Developing Crowdsourced Ontology Engineering Tasks: An iterative process

Jonathan M. Mortensen, Mark A. Musen, Natalya F. Noy

Stanford Center for Biomedical Informatics Research
Stanford University, Stanford CA 94305, USA

Abstract. It is increasingly evident that the realization of the Semantic Web will require not only computation, but also *human* contribution. Crowdsourcing is becoming a popular method to inject this human element. Researchers have shown how crowdsourcing can contribute to managing semantic data. One particular area that requires significant human curation is ontology engineering. Verifying large and complex ontologies is a challenging and expensive task. Recently, we have demonstrated that online, crowdsourced workers can assist with ontology verification. Specifically, in our work we sought to answer the following driving questions: (1) Is crowdsourcing ontology verification feasible? (2) What is the optimal formulation of the verification task? (3) How does this crowdsourcing method perform in an application? In this work, we summarize the experiments we developed to answer these questions and the results of each experiment. Through iterative task design, we found that workers could reach an accuracy of 88% when verifying SNOMED CT. We then discuss the practical knowledge we have gained from these experiments. This work shows the potential that crowdsourcing has to offer other ontology engineering tasks and provides a template one might follow when developing such methods.

1 Background

Research communities have begun using crowdsourcing to assist with managing the massive scale of data we have today. Indeed, certain tasks are better solved by humans than by computers. In the life sciences, Zooniverse, a platform wherein citizen scientists contribute to large scale studies, asks users to perform tasks such as classifying millions of galaxies or identifying cancer cells in an image [8]. In related work, Von Ahn and colleagues developed games with a purpose, a type of crowdsourcing where participants play a game, and as a result help complete some meaningful task. For example, in Fold.it, gamers assist with folding a protein, a computationally challenging task [4]. Further demonstrating the power of the crowd, Bernstein et al. developed a system that uses the crowd to quickly and accurately edit documents [1]. With crowdsourcing's popularity rising, many developer resources are now available, such as Amazon's Mechanical Turk, Crowdflower, oDesk, Houdini, etc. Finally, as evidenced by this workshop, CrowdSem, the Semantic Web community is beginning to leverage crowdsourcing. Systems such as CrowdMap, OntoGame, and ZenCrowd demonstrate how crowdsourcing can contribute to the Semantic Web [11, 2, 10]. Crowdsourcing enables the completion of tasks at a massive scale that cannot be done computationally or by a single human.

One area amenable to crowdsourcing is ontology engineering. Ontologies are complex, large, and traditionally require human curation, making their development an ideal candidate task for crowdsourcing. In our previous work, we developed a method for crowdsourcing ontology verification. Specifically, we sought to answer the following driving questions:

(1) Is crowdsourcing ontology verification feasible?
(2) What is the optimal formulation of the verification task?
(3) How does this crowdsourcing method perform in an application?

In this work, we briefly highlight each of the experiments we developed to answer our questions, and, with their results in mind, then discuss how one might approach designing crowdsourcing tasks for the Semantic Web. In previous work, we have published papers that explore each driving question in depth. The main contribution of this work is a unified framework that presents all of the experiments. This framework will enable us to reflect on current work and to ask new questions for crowdsourcing ontology engineering.

2 Ontology Verification Task

We have begun to reduce portions of ontology engineering into microtasks that can be solved through crowdsourcing. We devised a microtask method of ontology verification based on a study by Evermann and Fang [3] wherein participants answer computer-generated questions about ontology axioms . A participant verifies if a sentence about two concepts that are in a parent-child relationship is correct or incorrect. For example, the following question is a hierarchy-verification microtask for an ontology that contains classes Heart and Organ:

Is every Heart an Organ?

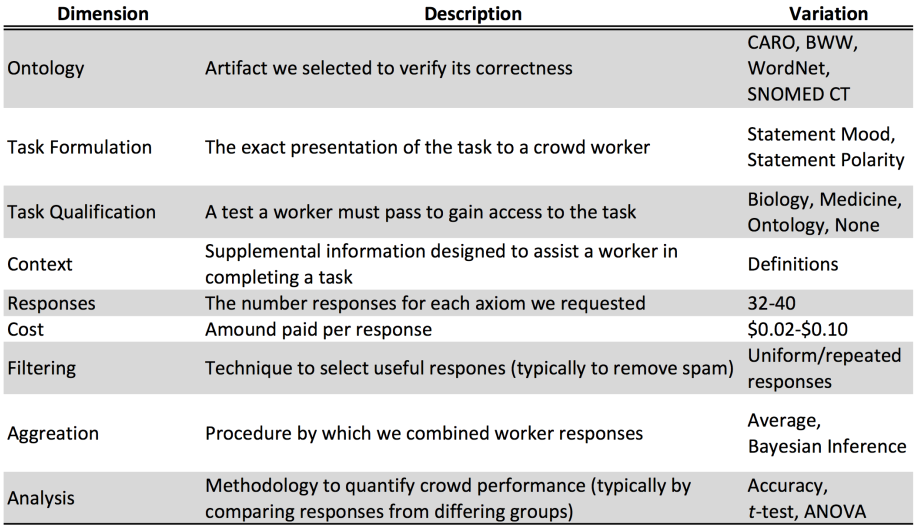
A worker then answers the question with a binary response of “Yes” or “No.”

This task is particularly useful in verifying ontologies because the class hierarchy is the main type of relationship found in many ontologies. For example, in 296 public ontologies in the BioPortal repository, 54% of these ontologies contained only SubClassOf relationships between classes. In 68% of ontologies, the SubClassOf relationships accounted for more than 80% of all relationships. Thus, verifying how well the class hierarchy corresponds to the domain will enable the verification of a large fraction of the relations in ontologies.

3 Protocol & Experimental Design

We developed various experiments that use the hierarchy verification task to answer our driving questions. Generally, each of these experiments follows the same basic procedure. First, we selected the ontology and axioms to verify. Next, we created a hierarchy-verification task formatted as HTML from these entities and submitted the task to Amazon Mechanical Turk. Finally, we obtained worker responses, removed spam, and compared the remaining responses to a gold standard using some analysis metric. Thus, in each experiment we used a standard set of basic components outlined in Table 1. Typically, we manipulated one of these components in each experiment. Figure 1 presents an example task as it appears to a worker on Amazon Mechanical Turk.

Table 1. Dimensions in which our crowdsourcing experiments vary.



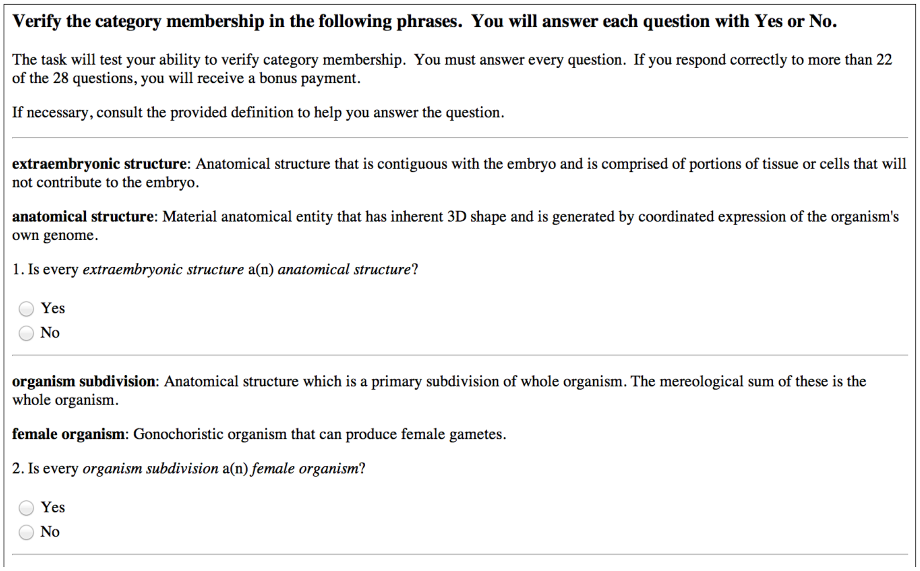
Dimension	Description	Variation
Ontology	Artifact we selected to verify its correctness	CARO, BWW, WordNet, SNOMED CT
Task Formulation	The exact presentation of the task to a crowd worker	Statement Mood, Statement Polarity
Task Qualification	A test a worker must pass to gain access to the task	Biology, Medicine, Ontology, None
Context	Supplemental information designed to assist a worker in completing a task	Definitions
Responses	The number responses for each axiom we requested	32-40
Cost	Amound paid per response	$0.02-$0.10
Filtering	Technique to select useful respones (typically to remove spam)	Uniform/repeated responses
Aggregation	Procedure by which we combined worker responses	Average, Bayesian Inference
Analysis	Methodology to quantify crowd performance (typically by comparing responses from differing groups)	Accuracy, *t*-test, ANOVA

4 Experiments

To answer the driving questions, we performed a series of experiments using the basic protocol. We began with the most basic question about feasibility of the method. Having shown that, we tested various parameters in order to optimize the method. Finally, we used the optimal method in verifying SNOMED CT. Table 2 summarizes these experiments and their parameters. In the following, we describe the specifics of each experiment and our conclusions for each.

4.1 Is crowdsourcing ontology verification feasible? [7]

In this first driving question, we wished to understand if it were possible for Amazon Mechanical Turk workers (turkers) to perform on par with other groups also performing the hierarchy-verification task.



Verify the category membership in the following phrases. You will answer each question with Yes or No.

The task will test your ability to verify category membership. You must answer every question. If you respond correctly to more than 22 of the 28 questions, you will receive a bonus payment.

If necessary, consult the provided definition to help you answer the question.

extraembryonic structure: Anatomical structure that is contiguous with the embryo and is comprised of portions of tissue or cells that will not contribute to the embryo.

anatomical structure: Material anatomical entity that has inherent 3D shape and is generated by coordinated expression of the organism's own genome.

1. Is every *extraembryonic structure* a(n) *anatomical structure*?

○ Yes
○ No

organism subdivision: Anatomical structure which is a primary subdivision of whole organism. The mereological sum of these is the whole organism.

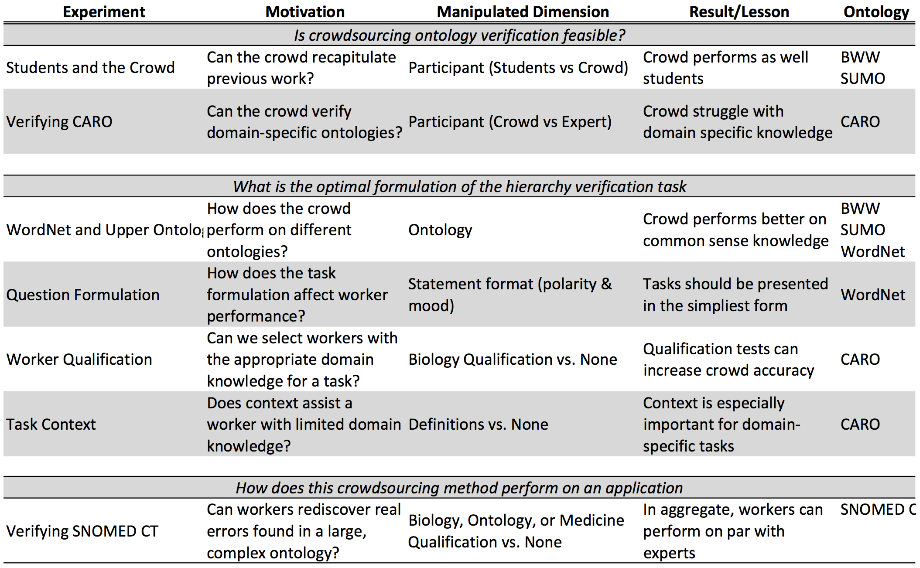
female organism: Gonochoristic organism that can produce female gametes.

2. Is every *organism subdivision* a(n) *female organism*?

○ Yes
○ No

Figure 1. Example task that a Mechanical Turk worker sees in a browser. In this example, workers are provided concept definitions and a Subsumption relation from the Common Anatomy Reference Ontology to verify.

Table 2. Experiments we performed in developing a method to crowdsource ontology verification.



Experiment	**Motivation**	**Manipulated Dimension**	**Result/Lesson**	**Ontology**
Is crowdsourcing ontology verification feasible?				
Students and the Crowd	Can the crowd recapitulate previous work?	Participant (Students vs Crowd)	Crowd performs as well students	BWW SUMO
Verifying CARO	Can the crowd verify domain-specific ontologies?	Participant (Crowd vs Expert)	Crowd struggle with domain specific knowledge	CARO
What is the optimal formulation of the hierarchy verification task				
WordNet and Upper Ontolo	How does the crowd perform on different ontologies?	Ontology	Crowd performs better on common sense knowledge	BWW SUMO WordNet
Question Formulation	How does the task formulation affect worker performance?	Statement format (polarity & mood)	Tasks should be presented in the simpliest form	WordNet
Worker Qualification	Can we select workers with the appropriate domain knowledge for a task?	Biology Qualification vs. None	Qualification tests can increase crowd accuracy	CARO
Task Context	Does context assist a worker with limited domain knowledge?	Definitions vs. None	Context is especially important for domain-specific tasks	CARO
How does this crowdsourcing method perform on an application				
Verifying SNOMED CT	Can workers rediscover real errors found in a large, complex ontology?	Biology, Ontology, or Medicine Qualification vs. None	In aggregate, workers can perform on par with experts	SNOMED C

Experiment 1: Students and the Crowd

Methods We determined whether turkers could recapitulate results from a study by Evermann and Fang [3]. In this study, after completing a training session, 32 students performed the hierarchy-verification task with 28 statements from the Bunge-Wand-Weber ontology (BWW) and 28 statements from Suggested Upper Merged Ontology (SUMO), where half of the statements were true, and half false in each. As an incentive to perform well, students were offered a reward for the best performance.

Knowing the results of that experiment, we asked turkers to verify the same statements. As in the initial study, we required turkers to complete a 12 question training qualification test. We asked for 32 turkers to answer each 28 question set and paid $0.10/set. Furthermore, we offered a bonus for good performance. After turkers completed the tasks, we removed spam responses from workers who responded with more than 23 identical answers. Finally, we compared the performance of the students with that of the turkers using a paired t-test

Results In both experiments, the average accuracy of student was 3–4% higher than the accuracy of the turkers. However, the difference was not statistically significant.

Conclusion Turkers recapitulated previous hierarchy-verification results and performed on par with students in the hierarchy-verification task.

Experiment 2: Verifying the Common Anatomy Reference Ontology (CARO)

Methods Verifying a domain ontology was the second component in showing the feasibility of our verification method. For this verification task, we used CARO, a well curated biomedical ontology. We selected 14 parent-child relations from the ontology as correct relations. Like with WordNet, we paired children with parents that were not in the same hierarchy to simulate incorrect relations. We then asked workers to verify these relations following the earlier experimental setup. In this situation, we had no qualification test. As a comparison, we asked experts on the obo-anatomy and National Center for Biomedical Ontology mailing lists to perform the same verification. Finally, we measured worker and expert performance, and compared the groups using a t-test.

Results With the proper task design of context and qualifications (addressed later), turkers performed 5% less accurately than experts, but there was not a statistically significant difference.

Conclusions Workers performed nearly as as well as experts in verifying a domain ontology. These results are quite encouraging. In addition, the results of this experiment led us to hypothesize that worker performance significantly depends on the task formulation. We address this next.

4.2 What is the optimal formulation of the hierarchy verification task? [5]

With the feasibility of crowdsourcing ontology verification established, we focused on formulating the task in an optimal fashion. There were four main parameters that we hypothesized would affect the method's performance: Ontology Type (i.e., the domain of the ontology being verified), Question Formulation (i.e., How should we ask a worker to verify a relationship?), Worker Qualification (i.e., How does the accuracy of a worker vary based on certain qualification?), and Context (i.e., What information should be provide to assist a worker in answering the question?).

Experiment 3: WordNet and Upper Ontologies

Methods Having shown that turkers perform similarly to students and domain experts, we then analyzed how turker performance varied based on ontology selection. To do so, we compared worker performance in verifying BWW and SUMO to verifying WordNet. We created a set of WordNet statements to verify by extracting parent-child relationships in WordNet and also generating incorrect statements from incorrectly paired concepts (i.e. pairing concepts in parent-child relationships that are not actually hierarchically related). We then asked workers to verify the 28 WordNet, SUMO and BWW statements following the same setup as the first experiment (including the training qualification), paying workers $0.10/set, giving a bonus, and removing spam.

Results Echoing the first experiment, workers performed only slightly better than random on BWW and SUMO, respectively. However, workers had an average accuracy of 89% verifying WordNet statements. There was a clear difference between worker performance on upper ontologies and WordNet.

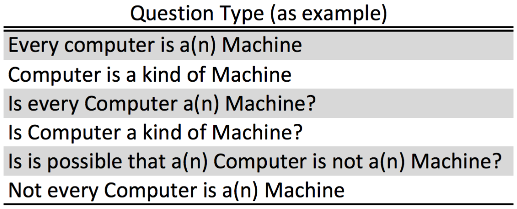
Conclusion While workers struggle with verifying conceptually difficult relationships, such as those contained in upper level ontologies, they perform reasonably well in tasks related to common-sense knowledge.

Experiment 4: Question Formulation

Methods We repeated the task of verifying 28 WordNet statements but varied the polarity and mood of the verification question we ask the workers. In this case, we did not require qualifications as with the earlier experiments. Table 3 shows the 6 different question styles through example.

Results Worker performance varied from 77% on negatively phrased statements to 91% with the positive, indicative mood (i.e., a True/False statement asserting the relationship). In addition, workers responded faster with positively phrased questions.

Table 3. Example question types we presented to users on Mechanical Turk.



Question Type (as example)
Every computer is a(n) Machine
Computer is a kind of Machine
Is every Computer a(n) Machine?
Is Computer a kind of Machine?
Is is possible that a(n) Computer is not a(n) Machine?
Not every Computer is a(n) Machine

Conclusion Generally for crowdsourcing, one should create tasks in the most cognitively simple format as possible. In this situation, asking the verification as simply as possible (i.e., Dog is a kind of Mammal. True or False?)

Experiment 5: Worker Qualification

Methods Having determined the optimal method to ask the verification question, we theorized that workers who could pass a domain-specific qualification test would perform better than a random worker on tasks related to that domain. We developed a 12 question high-school level biology qualification test. For turkers to access our tasks, they would have to pass this test. We assume that the ability to pass this test serves as a reasonable predictor of biology domain knowledge. We asked workers to complete the CARO verification (Experiment 3), but required them to first pass the qualification task, answering at least 50% of it correctly.

Results With qualifications, turkers improved their accuracy to 67% (from random without qualifications) when verifying CARO.

Conclusion When crowdsourcing, some method to select experts in the domain of the task is necessary to achieve reasonable performance. However, such low accuracy was not satisfying to the authors.

Experiment 6: Task Context

Methods With the increases in performance with proper question formulation and qualification requirements, we next proposed that concept definitions would assist workers in verifying a relation. In this experiment, we used CARO because the ontology has a complete set of definitions. We repeated Experiment 3, with qualifications and simply stated verification questions, varying whether users and experts were shown definitions.

Results With definitions, workers performed with an average accuracy of 82%. Experts performed with an average accuracy of 86%. So, when providing workers and experts with definitions, there was no statistically significant difference.

Conclusion In crowdsourcing, context is essential, especially for non-domain experts. While workers might not have very specific domain knowledge, with proper context or training, they can complete the task. This experiment revealed that in some situations, a properly designed microtask can indeed provide results on par with experts.

4.3 How does this crowdsourcing method perform on an application? [6]

The previous experiments were all synthetic – turkers only found errors that we introduced. With the optimal task formulation in hand, we shifted our focus to a true ontology verification task of verifying a portion of SNOMED CT. We selected SNOMED CT because it is a heavily studied, large and complex ontology, making it an ideal candidate for our work.

Experiment 7: Verifying SNOMED CT

Methods In 2011, Alan Rector and colleagues identified entailed SubClass axioms that were in error [9]. In our final experiment, we evaluated whether our method could recapitulate their findings. To do so, we asked workers to perform the hierarchy verification task on these 7 relations along with 7 related relations we already knew were correct. We used the optimal task formulation we determined in earlier experiments and provided definitions from the Unified Medical Language System. In addition, we posted the task with 4 different qualification tests: biology, medicine, ontology, and none. To note, instead of asking workers to complete the task of verifying all 14 relations in one go, as with earlier experiments, we instead broke up the task into smaller units, creating one task per axiom and paid unqualified workers and qualified workers $0.02 and $0.03 per verification, respectively. We then compared worker's average performance to their aggregate performance (i.e., when we combined all workers responses to one final response through majority voting [6]).

Results The aggregate worker response was 88% accurate in differentiating correct versus incorrect SNOMED CT relations. On average, any single worker performed 17% less accurately than the aggregate response. Furthermore, there was no significant difference in performance for tasks with differing qualification tests.

Conclusion Individually, workers did not perform well in identifying errors in SNOMED CT. However, as a group, they perform quite well. The stark difference between average worker performance and aggregate performance reinforces the fact that the power of the crowd lies in their combined response, not any worker alone.

5 Discussion

Each of the experiments we performed highlighted various lessons we learned in developing a method for crowdsourcing ontology verification. A few lessons are particularly useful for the Semantic Web community. First, many of our experiments focused on changing small components of the task. Even so, through this process we greatly improved crowd worker performance. It is clear that each task will be unique, but in most cases, extensive controlled trials will assist in identifying the best way to crowdsource a task. Following this strategy, we verified a complex ontology with relatively high accuracy. In addition, our current results only serve as a baseline – through additional iteration, we expect the increases in accuracy to continue.

Second, using the refined tasks, we showed that crowd workers, in aggregate, can perform on par with experts on domain specific tasks when provided with simple tasks and the proper context. The addition of context was the single biggest factor at improving performance. In the Semantic Web, a trove of structured data are available, all of which may provide such needed context (and maybe other elements, such as qualification tests). For example, when using the crowd to classify instance-level data, the class hierarchy, definitions, or other instance examples may all assist the crowd in their task.

Our results suggest that crowdsourcing might serve as method to improve other ontology engineering tasks such as typing instances, adding definitions, creating ontology mappings and even ontology development itself. In fact, Sarasua and colleagues used crowdsourcing to improve automated ontology mapping methods [10]. ZenCrowd follows a similar paradigm, using crowdsourcing to improve machine extracted links [2]. Indeed, crowdsourcing can serve as a human curated step in ontology engineering that acts in concert with automated methods (e.g., terminology induction supplemented with the crowd).

5.1 Future Work

The results thus far serve only as a baseline for crowdsourcing an ontology engineering task. We plan to focus research on other elements in the crowdsourcing pipeline, include entity selection (e.g., selecting the axioms for verification that will most likely be in error), generating context (e.g., how can we use the crowd to also supply context for workers downstream), and optimizing performance (e.g., developing aggregation strategies that maximize worker performance while minimizing task cost). We will also consider different incentives models including reputation or altruism, like the successful Zooniverse platform [8]. Finally, we will investigate how to integrate this method into a true ontology engineering workflow with the Protege ontology development platform.

6 Conclusion

Crowdsourcing is now another tool for the Semantic Web researcher and developer. In this work, we described various experiments we performed to refine a methodology to crowdsource ontology verification. In summary, we arrived at a highly accurate method through iterative, controlled development of the crowdsourcing task. In doing so, we gained valuable knowledge about method design for crowdsourcing. For example, providing task context is key to enabling accurate crowd workers. Finally, our results suggest that crowdsourcing can indeed contribute to ontology engineering.

Acknowledgements

This work is supported in part by Grant GM086587 from the National Institute of General Medical Sciences and by The National Center for Biomedical Ontology, supported by grant HG004028 from the National Human Genome Research Institute and the National Institutes of Health Common Fund. JMM is supported by National Library of Medicine Informatics Training Grant LM007033.

References

1. Bernstein, M., Little, G., Miller, R., Hartmann, B., Ackerman, M., Karger, D., Crowell, D., Panovich, K.: Soylent: a word processor with a crowd inside. In: The 23d annual ACM symposium on user interface software and technology. pp. 313–322. ACM (2010)
2. Demartini, G., Difallah, D.E., Cudré-Mauroux, P.: ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In: 21st World Wide Web Conference WWW2012. pp. 469–478. Lyon, France (2012)
3. Evermann, J., Fang, J.: Evaluating ontologies: Towards a cognitive measure of quality. Information Systems 35, 391-403 (2010)
4. Khatib, F., DiMaio, F., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., Zabranska, H., Pichova, I., Thompson, J., Popović, Z., Jaskolski, M., Baker, D.: Crystal structure of a monomeric retroviral protease solved by protein folding game players. Nat Struct Mol Biol 18(10), 1175–1177 (10 2011)
5. Mortensen, J.M., Alexander, P.R., Musen, M.A., Noy, N.F.: Crowdsourcing Ontology Verification. In: International Conference on Biomedical Ontologies (2013).
6. Mortensen, J.M., Musen, M.A., Noy, N.F.: Crowdsourcing the Verification of Relationships in Biomedical Ontologies. In: AMIA Annual Symposium. Accepted (2013)
7. Noy, N.F., Mortensen, J.M., Alexander, P.R., Musen, M.A.: Mechanical Turk as an Ontology Engineer? Using Microtasks as a Component of an Ontology Engineering Workflow. In: Web Science (2013)
8. Raddick, M.J., Bracey, G., Gay, P.L., Lintott, C.J., Cardamone, C., Murray, P., Schawinski, K., Szalay, A.S., Vandenberg, J.: Galaxy Zoo: Motivations of Citizen Scientists p. 41 (Mar 2013
9. Rector, A.L., Brandt, S., Schneider, T.: Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications. Journal of the American Medical Informatics Association 18(4), 432–440 (Apr 2011)
10. Sarasua, C., Simperl, E., Noy, N.F.: CrowdMAP: Crowdsourcing Ontology Alignment with Microtasks. In: 11th International Semantic Web Conference (ISWC). Springer, Boston, MA (2012)
11. Siorpaes, K., Hepp, M.: Games with a Purpose for the Semantic Web. IEEE Intelligent Systems 23(3), 50–60 (May 2008)