be accurate, well-defined, and easy-to-apply and can be used to compare which ontology is best-fitted for a specific context based on the measured aspects. Ontology-based quality attributes are for instance consistency, completeness, conciseness, expandability, and sensitiveness [McDaniel and Storey, 2019].

*Modularization evaluation* of ontologies consists of dividing the ontology into small independent pieces, which are then assessed separately on their syntax, semantics, and pragmatic quality. Self-contained modules can be later on of great importance for evolution and reuse of the ontology or parts of it [McDaniel and Storey, 2019].

To ease the process of selecting the right ontology for a specific domain the ontology should (1) not include errors, (2) be modular, (3), be stored in an ontology repository so it can be found easily, (4) have high score in specific assessment attributes (5) be applicable in the domain and for the needed task. As there are limitations and challenges associated with each evaluation research area, multiple approaches should be combined [McDaniel and Storey, 2019].

## 2.2 Human-Centric Ontology Evaluation Tasks

As identified by [McDaniel and Storey, 2019] semantic mistakes cannot always be automatically detected. Some research papers have already identified several ontology evaluation tasks that require human input. Most such tasks fall into the Ontology Evaluation areas Domain/Task Fit, and Error Checking, described in the previous section.

M. Poveda-Villaillal et al. [Villalón and Pérez, 2016] [Poveda-Villalón et al., 2014] present a catalog of bad practices in ontology engineering such as hierarchy circles and missing annotations and automates the detection of as many of such pitfalls as possible. In their work they manage to identify tasks for which the automation is not feasible, as they would require background information and human involvement. Examples are the detection of polysemous elements (e.g. theatre as a building but also as the performing art), misuse of ontology restrictions, and missing domain information (compared to a criteria specification document).

There is also some literature in which human-centric evaluation tasks have been approached. In [Teitsma et al., 2014] the authors present methods for developing ontologies for question answering systems. In order to evaluate how the structure of the ontologies complies with how people categorize their knowledge, a small experiment was conducted. The participants were asked to determine which concept is the outsider concept out of a set of three. The results are later used to compare the formal and cognitive semantic distance of concepts, which are used as measures to evaluate the ontologies [Teitsma et al., 2014].

Another example is the evaluation of the quality and cognitive soundness of Encyclopedic Knowledge Patterns, automatically extracted from Wikipedia [Nuzzolese et al., 2017]. The authors conduct an experiment, where the participants are asked to describe specific things of a particular domain or rate how important their correlations to other objects are. Based on the results Encyclopedic Knowledge Patterns are created and compared to those that were automatically detected [Nuzzolese et al., 2017].

Although human-centric ontology evaluation has already been addressed in the literature to some extent, there is still no systematic understanding of which tasks of the ontology evaluation require human involvement. This is one of the gaps this thesis aims to fill in and is discussed further in chapter 3 *Human-Centric Ontology Evaluation: Literature Review*.

A way to solve tasks which require human input would be to involve ontology engineering experts who would evaluate the correctness of the ontology. However, as already mentioned, such an evaluation incurs high costs and is time-consuming. A popular approach to solve specific parts of the ontology evaluation that require human contributions is the usage of Human Computation and Crowdsourcing techniques, which are discussed next.

### 2.3 Human Computation and Crowdsourcing

Using Human Computation (HC) methods means outsourcing specific tasks of a system, which cannot be fully automated to human participants and leveraging the human processing power to solve those tasks. The HC paradigm can therefore support those research areas of the Semantic Web that require human contributors [Sabou et al., 2018a]. One successful example of HC used for the construction of a knowledge base is Wikipedia [Roengsamut et al., 2015].

Crowdsourcing is a HC approach often used for tasks that require human intelligence [Roengsamut et al., 2015]. HC methods can involve a small number of contributors, however, crowdsourcing enables the leveraging of the "wisdom of the crowd" [Sabou et al., 2018a]. This makes Crowdsourcing a promising approach that can support (or replace) knowledge and domain experts in their work on Semantic Web tasks.

When approaching specific tasks with Human Computation & Crowdsourcing methods, one possibility is to split the assignment into micro-tasks which can be completed by a single contributor. In a micro-tasking online marketplace, such as Amazon Mechanical Turk<sup>1</sup> the requester would publish small tasks called Human Intelligence Tasks (HITs), which can be solved in several seconds or minutes. Contributors complete the HITs against a predefined monetary reward. The requester can also configure what qualification the contributors must have, how many responses are required for each HIT, and how long the tasks should be available for [Mortensen, 2013, Mortensen et al., 2015]. Once the required number of responses are gathered, the requester can aggregate and process the contributors' answers using different methods to calculate the final result.

In the next section, we discuss known usages of Human Computation & Crowdsourcing methods applied for human-centric ontology evaluation in particular.

### 2.4 Using Human Computation for Human-Centric Ontology Evaluation

Approaches towards conceptual model evaluation using Human Computation techniques have been mostly reported in Knowledge Engineering and Software Engineering literature.

In Software Engineering, Human Computation and Crowdsourcing techniques have been used for the verification of Enhanced entity-relationship diagrams based on the requirement specifications [Sabou et al., 2018b]. The crowd workers were provided with a model as well as the requirements for a particular model element. They were asked to verify this element and were able to select the detected defect out of a given defect taxonomy [Sabou et al., 2018b].

In [Hanika et al., 2014] and [Wohlgenannt et al., 2016] the authors investigate how crowdsourcing can be incorporated into the ontology engineering process with the goal

---

<sup>1</sup><https://www.mturk.com>

of outsourcing validation tasks to human participants during the ontology engineering process. They present a plugin for Protégé<sup>2</sup>, an open-source platform for ontology engineering, which delegates tasks to Games with a Purpose (GWAP) or paid-for crowdsourcing using CrowdFlower<sup>3</sup>. In their research among the tool development and usability details, the authors also address common specification and verification tasks from the ontology engineering process that can be crowdsourced such as "Verification of Relation Correctness" or "Specification of Relation Type".

There has already been some work on ontology evaluation using Human Computation techniques. However, it is not clear which human-centric ontology evaluation has already been addressed with HC methods and how to use HC to realise those HOETs that were not yet investigated. In chapter 3 *Human-Centric Ontology Evaluation: Literature Review* a review of previous research where Human Computation was used for Human-Centric Ontology Evaluation is provided. The aim of the thesis is to further contribute to the field, by exploring the use of Human Computation methods for a specific human-centric ontology evaluation task- the verification of ontology restrictions, which has not yet been approached using HC methods. We look into this specific HOET in detail in the next section.

## 2.5 Verification of Ontology Restrictions

One specific human-centric ontology evaluation task is the verification of ontology restrictions. There is some research on the misuse of the universal and existential restrictions during the ontology engineering process and a few outstanding papers on this topic are discussed next.

In [Rector et al., 2004] the authors describe common mistakes and patterns that they have observed while teaching ontology engineering to beginners. Amongst the mistakes are the misunderstanding of "some not" & "not some", the overspecialization of hierarchies and the misuse of the ontology quantifiers - tasks also identified by [Poveda-Villalón et al., 2014][Villalón and Pérez, 2016] in their pitfall catalog. [Rector et al., 2004] describe the ontology issues newcomers to OWL (Web Ontology Language) often have and what causes them. The authors discuss different representations of ontology models and argue that the representation has an impact on how well the ontology rules and axioms are perceived by humans. For instance, they consider an OWL model such as "*class Pizza restriction (hasTopping someValuesFrom Mozzarella)*", the meaning of which is not clear right away to beginners in ontology engineering. Then, they transform it to text by paraphrasing it into "*Pizzas have, amongst other things, some mozzarella topping*". [Rector et al., 2004] describes that this textual representation of the model makes the meaning of the pizza description much more understandable.

[Warren et al., 2019] also investigates difficulties ontology engineers experience in the modeling process - one of which is the correct usage of ontology restrictions. The paper

<sup>2</sup><https://protege.stanford.edu>

<sup>3</sup><https://appen.com>

extends the research done by [Rector et al., 2004] by investigating the perception of different representation languages in controlled experiments. Furthermore, the authors propose alternative ontology constructions that could further improve the comprehension of, amongst other things, ontology restriction axioms.

This thesis goes beyond the work covered in the addressed papers since it relies on the findings of the presented literature in order to implement a Human Computation solution for the verification of ontology restrictions.

To understand the difference between different axiom representations, the next sections look into possible ways how ontology restrictions can be formalised.

### 2.6 Ontology Axiom Representations

There are different possibilities how ontologies can be formalised. As we will see in section 3.3 *Evaluated Resource (SMS\_RQ1)* the most common formalisms in which ontologies are represented are OWL<sup>4</sup> and RDF<sup>5</sup>. However, it can sometimes be difficult to understand the meaning of axioms written in description logic languages [Rector et al., 2004]. Students or beginners in ontology engineering have proven to find it hard to understand the meaning of axioms represented in OWL, especially when it comes to the Open World Assumption and the logical meaning of the restrictions [Rector et al., 2004]. In the approach presented in this master thesis 3 different formalisms are considered in order to investigate which one is more effective.

One possibility is to rewrite the OWL restrictions into plain text. [Rector et al., 2004] suggests that the paraphrase should include the words "*amongst other things*", which explicitly states the Open World Assumption. Furthermore, the authors use the words "some" and "only" to refer to the existential ("owl:someValuesFrom") and universal ("owl:allValuesFrom") restriction as well as "and" and "or" rather than "owl:intersectionOf" and "owl:unionOf" to simplify the syntax. Figure 2.1 shows how a Margherita Pizza defined in OWL can be paraphrased to a plain text definition. The authors in [Rector et al., 2004] argue that the paraphrased text makes the understanding of the axioms more clear and intuitive. This type of paraphrasing is referred to as a *Rector formalism* in this thesis.

In [Warren et al., 2019] the authors also investigate how ontology restrictions can be represented to improve the comprehension of the axiom meaning. The authors suggest that the keyword "some" and "only" are not clear enough and replacing them with "at least one" and "no other than" could improve the usage of the ontology quantifiers and could minimize the mistakes in their usage. This alternative paraphrasing is addressed

---

<sup>4</sup>"The W3C Web Ontology Language (OWL) is a Semantic Web language designed to represent rich and complex knowledge about things, groups of things, and relations between things." [OWL, 2012]

<sup>5</sup>Resource Description Framework: "RDF is a standard model for data interchange on the Web." [RDF, 2014]



**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.

**Figure 2.1:** Example of an OWL axiom paraphrased into the Rector formalism [Rector et al., 2004]

as *Warren formalism* in the thesis. An example of how the Margherita Pizza would be defined is provided in Figure 2.2.

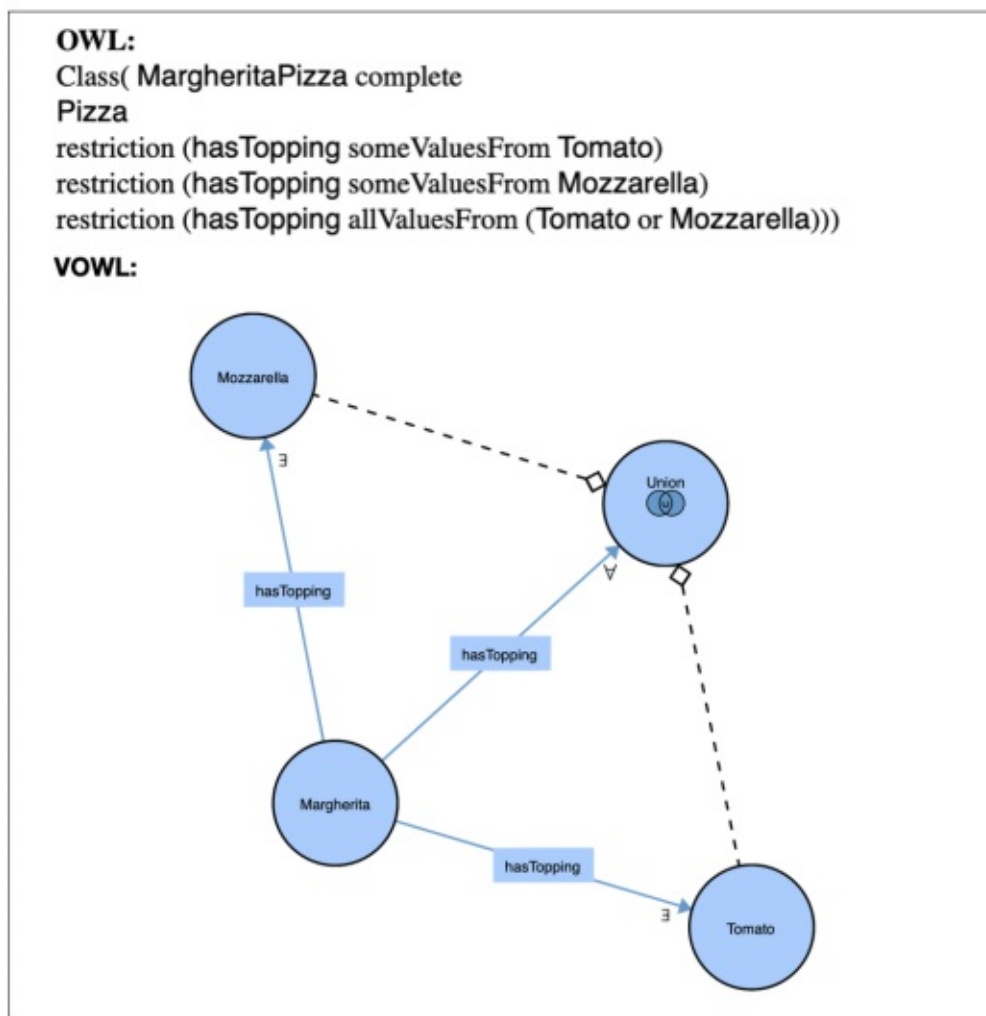
**OWL:**  
 Class( MargheritaPizza complete  
 Pizza  
 restriction (hasTopping someValuesFrom Tomato)  
 restriction (hasTopping someValuesFrom Mozzarella)  
 restriction (hasTopping allValuesFrom (Tomato or Mozzarella)))  
**Paraphrase:**  
 A margarita pizza is *any* pizza which, *amongst other things*, has *at least one* tomato topping and also *at least one* mozzarella topping and also has *no other than* mozzarella *and/or* tomato toppings.

**Figure 2.2:** Example of an OWL axiom paraphrased into the Warren formalism

While paraphrasing OWL axioms in a natural language could be easier to understand for some people, graphical representations might be helpful for those who have previous experience with model engineering. A well known visual representation of ontologies is the *VOWL*<sup>6</sup> (Visual Notation for OWL Ontologies) *formalism*. In Figure 2.3 it is shown how the Margherita Pizza axiom would be visualised in VOWL.

The thesis provides an experimental investigation of how the representation of ontology restriction axioms can influence how well they are understood, and more specifically whether the formulations proposed in [Warren et al., 2019] as alternative representations

<sup>6</sup><http://vowl.visualdataweb.org/v2/>



**Figure 2.3:** Example of an OWL axiom represented in the VOWL formalism

of the axioms are better understood than those originally introduced in [Rector et al., 2004]. The implementation of the HC approach as well as the results of the experiment are described later on in chapter 4 *Ontology Restrictions Verification: a Human Computation Approach* and chapter 6 *Evaluation*.

To conclude this chapter, we can summarise that the research area of Ontology Evaluation has been abundantly addressed in the literature from different angles. Human-centric tasks of the ontology evaluation process have made an appearance in various papers throughout the last years, for some of which a Human Computation approach has been proposed. Nevertheless, a systematic understanding of the types of tasks that require human-input is still missing. The next chapter aims to fill this gap by providing a

literature review of known HOETs and their characteristics as well as ways in which they have already been approached.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Human-Centric Ontology Evaluation: Literature Review

In this chapter, the first research question (RQ1) of the thesis is investigated.

**RQ1** Which ontology evaluation tasks cannot be yet reliably automated and need human involvement?

In order to determine which ontology evaluation tasks cannot be automated and need human involvement, a literature review is conducted on known human-centric ontology evaluation tasks and how they have been solved until now. Section 3.1 *Literature Review Method* describes the methodological process that was followed in order to investigate the research question. The next sections 3.5 - 3.6 provide an analysis of the findings of the literature review. Lastly, the main outcomes of the literature review are discussed in section 3.8 *Literature Review Summary*.

## 3.1 Literature Review Method

As aforementioned, the thesis literature review is part of a Systematic Mapping Study, in which several researchers participate. Known literature from previous studies was collected and categorized in a shared Mendeley<sup>1</sup> library. The identified papers were used as a starting point for this master thesis to gain insights into the topic, collect relevant keywords, and identify missing gaps in the literature.

The stages of the Systematic Mapping Study that this thesis is part of are shown in Figure 3.1. The first stage includes the planning of the study, in which a study protocol is defined of how the study should be conducted, what the research questions are, and what the scope is. After the protocol is defined and reviewed by the study researchers a pilot

---

<sup>1</sup><https://www.mendeley.com>

follows. For the pilot, a few papers are selected with which the defined steps in the study protocol are followed. Based on the outcomes of the pilot the study protocol is finalized and the execution of the study begins. In this stage papers to be included in the study are searched and extracted. Based on predefined exclusion and inclusion criteria the metadata of the found papers is investigated and a decision is made whether the papers should be included in the review or not. For the papers identified as relevant a second selection is performed, for which the full paper content is read, and if the paper is to be included in the literature review, the data of the paper is extracted in a structured format. Lastly, in the Analysis&Reporting stage, the previously extracted data is analysed and the results are reported. The main contributions of the thesis to each of the study phases are explained in the following sections.

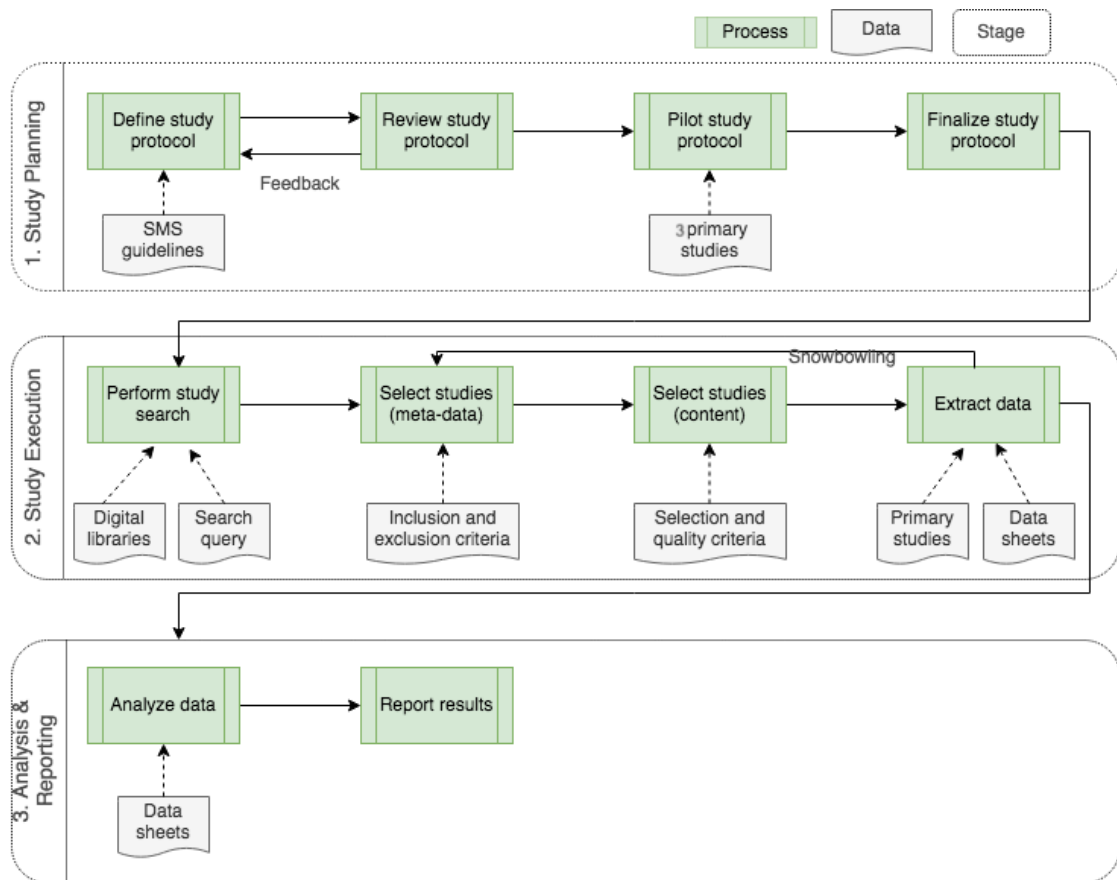


Figure 3.1: Overview of the Systematic Mapping Study process

### 3.1.1 Study Planning

As part of the Study Planning phase, the main contributions of the work in the thesis were to the definition and piloting of the study protocol, and are described below.

### Definition of a study protocol

In the Study protocol, amongst other things, goals for the SMS are identified, the team of researchers who will participate, and the methodology to be followed are defined.

The literature review included in the thesis is guided by the research questions of the SMS, which are shown in Table 3.1.

**Table 3.1:** Research Questions defined for the Systematic Mapping Study

ID	Research Aspect	Research Question
SMS_RQ1	What (evaluated resource)	What are the characteristics of the evaluated knowledge resource?
SMS_RQ2	What (evaluation goal)	What is the goal of the evaluation?
SMS_RQ3	Why (evaluation motivation)	In what context/setting does HET most often occur?
SMS_RQ4	How (evaluation population)	What are the characteristics of the evaluation population involved in verification approaches?
SMS_RQ5	How (method)	How is the evaluation concretely performed from a methodology/tooling perspective?

As part of the definition of the study protocol, the thesis contributed to the selection of a search string, with which the literature search is to be performed in scientifically established digital databases. Below the process of defining a search string and the selection of digital libraries is explained.

The search string consists of identified keywords in known relevant papers, terminology of the research area, and key terms of the research questions of the thesis. The search queries should further not be too restrictive to enable a broad overview of the topic. Based on known primary studies and surveys that focus on ontology-evaluation a list of keywords was extracted and the most common terms were outlined. To gain a better understanding of the number of papers to be later on reviewed, the digital search engines of the scientific databases, as well as to evaluate possible search queries the following digital libraries were searched using possible keywords combinations:

**Scopus**<sup>2</sup> In Scopus, the advanced search option was used and metadata fields such as the title, abstract, and keywords were searched.

**ISI Web of Science**<sup>3</sup> For this digital library, an advanced search was used as well. A focused search on the topic ( including title, abstract, and keyword metadata) was

<sup>2</sup><https://www.scopus.com/>

<sup>3</sup><https://apps.webofknowledge.com>

preferred.

**IEEE Xplore Digital Library**<sup>4</sup> A command search from the advanced search options was selected. Once again, the scope of the search was set to the title, abstract, or all metadata fields. However, the search string was limited to a maximum of 20 keyword terms.

**ACM Digital Library**<sup>5</sup> The ACM Digital Library offers an advanced search option as well. The scope of the search was limited to abstract, title, and keywords.

Based on the search strings that were used at this step the following conclusions could be outlined:

- Scoping the Search for the keyword "human" only to the title or abstract can exclude many papers, who looked into a human-centric evaluation, but this was not their main focus.
- Limiting the search to a specific research area is very exclusive and eliminates papers from other domains that could include ontology evaluation with human involvement.

From the drawn conclusions from me and the investigation of a senior researcher the finalized search strings were selected for the literature search. For the literature search the conjunction of three sub-queries  $Q = Q1SW \cap Q2HC \cap Q3Eval$  was used as a search string. The defined search sub-queries are listed in Table 3.2. The first sub-query Q1SW searches papers that focus on the overall research field of semantic web resources while sub-query Q3Eval limits the papers to those who are looking at their evaluation. Lastly, the sub-query Q2HC further restricts the results so that only papers that investigate human-centric ontology evaluation are searched.

**Table 3.2:** Sub-queries for the overall search query  $Q = Q1SW \cap Q2HC \cap Q3Eval$  used for the SMS study selection

Sub-query	String of relevant search keywords
Q1SW	"semantic web" or ontolog* or vocabular* or "knowledge graph*" or "knowledge base*" or "linked data" or RDF or OWL or SPARQL
Q2HC	"Human computation" OR "human in the loop" OR Crowd* OR Layman OR Laymen OR Participant* OR Manual OR Microtask* OR "expert sourcing" OR Game* or gamification OR user* OR GWAP OR "expert evaluation" OR "expert review"
Q3Eval	Assessment OR evaluat* OR validat* OR verif* OR Error* OR Pitfall* OR Defect* OR Bias OR Quality OR Anomaly OR Refinement

The scope of the search was further limited to papers published between 2010 and 2020, and only literature in English was considered. The evaluation of the search strategy was done by a test-set of 5 papers previously identified as relevant and had to appear in the results.

<sup>4</sup><https://ieeexplore.ieee.org/search/advanced>

<sup>5</sup><https://dl.acm.org/search/advanced>



### Piloting of the study protocol

As part of the piloting of the protocol for the SMS 4 primary studies were identified and each paper was read by me and a senior researcher. Each reviewer extracted the data from the papers in an extraction template, defined in the study protocol. The goal of the pilot was, amongst other things, to determine whether the extraction template was clear and whether more or less information than defined should be extracted. During online meetings with all reviewers the results of the extractions were discussed, compared and conflicts were resolved.

### 3.1.2 Study Execution

As part of the Study Execution stage of the SMS, the thesis' main contributions were to the selection of studies based on meta-data and the extraction of data from a portion of the total papers identified for review.

#### Study Selection

From literature pre-identified as relevant from the senior researchers participating in the Systematic Literature Review project, papers addressing human-centric ontology evaluation were selected and fully read. Moreover, several surveys on ontology evaluation were read as well to gain further understanding of recent work in the field.

From the digital library searches for the Systematic Mapping Study (SMS), one batch of 240 papers was selected. For those papers, the title, abstract, and keywords were inspected. Based on a set of inclusion and exclusion criteria, shown in Table 3.3, defined for the SMS, a decision was made whether the paper should be included in the study. Additionally, for another batch of 240 papers, a second-reviewer evaluation of the selection was done for the papers for which the first reviewer could not come to an inclusion decision. From the same batch, each 10th paper was checked to get some insights on the inter-rater agreement between the evaluators. In total, for 84 papers a second-reviewer evaluation was performed.

**Table 3.3:** Inclusion and Exclusion Criteria defined for the SMS Study Selection

Criteria	Inclusion Criteria	Exclusion Criteria
C1 Publication Type	Primary studies subject to peer review which includes published journal papers, papers published as part of conference proceedings or workshop proceedings, book chapters.	Studies that are not subject to peer review, secondary studies.
C2 Language	Studies written in English.	Studies written in a language other than English

*Continued on next page*

### 3. HUMAN-CENTRIC ONTOLOGY EVALUATION: LITERATURE REVIEW

Table 3.3 – Continued from previous page

Criteria	Inclusion Criteria	Exclusion Criteria
C3 Accessibility	Studies available in full-text.	Studies not available in full-text.
C4 Duplicates	If a study has been published in more than one paper and presents the same results, the latest version of the study will be included.	If a study has been published in more than one paper and presents the same results, older version of the study will be excluded.
C5 Ontology Evaluation	Studies with a focus on the evaluation/validation/completion of a semantic resource.	Studies for which the focus is not on the evaluation/validation/completion of a semantic resource.
C6 Human-Centricity	Studies which report on involving human effort for performing the semantic resource evaluation/validation/completion task.	Studies which do not report on involving human effort for performing the semantic resource evaluation/validation/completion task

For papers found through snow-bowling techniques, the title, abstract, and keywords were reviewed in order to decide if the full text should be read and whether the paper should be included in the master thesis.

#### Data Collection

For each study, which was fully read, information about the human-centric evaluation task was extracted in a structured form following a pre-defined template. The template was constructed based on the first primary studies read and the goals and research questions of the SMS. Fields of the template include bibliographic information, type of verified resource, verified aspect, evaluation method, evaluator role, a frame of reference for the evaluation, etc. The extraction template including each data item and the research questions this information is needed for are shown in Table 3.4.

Table 3.4: Systematic Mapping Study Data Extraction template

ID	Data Item	Description	RQ
<i>Bibliographic Information</i>			
D1	Publication Title	Title of paper	

*Continued on next page*

Table 3.4 – Continued from previous page

ID	Data Item	Description	RQ
D2	Publication Year	Calendar year	
D3	Publication Type	Journal, conference, workshop, book chapter	
D4	Publication Venue	Conference name, book title, journal title	
D5	Author Affiliations	Research institutes/organizations and countries	
D6	Keywords	Keywords assigned to the publication by the authors	
D7	Paper Summary	Short summary of the paper (this is not the original abstract but rather a summary in the reviewer's words that captures those points of the paper which are relevant for this study but are not necessarily present in the original abstract).	
<i>Study Information</i>			
D8	Type of Verified Semantic Resource	Which type of semantic resource is the object of the verification task? D8a: Specify the type of resource with the term used by the author (e.g. ontology, a knowledge graph, a linked data fragment, a knowledge pattern, a part of an ontology? D8b: If available, extract the definition of the verified resource as given by the authors of the paper. D8c: If the paper focuses on a concrete/well-known resource (e.g., DBpedia) please provide the name of that resource. If there are more resources, separate their name by a comma. D8d: Assign a resource type from your perspective and according to our glossary.	SMS_RQ1
D9	Size and/or Number of Verified Semantic Resource	What is the size of the verified semantic resource? (e.g., number of classes, properties, instances, triples?) Alternatively, how many semantic resources were verified?	SMS_RQ1
D10	Formalism of Verified Semantic Resource	In what formalism is the resource represented? (e.g., RDF-S, OWL)	SMS_RQ1

*Continued on next page*

### 3. HUMAN-CENTRIC ONTOLOGY EVALUATION: LITERATURE REVIEW

Table 3.4 – *Continued from previous page*

ID	Data Item	Description	RQ
D11	Verified Aspect	What aspect of the semantic resource is verified? (usability, fit for task, domain relevance, modeling correctness, cognitive soundness, domain completeness, completeness with respect to competency questions, bias)	SMS_RQ2
D12	Frame of Reference	What is the frame of reference against which the evaluation is performed? (human domain knowledge, gold standard resource)	SMS_RQ2
D13	Type of Identified Error	What is the type of identified defect? E.g., missing (domain) information, modeling error, etc.	SMS_RQ2
D14	Motivation for Evaluation	What is the motivation for performing the evaluation of the resource? E.g., select a semantic resource, verify automatically extracted information; make sure a task (e.g., question answering, browsing) can be performed sufficiently well with the semantic resource.	SMS_RQ3
D15	Application Domain	What is the application domain considered? (e.g., biology, medicine, engineering)	SMS_RQ3
D16	Evaluator Role	What roles do human participants play in the evaluation task? What do the human evaluators do? E.g., search for error candidates; verify errors; search for & verify errors.	SMS_RQ4
D17	Evaluation Population Size	How many evaluators are involved in the evaluation?	SMS_RQ4
D18	Population Demographics	What are the demographic characteristics of the evaluator population? E.g., age, gender, nationality, etc	SMS_RQ4
D19	Population Domain Expertise	What is the familiarity of the evaluators with the subject domain covered by the resource? E.g., layman, medium, domain expert	SMS_RQ4
D20	Population Knowledge Modeling Expertise	What is the expertise of the evaluators in terms of knowledge modeling? E.g., layman, medium, domain expert	SMS_RQ4

*Continued on next page*

Table 3.4 – *Continued from previous page*

ID	Data Item	Description	RQ
D21	Population Professions	What are the current professional roles of the evaluators? E.g., university employees, students, crowd-workers (with certain acceptance rate of their HITs), other employees	SMS_RQ4
D22	Population Motivations	How are evaluators motivated to participate in the evaluation? (e.g., monetary reward, compulsory participation, voluntary participation, etc)	SMS_RQ4
D23	Bias	Which potential biases are present (population bias, methodological bias)? For population bias: is the population composition likely to suffer from bias? And if yes, which type of bias is likely present? (e.g., gender bias, nationality bias, age-bias, etc) D23a: describe why you think there is a bias	SMS_RQ4
D24	Bias addressed	Does the paper discuss potential biases and steps taken to avoid such biases? (e.g., topic completely ignored, topic considered but not addressed, topic actively addressed)	SMS_RQ4
D25	Evaluation Method	Which methods/protocols are used during the evaluation task? (e.g., games with a purpose, crowdsourcing, face-2-face workshops, user-based study, focus group, questionnaire, interview)	SMS_RQ5
D26	Evaluation Methodology	Which evaluation methodology is followed? Do the authors refer to a well-known methodology? If yes, which? Or do the authors propose a custom evaluation methodology? If yes, please briefly summarize it (we are interested in the main steps followed for evaluating the semantic structure with human evaluators).	SMS_RQ5
D27	Evaluation Modality	Which tools/modalities are used to gather the evaluation data? (e.g., pen&paper, blackboard, excel table, crowdsourcing platform, custom-built interface, game interface)	SMS_RQ5
D28	Inter-Evaluator Agreement	Is inter-evaluator agreement on the collected data considered and measured?	SMS_RQ5
D29	Evaluation metrics	Which evaluation metrics are computed on the data collected from human evaluators?	SMS_RQ5

*Continued on next page*

Table 3.4 – Continued from previous page

ID	Data Item	Description	RQ
D30	Other relevant papers	For snowballing	SMS_RQ5

### 3.1.3 Analysis & Reporting

For the portion of papers, for which the data extraction was completed, an analysis of the findings guided by the SMS research questions is presented in sections 3.2 - 3.7 and more concretely:

- Section 3.2 *Included Papers* provides an overview of the reviewed papers.
- Section 3.3 *Evaluated Resource (SMS\_RQ1)* offers an analysis of common aspects of the ontology, which require human-involvement for their evaluation.
- Section 3.4 *Evaluation Goal (SMS\_RQ2)* looks into the goals of the ontology evaluations and common problems of ontologies that have to be verified as well as what evaluation tasks have to be performed so they can be evaluated.
- Section 3.5 *Evaluation Context (SMS\_RQ3)* provides some background information and motivations as context in which ontology evaluations were performed.
- Section 3.6 *Evaluation Population (SMS\_RQ4)* analyses the types of participants who have performed the discussed evaluations.
- Section 3.7 *Evaluation Method (SMS\_RQ5)* provides an overview of how the identified human-centric ontology evaluation tasks have been solved until now and gives insights into what metrics have been used in the literature to determine the quality of the ontology evaluations.

## 3.2 Included Papers

In this section, the papers included in the literature review are presented. In Table 3.5 the titles of the analysed literature are listed together with an ID used as a reference throughout this chapter for a better overview. Additionally, in Appendix A: Included Papers for HOETs Analysis, a summary of each paper is included for some background information and context.

**Table 3.5:** Overview of the included papers.

ID	Paper Title	Pub. Year	Reference
P1	Engineering ontologies for question answering	2014	[Teitsma et al., 2014]
P2	Aemoo: Linked data exploration based on knowledge patterns	2017	[Nuzzolese et al., 2017]
P3	Crowdsourcing Linked Data Quality Assessment	2013	[Acosta et al., 2013]
P4	Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of SNOMED CT	2015	[Mortensen et al., 2015]
P5	Is the crowd better as an assistant or a replacement in ontology engineering? An exploration through the lens of the Gene Ontology	2016	[Mortensen et al., 2016]
P6	Crowdsourcing the verification of relationships in biomedical ontologies	2013	[Mortensen et al., 2013]
P7	Crowdsourcing Ontology Verification	2013	[Mortensen, 2013]
P8	BetterRelations: Using a game to rate linked data triples	2011	[Hees et al., 2011]
P9	WhoKnows? Evaluating linked data heuristics with a quiz that cleans up DBpedia	2011	[Ketterl et al., 2011]
P10	Collaboratively Patching Linked Data	2012	[Knuth et al., 2012]
P11	Finding errors in a Chinese lexico-semantic resource using GWAP	2017	[Zhang et al., 2017]
P12	ACRyLIQ: Leveraging DBpedia for Adaptive Crowdsourcing in Linked Data Quality Assessment	2016	[Ul Hassan et al., 2016]
P13	Interactive Refinement of Linked Data: Toward a Crowdsourcing Approach	2015	[Roengsamut and Kuwabara, 2015]
P14	Toward gamification of knowledge base construction	2015	[Roengsamut et al., 2015]
P15	Knowledge Base Refinement with Gamified Crowdsourcing	2016	[Kurita et al., 2016]
P16	ACTraversal: Ranking Crowdsourced Commonsense Assertions and Certifications	2011	[Chang et al., 2011]
P17	Crowd-based ontology engineering with the uComp Protégé plugin	2016	[Wohlgenannt et al., 2016]
P18	TripleCheckMate: A Tool for Crowdsourcing the Quality Assessment of Linked Data	2013	[Kontokostas et al., 2013]
P19	User-driven quality evaluation of DBpedia	2013	[Zaveri et al., 2013]

*Continued on next page*

Table 3.5 – Continued from previous page

ID	Paper Title	Pub. Year	Reference
P20	Subjective vs. objective evaluation of ontological statements with crowdsourcing	2015	[Erez et al., 2015]
P21	Achieving Expert-Level Annotation Quality with CrowdTruth The Case of Medical Relation Extraction	2015	[Dumitrache et al., 2015]
P22	Improving geo-spatial linked data with the wisdom of the crowds	2013	[Karam and Melchiori, 2013]
P23	Ontology enhancement using crowdsourcing: a conceptual architecture	2020	[Kiptoo, 2020]
P24	Ontology Evaluation - a pitfall-based approach to ontology diagnosis	2016	[Villalón and Pérez, 2016]

In the sections to follow, human-centric ontology evaluation tasks discussed in the the outlined papers are analysed.

### 3.3 Evaluated Resource (SMS\_RQ1)

Human-centric ontology evaluation is a broad topic. Sometimes it is not feasible to evaluate the whole ontology at once, or some parts of the ontology can be verified with automated methods. Therefore, researchers usually investigate a user-centric evaluation on a small part of the ontology. Research question SMS\_RQ1 investigates what the characteristics of the evaluated resource are - of what type the resource is and in what formalism it is represented in.

In Table 3.6 a summary of the type of evaluated resource and its verified aspects follows as well as the papers where such evaluation was considered.

**Table 3.6:** Type of evaluated resource and verified aspects of it

Type of Resource	Verified Aspect	Paper Reference
Ontology	structure domain completeness compliance with human thought efficiency of the ontology construction	P1
	usage of correct modeling techniques	P24
Ontology classes	domain relevance	P17
	overspecialisation	P24
	polysemous elements	P24

*Continued on next page*



Table 3.6 – Continued from previous page

Type of Resource	Verified Aspect	Paper Reference
Ontology properties	wrong data type	P3 P12 P18 P19
	missing relationship	P3 P13 P15 P22
	missing relationship type	P17
Ontology relationships	factual correctness	P3 P10 P12 P13 P14 P18 P19 P20 P21 P22
	correctness of subsumption (subClassOf) relationship	P4 P5 P6 P7 P17 P23
	correctness of instanceOf relationship	P17
	domain relevance	P18 P19
	perception of consensus and subjectivity	P20 P21
Weights of ontology relationships	importance of relationships	P8
	incorrect weights for relationships	P15
Weights of ontology properties	perceived importance	P9
Encyclopedic Knowledge Patterns (EKPs)	cognitive soundness	P2
	fit for task	
Lexico-semantic resources	correctness of translation	P11
Assertions <sup>6</sup> & Certifications <sup>7</sup> from GWAP	ranking of Assertions & Certifications	P16
Ontology evaluators	reliability	P12

From Table 3.6 it can be seen that the majority of research is conducted on the verification of ontology relationships. Some papers such as [Mortensen et al., 2016] look into the correctness of taxonomic relationships that were not explicitly specified by the ontology engineers but were driven through reasoning.

A large portion of papers looks into the completeness of ontologies based on missing relationships between concepts and the accuracy of the modeled domain taking into consideration if factually correct relationships are modeled. For instance, the authors from [Acosta et al., 2013] verify RDF triples that were extracted automatically from various sources and translated into RDF. As some types of data can be challenging to transform in RDF, errors could appear or some information might be left out.

Multiple papers concentrate on the importance of particular attributes or connections, which is an important aspect of some applications. In [Kurita et al., 2016] an FAQ system

<sup>6</sup> "commonsense knowledge is comprised of assertions, which are defined as "subject-relation-object" triples"[Chang et al., 2011]

<sup>7</sup> "the evidences indicating partial-order of associated assertion's confidence level"[Chang et al., 2011]

is developed, where a keyword must be matched to the best-fitting entry in the system, however, this does not make other mappings incorrect but instead less relevant.

A small number of papers looks into aspects such as cognitive soundness [Nuzzolese et al., 2017], compliance of human thought [Teitsma et al., 2014], or controversial interpretations [Erez et al., 2015] [Dumitrache et al., 2015].

Figure 3.2 shows in what formalism the evaluated resource was represented. The two most common variations are OWL<sup>8</sup> and RDF<sup>9</sup>. There are several papers that do not include information about the formalism of the resource that is verified. In [Teitsma et al., 2014], 3 ontologies are presented, one is represented as OWL, one as SKOS<sup>10</sup> and for one the representation is not mentioned at all. Zhang et al. [Zhang et al., 2017] evaluate language mappings represented as UKC<sup>11</sup>, which accommodate multi-language lexico-semantic resources. The authors in [Dumitrache et al., 2015] evaluate a number of English sentences extracted from medical relations. [Chang et al., 2011] evaluates the AC graph consisting of assertions and certifications, represented as “subject-relation-object” triples and their confidence level.

## 3.4 Evaluation Goal (SMS\_RQ2)

Research question SMS\_RQ2 focuses on the goal of the evaluation. Important aspects are for instance the frame of reference against which the evaluation is performed and the type of problem to be identified, which are discussed in detail in the following sections.

### 3.4.1 Frame of Reference

In this section, the frame of reference against which the evaluation is performed is discussed. It is important to note that several of the papers use a combination of the different approaches as the verified resource is evaluated along with multiple aspects and criteria. Figure 3.3 shows percentage-wise what frame of reference was used in the papers and Table 3.7 lists exactly in which paper the frame of reference was applied.

Most of the papers compared the user-evaluation against experts’ domain knowledge or majority agreement amongst the participants. As discussed in section 3.3 *Evaluated Resource (SMS\_RQ1)* a large portion of the evaluations were carried out on ontology relationships and their correctness. It is very challenging and often not feasible to create a gold standard for the domain ontology as the resources are very large, complex, dynamically changing, or even controversial.

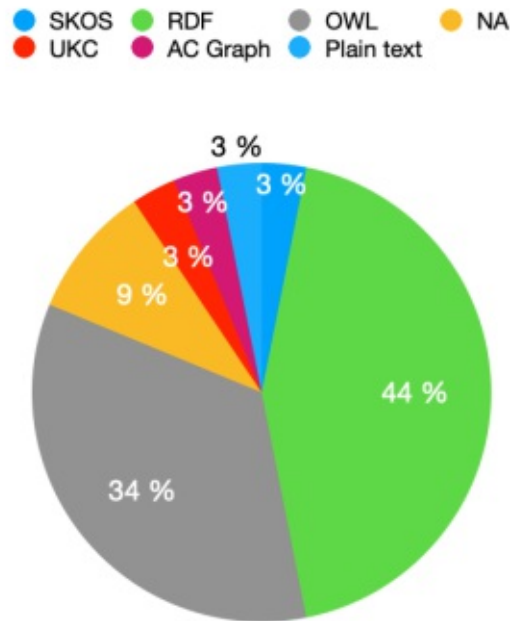
---

<sup>8</sup>"The W3C Web Ontology Language (OWL) is a Semantic Web language designed to represent rich and complex knowledge about things, groups of things, and relations between things." [OWL, 2012]

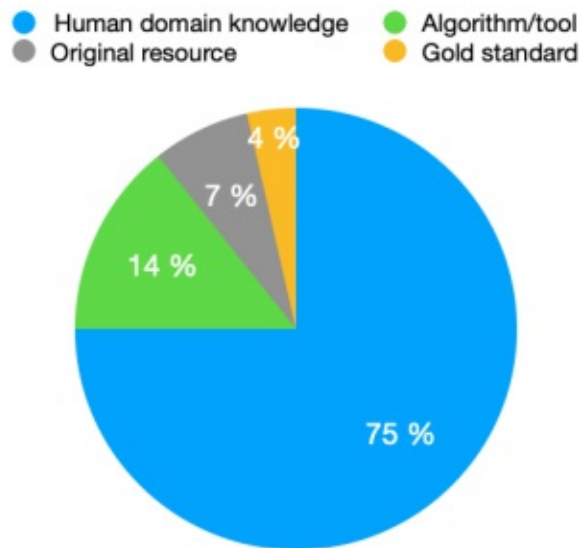
<sup>9</sup>Resource Description Framework:"RDF is a standard model for data interchange on the Web." [RDF, 2014]

<sup>10</sup>SKOS Simple Knowledge Organization System RDF Schema: "a common data model for sharing and linking knowledge organization systems via the Web. [SKO, 2009]"

<sup>11</sup>"The Universal Knowledge Core (UKC) is a psycholinguistic principles based multilingual, high quality, large scale, and diversity aware machine readable lexical resource." [UKC, 2020]



**Figure 3.2:** Formalism of the evaluated resources.



**Figure 3.3:** The frame of reference against which the evaluation is performed

However, in [Erez et al., 2015], a gold standard is created using scientific articles, government websites, and other verified resources. The gold standard was created by a

panel of information specialists, who also included different viewpoints in it.

On the other hand, [Mortensen, 2013] and [Wohlgenannt et al., 2016] used the original resource as a gold standard. This was possible because they planted the errors in the verified ontology manually.

Some authors compared or combined the results from the user evaluation with those of an automated algorithm or method. For instance, the performance of the exploratory search tool Aemoo[Nuzzolese et al., 2017], implemented using EKPs was compared to Google & RealFinder<sup>12</sup> search results.

[Acosta et al., 2013] combines and compares human domain knowledge results for verifying the correctness of ontology relationships with an implemented baseline algorithm which verifies if the original web page from which the resource was extracted contains some information about the relationship.

**Table 3.7:** The frame of reference against which the evaluation is performed.

Frame of Reference	Paper Reference
Original resource	P7 P17
Gold standard	P20
Algorithm/tool	P2 P3 P12 P16
Human domain knowledge	P1 P2 P3 P4 P5 P6 P8 P9 P10 P11 P12 P13 P15 P16 P17 P18 P19 P21 P22 P23 P24
Human modeling knowledge	P24

### 3.4.2 Ontology Quality Issues and Human-Centric Ontology Evaluation Tasks

Based on the included literature, a catalog of ontology issues is listed in Table 3.8 together with the **human-centric ontology evaluation tasks** (HOETs), which need to be performed to solve them. Included in Table 3.8 is also the human knowledge needed for completing the task and the role of the evaluators.

Based on the results, we can group the ontology problems to be identified into 4 main categories - incorrect information, missing information, cognitive defects, and incorrect modeling.

Most of the research papers focus on the first area and propose methods of how incorrect information (data types, relations, taxonomic structures, translations, etc.) can be detected. The most common assignments the contributors had to complete were True/False or multiple choice questions on a specified domain provided some definitions for context.

<sup>12</sup><http://www.visualdataweb.org/refinder.php>

Studies that focus on the second category - missing information, aim to either detect missing information in the evaluated semantic resources, for instance, by comparing a requirement document and the resource to be evaluated or to extend the resource further with the help of collective intelligence. In order to allow for missing information to be collected, evaluators had to define keywords or tags, provide relevant concepts for a domain of interest, define relation types, or rate the importance of relations.

A few papers give attention to semantic defects included in ontologies - identifying irrelevant information modeled, outlining concepts not compliant with how humans think, detecting viewpoints or controversial statements. Typical roles the evaluators had to perform so that the issue could be detected are judgments of whether a concept/relation relevant to a specific domain, finding outsider concepts in a provided set, or deciding whether a given statement is controversial.

All evaluation tasks from these 3 categories require the evaluators to have some domain knowledge to be able to perform the tasks successfully. The last problem category which focuses on incorrect modeling requires the participants to have modeling knowledge in addition to the needed domain knowledge. For this category, there is limited research available on how the problems can be tackled, however, the evaluators' role is to detect the incorrect modeling provided a specific resource.

**Table 3.8:** Ontology mistakes and how they can be solved

Problem	Evaluation Task	Evaluator Role	K	Paper Reference
<i>Quality Problem: Incorrect Information</i>				
incorrect subclasses	identify wrong taxonomic relationships	decide if a statement is true or false (provided some definitions for context)	DK	P5 P5 P6 P7 P17
		decide what is the parent of a sub-concept	DK	P23
wrong instanceOf relations	identify wrong instances	decide whether the instanceOf relation between a class and the individual is correct	DK	P17

*Continued on next page*

K = required knowledge for completing the task  
 DK = domain knowledge, MK = ontology modeling knowledge

### 3. HUMAN-CENTRIC ONTOLOGY EVALUATION: LITERATURE REVIEW

Table 3.8 – *Continued from previous page*

Problem	Evaluation Task	Evaluator Role	K	Paper Reference
incorrect domain information	identify wrong information modeled in the ontology	answer multiple choice questions (and report odd ratings)	DK	P9 P10 P13
		decide if a statement is true or false (provided some definitions for context)	DK	P3 P12 P16
		decide whether a triple contains wrong information	DK	P18 P19
	identify wrong tags	decide whether a value has a correct tag	DK	P12
wrong mappings of multi-language concepts	identify wrong translation of concepts	select the correct translation for a given concept	DK	P11
wrong weights of relations	identify wrong weights	decide on the most relevant answer for a multiple choice question	DK	P14 P15
wrong data types used	identify wrong data types	for a given triple decide whether it contains wrong data types	DK	P18 P19
<i>Quality Problem: Missing Information</i>				
missing domain information	define the gold standard concepts	write down relevant concepts after watching domain videos	DK	P1
	complete missing information	define keywords/tags for a statement	DK	P13 P15 P16 P17

*Continued on next page*

K = required knowledge for completing the task  
 DK = domain knowledge, MK = ontology modeling knowledge

Table 3.8 – Continued from previous page

Problem	Evaluation Task	Evaluator Role	K	Paper Reference
missing domain information	complete missing information	define the type and direction of a relationship	DK	P21
	identify missing information	given some system requirements identify missing information	DK	P24
missing weights of relations	find the perceived importance for relations	rate the importance of an object to describe a subject	DK	P2
		for a subject choose the more relevant object out of two	DK	P8
<i>Quality Problem: Cognitive Defects</i>				
domain irrelevant information	identify irrelevant information	given a concrete domain, decide whether a concept is relevant	DK	P17
		given a triple decide whether it contains irrelevant information	DK	P18 P19
concepts not used by humans	outline concepts not compliant with human thought	selecting the outsider concept out of 3	DK	P1
controversial domain information	identify controversial statements	given a statement decide whether it is true, false or controversial	DK	P20
	verify information is consistent	decide whether a statement is consistent	DK	P22

*Continued on next page*

K = required knowledge for completing the task

DK = domain knowledge, MK = ontology modeling knowledge

Table 3.8 – Continued from previous page

Problem	Evaluation Task	Evaluator Role	K	Paper Reference
polysemous elements used	detect polysemous elements	detect polysemous elements	DK	P24
<i>Quality Problem: Incorrect Modeling</i>				
wrong modeling technique used	identify the incorrect usage of restrictions	detect whether owl:allValuesFrom is used instead owl:someValuesFrom	DK & MK	P24
	detect incorrect usage of “some not” and “not some”	identify if "some not" is used in place of "not some"	DK & MK	P24
	detecting incorrect usage of classes	identify whether a primitive class is used instead of a defined one	DK & MK	P24
	detect duplication of a data type	identify if data type are created even though they were already included in the language	DK & MK	P24
	detect overspecialisation in hierarchies	detect whether "instanceOf" is used instead of "sub-ClassOf"	DK & MK	P24

K = required knowledge for completing the task

DK = domain knowledge, MK = ontology modeling knowledge

### 3.5 Evaluation Context (SMS\_RQ3)

Research question SMS\_RQ3 puts focus on the context in which the evaluation is performed - what is considered application domain and what is the motivation behind performing the evaluation.

To understand in which concrete context human-centric ontology evaluation tasks occur, the application domains considered in the included research are listed in Table 3.9. For understanding the scenarios of the ontology evaluation, the motivation behind it is included as well.



**Table 3.9:** Context for performing human-centric ontology evaluation tasks

Application domain	Evaluation Motivation	Paper Reference
General Knowledge	verify automatically extracted resources	P2 P9
	fit for task	P2 P9
	gather importance ratings as perceived by an average human	P8
	collecting issues in heterogeneous resource and tracking their origins	P10
	improving the precision of assessment results	P12 P16
Medical Science	maintenance of large and/or growing complex ontologies	P4 P5 P6 P7
	include viewpoints/interpretations in an ontology	P21
	extract information from user for ontology refinement	P23
	outsourcing verification tasks during the ontology engineering process	P17
Linguistics	improving the precision of assessment results	P12
	verify automatically extracted resources	P11
FAQ System	extract information from user for ontology refinement	P13 P14
	fit for task	P15
Crisis situation management	fit for task	P1
	choose an ontology	
Climate change	outsourcing verification tasks during the ontology engineering process	P17
Finance	outsourcing verification tasks during the ontology engineering process	P17
Wine	outsourcing verification tasks during the ontology engineering process	P17
Tennis	outsourcing verification tasks during the ontology engineering process	P17
Diet	include viewpoints/interpretations in an ontology	P20
Geo data	verify automatically extracted resources	P22
No concrete domain specified	verify automatically extracted resources	P3 P19
	assessing the quality of resources published on the Web	P18
	collecting common ontology pitfalls	P24

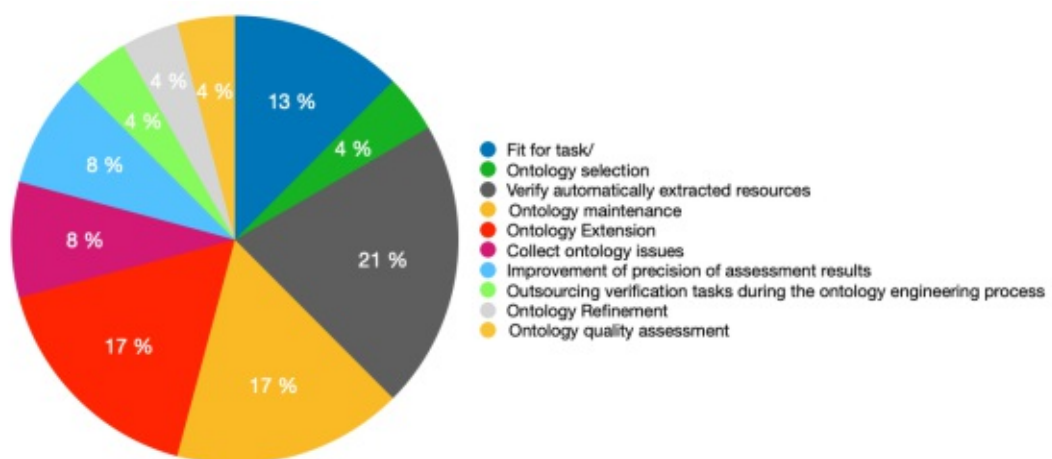
A portion of the papers looks into the medical science domain. There are several large

and complex medical ontologies, which consist of mainly taxonomic relationships. A part of the subsumption relationships are not specified explicitly but instead are created indirectly as a consequence of other modeled axioms. To ensure the quality of the ontology during its growth a user-centric evaluation is required. Expert-evaluations are not always feasible, especially, for large ontologies and therefore a crowdsourcing approach is preferred.

Many papers also consider a general knowledge domain. Some of the ontologies that model commonsense knowledge are extracted automatically and require verification. These types of resources can easily be evaluated by casual users, without the help of experts. Moreover, some ontologies need to be further enhanced with information that can only be obtained from human knowledge.

There are also some papers that did not define a specific application domain. In those papers, only the verification approach was discussed and no explicit study was conducted for a concrete domain.

To understand more about the motivation behind human-centric ontology evaluations Figure 3.4 also visualises the percentage of papers that are motivated by a specific aspect. We see that the most common reason why the evaluations are performed is the verification of automatically extracted resources. This shows again that a complete automation is not always reliable and therefore it is needed to verify such methods by including human participants in the process. With a very high percentage are also the ontology maintenance of complex and growing ontologies such as Medical Taxonomies and the ontology extension with some human domain knowledge such as interpretations or viewpoints.



**Figure 3.4:** Motivation for human-centric ontology evaluation tasks.

### 3.6 Evaluation Population (SMS\_RQ4)

A key aspect of the human-centric evaluation is the human population, performing it, which is investigated in order to answer research question SMS\_RQ4. A portion of the included papers (6 out of 24) only present an evaluation approach, tool, or method and do not report on an actual evaluation with human participants. Those papers will therefore be omitted from the population analysis, which follows.

To understand more about the population, and what characteristics it has, Table 3.10 summarises the most important facts. Also included is the type of motivation the participants had to join the experiments. Based on the information included in Table 3.10 it is possible to identify possible biases or specific viewpoints.

Many of the papers did not include much information about the evaluators - where they are from, what professions they have or what expertise they have. Especially, when working with crowd workers such information is not known as the tasks are distributed to anonymous users. Crowdsourcing platforms usually offer the possibility to restrict the origin or other characteristics of the workers, however, in almost no paper a restriction set-up is mentioned. Only in [Wohlgenannt et al., 2016], the authors say the crowd workers come from English-speaking countries. Some researchers overcome the problem of unknown domain expertise by requiring users to pass some qualification tests at the beginning.

Only one of the papers [Nuzzolese et al., 2017] considers bias when discussing the evaluation. In their scenario, EKPs were to be evaluated according to their usability for exploratory searches by using a custom-built tool. The authors compare the results against other search engines - Google and ReFinder. They argue that there might be some bias present as most participants have already had some experience with the other tools. They attempt to solve this issue, by assigning an equal number of users to start their tasks using Aemoo to the number of participants that start with one of the other search engines.

In [Wohlgenannt et al., 2016] only workers from English-speaking countries (United States, United Kingdom, and Australia) are considered. It is possible that as a consequence some bias was introduced into the evaluation.

**Table 3.10:** Evaluation population

Paper Reference	Size	Demo-graphics	Domain Expertise	Knowledge Modeling	Profession	Motivation
P1	21	NA	NA	NA	students	NA

*Continued on next page*

NA = the information is not mentioned in the paper

### 3. HUMAN-CENTRIC ONTOLOGY EVALUATION: LITERATURE REVIEW

Table 3.10 – Continued from previous page

Paper Reference	Size	Demo-graphics	Domain Expertise	Knowledge Modeling	Profession	Motivation
P2	17	NA <sup>13</sup>	NA	NA	NA	NA
	32	NA <sup>14</sup>	NA	NA	students	NA
P3	58	NA	NA	expert	members of Linking Open Data and DBpedia communities	monetary
	50	NA	NA	NA	crowd workers	monetary
P4	25	the paper reports that there was a diversity amongst the workers	NA	NA	crowd workers	monetary
P5	5	NA	expert	expert	5 of the paper authors	NA
	5	NA	NA	NA	crowd workers	monetary
P6	320	NA	layman/medium <sup>15</sup>	NA	crowd workers	monetary
P7	40	NA	medium <sup>16</sup>	novice <sup>17</sup>	crowd workers	monetary
P8	359	NA	NA	NA	NA	game points

*Continued on next page*

NA = the information is not mentioned in the paper

<sup>13</sup>the paper reports that the participants had different cultures and languages, as well as different skills

<sup>14</sup>the paper only reports that the participants study in Italy or in France

<sup>15</sup>depending on the setup, crowd workers had to pass a high school-level biology qualification task

<sup>16</sup>crowd workers to pass qualifications test in biology and medicine

<sup>17</sup>crowd workers had to pass qualification tests in Ontology modeling

Table 3.10 – Continued from previous page

Paper Reference	Size	Demographics	Domain Expertise	Knowledge Modeling	Profession	Motivation
P9	165	the paper reports a diversity in age, gender, origin, social background	NA	NA	NA	game points
P11	24	NA	layman/medium <sup>18</sup>	NA	NA	game points & language skills practice
P12	60	NA	NA	NA	NA	monetary
P14	35	NA	NA	NA	students	game rankings and scores
P16	46 <sup>19</sup>	NA	NA	NA	students	NA
P17	5-8	NA	layman	expert	NA	NA
	5	from English speaking countries (UK, USA, Australia)	NA	NA	crowd workers	monetary
P19	60	NA	NA	medium <sup>20</sup>	researchers	NA
P20	40	NA	NA	NA	crowd workers	monetary
	NA <sup>21</sup>	NA	NA	NA	information specialists	NA

*Continued on next page*

NA = the information is not mentioned in the paper

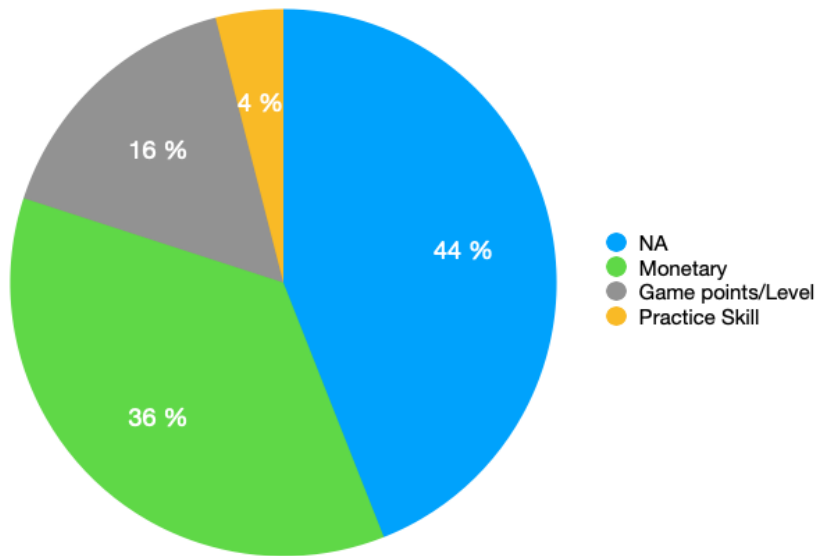
<sup>18</sup> English proficiency level of the participants is either College English Test (CET2) 4 or 6<sup>19</sup> 46 participants in total and at least 3 per statement<sup>20</sup> the participants were familiar with RDF<sup>21</sup> the authors only say it was a small group

### 3. HUMAN-CENTRIC ONTOLOGY EVALUATION: LITERATURE REVIEW

Table 3.10 – Continued from previous page

Paper Reference	Size	Demo-graphics	Domain Expertise	Knowledge Modeling	Profession	Motivation
P21	1	live in USA	inter-mediate <sup>22</sup>	NA	students	NA
	10-15	NA	NA	NA	crowd workers	NA
P23	30	NA	NA	NA	crowd workers	NA

NA = the information is not mentioned in the paper

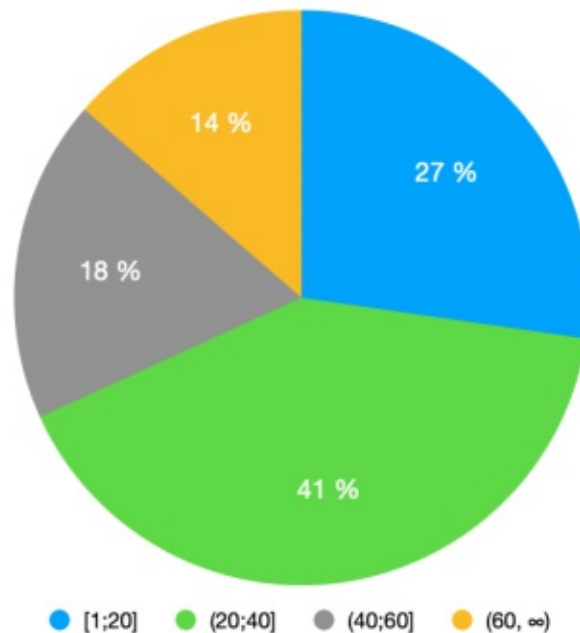


**Figure 3.5:** Contributors' motivation to participate in the evaluation tasks.

Based on the results presented in the above table a chart with the most frequent motivation of the contributors was created and can be seen in Figure 3.5. We see that in the studies where the motivation of the participants was disclosed, the most frequent motivator is monetary compensation. A portion of the evaluations, especially those which were conducted in the form of a game as we will discuss in the next section, were motivated by the earnings of game points or upgrading to a higher level in the game. Only one of the studies reported that the participants in the evaluation took part to practice and improve a specific skill.

<sup>22</sup>"medical students, in their third year at American universities that had just taken United States Medical Licensing Examination (USMLE)[Dumitrache et al., 2015] "

In Figure 3.6 the size of the evaluation populations is analysed. As it can be seen most evaluations (41%) were conducted with 20-40 participants, many studies performed the evaluation with even fewer contributors. On the other hand, using more than 40 participants is not very common judging by the analysed studies. Human Computation & Crowdsourcing rely on multiple replies from contributors to be able to extract the wisdom of the crowds, however, one of the goals of those techniques is to perform a cost-efficient evaluation since expert evaluations can become cost-intensive. As we previously saw evaluators are usually motivated by some monetary rewards, therefore, using too large of a crowd can become as expensive as an expert evaluation and can thus lose some of the benefits of the HC methodology. Therefore as shown in Figure 3.6 a typical population consists of no more than 40 contributors.



**Figure 3.6:** Size of the used evaluation population.

Figure 3.7 shows the frequency of the professions which the evaluators have. As mentioned above in many studies (25%) information on the contributors' background was not discussed and this can clearly be seen also from the chart. We also see that crowd workers are the participants with which most evaluations are conducted. Another commonly used population are student crowds, while experts such as researchers, study authors or Linked Data community members are rarely chosen.

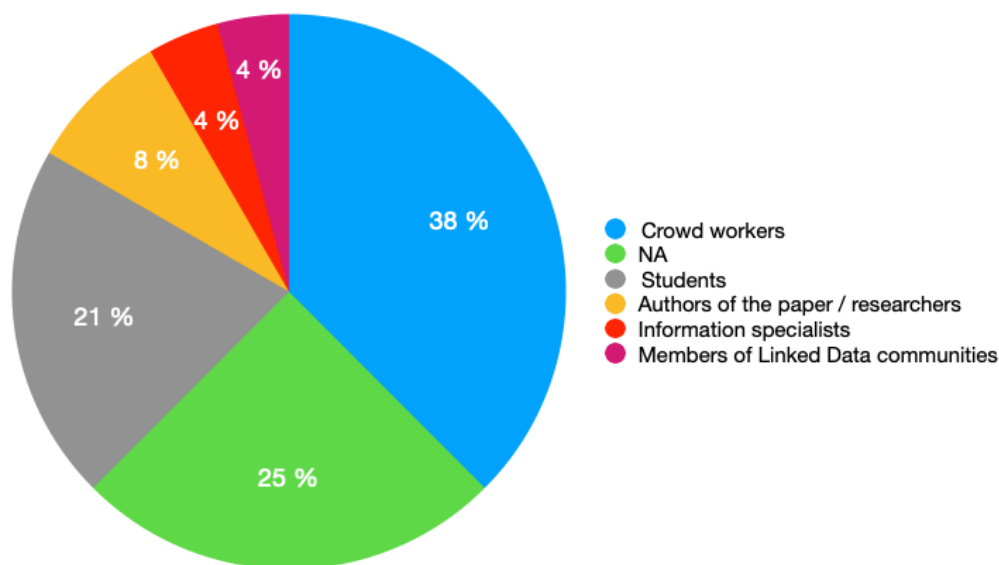


Figure 3.7: Professions of the participants performing the HOETs.

### 3.7 Evaluation Method (SMS\_RQ5)

The last research questions SMS\_RQ5 defined for the Systematic Mapping Study aims to answer how human-centric ontology evaluation has been previously performed - what methods were followed, what modalities were used, what metrics were computed, and was inter-rater agreement considered. Those aspects are described in detail in the next sections. As mentioned a portion of the reviewed papers only focuses on an evaluation approach, tool, or method and does not report on an actual evaluation with human participants. Therefore, those papers are not included in the evaluation method analysis.

#### 3.7.1 Evaluation methods and modalities

In this section, further analysis of the human-centric ontology evaluation tasks listed in Table 3.8 follows. Table 3.11 lists all methods and modalities that were already used in the literature to solve each task.

Many of the gathered human-centric ontology evaluation tasks have been approached using crowdsourcing platforms or game interfaces. A large number of the identified tasks have also been addressed multiple times in the literature and different solutions have been proposed. Furthermore, tasks focusing on semantics such as creating a gold standard or identifying concepts, not compliant with how humans think, have not been yet solved using Human Computation techniques- currently, only user studies have been conducted.

We see in Table 3.11 that the evaluation tasks of the incorrect modeling quality problem are not included. The reason for this is that such tasks have not been yet approached, but were only identified as issues requiring human input in the literature. Based on

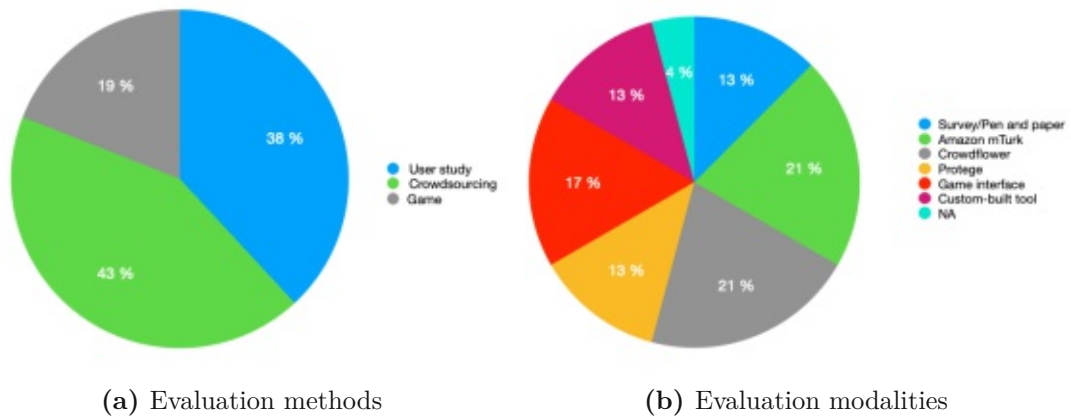


the professions of the contributors that we saw in the last section we can outline that the evaluations were mostly done with layman or novice crowds. These results could explain why only evaluation tasks were approached, which do not require the evaluators to have any modeling knowledge but only some domain or general knowledge. Modeling knowledge can be difficult to collect in a crowdsourcing platform and attracting qualified workers could become costly. This might be one of the reasons why tasks outlined in the previous sections that require modeling knowledge have not yet been solved using HC techniques.

**Table 3.11:** Ontology verification tasks and how they have been approached in the literature

Evaluation task	Evaluation Method	Evaluation Modality	Paper Reference
<i>Quality Problem: Incorrect Information</i>			
identify wrong taxonomic relationships	crowdsourcing & user study	Amazon Mechanical Turk & user surveys & CrowdFlower & Protege & custom-built-tool	P4 P5 P6 P7 P17 P23
identify wrong instances	crowdsourcing & user study	user surveys & CrowdFlower	P17
identify wrong information modeled in the ontology	crowdsourcing & game	Amazon Mechanical Turk & game interface & custom-built-tool	P3 P9 P10 P12 P13 P18 P19
identify wrong tags	crowdsourcing	Amazon Mechanical Turk	P12
identify wrong translation of concepts	game	game interface	P11
identify wrong weights	game	game interface	P14 P15
identify wrong data types	crowdsourcing	custom-built-tool & Amazon Mechanical Turk	P18 P19
<i>Quality Problem: Missing Information</i>			
define the gold standard concepts	user study	watching videos & pen and paper	P1
complete missing information	user study & crowdsourcing	CrowdFlower & Protege	P13 P15 P17 P21

*Continued on next page*



**Figure 3.8:** Applied evaluation methods and modalities for solving HOETs.

Table 3.11 – Continued from previous page

Evaluation task	Evaluation Method	Evaluation Modality	Paper Reference
find the perceived importance for relations	user study & game	pen and paper & game interface	P2 P8
<i>Quality Problem: Cognitive Defects</i>			
identify irrelevant information	user study & crowdsourcing	CrowdFlower & Protege	P17 P18 P19
outline concepts not compliant with human thought	user study	no information in the paper	P1
identify controversial statements	crowdsourcing & user study	CrowdFlower	P20
verify information is consistent	crowdsourcing	game interface & Amazon Mechanical Turk	P22

In Figure 3.8a the percentage of the most commonly used evaluation methods and evaluation modalities to solve HOETs are visualised. It can be seen that there are 3 ways how such tasks have been approached - with crowdsourcing, game sourcing, or a traditional user study. As also mentioned previously, crowdsourcing shows to be the widely used method for solving human-centric ontology evaluation tasks. Figure 3.8b also presented the commonly used modalities of the evaluation. The chart shows that Amazon Mechanical Turk and Crowdflower, which are both crowdsourcing platforms, is where over 40% of the tasks have been performed.

### 3.7.2 Evaluation metrics and Inter-evaluator agreement

To further understand how human-centric verification tasks are approached in the literature Table 3.12 lists for each of the included papers that perform a human-centric evaluation the evaluation metrics that were computed. An important metric is the inter-evaluator agreement, which shows the degree of agreement among evaluators, therefore it is presented in a separate column for a better overview.

**Table 3.12:** Evaluation metrics and inter-evaluator agreement

Paper Reference	Evaluation Metrics	Inter-evaluator agreement
P1	number of concepts, path lengths, coverage and precision, entropy, Ingve–Miller number, semantic distance <sup>23</sup> , efficiency (time to construct the ontology)	no
P2	Spearman’s rank correlation coefficient <sup>24</sup> between the assigned scores by the participants and those extracted automatically, number of correct solutions per minute given by the participants for each task and on average	yes
P3	precision against gold standard (& baseline algorithm), number of true positives and false positives	yes
P4	performance on the consensus standard, and estimated cost	yes
P5	performance of the crowd workers using Area Under the Receiver Operating Characteristic curve (AUC) <sup>25</sup> , t-tests with Benjamini-Hochberg false discovery rate correction, two-way ANOVA <sup>26</sup> , Wilcoxon rank-sum test <sup>27</sup>	yes (only between experts, but not for crowd workers)
P6	performance of the crowd workers using AUC, accuracy, sensitivity, and specificity, worker performance in comparison to random (using Fisher’s exact test )	no

*Continued on next page*

<sup>23</sup>“the match between the formal semantic distance of concepts within an ontology and the cognitive semantic distance of the same concepts as perceived by the participants”[Teitsma et al., 2014]

<sup>24</sup>“This measure gives us an indication on how precisely DBpedia wikilinks allow us to identify EKPs as compared to those drawn by the users”[Nuzzolese et al., 2017]

<sup>25</sup>“This measure ranges from 0 to 1 and captures how well the methodology performed at identifying the correct and incorrect relationships listed in the expert-based consensus standard at various probability thresholds”[Mortensen et al., 2016]

<sup>26</sup>used to “to understand the relative contributions of each factor has on the variability of the AUC”[Mortensen et al., 2016]

<sup>27</sup>“a non-parametric test, for “Google-ability” comparisons because”[Mortensen et al., 2016]

### 3. HUMAN-CENTRIC ONTOLOGY EVALUATION: LITERATURE REVIEW

Table 3.12 – *Continued from previous page*

Paper Reference	Evaluation Metrics	Inter-evaluator agreement
P7	worker performance in terms of time and accuracy, average performance of workers with and without qualification test and with and without the context	no
P8	Throughput, average lifetime play and expected contribution as, ratings based on relative decisions between pairs of items, comparison to a manually created gold standard	no
P9	property rankings	no
P11	users' playing pattern( average honest pattern, average accuracy)	no
P12	average accuracy, overhead costs, average reliability of workers	no
P14	performance of both the users and the game, enjoyment of playing the game scale from 1-5	no
P16	precision improvement after ACT <sup>28</sup> ranking, the efficiency of ACT	no
P17	time to perform the task, cost benefit, quality, usability of the plugin	yes
P19	number of distinct resources without issues, number of distinct resources with issues, number of distinct incorrect triples, number of distinct incorrect triples in the dbprop namespace, number of inter-evaluations; number of resources with evaluator disagreements, number of triples verified as correct, number of triples not evaluated correctly, percentage of correctly verified triples, Average number of issues per resource, average number of issues per resource in the dbprop namespace, percentage of affected triples, percentage of affected triples in the dbprop namespace"	yes
P20	accuracy (compared to a gold standard annotation, created based on scientific literature), percentage of correct judgments from crowd workers, accuracy of the majority vote and most popular vote among crowd workers, performance calculated with AUC	no

*Continued on next page*

<sup>28</sup>ACTraversal is a universal graph traversal aggregation for ranking common sense assertions and certification." [Chang et al., 2011]

Table 3.12 – *Continued from previous page*

Paper Reference	Evaluation Metrics	Inter-evaluator agreement
P21	CrowdTruth metrics <sup>29</sup> , sentence vector, sentence-relation score	yes
P23	amount of times a tag was made for a resource, number of missing features from the ontology that were identified from the crowds, number of features from the ontology not tagged by the crowd	no

As we see in the table most widely used metrics to evaluate the results from the contributors' verifications are performance in terms of accuracy of the results and time needed to perform the judgments, cost benefit, and influence of qualification tests on the results. Many researchers evaluate the performance in terms of true and false positives with AUC. For evaluations performed via a game-interface metrics used were also the enjoyment while playing, playing patterns, and time spent in the game.

Inter-rater agreement was considered as an evaluation metric in 35% of the discussed papers. However, some studies used it to analyse only a part of the evaluation method. For instance, when a combination of expert-evaluation and a crowdsourcing approach were combined, the inter-rater agreement was measured only amongst the experts. The most commonly used metrics to measure the inter-evaluator agreement were Kendall's coefficient and Fleiss' Kappa coefficient.

### 3.8 Literature Review Summary

In this chapter, we looked into known human-centric ontology evaluation tasks and their characteristics. In Table 3.8 a list of such tasks and the ontology aspect which they are related to were presented. We were able to outline 4 major quality issues that ontologies might have and require human input for their resolution - incorrect information, missing information, cognitive defects, and incorrect modeling. The most common motivation for the need for the evaluation of those issues is the verification of automatically extracted resources and as the frame of reference mostly, human domain knowledge was used.

The literature review also presented known approaches for solving the identified tasks. Crowdsourcing was found to be the most widely used method for the evaluations and based on the population demographics analysis the participants were mostly layman or novice crowds.

<sup>29</sup>"CrowdTruth metrics model quality at each vertex in relation to all the others"[Dumitrache et al., 2015]

**Table 3.13:** Ontology verification tasks that have not yet been approached using Human Computation techniques in the literature

Evaluation task	Method	K	Paper Reference
<i>Quality Problem: Missing Information</i>			
define the gold standard concepts	user-study	DK	P1
identify missing information	-	DK	P24
<i>Quality Problem: Cognitive Defect</i>			
outline concepts not compliant with human thought	user-study	DK	P1
detect polysemous elements	-	DK	P24
<i>Quality Problem: Incorrect Modeling</i>			
identify the incorrect usage of restrictions	-	DK&MK	P24
detect incorrect usage of “some not” and “not some”	-	DK&MK	P24
detect incorrect usage of classes	-	DK&MK	P24
detect duplication of a data type	-	DK&MK	P24
detect overspecialisation in hierarchies	-	DK&MK	P24

K = required knowledge for completing the task

DK = domain knowledge, MK = ontology modeling knowledge

[Villalón and Pérez, 2016] identifies multiple mistakes in ontology, requiring human verification, which have not been yet approached with Human Computation. Those tasks require not only some domain information but also require the evaluators to have some information modeling understanding as well. These tasks are listed in Table 3.13 together with tasks only approached with user-studies to provide a better overview of the task where a human-computation approach is still missing.

Since in the reviewed literature no evaluation task was approached for which both domain knowledge and modeling knowledge are required, the next chapters of the thesis will focus on one such task - the verification task of identifying incorrect usage of ontology restrictions, which was outlined in paper P24 [Villalón and Pérez, 2016]. In the next chapters, a Human Computation approach for solving the task is designed and implemented. To evaluate how the proposed method performs a student-experiment is conducted and the results are discussed.

# Ontology Restrictions Verification: a Human Computation Approach

This chapter investigates RQ2 in detail in order to determine what an appropriate Human Computation solution would be to solve one specific human-centric ontology evaluation task, not yet approached with such techniques.

**RQ2** How can Human Computation techniques be used to evaluate ontologies regarding the correct usage of restrictions?

The previous chapters already provided some insights into the importance of Ontology Evaluation. Wrongly represented facts or included biased information in ontologies could lead to failures of the systems using them. Many quality issues of ontologies can be identified via automated methods, however, some problems require domain and/or modeling knowledge and can only be solved with the help of some human input. One such pitfall is the misuse of the universal and existential quantifiers.

Firstly in section *4.1 Misuse of the Universal and Existential Restrictions*, the difference between the ontology restrictions is introduced and common mistakes in their usage are outlined. Afterwards, a Human Computation approach for verifying the correct usage of the ontology quantifiers is outlined in section *4.2 Ontology Restrictions Verification - a Human Computation Approach*.

## 4.1 Misuse of the Universal and Existential Restrictions

In order to implement a Human Computation solution for the verification task of evaluating the correct usage of ontology restrictions, one needs to first understand the

difference between the existential and universal quantifiers and investigate where possible mistakes might occur. The existential ( $\exists$ ) restriction indicates that there must be at least one property of the restricted type while other types are not restricted. On the other hand, the universal ( $\forall$ ) one indicates that all values of the restricted property must be of no other than a certain type but a property value does not have to exist.

Several studies have indicated that the use of these quantifiers is not trivial and is often linked to defects in ontologies. In the OOPS catalog [Villalón and Pérez, 2016], the authors identify the mistake that beginners often use the universal restriction as a default quantifier, instead of the existential restriction. [Rector et al., 2004] investigates the ignorance of the Open World assumption - the fact that information not explicitly stated is not incorrect unless it contradicts other axioms from the ontology. [Rector et al., 2004] and [Warren et al., 2019] also identify the common misconception that the universal restriction implies the existential restriction or in other words forgetting that when using the universal restriction the possibility of no property value existing is included as well. This often forgotten rule is also referred to as the “trivial satisfaction of the universal restriction”.

Below frequent mistakes in the usage of the ontology quantifiers are explained and examples are included.

- **Pitfall:** Misuse of the existential and universal ontology quantifiers
  - **Mistake 1:** Incorrectly assuming that the universal restriction implies the existential restriction
    - \* **Cause of mistake:** Trivial satisfaction of the universal restriction
    - \* **Outcome (Defect 1):** Using only the universal quantifier rather than a combination of both restrictions when this is needed
    - \* **Outcome (Defect 2):** Using the universal quantifier rather than the existential restriction as the default

For instance, the axiom “*PetLoverTypeA has only Cat pets.*” includes the universal restriction and can be satisfied by the following cases:

1. Instances of *PetLoverTypeA* have one or more *Cat* pets and no other types of pets.
2. Instances of *PetLoverTypeA* have no pets at all.

Often the second option - the trivial satisfaction of the restriction, is forgotten and it is incorrectly assumed that the universal restriction implies the existential restriction. As an outcome, using the universal quantifier rather than the existential restriction becomes the default for many newcomers to ontology modeling. [Rector et al., 2004] explains the error is pernicious as it often appears as if the result is working, however, at later phases of the ontology development problems start occurring.



Let us consider the example: “*A ProteinLoversPizza is any Pizza that, amongst other things, has only Meat toppings.*” If the Protein Lovers Pizza is modeled using the universal quantifier as in the example, the restriction can be trivially satisfied, meaning that the Protein Lovers pizza could have no toppings, which would classify the pizza as vegetarian in a classification system.

There are two possibilities to correct the modeling of the axiom depending on how we wish to model the Protein Lovers Pizza.

- \* adding an existential restriction: “*A ProteinLoversPizza is any Pizza that, amongst other things, has some Meat toppings and also has only Meat toppings.*”

This modeling explicitly states that the pizza must have Meat toppings and no other toppings.

- \* replacing the universal quantifier for an existential one: “*A ProteinLoversPizza is any Pizza that, amongst other things, has some Meat toppings.*”

The new modeling states that the pizza must have Meat toppings, however, other toppings are also allowed for instance Tomato or Cheese toppings.

- **Mistake 2:** Incorrectly assuming missing information is incorrect (i.e., if a fact  $f$  does not exist, assuming  $\text{not}(f)$  is true)

- \* **Cause of mistake:** "some" does not imply "some not" & the Open World Assumption

- \* **Outcome (Defect 3):** Forgetting the "closure restriction" for axioms

For instance, the axiom “*PetLoverTypeB has some Cat pets.*” includes the existential restriction and can be satisfied by the following cases:

1. Instances of PetLoverTypeB have one or more Cat pets and no other pets.
2. Instances of PetLoverTypeB have one or more Cat pets and also one or more pets of a type other than Cat.

In ontology engineering, we assume an Open World (OWA) - meaning that information that is not modeled is not incorrect unless it contradicts the model. In other areas (constraint languages, databases, etc. ), a Closed World Assumption (CWA) is used which in contrast to OWA assumes a fact is incorrect if not explicitly modeled. Because many newcomers to ontology modeling have previously worked with CWA systems, the second option from above is often forgotten or incorrectly assumed to be false [Warren et al., 2019][Rector et al., 2004].

Let us look at the example model of a Margherita pizza: “*A Margherita pizza is any pizza which, amongst other things, has some tomato toppings and also some mozzarella toppings.*” If the Pizza is modeled like above using only the existential quantifier, it would be correct to classify a Pizza with Mozzarella, Tomato, and Bacon toppings as a Margherita since the existential restriction alone does not exclude the possibility of other value types for the property, such as Bacon in this case.

The Margherita definition would become more clear after adding the universal quantifier as a “closure restriction” to restrict that the *hasTopping* relation can take values only from the desired classes, leading to a correct model: “A *Margherita pizza* is any pizza which, amongst other things, has some tomato topping and also some mozzarella topping and also has only mozzarella and/or tomato toppings.” [Rector et al., 2004].

A traditional approach to finding such defects in an ontology would be to involve ontology engineering experts, however, Human Computation & Crowdsourcing (HC&C) is a promising approach to outsource specific verification tasks to human participants, and has already been applied successfully in other domains. In the next sections, a HC&C approach for identifying the incorrect usage of universal and existential restrictions is proposed.

### 4.2 Ontology Restrictions Verification - a Human Computation Approach

A Human Computation approach was designated for evaluating the correct usage of ontology restrictions. The following sections describe the implemented solution, as well as design decisions made during its development.

#### 4.2.1 Data Preparation: Ontology Restrictions Extraction

The first part of the approach is centered on the preparation of the ontology so that the restrictions can be verified by multiple evaluators. For this to be established, all restrictions from the ontology which is to be evaluated are automatically extracted. Furthermore, the process also groups quantifiers on the same relation together forming axioms. Each axiom represents a small ontology that fully describes a specific relation and can be evaluated independently from the rest of the axioms.

Once the ontology is separated into axioms, the format of the axioms can be modified if needed to allow for effortless translation to a representational formalism of choice.

The next section describes the Human Intelligence Tasks (HITs) that the human contributors need to complete in order to evaluate the created axioms. In order to better present the HC&C solution, the examples shown are based on the Pizza Ontology<sup>1</sup>. Nevertheless, the proposed solution is generic and any other ontology can be used with little changes needed as long as it is possible to show an instance of the domain (e.g as an image), as discussed below in detail.

---


<sup>1</sup><https://protege.stanford.edu/ontologies/pizza/pizza.owl>

amazon **mturk**

Group A Part 1 (HIT Details)  Auto-accept next HIT Requestor: *State's* HITs: 10 Reward: \$0.80 Time Elapsed

See Instructions

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

Pizza Menu	Model
 <p><b>CHEESY PIZZA</b> contains cheese</p>	<p>Cheesy Pizza is any pizza that, amongst other things, has some Cheese topping.</p>

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)  
Use your text area to provide additional information.

Submit

**Figure 4.1:** Example of a HIT for the verification of ontology restrictions in Amazon Mechanical Turk

### 4.2.2 Task Design

Given a real-life domain entity (e.g. in the form of an image or natural language description) and a modeling of the entity in a formalism of choice (e.g. OWL, Rector, etc.) the evaluators would need to identify across defects.

In Figure 4.1 we see an example of one HIT, in which the domain entity is shown on the left as an image of a pizza menu item. On the right side of the shown task we have the model, which is represented in the Rector formalism, discussed previously in section 2.6 *Ontology Axiom Representations*. Based on the information from the menu item the evaluator needs to decide whether the model correctly represents the Cheesy Pizza instance and if not select the defect that makes the model incorrect.

In the example, we see that this type of pizza must include cheese and there are no further requirements on what toppings the pizza might or must have. In the modeling, the existential quantifier is used, which makes it correct and the worker would need to choose the first answer option.

#### Verification Task

The verification task has two variables to be set:

- (a) the representation of the real-world instance

Since there are different needs and possibilities related to different application domains, the task design supports various formats in which the domain instance can be presented in. The examples shown in the thesis make use of images to represent the domain entity, however, a description in a natural language can also be used.

(b) the representational formalism of the ontology axiom to be verified

The task design allows for an ontology representational formalism of choice such as OWL and RDF(s), which as shown in section 3.3 *Evaluated Resource (SMS\_RQ1)* are widely used in the literature, or as discussed in section 2.6 *Ontology Axiom Representations* alternative representations such as Rector, Warren and VOWL can be selected.

The verification of the ontology quantifier axioms is created as a semi-closed task. For the verification decision, workers are provided with answer options each of which presents a possible scenario of model changes. Each answer corresponds to a defect from a defined taxonomy, which is discussed below, to allow easy aggregation and evaluation of the results.

### Defect Taxonomy

Papers that focus on ontology restriction teaching and common mistakes in the usage of ontology quantifiers were searched and read in order to create a taxonomy of defects, which can be verified by users. Three common issues were found to be reoccurring when working with ontology quantifier axioms, as discussed in section 4.1 *Misuse of the Universal and Existential Restrictions*, and are summarized below.

- Incompleteness
  - missing existential restriction (corresponding to Defect 1 above)
  - missing universal restriction (corresponding to Defect 3 above)
- Misuse
  - universal restriction used instead of an existential restriction (corresponding to Defect 2 above)

To make the answer options symmetric a non-included in the predefined taxonomy defect is added as a verification option (*Defect 4: existential restriction used instead of a universal restriction*). This makes sure that the workers consider all possible modeling options and in a way also acts as a spam filter.

A free-text answer option is also added in the HIT design to allow for the possibility that new defects are identified or ambiguities in the question design or model representation are established.

Figure 4.2 shows one more example of an ontology axiom to be verified and how it is presented to the workers for verification. The workers see as already discussed a context entity (1) on the left and a model (2) on the right-side of the HIT pane. At the bottom of the HIT, there are the different answer possibilities (3), which allow for defect classification. The HIT from Figure 4.2 shows incorrect modeling of the "Polo Ad Astra" Pizza. The pizza has cajun spice, red onion, chicken, mozzarella, sweet pepper, and tomato as toppings as we see in the menu item. In the model, a combination of

The screenshot shows an Amazon Mechanical Turk HIT interface. At the top, it says 'amazon MECHANICAL TURK' and 'Group A Part 1 (HIT Details)'. There are buttons for 'See Instructions' (5) and 'Auto-accept next HIT'. Below this, a note says 'Please make sure you are familiar with the rules and examples provided in the Instructions before answering the question.' The main content is divided into two panels: 'Pizza Menu' and 'Model'. The 'Pizza Menu' panel (1) contains an image of a pizza and the text 'POLLO AD ASTRA' followed by a list of toppings: 'Cajun Spice, Chicken, Garlic, Mozzarella, Red Onion, Sweet Pepper, Tomato'. The 'Model' panel (2) contains the text: 'Pollo Ad Astra pizzas have, amongst other things, some Garlic topping, and some Cajun Spice topping, and some Red Onion topping, and some Chicken topping, and some Mozzarella topping, and some Sweet Pepper topping, and some Tomato topping.' Below the model is a question: 'Does the model represent the pizza menu item correctly?'. There are four radio button options (3): '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.', and 'For the model to correctly represent the menu item, one or more existential (some) restrictions need to be replaced by universal restrictions (only)'. Below the options is a 'Comment (optional)' field (4) with a placeholder 'place for any remarks' and a 'Submit' button. A 'See Instructions' button (5) is at the top left. Red circles and lines highlight these elements with labels: 'real-life entity for context (as text or image)', 'instructions on ontology restrictions', 'axiom modeling (in a selected formalism)', 'answer options correspondig to the defect taxonomy', and 'place for any remarks'.

**Figure 4.2:** Example of a HIT for the verification of ontology restrictions in Amazon Mechanical Turk

existential quantifiers is used. With the given model it would be possible to classify a pizza with cajun spice, red onion, chicken, mozzarella, sweet pepper, tomato **and bacon** as a "Polo Ad Astra" pizza, which makes it incorrect. From the possible answers in the HIT, the worker would need to choose to add an additional universal restriction (3rd answer), which would act as a closure axiom and would make additional topping adding prohibited.

An optional comment (4) is present as well so that ambiguosnesses in the question design or model representation can be found, for instance, if the image was not loading and a decision was not possible. In the example, we also see that there is an Instructions-button (5), which opens a pane with all the needed theoretical background needed to answer the question, as described further below.

### Context Information

The modeling theory behind ontology quantifiers is provided in an instructions panel and is available throughout the full verification task. The instructions contain definitions and descriptions adopted for the selected model formalism and also offer examples of correct and incorrect modeling choices with justifications. Figure 4.3 and 4.4 present instructions available for tasks represented in a VOWL formalism. There are 3 panes - one with a short summary of the task shown in Figure 4.3a, one with explanations of the correct usage of ontology restrictions as seen in Figure 4.3b and lastly one pane with examples presented in Figure 4.4.

Summary Detailed Instructions Examples

### Overview

In this job, you need to verify if a pizza menu item is correctly represented by a graphical model represented in the VOWL formalism.

---

### Steps

1. Examine the pizza menu item
2. Examine the model
3. Decide whether the model correctly represents the menu item
4. If the model is invalid, select what defect it includes

(a)

Summary Detailed Instructions Examples

### Overview

In this job, you need to verify if a pizza menu item is correctly represented by a graphical model represented in the VOWL formalism.

---

### Steps


1. Examine the pizza menu item
2. Examine the model
3. Decide whether the model correctly represents the menu item
4. If the model is invalid, select what defect it includes

---

### Rules & Tips


**Rules :**

- Existential restrictions indicate that there must be a property of the specified type. Other types are not restricted.  
Representation in a graphical model:



- Every instance of PetLoverTypeB has a Cat pet and may also have other pets.

- Universal restrictions indicate that all property values for the specified property must be of a certain type. All values must be of the type but a property value does not have to exist.  
Representation in a graphical model:



- If instances of PetLoverTypeA have pets, the pets are always dogs. Instances of PetLoverTypeA, however, may not have any pets.

**Tips:**

- The universal restriction does not imply the existential restriction.
- If some information is not included in the model, the missing information is not false unless it contradicts the model.

(b)

Figure 4.3: Provided instructions for the verification of ontology restrictions in Amazon Mechanical Turk for the VOWL formalism



**Figure 4.4:** Provided examples for context for the verification of ontology restrictions in Amazon Mechanical Turk for the VOWL formalism

Furthermore, previous research [Mortensen et al., 2013] [Mortensen, 2013] has shown that providing the evaluators with some domain information has positive effects on the verification accuracy. For this reason, a real-world domain entity is provided, based on which workers are to decide whether the provided axiom model is valid. Depending on the evaluated domain and ontology different formats can be used to present the context entity, in the shown examples from the thesis an image is used.

For the described verification task it is important to outline that the role of a human evaluator is of importance for several reasons:

- (1) It is not (yet) possible to reliably extract the semantic meaning of real-world entities automatically, such as menus as in the above examples, since they are complex and diverse.
- (2) An automated method could be subjective since engineers can (un)knowingly implement biased algorithms. Crowdsourcing techniques, on the other hand, harness the wisdom of the crowds. Moreover, for most domains there is no gold standard based on which an automated solution can be developed, instead, the collective intelligence of multiple evaluators needs to be used.
- (3) Ontology axioms need to be interpreted by evaluators with some modeling knowledge since elements can be nested in one another in various ways, which can change the axiom meaning. Most systems support only the basic structures and limited nesting.

### 4.3 Human Computation Approach Summary

In this chapter, we focused on the misuse of the ontology restrictions and the need for their verification requiring human contributions. For this purpose, a Human Computation approach was proposed, with which ontology quantifiers can be evaluated. The task design allows for some customization depending on the ontology and the domain in which the evaluations are performed. The modeling of the ontology axiom to be verified can be shown in a representational formalism of choice and a context entity is included in the task, which can be visualised in different formats to help the contributors with their decision. Included are also instructions on the correct modeling techniques to make sure that the workers have all the needed modeling knowledge to perform the evaluation. The task shows the contributors possible answer options amongst which they can choose, and each option corresponds to a defect from a defined taxonomy. To allow for remarks from the contributors on any misunderstandings a free-text field is also included.

We looked into the process of the task design and decisions made during the implementation of the HC solution. In order to evaluate the proposed approach, an experiment was constructed, which is described in detail in the next chapter.



# Setup of Evaluation Experiment

A Human Computation experiment was designed in order to evaluate the proposed in chapter 4 *Ontology Restrictions Verification: a Human Computation Approach* approach and enable the investigation of RQ3. The following sections describe the goals and set up of the experiment.

**RQ3** How suitable are Human Computation techniques for the evaluation task of verifying ontology restrictions?

## 5.1 Experiment Aims

The experiment aims to investigate how HC&C techniques can be used to solve the ontology verification task of identifying mistakes in the use of universal and existential quantifiers. The main goals are:

- Understanding the influence of different representational formalisms of universal and existential restrictions on the performance of the participants.
  - Human Computation task design aspect: representation of data
  - Hypothesis H1: The formalism in which axioms are represented has an influence on the performance/speed of the contributors.

It is important to investigate this design aspect so that in future it is clear how to best present the modeling of restrictions that needs to be checked to get the best possible results.

- Understanding the influence of prior modeling knowledge on the performance/speed of the contributors.
  - Human Computation task design aspect: choice of workers based on skills

- Hypothesis H2: Prior modeling knowledge has a positive influence on performance and time.

This effect has already been shown in [Warren et al., 2019] and the thesis aims to gather additional experimental data here. It is essential to investigate the design aspect so that in the future it is clear which skill the evaluating population should have. The results will be compared with the results from [Warren et al., 2019], where the hypothesis holds. The possibility exists that (only) some defects are better identified by more experienced participants. This would mean that some defects can be checked by juniors and others need the attention of a senior modeler. The results from the experiment will provide insights into the topic.

## 5.2 Experimental Setup

### Participants

The conducted human computation experiment relied on an internal student crowd rather than a layman crowd for several reasons. Firstly, the evaluation of the correct usage of ontology restrictions requires modeling knowledge of the differences between the quantifiers, but also general modeling knowledge for the understanding of graphical representations which can be difficult to collect in a crowdsourcing platform with layman contributors without ontology modeling knowledge. Secondly, working with any crowdsourcing platform means dealing with spammers. Since the students make the verifications as part of their studies and receive points for the quality of their results, they are motivated to work on the tasks and think about the correct answers. This is also a chance for them to practice and deepen their knowledge. Lastly, working with a student crowd allows a more controlled environment. Each student performed a self-assessment and took a qualification test which allows for a better understanding of the participants' prior knowledge.

The students who participated in the experiment took the course *Introduction to Semantic Systems* held in the 2020 winter term at Vienna University of Technology and are masters students of the following study programs: *Data Science*, *Business Informatics*, *Information & Knowledge Management*, *Medical Informatics*, *Software Engineering & Internet Computing*. In total 88 students registered and were awarded 10 bonus points for the course - 5 for participation and 5 for the quality of their results (score over 50%).

### Platform

Amazon Mechanical Turk<sup>1</sup> (mTurk) is a crowdsourcing platform that offers the possibility to harness the wisdom of the crowd. Requesters can implement their outsourced work assignments as Jobs. Each Job contains several HITs (Human Intelligence Tasks) which are simple and independent pieces of work that can be solved by a global workforce - the crowd workers. Usually, each HIT offers a monetary reward for the workers that complete it, which acts as motivation. Requesters have the ability to restrict to whom the HITs

---

<sup>1</sup><https://www.mturk.com>

will be available - so it is possible to require a specific qualification that the workers need to have in order to start working on the jobs. The platform offers an international and multicultural workforce and is, as we saw in chapter 3 section 3.7.1 *Evaluation methods and modalities*, one of the most widely used platforms for Crowdsourcing in the Ontology Evaluation domain. Amazon mTurk also offers a Sandbox<sup>2</sup>, where Jobs and HITs can be tested without additional fees before publishing them to the real platform. This is where the full Human Computation solution described in the previous chapter has been developed, tested, and evaluated.

### Selected Ontology

The ontology used for the experiment is the Pizza Ontology<sup>3</sup>. The ontology is a known good ontology and contains many structures with the universal and existential quantifiers. In [Rector et al., 2004] the authors argue that this is also the most successful ontology in teaching Western audiences about ontology restrictions and common good practices. They also describe the ontology as easy to work with and fun for beginners as pizzas are familiar and concrete, yet they are still diverse and one can illustrate important aspects with them.

### Experimental Data

In order to build the experimental data, universal and existential restrictions were extracted from the Pizza Ontology so that each pizza is described with a meaningful axiom. Several types of restriction structures were excluded:

- Restrictions on the spiciness value of pizzas or toppings of the type *"Sliced Tomato toppings have, amongst other things, some spiciness value Mild."* are excluded since they cannot be easily verified and do not make clear sense when taken out of context.
- Pizza axioms containing only one universal quantifier of the type *"Vegetarian Pizza is any pizza that, amongst other things, only has Vegetarian toppings."* are removed since there is no evidence of this structure being problematic or defect-prone.
- Pizza axioms containing a combination of existential and universal quantifiers, with the universal quantifier not acting as a "closure restriction" of the type *"Thin And Crispy Pizza is any pizza that, amongst other things, has some pizza base, and also only has a thin and crispy base."* are excluded since this structure is not proven to be defect-prone.

After extracting the axioms from the pizza ontology and removing the unproblematic structures, 30 meaningful pizza axioms are left, which are of the types:

- Pizza axioms containing only existential quantifiers  
Example: *"Meaty Pizza is any pizza that, amongst other things, has some Meat*

<sup>2</sup><https://requestersandbox.mturk.com>

<sup>3</sup><https://protege.stanford.edu/ontologies/pizza/pizza.owl>

*topping.*"

This structure is prone to the following mistakes:

- Misuse: a universal restriction instead of an existential restriction (Defect 2)
- Pizza axioms containing a combination of existential quantifiers and one universal (closure) restriction

Example: *"Margherita pizzas have, amongst other things, at least one Mozzarella topping, and at least one Tomato topping, and also no other than Mozzarella, and/or Tomato toppings."*

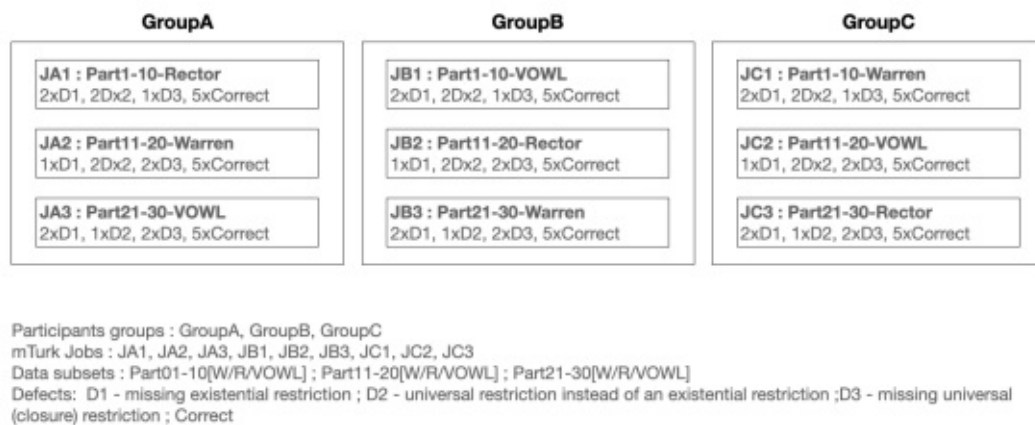
This structure is prone to the following mistakes:

- Incompleteness: missing existential restriction (Defect 1)
- Misuse: universal restriction instead of an existential restriction (Defect 2)
- Incompleteness: missing universal (closure) restriction (Defect 3)

For the experiment, half of the axioms were modified so that they include one of the defects mentioned above. The final experimental data contains 15 correct axioms as well as 15 incorrect ones, in which the three defect types are equally distributed.

### Data Split & Task Assignments

The 30 axioms of the experimental data were split into 3 sections: Part1-10, Part11-20, Part21-30. Each Part contains 5 defects, at least one from each type, and 5 correct statements. The exact distribution of the defects is shown in Figure 5.1. Each Part is also translated to the 3 representational formalisms explained in chapter 2 section 2.6 *Ontology Axiom Representations* - Rector, Warren and VOWL. Each data section presented in one of the formalisms was implemented as a separate Job on the Amazon mTurk Sandbox platform.



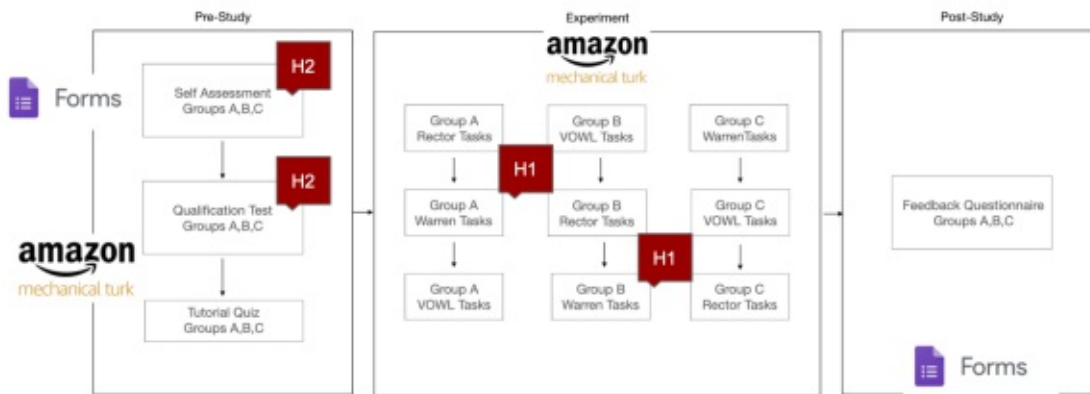
**Figure 5.1:** Overview of the group assignments and data split

The students were equally distributed into 3 groups: A, B, and C. To be fair and unbiased each group sees the same task sections in the same order, however, in a different formalism.

Tasks from the same modeling formalism are grouped together to lower the cognitive overhead of switching between different representations for the workers. As shown in Figure 5.1 Group A started working on Part1-10 in the Rector formalism, continued working on Part11-20 in the Warren formalism, and finished off with the VOWL formalism of Part21-30. In the meantime, Group B started with Part1-10, but sees those in the VOWL formalism and so on. The questions inside of each section are also automatically randomized by Amazon mTurk for each participant to make sure some questions are not overlooked and the sequence bias is avoided.

### 5.3 Experiment Overview

In this section, we look into the flow of the experiment and all included stages as shown in Figure 5.2.



**Figure 5.2:** Overview of the experiment workflow. H1 = Hypothesis H1, H2 = Hypothesis H2

The first phase of the experiment is the pre-study. Here the participants' knowledge is evaluated both from an objective and a subjective perspective in order to allow for the investigation of Hypothesis H2 (*Hypothesis H2: Prior modeling knowledge has a positive influence on performance and time.*).

The study participants first needed to complete a **Self-Assessment** Test, in which the participants rate themselves in regards to their modeling knowledge and English language skills. The participants are asked to categorize their knowledge in different areas (English skills, descriptive logic, modeling, OWL) into the categories: no/little/some/expert knowledge. The same knowledge categorization is used in [Warren et al., 2019]. This is an important part of the experiment as the authors in [Warren et al., 2019] have shown that previous knowledge of knowledge representation languages can have an effect on the accuracy/speed of interpreting logical propositions. To make the assessment of the participants' knowledge more objective, we provide a knowledge level scale to the participants based on which they can identify which category they belong to. Figure 5.3

shows the provided knowledge levels for ontology modeling skills, based on which the participants had to rate their knowledge. As a platform for the self-assessment Google Forms<sup>4</sup> was used. The self-assessment is included as *Appendix C: Self Assessment Test*.

The screenshot shows a Google Form titled "Ontology Modeling Skills". Below the title, there are four numbered levels of knowledge:
 

- 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.

 Below this, a question is asked: "Q5: How would you rate your knowledge in ontology modeling? \*". Underneath the question is a horizontal scale with four radio buttons labeled 1, 2, 3, and 4. The text "no knowledge" is positioned to the left of the scale, and "expert knowledge" is positioned to the right. The radio buttons are currently unselected.

**Figure 5.3:** Example of a knowledge level scale from the Self-Assessment Test on Google Forms

With the purpose of evaluating the knowledge of the participants objectively a **Qualification Test** was designed in Amazon Mechanical Turk. The test only targets the ontology modeling knowledge of 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 might have a subjective view of their competencies. The test includes questions from different difficulty levels, based on which participants can be sorted into the same categories - no/little/some/expert knowledge. Figure 5.4 shows one of the examples in the Qualification Test. To exclude any bias in regards to the representation of the restrictions, each question includes all 3 representations (Rector, Warren & VOWL) of the axioms. The qualification test is added as *Appendix D: Qualification Test*.

A **Tutorial** job was created in order to allow the students to get to know the question format as well as the Amazon Mechanical Turk platform better before working on the verification jobs. The job had the same structure as the original job, however, the data was from a different domain - the Wine domain, rather than the Pizza domain used for the main tasks. This makes sure that all participants acquire the basic knowledge of using the crowdsourcing system prior to the actual experiment. Figure 5.5 shows one HIT from the tutorial Job. The instructions pane is as described in the last chapter, however, it has been adapted to the Wine domain.

<sup>4</sup><https://workspace.google.com/products/forms/>

**Qualification Test**

Make sure to accept the HIT before you start answering the questions.

Please enter your Student ID:

**Section 2: Some Knowledge**

This section tests your understanding of the implications of ontology axioms and restrictions.

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

**VOWL**

**Rector**

PetLoverTypeA has, amongst other things, only Dog pets.

---

**Warren**

PetLoverTypeA has, amongst other things, no other than Dog pets.

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.

**Figure 5.4:** Example question from the Qualification Test on Amazon Mechanical Turk

**Wine sort**

**WHITE BORDEAUX**  
made from Semillon  
Grape, Sauvignon Blanc  
Grape

**Model**

Does the model represent the wine sort correctly?

- The model correctly represents the wine sort.
- For the model to correctly represent the wine sort, one or more existential (∃) restrictions need to be added.
- For the model to correctly represent the wine sort, one or more universal (∀) restrictions need to be added.
- For the model to correctly represent the wine sort, one or more universal (∀) restrictions need to be replaced by existential (∃) restrictions.
- For the model to correctly represent the wine sort, one or more existential (∃) restrictions need to be replaced by universal (∀) restrictions.

Comment (optional):  
If case you have any remarks please add them here.

**Figure 5.5:** Example of a HIT from the Tutorial Job from Amazon Mechanical Turk

The next phase is the actual experiment, where the students perform the verification of the ontology restrictions axioms. Each group of students performs **3 Verification Jobs**, containing 10 HITS each and using a different formalism to represent the axioms. This setup is required so that we can get insights into Hypothesis H1 (*H1: The formalism in which axioms are represented has an influence on the performance/speed of the contributors.*)

Lastly, as a post-study, a **Feedback Questionnaire**, implemented in Google Forms, had to be filled in by each participant. The post-study questionnaire aims to determine whether the experiment was designed well and how it was perceived by the participants. Based on the results from the questionnaire possible improvements can be outlined. In the feedback form, students also gave their opinion, which formalism (Rector|Warren|VOWL)

they understood best, which is later on also compared with their verification results. The feedback form is included as *Appendix E: Feedback Questionnaire*.

To complete all parts of the experiment (pre-study, experiment, and post-study) the participants were given 2 hours. They were asked to create all needed accounts (a Google Account, an Amazon Account, and an Amazon Worker Sandbox Account) beforehand.

The week before the experiment, the participants received an email with the group assignment and a tutorial was presented which explained in detail what parts the experiment will consist of as well as an intro and tips about the used platform. On the day of the experiment, the students received a link to an overview page, which included links to all parts of the experiment in the order in which they should be completed, as well as again the tutorial slides and useful tips. The overview page for Group A is included as *Appendix B: Experiment Starting Point for the Participants from Group A*. Each student performed the tasks at home at the given time. During the experiment, there was an active Zoom<sup>5</sup> meeting where organizational questions were answered and technical issues with the platform were solved.

### 5.4 Evaluation Setup Summary

The presented experiment aims at investigating the results which can be achieved with the proposed solution as well as several aspects of the task design, which are needed for future experiment developments. The two hypothesis that the experiment aims to investigate are:

- Hypothesis H1: The formalism in which axioms are represented has an influence on the performance/speed of the contributors
- Hypothesis H2: Prior modeling knowledge has a positive influence on performance and time.

To allow for the analysis of the first hypothesis H1, the experimental data includes axioms in each of the representational formalism, discussed above. In order to allow for the investigation of H2, a pre-study is conducted in which the background modeling knowledge of the participants is rated both subjectively and objectively.

The results of the conducted experiment and the evaluation of the HC solution are analysed in the next chapter.

---

<sup>5</sup><http://zoom.us>



# Evaluation

This chapter outlines the results of the conducted Human Computation experiment and provides insights into the usefulness of the proposed verification approach in order to answer RQ3.

**RQ3** How suitable are Human Computation techniques for the evaluation task of verifying ontology restrictions?

In section *6.1 Evaluation Methods* the methodological process followed to evaluate the proposed Human Computation solution is outlined. Section *6.2 Participants Background Knowledge* discusses the previous knowledge that participants in the experiment had evaluated in the pre-study stage. The results of the verifications and the post-study are analysed in section *6.3 Experiment Results* and lastly a summary of the evaluation is provided in section *6.4 Evaluation Summary*.

## 6.1 Evaluation Methods

In order to evaluate the proposed HC&C verification task design, a mixture of qualitative and quantitative evaluation was applied, as the focus lays not only on the percentage of the tasks the workers managed to complete successfully but also on analyzing the verification process.

As explained in the previous chapter in detail an experiment was designed with which the approach can be evaluated. At a first stage of the evaluation, a focus group was gathered with which a pilot of the experiment was conducted. Participants were 5 experts of the Semantic Systems Research Group at the Vienna University of Technology as well as 3 Masters students writing their thesis with the group. The goal of the pilot was to evaluate whether the task design was clear and easily understandable. Further goals were outlining any technical issues or difficulties with the used platforms, gathering

improvement suggestions for the task formulations and assessing the quality of the task design.

The pilot consisted of three parts: (1) an initial stage in which a Zoom meeting was held with the focus group and the tasks they had to complete were explained, (2) a verification stage, in which participants did the evaluation tasks on their own in a 2-hour-time window, and (3) a feedback stage, in which participants shared their experience via email and were further discussed during a final Zoom call. Based on the information gathered during the pilot some improvements in regards to the task formulation, order of tasks and technical guides were made before the actual experiment started.

While the goal of the pilot was to gather some qualitative insights to the appropriateness of the implemented approach, the experiment itself aimed to gather some quantitative data in order to investigate the two hypotheses described in the last chapter. Main metrics used are:

- the accuracy of the results (percentage of correct responses) and the speed of the verification (the evaluation response time)

As we saw in section 3.7.2 *Evaluation metrics and Inter-evaluator agreement* those are commonly used metrics to evaluate HC evaluation approaches. The metrics are calculated for the tasks is each representational formalism, which allows us to gain insights into which formalism resulted in the best verification quality.

- point-based correlation between the results and background knowledge of the participants.

These metrics are needed in order to understand how prior background knowledge of the participants influences the quality of the contributors' evaluations.

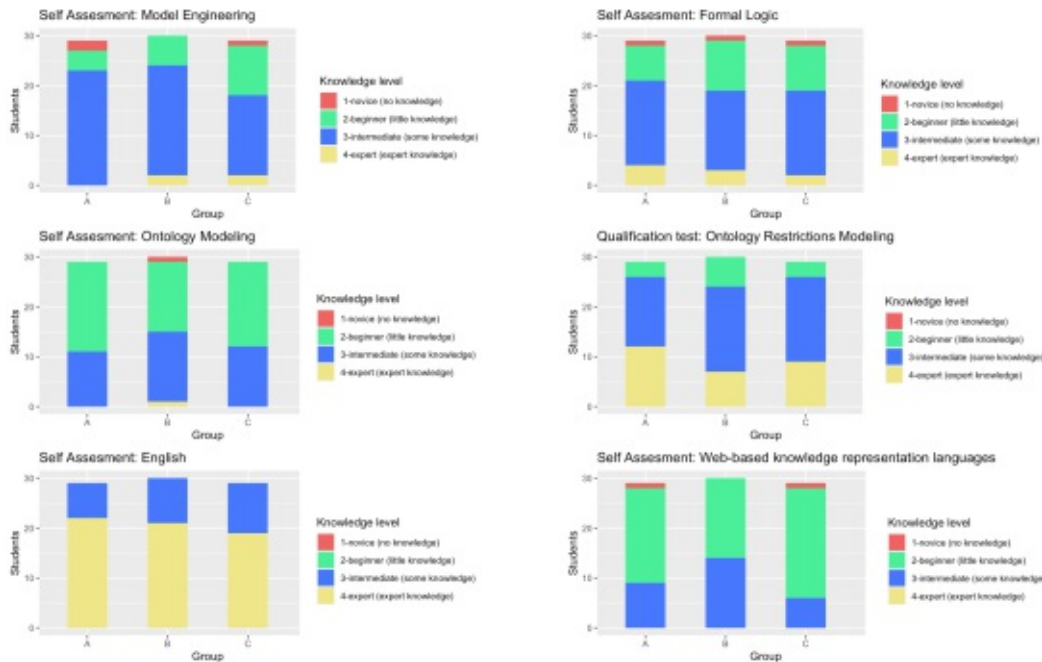
In the next sections, the results from the conducted experiment are analysed. Participants' feedback provided in the form of comments in the verification tasks or in the feedback questionnaire are included at the end.

## 6.2 Participants Background Knowledge

As discussed in the previous chapter 5 *Setup of Evaluation Experiment*, the experiments was performed with an internal student crowd, consisting of 88 masters students who took the course *Introduction to Semantic Systems* held in the 2020 winter term at Vienna University of Technology.

In order to understand the results of the experiment and how they were achieved it is important to first look into the qualification of the evaluator population prior to the study. For this reason, the results of the pre-study (Self-Assessment & Qualification test), as shown above in Figure 5.2, are analysed.

Figure 6.1 shows the participants' knowledge in each of the groups that the students were separated in, 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 favorable for the results interpretations.



**Figure 6.1:** Background knowledge of the participants in each study group

Since the study was conducted in English, it was essential to include a question on English Level knowledge skills in the Self-Assessment. All students who participated in the experiment rated their English knowledge in the two highest categories provided, meaning that all of participants were proficient in the language and were able to understand and work with English texts and the ideas conveyed in them.

In a previous study [Warren et al., 2019] it was shown that background knowledge in model engineering or formal logics impacts the understanding of ontology restrictions, which is why those areas were considered in the Self-Assessment. As we see in Figure 6.1 most students already had some (intermediate) knowledge in this area, which was expected since the participants are master students in the field of informatics.

Because the experiment's focus is on ontology evaluation, the qualification of the participants in ontology engineering is considered both in the Self-Assessment as well as in a separate qualification test on ontology restrictions. The majority of the students categorized themselves as beginners or intermediates in ontology modeling and from the results of the Qualification Test it can be seen that they already had some knowledge on ontology

quantifiers as well. Lastly, knowledge on Web-based representation languages such as OWL or RDF(s) was also assessed and most participants fall into the beginner-category.

The contributors' characteristics are further considered when looking into whether background knowledge had impact on the accuracy of the results in section 6.3.2 *H2 Results: Performance Based on Evaluators' Prior Knowledge*.

### 6.3 Experiment Results

In total 2629 student responses were gathered and each verification task had approximately 28-30 responses. Overall 92.58% of those responses were correct and one single judgement took on average about a minute (56.79 seconds). When aggregating the results with the majority voting strategy a 100% accuracy is reached, meaning that each axiom from the experimental data was evaluated correctly. These results already show that the proposed solution can be used to gather high performance results on the task of verifying ontology restrictions using human computation.

Table 6.1 shows the mentioned measurements for each verified axiom in each of the formalisms it was represented in. As it can be seen that some axioms (e.g axioms 8,15,24) show a high correctness of verification in every formalism. However, when looking at, for instance, axiom 2 we see the evaluation task in the Rector formalism only achieved 60% of correctness while the same axiom in the Warren formalism had more than 90% correct responses. How the representation of the axiom influences the performance of the evaluators in terms of speed and accuracy of their evaluations is discussed further in section 6.3.1 *H1 Results: Influence of Representational Formalism*.

**Table 6.1:** Performance of the evaluators in verifying ontology quantifiers for each axiom for the Rector/VOWL/Warren formalism

axiom ID	number of responses	correctness of responses	response time (s)	accuracy (majority vote)
	R/VOWL/W	R/VOWL/W	R/VOWL/W	R/VOWL/W
1	29/30/28	0.72/0.8/0.79	65.66/84.6/56.64	1/1/1
2	30/29/29	0.6/0.83/0.93	53.2/70/38.31	1/1/1
3	29/29/29	0.76/0.76/0.72	57.97/84.59/59.76	1/1/1
4	29/30/29	0.86/0.83/0.9	89.48/69.37/51.24	1/1/1
5	29/29/29	0.79/0.93/0.76	47.41/35.34/47.07	1/1/1
6	29/29/30	0.97/0.86/0.83	54.48/37.97/36.9	1/1/1
7	29/29/30	0.9/0.66/0.87	73.79/145.48/102.33	1/1/1
8	29/30/29	0.97/0.93/0.9	72.66/45.4/103.07	1/1/1
9	30/29/29	0.83/0.93/0.97	47.7/33.9/61.21	1/1/1
10	28/30/29	0.93/0.93/0.9	59.71/61.57/68.14	1/1/1
11	30/29/29	0.8/0.97/0.9	41.1/35.86/51.45	1/1/1

*Continued on next page*

Table 6.1 – Continued from previous page

axiom ID	number of responses	correctness of responses	response time (s)	accuracy (majority vote)
	R/VOWL/W	R/VOWL/W	R/VOWL/W	R/VOWL/W
12	29/29/29	1/1/0.9	40.52/56.76/37.76	1/1/1
13	29/30/29	0.97/0.97/0.86	55.34/61.43/76.62	1/1/1
14	29/29/29	0.97/1/0.9	38/56.48/34.59	1/1/1
15	29/30/29	0.93/0.93/0.97	86.45/64.33/86.55	1/1/1
16	30/29/29	0.97/1/0.83	55.27/34.76/45.86	1/1/1
17	29/29/29	1/1/0.83	34.55/34.31/58.24	1/1/1
18	29/30/29	0.93/0.87/0.97	90.1/88.07/90.55	1/1/1
19	29/29/29	1/0.97/0.9	43.62/47.24/42.24	1/1/1
20	29/29/30	1/0.97/0.97	39.9/23.21/43.03	1/1/1
21	30/29/29	0.9/1/1	36.9/29.07/35.86	1/1/1
22	29/30/29	1/1/0.97	55.97/43.33/79.31	1/1/1
23	29/30/29	0.97/1/1	65.66/87.1/65.93	1/1/1
24	29/30/29	1/1/1	69.79/51.8/88.24	1/1/1
25	30/29/29	1/1/1	67.33/53.83/55.03	1/1/1
26	30/29/29	1/1/1	36.2/39.59/50.52	1/1/1
27	29/29/28	0.97/1/1	58.14/46.17/54.43	1/1/1
28	29/29/29	1/1/1	55.1/34/35.69	1/1/1
29	29/29/30	1/1/1	39.45/25.93/31.93	1/1/1
30	29/29/30	0.97/1/1	39.14/34.9/45.27	1/1/1

Another perspective that was analysed is the difference in the performance for the different axiom structures as well as different defect types. Ontology axioms build of a combination of existential and universal quantifiers (Structure 1) make up around 80% of the experimental data and are the base of the examples provided in the instructions during the verification tasks. Such axioms were verified with correctness of 94%. On the other hand, axioms including a single restriction type (Structure 2) build the rest of the experimental data and were evaluated with slightly worse correctness of 83%. However, even though there were no examples given for the Structure-2-axioms, the HC task design still reaches 100% verification accuracy with the majority voting system and thus we can argue that the proposed HC solution does perform very well for the verification task of evaluating ontology restrictions.

87% of the verified axioms which contained Defect 1 (missing existential restriction) were identified correctly. Similarly, Defect 2 (universal quantifier used instead of an existential one) was detected correctly in 89% of the responses. The highest correctness of 95% was reached in verifications of axioms including Defect 3 (missing universal restriction). The same percentage of correctness is also achieved for axioms which were correctly identified as correct.

In order to provide further insights into how suitable the proposed HC approach was, the influence of different HC task design aspects is investigated. The next sections provide an analysis of the dependencies between task representation or background knowledge of the contributors and the quality of the achieved results.

### 6.3.1 H1 Results: Influence of Representational Formalism

**Hypothesis H1:** The formalism in which axioms are represented has an influence on the performance/speed of the contributors.

From the gathered results it can be deduced that the representational formalism, in which the axiom was presented to the contributors, did in fact have influence on how accurately and how fast the axioms were evaluated.

Table 6.2 shows the results from the evaluation tasks based on the representation of the axioms. For each formalism, the percentage of correct responses overall, the average percentage of correct responses per HIT, and the average verification time per HIT is provided. It can be seen that while the results are very similar, the verifications performed in the VOWL formalism have a slightly higher accuracy and the average time needed for evaluating an axiom is lower than in the textual representations. Figure 6.2a shows a graphical representation of the results.

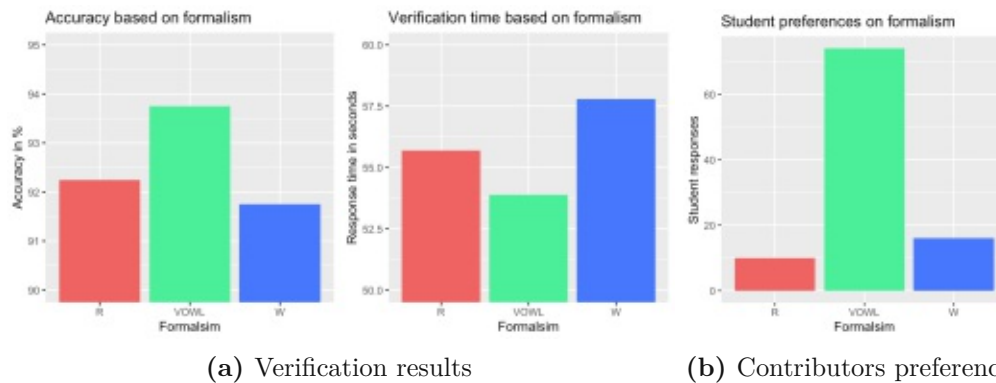
**Table 6.2:** Performance of the evaluators in verifying ontology quantifiers based on representational formalism

formalism	correct overall	avg correctness per HIT	avg response time per HIT
Rector	92.24%	92.28%	55.69s
VOWL	93.75%	93.76%	53.88s
Warren	91.75%	91.74%	57.79s

Another important factor to consider is what formalism was easiest to understand from the perspective of the evaluators. To gather insights into the aspect, the results from the post-study survey, where workers were able to provide their feedback, are analysed. Figure 6.2b shows that the majority (74%) of the participants preferred the VOWL formalism against the textual paraphrasing of the axioms.

From the presented results, it can be concluded that VOWL representation not only resulted in the fastest and best quality results overall but it was also preferred by the evaluators.

To achieve a better overview of the impact of the representation of the axiom on the results of the evaluations, an analysis of its influence on different axiom structures and defect types follows.



**Figure 6.2:** Influence of different representations of universal and existential restrictions on the performance of the contributors.

#### Performance based on representational formalism for each defect type

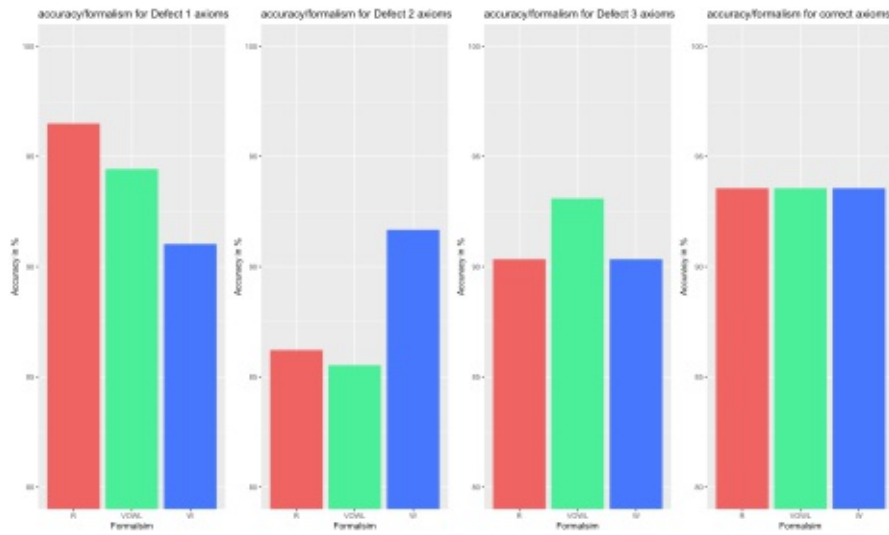
In order to understand the influence of the axiom representation on the speed and accuracy of its verification in terms of the correct usage of ontology quantifiers, the performance for each defect type is analysed. Figure 6.3 shows the average accuracy and speed of responses in the different formalisms for each defect from the defect taxonomy introduced in the previous chapter. While the verification of known correct axioms does not show any differences for the different representations in regards to performance, some defects were better identified in a specific formalism. For axioms including Defect 1 (missing existential restriction), the representation in the Rector formalism achieves best results in term of accuracy, and the fastest verification in VOWL. For Defect 2 axioms (universal quantifier used instead of an existential one), the best performance was attained with the Warren phrasing both in terms of accuracy and speed of verifications. And lastly, for Defect 3 (missing closure (universal) restriction for an axiom), evaluations for VOWL tasks had the highest accuracy while the speed was best in the Rector formalism.

These results reveal that while overall the graphical representation appeared to reach the highest performance of verifications, some defect types were actually easier to detect in a textual representation of the axioms.

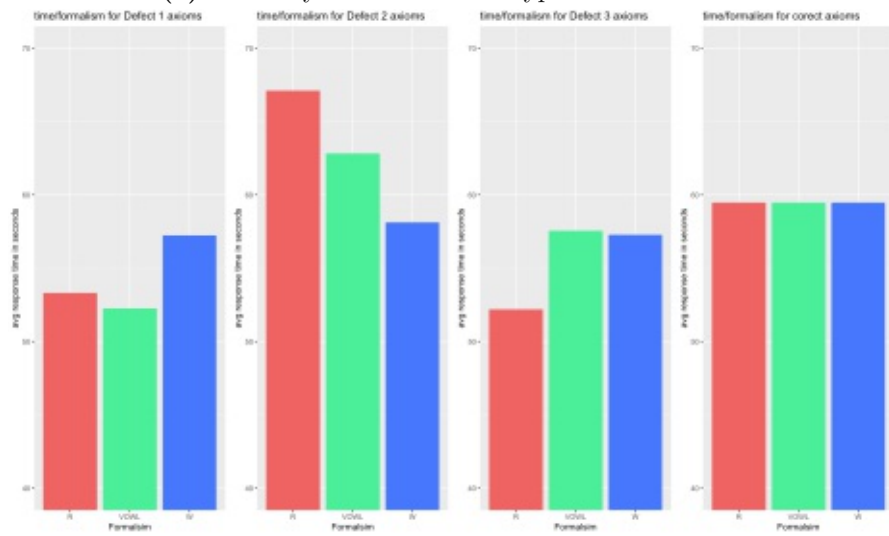
#### Performance based on representational formalism for each axiom structure

The influence of the axiom representation on the verification results for each axiom structure is shown in Figure 6.4. It can be seen that for evaluating axioms of type Structure 1 (combination of universal and existential restrictions) the Rector formalism is the best both in terms of accuracy of the results and speed of the verifications. On the other hand, Structure-2-axioms (including only one quantifier type) represented with the Rector paraphrasing were verified correctly with the lowest performance amongst the formalisms. For this structure, the Warren representation shows slightly better results than VOWL.

Like in the results for the performance based on defect type, here the representation of



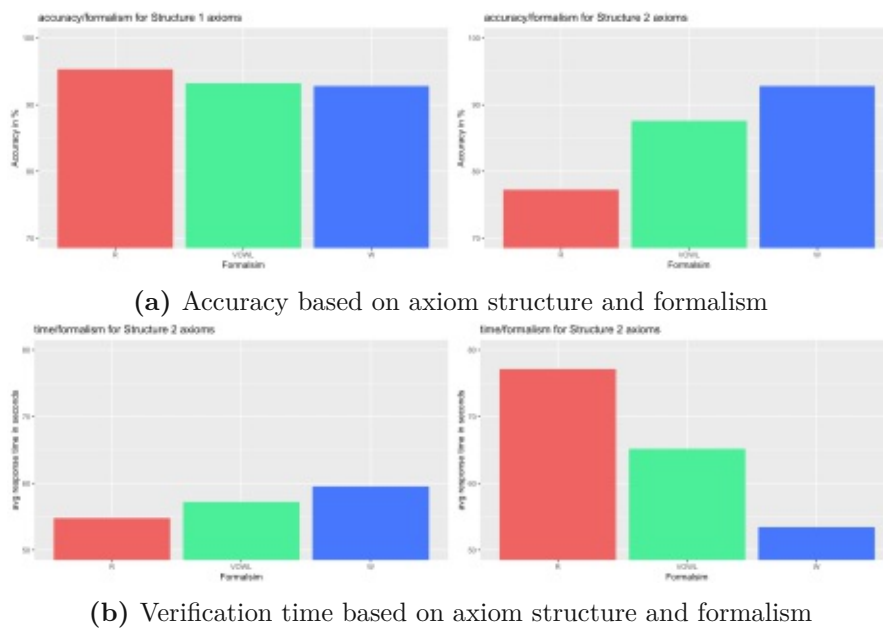
(a) Accuracy based on defect type and formalism



(b) Verification time based on defect type and formalism

**Figure 6.3:** Influence of axiom representation on the performance of the contributors in identifying different defects in ontology restrictions.





**Figure 6.4:** Influence of axiom representation on the performance of the contributors in identifying defects in different axiom structures.

formalism also has a different impact on the different structures. In case the structure types are known before the evaluation, this analysis can be used to choose the best representational formalism depending on the axioms at hand. However, it is more likely that information on the defect types or correct structure types would be unknown beforehand. For such cases, OWL seems to offer high and stable results throughout different setups and would thus be the best choice.

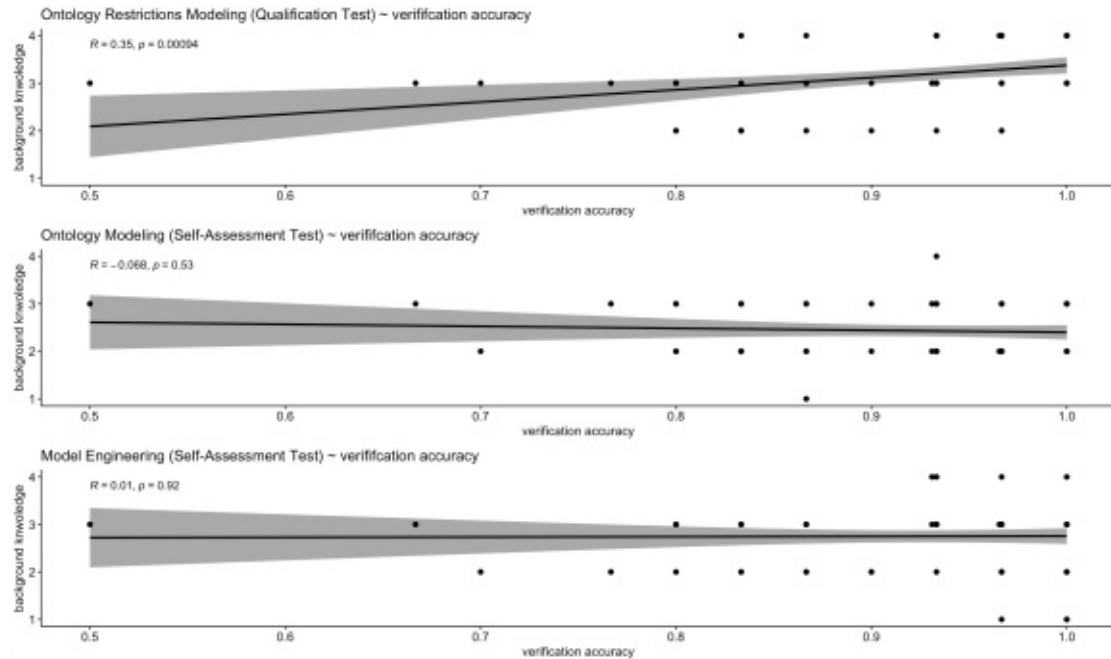
### 6.3.2 H2 Results: Performance Based on Evaluators' Prior Knowledge

**Hypothesis H2:** Prior modeling knowledge has a positive influence on performance and time.

The experiment results, as we will see further in this section, show that prior modeling knowledge can have a positive impact on the results of the contributions. However, the knowledge measured in the Self-Assessment Test and the knowledge categorisation via the Qualification Test show different significance of the influence on the verifications.

Figure 6.5 shows the point-biserial correlation between the verification accuracy and the background knowledge in each of the following areas: ontology restriction modeling, ontology modeling, and model engineering. Looking at the graphic, it is clear that the prior knowledge obtained during the Qualification Test has the highest positive effect on the evaluations of the contributors. Judging by the p-value for this correlation, we can say

that the effects are also statistically significant. On the other hand, the Self-Assessment results do not seem to have any significant impact on the correctness of the evaluations and therefore we can conclude that an objective qualification test proves to be a more effective way to predict the performance of the contributors, rather than a subjective Self-Assessment.

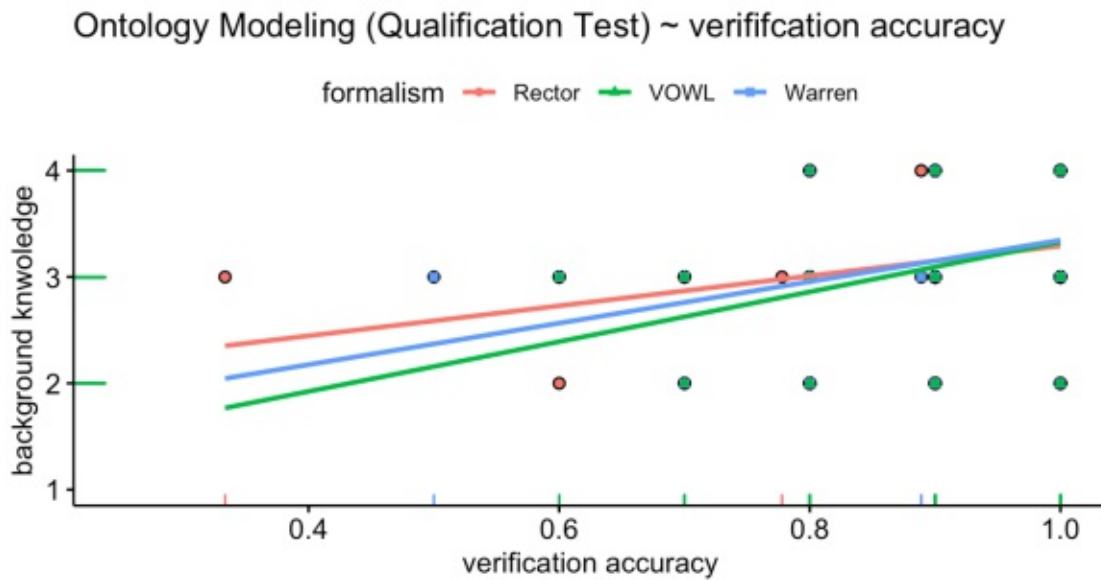


**Figure 6.5:** Point-biserial correlation between prior modeling knowledge and the accuracy of verifications

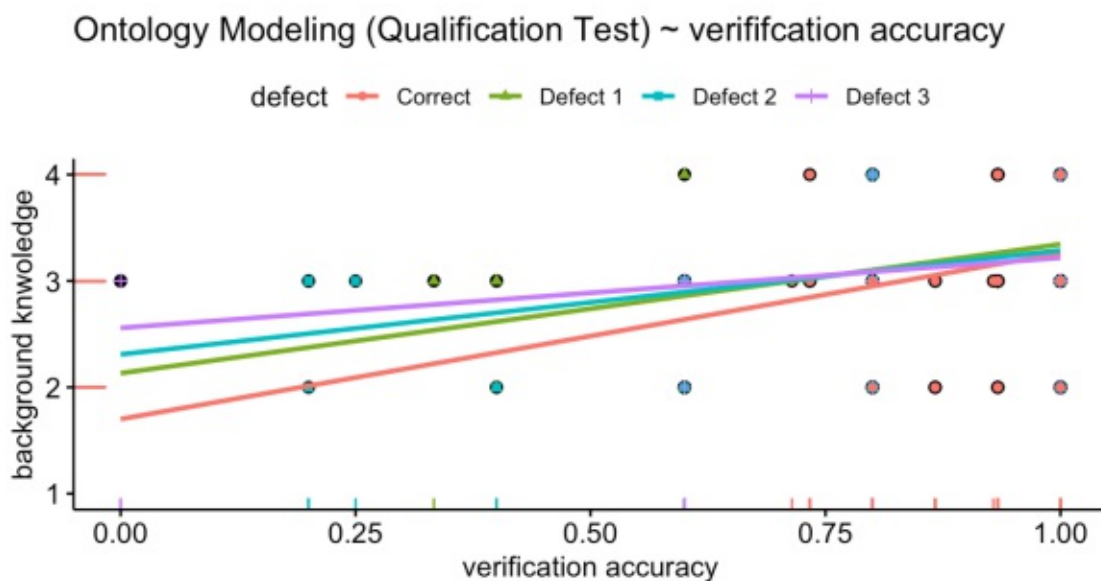
In order to look deeper into the correlation, Figure 6.6 shows the influence of the Qualification Test results on the verification accuracy in each representational formalism separately. As observed, the positive effects of the background knowledge are recognisable in each representation, however the correlation is the strongest for the VOWL formalism.

To further understand whether only specific defects were better identified from contributors with prior background knowledge, Figure 6.7 shows the correlation between the Qualification Test score and the verification accuracy for each defect type. It can be seen that while the positive influence of prior knowledge on the results is easily recognisable for Defects 1 and 2 as well as known correct axioms, the correlation seen for Defect 3 is not very strong. Based on the results it can be concluded that in future the evaluations can be distributed to contributors with different modeling knowledge levels depending on the defect types that might be included.

Figure 6.8 visualises the influence of previous knowledge on the time needed to complete the tasks. We see that the results of the Qualification test that had a positive impact on the verification accuracy do not have any significant effect on the time needed to



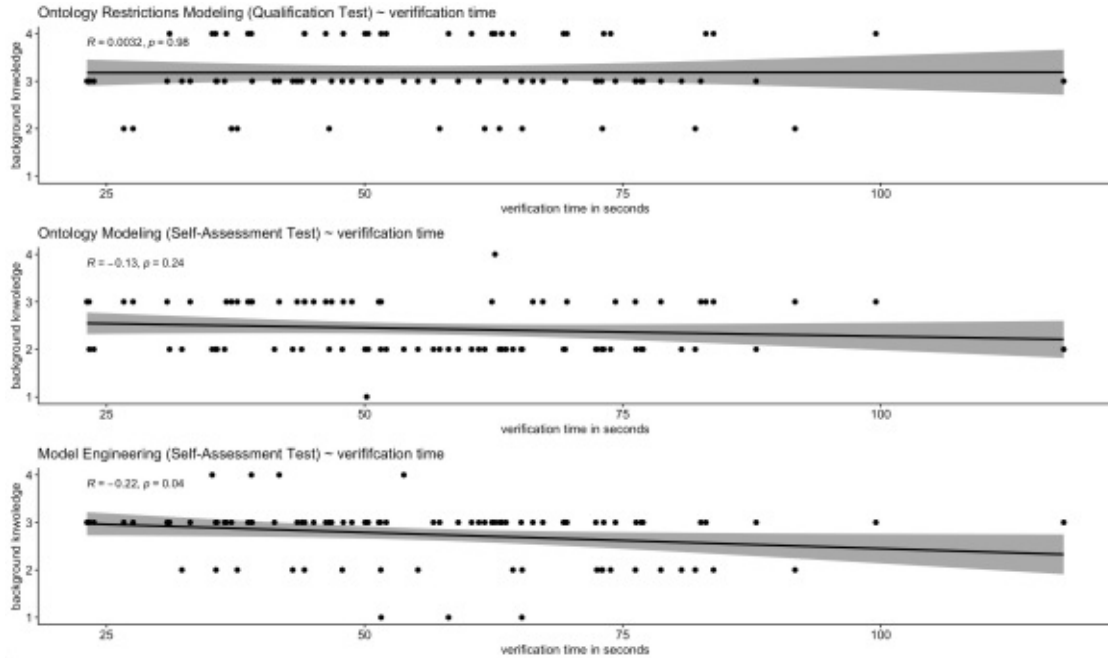
**Figure 6.6:** Point-biserial correlation between Qualification Test results and the accuracy of verifications in each representational formalism



**Figure 6.7:** Point-biserial correlation between Qualification Test results and the accuracy of verifications of each defect type

perform the verifications. At the same time the Self-Assessment Test managed to obtain better the background knowledge needed for faster responses. Prior knowledge in Model Engineering has the highest influence on the verification time and the effects of the

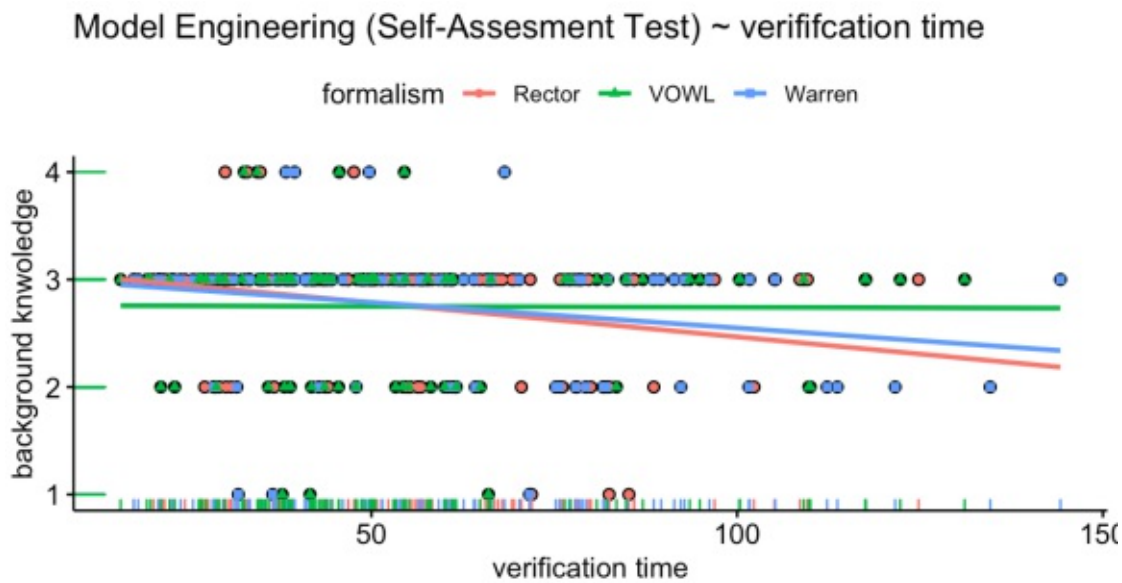
negative correlation are seen as statistically significant for the conventional significance level of 5%.



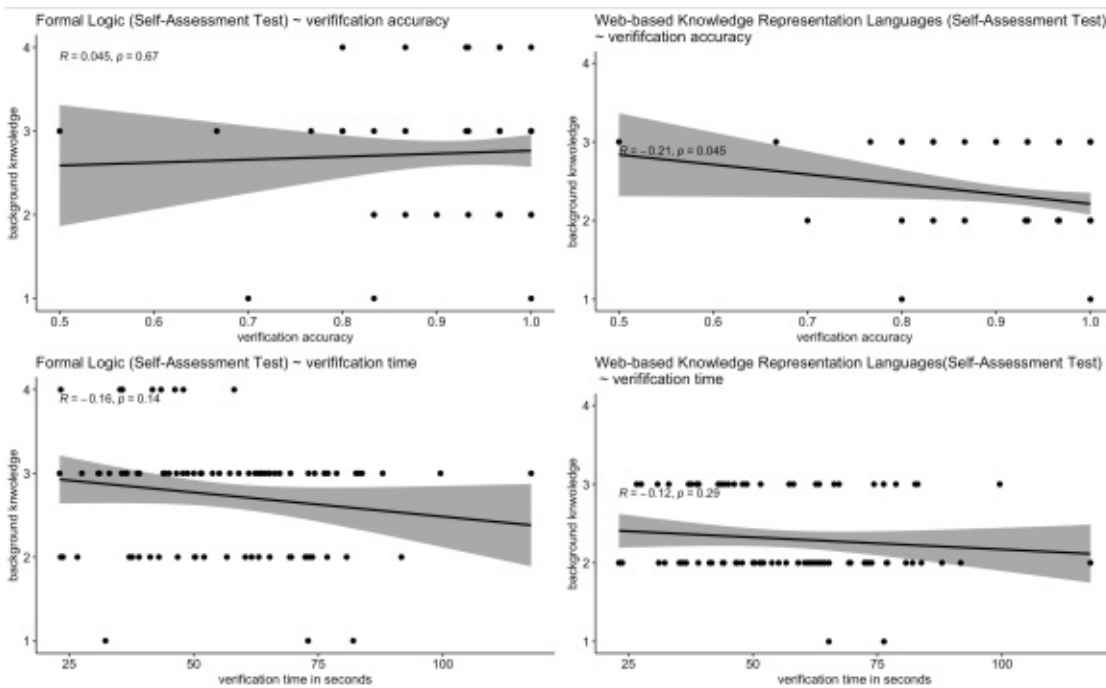
**Figure 6.8:** Point-biserial correlation between prior modeling knowledge and the time needed to perform a single verification

To understand the correlation better, Figure 6.9 shows the influence of prior knowledge in Model Engineering on the verification time in each representational formalism separately. It can be seen that the negative correlation holds only for both textual representations (Rector and Warren). Prior knowledge in Model Engineering does not reduce the time needed to verify axioms in the VOWL representation.

From the experiment results it was also possible to analyse the influence of other background knowledge areas on the performance of the contributors. In the Self-Assessment Test participants were asked to rate their knowledge in Formal Logics. However, as seen in Figure 6.10 no significant correlation was found for this background knowledge aspect neither for the accuracy nor time needed for the evaluations. Another area that was included in the Self-Assessment was the knowledge in web-based representation languages such as OWL or RDF(s). Surprisingly, students who rated their knowledge in a higher category performed worse on the evaluations. As mentioned beforehand, the self-assessment test can be very subjective, therefore, further experimental investigations are needed where these areas are evaluated objectively.



**Figure 6.9:** Point-biserial correlation between prior knowledge in Model Engineering and the time needed to perform a single verification



**Figure 6.10:** Point-biserial correlation between prior knowledge in further areas and evaluation performance

### 6.3.3 Changes of the Contributors' Performance over Time

An interesting aspect to be analysed based on the experiment results is whether the contributors learned the "patterns" of the axioms and included defects over time and whether this changed their performance.

Plotted in Figure 6.11 for each HIT in chronological order is the correctness of each participant's n-th verification. There is no significant tendency of the verifications getting better since the accuracy is already very high on the first tasks. However, it is observed that the time students needed for the verification of each following task gets significantly smaller.

These results show that the contributors got better with time since they identified the defects faster and their results stayed high.

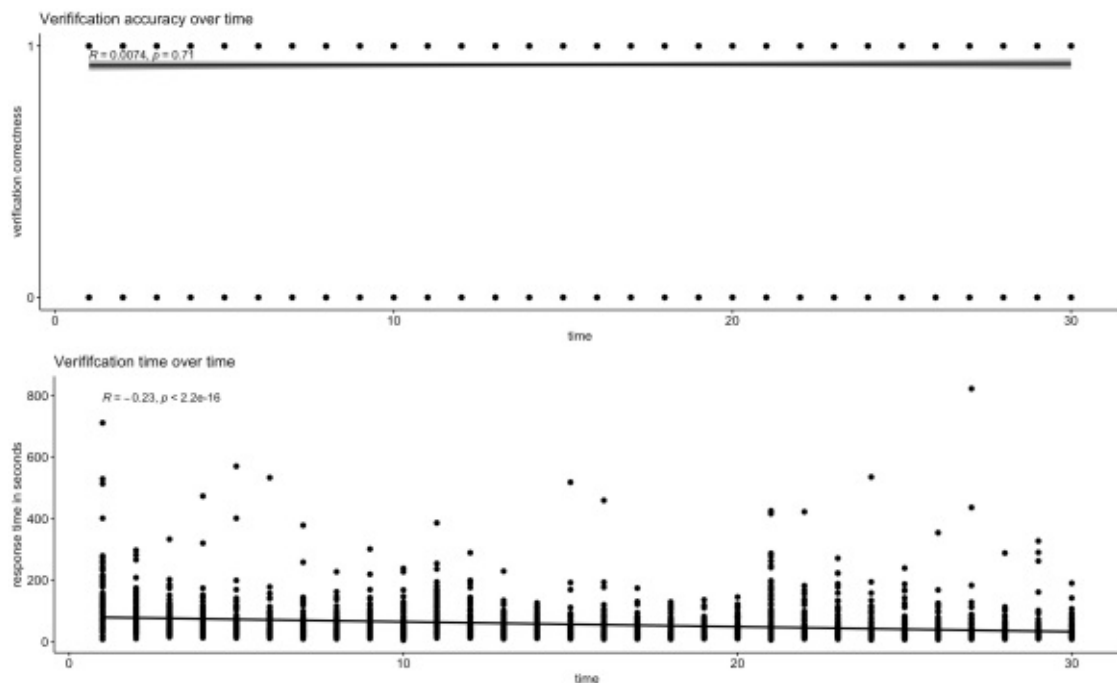


Figure 6.11: Performance of contributors' verifications for HITs over time

### 6.3.4 HIT Comments and Feedback Analysis

Part of the verification HIT design was to include an option for a comment from the contributors. In total 71 comments were gathered from the collected 2629 responses. A big portion of the comments included an explanation why a specific option was chosen. Since the correctness of such responses was very high, it can be argued that the instructions provided to the participants were clear and an understanding of the restrictions was achieved.

From the HIT comments and feedback from the post-study it was also possible to identify some unclarities amongst the evaluators:

- The participants did not completely understand the wording "amongst other things" from the textual representation.  
Many comments point out that they "assume" that a pizza also has other ingredients, which are not defined with the restricted "hasTopping"-relation such as dough or spices. This assumption is in fact correct, however, the wording causes some confusion.
- Axioms following Structure 2 were also more ambiguous than Structure-1-axioms. We saw in the previous sections that the verification of this structure also had a slightly lower accuracy. Many participants who noted this in the comments wished that there were examples of such structures in the instructions as well.
- Axioms that included other modeling properties in combination with the restrictions were hard to understand.  
For instance, when the axioms included a negation for the restriction, the participants did not understand the meaning completely. [Rector et al., 2004] and [Warren et al., 2019] identify the mistake that novice ontology engineers often misunderstand the difference between "some not" and "not some". Since the instructions only included explanations on the ontology quantifiers, some participants had troubles verifying these advanced axioms.
- Some contributors did not understand the textual representation for a union as "and/or".  
Axioms which included an union of multiple restrictions were not very clear in the textual representation. Some students noted that "and/or" was confusing for them.

The results of the comments & feedback analysis shows that the textual representation formalisms Rector & Warren were harder for students to understand and caused some unclarities because of different understanding of the used wording. These findings show again that using VOWL for representing ontology axioms can deliver the best results.

## 6.4 Evaluation Summary

In this chapter we looked into the usefulness of the approach, proposed in chapter 4 *Ontology Restrictions Verification: a Human Computation Approach*. For this purpose the results from the conducted experiment, explained in detail in chapter 5 *Setup of Evaluation Experiment*, were analysed and discussed.

Based on the experiment results we evaluated that overall 92.58% of the responses gathered in the experiment were correct and a single verification was completed on average in about a minute.

In order to further understand how well the HC approach was designed and whether it is fitted for the task of verifying ontology restrictions, we looked into the two hypotheses defined in section 5.1 *Experiment Aims*:

- Hypothesis H1: The formalism in which axioms are represented has an influence on the performance/speed of the contributors.

The investigation in the thesis showed that the axiom representation in which contributors see the ontology quantifiers has influence on the accuracy of their verification results. Presenting the axioms in the VOWL formalism can produce the best quality verification results overall, however, for some defect types, the textual representations proved to be more beneficial in terms of the quality of the evaluations. We can conclude that in future research, in case it can be detected based on the axiom structure what possible defects could appear, axioms can be translated into the formalism with which the highest quality of results can be reached, depending on the defects they might include.

- Hypothesis H2: Prior modeling knowledge has a positive influence on performance and time.

With the results shown in this chapter we report rejection of the null hypothesis. Based on the experimental investigation, we can say that using a crowd of intermediate ontology engineers, whose prior modeling knowledge was evaluated using an objective qualification rather than a subjective self-assessment, can be used to achieve the best quality of evaluations in each of the tested representational formalisms. We found out that the positive influence of prior modeling knowledge on the verification results appears only in the verification of specific defect types, meaning that in future experiments, the tasks can be distributed to participants with different modeling backgrounds so that more advanced contributors focus only on the tasks where their prior knowledge is actually beneficial.

From the comments and feedback analysis, the semi-closed design of the HC evaluation tasks proves to be very important. Providing contributors with the possibility to give their own notes on the tasks helped with identifying several limitations, such as unclarities in the textual representations of the union shown as "and/or" as suggested by [Rector et al., 2004]. Based on the identified unclarities amongst the contributors the task design can be improved, so that in future even higher percentage of verifications can be reached.

At this point, it is important to address again the contributors, who joined in the experiment and to understand that the gathered results were achieved when working with novice ontology engineers. When using a layman crowd the hypotheses investigations could have different outcomes. For instance, for master's students from a technical university, a graphical representation is easy to understand. For others, however, who do not have any modeling experience the designed HC task could be more challenging and a textual representation might result in better verifications quality. Therefore,



investigations of different setups are needed where the hypotheses can be further tested. Moreover, the experiment should in future be repeated with a more diverse crowd to avoid bias that could be introduced when the participants share the same demographic characteristics.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

## Conclusion and Future Work

Ontology Evaluation is a crucial research area to be addressed considering that the correctness of an ontology has a major impact on the quality of the system making use of it. Many tasks of the evaluation process can be solved by automated methods, however, some assessments require an expert's input, which often involves high costs and long evaluation times. Human Computation is a promising paradigm that allows the outsourcing of specific tasks to human participants at a lower cost, for instance, by combining it with Crowdsourcing techniques.

[Poveda-Villalón et al., 2014] and [Villalón and Pérez, 2016] identify several ontology pitfalls the detection and correction of which cannot be (yet) fully automated and require human intervention to be solved. This master thesis goes beyond this research by looking into specific **human-centric ontology evaluation tasks** (HOETs) and their characteristics. A Literature Review of known HOETs and how they have been approached in the literature was performed. As a result, several human-centric ontology evaluation tasks were outlined, which have not yet been approached with Human Computation techniques. Moreover, a new approach for solving one of those tasks - the verification of ontology quantifiers, using Human Computation and Crowdsourcing methods, was designed and implemented in the crowdsourcing platform Amazon Mechanical Turk. For the evaluation of the proposed solution, an experiment was carried out, the results of which show that the developed methodology can achieve excellent verification results.

In the following sections, the results of the research questions and main contributions to the field of research are described. Furthermore, limitations of the thesis are discussed and directions for future work are identified.

### 7.1 Answers to Research Questions and Discussion

In this section, we look into the research questions that this thesis was guided by and outline the main results obtained during the investigation of each of them.

- **RQ1** Which ontology evaluation tasks cannot be yet reliably automated and need human involvement?

After a thorough literature review of papers, whose focus lays on human-centric ontology evaluation, a list of human-centric ontology evolution tasks was built. For each of those tasks, various angles were investigated such as the evaluated ontology aspect, the motivation behind the evaluation, and the used evaluation method. The results show that HOETs are mostly solved by making use of human intelligence such as domain-specific knowledge or generally applicable knowledge, which cannot be yet made machine-readable.

The literature review also outlines that the tasks, which cannot be (yet reliably) solved using automated methods and have still not been approached using HC&C methods are assignments that require both domain knowledge and ontology modeling knowledge. Therefore, one such task was identified for the second part of the thesis, so that first investigations of HOETs requiring modeling knowledge can be carried out.

- **RQ2** How can Human Computation techniques be used to evaluate ontologies regarding the correct usage of restrictions?

Out of the list of HOETs not yet approached using HC&C, one task was selected - the verification of ontology quantifiers. A HC approach for the evaluation of ontology restrictions was designed and implemented on Amazon Mechanical Turk Sandbox. The main characteristics of the task design are

- (1) the extraction of ontology restrictions into axioms,
- (2) presenting contributors with the verification task of comparing a model and a real-world domain entity for context,
- (3) allowing diverse formats for the domain entity and model representational formalism to allow for task customization based on the application domain and worker crowd,
- (4) giving contributors predefined options to choose from but still allowing for an open-ended comment from their side, and
- (5) providing the contributors with needed modeling background information in the form of task instructions.

- **RQ3** How suitable are Human Computation techniques for the evaluation task of verifying ontology restrictions?

In order to allow for the evaluation of the proposed approach, an experiment was designed, piloted with experts, and executed with a student crowd of novice ontology engineers. The experiment had 3 main stages:

- (1) a pre-study where prior knowledge of the contributors was evaluated using a Self-Assessment Test and a Qualification Test, which was important for the investigation of the influence of prior modeling knowledge on the verification results,
- (2) a main stage, where contributors had to complete ontology quantifier verification tasks given a model in different representational formalisms to explore whether the formalism in which the model is shown influences the quality of the evaluation and
- (3) a post-study where feedback from the contributors was collected.

The pilot of the experiment aimed at identifying limitations of the task design while the execution of the experiment with a student crowd gathered quantitative data, which was then analysed to gain insights into the usefulness of the proposed HC approach and investigate different task design aspects.

The results from the experiment point out that the representational formalism in which ontology axioms are presented to the contributors can have an impact on their performance. VOWL stands out as a preferred formalism amongst the contributors and as the best performing one overall, however, some specific defects were easier to detect in the textual representations. Further analysis showed that prior modeling knowledge of the participants, evaluated with a qualification test, can significantly improve the verification results. Moreover, contributors, who assessed themselves as having some/expert knowledge in Model Engineering were able to complete the evaluations at a higher speed.

The discussed findings from the analysis of the experiment results can be used for future research where a representational formalism of ontology quantifiers has to be chosen and a participants crowd needs to be selected.

To summarise, the main contributions of the thesis, are (1) the analysis of known human-centric ontology evaluation tasks and their characteristics as well as the outline of HOETs not yet approached using HC&C, (2) the design and implementation of a Human Computation solution for the HOET of evaluating ontology restrictions, and (3) an experimental investigation of the proposed solution, which provides insights into main HC task design aspects.

## 7.2 Limitations & Future Work

Lastly, to conclude this master thesis, we look into the limitations of the conducted research and identify directions for future work.

To start with, the performed literature review of human-centric ontology evaluation tasks provided in chapter 3 *Human-Centric Ontology Evaluation: Literature Review* does not offer a complete overview of papers looking into HOETs. As mentioned the work done in this thesis is only a part of a larger ongoing study, which aims at providing a complete catalog of HOETs and their characteristics.

## 7. CONCLUSION AND FUTURE WORK

---

As discussed above, the proposed Human Computation solution proved to be useful for gathering high-quality verifications of ontology restrictions. However, the approach can be further improved based on the evaluation results so that the achieved accuracy can be even further increased for future evaluations. Furthermore, the approach can be extended to support multiple verification tasks in order for several aspects of the ontology evaluation to be covered. At the moment, there are still several HOETs, which have not been yet approached using Human Computation techniques. Such tasks should be investigated so that later on a generic approach can be provided for solving different types of HOETs.

A future aspect to be considered for research is also whether the same hypotheses results can be achieved when using a layman crowd for the experiment. Representing axioms in VOWL proved to be best fitted when the contributors were novice ontology engineers, however, participants with no modeling background knowledge might find another representation easier to understand. Therefore, further investigations of the hypotheses in different setups are needed.

Moreover, an extension of the developed Human Computation approach of verifying ontology restrictions using crowdsourcing techniques should be explored, where different contributions aggregation strategies are investigated to determine the method which results in the highest accuracy of the verifications.

Last but not least, the proposed Human Computation solution also proved to be very useful as part of distance learning and helped improve the learning process of ontology engineering for the students who participated in the experiment. Specifically, the difference between the ontology quantifiers was practiced, and real-world examples of good and bad modeling techniques were made clear. It is planned to repeat the experiment in the following years again as a quiz, part of teaching the course "Introduction to Semantic Systems". Therefore, an extended version of the proposed approach, where multiple ontology verification tasks are supported, would even further support the distance learning approach at Vienna University of Technology.

# List of Figures

1.1	Example of a human-centric ontology evaluation task, reproduced from [Vilalón and Pérez, 2016]. . . . .	3
1.2	Methods applied in the master thesis . . . . .	4
2.1	Example of an OWL axiom paraphrased into the Rector formalism [Rector et al., 2004] . . . . .	13
2.2	Example of an OWL axiom paraphrased into the Warren formalism . . . . .	13
2.3	Example of an OWL axiom represented in the VOWL formalism . . . . .	14
3.1	Overview of the Systematic Mapping Study process . . . . .	18
3.2	Formalism of the evaluated resources. . . . .	31
3.3	The frame of reference against which the evaluation is performed . . . . .	31
3.4	Motivation for human-centric ontology evaluation tasks. . . . .	38
3.5	Contributors' motivation to participate in the evaluation tasks. . . . .	42
3.6	Size of the used evaluation population. . . . .	43
3.7	Professions of the participants performing the HOETs. . . . .	44
3.8	Applied evaluation methods and modalities for solving HOETs. . . . .	46
4.1	Example of a HIT for the verification of ontology restrictions in Amazon Mechanical Turk . . . . .	55
4.2	Example of a HIT for the verification of ontology restrictions in Amazon Mechanical Turk . . . . .	57
4.3	Provided instructions for the verification of ontology restrictions in Amazon Mechanical Turk for the VOWL formalism . . . . .	58
4.4	Provided examples for context for the verification of ontology restrictions in Amazon Mechanical Turk for the VOWL formalism . . . . .	59
5.1	Overview of the group assignments and data split . . . . .	64
5.2	Overview of the experiment workflow. H1 = Hypothesis H1, H2 = Hypothesis H2 . . . . .	65
5.3	Example of a knowledge level scale from the Self-Assessment Test on Google Forms . . . . .	66
5.4	Example question from the Qualification Test on Amazon Mechanical Turk . . . . .	67
5.5	Example of a HIT from the Tutorial Job from Amazon Mechanical Turk . . . . .	67
		91

6.1	Background knowledge of the participants in each study group . . . . .	71
6.2	Influence of different representations of universal and existential restrictions on the performance of the contributors. . . . .	75
6.3	Influence of axiom representation on the performance of the contributors in identifying different defects in ontology restrictions. . . . .	76
6.4	Influence of axiom representation on the performance of the contributors in identifying defects in different axiom structures. . . . .	77
6.5	Point-biserial correlation between prior modeling knowledge and the accuracy of verifications . . . . .	78
6.6	Point-biserial correlation between Qualification Test results and the accuracy of verifications in each representational formalism . . . . .	79
6.7	Point-biserial correlation between Qualification Test results and the accuracy of verifications of each defect type . . . . .	79
6.8	Point-biserial correlation between prior modeling knowledge and the time needed to perform a single verification . . . . .	80
6.9	Point-biserial correlation between prior knowledge in Model Engineering and the time needed to perform a single verification . . . . .	81
6.10	Point-biserial correlation between prior knowledge in further areas and evaluation performance . . . . .	81
6.11	Performance of contributors' verifications for HITs over time . . . . .	82



## List of Tables

3.1	Research Questions defined for the Systematic Mapping Study . . . . .	19
3.2	Sub-queries for the overall search query $Q = Q1SW \cap Q2HC \cap Q3Eval$ used for the SMS study selection . . . . .	20
3.3	Inclusion and Exclusion Criteria defined for the SMS Study Selection . . . . .	21
3.4	Systematic Mapping Study Data Extraction template . . . . .	22
3.5	Overview of the included papers. . . . .	27
3.6	Type of evaluated resource and verified aspects of it . . . . .	28
3.7	The frame of reference against which the evaluation is performed. . . . .	32
3.8	Ontology mistakes and how they can be solved . . . . .	33
3.9	Context for performing human-centric ontology evaluation tasks . . . . .	37
3.10	Evaluation population . . . . .	39
3.11	Ontology verification tasks and how they have been approached in the literature	45
3.12	Evaluation metrics and inter-evaluator agreement . . . . .	47
3.13	Ontology verification tasks that have not yet been approached using Human Computation techniques in the literature . . . . .	50
6.1	Performance of the evaluators in verifying ontology quantifiers for each axiom for the Rector/VOWL/Warren formalism . . . . .	72
6.2	Performance of the evaluators in verifying ontology quantifiers based on representational formalism . . . . .	74



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Bibliography

- [SKO, 2009] (2009). Skos simple knowledge organization system reference. <https://www.w3.org/TR/skos-reference/>. Accessed: 2020-09-20.
- [OWL, 2012] (2012). Web ontology language (owl). <https://www.w3.org/OWL/>. Accessed: 2020-09-20.
- [RDF, 2014] (2014). Resource description framework (rdf). <https://www.w3.org/RDF/>. Accessed: 2020-09-20.
- [UKC, 2020] (2020). Ukc universal knowledge core. <http://ukc.disi.unitn.it>. Accessed: 2020-09-20.
- [Acosta et al., 2013] Acosta, M., Zaveri, A., Simperl, E., Kontokostas, D., Auer, S., and Lehmann, J. (2013). Crowdsourcing linked data quality assessment. In *International semantic web conference*, pages 260–276. Springer.
- [Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific american*, 284(5):34–43.
- [Brank et al., 2005] Brank, J., Grobelnik, M., and Mladenic, D. (2005). A survey of ontology evaluation techniques. In *Proceedings of the conference on data mining and data warehouses (SiKDD 2005)*, pages 166–170. Citeseer Ljubljana, Slovenia.
- [Chang et al., 2011] Chang, T.-H., Kuo, Y.-L., and Hsu, J. Y.-j. (2011). Actraversal: ranking crowdsourced commonsense assertions and certifications. In *International Conference on Principles and Practice of Multi-Agent Systems*, pages 234–246. Springer.
- [Dresch et al., 2014] Dresch, A., Lacerda, D., and Antunes, J. (2014). *Design Science Research: A Method for Science and Technology Advancement*. Springer.
- [Dumitrache et al., 2015] Dumitrache, A., Aroyo, L., and Welty, C. (2015). Achieving expert-level annotation quality with crowdtruth. In *Proc. of BDM2I Workshop, ISWC*.
- [Erez et al., 2015] Erez, E. S., Zhitomirsky-Geffet, M., and Bar-Ilan, J. (2015). Subjective vs. objective evaluation of ontological statements with crowdsourcing. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

- [Hanika et al., 2014] Hanika, F., Wohlgenannt, G., and Sabou, M. (2014). The ucomp protege plugin for crowdsourcing ontology validation. In *International Semantic Web Conference (Posters & Demos)*, pages 253–256.
- [Hees et al., 2011] Hees, J., Roth-Berghofer, T., Biedert, R., Adrian, B., and Dengel, A. (2011). Betterrelations: using a game to rate linked data triples. In *Annual Conference on Artificial Intelligence*, pages 134–138. Springer.
- [Hevner, 2007] Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian journal of information systems*, 19(2):4.
- [Hevner et al., 2004] Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). Design science in information systems research. *MIS quarterly*, pages 75–105.
- [Karam and Melchiori, 2013] Karam, R. and Melchiori, M. (2013). Improving geo-spatial linked data with the wisdom of the crowds. In *Proceedings of the joint EDBT/ICDT 2013 workshops*, pages 68–74.
- [Kehagias et al., 2008] Kehagias, D. D., Papadimitriou, I., Hois, J., Tzouvaras, D., and Bateman, J. (2008). A methodological approach for ontology evaluation and refinement. In *ASK-IT Final Conference. June.(Cit. on p.)*, pages 1–13.
- [Ketterl et al., 2011] Ketterl, M., Knipping, L., Ludwig, N., Mertens, R., Waitelonis, J., Knuth, M., and Sack, H. (2011). Whoknows? evaluating linked data heuristics with a quiz that cleans up dbpedia. *Interactive Technology and Smart Education*.
- [Kiptoo, 2020] Kiptoo, C. C. (2020). Ontology enhancement using crowdsourcing: a conceptual architecture. *International Journal of Crowd Science*.
- [Kitchenham et al., 2011] Kitchenham, B. A., Budgen, D., and Brereton, O. P. (2011). Using mapping studies as the basis for further research—a participant-observer case study. *Information and Software Technology*, 53(6):638–651.
- [Kitchenham and Charters, 2007] Kitchenham, B. A. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering technical report. *Software Engineering Group, EBSE Technical Report, Keele University and Department of Computer Science University of Durham*, 2.
- [Knuth et al., 2012] Knuth, M., Hercher, J., and Sack, H. (2012). Collaboratively patching linked data. *arXiv preprint arXiv:1204.2715*.
- [Kontokostas et al., 2013] Kontokostas, D., Zaveri, A., Auer, S., and Lehmann, J. (2013). Triplecheckmate: A tool for crowdsourcing the quality assessment of linked data. In *International Conference on Knowledge Engineering and the Semantic Web*, pages 265–272. Springer.

- [Kurita et al., 2016] Kurita, D., Roengsamut, B., Kuwabara, K., and Huang, H.-H. (2016). Knowledge base refinement with gamified crowdsourcing. In *Asian Conference on Intelligent Information and Database Systems*, pages 33–42. Springer.
- [LaToza and Van Der Hoek, 2015] LaToza, T. D. and Van Der Hoek, A. (2015). Crowdsourcing in software engineering: Models, motivations, and challenges. *IEEE software*, 33(1):74–80.
- [McDaniel and Storey, 2019] McDaniel, M. and Storey, V. C. (2019). Evaluating domain ontologies: clarification, classification, and challenges. *ACM Computing Surveys (CSUR)*, 52(4):1–44.
- [Mortensen, 2013] Mortensen, J. M. (2013). Crowdsourcing ontology verification. In *International Semantic Web Conference*, pages 448–455. Springer.
- [Mortensen et al., 2015] Mortensen, J. M., Minty, E. P., Januszyk, M., Sweeney, T. E., Rector, A. L., Noy, N. F., and Musen, M. A. (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):640–648.
- [Mortensen et al., 2013] Mortensen, J. M., Musen, M. A., and Noy, N. F. (2013). Crowdsourcing the verification of relationships in biomedical ontologies. In *AMIA Annual symposium proceedings*, volume 2013, page 1020. American Medical Informatics Association.
- [Mortensen et al., 2016] Mortensen, J. M., Telis, N., Hughey, J. J., Fan-Minogue, H., Van Auken, K., Dumontier, M., and Musen, M. A. (2016). Is the crowd better as an assistant or a replacement in ontology engineering? an exploration through the lens of the gene ontology. *Journal of biomedical informatics*, 60:199–209.
- [Nuzzolese et al., 2017] Nuzzolese, A. G., Presutti, V., Gangemi, A., Peroni, S., and Ciancarini, P. (2017). Aemoo: Linked data exploration based on knowledge patterns. *Semantic Web*, 8(1):87–112.
- [Poveda-Villalón et al., 2014] Poveda-Villalón, M., Gómez-Pérez, A., and Suárez-Figueroa, M. C. (2014). Oops!(ontology pitfall scanner!): An on-line tool for ontology evaluation. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 10(2):7–34.
- [Rector et al., 2004] Rector, A., Drummond, N., Horridge, M., Rogers, J., Knublauch, H., Stevens, R., Wang, H., and Wroe, C. (2004). Owl pizzas: Practical experience of teaching owl-dl: Common errors & common patterns. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 63–81. Springer.
- [Roengsamut and Kuwabara, 2015] Roengsamut, B. and Kuwabara, K. (2015). Interactive refinement of linked data: toward a crowdsourcing approach. In *Asian Conference on Intelligent Information and Database Systems*, pages 3–12. Springer.

- [Roengsamut et al., 2015] Roengsamut, B., Kuwabara, K., and Huang, H.-H. (2015). Toward gamification of knowledge base construction. In *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*, pages 1–7. IEEE.
- [Sabou et al., 2018a] Sabou, M., Aroyo, L., Bontcheva, K., Bozzon, A., and Qarout, R. K. (2018a). Semantic web and human computation: The status of an emerging field. *Semantic Web*, 9(3):291–302.
- [Sabou et al., 2018b] Sabou, M., Winkler, D., Penzerstadler, P., and Biffl, S. (2018b). Verifying conceptual domain models with human computation: A case study in software engineering. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.
- [Studer et al., 1998] Studer, R., Benjamins, V. R., and Fensel, D. (1998). Knowledge engineering: principles and methods. *Data & knowledge engineering*, 25(1-2):161–197.
- [Teitsma et al., 2014] Teitsma, M., Sandberg, J., Schreiber, G., Wielinga, B., and van Hage, W. R. (2014). Engineering ontologies for question answering. *Applied Ontology*, 9(1):1–25.
- [Ul Hassan et al., 2016] Ul Hassan, U., Zaveri, A., Marx, E., Curry, E., and Lehmann, J. (2016). Acryliq: Leveraging dbpedia for adaptive crowdsourcing in linked data quality assessment. In *European Knowledge Acquisition Workshop*, pages 681–696. Springer.
- [Villalón and Pérez, 2016] Villalón, M. P. and Pérez, A. G. (2016). Ontology evaluation: a pitfall-based approach to ontology diagnosis. *PhD Tesis, Universidad Politecnica de Madrid, Escuela Tecnica Superior de Ingenieros Informaticos*.
- [Warren et al., 2019] Warren, P., Mulholland, P., Collins, T., and Motta, E. (2019). Improving comprehension of knowledge representation languages: A case study with description logics. *International Journal of Human-Computer Studies*, 122:145–167.
- [Wohlgenannt et al., 2016] Wohlgenannt, G., Sabou, M., and Hanika, F. (2016). Crowd-based ontology engineering with the ucomp protégé plugin. *Semantic Web*, 7(4):379–398.
- [Zaveri et al., 2013] Zaveri, A., Kontokostas, D., Sherif, M. A., Bühmann, L., Morsey, M., Auer, S., and Lehmann, J. (2013). User-driven quality evaluation of dbpedia. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 97–104.
- [Zaveri et al., 2016] Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., and Auer, S. (2016). Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93.
- [Zhang et al., 2017] Zhang, H., Ojha, S. R., and Giunchiglia, F. (2017). Finding errors in a chinese lexico-semantic resource using gwap. In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, pages 77–80. IEEE.

# Appendices

## Appendix A: Included Papers for HOETs Analysis

In this section, the papers included in the literature review are summarised to provide a better overview of the scope and some context for the analysis.

### **P1 Engineering ontologies for question answering** [Teitsma et al., 2014]

In this paper, the authors present methods for developing ontologies for question answering systems - a pragmatic (developed from a categorization used at an emergency call center), an expert-based (constructed by a knowledge engineer), and a basic-level ontology (constructed by ordinary people), as well as an evaluation framework for evaluating and comparing them. The idea of such a system is to enable gathering information from ordinary people witnessing a crisis situation by generating questions to assess the happening and providing possible answers to determine further details. The different ontologies are evaluated based on multiple criteria - structure, efficiency of construction, completeness, and cognitive semantics, for which some user-experiments were conducted.

### **P2 Aemoo: Linked data exploration based on knowledge patterns** [Nuzzolese et al., 2017]

The authors present a tool (Aemoo) for supporting users in exploratory searches. The main topic are Encyclopedic Knowledge Patterns (EKP), which can be described as small ontologies, automatically extracted from Wikipedia. EKPs organize and visualize knowledge, and are used as criteria to determine the relevance of connections for entities, which can be particularly useful in exploratory searches. The paper also describes user-based controlled experiments, conducted to evaluate the cognitive soundness of the EKPs and their performance when implemented in the Aemoo tool.

### **P3 Crowdsourcing Linked Data Quality Assessment** [Acosta et al., 2013]

Acosta et al. look at common problems with the quality of Linked Data and to what extent they could be solved by crowdsourcing. The authors first develop a contest for experts and enthusiasts with the goal of finding incorrect RDF triples and classifying them based on the quality problem. The resulting set of the erroneous triples was provided to *layman* crowds, who were then asked to identify the quality issue out of a predefined taxonomy of errors, provided some Wikipedia links for context. It becomes clear that the two crowdsourcing approaches (expert-based and layman) complement each other and can be used for enhancing the quality of Linked Data resources in an affordable way.

#### **P4 Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of SNOMED CT** [Mortensen et al., 2015]

The paper presents a crowd-based verification method for a subset of the SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) Ontology, "*an ontology that the US government now mandates for use in the clinic as part of 'Meaningful Use' of electronic health records*" [Mortensen et al., 2015]. The performance of the crowd was compared to the results of a panel of experts that were given the same tasks. The authors determine that the crowd performed almost as well as a single expert in identifying errors in the relationships of the ontology and can thus be used in ontology verification tasks when the budget is limited or an expert is not available.

#### **P5 Is the crowd better as an assistant or a replacement in ontology engineering? An exploration through the lens of the Gene Ontology** [Mortensen et al., 2016]

Mortensen et al. compare how crowds perform in different set-ups. Firstly, crowds and experts are asked to identify ontology errors in a subset of the Gene Ontology (a large biomedical ontology). In comparison to the experts, the crowd did not perform very well. The authors argue that the reason for this could be that there are not many Google search results for the Gene Ontology, compared to the SNOMED CT Ontology, for which crowds showed much higher precision as described in a previous study [Mortensen et al., 2015]. However, in the second experiment experts were asked to identify relationships that could be easily understood by non-experts. The crowds then only acted as assistance in the verification process, which showed performance improvements while allowing experts to focus on more difficult tasks.

#### **P6 Crowdsourcing the verification of relationships in biomedical ontologies** [Mortensen et al., 2013]

The authors use crowdsourcing to verify a small subset of the SNOMED CT Ontology, consisting of taxonomic relationships. The correctness of the relationships is predefined - one half contains errors while the other half is correct. The crowds manage to identify 85% of the errors and the authors discuss that when the crowd is provided with enough context no prior domain qualification tests are needed.

#### **P7 Crowdsourcing Ontology Verification** [Mortensen, 2013]

In this paper, the authors report on various experiments with different crowdsourcing set-ups for the verification of hierarchical ontology relationships. It is investigated how task formulations, provided context, and qualification of workers can influence the evaluation results. Mortensen shows that the best performance can be established when the tasks are formulated as simply as possible, concept definitions are provided as context and the workers have passed a domain qualification test.

#### **P8 BetterRelations: Using a game to rate linked data triples** [Hees et al., 2011]

Hees et al. identify that triple importance is often missing from Linked Data. They develop a two-player game, called BetterRelations, in which users can rank the importance of relationships. This is done by asking the players to choose the fact that their partner



is most likely to know and choose as well. The topics are extracted from the most visited Wikipedia pages to ensure that they are relevant and well-known.

### **P9 WhoKnows? Evaluating linked data heuristics with a quiz that cleans up DBpedia** [Ketterl et al., 2011]

The authors of the paper develop a quiz-like game in order to motivate users in a playful way to rank property importance and to help cleanse DBpedia resources of inconsistencies that often occur when using automated extraction methods. The players are asked questions about a topic and need to select the correct answer(s). Immediately afterward they see the result and can submit a complaint in case they think the answer was inconsistent or includes some kind of conflict. The game results contribute to calculating the property ranking as well as identifying inconsistencies in the data.

### **P10 Collaboratively Patching Linked Data** [Knuth et al., 2012]

The authors propose a method for keeping track of inconsistencies in a data set by creating patch requests in a suggested PatchR vocabulary, which also includes some provenance information and storing them in a centralized repository. To show a use case of how this can be implemented, the authors extend the game WhoKnows? that has been described already above. When a player submits a complaint about a specific question, they can now also specify why they are unhappy with the provided answer by selecting a reason out of a predefined set of inconsistencies. After the user votes, a patch request is automatically created based on the type of the identified problem.

### **P11 Finding errors in a Chinese lexico-semantic resource using GWAP** [Zhang et al., 2017]

The authors present a game with a purpose (GWAP) that aims to find errors in multilingual resources that occurred during an automatic translation. The players are presented with an English question and some answers in Chinese. The correct answers are assumed to be those included in the original resource, however, if most players select a different option there could be an incorrect translation and this is taken into consideration.

### **P12 ACryLIQ: Leveraging DBpedia for Adaptive Crowdsourcing in Linked Data Quality Assessment** [Ul Hassan et al., 2016]

The authors look into crowd workers' reliability and proper ways to assign each task to the best-suited worker. They propose a method for automatically generating test questions from DBpedia to assess the knowledge of workers about the assessment topic. Based on the results from these tests the actual tasks are assigned to the most reliable workers to improve the accuracy of the results.

### **P13 Interactive Refinement of Linked Data: Toward a Crowdsourcing Approach** [Roengsamut and Kuwabara, 2015]

In the paper, a multilingual rental apartment Frequently Asked Questions (FAQ) system is presented, which refines data interactively from casual users. The goal is to make it easier for non-native students to get answers in case something in their rented apartment is not working or is broken. Each question entry is connected to a part of a floor plan based on keywords from the question. In order to overcome mismatched mappings,

non-experts would be asked to select the most related part of the floor plan to a specific keyword, which is saved in an ontology. Roengsamut et al. also present the idea of having the students upload a picture of their problem in case they do not know how to describe it in words. Non-experts would then be asked to assign keywords to the image.

**P14 Toward gamification of knowledge base construction** [Roengsamut et al., 2015]

Roengsamut et al. further develop the idea from the previously described paper ([Roengsamut and Kuwabara, 2015]) by developing the FAQ system as a game. The players answer a quiz where they either need to add tags to an image or answer questions about a keyword and a section of the floor plan as well as the relationship between the two. Based on the submissions from players the underlying ontology is being updated.

**P15 Knowledge Base Refinement with Gamified Crowdsourcing** [Kurita et al., 2016]

This paper shows a simulation model for the game presented in ([Roengsamut et al., 2015]) and already discussed above. The authors investigate whether the game would serve the purpose of database refinement. In the simulation experiment changes in the weights of some relationships are examined as well as the reliability of users. Based on the experiment the authors come to the conclusion that the game would indeed help correct errors in the data set.

**P16 ACTraversal: Ranking Crowdsourced Commonsense Assertions and Certifications** [Chang et al., 2011]

The authors propose a method for data verification in which text mining algorithms and GWAP results are combined to achieve higher precision. The paper looks into an algorithm that ranks the results from verification games based on mining techniques for building a commonsense database of higher quality.

**P17 Crowd-based ontology engineering with the uComp Protégé plugin** [Wohlgenannt et al., 2016]

The paper presents a plugin for Protégé<sup>1</sup> which allows for outsourcing specific tasks to human participants during the ontology engineering process. To enable the task validation the plugin sends tasks to Games with a Purpose (GWAP) or paid-for crowdsourcing using CrowdFlower<sup>2</sup>. The authors also outline specific verification tasks for which the plugin would be of importance.

**P18 TripleCheckMate: A Tool for Crowdsourcing the Quality Assessment of Linked Data** [Kontokostas et al., 2013]

Kontokostas et al. propose an approach for assessing the quality of Linked Data resources, which were extracted automatically or semi-automatically. They use TrippleCheckMate, a tool that provides users with an individual resource and asks users to verify its correctness. In case the participant thinks there is an error included, they specify the exact type of inconsistency from a predefined taxonomy of quality problems.

---

<sup>1</sup><https://protege.stanford.edu>

<sup>2</sup><https://appen.com>

### **P19 User-driven quality evaluation of DBpedia** [Zaveri et al., 2013]

This paper is a prestudy of the paper described above [Kontokostas et al., 2013]. Here the authors identify common quality problems in Linked Data resources and represent them in a taxonomy. They categorize those into the following categories: Accuracy, Relevancy, Representational consistency, and Interlinking. The second half of the paper describes a crowdsourcing experiment done using TripleCheckMate, where users verified the correctness of RDF triples and assign them to one of the issues included in the taxonomy if they were problematic.

### **P20 Subjective vs. objective evaluation of ontological statements with crowdsourcing** [Erez et al., 2015]

The authors aim to explore subjectivity modeled in ontologies. They argue that experts tend to build ontologies based on their personal beliefs and experience. Using crowdsourcing Erez et al. investigate how workers perceive different viewpoints and controversial facts. They do this by providing the crowd with statements from the ontology and asking them to classify them into one of the categories - correct, incorrect, or controversial.

### **P21 Achieving Expert-Level Annotation Quality with CrowdTruth The Case of Medical Relation Extraction** [Dumitrache et al., 2015]

Dumitrache et al. present a method for creating a ground truth for training models based on user annotations. They allow capturing of different interpretations by taking into consideration disagreements among the annotators. The authors perform two experiments where the users had to, among others, annotate the relationship between two terms. In the approach both experts and crowd workers participated in the annotation process.

### **P22 Improving geo-spatial linked data with the wisdom of the crowds** [Karam and Melchiori, 2013]

In the paper, an approach to organize user corrections of geo-spatial linked data is discussed, driven by the fact that such data often includes conflicts or low-quality entries. The authors develop a framework where it is possible to store corrections and completions submitted by users and also rank them to achieve better accuracy and completeness of the data.

### **P23 Ontology enhancement using crowdsourcing: a conceptual architecture** [Kiptoo, 2020]

This paper concentrates on ontology enhancement achieved by crowdsourcing and investigates how non-experts can improve and complete ontology taxonomic knowledge. This is done in the form of fruit fly identification via crowd workers. The participants have the task to tag the features that they see out of a predefined set, provided an image of a fruit fly.

### **P24 Ontology Evaluation - a pitfall-based approach to ontology diagnosis** [Villalón and Pérez, 2016]

In this paper, a collection of bad practices, which are often to be seen in ontology development, are presented. The author goes into detail for each problem and provides examples, ways to detect such errors, and ways to solve them. The detection of errors

was automated as far as this was possible. The paper also outlines several tasks that require human-domain knowledge and thus require a user-centric evaluation.

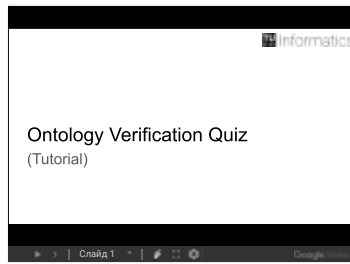
In the following sections, the human-centric ontology evaluation tasks from the described papers are analysed.

# Appendix B: Experiment Starting Point for the Participants from Group A

## Ontology Verification

### Quiz

The main focus is the validation of ontology restrictions by detecting common mistakes in the use of the universal and existential quantifiers.



- To complete each phase of the assignment, please follow the links in the provided order.
- Once you have completed a task, go back to this page and continue with the next link
- Preferably, use Safari, Google Chrome, Firefox or Edge as a web browser. The tasks on Amazon Mechanical Turk will not load in Internet Explorer.
- Please also note that you need to complete all phases of the quiz to get the assignment points.

### Important Hints

Check the instructions for the Tutorial and each part of the Verification Quiz.

### Prerequisites

- a Google Account: [Google Account Creation](#).
- an Amazon Account: [Amazon Account Creation](#)
- an Amazon Mechanical Turk Sandbox Worker Account: [MTurk Sandbox Worker Account Creation](#).

### Phase 1 : Preparation & Pre-Study (approx. 50 min)

- Complete the [Self Assessment Test](#) (Google Forms) (approx. 15 min)
- [Qualification Test](#) (MTurk Sandbox) (approx. 15 min)  
Don't forget to accept the HIT before you start answering the questions.  
Make sure to use the **Submit**-button after you complete the last section (bottom left of the screen). If you return the HIT (bottom/top right of the screen) your answers will not be submitted and the rest of the quiz cannot be matched to your StudentID.
- [Tutorial](#) (MTurk Sandbox). (approx. 20 min)  
Don't forget to use the Instructions-button to view rules and tips about the tasks in the different formalism.

--- Optional Break (approx 20 min) ---

### Phase 2 : Verification Quiz (approx. 60 min)

- [Part 1 - Rector Formalism](#)  
Don't forget to use the Instructions-button to view rules and tips about the tasks in the Rector formalism.
- [Part 2 - Warren Formalism](#)  
Don't forget to use the Instructions-button to view rules and tips about the tasks in the Warren formalism.
- [Part 3 - VOWL Formalism](#)  
Don't forget to use the Instructions-button to view rules and tips about the tasks in the VOWL formalism.

If you encounter any issues, support will be available on 11.12.2020 from 13:00 to 15:30 in the [Quiz Support Zoom Meeting](#).

[Quiz Support Zoom Meeting Details](#)

## Appendix C: Self Assessment Test

### Self-Assessment Test

Your answers will be treated anonymously. Your name will only be used to connect your answers to your inspection records. No individual information will be made public in any form.

**\*Required**

Name \*

Your answer

Student ID (Matrikelnummer) \*

Your answer

#### English-Language Skills

For the question below, please consider the following levels:

1 - no understanding: My understanding of English is very basic.

2 - little understanding: I can understand and use familiar everyday expressions and very basic phrases.

3 - some understanding: I can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc.

4 - expert understanding: I can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in my field of specialization, and can recognize implicit meaning.

Q1: How would you rate your understanding of English documents? \*

no understanding      1      2      3      4      expert understanding

### Experience with Formal Logics

For the question below, please consider the following levels:

- 1 - no knowledge: I don't have experience in the area.
- 2 - little knowledge: I am aware of the basic symbolic notation and understand the meaning of logical conjunctions and quantifiers.
- 3 - some knowledge: I have an understanding of logical axioms and can derive the conveyed implications from them.
- 4 - expert knowledge: I fully understand logical axioms and can derive explicit and implicit implications from them.

Q2: How would you rate your knowledge of formal logic? \*

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

### General Modeling Skills

For the question 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 model components and can recognise them in graphical and textual representations.
- 3 - some knowledge = I have performed modeling as part of my education/study assignments.
- 4 - expert knowledge = I have performed extensive modeling during my professional employment.

Q3: How would you rate your knowledge in Model-Driven Engineering? \*

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

Q4: Select the areas in which you already have at least some experience:

- Entity–Relationship (ER) diagrams
- Unified Modeling Language (UML)
- Stock and Flow diagram (SFD)
- Causal Loop Diagram (CLD)
- Business Process Model and Notation (BPMN)
- Event-Driven Process Chain (EPC)
- Other:

### 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

Q6: How would you rate your knowledge of web-based knowledge representation languages (e.g., RDF(S), OWL)? \*

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



Q7: Select the areas in which you already have some experience:

- Web Ontology Language (OWL)
- RDF Schema (RDFS)
- Simple Knowledge Organization System (SKOS)
- Other:

### Crowdsourcing

Q8: Do you have any experience with Crowdsourcing platforms? \*

- Yes
- No

**Thank you for your participation!**

Submit

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google. [Report Abuse](#) - [Terms of Service](#) - [Privacy Policy](#)

Google Forms



## Appendix D: Qualification Test

### Qualification Test

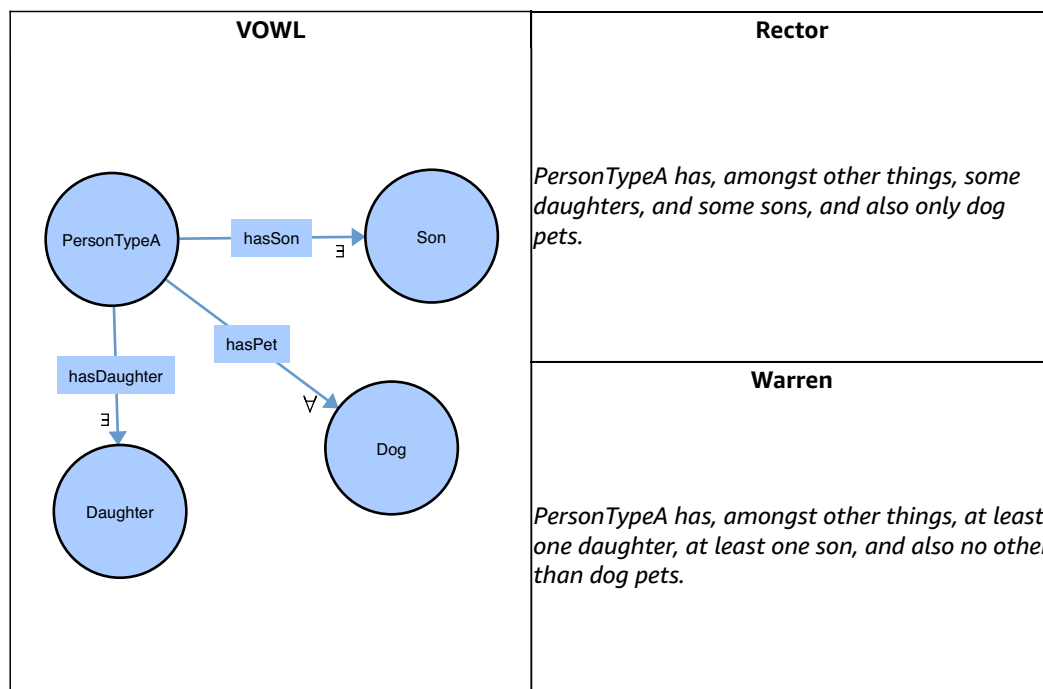
Make sure to accept the HIT before you start answering the questions.

Please enter your Student ID:

#### Section 1: Little Knowledge

This section tests your understanding of basic ontology components and the ability to recognise them in graphical and textual representations.

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?

#### 2. Identifying the different quantifiers from the model

How many universal restrictions (owl:allValuesFrom) can you identify in the model?

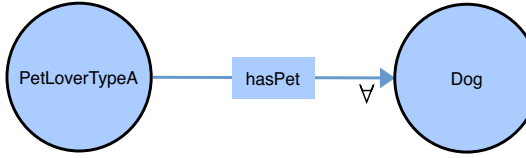
How many existential restrictions (owl:someValuesFrom) can you identify in the model?

Continue to Section 2

## Section 2: Some Knowledge

This section tests your understanding of the implications of ontology axioms and restrictions.

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

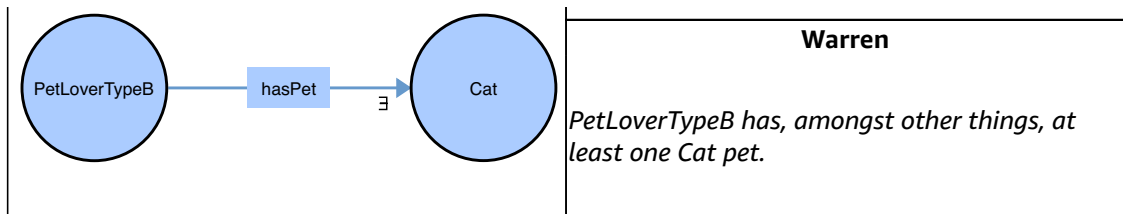
<b>VOWL</b>  	<p style="text-align: center;"><b>Rector</b></p> <p><i>PetLoverTypeA has, amongst other things, only Dog pets.</i></p> <hr/> <p style="text-align: center;"><b>Warren</b></p> <p><i>PetLoverTypeA has, amongst other things, no other than Dog pets.</i></p>
---	--

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.

**Consider the model, represented in 3 equivalent formalisms (VOWL | Rector | Warren) and answer question 4 below.**

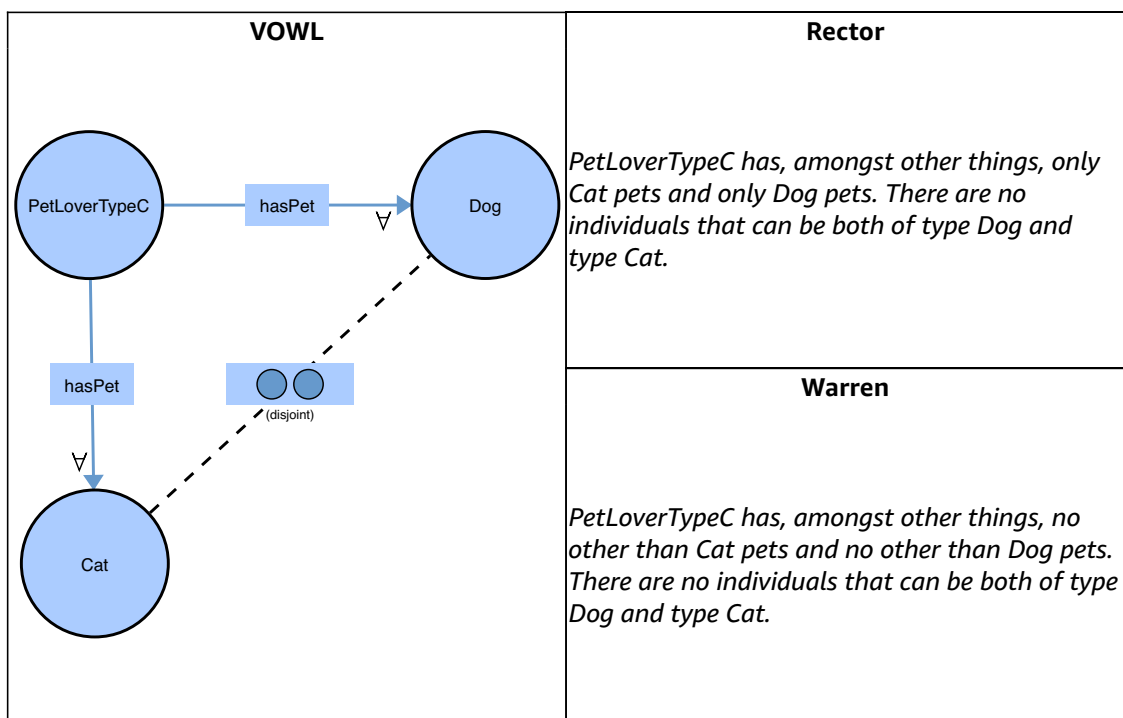
<b>VOWL</b>	<p style="text-align: center;"><b>Rector</b></p> <p><i>PetLoverTypeB has, amongst other things, some Cat pets.</i></p>
-------------	--



4. Select the statement that describes instances of PetLoverTypeB correctly.

- Instances of PetLoverTypeB must have a Cat pet and cannot have other types of pets.
- Instances of PetLoverTypeB might not have a Cat pet and cannot have other types of pets.
- Instances of PetLoverTypeB must have a Cat pet and can also have other types of pets.
- Instances of PetLoverTypeB might not have a Cat pet and can also have other types of pets.

Consider the model, represented in 3 equivalent formalisms (VOWL | Rector | Warren) and answer question 5 below.



5. Select the statement that correctly represents instances of PetLoverTypeC.

- Instances of PetLoverTypeC must have 2 pets - a Dog and a Cat.
- Instances of PetLoverTypeC might have 2 pets - a Dog and a Cat but also might not have any pets.

- Instances of PetLoverTypeC cannot have any pets.
- Instances of PetLoverTypeC could have 0 to n pets from type Cat or 0 to n pets from type Dog but not both.

Continue to Section 3

### Section 3: Expert Knowledge

This section tests your ability to reason with ontology models, as well as compare and relate them to each other.

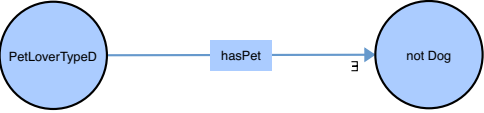

Consider models A and B both describing PetLoverTypeE, each represented in 3 equivalent formalisms (VOWL | Rector | Warren) and answer question 6 below.

Model A: PetLoverTypeE	Model B: PetLoverTypeE
<p><b>VOWL</b></p>	<p><b>VOWL</b></p>
<p><b>Rector</b></p> <p><i>PetLoverTypeE has, amongst other things, some Cat pets and some Dog pets and also only Cat and Dog pets.</i></p>	<p><b>Rector</b></p> <p><i>PetLoverTypeE has, amongst other things, some Cat pets and some Dog pets.</i></p>
<p><b>Warren</b></p> <p><i>PetLoverTypeE has, amongst other things, at least one Cat pet and at least one Dog pet and also no other than Cat and Dog pets.</i></p>	<p><b>Warren</b></p> <p><i>PetLoverTypeE has, amongst other things, at least one Cat pet and at least one Dog pet.</i></p>

6. Select the correct statement about models A and B describing PetLoverTypeE.

- Model A allows for instances of PetLoverTypeE to have a pet that is neither a Dog nor a Cat.
- Model B allows for instances of PetLoverTypeE to have a pet that is neither a Dog nor a Cat.
- None of the models allow for instances of PetLoverTypeE to have a pet that is neither a Dog nor a Cat.
- Both models allow for instances of PetLoverTypeE to have a pet that is neither a Dog nor a Cat.

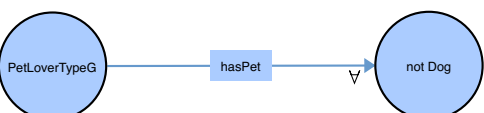
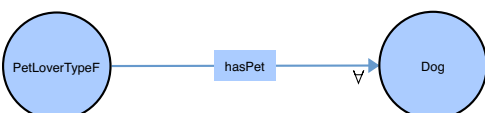
Consider models A and B describing PetLoverTypeD and PerLoverTypeF, each represented in 3 equivalent formalisms (VOWL | Rector | Warren) and answer question 7 below.

Model A: PetLoverTypeD	Model B: PerLoverTypeF
<b>VOWL</b> 	<b>VOWL</b> 
<b>Rector</b> <i>PetLoverTypeD has, amongst other things, some pets that are not Dogs.</i>	<b>Rector</b> <i>PetLoverTypeF has, amongst other things, only Dog pets.</i>
<b>Warren</b> <i>PetLoverTypeD has, amongst other things, at least one pet that is not a Dog.</i>	<b>Warren</b> <i>PetLoverTypeF has, amongst other things, no other than Dog pets.</i>

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

- Yes  
 No

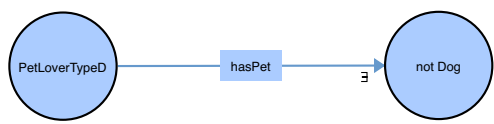
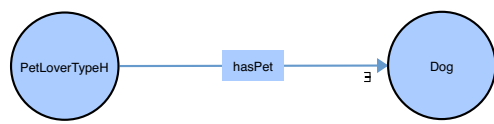
Consider models A and B describing PetLoverTypeG and PerLoverTypeF, each represented in 3 equivalent formalisms (VOWL | Rector | Warren) and answer question 8 below.

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

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

Consider models A and B describing PetLoverTypeD and PerLoverTypeH, each represented in 3 equivalent formalisms (VOWL | Rector | Warren) and answer question 9 below.

Model A: PetLoverTypeD	Model B: PerLoverTypeH
<p><b>VOWL</b></p> 	<p><b>VOWL</b></p> 
<p><b>Rector</b></p> <p><i>PetLoverTypeD has, amongst other things, some pets that are not Dogs.</i></p>	<p><b>Rector</b></p> <p><i>PetLoverTypeH has, amongst other things, some Dog pets.</i></p>
<p><b>Warren</b></p> <p><i>PetLoverTypeD has, amongst other things, at least one pet that is not a Dog.</i></p>	<p><b>Warren</b></p> <p><i>PetLoverTypeH has, amongst other things, at least one Dog pet.</i></p>

9. Is it true that PetLoverTypeD is disjoint to PetLoverTypeH? That is, there can be no instance that is at the same time of type PetLoverTypeD and PetLoverTypeH.

- Yes  
 No

Submit

## Appendix E: Feedback Questionnaire

### Post-Study Questionnaire

Your answers will be treated anonymously. Your name will only be used to connect your answers to your inspection records. No individual information will be made public in any form.

**\*Required**

Name \*

Your answer

Student ID (Matrikelnummer) \*

Your answer

#### Feedback Questions

Q1: Was the tutorial example useful for understanding the difference between the existential & universal restriction? \*

Yes

No

Q2: What could have made the difference between the existential & universal restriction clearer? \*

Your answer



Q3: Was it easy to recognize invalid representations of menu items? \*

- Yes
- No

Q4: Which formalism was easiest for you to understand? \*

- VOWL- Graphical representation
- Rector - Text representation using "some" and "only" as keywords.
- Warren - Text representation using "at least one" and "no other than" as keywords.

Q5: Did you use the provided instructions while performing the judgments? \*

- never
- sometimes
- most of the time
- all the time

**Thank you for your participation!**

Submit

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google. [Report Abuse](#) - [Terms of Service](#) - [Privacy Policy](#)

Google Forms

