A Comparison of Domain Experts and Crowdsourcing Regarding Concept Relevance Evaluation in Ontology Learning

Gerhard Wohlgenannt

Vienna Univ. of Economics and Business, Welthandelsplatz 1, 1200 Wien, Austria
{gerhard.wohlgenannt}@wu.ac.at
http://www.wu.ac.at

Abstract. Ontology learning helps to bootstrap and simplify the complex and expensive process of ontology construction by semi-automatically generating ontologies from data. As other complex machine learning or NLP tasks, such systems always produce a certain ratio of errors, which make manually refining and pruning the resulting ontologies necessary. Here, we compare the use of domain experts and paid crowdsourcing for verifying domain ontologies. We present extensive experiments with different settings and task descriptions in order to raise the rating quality the task of relevance assessment of new concept candidates generated by the system. With proper task descriptions and settings, crowd workers can provide quality similar to human experts. In case of unclear task descriptions, crowd workers and domain experts often have a very different interpretation of the task at hand – we analyze various types of discrepancy in interpretation.

Keywords: ontology learning, evaluation, crowdsourcing, human computation

1 Introduction

With the emergence of Web technologies, *knowledge creation* has evolved into a distributed process that integrates groups of users with different levels of expertise [5]. Recent approaches further broaden the knowledge creation process to include large populations of non-experts by using crowdsourcing techniques [6]. Crowdsourcing, in the form of gamification, and esp. in the form of paid microtask crowdsourcing, has become a popular means to solve tasks that computers cannot solve yet. It is often used to create training data for supervised machine learning, or for annotation and evaluation tasks.

Ontology engineering is a crucial knowledge acquisition process in the area of the Semantic Web. Ontologies are the vocabulary and therefore the backbone of the Semantic Web. Ontology construction is a complex and expensive task, therefore ontology learning systems, which (semi-)automatically generate ontologies from existing data (eg. unstructured domain text corpora), have been proposed. As the automatically generated ontological constructs need re-design

and pruning, we apply crowdsourcing and domain experts for evaluating various parts of the ontologies.

More specifically, in our ontology learning system [13, 16], we have applied both domain expert evaluation, as well as paid crowd workers to rate the domain relevance of domain concept candidates generated by the system. The ontology learning system learns lightweight ontologies from scratch in monthly intervals – in various knowledge domains. Therefore, the system has accumulated a lot of data which we will use in this publication to compare the characteristics and quality of domain expert judgments versus ratings by crowd workers.

We want to give some insights and lessons learned about the following questions: what are the quality and characteristics of crowd worker judgments in a task setting like judging the domain relevance of concept candidate labels? What are the differences in task setup and task description between crowd workers and domain experts? What influence do task description and settings in the crowdsourcing platform have on the resulting quality? And, in general, how well suited is crowdsourcing for domain specific knowledge acquisition jobs?

To address the research questions, we first compared the original ratings of crowdsourcing and domain experts for the data collected between 2013–2016. Then, we had another domain expert evaluation using a clearer task description in order to create a gold standard. We also repeated the crowdsourcing rating process with an extended task description and a careful selection of crowd workers, with the goal to improve the quality of the crowdsourcing results as far as possible – with regards to the gold standard. In a nutshell, our experiments show that crowd workers can provide quality similar to domain experts, if measures to raise quality are taken. But, obviously, the more specialized and complex the domain and the tasks, the harder it is to maintain good quality.

2 Related Work

There are three main types of crowdsourcing methods: paid-for crowdsourcing, games with a purpose, and altruism. Games with a purpose include human computation tasks as a side effect into playing (online) games [1, 7]. In paid-for crowdsourcing, more precisely *Mechanized Labor*, contributors carry out small tasks for a small amount of money, this is also call micro-task crowdsourcing. Two popular marketplaces that bring together crowd workers and customers are Amazon Mechanical Turk and CrowdFlower.

In the realm of ontology engineering, paid crowdsourcing has been used for various tasks. Eckert et al. [4] build concept hierarchies in the philosophy domain using Amazon Mechanical Turk. They use crowdsourcing to judge the relatedness of concept pairs, and to find taxonomic structures. An important aspect of ontology creation is taxonomy building, Noy et al. [10] verify the correctness of taxonomic relations with paid micro-task crowdsourcing. Wohlgenannt et al. [15] build and evaluate a crowdsourcing plugin for the Protege ontology editor. The authors focus on the tight integration of crowdsourcing (paid crowdsourcing, and

games with a purpose) into the knowledge engineering workflow, and analyze the benefits of crowdsourcing in terms of cost, time and scalability as compared to domain experts.

A closely related field of research is ontology alignment, where Sarasua et al. [11] use crowd workers to evaluate the correctness of *sameAs* relations, and to choose relations between terms. ZenCrowd [3] verifies the output of automatic entity linking algorithms. For a given term, crowd workers select the best fitting DBpedia URL that represents the entity. Rather recently, Amazon Mechanical Turk was used to generate Semantic Web benchmark data in the the Conference track of the Ontology Alignment Evaluation Initiative (OAEI).

Most previous work which studies the quality of annotations generated by crowd workers in the field of knowledge acquisition comes to the conclusion that the quality delivered by crowd workers is similar to domain experts, esp. when the complexity of the task is moderate [2]. Here, we want to provide more detailed insights into differences between domain expert and crowd worker judgements, and on task setup and task description for improved crowd worker quality.

3 System description

This section includes a short description of the ontology learning system used to generate the concept candidates. More details about the system can be found for example in [9, 13]. The goal of the system is to learn lightweight ontologies from so-called sources of evidence. These evidence sources include (domain-filtered) text collected from news media Web sites, social sources such as text from Twitter and Facebook, and structured sources like DBpedia and WordNet. Figure 1 provides an overview of the system. The starting point is a seed ontology. The

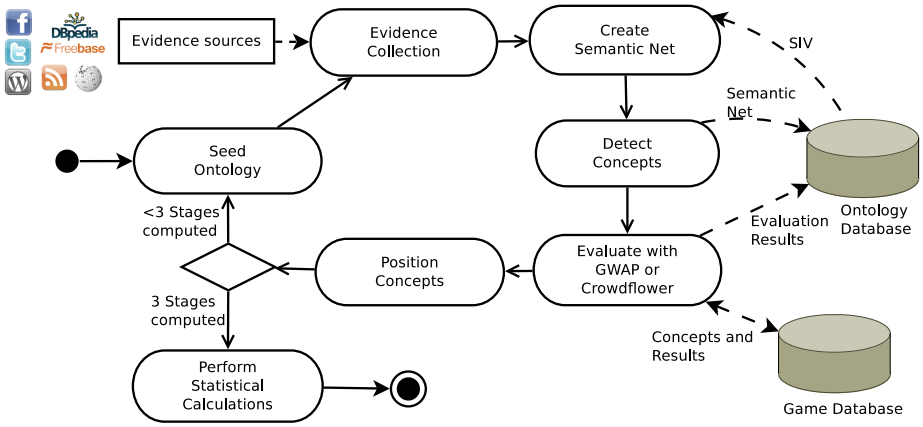

Fig. 1. The ontology learning process.

seed ontology typically contains 2–3 root concepts in the respective domain. For

these seed concepts, the system collects new evidence for related terms from the evidence sources. All this new evidence is stored into a Semantic Network. The neural networking algorithm of spreading activation then helps to select the most important concept candidates from the Semantic Network [14]. These concept candidates are then manually verified for domain relevance by either domain experts or crowd workers. Finally, the verified concepts are positioned in the existing ontology, resulting in an *extended ontology*. The extended ontology now serves as new seed ontology, and the extension process starts over. Usually we do three extension runs.

For this work, the most important step is the selection of concept candidates – as these will be evaluated in the remainder of the paper. From the plethora of terms in the Semantic Network (typically many thousands of terms), our system selects the 25 most promising concept candidates, according to their activation levels from the spreading activation algorithm.

4 Evaluation setup

All data used in the experiments in this paper stems from ontology learning experiments conducted from October 2013 to December 2015. In every month the system [13] computes ontologies in various domains, in each for various system settings, from scratch. Each month we only use the corpus data (for example from news media sites) collected in that respective month, which leads to an evolution of ontologies.

Fig. 2. Screenshot of an evaluation question on CrowdFlower.

The ontology system generates (among other things) concept candidate labels, which are then manually evaluated for domain relevance using either domain experts (DE) or crowd workers (CW). In this research we evaluate 4 domains, but only one domain (*Tennis*) has concept candidate relevance ratings by both DE and crowdsourcing. In crowdsourcing, we collected five votes for each

concept candidate label, and used majority voting to make a decision. Figure 2 shows a screenshot of what a crowd worker is presented when doing the evaluation task. We did not include the task instructions in the screenshot, as they would take up too much space.

5 Results

This section presents the evaluation results. In Section 5.1 we give an overview of the evaluation data collected, and then compare the original crowdsourcing and domain expert evaluations. In Section 5.2 we do a re-evaluation of the concept candidates with domain experts – using more precise task descriptions. This gives us a gold standard evaluation, which we then compare to the original evaluation of DE and CW. Section 5.3 presents the results from repeating the CW evaluations with improved task descriptions and settings, and compares them to the gold standard. And finally, in Section 5.4 we compare the original to the repeated CW evaluations.

5.1 Analysis of original data

Firstly, we present an aggregated view on the evaluation data. Table 1 shows the ratio of concept candidates judged as *relevant* in our four domains. In this table the statistics are referring to *distinct* concept labels. The concepts are counted only once, even if the concept candidates are occurring in various ontologies for the specific domain over time. We distinguish between results from *domain experts* and *crowd workers*.

Domain	Domain Experts	Crowd Workers
Tennis	137 of 647 (21.17%)	157 of 291 (53.95%)
Climate Change	304 of 889 (34.19%)	
NOAA	147 of 358 (41.06%)	
Middle-east crisis (DE)		322 of 570 (56.49%)

Table 1. Concept candidate labels judged relevant compared to total number of candidates automatically generated by the system in the domain; this table uses distinct concept labels, disregarding repeated occurrence of a label.

The most obvious observation in Table 1 is that CW were by far less strict with judging concept candidates. Crowd workers rated between 53%-57% of distinct concepts as relevant, whereas the ratio was only between 21% and 41% for DE, depending on the domain. So, domain experts naturally tend to have a stricter view on the world, and CW are more likely to accept a term if in doubt – esp. if they are not given precise task instructions.

Table 2 is similar to Table 1, but now we take multiple occurrences of the same concept candidates in different ontologies into account. As the underlying

data changes, the light-weight ontologies evolve, but obviously many domain concept candidates often re-occur.

Domain	Domain Experts	Crowdworkers
Tennis	1675 of 4165 (40.22%)	632 of 1030 (61.36%)
Climate Change	11508 of 16332 (70.46%)	
NOAA	578 of 1124 (51.42%)	
Middle-east crisis (DE)		592 of 1025 (57.75%)

Table 2. Ratio of concept candidate labels judged relevant; this table takes into account repeated occurrence of a concept label.

When we take repeated occurrence of concepts into account, the ratio of relevant concepts raises drastically for both judgments from DE and CW. The reason is the following: the concepts that re-occur over time and under different settings are typically the concepts which the ontology learning system regards the most relevant, whereas the concept candidates that are generated very rarely over time are likely to be not as important to the domain or the result of rather random appearance in the underlying text corpora. In domains where we have ratings both from CW and DE, ie. *Tennis*, we still have the situation of DE being more strict with their judgments.

The percentage of concepts ranked positive in the *climate change* domain is very high, although the concept labels have been ranked by DE. Our interpretation is the following: In the *climate change* domain we apply high quality corpora, not only mirrored from general news media sites, but also from domain-specific Websites of environmental NGOs. The corpora are larger and of better quality than for the other domains. (ii) The *climate change* domain is more stable than the other domains, and has a number of relevant concepts which re-occur in most generated ontologies, such as "global warming" or "climate". Savenkov and Wohlgenannt [12] evaluate the ontology learning data, which is also underlying this work, regarding ontological volatility, and find that the *Tennis* domain is more volatile than *climate change*.

In order to better interpret the observation that DE are much stricter in judgment, we analyzed the concept candidates that are overlapping, ie. which both appear in ontologies in the *Tennis* domain – which were evaluated using either by DE or CW. In total we have 691 distinct concept candidates in the *Tennis* domain, 247 of these overlapping between the evaluation methods (DE, CW). Taking into account repeated occurrence, the numbers change to 5195 total candidate concepts, and 4380 overlapping.

Similarly to a confusion matrix, Table 3 separates the overlapping concepts into four classes: (a) where both CW as well as DE judged the label as relevant to the domain, (b) DE say the concept is relevant, CW say it is not, (c) DE rate as non-relevant, but CW rate as relevant, (d) where both evaluator types come

to the same conclusion that the concept candidate is not relevant to the domain (of *Tennis*).

	CW: relevant	**CW: non-relevant**
DE: relevant	1889 (69)	108 (6)
DE: non-relevant	966 (76)	1417 (96)

Table 3. Relevance judgments for overlapping concept labels between Domain Experts (DE) and Crowd Workers (CW). Gives the numbers for repeated occurrence of a concept label in different ontologies, and in parenthesis counts for distinct concept labels.

Taking multiple occurrence of concept candidates into account, we can see the DE and CW agree in most cases (1889 plus 1417), which is an agreement on about 75% of candidates. But there is also a big group of concept candidates where crowd workers are less strict with their judgment (966). Only 108 of the 4380 overlapping concept labels DE judge as relevant, while CW rate as non-relevant.

Looking closer at the data, we can see that many of the terms in group *CW: relevant / DE: non-relevant* fall in one of three categories: (a) terms which have some relevance, but where the DE were more strict, eg. `baseline`, `field`, `loss` or `summer`, (b) wrong judgments by CW, for example `test`, `table tennis`, or `dog`, (c) terms which are not proper English words or phrases, such as `world no` or `ausopen`.

In order get a clearer picture of the differences between DE and CW, we had another domain expert re-evaluate (rate) all the 691 distinct concept candidates in the *Tennis* domain. This time we gave very clear instructions on how to measure domain relevance, most importantly: (a) only accept proper English phrases and abbreviations as relevant. For example `ATP` or `us open` is relevant, but not `usopen`; (b) if in doubt if a term is (closely) related to the domain rate as relevant.

5.2 Re-evaluation with an additional domain expert

In this section, we used an additional evaluation made by a domain expert over all 691 concept candidate labels as a gold standard – in order to analyze and interpret the results from the original DE and CW judgments.

First, we evaluated the accuracy, ie. the ratio of judgments, where CW and DE agree with ratings given in the gold standard. Table 4 shows the percentage of agreement, split into three categories: concept candidates rated *relevant* by the gold standard evaluation, concepts rated not relevant, and the sum of the two (*total*). In the table we distinguish, again, between taking the number of occurrences of concept candidates into account, and using distinct concepts only (in parenthesis).

	relevant	non-relevant	total
Accuracy DE	72.93% (58.51%)	96.46% (94.09%)	**84.00%** (83.72%)
Accuracy CW	85.38% (85.71%)	66.12% (58.72%)	**77.37%** (69.75%)

Table 4. Accuracy of DE and CW ratings with regards to the GS, for concept candidates incl. re-occurring candidates; distinct values in parenthesis

We see some interesting facts, for example that most concepts rated non-relevant by the gold standard evaluation were also rated non-relevant by DE evaluations, with 96.46%. So there is a strong agreement on non-relevant concepts, but for concepts rated *relevant* by the gold standard (GS) evaluation, agreement is much lower. This shows that the original DE were a lot stricter with accepting concept candidates as relevant. For CW evaluations the data shows the opposite effect, strong overlap for rating concepts as relevant, but only for 66.12% of the concepts rated as negative by the gold standard evaluator the CW agreed.

In total, as expected, the ratio of overlap is higher between DE evaluations and the gold standard evaluations than for CW versus GS, although the differences are rather small (84.00% versus 77.37%).

We also applied Cohen's kappa as prominent measure to compute inter-annotator agreement. The kappa value for CW and GS is 0.53 (again taking the number of occurrences into account), for DE and GS it is 0.68. According to the interpretation of Landis and Koch [8] we see *substantial* agreement between DE and GS, and moderate agreement between CW and GS.

A closer look at the data reveals some of the causes for the observation made. In a number of cases, GS evaluations rated concepts as relevant, whereas DE did not. Examples are: `draw, lawn, history, baseline`. All these examples are at least remotely relevant for the *Tennis* domain. The relatively high number of disagreement about *relevant* concepts clearly stems from the instructions given to the gold standard evaluator, namely to judge as relevant when in doubt. In contrast to DE versus GS, for CW compared to the GS, there was a lot of disagreement on concepts judged to be non-relevant by the GS evaluations. As in the last section, this often concerns concepts labels with are not English terms (eg. `womenstennis, andymurray`), plainly wrong ratings by CW (eg. `hair, garden, inning`), or terms too remotely relevant for the domain (eg. `foot, qualifier, livescore`).

5.3 Repetition of Crowdsourcing with new settings

With the lessons learned from our original experiments, in May 2016, we repeated the CrowdFlower evaluations for all the 291 (overlapping) concept candidates in the Tennis tennis domain. In this new crowdsourcing job, we gave preciser *task instructions*, which included a number of examples. The instructions were similar

to the instructions given to the gold standard domain expert evaluator, but a bit more detailed. Furthermore, we tried to get high quality results by only accepting the best workers (*level 3 workers*), and restricting worker residence to English speaking countries such as UK or US. Finally we used carefully designed test questions, which are called "gold units" in CrowdFlower, and crowd workers needed to pass at least 80% of the gold units.

Again, we use the GS evaluation as a baseline, and compare the newly gathered crowdsourcing results (*CW-new*) to the gold standard. Table 5 presents the results, distinguishing between agreement on *relevance* resp. *non-relevance* of candidate concepts for the Tennis domain.

	relevant	**non-relevant**	**total**
Number of terms	2157 of 2631	1591 of 1801	3748 of 4432
Number of terms (distinct)	93 of 119	153 of 172	246 of 291
Accuracy CW	81.98% (78.15%)	88.34% (88.95%)	**84.56%** (84.53%)

Table 5. Accuracy of repeated (modified) CW evaluation with regards to the GS, for concept candidates including re-occurring candidates; distinct values in are given parenthesis.

As we can see, the accuracy of the new CW evaluation is substantially higher than for the original CW data. Cohen's kappa is now 0.68 for distinct candidates, and 0.69 when taking the number of occurrences into account – as compared to 0.53 in the original CW evaluation. We attribute the improvement mainly to the updated task instructions. In *CW-new* we now rarely see unwanted terms such as `andymurray` (which is not a proper entity), but obviously there are still problems. For example, *CW-new* rated `quarterback` as relevant, which hints at domain knowledge missing. Another example is `world no`, which is a bi-gram fragment, and should not be rated as relevant. On the other hand, some terms which are relevant according to the GS evaluation, such as `ball games, fight, defender`, which were rated as non-relevant by *CW-new*.

Improved task instruction will never solve uncertainty with borderline cases regarding domain relevance, but should help to reduce other reasons for wrong judgements. We analyzed our two main reasons for wrong judgements: (i) concept candidates which are clearly not relevant to the domain, and (ii) terms which are not proper English concept labels (eg. hashtags, etc). For group (i) the number of errors was reduced from *21* to *3*, and for group (ii) from *14* to *3*. This clearly shows that the improved task descriptions and settings helped with these sources of errors.

Despite the occasional errors, the quality of *CF-new* has the same level of agreement to the GS as the original DE evaluation. In an attempt to further improve quality, we did another set of CrowdFlower evaluations of the 291 terms, where we only accepted workers that had at least 99% of the test questions (gold units) correct. From this additional crowdsourcing evaluation we expected an in-

crease in accuracy, but actually it stayed about the same. We archived a total accuracy of 84.2% on distinct candidates, and 82.1% when taking the number of occurrences into account. Cohen's kappa was 0.66 and 0.65, respectively. This shows the limits of human evaluation, which are caused by two main reasons: Most importantly, for some terms domain relevance is not clear, they are borderline cases. Furthermore, humans make mistakes in judgment, either because they did not understand the task instructions in all details, or they lack knowledge of parts of the domain.

5.4 Comparing the crowdsourcing evaluations

Finally, we investigate the differences between the results of the original CW evaluation, and the new CW evaluation. Table 6 shows the agreement on relevant and non-relevant concept candidates, as well as the differences between the two evaluations.

	CW-orig: relevant	**CW-orig: non-relevant**
CW-new: relevant	1785 (98)	582 (49)
CW-new: non-relevant	430 (81)	1635 (145)

Table 6. Relevance judgements comparing the original CW evaluation (*CW-orig*) and the re-evaluation with new settings and task descriptions (*CW-new*). Given are the numbers counting repeated occurrence of a concept label in different ontologies, and in parenthesis the numbers for distinct concept labels.

There is a large discrepancy between *CW-orig* and *CW-new*, the agreement is only *moderate*, with a Cohen's kappa value of 0.54. For distinct terms, the kappa is even lower, at 0.30. Again, the differences mostly come from the improved and clearer task descriptions. The original CW evaluation led to a more open view on the domain, and as already mentioned, included many terms which are not proper phrases or entities as relevant.

5.5 Discussion

One of the key learnings regarding *task setup* is that crowd workers will have a different interpretation of task definitions than domain experts. Therefore task definitions for crowd workers must be very precise and be backed up by more examples on how to solve a task. So extensive and precise *task descriptions* are crucial when using crowd workers, in addition to traditional measures to improve worker quality, such as using gold units (with CrowdFlower), allowing only workers with the highest skill level (as recorded in previous tasks), and only native English speakers. A discussion of the detailed results for various evaluation setups is included in the sections above already.

All the data itself, examples of task descriptions, gold units, and the code we used to analyze the data can be found online[1].

6 Conclusion and Future Work

In this paper, we compare micro-task crowdsourcing to the use of domain experts for the task of domain relevance assessment of concept candidates in ontology learning. First, we compared the data from our original crowdsourcing evaluation to the original domain expert evaluation. Then, we repeated the domain expert evaluation with improved task descriptions to create a gold standard. We also repeated the crowdsourcing evaluations with improved task descriptions and settings with the goal to raise crowd worker quality, and then compared that data to the gold standard.

We found that a very precise task description, including a number of examples, as well as strict worker selection and the use of gold units are crucial to ensure high quality results from crowd workers. Using these measures allows to deliver quality similar to human experts. But esp. in complex domains, crowd worker quality will vary, so we advise to explore the results with experiments with different settings and task descriptions in such cases. A limitation of all evaluation types (crowdsourcing and domain experts) are cases which cannot be judged clearly – in our case concept candidates with only moderate domain relevance. Disagreement in judgment among crowd users helps to detect these cases.

The main contributions of this work are the following: i) evaluating the suitability of crowdsourcing for a specific ontology learning task, ii) comparing the quality of crowd worker assessment to domain experts – based on extensive experiments which used different settings and variations, iii) doing a detailed analysis of the effects of evaluation strategies on the quality of results and the types of errors, iv) giving guidelines on how to set up crowdsourcing tasks in order to improve the evaluation quality.

In future work we will have a closer look at other domains used in our system, such as *climate change*, and also re-evaluate its respective concept candidates with updated CrowdFlower settings and task description. Furthermore, our system keeps relevance judgements about concepts only for a given time period – in order to facilitate the evolution of the domain by data-driven change. Concept candidates which re-appear will be judged again in a 6 month interval. For the data collected we will study how concept relevance understanding changes over time.

Acknowledgments. The work presented in this paper was created based on results from project uComp. uComp received the funding support of EPSRC EP/K017896/1, FWF 1097-N23, and ANR-12-CHRI-0003-03, in the framework of the CHIST-ERA ERA-NET.

[1] `https://aic.ai.wu.ac.at/~wohlg/miwai2016`

References

1. von Ahn, L., Dabbish, L.: Designing games with a purpose. Commun. ACM 51(8), 58–67 (2008), http://doi.acm.org/10.1145/1378704.1378719
2. Cheatham, M., Hitzler, P.: Conference v2.0: An uncertain version of the oaei conference benchmark. In: Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandecic, D., Groth, P., Noy, N., Janowicz, K., Goble, C. (eds.) The Semantic Web - ISWC 2014, Lecture Notes in Computer Science, vol. 8797, pp. 33–48. Springer International Publishing (2014)
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 Int. Conf. on World Wide Web. pp. 469–478. ACM (2012)
4. Eckert, K., Niepert, M., Niemann, C., Buckner, C., Allen, C., Stuckenschmidt, H.: Crowdsourcing the Assembly of Concept Hierarchies. In: Proc. of the 10th Annual Joint Conf. on Digital Libraries. pp. 139–148. JCDL '10, ACM (2010)
5. Gil, Y.: Interactive knowledge capture in the new millennium: how the Semantic Web changed everything. The Knowledge Engineering Review 26(1), 45–51 (2011)
6. Howe, J.: Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business (2009), http://crowdsourcing.typepad.com/
7. Krause, M., Smeddinck, J.: Human Computation Games: a Survey. In: Proc. of 19th European Signal Processing Conference (EUSIPCO 2011) (2011)
8. Landis, J., Koch, G.: The measurement of observer agreement for categorical data. Biometrics 33(1), 159–174 (1977)
9. Liu, W., Weichselbraun, A., Scharl, A., Chang, E.: Semi-automatic ontology extension using spreading activation. Journal of Universal Knowledge Management 0(1), 50–58 (2005)
10. Noy, N.F., Mortensen, J., Musen, M.A., Alexander, P.R.: Mechanical Turk As an Ontology Engineer?: Using Microtasks As a Component of an Ontology-engineering Workflow. In: Proc. of the 5th Annual ACM Web Science Conf. pp. 262–271. WebSci '13 (2013)
11. Sarasua, C., Simperl, E., Noy, N.F.: Crowdmap: Crowdsourcing ontology alignment with microtasks. In: Cudré-Mauroux, P., Heflin, J., et al. (eds.) International Semantic Web Conference (1). Lecture Notes in Computer Science, vol. 7649, pp. 525–541. Springer (2012)
12. Savenkov, V., Wohlgenannt, G.: Similarity metrics in ontology evolution. In: Klinov, P., Mourmotsev, D. (eds.) KESW 2015, Posters and Position papers. Moscow, Russia (October 2015)
13. Wohlgenannt, G.: Leveraging and balancing heterogeneous sources of evidence in ontology learning. In: et al., F.G. (ed.) ESWC,Portoroz, Slovenia. LNCS, vol. 9088, pp. 54–68. Springer (2015)
14. Wohlgenannt, G., Belk, S., Schett, M.: Computing semantic association: Comparing spreading activation and spectral association for ontology learning. In: Ramanna, S., Lingras, P., Sombattheera, C., Krishna, A. (eds.) MIWAI. Lecture Notes in Computer Science, vol. 8271, pp. 317–328. Springer (2013)
15. Wohlgenannt, G., Sabou, M., Hanika, F.: Crowd-based ontology engineering with the ucomp protege plugin. Semantic Web Journal (SWJ) p. Accepted/Scheduled for publication (2015)
16. Wohlgenannt, G., Weichselbraun, A., Scharl, A., Sabou, M.: Dynamic integration of multiple evidence sources for ontology learning. Journal of Information and Data Management 3(3), 243–254 (2012)