World Wide Web (2014) 17:85–104
DOI 10.1007/s11280-012-0193-1

Assessing the suitability of an honest rating mechanism for the collaborative creation of structured knowledge

Conny Kühne · Klemens Böhm

Received: 19 October 2011 / Revised: 6 November 2012 /
Accepted: 12 November 2012 / Published online: 29 November 2012

Abstract Creating and maintaining semantic structures such as ontologies continues to be an important issue. The approach investigated here is to let members of an online community create structured knowledge collaboratively and to use ratings to evaluate the data created. Obviously, such ratings have to be of high quality. Honest rating mechanisms (HRMs) known from literature are a promising means to gain such high-quality ratings. However, the design of such mechanisms for collaborative knowledge creation and their effectiveness have not been studied so far. To evaluate the effects of an HRM on rating quality in this context, we have conducted several experiments with online communities. We find that an HRM increases rating quality and "punishes" rating errors. We also find that rating-based rewards increase the quality of the structured knowledge created.

Keywords incentives · structured knowledge · peer production

1 Introduction

Structured knowledge, e.g., in the form of ontologies, has become increasingly important. The question of how to create it on a larger scale, however, continues to be a fundamental research issue. The approach investigated in this paper is web-based peer production [4]. It is a decentralized production process where contributors work on a project without a hierarchical organization. It is both scalable when adding further users and robust because individual users can be replaced.

C. Kühne (✉) · K. Böhm
Institute for Program Structures and Data Organisation, Karlsruhe Institute of Technology, Am Fasanengarten 5, Building 50.34, 76131 Kalrsruhe, Germany
e-mail: conny.kuehne@kit.edu

K. Böhm
e-mail: klemens.boehm@kit.edu

When a community of peers, without a coordinating authority, creates and maintains structured knowledge, two questions arise: (i) How to motivate users? Many online communities suffer from under-contribution [3], i.e., only a minority contributes. (ii) How to ensure and assess data quality? Compared to, say, Wikipedia, quality assurance may be even more important for structured knowledge, which is used for query processing or automated reasoning.

In this paper, we study how rating-based incentive mechanisms can assure the quality of the structured knowledge created collaboratively. We envision the following real-world scenario: A community creates structured knowledge collaboratively. Its members review the contributions of each other and rate them according to the quality perceived. For instance, think of a project with partners who are geographically distributed, and who all contribute to a common knowledge base. To motivate users to contribute, they receive rewards according to the number and quality of their contributions. The quality of contributions, in turn, is computed based on the ratings. Finally, the points gathered are converted into external rewards, e.g., gift coupons as with Epinions or system privileges as with Slashdot. Ratings have to be of high quality as well. A promising approach to gain high-quality ratings are *honest rating mechanisms* (HRMs) [18, 23, 27]. An HRM rewards subjective truthfulness in scenarios where no objective truth criterion is available. To the best of our knowledge, these mechanisms have not been studied yet in the context of collaborative knowledge creation.

In this paper, we study the following research questions regarding the creation of structured knowledge.

1. How do reward mechanisms for contributions as well as for ratings influence user behavior?
2. Does an HRM lead to ratings of higher quality, compared to a fixed reward per rating?
3. How do rewards for contributions dependent on ratings influence the quality and the number of contributions, compared to rewards that are fixed?

To this end, we have developed a platform for the collaborative creation of structured knowledge called *Consensus Builder* (CB). It features fine-grained rating and incentive mechanisms, in particular an HRM, and is operational in a real-world environment. Based on the research questions, we formulate a number of hypotheses and test them in a series of controlled field experiments. The setups are close to the real-world scenario envisioned. Controlled experiments allow us to gain insights into causal effects of the tested mechanisms. This cannot be achieved by observational studies. For each experiment, we have recruited a different community. Participants have used CB from home or from their workplace to create structured data.

We make the following contributions:

– We describe the features of Consensus Builder, our platform for the collaborative construction of structured knowledge. It contains reward mechanisms for contributions and for honest ratings.
– We discuss how reward mechanisms can be applied to the creation of structured knowledge.
– Based on our research questions, we formulate hypotheses and design experiments to test them.

– We extensively test the reward mechanism w.r.t. the creation of structured knowledge in different settings and with participants of different backgrounds.

A main finding of ours is that an HRM leads to ratings of equal or higher quality, compared to static rewards. Further, we find that rating-dependent rewards for contributions do improve contribution quality, but result in fewer contributions compared to a fixed reward per contribution.

Paper outline: Section 2 reviews related work. Section 3 presents our collaboration platform. We discuss the HRM in Sect. 4. Section 5 states our hypotheses. Section 6 discusses the experimental setup and sections 7 and 8 present the results of the experiments. Section 9 concludes.

2 Related work

There exists a large body of work on motivation and collaboration in online communities. For example, the GroupLens research group[1] has conducted various experiments in online communities for movie recommendations. To study the question how to collect more contributions, they have deployed theories from social science [3, 9]. Others investigate collaboration and motivation in Wikipedia [20, 36, 40]. None of these publications investigates incentives to rate the quality of contributions.

Several well-known methodologies for ontology engineering incorporate collaboration mechanisms [15, 22, 24, 39]. The HCOME methodology for instance proposes a decentralized engineering model where knowledge workers develop ontologies individually and, in a later step, share it with the community for revision and merging [22]. However, neither of these methodologies is explicitly concerned with mechanisms for user motivation. Braun and colleagues introduce a more lightweight Web 2.0 based methodology for ontology engineering [6]. Their community-driven approach relies on the maturing of ideas and concepts from tags into formal ontologies. Further, they discuss intrinsic motivations users might have to engage in the maturing activities, such as future retrieval or contribution and sharing. Their approach might benefit from explicitly incentivizing users, e.g., using the mechanisms presented and evaluated here.

Various techniques have been proposed for creating ontologies from text sources [7] as well as from semi-structured sources such as UML class diagrams [42] or tables [34] automatically. Even though such automated approaches to ontology learning can support the ontology engineering process to a large degree, ontology development remains a human-driven process [31].

Various tools for the collaborative creation of structured knowledge have been proposed, ranging from full-fledged ontology editors with collaborative features [33, 35] over Wiki-based approaches for semantic data [2, 8, 28, 32, 37] to tools that support tagging folksonomies [17, 43]. Some of these tools feature rating mechanisms. All this work does not include systematic attempts to evaluate the effect of ratings on knowledge quality for these tools. There also exist commercial tools for the collaborative creation of structured knowledge, notably Freebase [5]. For a detailed overview of tools for the collaborative construction of structured knowledge,

[1] www.grouplens.org

as well as a discussion of the general requirements and alternatives for such tools, see [25]. Further, Siorpaes and Simperl analyze tools and methodologies for semantic content creation and identify tasks that are inherently human driven [31].

[16] studies rating based incentive mechanisms for the collaborative construction of structured knowledge. The authors conduct two controlled experiments and find that ratings are a reliable measure of contribution quality. Further, they find that the presence of ratings increases the quality of contributions compared to a setting without. However, they do not test the effect of incentive mechanisms for ratings on rating quality and user behavior.

Eckert et al. use input from Amazon Mechanical Turk (AMT) workers to construct *is-a* relations between pre-selected terms in a philosophy knowledge domain [12]. They propose a redundancy-based method to achieve high-quality results from the input of AMT workers. As opposed to our approach, they included concept pairs for which they could objectively determine a correct answer, i.e., a gold standard, to identify well-performing AMT workers. Further, their method, as any AMT-based scheme, is only applicable to domains known to the general public. Only such domains can attract a sufficient number of AMT workers with the necessary understanding.

Von Ahn uses games to motivate users to perform useful tasks [38]. For example, users in the ESP game are randomly paired to create tags for an image and receive points whenever their tags match. Siorpaes and Hepp [30] use this principle to build ontologies from Wikipedia entries by categorizing entries as either classes or instances. Users receive points when they agree on the categorization. Again, these approaches are better suited for knowledge domains that are known to the general public and where data is already available, e.g., in form of Wikipedia entries. Next, the game of assigning Wikipedia entries to predefined categories and rewarding users based on answers by other users is essentially an HRM setting. Thus, scaling the rewards for agreement by means of an HRM could give way to better results with these games. Similarly to the games scenario of von Ahn, Kochhar et al. collect human input for decisions in the Freebase knowledge base that an automated process cannot decide [21]. They make use of paid contractors and to a lesser degree of volunteers to act as judges for these non-automatable decisions. To increase the quality of the judgments, they cultivate long-standing relationships with their judges. In addition, they identify judges with id and profile. The authors describe these relationships as the main reason for the good quality of the judgments. Thus, their results are unlikely to be transferable to typical online communities with an anonymous character. Like the games scenario described above the judge decision could potentially benefit from the use of an HRM.

3 Creating structured knowledge with Consensus Builder

This section describes our tool Consensus Builder (CB) that allows for the collaborative creation of structured knowledge.[2]

[2]See http://consensus.ipd.uka.de/wwwj-demo for a demo.

Data format CB uses a data format similar to Topic Maps [26] and Entity-Relationship Models with some slight deviations for better usability. Data can be created on the type and on the instance level. I.e., users can specify the schema and create instance data. Topics represent entities of the real world, e.g., `Harrison Ford` or `Indiana Jones`. Topics can have one or more types, e.g., `Harrison Ford` is of type `Person` and of type `Actor`. Types contain attributes and association types. Attributes describe simple data, like 'date of birth', and are constrained by data types, e.g., integer, string. Association types describe associations between topic types, e.g., `Actor <acts in> Movie`.

The objective of this paper is the design and deployment of incentive mechanisms for the collaborative construction of structured knowledge. In this specific context, the data format to encode the knowledge is of secondary concern. We have mainly chosen the data format described above because of its ease of use for non-expert users. The functionality of CB is applicable to other formats for structured knowledge as well, such as those specific to the Semantic Web. Furthermore, the data format currently used can be mapped to OWL in a straightforward way: Topic types are mapped to OWL classes, attributes to data-type properties, association types to object properties (as well as to their inverse properties) with domains and ranges restricted to the respective classes or data types, topics are mapped to instances, attribute values to literals, etc.

Collaborative editing and rating Users can create and change all parts of the data model collaboratively. They can create single data elements like attributes or topics and change elements created by others. When a type is added to a topic, it inherits the attributes and association types of that type. Users can then set attribute values and add associations. CB takes care that users can create only attribute values and associations that are valid regarding the type level. In addition, CB provides various functions for browsing and searching, comments to discuss topics and topic types, and statistics such as user scores.

CB features a fine-grained association between ratings and contributions, called *rating scheme*, that lets users assess the quality of contributions on a very detailed level. Users can rate every element on the type as well as on the instance level that can be manipulated, e.g., topic names, attributes and association types, and attribute values.

A rating of a contribution c refers to the correctness of c. The specific semantics differ depending on the contribution rated. For example, rating a type-instance relationship between type `Person` and topic `Harrison Ford` means evaluating the correctness of the statement "Harrison Ford is a Person". Rating the attribute `Date of birth` of type `Person` evaluates the correctness of the statement "Date of birth is an attribute of Person".

CB supports rating scales of arbitrary granularity. Since our rating scheme is very fine-grained, we prefer rating scales of lower granularity in order not to overburden the user. In general, we either use the well-known five-star rating scale or a binary rating scale ('low' vs. 'high'). Cosley et al. show that, even though users prefer a finer-grained rating scale, the granularity of the rating scale does not have an adverse effect on user behavior [10]. The functions for rating, editing, and displaying of the data are tightly integrated in the user interface (cf. Figure 1).

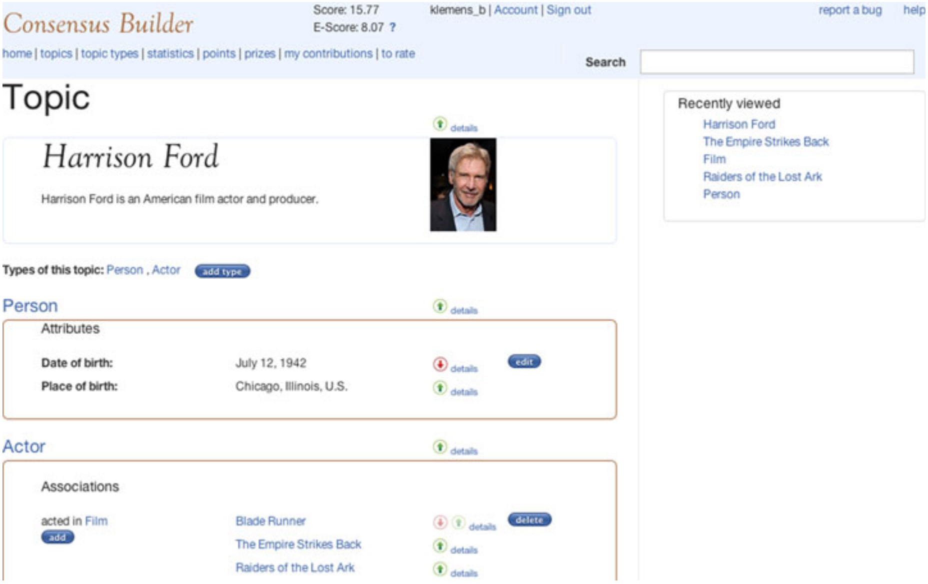

Figure 1 Details for topic ‘Harrison Ford’ in Consensus Builder.

Users can change and delete individual contributions. However, change operations are not trivial in any setting where data items depend on each other. For instance, what happens with other contributions and ratings associated with a contribution just deleted? Think of the deletion of a type that has associated topics and has received high ratings. To address these issues, we have made the following design decision. A contribution can only be changed if it satisfies two conditions. First, it must not be associated with other contributions, e.g., an attribute can only be changed if there are no attribute values associated with it. Second, it either must not have received any ratings, or its average rating value must be below a certain threshold. Consequently, only contributions deemed low quality can be changed. The user who has changed the contribution becomes its new owner.

Commission: rating-based remuneration To motivate users to create and maintain the data we reward data operations with points based on ratings by other users. We refer to the rating-based remunerations as *commission*. A user obtains a commission every time another user issues a positive rating for a contribution that first user is the owner of. The value of the commission is computed as $\kappa \cdot c(operation)$, where κ is a constant factor depending on the scenario, and $c(operation)$ depends on the operation. For instance, $c(\text{create topic}) = 3.0$, and $c(\text{set attribute value}) = 2.0$.

4 Honest rating mechanism

We want to elicit high-quality ratings, as opposed to ratings that are uninformed or simply copy the majority opinion. (This does not exclude the majority opinion from being correct.) However, a simple reward, e.g., one point per rating, does not suffice.

It does not provide an incentive for the rater to gather information before issuing her rating and to respond truthfully.

To address these challenges, we use an incentive mechanism that rewards truthful ratings [23]. The mechanism has originally been designed for online product ratings where buyers perceive a noisy signal of the quality of a product and rate it according to their perception. However, the mechanism is not constrained to product ratings. It is applicable to any scenario where agents receive a noisy signal of some states of the world. In our setting, the relevant states are the quality values of the contributions, i.e., their correctness. In line with the literature, we refer to the quality of the contribution as its type. The mechanism assumes that the type t of a given contribution is fixed, i.e., each contribution has a true quality that does not change over time. The number of types is finite and indexed by $t = 1, \ldots, T$. The mechanism assumes that all raters share a common prior belief regarding the distribution of types $Pr(t)$. The true type is hidden from the rater, who can only perceive a noisy signal. The rater i receives signal S^i from the set $S = \{s_1, \ldots, s_M\}$ of possible signals. Think of the signal as the relevant information the rater uses to form her subjective opinion about the quality of the contribution. Let $Pr(s_m|t) = Pr(S^i = s_m|t)$ be the probability that a rater receives signal s_m when the true type of the product is t. In other words, each type induces a distribution of signals. The mechanism assumes that different types induce different signal distributions. E.g., a low-quality type might induce low-quality signals with higher probability than a high-quality type. Conditional on the product's type, the signals of the raters are independent and identically distributed. The mechanism assumes $Pr(.|.)$ to be common knowledge, and $\sum_{j=1}^{M} Pr(s_j|t) = 1$.

After all raters have received the signals, the mechanism asks them to submit ratings according to their signals simultaneously. Let $r^i = (r^i(1), \ldots, r^i(M))$ denote the rating strategy of rater i. That is, i announces $r^i(j)$ whenever she observes signal s_j. The honest reporting strategy is $\bar{r}$ with $\bar{r}(j) = s_j$ for all $j \in \{1, \ldots, M\}$, i.e., the rater always reports the truth. Subsequently, the mechanism scores every rating submitted by comparing it to the rating of another rater, $ref(i)$, called the reference rater of i. Let $\tau(r^i(j), r^{ref(i)}(k))$ be the payment received by i when she announces $r^i(j)$ and $ref(i)$ announces $r^{ref(i)}(k)$.

The expected payment of rater i depends on her prior belief and on the signal s_j she has received:

$$\begin{aligned} V(r^i, r^{ref(i)}|s_j) &= E_{s_k \in S}(\tau(r^i(j), r^{ref(i)}(k))) \\ &= \sum_{k=1}^{M} Pr(S^{ref(i)} = s_k | S^i = s_j)(\tau(r^i(j), r^{ref(i)}(k))) \end{aligned}$$

The conditional distribution that $ref(i)$ receives the signal s_k can be computed as:

$$Pr(s_k|s_j) = \sum_{t=1}^{T} Pr(s_k|t) \cdot Pr(t|s_j) \tag{1}$$

The posterior probability $Pr(t|s_j)$ of type t given the perception of signal s_j, can be computed using Bayes' Law.

The honest reporting strategy $\bar{r}$ is a Nash equilibrium if and only if for all signals $s_j \in S$ and all reporting strategies $\hat{r} \neq \bar{r}$:

$$V(\bar{r}, \bar{r}|s_j) \geq V(\hat{r}, \bar{r}|s_j) \tag{2}$$

Miller et al. prove that payment schemes $\tau(.,.)$ for $V(.,.|.)$ which satisfy (2) exist. For example, one such payment scheme is the logarithmic scoring rule $\tau(r^i(j), r^{ref(i)}(k)) = ln(Pr(s_k|s_j))$. We use a linear program described in [18] to compute the payments.

The mechanism uses i's rating to update the probability distribution that predicts $ref(i)$'s rating. The score reflects the quality of the reference rating relative to the updated distribution. If the rating of the reference rater is honest, a rater can maximize her expected payment by announcing her subjective beliefs.

For illustration, consider the following simple example. There are two possible types *Bad* (B) and *Good* (G), with prior $Pr(G) = 0.7$, and two possible signals *low* (l) and *high* (h), with signal distributions $Pr(h|B) = 0.4$ and $Pr(h|G) = .8$. Therefore, $Pr(h) = Pr(h|B) \cdot Pr(B) + Pr(h|G) \cdot Pr(G) = 0.68$. Suppose that rater i has received a *high* signal. In this case, according to (1), her expectation of $ref(i)$ receiving a *high* signal is $Pr(S^{ref(i)} = h|S^i = h) = 0.73$. Using the logarithmic scoring rule, the resulting expected payments $V(r^i, r^{ref(i)}|h)$ for i is -0.58 if she announces a *high* signal and -0.64 if she announces a low signal. So telling the truth, i.e., announcing the *high* signal maximizes her expected payoff. Conversely, had i instead received a *low* signal her expectation of others receiving a *high* signal would become $Pr(S^{ref(i)} = h|S^i = l) = 0.575$. Her expected payoffs in this case are -0.68 and -0.74 for announcing *low* and *high*, respectively.

Note that the inventors of the HRM claim, that it is not necessary for users to do the rather complicated computations [23]. As long as they trust the mechanism, users will prefer to report honestly.

Even though honest reporting is the desired equilibrium strategy of the mechanism, it is not the only one. Other equilibria like rating always high or rating always low exist as well. [19] develops countermeasures against such lying coalitions. The countermeasures are based on increasing the number of reference ratings per rating and increasing the budget of the mechanism to offset incentives for lying coalitions. Systematic taste differences among raters also pose a potential threat to the proper functioning of the mechanism. For example, some raters might have contrarian views or might generally be harsher in their assessment of quality. Whether the problems of lying equilibria or taste differences occur without countermeasures in reality is an interesting question that our experiments will address as well.

In the following we describe the design decisions behind our implementation of the HRM: Let $Bin(n|q, N)$ denote a binomial distribution of getting n successes in N Bernoulli trials with success probability q. We assume that each type t generates a binomial distribution of signals with success probability $q = t/(T+1)$. I.e., type t generates a signal distribution with $Pr(s_m|t) = Bin(m-1|t/(T+1), M-1)$. For each contribution we maintain one type distribution $Pr(t)$. We also maintain a global type distribution that we use to initialize $Pr(t)$ of newly created items. We update $Pr(t)$ with the new ratings submitted. If someone changes an item, we reset its rating history and initialize its prior distribution with the global distribution at the time of change.

We put subsequent ratings of a contribution into groups of small size (typically 3 or 4) and score ratings in a group against each other. To motivate the user further, we display the sum of the expected scores for her unscored ratings.

5 Hypotheses

We formulate three hypotheses. They refer to the quality of either ratings or contributions. We measure quality with respect to a gold standard. I.e., a high quality rating coincides with the gold standard, whereas a low quality rating or rating error does not.

$\mathbf{H_{rate}}$ An HRM improves rating quality.

We deem high-quality ratings essential for the creation of high-quality knowledge. In previous work [16] we have shown that the presence of ratings has a positive impact on the quality of the structured knowledge. Here, we are explicitly concerned with the quality of ratings, for two reasons: (1) Ratings of high quality do allow assessing the quality of contributions. (2) If the ratings are of high quality it is relatively easy to filter out bad contributions and thereby increase the quality of the knowledge created. This cannot be done if ratings are of low or unknown quality.

$\mathbf{H_{comm}}$ Commissions improve contribution quality and reduce quantity, compared to fixed remunerations.

In the case of static rewards users are rewarded for every contribution, regardless of its quality. This is why we expect contributions to be fewer, but of higher value when they are rewarded contingent on ratings of other users.

$\mathbf{H_{err}}$ Rating errors receive lower scores from the HRM.

If this hypothesis is true the HRM functions correctly. In this case truthful and well-informed ratings lead to high scores. Additionally, if H_{err} is verified, this is evidence that there are no systematic taste differences and lying coalitions.

6 Experiments

To test our hypotheses we have conducted three experiments with CB. In each experiment, we randomly assigned participants to the experimental group (EG) or the control group (CG). We use the EG to evaluate the effects of the mechanisms in questions. The CG serves as the baseline. In the following, we first present the individual experiments. We then describe characteristics of the experimental setup common to all experiments, the different gold standards we use for quality assessment, and the statistical methods we use in our analysis.

RATEONLY In this experiment we focus exclusively on the HRM. We recruited participants among students of our chair and instructed them to rate 127 contributions. We had preselected these 127 contributions from a knowledge base which students had created for the domain “Karlsruhe Institute of Technology” in a creation phase prior to the experiment itself. The selected contributions remained embedded in the other contributions from the creation phase. But since we wanted to test the rating mechanism in isolation for this experiment, we disabled ratings for these other contributions, as well as data manipulations. This reduces effects not related to the HRM. For three days, participants rated the contributions using CB. To test H_{rate}, we scored participants in the EG with the HRM, while the CG was scored statically with one point per rating.

HONSTUDENTS We tested H_{rate} in a setting with the full functionality of CB. We invited students of the lecture "Machine Design" of the department of mechanical engineering. We told them to create topics and types which represent the content of the lecture and to rate the contributions of others. Again, we rewarded the EG by means of the HRM and the CG with one point per rating. To allow for comparing ratings of CG and EG later on, both groups had to rate contributions from the same set. For this reason, participants of both groups worked together on the same data. If the groups had used separate data, it would have been hard to say whether differences in rating quality result from the rating mechanism or from differences in the nature of the contributions.

HONSTAFF We repeated the experiment HONSTUDENTS to test H_{rate} in a setting close to that of a community within a company. For this run we invited researchers from the Institute of Product Engineering. As knowledge domain we used a model for the engineering design process developed by this institute [1]. We advised participants to use the elements of that engineering model as topic types and concrete instances as topics.

COMMISSION We designed the next experiment to test H_{comm}. Testing it in the experiments just described would have been difficult. This is because testing H_{comm} requires the CG and the EG to operate on separate data in order to eliminate potential influences between the groups and to allow for an unambiguous quality assessment of the contributions of each group. Such undesired influences are likely to occur in shared knowledge bases because data entries depend on each other, e.g., the contributions on the instance level depend on the schema level. Furthermore, according to H_{comm}, we expect a higher number of low quality contributions in the CG. This might affect the results even more if both groups operated on shared data. COMMISSION took place in a real world setting as well. Participants were students of the lecture "Database Systems". Again, we asked participants to model the content of the lecture and related information on the schema and instance level and to rate contributions of each other. To test H_{comm}, we choose *usage of commissions* as independent variable: The CG received a fixed amount per operation depending on the operation category, as specified by $c(operation)$. To prevent potential exploitation, this amount was deducted when the contribution was deleted. The EG received commissions contingent on the operation category and on ratings of other participants, i.e., the current owner received $\kappa \cdot c(operation)$ for every positive rating the contribution had received and 0 points for negative ratings. We set κ to 0.2. We rewarded ratings of both the CG and the EG by means of the HRM.

In each experiment described so far, at least one of the experimental groups used the HRM. This allows us to test H_{err}.

Experimental setups—discussion Our experiments go well beyond vanilla laboratory experiments, in several respects: They take place in real-world settings, within online communities where participants are not restricted by laboratory conditions. Unlike toy domains, the knowledge domains used were complex and had real-world significance. The participants used CB in an asynchronous fashion from home or from their workplace. We put attention to not letting them know that an experiment was taking place nor that there were different experimental groups, by announcing the experiments as "beta test and user study". (We introduced experimental features by means of announcements within CB.) Further, the assignment to groups was transparent to the participants, i.e., there were no indicators (e.g., specific URLs, etc.) that made the group explicit. The experiments lasted up to several weeks. This has blurred the distinction between real world and experiment further. Finally, participants remained anonymous to each other throughout the experiment and had no information about how many members their respective communities had. The two university courses we recruited participants from had an anonymous character as well. Both had a high number of students (up to 600). Further, one of the lectures (HONSTUDENTS) has been recorded and broadcast on the Internet, and most students chose to watch it from home. Thus, even though participants in these experiments were from the same course, we have not been aware of any personal interaction regarding the creation of structured knowledge. This impression was confirmed in personal discussions with participants who have come to our offices to collect their remunerations for participating. Thus, regarding the aspects that are relevant to the character of our study, the settings do not differ much from large online communities.

Further characteristics of the experiments To allow for a comparison with the CG, which was rewarded 1 point per rating in the experiments that test H_{rate}, we scaled the expected score of the HRM to 1.5 points. (Assuming risk-averse participants, we use 1.5 points instead of 1.) To motivate students to participate in COMMISSION and HONSTUDENTS, each *active* participant received a guaranteed compensation of 5 Euro. A participant was considered active if he had reached at least 30 points. For the N_e active participants of each experiment e we conducted a lottery over $N_e \cdot 10$ Euro. The lottery consisted of $2 \cdot N_e$ draws of 5 Euro each. Every full point counted as an individual lottery ticket. To motivate the researchers to participate in HONSTAFF we raffled off two digital cameras (Canon IXUS 105) and eight USB-Sticks.

For RATEONLY we paid each participant of the CG a fixed compensation of 6 Euro. Each participant i of the N participants in the EG received $(3 + \frac{points(i)}{\sum_i points(i)} \cdot N) \cdot 7$ Euro, where $points(i)$ is the overall score of i.

Prior to each experiment, participants could take part in a "training phase" to get accustomed with CB. The purpose of the training was to remove potential distortions of the experiments due to the learning curve. To aid learning we provided screenscasts that explained the usage of CB and the details of data modeling in particular. Next, participants could enter an experiment while it was already running. The domain of the email address constrained the registration to guard against sybil attacks, i.e., against single users creating multiple accounts to gain unfair advantages [11]. An algorithm based on biased coin randomization [13] assigned participants to either the CG or the EG, while keeping the numbers of members of the groups balanced.

After each experiment, we invited the participants to complete a questionnaire. It elicited feedback on rewards, ratings, rating mechanisms, the behavior of other participants, and the usability of CB. The number of questions per questionnaire ranged up to 24, dependent on the configuration for the respective participant. We used a five point Likert scale response format ('strongly disagree' to 'strongly agree') for most questions.

Rating scale and HRM parameters For all experiments, we use a binary rating scale, i.e., ratings are either low or high, in order not to overload the user. Accordingly, we have modeled signals to be in the set $\{l, h\}$. I.e., the binomial distribution we use to model the signal distribution for a given type degenerates to a Bernoulli distribution. We use types in the set $\{1, \ldots, 9\}$ because this resolution is sufficient for our purposes.

Gold standards for quality assessment Assessing the quality of contributions and ratings required several different gold standards. For HonStudents, HonStaff, and Commission we let domain experts rate a subset of contributions and used their ratings as gold standard. The subset depended on the hypothesis to test. Testing H_{rate} required comparing the rating quality between the CG and the EG. We randomly picked 150 contributions that had received at least one rating from both experimental groups. To test H_{comm} we simply picked 50 contributions randomly from each group. The experts used the same CB user interface as the participants to issue ratings. To understand the context, experts could see all contributions created in the respective experiment. For Commission, in addition to the detail ratings for randomly selected contributions such as attributes, associations etc., we let the experts assess the 'overall quality' and 'overall adequateness' of the topic or topic type associated with the respective contribution as a whole. This allows for a comprehensive quality assessment of the contributions. For the overall ratings we used a five-star rating scale instead of a binary one. We chose the following domain experts for the various experiments: the teaching assistant of the respective course for Commission and for HonStudents, and a scientist whose research topic is the engineering model that served as the domain for HonStaff. Since both domain experts had limited experience in data modeling for the latter two experiments, a database expert supported them with the data modeling.

For RateOnly we selected 127 contributions manually, out of the more than 5000 contributions created during the data-creation phase. The contributions selected were unambiguously correct, as confirmed by information publicly available on websites. We manipulated 34 out of the 127 contributions (all on instance level) so that they were false. The manipulated contributions together with the remaining manually selected ones served as the gold standard.

We classified the manipulations in three categories according to the effort needed to verify the respective errors:

1. *Easy to verify.* These are blatant errors, like a building having 666 elevators or a paper on sensor networks published in 1920.
2. *Medium effort to verify.* This category contained plausible-looking errors, like changes in room numbers or changes in co-authors of a paper. They could be detected by internet search.
3. *Hard to verify.* These manipulations were subtle and could only be verified with high effort, for example, the number of floors in a remote building.

We expect that the HRM has an effect on errors of Category 2 only. Both groups should recognize errors in Category 1. Category 1 allows checking whether

Table 1 Overview of experimental setups.

	RATEONLY	HONSTUDENTS	HONSTAFF	COMMISSION
Static ratings	CG	CG	CG	–
Honest ratings	EG	EG	EG	both
Static contrib.	–	–	–	CG
Commission	all	all	all	EG
Duration	3 days	3 weeks	2 weeks	3 weeks
Shared data CG/EG	yes	yes	yes	no
Gold standard	Manipulation	Experts	Experts	Experts

participants made any effort at all. For Category 3, the effort for error detection exceeds the benefit from honest ratings by much. It serves as an extra check to exclude the possibility that the EG was more motivated than the CG a priori.

Table 1 shows an overview of the different setups.

Statistical methods We use Pearson's χ^2 test to evaluate associations between binary variables, e.g., between classification errors and usage of the HRM. (For directional associations we use the one-tailed χ^2 test[14].) We use Student's *t*-test to compare the mean number of contributions. We compare the five-star ratings for overall quality by means of the Mann-Whitney U test [41]. We use Pearson's correlation to test the point-biserial correlation between binary rating errors and rating scores. Finally, we use Spearman's ρ to test the correlation between Likert responses from the questionnaire and experiment results.

7 Results

We present the results of our experiments, including the evaluation of the hypotheses and of the questionnaire. Table 2 shows a list of quantitative characteristics of the four experiments.

7.1 H_{rate}: an HRM improves rating quality

To test H_{rate} we compare the error rates of ratings rewarded with the HRM (EG) to those rewarded statically (CG). In the following, let $r \in \{0, 1\}$ denote a given rating,

Table 2 Overview of the experiments.

	RATEONLY		HONSTUDENTS		HONSTAFF		COMMISSION	
	CG	EG	CG	EG	CG	EG	CG	EG
Participants	3	6	8	12	10	10	11	14
Contributions	127*	127*	151	1052	136	162	802	206
Ratings	381*	762*	943	456	419	555	180	151
Ratings per contribution	3*	6*	0.78	0.38	1.4	1.86	0.22	0.73

*Number of contributions and ratings fixed.

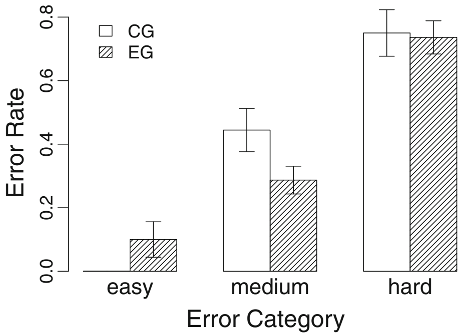

Figure 2 Error rate of ratings by error category.

and let $g(r) \in \{0, 1\}$ denote the gold standard of that rating.[3] We define the rating error $re(r)$ as an indicator variable

$$re(r) = \begin{cases} 1 & \text{if } r \neq g(r) \\ 0 & \text{otherwise.} \end{cases}$$

The *error rate of ratings (ERR)* is the proportion of rating errors out of all ratings, i.e., $\mathrm{ERR} = \frac{1}{|G|} \sum_{r \in G} re(r)$, where G is the set of ratings which we have a gold standard for.

For HonStudents there was a highly significant association between rating errors and the usage of the HRM ($\chi^2(1) = 71.52$, $p < 0.01$). The ERR was much higher for the CG (0.57) than for the EG (0.11). The odds ratio of making a rating error when using the HRM was 0.09. We conclude for this experiment that the mechanism improved rating quality.

For HonStaff we found no statistically significant association between rating errors and the usage of the HRM ($\chi^2(1) = 1.9071$, $p = 0.17$). The CG showed slightly better results regarding rating quality (ERR = 0.16) than the EG (ERR = 0.22). For HonStaff we conclude that there is no significant effect of the HRM on rating quality. A possible reason for this is that the researchers already had a high intrinsic motivation to create high-quality data since they wanted to use it in their research later on. Further, even though they did not interact outside of CB for the knowledge-creation task, they might have felt stronger obligations towards their relatively close-knit group. This high intrinsic motivation might have diminished the effects of the HRM.

Figure 2 shows the ERR for the three error categories of RateOnly for CG and EG respectively. The participants of the EG made significantly fewer errors in the "medium effort to verify" category (odds ratio = 0.5, $\chi^2(1) = 3.3$, one-tailed $p < 0.05$). For the other two error categories we found no significant association between errors and the usage of the HRM. The ERR for ratings of non-manipulated contributions was very low in both groups (CG: 0.054, EG: 0.043) and the association not statistically significant (two-tailed $\chi^2(1) = 0.27$, $p = 0.6$). We conclude that, for this experiment, the HRM increases rating quality.

Summing up, we find that two out of three experiments support $\mathrm{H}_{\mathrm{rate}}$.

[3]The gold standard of a rating simply is the gold standard of the contribution the rating belongs to.

7.2 H_{comm}: commissions improve contribution quality and reduce quantity

To test the "quality" part of H_{comm} we compare the five-star expert ratings for selected contributions rewarded with commission (EG) to those rewarded statically (CG). Figure 3 shows the distribution of expert ratings for quality and adequateness of CG and EG. The quality ratings for the contributions rewarded statically were significantly lower than for contributions rewarded with commissions (Mann-Whithney $U = 684$, $p < 0.05$). The ratings for adequateness of contributions rewarded statically were significantly lower as well, compared to those rewarded with commissions (Mann-Whithney $U = 549$, $p < 0.01$).

There were much fewer contributions per participant in the group using commissions ($mean = 27.0$, $se = 7.96$) than in the one without ($mean = 126.5$, $se = 67.27$). However, the difference was not statistically significant ($t(5.14) = 1.47$, $p = 0.10$). Interestingly we also found that participants remunerated with commissions seem to rate more critically. The ratio of negative ratings was significantly higher in the EG (0.258) than in the CG (0.039) ($\chi^2(1) = 31.2$, $p < 0.01$), even though the experts rated the contributions of the EG more favorably.

7.3 H_{err}: rating errors receive lower scores from the HRM

To test H_{err} we calculate the correlation between the rating error and the rating score by the HRM. A negative correlation means that the HRM scored rating errors lower than correct ratings. For RATEONLY we calculate the correlation per user, since every participant issued the same number of ratings. For the other experiments we calculate the correlation coefficients per rating.

Indirectly, H_{err} measures the average agreement of the raters with the gold standard. If the community, more precisely, the raters using the HRM, is sufficiently in agreement with the gold standard, punishment in the form of lower scores *should* follow. If, on the other hand, the community disagrees with the gold standard on average, rating errors should yield higher scores. Such a disagreement could happen for reasons of systematical differences in perception, e.g., due to different tastes or incompetence, or because of collusion attacks.

We find rather strong negative correlations for COMMISSION and for RATEONLY, and weak ones for HONSTUDENTS and for HONSTAFF (cf. Table 3).

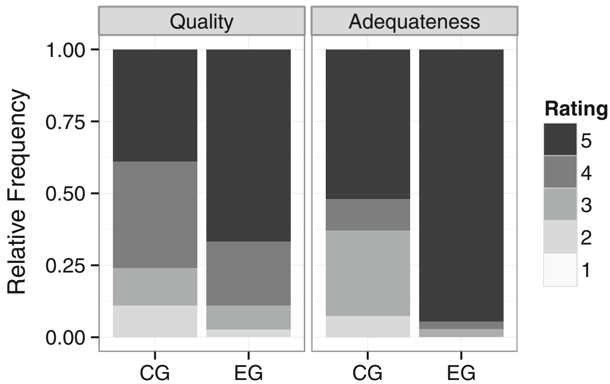

Figure 3 Distribution of expert ratings for quality and adequateness. 1 is the lowest rating, 5 is the highest rating a contribution could receive.

Table 3 Pearson correlation between rating error and score for ratings scored with the HRM.

Experiment	Correlation	p
HONSTUDENTS	−0.18	0.38
HONSTAFF	−0.26	< 0.01
COMMISSION	−0.97	< 0.01
RATEONLY	−0.83	< 0.05

Potential reasons for the two strong correlations could be that the manipulations used as gold standard in RATEONLY are more precise than expert ratings. For COMMISSION, we speculate that the strong correlation results from the better knowledge of participants regarding data-modeling techniques, and therefore a higher correlation with the expert ratings, compared to participants of HONSTUDENTS and HONSTAFF.

Overall, we conclude that the HRM punished rating errors to different degrees. The communities seem to have been in consensus with the experts, i.e., there were no systematic differences in perception.

7.4 Evaluation of the questionnaire

27 out of the 46 users invited answered the questionnaire. Figure 4 shows the results for selected questions.

We asked participants which rating strategy they used to maximize their rating score. Some stated an altruistic attitude "I did not intend to get as many points as possible, but tried to increase the quality of contributions by rating pointless or bad contributions as bad.", "I tried to rate as much as possible as honestly as possible." (both rewarded by the HRM). Others said they tried to maximize their scores, although with different rating strategies, dependent on their respective scoring mechanism, namely "Rating many items. But only those whose quality was easy to decide." (HRM), and "Simply rated everything, no matter how." (static reward for ratings).

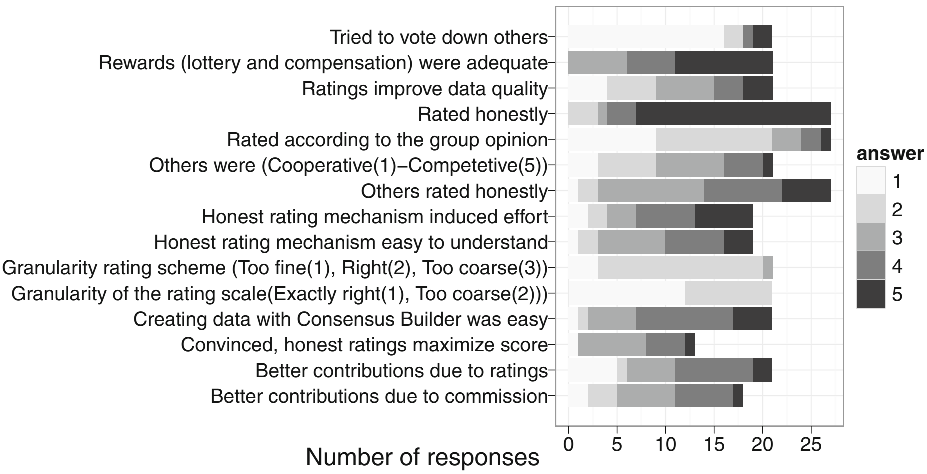

Figure 4 Number of responses for selected questionnaire questions. answers range from 1: 'strongly disagree' to 5: 'strongly agree', unless noted otherwise.

Finally, we analyzed the correlations of experimental results of the participants and their questionnaire answers. Not surprisingly, we find a positive correlation between the understanding of the HRM and the number of ratings ($\rho = 0.52$, $p <$ .05). There is a strong correlation between the number of ratings and the stated strategy of rating the contributions of others badly in order to keep their scores low, but only for raters whose ratings were scored statically ($\rho = 0.764$, $p < 0.5$), while for participants using the HRM this correlation was negligible ($\rho = 0.16$, $p = 0.59$). We find a weak correlation between the understanding of the HRM and the score received per rating ($\rho = 0.31$, $p = .18$). In other words, it is not necessary to understand the mechanism in order to profit from it.

8 Discussion and lessons learned

Participants have been intrinsically motivated to some degree. They made good contributions and gave high-quality ratings even when they did not receive an extra reward for it. However, one of our results is that contributions and ratings are at least as good or better in the presence of commissions and the HRM, respectively. We speculate that, at least to some degree, the intrinsic motivation resulted from the fixed monetary compensation for participation, insofar as participants felt they had to offer at least some effort. In fact, when planning the experiments, given a fixed total budget, we were confronted with a tradeoff between two quantities: on the one hand, the guaranteed compensation, on the other, the score-dependent compensation. A low guaranteed compensation results in fewer participants. A high guaranteed compensation provides less incentive from rating dependent rewards.

Despite the rewards offered, ratings are sparse in both CG and EG in the experiments with a variable number of ratings, i.e., there are not nearly as many ratings per contribution as there are participants per group (cf. Table 2). Questionnaire responses offer some explanation for this: Some participants do not like to rate data items they do not understand well enough even if they receive a guaranteed score. One participant of Rateonly dropped out after the creation phase because she did not feel sufficiently familiar with the knowledge domain. Another reason might be the guaranteed compensation, as discussed above.

The results show a weak correlation between rating scores and understanding of the mechanism. An interesting question is whether a fake HRM would have the same effects as the real one. We speculate that telling participants that an HRM is used while scoring with some fake mechanism (for example, randomly) would still yield comparable results, at least in the short run. This could be tested experimentally by comparing the alternatives ‘no mechanism’, ‘real HRM’, and ‘fake HRM’. Note that, even if a fake mechanism yielded results similar to those of the real mechanism, the real mechanism would still be at least as good (or better in case participants realized the fake).[4]

[4]Evidence that a fake mechanism could work in the short run comes from Shaw et al. [29]. They find that telling crowd workers the idea behind the Bayesian Truth Serum HRM (”You will receive a bonus payment if your answer is more common than collectively predicted.”), without performing the actual computation of the score, is sufficient to incentivize the workers to answer survey questions more accurately. Their task was rather short lived, so the same reasoning as above applies: In the long run, the fake would likely be discovered.

Next, it turned out to be difficult to find domains with all of the following characteristics: (a) They are sufficiently controversial to generate variance in the ratings. (b) The experimenters, but not the participants, have access to the gold standard. For example, in the creation phase before the RATEONLY experiment, participants kept the schema extremely simple and almost exclusively copied data publicly available on the web. Since the contributions were almost completely correct, there were no negative ratings and hence no variance in the rating values. This means that we could not have measured the effects of our mechanisms on either contribution or rating quality of the creation phase meaningfully.

Finally, the quality of the schema created by participants not familiar with data modeling was surprisingly good. Despite some beginner mistakes (confusion of normal associations and type associations, creation of topic type 'Properties') the quality of the schema level was good and detailed.

9 Summary

In this paper we have investigated how rating-based reward mechanisms can improve the quality of structured knowledge created collaboratively. In particular, we have discussed how mechanisms for honest ratings known from literature can be applied to this scenario. We have presented a community platform that features such reward mechanisms. We have formulated hypotheses and designed experiments to evaluate the effects of reward mechanisms for the collaborative creation of structured knowledge. We have carried out the experiments with different online communities. The communities constructed complex knowledge domains in settings close to real-world scenarios. We find that an honest rating mechanism improves the quality of ratings in two out of three experiments and that it "punishes" rating errors with lower scores. Further we find that rewards for good contributions increase the quality of the contributions.

References

1. Albers, A., et al.: A new perspective on product engineering overcoming sequential process models. In: The Future of Design Methodology. Springer London (2011)
2. Auer, S., Dietzold, S.: OntoWiki—a tool for social, semantic collaboration. In: Proceedings of the 5th International Semantic Web Conference ISWC (2006)
3. Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P., Kraut, R.E.: Using social psychology to motivate contributions to online communities. In: Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work, CSCW '04, pp. 212–221. ACM, New York, NY, USA (2004)
4. Benkler, Y.: The wealth of networks: how social production transforms markets and freedom. Yale University Press, New Haven, CT, USA (2006)
5. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD '08, pp. 1247–1250. ACM, New York, NY, USA (2008)
6. Braun, S., Schmidt, A., Walter, A., Nagypal, G., Zacharias, V.: Ontology maturing: a collaborative Web 2.0 approach to ontology engineering. In: Proceedings of the Workshop on Social and Collaborative Construction of Structured Knowledge at the 16th International World Wide Web Conference (WWW 07), Banff, Canada (2007)

7. Buitelaar, P., Cimiano, P.: Ontology Learning and Population: Bridging the Gap Between Text and Knowledge. Frontiers in Artificial Intelligence and Applications. Ios Press (2008)
8. Campanini, S.E., Castagna, P., Tazzoli, R.: Platypus wiki: a semantic wiki wiki Web. In: 1st Italian Semantic Web Workshop (2004)
9. Cosley, D., Frankowski, D., Kiesler, S., Terveen, L., Riedl, J.: How oversight improves member-maintained communities. In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems, CHI '05, pp. 11–20. ACM, New York, NY, USA (2005)
10. Cosley, D., et al.: Is seeing believing?: how recommender system interfaces affect users' opinions. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2003)
11. Douceur, J.R.: The sybil attack. In: IPTPS (2002)
12. Eckert, K., et al.: Crowdsourcing the assembly of concept hierarchies. In: Proceedings of the 10th Annual Joint Conference on Digital Libraries - JCDL '10 (2010)
13. Efron, B.: Forcing a sequential experiment to be balanced. Biometrika **58**(3), 403–417 (1971)
14. Fleiss, J.L.: Statistical Methods for Rates and Proportions, 2nd edn. Wiley, New York (1981)
15. Holsapple, C.W., Joshi, K.D.: A collaborative approach to ontology design. ACM Commun. **45**(2), 42–47 (2002)
16. Hütter, C., Kühne, C., Böhm, K.: Peer production of structured knowledge—an empirical study of ratings and incentive mechanisms. In: Proceeding of the 17th ACM Conference on Information and Knowledge Mining - CIKM '08 (2008)
17. Jäschke, R., et al.: Organizing publications and bookmarks in bibsonomy. In: CKC (2007)
18. Jurca, R., Faltings, B.: Minimum payments that reward honest reputation feedback. In: Proceedings of the 7th ACM Conference on Electronic Commerce - EC '06 (2006)
19. Jurca, R., Faltings, B.: Collusion-resistant, incentive-compatible feedback payments. In: Proceedings of the 8th ACM Conference on Electronic Commerce, EC '07, pp. 200–209. ACM, New York, NY, USA (2007)
20. Kittur, A., Kraut, R.E.: Harnessing the wisdom of crowds in Wikipedia: quality through coordination. In: Proceedings of the 2008 ACM Conference on Computer Supported Cooperative work, CSCW '08, pp. 37–46. ACM, New York, NY, USA (2008)
21. Kochhar, S., Mazzocchi, S., Paritosh, P.: The anatomy of a large-scale human computation engine. In: Proceedings of the ACM SIGKDD Workshop on Human Computation, HCOMP '10, pp. 10–17. ACM, New York, NY, USA (2010)
22. Kotis, K., Vouros, A.: Human-centered ontology engineering: The hcome methodology. Knowl. Inf. Syst. **10**(1), 109–131 (2006)
23. Miller, N., Resnick, P., Zeckhauser, R.: Eliciting informative feedback: the peer-prediction method. Manag. Sci. **51**(9), 1359–1373 (2005)
24. Moor, A.D., Leenheer, P.D., Meersman, R., Starlab, V.: Dogma-mess: a meaning evolution support system for interorganizational ontology engineering. In: In Proc. of the 14th International Conference on Conceptual Structures, ICCS 2006, pp. 189–203. Springer-Verlag (2006)
25. Noy, N., Chugh, A., Alani, H.: The ckc challenge: exploring tools for collaborative knowledge construction. IEEE Intell. Syst. **23**(1), 64–68 (2008)
26. Pepper, S., Architect, S.I., Infotek, S.: The TAO of Topic Maps—Finding the Way in the Age of Infoglut. In: Proceedings of XML Europe (2000)
27. Prelec, D.: A Bayesian truth serum for subjective data. Science (New York, N.Y.) **306**, 462–466 (2004)
28. Schaffert, S.: IkeWiki: a semantic wiki for collaborative knowledge management. In: Tolksdorf, R., Simperl, E.P.B., Schild, K. (eds.) In: Proceedings of the First International Workshop on Semantic Technologies in Collaborative Applications STICA 06, pp. 388–396 (2006)
29. Shaw, A.D., Horton, J.J., Chen, D.L.: Designing incentives for inexpert human raters. In: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative work, CSCW '11, pp. 275–284. ACM, New York, NY, USA (2011)
30. Siorpaes, K., Hepp, M.: Games with a purpose for the semantic Web. IEEE Intell. Syst. **23**(3), 50–60 (2008)
31. Siorpaes, K., Simperl, E.: Human intelligence in the process of semantic content creation. World Wide Web **13**(1–2), 33–59 (2010)
32. Souzis, A.: Building a semantic wiki. IEEE Intell. Syst. **20**, 87–91 (2005)
33. Sunagawa, E., Kozaki, K., Kitamura, Y., Mizoguchi, R.: An environment for distributed ontology development based on dependency management. In: International Semantic Web Conference, pp. 453–468 (2003)
34. Tijerino, Y.A., Embley, D.W., Lonsdale, D.W., Ding, Y., Nagy, G.: Towards ontology generation from tables. World Wide Web **8**(3), 261–285 (2005)

35. Tudorache, T., Noy, N.F.: Collaborative protege. In: CKC (2007)
36. Viegas, F.B., Wattenberg, M., Kriss, J., van Ham, F.: Talk before you type: coordination in Wikipedia. In: Proceedings of the 40th Annual Hawaii International Conference on System Sciences, HICSS '07, p. 78. IEEE Computer Society, Washington, DC, USA (2007)
37. Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., Studer, R.: Semantic Wikipedia. In: WWW, pp. 585–594 (2006)
38. von Ahn, L.: Games with a purpose. IEEE Comput. **39**(6), 92–94 (2006)
39. Vrandecic, D., Pinto, H.S., Tempich, C., Sure, Y.: The diligent knowledge processes. J. Knowl. Manag. **9**(5), 85–96 (2005)
40. Wilkinson, D.M., Huberman, B.A.: Assessing the value of cooperation in Wikipedia. First Monday **12**(4) (2007)
41. de Winter, J.C., Dodou, D.: Five-point Likert items: t test versus Mann-Whitney-Wilcoxon. Pract. Assess., Res. Eval. **15**(11) (2010)
42. Xu, Z., Ni, Y., He, W., Lin, L., Yan, Q.: Automatic extraction of OWL ontologies from UML class diagrams: a semantics-preserving approach. World Wide Web **15**(5–6), 517–545 (2011)
43. Zacharias, V., Braun, S.: Soboleo—social bookmarking and lighweight engineering of ontologies. In: CKC (2007)