Crowdsourcing Ontology Verification

Jonathan M. Mortensen*, Paul R. Alexander, Mark A. Musen, and Natalya F. Noy

Stanford Center for Biomedical Informatics Research Stanford University, Stanford, CA 94305-5479 USA

ABSTRACT

Biomedical ontologies are becoming increasingly large and complex. A single user cannot easily develop or maintain them. Researchers have developed various automated techniques to assist with ontology development and engineering at scale. However, these solutions are not always complete. Microtask crowdsourcing, wherein workers are paid small amounts to complete simple, short tasks, may be one technique to alleviate some of the development difficulties. Previously, we developed a method to verify an ontology hierarchy using microtask crowdsourcing. In this work, we investigated the finer details of the design and configuration of a hierarchyverification task. For example, when we provided definitions and required gualifications, workers performed with 82% accuracy on the hierarchy-verification task, compared to 50% without. We showed that to achieve reasonable performance on such a task, workers require context via definitions, tasks require qualifications that select a worker with proper domain knowledge, and a question must be phrased with the least cognitive load (i.e., in the simplest way).

1 INTRODUCTION

Ontology engineering is a labor-intensive and knowledge-intensive task. As the size, number, and complexity of ontologies grow, ontology engineering and maintenance becomes increasingly difficult. Large biomedical ontologies containing tens of thousands of classes, such as the Gene Ontology (GO) (GOConsortium, 2001), are possible only through collaborative development. BioPortal—a repository of ontologies in biomedical domain—has more than 320 entries at the time of this writing (Whetzel *et al.*, 2011).

Researchers have worked to develop methods that automatically perform various ontology-engineering tasks, including evaluating ontology quality, performing alignment, and generating new ontologies. For instance, the best automatic ontology alignment tools now have precision as high as 78–90%, depending on the task (Euzenat *et al.*, 2011). Researchers in ontology learning from text have succeeded in extracting ontology terms with 97% precision (Kozareva and Hovy, 2010). However, in many cases, a fully automatic solution is not feasible. For example, the subsumption hierarchy that the ontology-learning tools induce is limited and the recall is low, with the best tools achieving 40% recall at 95% precision. Similarly, verifying that an ontology corresponds to an experts model of the scientific domain is a time-consuming task that requires checking every relationship in the ontology (Evermann and Fang, 2010).

One avenue to overcome the difficulties of developing large, complex ontologies is crowdsourcing, or human computation. In this model, humans perform small tasks to help solve challenging problems. Incentives can range from small payments to public recognition and social reputation to the desire to help scientific progress (Raddick et al., 2009). There are various platforms, such as Amazon's Mechanical Turk, that enable a specific form of human computation-microtask crowdsourcing. When scientists deploy microtask crowdsourcing, they break their problem into small tasks, each of which takes from a few seconds to a few minutes to complete. They post these tasks in a virtual market, and workers on this platform perform the tasks and collect small payments. These tasks can involve, for example, identifying components in an image, answering questions about a web page and finding errors in a text. Researchers have shown that, if the tasks are designed correctly, the workers can be very efficient in evaluating user interfaces, finding grammatical errors in text, and re-writing text (Bernstein et al., 2010). Recently, researchers have developed special-purpose platforms such as Zooniverse (Raddick et al., 2009) to involve citizen scientists in solving complex scientific problems and fold.it (Cooper et al., 2010) to allow anybody to help predict structures of proteins by playing games.

In previous work, we developed methods to crowdsource ontology alignment and evaluation. First, we devised a workflow for using microtasking as part of ontology alignment (Sarasua et al., 2012). We generated candidate mappings with AROMA, an automated ontology mapping tool that performed well in the 2011 Ontology Alignment Evaluation Initiative (OAEI) (David et al., 2007). These candidate mappings initially had 35% precision and 46% recall. We then asked crowdsourced workers to verify these candidates through microtasking. After the workers completed the microtasks, precision increased to 75%, with none of the correct mappings eliminated (i.e., no loss in recall). We also performed preliminary studies to evaluate the feasibility of crowdsourcing in verifying the subsumption structure of ontologies (the hierarchyverification task) (Noy et al., 2013). Workers achieved 89% accuracy when verifying the hierarchy in WordNet and 81% when verifying CARO, an anatomy ontology, thus performing similarly to experts. These initial experiments show that workers in the crowd can help us develop and maintain ontologies, even in specialized domains.

We can present such tasks to workers in many forms. In fact, Kittur and colleagues describe a significant change in the performance of Mechanical Turk workers with only minor modifications in experimental configuration (Kittur *et al.*, 2008). In psychology studies, for example, researchers consistently demonstrated that participant selection, priming, question tone, and context all affect performance and reliability (Schwarz, 1999; Tanur, 1992). Thus, in this work, we focus on understanding how best to configure the microtasks in order to improve the workers performance in verifying biomedical ontologies. Specifically, we examine the effect of different ways of asking questions, providing context about ontology classes, and ways to select the workers who are qualified to answer questions about biomedical ontologies.

^{*}To whom correspondence should be addressed: jmort@stanford.edu

These studies will help us operationalize the use of microtask crowdsourcing as a viable component in the workflow of developing biomedical ontologies.

2 BACKGROUND

In this section, we provide background on microtask crowdsourcing and its recent use for managing structured data. We then define the specific task—the hierarchy-verification task—on which we focus in this paper.

2.1 Microtask Crowdsourcing

In a microtasking platform, the requesters, who need to have a certain task performed, divide the task into microtasks, with each microtask usually requiring a few seconds to a few minutes to complete. The requesters publish the microtasks in an online marketplace, such as Mechanical Turk. The workers on the platform find the tasks that they want to perform, and get paid to do the work. When publishing a microtask, a requester specifies a number of configuration parameters, such as the number of answers that she needs for each microtask, the time to complete the microtask, and any restrictions on the profile of the workers (e.g., geographical location, knowledge of a specific natural language). In addition, the requester can ask the workers to take a *qualification test* in order to gain access to her tasks. Only workers who pass the test by answering the percentage of questions that the requester specifies, can access the tasks.

Upon completion of the tasks by the workers, the requester collects and assesses the responses, and rewards the accepted responses according to a pre-defined remuneration scheme. For most platforms, the requester can automate the interaction with the system via an API, while the workers undertake their tasks using a Web-based interface generated by the requester. The overall effectiveness of crowdsourcing can be influenced dramatically by the way in which a requester packages a given problem as a series of microtasks (Kittur et al., 2008; Franklin et al., 2011). Because multiple workers can perform the same microtask, the requester can implement different methods for pooling the results (Ipeirotis et al., 2010). For example, the requester can use majority voting (take the solution on which the majority of workers agree) or more sophisticated techniques that take into account such factors as the (estimated) expertise of specific workers, or the probabilistic distribution of accuracy of the answers of a given worker.

2.2 Managing Structured Data

Researchers have successfully used human computation in managing structured data (Quinn and Bederson, 2011). For example, they have used so-called "games with a purpose" for tasks ranging from image tagging (von Ahn and Dabbish, 2004) to ontology alignment (Thaler *et al.*, 2011) and identity resolution (Markotschi and Völker, 2010). In management of structured and linked data, ZenCrowd, for example, combines the results of automatically generated answers with the answers by workers in order to link entities recognized in a text with entities in the Linked Open Data cloud (Demartini *et al.*, 2012). Simperl and colleagues discuss the use of crowdsourcing for querying semantic data (Simperl *et al.*, 2011). Sarasua and colleagues studied the use of microtask crowdsourcing to improve ontology alignment (Sarasua *et al.*, 2012). In CrowdDB (Franklin *et al.*, 2011), workers fill out

missing information in a database table that is needed to answer a query.

2.3 Hierarchy Verification Task

In most biomedical ontologies, the class hierarchy not only constitutes the backbone of the structure, but also is the only semantic relationship between classes that ontology developers have defined. For example, we analyzed 296 public ontologies in the BioPortal repository, which had at least one relation between classes defined. In 54% of these ontologies, the subclass–superclass relationship was the *only* relationship between classes. In 68% of ontologies, the subclass–superclass relationships constituted more than 80% of all relationships.

Thus, verifying how well the class hierarchy corresponds to the domain will account for verification of a large fraction of the relationship in biomedical ontologies. We have previously explored applying crowdsourcing to the hierarchy-verification task. Based on a study developed by Evermann and Fang (2010), we created a crowdsourcing method of ontology verification, wherein workers answer computer-generated questions based on ontology axioms. 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."

However, we can ask the same question in many different ways, decide whether or not to provide context (such as the class definition), or show only the class label. Furthermore, because verifying statements from biomedical ontologies requires at least some domain knowledge, we can also control who gets access to our tasks. In this work, we study the effect of these parameters on the accuracy of the workers' performance.

3 METHODS

In order to determine the effect of various task configurations, we generated a hierarchy-verification task, like the above example, in various configurations. We then compared worker performance for each configuration.

In summary, we performed three experiments varying the following configuration elements to quantify their effect on worker performance:

- 1. Question Formulation What is the best way to ask a hierarchy verification question? (Section 3.2)
- 2. Context How can additional information improve worker performance? (Section 3.3)
- 3. Qualification Tests How can we select the appropriate users for the ontology domain? (Section 3.4)

3.1 Base Protocol

In each experiment that follows, we used this base protocol.

3.1.1 Ontology As a representative biomedical ontology, we used the most recent version of CARO (ver. 12/14/2011) that we obtained from BioPortal.

3.1.2 Verifying the Hierarchy To create the verification questions like the example above, we extracted pairs of classes that had a subclassOf (parent-child) relationship defined between

them. CARO has 49 classes and 48 directly asserted subclassOf relationship. We chose pairs of classes for the microtasks randomly from these 48 pairs. To generate pairs of classes that were not actually in a subclass–superclass relationship, we randomly chose pairs of classes from CARO and verified that they were not in those 48. While these negative relationships are not explicitly stated in the ontology, they are non-sensical and very unlikely to be true.

Table 1 shows the complete set of pairs. Because CARO is heavily curated, we used the ontology itself as the gold standard, assuming that all the pairs of classes that it asserts to be related were indeed related.

3.1.3 Creating a Verification Task We generated 28 questions from the CARO concept pairs we extracted in the previous step. Figure 1 shows an example task.

3.1.4 Cost After creating the questions, we submitted the task to Amazon Mechanical Turk. We paid 0.10 to a worker when they completed each task composed of 28 hierarchy-verification questions. If a worker performed extremely well (with at least 75% accuracy), we gave them a 0.10 bonus. We advertised this bonus with the task, as researchers have shown that a potential bonus increases the quality of the responses (Wang *et al.*, 2012).

3.1.5 Number of Responses Our goal was to collect 32 qualified responses in order to achieve statistical validity for the test. We continued requesting responses until we had 32 non-spam responses (see Section 3.1.6)

3.1.6 Spam Crowdsourced workers have an incentive to do the least amount of work possible so they can get paid for the largest number of tasks. Thus, many users are likely to give spam responses. First, we require that users answer all questions. By doing so, the effort to select any answer is almost the same as truly answering a question. Second, we disqualified all the workers who had more than 23 identical answers out of 28 (i.e., selecting TRUE or FALSE more than 23 times), removing their responses from the analysis. This filtering step allowed us to remove the workers who appeared to have performed the task by purposefully selecting the same response for every verification question.

3.1.7 Analysis After workers completed a task, we downloaded and analyzed their responses. For each task configuration, we measured the accuracy of each worker using the reference CARO pairs. We compared worker accuracy between different task configurations using the student's *t*-test, which compares the performance distribution of two groups, or ANOVA, which allows for comparison between more than two groups. Because there are many configurations, we used Bonferroni correction with the statistical tests.

3.2 Verification Question Formulation

To determine the effect of question formulation, we varied the question grammatical polarity as either positive or negative and the mood as either interrogative (YES or NO) or indicative (TRUE or FALSE). Table 2 presents the question forms. In this experiment, we

	<u></u>			
Ontology	: CARO			
Child	Parent			
TRUE statements				
extraembryonic structure	anatomical structure			
simple cuboidal epithelium	unilaminar epithelium			
portion of tissue	anatomical structure			
anatomical structure	material anatomical entity			
multi-cell-component structure	anatomical structure			
unilaminar epithelium	epithelium			
hermaphroditic organism	multi-cellular organism			
protandrous hermaphroditic organism	sequential hermaphroditic organism			
sequential hermaphroditic organism	hermaphroditic organism			
anatomical point	immaterial anatomical entity			
anatomical line	immaterial anatomical entity			
acellular anatomical structure	anatomical structure			
compound organ component	multi-tissue structure			
male organism	gonochoristic organism			
FALSE st	atements			
organism subdivision	female organism			
asexual organism	multi-cell-component structure			
portion of tissue	anatomical space			
acellular anatomical structure	simple organ			
single cell organism	epithelial cell			
compound organ	cell component			
male organism	acellular anatomical structure			
female organism	cavitated compound organ			
portion of cell substance	simple columnar epithelium			
basal lamina	anatomical surface			
anatomical cluster	sequential hermaphroditic organism			
anatomical point	material anatomical entity			
neuron projection bundle segment	solid compound organ			
extraembryonic structure	hermaphroditic organism			

 Table 1. The pairs of term for the sentence verification tasks for CARO with definitions. The table shows the data both for TRUE and FALSE statements.

Verify the cate	gory membership in the following phrases. You will answer each question with Yes or No.
	your ability to verify category membership. You must answer every question. If you respond correctly to more than 22 ns, you will receive a bonus payment.
If necessary, con-	sult the provided definition to help you answer the question.
extraembryonic not contribute to	structure: Anatomical structure that is contiguous with the embryo and is comprised of portions of tissue or cells that will the embryo.
anatomical strue own genome.	cture: Material anatomical entity that has inherent 3D shape and is generated by coordinated expression of the organism's
1. Is every extrae	mbryonic structure a(n) anatomical structure?
O Yes	
🔘 No	
organism subdiv whole organism.	ision: Anatomical structure which is a primary subdivision of whole organism. The mereological sum of these is the
female organism	a: Gonochoristic organism that can produce female gametes.
2. Is every organ	ism subdivision a(n) female organism?
O Yes	
No	

Fig. 1. A CARO hierarchy-verification task with definitions that workers are paid to complete.

use parent–child relationships from WordNet,¹ following a similar procedure to generate the concept pairs as we did for CARO. Finally,

¹ Due to time constraints, we have used WordNet concepts instead of CARO concepts. We believe that the results from using WordNet concepts are useful and are generally comparable to CARO concepts.

we used ANOVA to determine if the difference in worker accuracy and worker completion time is significant.

3.3 Adding Context with Concept Definitions

To determine the effect of adding context to a task, we retrieved concept definitions and added them with the question text. We then compared worker accuracy on tasks with and without definitions. Specifically, we extracted these definitions from the "def" annotation in CARO. For each concept in a question, we provide the definition above the verification question. Figure 1 shows a task with definitions added. We then follow the base protocol with the added definitions to the task.

3.4 Qualifications to Select Knowledgeable Workers

To select workers who would be more qualified to answer questions based on biomedical ontologies and to improve the quality of the answers, we use qualification tests asking workers to respond to a number of domain-specific questions before they can work on our tasks. For this configuration, we developed a 12-question qualification test based on high-school biology questions. Figure 2 shows the beginning of this test.

	correct. he task will test your knowledge of biology.				
Which of the fo	llowing best describes the result of a mutation in an organism's DNA?				
The mutation n	nay produce a zygote.				
The mutation n	nay cause phenotypic change.				
The mutation c	auses damage when it occurs.				
The mutation c	reates entirely new organisms				
to digest protei	mary function of the large intestine?				
to absorb nutri	ents				
	complex carbohydrates				
to break down					
0	er from undigested waste				
to remove wate	-				
to remove wate	r from undigested weste				

Fig. 2. A biology qualification test that users must complete to gain access to a hierarchy-verification task.

4 RESULTS

In total, we had 320 qualified responses. We paid the workers \$32.00 dollars for the tasks and \$20.30 in bonuses.

4.1 Question Formulation

First, we compared worker performance in terms of time and accuracy when they were given different question formulations. Table 2 presents questions and worker performance . These results show that questions formulated with a positive polarity in the indicative mood elicit the best results in terms of accuracy (91%, p<0.005). However, the positive, interrogative form leads to lowest response times (114.8s, p<0.005). The results show that negative questions do not elicit high quality or quick responses.

Example question	Polarity	Mood	Time* (sec)	Accuracy*
Is Computer a kind of Machine?	+	Q	114.8	0.87
Is every Computer a Machine?	+	Q	123.7	0.87
Computer is a kind of Machine	+	S	121.1	0.91
Every Computer is a Machine	+	S	129.1	0.91
Is is possible that a Computer is not a Machine?	-	Q	164.3	0.82
Not every Computer is a Machine	-	S	138.8	0.77

Q=Interrogative mood, S=indicative mood

 Table 2. Average accuracy of workers on the same WordNet pairings with verification task posed in different question forms.

Avg. Accuracy of Turkers vs Experts					
	No Qualifications	Qualifications			
No concept definitions	0.494	0.670			
With concept definitions	0.640	0.818			
	6, 1 1	1.1.4. 1.1			

 Table 3. Average accuracy of turkers when validating hierarchical relations in CARO, an anatomy ontology. All pairs significant at p<0.008 via student's *t*-test, except the task with qualification and no definitions compared with the task with no qualification and definitions

4.2 Concept Definitions & Task Qualifications

Table 3 presents results for two other experiments. The table compares the average performance of workers with and without qualification test and with and without the context (i.e., definition of the class). Clearly, the workers who achieved the highest performance (82% accuracy) were the workers who had to pass the qualification test and who were subsequently given the definitions of the classes and not only their labels. Without qualification questions and even with the context, the workers performance is only slightly better than guessing.

5 DISCUSSION

This experiment highlights the effect of question formulation, concept definitions, and qualification tests on the performance of Mechanical Turk workers when performing hierarchy verification. We showed that the best performing workers had tasks with questions formulated in the most basic form, a domain-specific qualification, and concept definitions for context.

5.1 Question Formulation

We observed a significant difference in accuracy and speed of responses based on the formulation of the questions. While we focused only on hierarchy-verification questions, our results indicate that, when we apply similar methods to other types of ontology-management tasks, we will need to account for the possible performance differences due to question formulation. We introduced the negative questions in an attempt to direct the workers to think of exceptions to a hierarchical relation. However, it appears that the additional cognitive load of a complex question reduced performance. In fact, this result corroborates similar conclusions from psychology research (Clark and Chase, 1972; Just and Carpenter, 1971). In addition, we found that when verifying true relationships, workers performed better with positively formed questions and with false relationships they performed better with negatively formed questions. This finding makes intuitive sense, but is not useful in practice because a system creating tasks would not know ground truth. Our results indicate that it is worthwhile to perform pilot experiments for new types of tasks to determine the most optimal configuration. Furthermore, careful consideration of study design like that of psychological test questionnaires is necessary when designing microtasks. In our future research, we plan to explore whether our finding that positive-polarity and indicative-mood questions produce the best result can be generalized to other smiler tasks.

5.2 Concept Definitions

As one might expect, adding concept definitions improved worker performance. Workers who pass qualification tests likely have the general knowledge but may not know a particular concept from the ontology. These definitions provide the necessary context. Generally, these results indicate that workers need context to perform an ontology management task effectively. Therefore, when designing a task, one must consider how much context to present a worker without overwhelming her. In the future, we will consider adding additional context clues, such as hierarchy, synonyms, and references. To note, a definition may directly describe the relationship between two concepts (e.g. "A skin cell is a cell"). We will also investigate strategies to correctly verify an ontology via indirect hierarchy verification so that definitions do not directly provide answers a hierarchy verification task.

5.3 Task Qualifications

We envisioned that qualification questions can serve three purposes. First, they help filter out spammers, by requiring workers to answer "free" questions before gaining access to the questions for which they will be paid. Second, they determine whether a worker has the necessary general knowledge to answer the hierarchy-verification questions in a specific domain. Our results demonstrate that qualification questions do indeed improve the quality of responses. However, whether or not we had qualification questions had a dramatic effect on the time that it took us to collect the responses. In all cases, not having qualification questions produced the required number of responses in a matter of minutes. With qualification questions, regardless of whether they were simple or not, we had to wait between 3-4 days to a week or two before we obtained the required number of responses. However, ontology verification, and ontology evaluation in general, is usually not a time-sensitive task. Thus, in many cases, ontology developers might be able to request that a module in their ontology gets verified by the crowd and come back several days later to see the results.

5.4 Cost

A potential barrier to crowdsourcing ontology management is cost. We have found that the cost is, in fact, reasonable. Based on our current method, we pay workers between \$0.003 and \$0.004 per verification question (or subClassOf axiom). An ontology such as SNOMED CT contains approximately 600,000 subClassOf relations. Thus, verifying such a large ontology would cost \$2,500 for each question to be verified once. To reach confidence about responses, each axiom should be verified multiple times. Even so,

the task cost is much lower than that of a trained expert, and likely can be completed more quickly.

5.5 Worker Spam

Crowdsourced workers have an incentive to do the least amount of work possible so they can get paid for the most amount of tasks. Thus, many users are likely to rapidly complete tasks and give random responses which we consider to be spam. From our experience, approximately 25% of responses are spam. This work was not focused on spam removal, but as a naive approach to spam removal, we first required that users answer all questions. By doing so, the effort to select any answer is almost the same as truly answering a question. Second, we disqualified all the workers who had more than 23 identical answers out of 28 (i.e., selecting TRUE or FALSE more than 23 times), removing their responses from the analysis. This filtering step allowed us to remove the workers who appeared to have performed the task by randomly selecting responses.

5.6 Crowdsourcing as part of ontology development

As we continue to understand both the feasibility and the best conditions to perform ontology-management tasks through microtask crowdsourcing, we envision that crowdsourcing can become a natural component of an ontology-development workflow. For instance, an ontology worker may select a portion of an ontology for which she needs additional quality assurance, generate microtasks for the crowd, and get the results, which will highlight potentially problematic areas of the ontology.

6 CONCLUSION

Ontology management via crowdsourcing provides a method to alleviate ontology development difficulties, especially with large ontologies. As we incorporate crowdsourcing components in ontology-management platforms, we need to understand which tasks are most amenable to crowdsourcing and how to configure the tasks in order to achieve the best performance. In this paper, we have demonstrated that when using microtask crowdsourcing for ontology verification, qualification tests, definitions, and question formulation significantly affect worker performance. We suggest that such configuration parameters will be important in other ontology management tasks.

ACKNOWLEDGEMENTS

This work has been 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

- Bernstein, M., Little, G., Miller, R., Hartmann, B., Ackerman, M., Karger, D., Crowell, D., and Panovich, K. (2010). Soylent: a word processor with a crowd inside. In *The 23d annual ACM symposium on user interface software and technology*, pages 313–322. ACM.
- Clark, H. H. and Chase, W. G. (1972). On the process of comparing sentences against pictures. Cognitive Psychology, 3(3), 472 – 517.

- Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., and Popovi?, Z. (2010). Predicting protein structures with a multiplayer online game. *Nature*, **466**(7307), 756–760.
- David, J., Guillet, F., and Briand, H. (2007). Association rule ontology matching approach. International Journal on Semantic Web and Information Systems, 3(2), 27–49.
- Demartini, G., Difallah, D. E., and Cudr-Mauroux, P. (2012). Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In 21st World Wide Web Conference WWW2012, pages 469–478, Lyon, France.
- Euzenat, J., Meilicke, C., Stuckenschmidt, H., Shvaiko, P., and Trojahn, C. (2011). Ontology alignment evaluation initiative: six years of experience. *Journal on data semantics XV*, pages 158–192.
- Evermann, J. and Fang, J. (2010). Evaluating ontologies: Towards a cognitive measure of quality. *Information Systems*, 35, 391403.
- Franklin, M., Kossmann, D., Kraska, T., Ramesh, S., and Xin, R. (2011). Crowddb: answering queries with crowdsourcing. In *International Conference on Management* of Data SIGMOD 2011, pages 61–72.
- GOConsortium (2001). Creating the Gene Ontology resource: design and implementation. Genome Res, 11(8), 1425–33.
- Ipeirotis, P., Provost, F., and Wang, J. (2010). Quality management on amazon mechanical turk. In ACM SIGKDD Workshop on Human Computation, pages 64–67.
- Just, M. A. and Carpenter, P. A. (1971). Comprehension of negation with quantification. Journal of Verbal Learning and Verbal Behavior, 10(3), 244 – 253.
- Kittur, A., Chi, E., and Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In 26th annual SIGCHI conference on human factors in computing systems, pages 453–456.
- Kozareva, Z. and Hovy, E. (2010). A semi-supervised method to learn and construct taxonomies using the web. In *Conference on Empirical Methods in Natural Language Processing*, pages 1110–1118, Cambridge, MA. Association for Computational Linguistics.
- Markotschi, T. and Völker, J. (2010). GuessWhat?! Human Intelligence for Mining Linked Data. In Proceedings of the Workshop on Knowledge Injection into and Extraction from Linked Data at EKAW.

- Noy, N. F., Mortensen, J., Alexander, P. R., and Musen, M. A. (2013). Ontology engineering through microtask crowdsourcing. *Under review*.
- Quinn, A. and Bederson, B. (2011). Human computation: a survey and taxonomy of a growing field. In Annual Conference on Human Factors in Computing Systems (CHI 2011), pages 1403–1412, Vancouver, BC. ACM.
- Raddick, M., Bracey, G., Gay, P., Lintott, C., Murray, P., Schawinski, K., Szalay, A., and Vandenberg, J. (2009). Galaxy zoo: exploring the motivations of citizen science volunteers. arXiv preprint arXiv:0909.2925.
- Sarasua, C., Simperl, E., and Noy, N. F. (2012). Crowdmap: Crowdsourcing ontology alignment with microtasks. In 11th International Semantic Web Conference (ISWC), Boston, MA. Springer.
- Schwarz, N. (1999). Self-reports: How the questions shape the answers. American Psychologist, 54(2), 93–105.
- Simperl, E., Norton, B., and Vrandecic, D. (2011). Crowdsourcing tasks in linked data management. In 2nd workshop on consuming Linked Data COLD2011 co-located with the 10th International Semantic Web Conference ISWC 2011, Bonn, Germany.
- Tanur, J. M. (1992). Questions about Questions: Inquiries Into the Cognitive Bases of Surveys. Russell Sage Foundation Publications.
- Thaler, S., Siorpaes, K., and Simperl, E. (2011). SpotTheLink: A Game for Ontology Alignment. In 6th Conference for Professional Knowledge Management.
- von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. In SIGCHI conference on Human factors in computing systems, pages 319–326. ACM Press New York, NY, USA.
- Wang, J., Ghose, A., and Ipeirotis, P. (2012). Bonus, disclosure, and choice: What motivates the creation of high-quality paid reviews? In *Thirty Third International Conference on Information Systems (ICIS)*, Orlando, FL.
- Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C. I., Tudorache, T., and Musen, M. A. (2011). Bioportal: Enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Research (NAR)*, **39**(Web Server issue), W541– 5.