Undefined 1 (2009) 1–5
IOS Press



# The Thin Red Line - To What Extent can Crowd Workers Emulate the Performance of Experts?

**Editor(s):** Name Surname, University, Country
**Solicited review(s):** Name Surname, University, Country
**Open review(s):** Name Surname, University, Country

Ujwal Gadiraju [a,*], Patrick Siehndel [a], Besnik Fetahu [a], Stefan Dietze [a], Thomas Krijnen [b] and Jakob Beetz [b]

[a] *L3S Research Center, Appelstr. 9a, 30167 Hanover, Germany*
*E-mail: {gadiraju,siehndel,fetahu,dietze}@l3s.de*
[b] *Design Systems Group, Department of Built Environment, TU Eindhoven Eindhoven, The Netherlands*
*E-mail: {t.f.krijnen,j.beetz}@tue.nl*

**Abstract.**
Crowdsourcing has seen an increasing popularity for solving a variety of tasks. Specifically for tasks which address Web data interoperability, such as entity interlinking or schema mapping, crowdsourcing has been widely adopted in research and practice. In earlier work, we have investigated the behavioral pattern exhibited by workers, specifically, with respect to malicious activity and its detection. However, little research has been conducted so far to investigate the capability of crowd workers to substitute expert-based judgements, or, the extent to which worker performance is influenced by the task environment. In this work, we present results from a study which assesses the impact of the audience, crowd workers versus experts, and the specific task design and environment on the worker performance. Our study has been conducted on the specific tasks of link prediction and schema mapping, where tasks have been carried out by crowd workers and predefined experts. We compare the performance of crowd workers and experts in environments that differ with respect to the level of contextual information provided about the task at hand. Our results show that it is feasible to attain high quality results from crowd workers, that are comparable to experts when adequate context is presented to the crowd. In addition, we present a detailed analysis of the errors that the crowd and experts tend to commit in the task of link prediction and schema mapping. Finally, we present a novel method to assess the inherent *difficulty* of the link prediction task, and the impact of this task difficulty on the quality of the results.

Keywords: Crowdsourcing, Evaluation, Experts, Crowd Workers, Link Prediction, Schema Mapping

## 1. Introduction

Over the last decade, crowdsourcing has gained vast popularity amongst researchers and practitioners, specifically to address tasks related to Web data linking and integration. In more recent times, crowdsourcing has found remarkable applications ranging from biomolecule designing[1] to aiding disaster relief operations [28]. Due to this widespread popularity and

*Corresponding author. E-mail: gadiraju@l3s.de.

[1] http://www.nature.com/news/victory-for-crowdsourced-biomolecule-design-1.9872

the vast potential in the crowdsourcing paradigm, researchers have taken interest in various aspects such as improving the quality of the crowdsourced results, increasing the efficiency and cost-effectiveness of crowdsourcing, optimizing the throughput, and so forth.

While previous works have investigated the applicability of crowdsourcing for different purposes and in varying domains, little work has been accomplished with respect to investigating the extent to which crowdsourced results can compete with those acquired through experts. James Suroweicki pointed out in his seminal book that the 'wisdom of crowds' can replicate expertise under certain conditions that facilitate diversity in the crowd and independence in their judgments [26]. The notion of crowdsourcing is built around the thought that the 'whole is greater than the sum of its parts', i.e., small and independent contributions from a large number of workers can be accumulated to attain an adequate result overall. However, not all tasks are fit for crowdsourcing. On the one hand, this may be simply due to the complex nature of some tasks that cannot be decomposed into smaller units for the consumption of various workers in a crowd. On the other hand, tasks may require domain-specific expertise.

In this work, we aim to study the extent to which crowd workers and their collective wisdom can emulate that of experts in a task specific setting. For this reason, we consider the task of *link prediction* and *schema mapping*, which aim at the identification of explicit links between resources on the Semantic Web at the instance and schema level. Different types of predicates, such as `owl:sameAs`, `skos:relatedMatch`, `skos:narrowMatch` and so forth, can be used for linking various resources. This task is challenging for a variety of reasons, including (i) complexity of contextual information that can potentially aid in determining appropriate relations, i.e., the available semantic context of entities and concepts, and (ii) possible ambiguity between resources, requiring disambiguation. Understanding the semantic context of a given set of entities or concepts is specifically challenging, when considering cross-dataset links, which require the consideration and comprehension of distinct knowledge graphs.

Our findings bear important implications in the realms of crowdsourcing domain-specific microtasks (such as the one adopted in this paper pertaining to a real Semantic Web task that requires human computation). The main contributions of our work are listed below.

**1.** Through extensive experiments, we investigate the impact of the target community, i.e., crowd workers and experts, and the impact of the task environment on the performance in the specific task of link prediction and schema mapping.
**2.** We present a detailed analysis on the types of errors that crowd workers as well as experts tend to commit in the task of link prediction and schema mapping.
**3.** We present findings regarding the extent to which crowd workers can replicate the performance of experts in two distinct settings; one that is more suitable to crowd workers by virtue of a high context, and another that is suitable for experts in the field by virtue of limited context. These insights provide clues about suitability of tasks and appropriate task design.
**4.** We propose a novel method to assess the expertise that is required to accurately identify the apt predicate between a pair of URIs. Thus, we present a way to comprehend the *difficulty* of a task and its influence on the performance of crowd workers and experts, while at the same time demonstrating how tailored task design can improve the performance of the crowd for more challenging tasks.

The remainder of this paper is structured as follows. In the next section, we present relevant literature and position our work in the light of exisiting research. In Section 3, we present our goals and approach. Section 4 describes our experimental setup, followed by Section 5 and 6 that elucidate our analysis and findings. Finally, we present a discussion, draw conclusions and set precedents for future work in Section 7.

## 2. Background - Crowdsourcing Tasks and Task Design

In this section, we present key findings from our previous works that serve as background and motivation for the work presented in this paper. As part of earlier work, we delved into means of making crowdsourcing microtasks more effective through optimal task design, and based on our findings from analysis of workers behavior. We draw important considerations from our previous observations, and shape our crowdsourcing tasks by encorporating the recommended guidelines accordingly.

We studied two primary aspects that influence the effectiveness of crowdsourcing: (i) microtask design [7], and (ii) workers behavior [8]. We hypothesized that leveraging the dynamics of tasks that are crowdsourced on one hand, and accounting for the behavior

of workers on the other hand, can help in designing tasks efficiently. However, that requires a thorough understanding of the typical behavioral pattern exhibited by crowd workers, as opposed to dedicated experts.

In order to gain an understanding of the landscape of microtasks that are popularly crowdsourced, based on an extensive study of 1000 workers on CrowdFlower[2], an established microtask crowdsourcing platform, we proposed a two-level categorization scheme for tasks as shown in Table 1.

The top level in the taxonomy consists of goal-oriented classes. The second level contains sub-classes of these top level classes, that are based on the work-flow of the microtasks. The top level describes the overall objective of a given microtask, while the second level describes the process that a crowd worker has to go through in order to complete the task successfully and help the requester or task administrator achieve his goal. It is noteworthy that in this taxonomy work-flow oriented sub-classes can belong to multiple goal-oriented top level classes. This categorization scheme serves as the first step towards establishing task-specific guidelines for the efficient design of microtasks.

Additionally, we studied the behavioral patterns of workers (especially malicious workers) based on their responses in *surveys*, which is a popular top level type of task from the proposed taxonomy. We relied on the following factors to determine the behavior topology proposed in our work; (i) eligibility of a worker to participate in a task, (ii) adherence of responses to pre-stated rules, and (iii) the extent to which responses satisfy the expectations of the administrator. Based on this study, we could identify the following characteristic types of malicious workers [8].

- **Ineligible Workers (IE)**. Task administrators present instructions to the workers that they should follow to complete a given task successfully. The workers who do not qualify as per such priorly stated requisites belong to this category.
- **Fast Deceivers (FD).** Malicious workers are characterized by their behavior that is highly suggestive of an underlying motivation to earn quick money by exploiting microtasks. This is apparent from some workers who adopt the 'fast-response-first' approach such as copy-pasting the same response for instance. Such workers belong to the class of fast deceivers.
- **Smart Deceivers (SD).** Some eligible workers who are malicious, attempt to deceive task administrators by cleverly adhering to the rules. Such workers mask their real objective by simply not violating or triggering implicit validators, and belong to this category.
- **Rule Breakers (RB).** A behavior prevalent among malicious workers is their lack of conformation to clear instructions with respect to each response. Data collected as a result of such behavior has little value, since the resulting responses may not be useful to the extent intended by the task administrator.
- **Gold Standard Preys (GSP).** Some workers who abide by the instructions and provide valid responses, surprisingly fall short at the gold standard questions. Although these workers exhibit non-malicious behavior, they stumble at one or more of the gold standard test questions due to to their inattentiveness, fatigue or boredom.

Based on our study, involving 1000 workers, 568 passed the gold standard test and have been classified as *trust-worthy workers*, where only 335 workers could be considered *elite workers*, i.e. providing perfect answers according to our ground truth. Non-elite workers, trustworthy as well as un-trustworth ones, have been classified into the above categories, leading to the distribution shown in Figure 1.

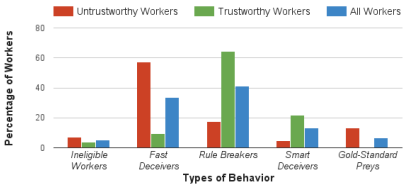



Fig. 1. Distribution of non-elite workers as per their behavior.

While this preliminary analysis is not the focus of this paper and we refer to further details in [8], it serves to show the specific nature of crowd workers, specifically when compared to experts. As these behavioral parttern affect workers' performance and the overall quality of results, we proposed guidelines that can be used to tackle each of the malicious behavioral patterns.

- To restrict the participation of ineligible workers, task administrators should employ pre-screening methods.

[2] http://www.crowdflower.com/

Table 1
A two-level taxonomy for typically crowdsourced microtasks.

| **Information Finding (IF)** | **Verification & Validation (VV)** | **Interpretation & Analysis (IA)** | **Content Creation (CC)** | **Surveys (S)** | **Content Access (CA)** |
|---|---|---|---|---|---|
| Data / Metadata finding | Content Verification | Classification | Media Transcription | Feedback/Opinions | Testing |
| | Content Validation | Categorization | Data Enhancement | Demographics | Promoting |
| | Spam Detection | Media Transcription | Translation | | |
| | Data matching | Ranking | Tagging | | |
| | | Data Selection | | | |
| | | Sentiment Analysis | | | |
| | | Content Moderation | | | |
| | | Link Prediction | | | |

– An important guideline to enforce is to curtail malicious activity from fast deceivers. Stringent validators should be used in order to ensure that workers cannot bypass open-ended questions by copy-pasting identical or irrelevant material as responses.
– Rule breakers can be curtailed by ensuring that basic response-validators are employed, so that workers cannot pass off inaccurate responses, or nearly fair responses. Lexical validators can enforce workers to meet the exact requirements of the task and prevent ill-fitting responses.
– Since smart deceivers take special precautions to avoid being detected, they present the biggest hindrance in overcoming. These workers can be restricted by using psychometric approaches (for instance, repeating or rephrasing the same question(s) periodically and cross-checking whether the respondent provides the same response).
– Finally, we note that depending on the type of task, there may be a fair number of gold standard preys. We recommend post-processing step that can be accommodated in order to identify such workers and consider their acceptable responses to boost the reliability and quality of results.

In addition to following the lessons learned from these works, we conducted a series of experiments in order to deduce an optimised task design for the *Interpretation and Analysis* (**IA**) task type. Since there are no established guidelines or tangible recommendations for task design with respect to key parameters such as *length*, *monetary incentive* and *time required* for task completion, we studied the tuning of these parameters based on our findings from extensive experiments and analysis of 'categorization' tasks. We established that the task completion time of a worker is strongly and positively correlated to the worker's accuracy. Therefore, care must be taken to ensure adequate time is provided to the crowd for task completion. For optimal results, it is therefore safer to err on the higher side of the time required. We also found that the accuracy of workers decreases as they proceed in a task, more so towards the end of longer tasks. This shows that a task administrator can profit by splitting a relatively long task into shorter ones before deploying it to the crowd.

In the context of comparing crowd workers and experts, crowd workers have the following distinct characteristic features:

**Varying Demographics** Through our previous works [7,8], we have confirmed that in the absence of restrictive and controlled environments, crowd workers that participate in microtasks through established crowdsourcing platforms such as CrowdFlower depict varying demographic features, more so in terms of the geographical landscape that they span.

**Incentives and Monetary Compensation** A particulary prominent feature of crowd workers that has been studied widely in previous works is their motivation to participate and contribute to completing crowdsourced microtasks [19,13]. We found that the primary motivation for crowd workers is to earn the monetary rewards typically attained on successful task completion [7].

**Malicious Activity** It has been shown that malicious activity is prevalent in the crowd [5,8]. While there are trustworthy workers who genuinely attempt to perform to the best of their ability and ensure a high quality as a result of that, there are malicious workers who aim to complete as many tasks as they can in a short span of time with an aim to maximize their earnings.

Due to these factors and the inherent characteristics of crow behavior, achieving comparable performances

to an extent where experts can be replaced by crowd workers in a particular task can be challenging. In addition, studies are required which provide a more structured investigation of the performance of crowd workers, when compared to expert workers, and strategies on how to compensate apparent performance gaps.

## 3. Objectives and Methodology

In this section, we describe the methodology and experimental setup in which we investigate the performance of crowd workers in comparison to that of *experts*. Through our work we aim to study the reliability of using crowdsourcing for a domain-specific *Interpretation and Analysis* (**IA**) task. The **IA** task is one of the 6 top-level microtask types in the crowdsourcing microtask taxonomy introduced in our previous work [7]. As mentioned earlier, we consider the task of link prediction and schema mapping.

### *3.1. Objectives*

We aim to assess whether exploiting the collective wisdom of a crowd can be reliably used in order to determine accurate relations between resource URIs, in comparison to the judgment of experts in the field of Semantic Web and Linked Data. To this end, we aim to tackle the following research questions.
**RQ#1.** Can the tasks of link prediction and schema mapping be reliably crowdsourced?
**RQ#2.** Is there a significant performance gap between experts and crowd workers in the link prediction and schema mapping task?
**RQ#3.** What is the impact of the task enviroment on the performance of crowd workers and experts?
**RQ#4.** How is the task performance of crowd workers affected by varying task difficulty, when compared to the performance of experts?

### *3.2. Methodology*

In order to address the research questions presented in the earlier section, we investigate performance on the link prediction and schema mapping tasks in four different setups. Setups vary with respect to their target *participants* and the level of *contextual information* presented. With regards to the participants, we consider the following categories.

**Experts.** We consider *experts* to be our participants who are familiar with the task of link prediction and schema mappings by virtue of their profession. They are either students, PhD candidates, or Post-doctoral researchers with an interest in the field of the Semantic Web and Linked Data.

**Crowd.** We refer to the participants that respond to our microtasks deployed on the crowdsourcing platform through different crowdsourcing channels, as *crowd workers* or simply the *crowd*. Note that crowd workers inherently depict a high diversity with respect to their demographic characteristics and knowledge, as shown in the previous section.

In order to vary the quantity and quality of contextual information, we use two distinct enviroments for harvesting input from the participants.

**Environment I (high context).** We use CrowdFlower in order to deploy the task of link prediction and schema mapping, and consequently gather judgments. The task is designed such that it is suitable for crowdsourcing. Specifically, we provide access to background information about involved concepts through their URIs. Workers and experts can follow those links, in order to determine the apt relation between URI paris. We represent the crowd workers participating here with *Crowd I*, and the experts with *Experts I*. Note that we followed the applicable guidelines during the task design phase, as described in Section 2.

**Environment II (low context).** We built an interface (henceforth referred to as the *Interlink Interface*) that is conducive for acquiring judgements from crowd workers and experts. This setup consists of a graphical representation of URI pairs and is designed such that one can infer the concepts portrayed by the URI pairs immediately. In comparison to *Environment I* described above, lesser context regarding the URI pairs is available directly. We represent the crowd workers participating here with *Crowd II*, and the experts with *Experts II*.

Finally, this results in four different configurations (*Crowd I*, *Experts I*, *Crowd II*, *Experts II*) which are investigated in the forthcoming sections.

## 4. Experimental Setup

### *4.1. Task and Dataset*

We assess the performance of the crowd and experts pertaining to the task of identifying apt relationship links between URI pairs, as defined

priorly within the publicly available DBpedia[3] dataset. These links between URI pairs are defined as *predicates*. Here, we specifically consider two schema-level predicates (`owl:equivalentClass`[4], `rdfs:subClassOf`[5]), and two instance-level predicates (`owl:sameAs`[6], `skos:related`[7]) for the sake of our experiments.

Since the task to be performed consists of assigning one of these predicates to a given pair of resources, either entities or concepts, we first create a ground truth dataset. The ground truth consists of randomly selected triples (of the form: `subject`, `predicate`, `object`), where predicates belong to one of the aforementioned categories. For each of these four distinct predicates, we randomly select a set of 150 triples, resulting in a ground truth of 600 triples. As the subject and object are uniquely identified through URIs, we refer to them as a *URI pair*, where both URIs of a single pair are either a DBpedia entity, a DBpedia concept (or type) or a DBpedia category.

While crowdsourcing is a relatively cheap paradigm and judgments from workers are readily available, acquiring judgments from experts requires more resources (in terms of time). For this reason, we have limited our experiment to this sample size.

### 4.2. Environment 1 - Crowdsourcing Microtask with added contextual Information

In the crowdsourcing microtask setup, we model the task of link prediction and schema mapping in order to make it conducive for consumption by crowd workers. Figure 2 shows an example pair of URIs within the crowdsourced microtask. Workers are asked to select the most accurate relation between the pairs of resources in each case.

As part of the instructions and guidelines for the microtask, we explain the *context*, i.e. the underlying meaning and distinctions between the 4 considered predicates for our experiments. Note that we use simplified representations of the predicates, such as *same* instead of `owl:sameAs`, *related* instead of `skos:related`, *equivalent* instead of `owl:equivalentClass` and *subclass* instead of `rdfs:subClassOf`, in order to make the task more comprehensible for crowd workers. We provide examples of each relation in order to ensure that there is no artificial bias created in the responses of workers, due to the misinterpretation of the relations between URI pairs.

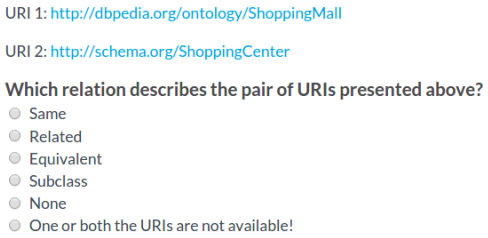



Fig. 2. Example URI pair from different schemas (DBpedia and Schema.org), in the crowdsourcing microtask on CrowdFlower.

We do not enforce any restrictions on the workers based on geographical boundaries. For each pair of URIs, we gather 5 judgments from independent crowd workers, resulting in 3,000 responses. In order to receive reliable responses, we restrict the participation of the crowd to the *Level-1* group of workers. These are the workers with the best quality and reputation ratings (top level) as determined by the CrowdFlower platform. For every set of 20 responses that a worker submits corresponding to URI paris, we provide a monetary compensation of 2 USD cents. Drawing from the guidelines from prior work, we do not restrict the time available for completion of the task with an aim to acquire good quality responses.

For the same dataset consisting of 600 URI pairs, we also acquire judgments from experts. In total, around 20 experts participated in the task. We use the same task format and design as in case of the crowd workers, by leveraging the CrowdFlower 'internal workforce' facility. This helps us to recruit our own 'crowd' by sharing a direct and internal link to the microtask. In this way, for each pair of URIs, we gather 3 independent judgments from experts, resulting in 1,800 responses. Note that we do not provide experts with a monetary compensation.

### 4.3. Environment 2 - Visual Interlinking Interface

In addition to using CrowdFlower to host the task in the form of a crowd-friendly microtask, we developed an interface (`http://crowds-vs-experts.duraark.eu/`) to gather responses from both crowd workers and experts. As opposed to Environment 1, no

[3] `http://dbpedia.org/`
[4] `http://www.w3.org/TR/owl-ref/#equivalentClass-def`
[5] `http://www.w3.org/TR/owl-ref/#subClassOf-def`
[6] `http://www.w3.org/TR/owl-ref/#sameAs-def`
[7] `http://www.w3.org/TR/skos-reference/#semantic-relations`

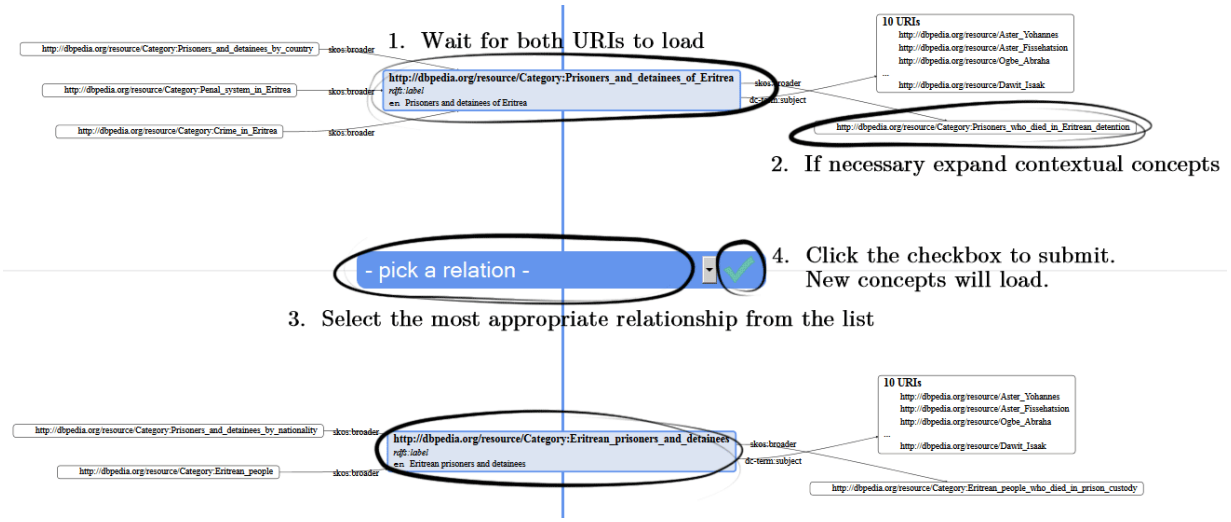



Fig. 3. The *Interlink Interface* and steps suggested for gathering responses regarding the relationships between URI pairs from Crowd Workers and Experts.

added contextual information is provided here, but understanding of used concepts and predicate types is assumed. Our assumption is that this will imply a less crowd-friendly, but expert-oriented environment.

Figure 3 depicts the *Interlinking* interface. As shown in the figure, the interface presents additional relationships that correspond to each of the URIs in order to facilitate an understanding of what the resource represented by the URI really conveys. Workers and experts can use these particulars to assess the most apt relation between the two given URIs. In addition to this, one can also attain additional contextual information by clicking on any of the URIs presented on the screen. We hypothesize that this minimalistic representation of the URI pairs, embedded in additional related resource URIs that provide a brief context, are typically sufficient for experts to make their decisions.

Note that in case of Environment 2, the URI pairs were limited to those that were both within the DBpedia namespace due to interface constraints. This resulted in 363 URI pairs. Once again, we gathered 5 independent judgments from crowd workers and 3 independent judgments from experts for each pair resulting in 2,904 responses. We acquire 5 judgments from workers and 3 from experts. This is due to the fact that replication and redundancy in gathering responses is a necessary instrument to harness the 'wisdom of a crowd' [26], as opposed to the case of expert judgments. As mentioned earlier, due to the costly nature of expert judgments (especially in terms of time) we restrict the number of responses from independent experts for each URI pair. Nevertheless, we collect 3 judgments from experts for each URI pair to account for the agreement between different experts.

## 5. Results: Crowds versus Experts

In this section we present, discuss, and compare the results we obtained from the two different experimental setups. We investigate the accuracy of crowd workers and experts, as well as the errors they are prone to committing.

### 5.1. Accuracy and Agreement

In order to investigate **RQ#1**, **RQ#2** and **RQ#3**, we assess the accuracy of judgements and the inter-annotator agreement. In the Environment I, we note that experts achieve a higher accuracy in determining the apt relations between the given pairs of URIs. The crowd workers perform with an average accuracy of around 85%, while the experts achieve an accuracy of nearly 90%. The inter-annotator agreement (according to percent pairwise agreement) is also higher in case of the experts (over 94%) when compared to that of the workers (75%).

In the Environment II, we observe that experts vastly outperform crowd workers. While the experts perform with an accuracy of nearly 75%, the crowd workers achieve an accuracy of only around 45%. In addition, we see that the experts exhibit a very high inter-

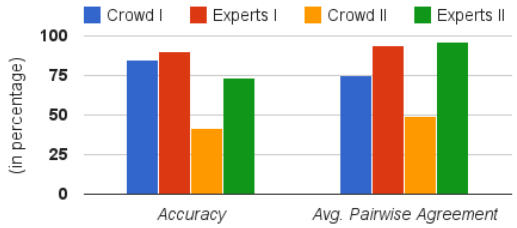



Fig. 4. A comparison between the accuracy and pairwise agreement of Crowd Workers and Experts in the Environments I and II.

annotator agreement (according to percent pairwise agreement) of nearly 95% in comparison to the 50% agreement in case of the crowd workers. We discuss the implications of these findings in Section 8.

### 5.2. *Errors in Judgment*

While investigating **RQ#4**, i.e. the impact of task difficulty on the performance in varying setups, we hypothesize that some relationships between URI pairs are easier to determine than others.

We investigate the performance of crowd workers and experts with respect to each of the relationships considered. Herein, we analyze the errors that workers and experts commit from two different standpoints.

**Predicates Erred-On.** Here we observe the types of relationships between URI pairs which workers and experts fail to identify accurately.

**Predicates Erred-As.** In this case, we analyze the predicates that are wrongly attributed to URI pairs by crowd workers and experts.

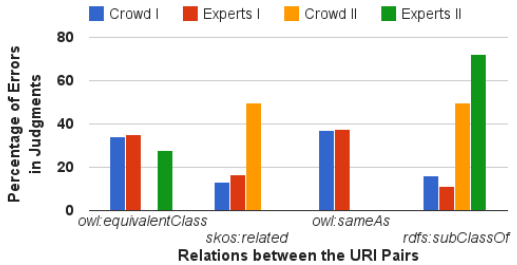



Fig. 5. Distribution of *Predicates Erred-On* by Crowd Workers and Experts in the Environments I and II.

Figure 5 presents the distribution of relationships erred-on by crowd workers and experts in Environment I (Crowd I, Experts I) and Environment II (Crowd II, Experts II).

In the Environment I, we find a similar distribution between crowd workers and experts. Herein, both Crowd I and Experts I err most in judging URI pairs with the `owl:samAs` relationship, followed by those with `owl:equivalentClass` relationship. We believe that this is due to the subtle difference conceptually between the class-level and instance-level predicates between the resources. The smallest portion of the errors committed by crowd workers (around 12% of the errors) pertain to the `skos:related` relation, while only 10% of the errors in judgments by experts correspond to the `rdfs:subClassOf` relation.

On the contrary, in Environment II crowd workers and experts err on varying predicates. While Crowd II err equally on the `skos:related` and `rdfs:subClassOf` predicates, around 75% of the errors committed by Experts II correspond to the `rdfs:subClassOf` predicate and 25% correspond to the `owl:equivalentClass` predicate. Neither the Crowd II, nor Experts II err on the URI pairs with the `owl:samAs` predicate.

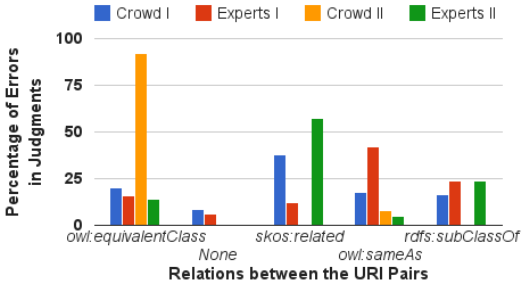
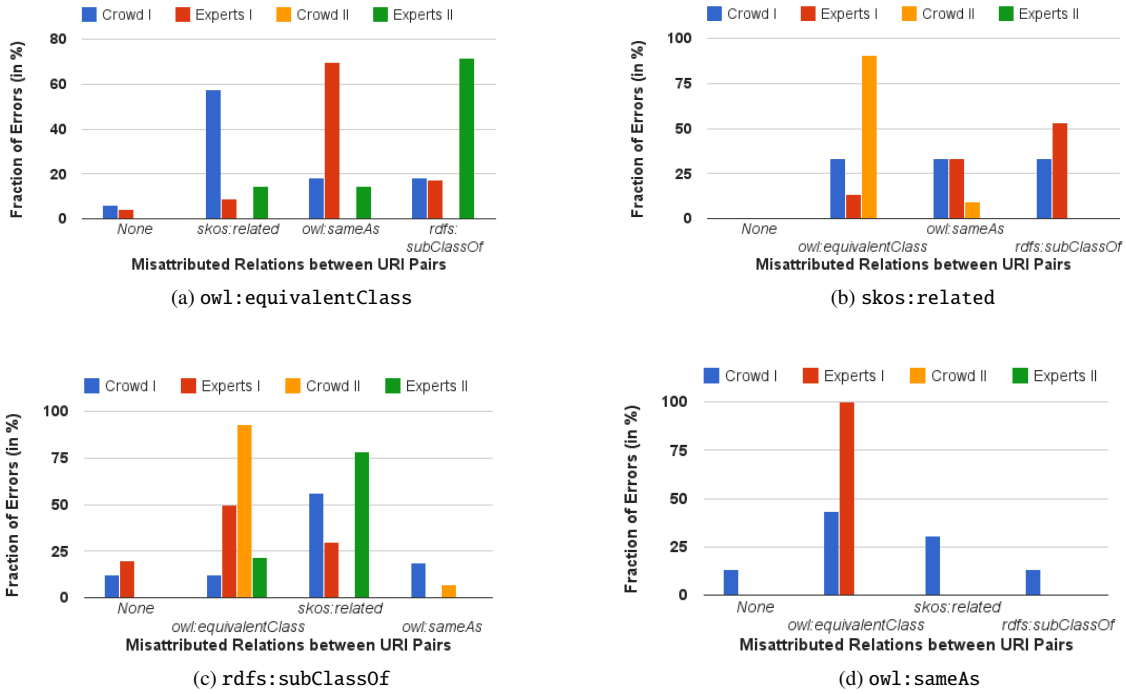



Fig. 6. Distribution of *Predicates Erred-As* by Crowd Workers and Experts.

Figure 6 shows the distribution of relionships erred-as by workers and experts in both environments, i.e., the relationships that are misattributed to URI pairs. Of the errors in judgment committed by Crowd I, around 37.5% of relationships between URI pairs are misattributed to `skos:related`, while nearly 20% of the errors correspond to attributing the `owl:equivalentClass` relation to URI pairs. In case of the Experts I, around 40% of the errors in their judgments are a consequence of identifying URI pairs as being related by the `owl:sameAs` relationship, followed by nearly 25% corresponding to the `rdfs:subClassOf` relation. A small percentage of the errors arise from workers and experts identifying no relationship between URI pairs in the Environment I, as depicted by the relation `None` in Figure 6.




(a) `owl:equivalentClass` (b) `skos:related`

(c) `rdfs:subClassOf` (d) `owl:sameAs`

Fig. 7. Distribution of misattributed URI pairs for each of the considered predicates by Experts and Crowd Workers, in Environments I and II.

In Environment II, nearly 90% of the errors committed by Crowd II correspond to wrongly attributing URI pairs with the `owl:equivalentClass` predicate. Another 10% of the errors by Crowd II result from misattributing the URI pairs with the `owl:sameAs` relationship. Experts II misattribute over 50% of their errors in judgments to `skos:related`, nearly 25% to the `rdfs:subClassOf`, around 20% to `owl:equivalentClass`, and almost 5% to the `owl:equivalentClass` predicate.

### *5.2.1. Misattributed Relations*

Next, we take a closer look at the errors in judgments by workers and experts in both the experimental environments. We study the nature of the misattributed relations, i.e., for each type of relationship between the URI pairs, we study the relations that are wrongly attributed to the pairs. Figure 7 presents the distribution of the misattributed URI pairs across the different relations considered.

In case of the pairs with the `owl:equivalentClass` relationship (see Figure 7a), a large proportion of the mistakes by the Experts I (nearly 70%) attribute `owl:sameAs` relation to the pairs, while nearly 60% of the errors committed by the Crowd I in this case, attribute `skos:related` relation to the pairs. On the contrary, in Environment II crowd workers (i.e., Crowd II) do not misattribute the `owl:equivalentClass` predicate. Over 70% of the mistakes made by the Experts II in this case, misattribute the URI pairs to the `rdfs:subClassOf` predicate.

Figure 7b shows that in case of URI pairs which actually share the `skos:related` relationship, in Environment I experts wrongly identify 60% of them to be related by `rdfs:subClassOf` relation, followed by around 33% of them to be related by the `owl:sameAs` relation. Crowd workers misattribute the pairs with `skos:related`, in equal proportions with `owl:equivalentClass`, `rdfs:subClassOf`, and `owl:sameAs` relations. In contrast to this, in Environment II the experts (i.e., Experts II) do not err on these URI pairs with `skos:related` predicate. Nearly 90% of the errors committed by Crowd II in this regard, misattribute the `owl:equivalentClass` predicate to the URI pairs, while almost 10% do so with the `owl:sameAs` predicate.

From Figure 7c, we see that in Environment I crowd workers err on pairs with the `rdfs:subClassOf` relation, by judging that they share the `skos:related` predicate (over 55%), while experts primarily misattribute the `owl:equivalentClass` predicate to these

pairs (around 50%). In Environment II on the other hand, around 90% of the misattributions by Crowd II pertain to `owl:equivalentClass` predicate, while over 75% of the misattributions by Experts II correspond to the `skos:related` predicate.

Finally, we find that of all the misattributed relations between pairs that actually share the `owl:sameAs` relation, in Environment I experts err by identifying them to be related by the `owl:equivalentClass` relationship, and crowd workers depict the same albeit to a lesser extent (nearly 50% of the misattributed relations). We believe that this is indicative of the difficulty in distinguishing class and instance-level concepts in case of both experts and workers. In Environment II however, we observe that neither crowd workers (i.e., Crowd II) nor experts (i.e., Experts II) commit mistakes when it comes to judging the apt predicate between URI pairs sharing the `owl:sameAs` relation.

## 6. Impact of Concept Familiarity

An additional investigation with regard to **RQ#4** builds on the assumption that predicates can be easier to adjudicate if the concepts or entities involved are familiar to participants and better understood. Here, we hypothesize that *more popular* and *less granular* resources are more likely to be familiar to the worker, whether crowd or expert, leading to better results and a lesser dependence on added context information (Environment 1).

As indicators for popularity and granularity, we rely on features extracted from Wikipedia. For this reason, we considered a subset of 363 URI pairs where both subject and object URIs are within the DBpedia namespace, and hence, have a corresponding Wikipedia page. In the first series of experiments we used features such as *in-links*, *out-links* and the *positon* of an associated concept within the Wikipedia category graph. In previous works, these features have been shown to indicate the popularity of Wikipedia pages, and hence, we suppose are reliable indicators of how well an entity is understood by the general public.

We are therefore led to believe that these features will help us determine whether workers are more capable of responding accurately in case of certain concepts when compared to others. In the second series of experiments, we took a closer look at the different categories of the content of the links, and how these influence the results of our experiments.

Table 2
Features of correctly and incorrectly determined predicates based on responses from the crowd workers in Environment I.

| | Depth | In-links | Out-links | Text Length |
|---|---|---|---|---|
| Correct Entities | 5.74 | 2556.6 | 132.93 | 27,609 |
| False Entities | 6.01 | 1161.77 | 76.79 | 16,408 |

### 6.1. Impact of Entity Familiarity on Performance

The experimental setup for this section is as follows. For each given resource from DBpedia represented by a URI, we retrieve the corresponding Wikipedia article. For each resource, we analyzed the following features:

**Depth** The *depth* describes the position of a given concept in the Wikipedia category graph. As we traverse deeper down into the Wikipedia category graph, the concepts tend to be become less general and more specific.

**In-links** This describes the number of distinct Wikipedia articles pointing to a given article. Articles with many in-links are related to many other articles.

**Out-links** This describes the number of other distinct Wikipedia articles that a given article links to. The number of Out-links also describes the relation to other articles.

**Text Length** This feature describes the length in characters of the text in a Wikipedia article. We find that articles regarding resources that are of public interest, tend to be longer.

Given this set of features and the responses provided by the crowd workers and experts, we analyzed how the features relate to the inter-annotator agreement and whether or not the correct relationship predicate was determined. Table 2 shows the averages of the mentioned features, corresponding to correctly and incorrectly determined predicates by crowd workers (Crowd I) in the high-context Environment I.

We can see that the entities involved in answers that were correctly determined ('Correct Entities' in Table 2) by Crowd I tend to have a lower *depth* in the category Graph. This indicates that relationship predicates, between URI pairs represesnting more general topics are easier to establish in this specific environment. In the case of *in-links* and *out-links* we can see some major differences between the entities of correctly and incorrectly established predicates. The entities corresponding to correctly determined predicates have nearly twice the amount of in and out-links on

Table 3
Features of correctly and incorrectly determined predicates based on responses from the experts in Environment I.

| | Depth | In-links | Out-links | Text Length |
|---|---|---|---|---|
| Correct Entities | 5.58 | 2554.13 | 128.81 | 26,707 |
| False Entities | 5.87 | 380.97 | 61.07 | 13,579 |

Table 4
Features of correctly and incorrectly determined predicates based on responses from the crowd workers in Environment 2.

| | Depth | In-links | Out-links | Text Length |
|---|---|---|---|---|
| Correct Entities | 3.88 | 1171.3 | 115 | 13,638 |
| False Entities | 5.64 | 2845.65 | 146.73 | 33,178 |

average. In addition, we find the corresponding article size ('Text length' in Table 2) of these entities was generally larger. The differences between these sets are also statistically significant with $p < 0.05$ for depth, in-links and text length, with $p < 0.1$ for the out-links.

When looking at the same set of features calculated based on the responses from the experts in Environment I, we observe very similar patterns. These findings are presented in Table 3. Here in case of the Experts I, the differences between the features corresponding to the entities whose predicates have been either correctly or incorrectly determined, are even larger. Again we find that the differences for depth, in-links and text length are statistically significant with $p < 0.05$, while that for out-links is not significant.

We conducted the same series of experiments with respect to the responses received from crowd workers and experts in the low context Environment II. Table 4 presents our findings corresponding to Crowd II.

We observe a similar pattern as in case of Crowd I, wherein the *depth* corresponding to entities whose predicates are correctly determined ('Correct Entities' in Table 4) is lesser than that corresponding to the incorrectly established predicates ('False Entities' in Table 4). These differences are also found to be statistically significant with $p < 0.05$. In case of *text length* and the *in-links* we also found significant differences between Correct Entities and False Entities. However, in contrast to our findings in Environment I, on average the False Entities correspond to longer texts and a higher number of in-links. We observe that this is caused by the fact that even relatively easy to establish URI predicates (as indicated by the *depth*), were incorrectly determined by crowd workers in the low context Environment II.

The results from the experts in Environment II are reported in Table 5. In case of Experts II, we see

Table 5
Features of correctly and incorrectly determined predicates based on responses from the experts in Environment 2.

| | Depth | In-links | Out-links | Text Length |
|---|---|---|---|---|
| Correct Entities | 5.5 | 3212.64 | 167.71 | 37,509 |
| False Entities | 5.33 | 1993.37 | 121.37 | 25,421 |

that most of the features depict lesser differences between entities whose predicates are correctly and incorrectly established. We found statisically significant differences for the out-links and the text length. Once again we observe that longer Wikipedia articles corresponding to the entities, with more out-links seem to be more general and therefore the accurate predicates are easier to identify.

### 6.2. Familiarity Estimation for different Topics

In the experiments described in the last section we focused on relatively general features corresponding to the entities represented by the URI pairs. In this section, we further analyze how the categories covered by these entities influence the quality of the responses from crowd workers and experts.

In order to carry out this analysis, we generated profiles for each entity describing the topics, respectively *DBpedia categories*, the entity is related to. We used the Wikipedia/DBpedia category graph to generate profiles which relate entities to top level categories. An entity is considered to be related to a category if there exists a direct path, following parent category relations, from the categories where the article belongs to, up to the top categories. These categories to which the entity is thereby related to, are reffered to as *topics*.

Since the Wikipedia category graph is very dense we only follow links to parent categories where the distance to the root of the category graph is not getting longer. Additionally we took the number of parent categories and the number of steps into account when calculating the weight for each category. A more detailed description on how these entity profiles are built can be found in our previous works ([11] and [22]).

For the sake of understanding the topical variance pertaining to different entities, let us consider two example entities, namely *Bees* and *Music festival*. In Figure 8 we can see that both these entities depict distinct patterns with respect to their topical relevance; while Bees are more related to the categories *Life* and *Agriculture* the entity *Music festival* is mainly related to *People*, *Culture*, *Arts* and *Society*.

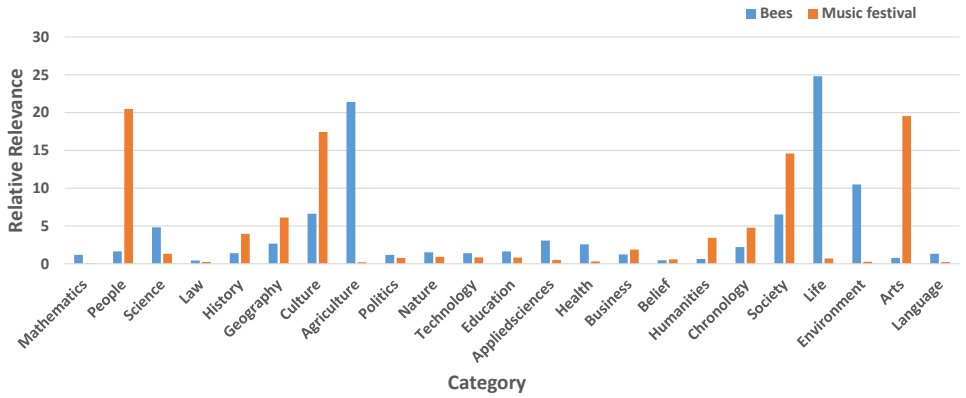



Fig. 8. Two example enitites considered ('bees' and 'music festival'), and their corresponding weights for different topics.

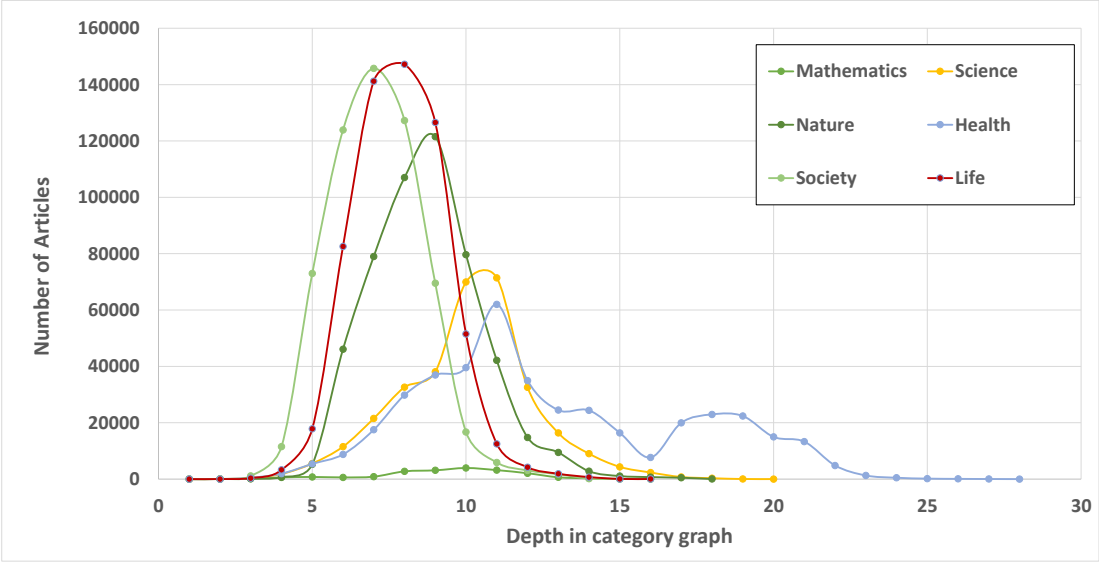



Fig. 11. Comparison of different Wikipedia categories and the depth of related articles in the category graph

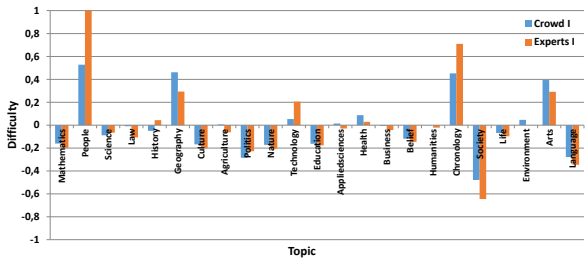



Fig. 9. Difference between *topics* of correctly and incorrectly resolved URI pairs in Environment I, and the corresponding *difficulty*.

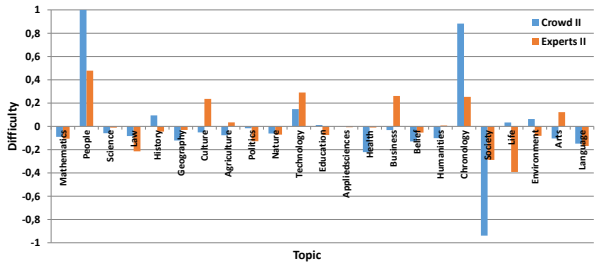



Fig. 10. Difference between *topics* of correctly and incorrectly resolved URI pairs in Environment II, and the corresponding *difficulty*.

With an aim to analyze how the topics may influence the quality of responses gathered from the crowd workers and the experts, we separated the correctly and incorrectly determined URI pairs and generated average profiles over the corresponding entities. Based on these profiles we calculated the difference between

correctly and incorrectly resolved URI pairs. We normalize this difference to fit the scale of $[-1, 1]$, where '-1' represents a topic wherein predicates between the URIs are very easy to resolve, and '1' represents the topic corresponding to the highest *difficulty* in resolving entities. The result is shown in Figure 9. We can see that for some categories there are major differences between the correctly and incorrectly established predicates. The most difficult topics are *People, Geography, Chronology and Arts*. On the other hand, the entities related to *Society* and *Language* seem to be easier and therefore predicates pertaining to such URI pairs were predominantly correctly resolved. It is interesting to note that the performance of the crowd workers and the experts are very similar. The correlation between the two profiles is 0.93 indicating that both groups incorrectly resolved predicates for entities related to the same categories.

We carried out the same experiments with respect to Environment II. Our findings are presented in Figure 10. We find that in case of the results from the crowd workers (Crowd II), the difficult topics are again *People* and *Chronology* while the easier one appears to be *Society*. In case of the experts (Experts II) however, the results differ when compared to Environment I. While *People* still is one of the most difficult topics, other categories seem to be similarly difficult in Environment II. This observation might be explained with the fact that Environment II does not show contextual information, increasing difficulty for certain kind of entities.

A closer look at the data reveals that while in Environment I, the differences for all topics with values larger than 0.1 are significant with $p < 0.05$ for both Crowd I and Experts I, in Environment II only *Language* and *Mathematics* show significant differences in case of the experts. We find statistically significant differences corresponding to the Crowd II with respect to the Values of *Society, Life, Chronology and People* corresponding to the correctly and incorrectly resolved predicates between URI pairs.

Besides the fact that some *topics* may be more difficult than others, the features may also vary for different topics. Within Wikipedia different topics are covered in different *depths* and due to this reason features such as *in-links*, *out-links* or the *depth* in the category graph may vary over the different topics.

Figure 11 shows an example of 6 different Wikipedia top categories and how deep the related articles are within the Wikipedia category graph. The 6 example categories show that the distribution of the articles can vary from topic to topic. While articles related to *Society* or *Life* show relatively similar distributions with an average depth of 7.1 and 7.87, other categories show distributions that have a wider spread over the category graph. Wikipedia articles representing entities related to *Science* or *Health* have an average depth of 10.03 and 12.7 respectively.

## 7. Related Literature

In this section we first discuss the confluence of the fields of Crowdsourcing and the Semantic Web in order to motivate our work and build a vision for the implications of our work. Then we describe the relevant previous works within two distinct realms; (a) Link Prediction and Schema Mapping, since in our work we focus on this particular task, and (b) Crowds and Experts, since our work aims to investigate the extent to which crowds can replicate the performance of experts in the Semantic Web related task considered here.

### *7.1. Crowdsourcing and the Semantic Web*

Due to the nature of Semantic Web technologies, the need for human input or intervention is apparent. In recent times, *crowdsourcing* and *gamification* have been adopted as a means to solve several problems emerging in the Semantic Web and engage workers within such tasks.

Ontology alignment is one domain that has profited from dedicating atomic microtasks to the crowd. Demartini et al. leverage crowdsourcing techniques with an aim to achieve large-scale entity linking [3]. Sarasua et al. introduced CrowdMap [20], a model to acquire human input via crowdsourcing in order to solve the problem of ontology alignment. In other works, researchers have used gamification in order to receive human input. Thaler et al. developed a game called SpotTheLink with an aim to solve the problem of ontology alignment [27]. Thaler et al. also used gamification for image annotation and interlinking. Keeping up with the tradition of 'semantic games with a purpose' [23], Markotschi et al. developed a game in order to receive contributions to the creation of formal domain ontologies from Linked Open Data [12]. Complementing these existing works on one hand, and in contrast to their implications on the other hand, in our work we focus on the comparison between employing crowd workers as opposed to experts in order to acquire human input in the task of *link prediction* and

*schema mapping*. Acquiring input from experts is an expensive endeavor. Hence, we investigate the feasibility of replicating expert judgments by exploiting the collective wisdom of crowd workers.

### *7.2. Link Prediction and Schema Mapping*

As described earlier, the task of *link prediction* and *schema mapping* refers to the creation of explicit links between resources on the Semantic Web at the instance and schema level. We discussed that owing to the following two primary reasons; (i) sparsity of additional metadata that can potentially aid in determining appropriate relations (i.e., the available context), and (ii) possible ambiguity between resources requiring disambiguation, human input is crucial with respect to the link prediction and schema mapping task.

In the last few years, there has been considerable work in the field of link discovery and prediction [14,15,16]. While these state-of-the-art works by Ngomo et al. tackle time-efficient execution of link specifications [14,15], the authors acknowledge that such automatic approaches for link prediction have a scope for improvement in terms of the accuracy [16]. This indicates the necessity of human input in order to further improve the accuracy in link prediction. We thereby choose to investigate the quality of human annotations that can be acquired in this regard, by using crowd workers on one hand, and experts in the field of Linked Data and Semantic Web on the other.

A quite established and excessively related field to the implications that our work has, comes from the *ontology alignment* [6] or *schema mapping* field. The main studied concepts in this field, range from finding direct mappings of type `1-1, 1-n, n-1` etc., that is one to one, one to many and many to one. An example of such a mapping might be matching literals that are assigned to a predicate `name` to a multiple predicates of another resource, i.e., *first name* and *last name*. In [17] crowdsourcing is used to determine which similarity measure to use for the alignment task. An interesting work in this direction has been carried by Albagli et al. [1]. The approach uses Markov Networks to combine the different factors that can be included on a ontology alignment task, e.g. *human experts, existing mappings* etc. A non-negligible aspect in the ontology alignment task, that of scalability, has been addressed partially by Duan et al. [4]. The authors use the local sensitivity hashing technique to group instances coming from an ontology, such that the number of considered pairs for the alignment is drastically reduced. However, in all cases the related work in this field even though it tackles a similar problem, the main differences lie in that fact that we generate *typed* mappings between instances. The attached semantics to the matched instances that range from `rdfs:subClassOf, skos:broader` and so forth, require a considerably large human input. This will help in creating automated approaches to tackle the problem of schema mapping automatically, and with a higher accuracy.

### *7.3. Crowds and Experts*

Over the last decade, there has been some work in the direction of combining the 'wisdom of the crowd' with knowledge from experts in order to overcome problems emerging from noisy crowdsourced labels apart from other machine learning settings [10,18]. Similarly, Sordo et al. use expert based classifications and crowd wisdom in order to improve music genre classification [25]. *Nichesourcing* has been introduced by de Boer et al., to combine the strengths of the crowd with those of professionals and experts, thereby improving the outcome of human-based computation for certain appropriate tasks [2]. Our work differs from such previous works in that we do not aim to combine the input from crowd workers and experts, instead we investigate the extent to which expert-level judgments can be acquired from the crowd, in the specific task of link prediction and schema mapping. While the previous works mentioned here, have credibly leveraged the combination of expert knowledge and the collective wisdom of the crowd, we acknowledge the fact that expert judgments are not readily available and require more resources in terms of time and costs.

Hill and Ready-Campbell show that user-generated content can be used to extract crowd wisdom in order to forecast stocks [9]. In contrast to our work, the authors identify and rank 'experts' within the crowd in order to facilitate better stock picking decisions. The authors consider users contributing to online stock voting markets as crowd workers and thereby do not consider an incentivized crowd, as in a typical crowdsourcing setting. In our work, we consider a distinct group of experts and incentivized crowd workers. In a closely related work, Sjöberg compared political election forecasts determined by experts (journalists, political scientists, and editors of readers' letters) and the crowd (general public) [24]. This work however, is limited to identifying that the public was most successful in forecasting the election results, thereby val-

idating the 'wisdom of the crowd'. In addition, while the work by Sjöberg considers a non-incentivized setting for the crowd, we compensate the crowd workers monetarily for their contributions. An incentivized crowd is a realistic preamble for crowdsourcing Semantic Web tasks such as the one tackled in our work here. In a different domain, Shankar et al. show that their crowdsourced location-based service performs comparably to expert-based state-of-the-art approaches [21]. Unlike in our work, here again, authors consider the users of mobile social networks such as Twitter, FourSquare and Facebook Places as the crowd and the user-generated content as the crowdsourced results. Importantly in our work, we quantify the expertise of crowd workers in order to identify the extent to which the crowd can replicate the performance of experts in a specific scenario.

## 8. Discussion and Limitations

From our experiments across the two setups, we note that the highest accuracy is achieved by experts in the Environment I (see Figure 4). The highest inter-annotator agreement is observed between the experts in the Environment II, closely followed by the experts in Environment I. With respect to **RQ1** and **RQ2**, Figure 4, illustrates that the performance of crowd workers is comparable to that of experts under circumstances that are conducive for crowd workers, i.e., as in Environment I with a high context that is immediately available at the workers' disposal.

We observe that even in this case, experts slightly exceed the performance of crowd workers. Crowd workers fail to perform well in the case where the setup presents limited information on which workers can directly rely upon (low context in Environment II). While the experts exhibit high agreement in both environments, crowd workers exhibit poor inter-annotator agreement in Environment II. This also underlines the impact of the task environment (**RQ3**) on the performance of crowd workers.

Through our experimental results in Environment I and II, we observe that `owl:sameAs` is cumulatively the least erred-on predicate, i.e., on average crowd workers (Crowd I, Crowd II) and experts (Experts I, Experts II) can most easily resolve the URI pairs with `owl:sameAs` relationship predicate. In addition, we see that the URI pairs with `owl:sameAs` predicate are typically erred-on by crowd workers and experts by misattributing the `owl:equivalentClass` predicate to the URI pair instead. On the other hand, we note that `owl:equivalentClass` is cumulatively the most erred-on predicate, i.e., on average crowd workers (Crowd I, Crowd II) and experts (Experts I, Experts II) cannot easily resolve the URI pairs with `owl:equivalentClass` relationship predicate. However, we do not observe a clear trend in the nature of the errors committed on URI pairs with `owl:equivalentClass` predicates, with respect to their distribution across the considered relationship predicates (the errors seem to be evenly distributed across the other 3 predicates considered in the experiments).

Through our experiments regarding **RQ4**, i.e. the estimation of the *difficulty* in accurately identifying the predicates corresponding to a URI pair, we find statistically significant patterns that explain the influence of the topic that the URIs represent on the performance of the crowd workers and experts. We find that in the high context Environment I, the crowd workers (Crowd I) and experts (Experts I) rely on the context available to adjudicate the apt predicates corresponding to URI pairs. Their familiarity with the entities thereby plays a key role in their performance. In the Environment I, features that are indicators of low *difficulty* are (i) low depth in the Wikipedia category graph of the entities (since a lower depth in the Wikipedia category graph indicates that the topic is more general and less specific), (ii) high number of in-links and out-links (since a high number of in-links and out-links indicates the popularity of the entity, and thereby the potential familiarity with it), and (iii) high text length (since popular entities tend to have longer corresponding Wikipedia articles describing them). In the low context Environment II however, we note that a similar pattern does not hold across the Crowd II and Experts II. We believe that this is due to the low context nature of Environment II.

## 9. Conclusions and Future Work

In this work, we have investigated the performance of crowd workers and experts in the task of link prediction and schema mapping. We delved into the impact of the task environment (in terms of context that is directly presented to a participant), and difficulty of the task, on the performance of crowd workers and experts. In addition, we present a detailed analysis on the types of errors that crowd workers and experts are

prone to committing, in the specific task of link prediction and schema mapping.

We show that the task of link prediction and schema mapping can be reliably crowdsourced. We find that the crowd workers perform comparably with respect to the experts, when sufficient context regarding the URI pairs is provided (as in Environment I). In the absence of adequate context however, we find that the crowd workers perform poorly in comparison. Therefore, the task design plays a crucial role in the quality of the results produced by the crowd workers in comparison to the experts. Although the experts outperform the crowd workers in both environments that we have considered in our experiments, we note that the crowd workers are capable of producing high quality results as well. This implies that one can save valuable resources in terms of time and costs, by employing crowdsourcing for this task. We conclude that it is feasible to use crowd workers for the task of link prediction and schema mapping, by paying due importance to the design of the task. We thereby answer the research questions **RQ#1**, **RQ#2**, and **RQ#3** that we set out to solve through this work.

Finally, we propose a method to quantify the expertise that is required to accurately identify the apt relationship predicate between a pair of URIs. We leverage the Wikipedia category graph and features emerging from it, in order to determine the level of *difficulty* in resolving the apt predicate for a given URI pair. We show that there is a strong influence of the *difficulty* on the performance of crowd workers and experts. In cases where the difficulty is high, crowd workers and experts tend to err. We thereby answer the **RQ#4**.

In the imminent future, we plan to extend our work by quantifying the expertise of crowd workers as a function of the *difficulty* in solving a particular task. In addition, we plan to empirically extend our work presented here to assess the extent to which experts can be replaced by crowd workers in different tasks within the taxonomy of crowdsourcing microtasks [7], in order to understand task specific influences.

## 10. Acknowledgements

This work has been carried out in the context of DURAARK, funded by the European Commission within the 7th Framework Programme (Grant Agreement no: 600908).

## References

[1] S. Albagli, R. Ben-Eliyahu-Zohary, and S. E. Shimony. Markov network based ontology matching. In *IJCAI 2009, Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, California, USA, July 11-17, 2009*, pages 1884–1889, 2009.

[2] V. De Boer, M. Hildebrand, L. Aroyo, P. De Leenheer, C. Dijkshoorn, B. Tesfa, and G. Schreiber. Nichesourcing: Harnessing the power of crowds of experts. In *Knowledge Engineering and Knowledge Management*, pages 16–20. Springer, 2012.

[3] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *WWW*, pages 469–478, 2012.

[4] S. Duan, A. Fokoue, O. Hassanzadeh, A. Kementsietsidis, K. Srinivas, and M. J. Ward. Instance-based matching of large ontologies using locality-sensitive hashing. In *The Semantic Web - ISWC 2012 - 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I*, pages 49–64, 2012.

[5] C. Eickhoff and A. P. de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval*, 16(2):121–137, 2013.

[6] J. Euzenat and P. Shvaiko. *Ontology Matching, Second Edition*. Springer, 2013.

[7] U. Gadiraju, R. Kawase, and S. Dietze. A taxonomy of microtasks on the web. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 218–223. ACM, 2014.

[8] U. Gadiraju, R. Kawase, S. Dietze, and G. Demartini. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of CHI'15*, 2015. To appear.

[9] S. Hill and N. Ready-Campbell. Expert stock picker: the wisdom of (experts in) crowds. *International Journal of Electronic Commerce*, 15(3):73–102, 2011.

[10] H. Kajino, Y. Tsuboi, I. Sato, and H. Kashima. Learning from crowds and experts. In *Proc of the Human Computation Workshop*, pages 107–113, 2012.

[11] R. Kawase, P. Siehndel, B. Pereira Nunes, E. Herder, and W. Nejdl. Exploiting the wisdom of the crowds for characterizing and connecting heterogeneous resources. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media, Santiago, Chile*, pages 1–4, 2014.

[12] T. Markotschi and J. Völker. Guess what?! human intelligence for mining linked data. 2010.

[13] W. Mason and D. J. Watts. Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter*, 11(2):100–108, 2010.

[14] A.-C. N. Ngomo. On link discovery using a hybrid approach. *Journal on Data Semantics*, 1(4):203–217, 2012.

[15] A.-C. N. Ngomo and S. Auer. Limes: a time-efficient approach for large-scale link discovery on the web of data. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three*, pages 2312–2317. AAAI Press, 2011.

[16] A.-C. N. Ngomo and K. Lyko. Eagle: Efficient active learning of link specifications using genetic programming. In *The Semantic Web: Research and Applications*, pages 149–163. Springer, 2012.

[17] P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang,

J. Z. Pan, I. Horrocks, and B. Glimm, editors. *The Semantic Web - ISWC 2010 - 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I*, volume 6496 of *Lecture Notes in Computer Science*. Springer, 2010.

[18] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *The Journal of Machine Learning Research*, 11:1297–1322, 2010.

[19] J. Rogstadius, V. Kostakos, A. Kittur, B. Smus, J. Laredo, and M. Vukovic. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *ICWSM*, 2011.

[20] C. Sarasua, E. Simperl, and N. F. Noy. Crowdmap: Crowdsourcing ontology alignment with microtasks. In *International Semantic Web Conference (1)*, pages 525–541, 2012.

[21] P. Shankar, Y.-W. Huang, P. Castro, B. Nath, and L. Iftode. Crowds replace experts: Building better location-based services using mobile social network interactions. In *Pervasive Computing and Communications (PerCom), 2012 IEEE International Conference on*, pages 20–29. IEEE, 2012.

[22] P. Siehndel and R. Kawase. Twikime! user profiles that make sense. In *Proceedings of the ISWC 2012 Posters Demonstrations Track: ISWC 2012 Posters Demos*, pages 61–64, 2012.

[23] K. Siorpaes and M. Hepp. Games with a purpose for the semantic web. pages 50–60, 2008.

[24] L. Sjöberg. Are all crowds equally wise? a comparison of political election forecasts by experts and the public. *Journal of Forecasting*, 28(1):1–18, 2009.

[25] M. Sordo, O. Celma, M. Blech, and E. Guaus. The quest for musical genres: Do the experts and the wisdom of crowds agree? In *ISMIR*, pages 255–260, 2008.

[26] J. Surowiecki. *The wisdom of crowds*. Anchor, 2005.

[27] S. Thaler, E. P. B. Simperl, and K. Siorpaes. Spotthelink: A game for ontology alignment. In *Wissensmanagement*, pages 246–253, 2011.

[28] M. Zook, M. Graham, T. Shelton, and S. Gorman. Volunteered geographic information and crowdsourcing disaster relief: a case study of the haitian earthquake. *World Medical & Health Policy*, 2(2):7–33, 2010.