Evaluating Knowledge Graphs with Hybrid Intelligence

Stefani Tsaneva*[1,2][0000−0002−0895−6379]

[1] Vienna University of Economics and Business, Austria
[2] TU Wien, Austria
stefani.tsaneva@wu.ac.at

Abstract. Knowledge graphs (KGs) enable the conceptualization of knowledge about the world in a machine-readable format and serve as a foundation to many advanced intelligent applications, such as conversational agents. Ensuring the correctness and quality of KGs is essential for the prevention of invalid application outputs and biased systems, which can result from incorrectly or incompletely represented information. While certain KG quality issues can be automatically detected, others require human involvement, including the identification of incorrectly modeled statements or the discovery of concepts not compliant with how humans think. Human computation and crowdsourcing (HC&C) techniques have been used as a promising method for outsourcing human-centric tasks to human contributors at a reduced cost. Nevertheless, there is no clear guideline on how human-centric KG evaluations should be prepared and scalable evaluation of large KGs utilizing HC&C techniques alone remains a challenge. In this thesis, we investigate a human-centric KG evaluation approach, relying on hybrid (human-AI) intelligence, which leverages techniques from the semantic web, HC&C and multi-agent systems communities for ensuring an efficiently planned, scalable, well-coordinated, and thus transparent KG evaluation process.

Keywords: Semantic Web · Knowledge Graph Evaluation · Human Computation · Crowdsourcing · Hybrid Intelligence

1 Introduction

Knowledge graphs enable the representation of domain-specific and domain-independent information in a machine-readable format and are commonly used as a backbone to many information systems and advanced intelligence applications, which rely on human knowledge [20]. KGs are often curated by extracting information from semi-structured data sources or through crowdsourcing campaigns. Since an automated extraction of knowledge is rarely impeccable, the quality of the resulting KGs should be evaluated. Moreover, KGs are often reused and extended over time, thus it is essential to ensure that they remain up-to-date and accurately reflect evolving knowledge through ongoing maintenance [17].

* Early Stage Ph.D.

Many quality issues related to knowledge graphs can be automatically detected (e.g., logical inconsistencies), while others require a human-centric evaluation. An example is the identification of concepts not compliant with human cognition, inaccurately represented facts and controversial statements modeled from a single perspective [23,4,14]. The traditional approach for addressing such issues relies on domain-expert-evaluations. However, this is a costly and time-intensive process, particularly when dealing with large KGs.

Human computation (HC), a method of outsourcing unautomatable tasks of a system to human participants, can reduce evaluation costs by replacing domain experts with crowd workers. HC is widely adopted for various tasks in the semantic web (SW) research community [24]. A recent systematic mapping study (SMS) [25] showed that 40% of papers, discussing a human-centric evaluation of SW resources, rely on HC&C methods. For example, HC&C techniques were utilized for verifying large biomedical ontologies [15], and evaluating the quality of linked data as a collaborative effort between experts and the crowd [1]. Yet, several issues in the human-centric KG evaluation domain remain:

P1: Lack of methodology and tools. The SMS [25] highlighted that while human-centric KG evaluation has been abundantly addressed there is currently no standardized methodology and tools supporting the knowledge engineers in the preparations of such evaluations. This results in significant amount of manual efforts for the engineers planing an evaluation campaign.

P2: Scalability Issue. Large KGs present a scalability challenge even for crowd-sourced evaluations. In [21] the authors calculated that applying the crowdsourcing assessment, proposed in [1], would require 3,000 years to validate DBpedia, a large KG curated using automated extraction methods.

P3: Lack of transparency. To ensure a transparent KG evaluation process, especially when both human and AI agents perform the evaluation, the question arises of how such an evaluation approach should be coordinated, so that each evaluation can be traced back to its origin across the process.

In this thesis we aim to establish a typical process of human-centric KG evaluation and implement a tool supporting the preparation of such evaluation campaigns (contribution C1, addressing P1). Additionally, we intend to implement a human-centric knowledge graph evaluation system, relying on hybrid (human-AI) intelligence to enable an efficient evaluation of large KGs (C2,P2). To ensure transparency in the KG evaluation process, we further formalize a coordination framework within the hybrid approach (C3,P3).

To address the outlined objectives we employ a design science methodology [7] and adhere to principles from experimental software engineering [31]. We utilize prior research on hybrid intelligence systems proposed for SW tasks (i.e., ontology alignment [27] and entity linking [3]) and multi-agent system approaches aiming at crowd coordination[13,5]. Moreover, we specify two concrete use cases for the evaluation of the established artifacts, namely evaluating the *Computer Science Ontology* and *WebIsALOD*.

We continue by providing introductory definitions and related work in Sect. 2. The problem statement and a discussion of the formulated research questions

follow in Sect. 3. We outline the followed research methodology and the planned evaluation of the proposed approach in Sect. 4 and 5 respectively. Preliminary results are included in Sect. 6 and we conclude with a summary in Sect. 7.

2 Background and Related Work

We start by defining introductory notions in Sect. 2.1, continue with a discussion of related work in the human-centric KG evaluation problem space (Sect. 2.2) and an overview of current approaches in the solution space covering hybrid-intelligence and multi-agent systems (Sect. 2.3).

2.1 Definitions

Knowledge Graph. We refer to the definition of knowledge graphs recently proposed by Hogan et al., which defines a KG as "a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent potentially different relations between these entities"[8]. This definition is broad and as such also encompass other types of semantic resources, such as ontologies and linked data.

Knowledge Graph Evaluation. In this thesis we view knowledge graph evaluation as the refinement of KGs, defined by Paulheim as the improvement of KGs by means of the identification and correction of errors or by completion of missing information [20].

Human-centric Knowledge Graph Evaluation. Combining the definition above with Mortensen's argument that "only domain experts can interpret the symbols in an ontology and determine whether they reflect their understanding of the domain."[16], we define human-centric KG evaluation as the *improvement of knowledge graphs by means of dealing with errors which require human judgment to be identified and corrected and the completion of missing information by leveraging human (domain/general) knowledge.*

2.2 State of the Art in Human-centric Knowledge Graph Evaluation

Evaluating the correctness of KGs has been extensively studied for more than 20 years. McDaniel and Storey [14] reviewed research in ontology assessment from the last 20 years and identified that semantic mistakes cannot (yet) be fully automatically detected. While automatic methods are fast and scalable, they have limitations that require human involvement to be addressed.

Recently, Sabou et al. conducted a systematic mapping study of 100 papers from the last decade (2010-2020) dealing with human-centric evaluation of various semantic resources, corresponding to our definition of KGs [25]. The study showed that human-centric evaluations have been applied in a variety of

domains and verification tasks. For instance, Acosta et al. proposed a find-fix-verify workflow for assessing linked data quality issues, where domain experts identify potential errors and crowd workers verify them [1]. Mortensen et al. presented a crowd-based verification of taxonomic relationships from a medical ontology [15], while in [4] crowdsourcing was used to investigate humans' perception on viewpoints and controversial facts modeled in ontologies. Ontology enhancement achieved by crowdsourcing was investigated in [11] and a validation of enriched ontologies was explored in [9].

Several methods have been proposed for the evaluation of SW resources, aligning with our KG definition, such as *linked data triples quality evaluation* through crowdsourcing and the TripleCheckMate tool [12], a *task-based ontology evaluation* methodology [22], and a *Protégé plugin* that outsources certain tasks of the ontology engineering process to games with a purpose or a crowdsourcing platform [30]. Nevertheless, these methods lack important details and have been established in an ad-hoc manner, rather than using a structured approach.

Despite the abundance of human-centric KG evaluation approaches, there is yet no agreed upon methodology for conducting KG evaluation campaigns and no tool supporting the knowledge engineers preparing them. Additionally, current HC&C approaches have limitations when evaluating large KGs, and evaluations are often not transparent. Thus, we next look into approaches of hybrid human-AI and multi-agent systems addressing similar challenges in related fields.

2.3 Related Work on Hybrid-Intelligence and Multi-Agent Systems

Hybrid Human-AI Workflows for Semantic Web Tasks. Evaluations performed with human involvement achieve high accuracy, but can become costly when verifying large-scale knowledge graphs. Hybrid human-machine workflows, where automated methods are supplemented by human input when the confidence score is low, have been successfully applied for semi-automatic entity-linking [3] and ontology alignment [27] tasks. Nevertheless, such a hybrid solution has not yet been approached for the human-centric evaluation of KGs.

Coordination in Human-AI Collaborations. Workflow coordination known from crowd coordination theory is mainly focused on the self-organization of the crowd, while the coordination of hybrid systems has different requirements. Previous studies [10,2] have identified that methods known from multi-agent systems (MAS) can be utilized to solve crowd coordination challenges.

There has been limited research on how MAS methods can be used to coordinate a hybrid process including both human agents and algorithms. It has been shown that MAS algorithms can support and improve the performance of crowd workers in tasks such as constraint satisfaction problems [13]. Additionally, the combination of crowdsourcing and MAS has been investigated in a sustainable transportation use case, where best route calculations guide delivery drivers [5]. Yet, there has not been an investigation of how MAS can be used to support KG evaluation campaigns and their transparency.

3 Problem Statement and Contributions

This thesis aims at investigating scalable and transparent evaluation of large knowledge graphs. The following research questions are formulated and their connection to specific challenges and contributions are visualized in Fig. 1:

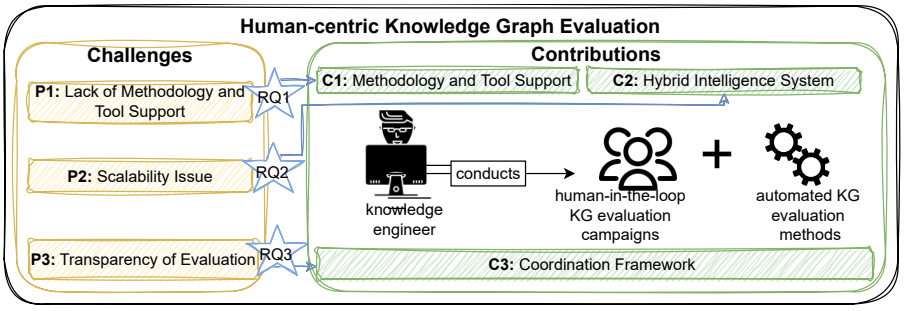

Fig. 1. Challenges in human-centric KG evaluation (P1-P2), formulated research questions (RQ1-RQ3) and expected contributions (C1-C3).

RQ1. *What is a typical process of human-centric knowledge graphs evaluation?*

Currently, the process of managing a (large-scale) KG evaluation campaign involving human participants requires high efforts of the knowledge engineer conducting the evaluation (P1 in Fig. 1) as a result of a lack of clear design guidance (e.g., how to manage the qualification of the participants or how to display the KG segments to the evaluators) and missing methodology (i.e., what part of the evaluation should be designed at which stage). To minimize the organizational efforts of knowledge engineers, clear steps to be followed should be outlined and a tool supporting this methodology should be implemented (C1 in Fig. 1).

RQ2. *How can hybrid intelligence be applied for achieving a scalable human-centric knowledge graph evaluation process?*

Large KGs pose a challenge for current human-centric evaluation approaches (P2 in Fig. 1). This thesis will investigate how the strengths of state-of-the art algorithms and human-in-the-loop approaches can be combined to reduce human efforts and costs of human-centric KG evaluations to ensure a scalable solution. We will explore methods, reducing the tasks assigned to human participants (e.g., graph-based defect candidate detection, link prediction), and requirements (e.g., possible human-AI interaction workflows) for an efficient hybrid intelligence system by looking at concrete KG evaluation use cases. The investigations performed will lead to the implementation of a hybrid intelligence system for human-centric KG evaluation (C2 in Fig. 1).

RQ3. *How can a human-AI knowledge graph evaluation campaign be coordinated to ensure a transparent evaluation process?*

A hybrid human-machine framework requires the coordination of complex workflows between human and machine (algorithmic) agents. A clear transparent process should be followed for delegating the tasks and coordinating them between all agents to ensure the traceability of potential mistakes for the ease of their correction (P3 in Fig. 1). The process should also allow for information of the evaluation to be saved so that in case a re-evaluation is needed, e.g., after a KG modification, only these KG elements are verified that are affected by the implemented changes. Therefore, the result of RQ3 would be a coordination framework for human-centric KG evaluations (C3 in Fig. 1).

4 Research Methodology and Approach

In this thesis, we follow the *design science methodology* for information systems research [7] to establish the following information artifacts: a set of task design guidelines for human-centric KG evaluation tasks (C1); a methodology and tool support for carrying out human-centric KG evaluation campaigns (C1); a hybrid intelligence system for conducting KG evaluation studies (C2); and a coordination framework for hybrid intelligence KG evaluations (C3).

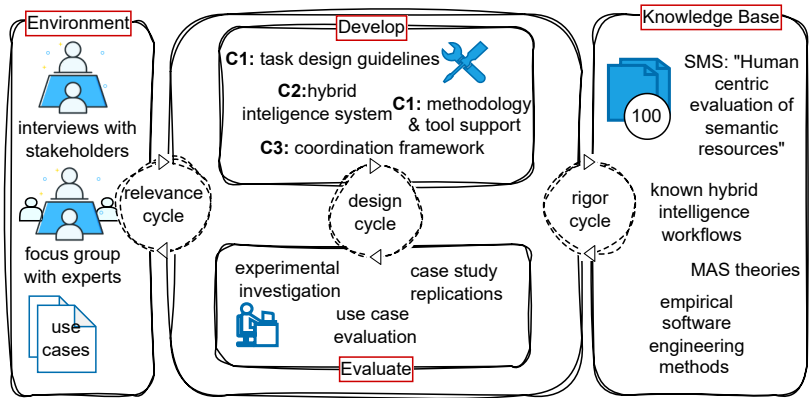

Fig. 2. An overview of the design-science-based methodology followed in this thesis.

Figure 2 visualizes how the relevance, rigor and design cycles are addressed. By involving key stakeholders in need of such artifacts (i.e., knowledge engineers) and focusing on two concrete use cases we address the *relevance cycle* of the design science methodology. We ensure the rigor cycle by incorporating knowledge from existing literature, among others, the results of a large scale SMS [25], developed workflows for hybrid intelligent systems, methods from MAS, and empirical principles of software engineering followed in the planned evaluations.

The artifacts resulting from the investigation of each research question will rely on (several) *evaluation cycles* as described in Sect. 5. We plan two eval-

uation use cases, which allow us to test the hybrid system for e.g., evaluating the correctness of hierarchical relations, in both a domain-specific and a general setting:

Evaluating the Computer Science Ontology (CSO). CSO structures computer science knowledge extracted automatically from 16M publications [18] and enables novel scientometrics tasks such as identifying research communities [19] and forecasting research trends [26]. Its current verification relies on domain experts, who search through the CSO Portal, rate topics and relations as (in)correct, and provide alternative viewpoints, aggregated by an editorial team.

Evaluating WebIsALOD. WebIsALOD is a large KG containing 400M hypernymy relations, automatically extracted from the CommonCrawl web corpus, describing generic knowledge [6]. The current verification of the resource relies on machine learning models trained with a set of 500 relations, validated by crowd workers, to determine confidence scores related to relation correctness.

5 Evaluation Plan

Several information systems artifacts are to be developed in the course of this thesis, each requiring a different evaluation strategy.

Controlled Experiments. The human-centric KG evaluation *task design guidelines*, resulting from RQ1, will be evaluated adhering to the methodology for experimental investigation in software engineering by Wohlin [31], i.e., by hypothesis testing.

Case Studies. The designed *methodology*, developed as part of RQ1, will be evaluated by replicating human-centric KG evaluation approaches, previously performed without following a concrete methodology. Thus, a comparative analysis of the time efforts and re-usability of the process is enabled. The *tool support* developed for the human-centric KG evaluation process will indirectly be evaluated in these replication studies. However, we also plan to conduct *interviews* with domain experts and software architects to further evaluate the tool.

For the evaluation of the *hybrid intelligence system*, designed as part of RQ2, we will apply the system in the concrete use case of the Computer Science Ontology. We plan to organize an evaluation with computer science researchers, where one control group will use the CSO Portal as a baseline and the other(s) the implemented hybrid system. The evaluation will consist in assessing differences in terms of the quantity and range of identified defects and viewpoints as well as the time needed to perform the evaluation.

Lastly, to evaluate the *coordination framework* (RQ3) for the hybrid intelligence system we plan an active-learning evaluation approach of WebIsALOD. We will test different strategies to select tasks to be sent to human agents (e.g., based on confidence scores, outlier detection, etc.). Verified items will be used to update the classification model and the evaluation will consist in iterating through the process until no significant improvements in the classification accuracy are observed when refining the model. Additionally, the transparency of the

framework will be evaluated along several dimensions such as understandability, conciseness, provenance, etc.

6 Preliminary Results

HC&C methods for ontology verification. As an entry point to this thesis, human-centric tasks and their solution was investigated in parallel to the conducted SMS [25]. In [29], we proposed a HC solution for the verification of ontology restrictions by means of universal and existential quantifiers and reported on a controlled experiment to study two core task design aspects: (i) the formalism to represent ontology axioms in the HC task and (2) participant qualification testing. We found that visual axiom representation and prior knowledge of ontology restriction models lead to best results while prior modeling knowledge reduces the evaluation times. In a future publication, we will discuss how the qualification test was set up and propose implementation guidelines that could be used for ontology-related tasks, contributing to answering RQ1.

HERO - a Human-Centric Ontology Evaluation PROcess. RQ1 was also partially addressed in [28] where we formalized a process for conducting human-centric ontology evaluation. In addition, an initial framework was developed to support this process by (semi-)automating a portion of the activities. HERO is a process model for human-centric KG evaluation, targeted toward micro-tasking environments such as crowdsourcing platforms and focusing on batch-style evaluations. At a high-level the process and its activities can be structured into the stages of preparation, execution and follow-up. The process was derived by analyzing steps discussed in literature, semi-structured interviews with experts and an expert focus group discussion. For the evaluation, we replicated our previous manual evaluation approach from [29] with the support of the HERO artifacts and compared the time effort in both approaches. We found that HERO could decrease manual effort up to 88% for the preparation activities involved in human-centric ontology evaluation campaigns. In [28] we focused on describing the process that knowledge engineers follow when conducting human-centric evaluation, while in a future publication the implemented tool will be discussed in detail.

7 Summary

Knowledge graphs are used as a skeleton of many AI applications having high impact on human society. Since automated methods have their limitations, human involvement is a requirement for the evaluation of KGs. Assessing and improving low-quality KGs deals with incorrectly modeled information, and can thus prevent biased and discriminating systems resulting from knowledge graph quality issues. The proposed hybrid intelligence approach would enable human-centric KG evaluation of large-scale KGs, which are a problem for currently available evaluation methods. By providing a transparent evaluation process, bias sources

can easily be identified and corrected, and the overall quality of the KG, and the system using it, can be improved. The outlined work holds the potential to bring novel contributions that can impact not only the SW, HC&C, and MAS communities, but also offer valuable insights for human-AI collaborations across various domains.

Acknowledgments I would like to thank my supervisor, Prof. Marta Sabou, for her valuable advice, expert insights and guidance. The work presented in this paper is partly supported by the FWF HOnEst project (V 754-N).

References

1. Acosta, M., Zaveri, A., Simperl, E., Kontokostas, D., Auer, S., Lehmann, J.: Crowdsourcing linked data quality assessment. In: Int. Semantic Web Conf. pp. 260–276. Springer (2013)
2. Das, R., Vukovic, M.: Emerging theories and models of human computation systems: a brief survey. In: Proc. of the 2nd Int. Workshop on Ubiquitous Crowdsourcing. pp. 1–4 (2011)
3. Demartini, G., Difallah, D.E., Cudré-Mauroux, P.: Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In: Proc. of the 21st Inf. Conf. on World Wide Web. pp. 469–478 (2012)
4. Erez, E.S., Zhitomirsky-Geffet, M., Bar-Ilan, J.: Subjective vs. objective evaluation of ontological statements with crowdsourcing. In: Proc. of the Assoc. for Inf. Science and Technology. pp. 1–4 (2015)
5. Giret, A., Carrascosa, C., Julian, V., Rebollo, M., Botti, V.: A crowdsourcing approach for sustainable last mile delivery. Sustainability **10**(12), 4563 (2018)
6. Hertling, S., Paulheim, H.: Webisalod: providing hypernymy relations extracted from the web as linked open data. In: Int. Semantic Web Conf. pp. 111–119. Springer (2017)
7. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. MIS quarterly pp. 75–105 (2004)
8. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G.d., Gutierrez, C., Kirrane, S., Gayo, J.E.L., Navigli, R., Neumaier, S., et al.: Knowledge graphs. ACM Computing Surveys **54**(4), 1–37 (2021)
9. Iyer, V., Sanagavarapu, L.M., Raghu Reddy, Y.: A framework for syntactic and semantic quality evaluation of ontologies. In: Int. Conf. on Secure Knowledge Management In Artificial Intelligence Era. pp. 73–93. Springer (2021)
10. Jiang, J., An, B., Jiang, Y., Lin, D., Bu, Z., Cao, J., Hao, Z.: Understanding crowdsourcing systems from a multiagent perspective and approach. ACM Transactions on Autonomous and Adaptive Syst. **13**(2), 1–32 (2018)
11. Kiptoo, C.C.: Ontology enhancement using crowdsourcing: a conceptual architecture. Int. J. of Crowd Science (2020)
12. Kontokostas, D., Zaveri, A., Auer, S., Lehmann, J.: Triplecheckmate: A tool for crowdsourcing the quality assessment of linked data. Communications in Computer and Information Science **394**, 265–272 (2013)
13. Mao, A., Parkes, D.C., Procaccia, A.D., Zhang, H.: Human computation and multiagent systems: an algorithmic perspective. In: Proc. of the 25th AAAI Conf. on Artificial Intelligence. Citeseer (2011)

14. McDaniel, M., Storey, V.C.: Evaluating domain ontologies: clarification, classification, and challenges. ACM Comp. Surveys **52**(4), 1–44 (2019)
15. Mortensen, J.M., Minty, E.P., Januszyk, M., Sweeney, T.E., Rector, A.L., Noy, N.F., Musen, M.A.: Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of SNOMED CT. J. of the American Medical Informatics Assoc. **22**(3), 640–648 (2015)
16. Mortensen, J.M., Telis, N., Hughey, J.J., Fan-Minogue, H., Van Auken, K., Dumontier, M., Musen, M.A.: Is the crowd better as an assistant or a replacement in ontology engineering? an exploration through the lens of the gene ontology. J. of Biomedical Informatics **60**, 199–209 (2016)
17. Nishioka, C., Scherp, A.: Analysing the evolution of knowledge graphs for the purpose of change verification. In: IEEE 12th Int. Conf. on Semantic Computing. pp. 25–32. IEEE (2018)
18. Osborne, F., Motta, E.: Klink-2: integrating multiple web sources to generate semantic topic networks. In: Int. Semantic Web Conf. pp. 408–424. Springer (2015)
19. Osborne, F., Scavo, G., Motta, E.: Identifying diachronic topic-based research communities by clustering shared research trajectories. In: Eur. Semantic Web Conf. pp. 114–129. Springer (2014)
20. Paulheim, H.: Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web J. **8**(3), 489–508 (2017)
21. Paulheim, H., Bizer, C.: Improving the quality of linked data using statistical distributions. Int. J. on Semantic Web and Information Systems **10**(2), 63–86 (2014)
22. Pittet, P., Barthélémy, J.: Exploiting users' feedbacks: Towards a task-based evaluation of application ontologies throughout their lifecycle. In: Int. Conf. on Knowledge Engineering and Ontology Development. vol. 2 (2015)
23. Poveda-Villalón, M., Gómez-Pérez, A., Suárez-Figueroa, M.C.: Oops!(ontology pitfall scanner!): An on-line tool for ontology evaluation. Int. J. on Semantic Web and Inf. Syst. **10**(2), 7–34 (2014)
24. Sabou, M., Aroyo, L., Bontcheva, K., Bozzon, A., Qarout, R.K.: Semantic web and human computation: The status of an emerging field. Semantic Web J. **9**(3), 291–302 (2018)
25. Sabou, M., Fernandez, M., Poveda-Villalón, M., Suárez-Figueroa, M.C., Tsaneva, S.: Human-centric evaluation of semantic resources: A systematic mapping study, In preparation
26. Salatino, A.A., Osborne, F., Motta, E.: Augur: forecasting the emergence of new research topics. In: Proc. of the 18th ACM/IEEE on Joint Conf. on Digital Libraries. pp. 303–312 (2018)
27. Sarasua, C., Simperl, E., Noy, N.F.: Crowdmap: Crowdsourcing ontology alignment with microtasks. In: Int. Semantic Web Conf. pp. 525–541. Springer (2012)
28. Tsaneva, S., Käsznar, K., Sabou, M.: Human-centric ontology evaluation: Process and tool support. In: Int. Conf. on Knowledge Engineering and Knowledge Management. pp. 182–197. Springer (2022)
29. Tsaneva, S., Sabou, M.: A human computation approach for ontology restrictions verification. In: AAAI Conf. on Human Computation and Crowdsourcing (2021), www.humancomputation.com/2021/assets/wips_demos/HCOMP_2021_paper_90.pdf
30. Wohlgenannt, G., Sabou, M., Hanika, F.: Crowd-based ontology engineering with the ucomp protégé plugin. Semantic Web **7**(4), 379–398 (2016)
31. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experimentation in software engineering. Springer Science & Business Media (2012)