Measuring the Effectiveness of Gamesourcing Expert Oil Painting Annotations

Myriam C. Traub, Jacco van Ossenbruggen, Jiyin He, and Lynda Hardman

Centrum Wiskunde & Informatica, Science Park 123, Amsterdam, The Netherlands
firstname.lastname@cwi.nl

Abstract. Tasks that require users to have expert knowledge are difficult to crowdsource. They are mostly too complex to be carried out by non-experts and the available experts in the crowd are difficult to target. Adapting an expert task into a non-expert user task, thereby enabling the ordinary "crowd" to accomplish it, can be a useful approach. We studied whether a simplified version of an expert annotation task can be carried out by non-expert users. Users conducted a game-style annotation task of oil paintings. The obtained annotations were compared with those from experts. Our results show a significant agreement between the annotations done by experts and non-experts, that users improve over time and that the aggregation of users' annotations per painting increases their precision.

Keywords: annotations, crowdsourcing, expert tasks, wisdom of the crowd.

1 Introduction

Cultural heritage institutions place great value in the correct and detailed description of the works in their collections. They typically employ experts (e.g. art-historians) to annotate artworks, often using predefined terms from expert vocabularies, to facilitate search in their collections. Experts are scarce and expensive, so that involving non-experts has become more common. For large image archives that have been digitized but not annotated, there are often insufficient experts available, so that employing non-expert annotations would allow the archive to become searchable (see for example ARTigo[1], a tagging game based on the ESP game[2]).

In the context of a project with the Rijksmuseum Amsterdam, we take an example annotation task that is traditionally seen as too difficult for the general public, and investigate whether we can transform it into a game-style task that can be played directly, or quickly learned while playing, by non-experts. Since we need to compare the judgements of non-experts with those of experts, we picked a dataset and annotation task for which expert judgements were available.

[1] http://www.artigo.org/

[2] http://www.gwap.com/gwap/gamesPreview/espgame/

M. de Rijke et al. (Eds.): ECIR 2014, LNCS 8416, pp. 112–123, 2014.

We conducted two experiments to investigate the following research questions. First, we want to know how the choices of non-expert users compare to those of experts in order to estimate the suitability of the non-expert annotations as part of a professional workflow.

Second, whether users perform better later in the game, and, if so, if they do this only on repeated images or also on new images. Third, how the partial absence of the correct answer affects the user performance in order to determine whether purely non-expert input is reliable.

2 Related Work

Increasing numbers of cultural heritage institutions initiate projects based on crowdsourcing to either enrich existing resources or create new ones [1]. Two well-known projects in this field are the Steve Tagger[3] and the Your Paintings Tagger[4]. Both constitute cooperations between museum professionals and website visitors to engage visitors with museum collections and to obtain tags that describe the content of paintings to facilitate search.

A previous study, [7], suggests that expert vocabularies that are used by professional cataloguers are often too limited to describe a painting exhaustively. This gap can be closed by making use of external thesauri from domains other than art history (e.g. WordNet, a lexical, linguistic database[5]). The interface for this task, however, targets professional users.

Steve Tagger and the Your Paintings Tagger focus on enriching their artwork descriptions with information that is common knowledge (e.g. Is a flower depicted?). The SEALINCMedia project[6] focuses on finding precise information (e.g. the Latin name of a plant) about depicted objects. To achieve this, the crowd is searched for experts who are able to provide this very specific information [2] and a recommender system selects artworks that match the users' expertise.

Another example for crowdsourcing expert knowledge is "Umati". Heimerl et al. [6] transformed a vending machine into a kiosk that returns snacks for performing survey and grading tasks. The restricted access to "Umati" in the university hallway ensured that the participants possessed the necessary background knowledge to solve the presented task. While their project also aims at getting expert work done with crowdsourcing mechanisms, their approach is different from ours. Whereas they aim at attracting skilled users to accomplish the task, we give non-experts the support they need to carry out an expert task.

Since most of these approaches target website visitors or passers-by, rather than paid crowd workers on commercial platforms, they need to offer an alternative source of motivation for users. Luis von Ahn's ESP Game [9] inspired

[3] http://tagger.steve.museum/

[4] http://tagger.thepcf.org.uk/

[5] http://wordnet.princeton.edu/

[6] http://sealincmedia.wordpress.com/

several art tagging games developed by the ARTigo project[7]. These games seek to obtain artwork annotations by engaging users in gameplay.

Goldbeck et al. [4] showed that tagging behavior is significantly different for abstract compared with representational paintings. Users were allowed to enter tags freely, without being limited to the use of expert vocabularies. Since our set of images showed a similar variety in styles and periods, we also investigated whether particular features of images had an influence on the user behavior.

He et al. [5] investigated if and how the crowd is able to identify fish species on photos taken by underwater cameras. This task is usually carried out by marine biologists. In the study, users were asked to identify fish species by judging the visual similarity between an image taken from video and images showing already identified fish species.

A common challenge of tagging projects lies in transforming the large quantity of tags obtained through the crowd to high quality annotations of use in a professional environment. As Galton proved in 1907, the aggregation of the *vox populi* can lead to surprisingly exact results that are "correct to within 1 per cent of the real value" [3]. Such aggregation methods can improve the precision of user judgements [8], a feature that can potentially be used to increase the agreement between users and experts of our tagging game.

3 Experimental Setup

We investigated the categorization of paintings into subject types (e.g. landscapes, portraits, still lifes, marines), which is typically considered to be an expert task. We simplified the task by changing it into a multiple choice game with a limited, preselected set of candidates to choose from. Each included the subject type's label, a short explanation of its intended usage and a representative example image. To investigate the influence of the preselection of the candidates on the performance of the users, we carried out two experiments: a baseline condition, which always had a correct answer among the presented candidate answers, and, to simulate a more realistic setting, a condition where in 25% of the cases the correct answers had been deliberately removed.

3.1 Procedure

Users were presented with a succession of images (referred to as *query images*) of paintings that they were asked to match with a suitable subject type (see Figure 1). We supported users by showing them a pre-selection of six *candidates*. Five of these candidates represented subject types and one of them (labeled "others") could be used if the assumed correct subject type was not presented. To motivate users to annotate images correctly and to give them feedback about the "correctness"[8] of their judgements, they were awarded ten points for judgements that agree with the expert and one point for the attempt (even if incorrect).

[7] `http://www.artigo.org/`

[8] By "correct" we mean that a given judgement agrees with the expert.

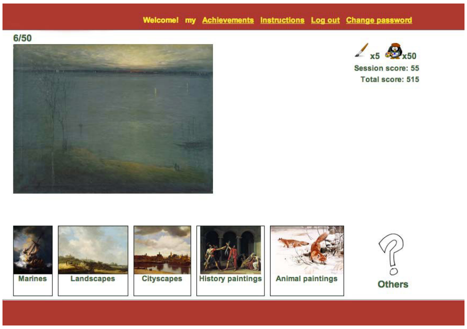


Fig. 1. Interface of the art game with the large query image on the upper left. The five candidate subject types are shown below, together with the *others* candidate.

The correct answer was always presented and users got direct feedback on every judgement they made. With this experiment we wanted to find out whether (and how well) users learn under ideal conditions. We use the data of the first experiment as a baseline for comparing the results of the second experiment.

In the second experiment, the correct answer is not always presented.

3.2 Experiments Conducted

We adapted the online tagging game used for the Fish4Knowledge project [5]. On the login page of the game, we provide a detailed description of the game including screenshots, instructions and the rules of the game.

Baseline Condition - For each query image, we selected one candidate that, according to the expert ratings, represents a correct subject type and three candidates representing related, but incorrect, subject types. One candidate was chosen randomly from the remaining subject types. For cases, when there were only two related but incorrect subject types available, we showed two incorrect random ones, so the total number of candidates would remain six (including the *others* candidate). The categorization of similar subject types was done manually and is based on their similarity. An example of related subject types is *figure*, *full-length figure*, *half figure*, *portrait* and *allegory*.

Imperfect Condition - In this setting, the correct candidate is not presented in 25% of the cases. This is used to find out how good the learning performance of users is when the candidate selection is done by an automated technique that may fail to find a correct candidate in its top five. The selection of the candidates was the same as in the baseline experiment, for the missing correct candidate we added another incorrect candidate.

3.3 Materials

The expert dataset [10] provides annotations of subject types for the paintings of the Steve Tagger project by experts from the Rijksmuseum Amsterdam.

Table 1. Used subject types and the number of expert annotations

Subject type	Annotations
full-length figures	40
landscapes	33
half figures	13
allegories, history paintings, portraits, animal paintings, genre, kacho, figures	8
townscapes	6
flower pieces	5
marines, cityscapes, maesta, seascapes, still lifes	3

We selected 168 expert annotations for 125 paintings (Table 1). The number of annotations per painting ranged from four (for one painting) to only one (for 83 paintings). These multiple classifications are considered correct: a painting showing an everyday scene on a beach[9] can be classified as *seascapes*, *genre*, *full-length figure* and *landscapes*. This, however, makes our classification task more difficult.

Query Images - The images used as query images are a subset of the thumbnails of paintings from the Steve Tagger[10] data set. The paintings are diverse in origin, subject, degree of abstraction and style of painting. Apart from the image, we provided no further information about the painting. Within the first ten images that were presented to the user, there were no repetitions. Afterwards, images may have been presented again with a 50% chance. The repetitions gave us more insight on the performance of the users.

Candidates - A candidate consists of an image, a label (subject type) and a description. For each subject type we selected one representative image from the corresponding Wikipedia page[11]. The main criterion for the selection was that the painting should show typical characteristics. The candidates were labeled with the names of the subject types from the Art & Architecture Thesaurus[12] (AAT) which comprises in total more than 100 subject types. The representative images were intended to give users a first visual indication of which subject type might qualify and it made it easier for users to remember it. If this was not sufficient for them to judge the image, they could verify their assumption by displaying short descriptions taken from the AAT, for example:

[9] `http://tagger.steve.museum/steve/object/280`

[10] `http://tagger.steve.museum/`

[11] E.g.: `http://en.wikipedia.org/wiki/Maesta`

[12] `http://www.getty.edu/research/tools/vocabularies/aat/index.html`

Marines
"Creative works that depict scenes having to do with ships, shipbuilding, or harbors. For creative works depicting the ocean or other large body of water where the water itself dominates the scene, use 'seascapes'. "[13]

The descriptions of the subject types are important, as the differences between some subject types are subtle.

3.4 Participants

Participants were recruited over social networks and mailing lists. For the analysis we used 21 for the first experiment and 17 in the second one, in total 38, after removing three users who made fewer than five annotations. The majority of the participants have a technical professional background and no art-historical background. In the baseline condition, users who scored at least 400 received a small reward.

3.5 Limitations

Our image collection comprised 125 paintings, and compared with a museum's collection this is a small number. Because of the repetitions, the number of paintings that the user saw only increased gradually over time, which would have made it possible to successively introduce a larger number of images to the users. This, however, would have made it difficult to obtain the necessary ground truth.

In the available ground truth data, each painting was judged by only one expert, which prevents us from measuring agreement among experts. This measurement might have revealed inconsistencies in the data that influenced users' performance.

In realistic cases, ground truth will be available for only a small fraction of the data. To apply to such datasets, our setting needs other means of selecting the candidates. This can be realised, for example, by using the output of an imperfect machine learning algorithm, or by taking the results of another crowdsourcing platform. We think it is realistic to assume that in such settings the correct answer is not always among the results, and acknowledge that the frequency of this really happening may differ from the 25% we assumed in our second experiment.

The game did not go viral, which can mean that incentives for the users to play the game and/or the marketing could be improved.

4 Results

An overview of the results of all users of both experiments shows a large variation in number of judgements and precision (Fig. 2). Users who judged more images

[13] `http://www.getty.edu/vow/AATFullDisplay?find=marines&logic=AND&note=&english=N&prev_page=1&subjectid=300235692`

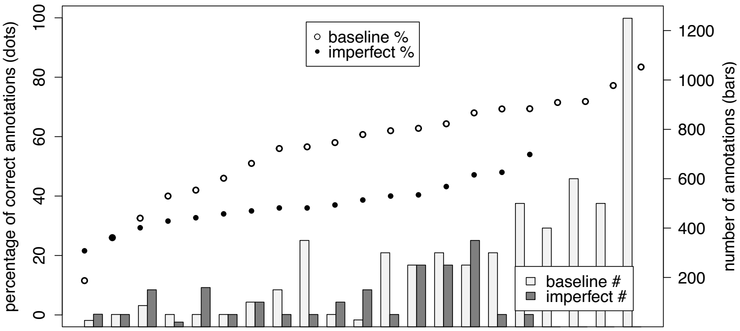


Fig. 2. Percentage of correct annotations per user (dots) and the number of annotations (bars). The users are ordered by increasing precision from left to right.

also tend to have higher precision. This might suggest that users indeed learn to better carry out the task or that well-performing users played more.

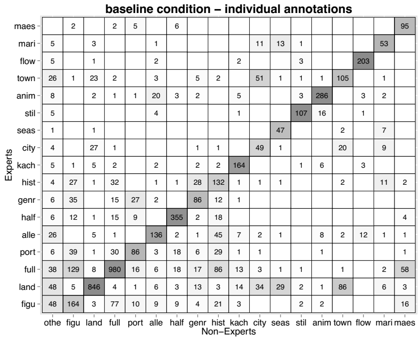
In both conditions, all users who finished at least one round of 50 images performed much better than a random selection of the candidates (with a precision of 17%), suggesting that we do not have real spammers amongst our players. On average, the precision of the users in the baseline condition (56%) is higher than in the imperfect condition (37%). This indicates that the imperfect condition is more difficult. This is in line with our expectations: in order to agree with the expert, users in the imperfect condition sometimes need to select the *other* candidate instead of a candidate subject type that might look very similar to the subject type chosen by the expert.

4.1 Agreement Per Subject Type

To understand the agreement between experts and users, we measure precision and recall per subject type. *Precision* is the number of agreedupon judgements for a subject type divided by the total judgements given by users for that subject type. *Recall* is the number of agreed-upon judgements for a subject type divided by the total judgements given by the expert for that subject type.

Both measures are visualized in confusion heatmaps (Fig.s 3 - 6). The rows represent the experts' judgements, while the columns show how the users classified the images. The shade of the cells visualizes the value of that cell as the fraction of the users' total votes for that specific subject type. Darker cells on the diagonal indicate higher agreement, while other dark cells indicate disagreement.

Some subject types score low on precision: *cityscapes* is frequently chosen by non-experts when the expert used *landscapes* or *townscapes*, while users select *history paintings* where the expert sees *figures* (Fig. 3). On the other hand, *flower pieces* and *animal paintings* score high on both precision and recall. Selecting the *others* candidate did not return points in the baseline condition, and some players reported to have noticed this and did not use this candidate afterwards. With 243 *others* judgements out of a total of 5640, it received relatively few



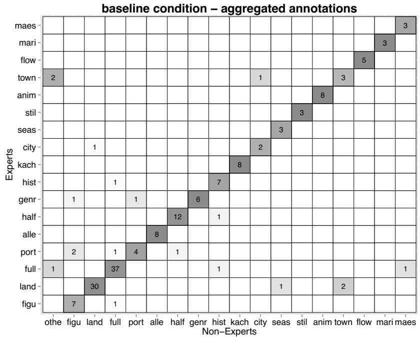


Fig. 3. Despite many deviations, the graph shows a colored diagonal representing an agreement between non-experts and experts. The task therefore seems to be difficult but still manageable for users.

Fig. 4. The "Wisdom of the Crowd" effect eliminates many deviations of the non-experts' judgements from the experts' judgements. However, there are still deviations for similar subject types such as *cityscapes* and *townscapes.*

clicks. The agreement between users and experts is substantial (Cohen's Kappa of 0.65), we see a clear diagonal of darker color.

Aggregating user judgements by using majority voting (Fig. 4), removes some deviations from the experts' judgements (Cohen's Kappa of 0.87) to almost perfect agreement. For example, all *cityscapes* judgements by users for cases where expert judged *landscapes* are overruled in the voting process and this major source of disagreement in Fig. 3 disappears. There is only one case where the expert judged *townscapes* and the majority vote of the users remained *cityscapes.* The painting description states that it shows "a dramatic bird's eye view of Broadway and Wall Street"[14] in New York. Therefore, *townscapes* cannot be the correct subject type and users were right to disagree with the expert. Most *others* judgements are largely eliminated by the majority voting. However, three paintings remain classified as *others* by the majority which indicates a very strong disagreement with the experts' judgement. One of these paintings does not show a settlement, but in an abstract way depicts a bomb store in the "interior of the mine"[15]. The other two show a carpet merchant in Cairo[16] and the "Entry of Christ into Jerusalem"[17], both being representations of large cities and therefore incorrectly categorized as *townscapes* by the expert.

In the imperfect condition, the confusion heatmaps are similar, however, the disagreement between users and experts is higher. The *others* candidate was the correct option in 25% of the cases. The users made more use of it, as shown by the higher numbers in the first column of Fig. 6. The agreement in the *allegories* column is, with 13%, even below chance. Majority voting increases the

[14] http://www.clevelandart.org/art/1977.43

[15] http://www.tate.org.uk/art/artworks/bomberg-bomb-store-t06998

[16] https://collections.artsmia.org/index.php?page=detail&id=10361

[17] http://tagger.steve.museum/steve/object/172

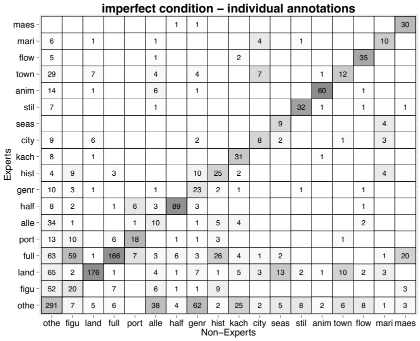


Fig. 5. The *others* candidate attracted many user votes. Compared to the baseline condition, the diagonal is less prominent, meaning that the agreement is lower in most cases.

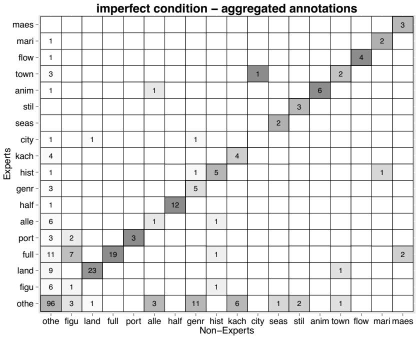


Fig. 6. The aggregation of user votes could compensate some of the deviations from agreement, however the additional *others* candidate had a negative effect on the agreement for allegories, genre and kacho

precision, but only to 20%. The AAT defines this subject type to "express complex abstract ideas, for example works that employ symbolic, fictional figures and actions to express truths or generalizations about human conduct or experience". Therefore, it is very difficult to recognize an *allegory* as such without context information about the painting. User judgements diverging from the expert's judgements are largely removed by majority vote. The "Wisdom of the Crowd" effect, however, is not as strong as in the baseline condition. It raised the Cohen's Kappa from 0.47 to a (still) moderate agreement of 0.55.

We further analyzed the agreement of the non-experts and the experts on image level in the baseline condition. The broad range from 2% to 98% indicates very strong (dis-)agreement for some cases. In the images with the highest agreement, the relation between the depicted scenes and the subject type is intuitively comprehensible: the images with 98% agreement show flowers (*flower pieces*), monkeys (*animal painting*) and a still life (*still lifes*). An entirely different picture emerges, when we look at the images with low agreement. We presented the most striking cases to an expert from the Rijksmuseum Amsterdam to re-evaluate the experts' judgements and we identified two main reasons for disagreement: users would have needed additional information, such as the title, to classify the painting correctly; the expert annotations were incomplete or incorrect.

4.2 Performance over Time

The improvement of the users' precision over time does not necessarily mean that they have learned how to solve the problem (generalization), but that they "only" have learned the correct solution for a concrete problem (memorizing).

Memorizing - A learning effect is evident in the performance curve of the users for repeated images (Fig. 7). In the baseline condition, users had an initial

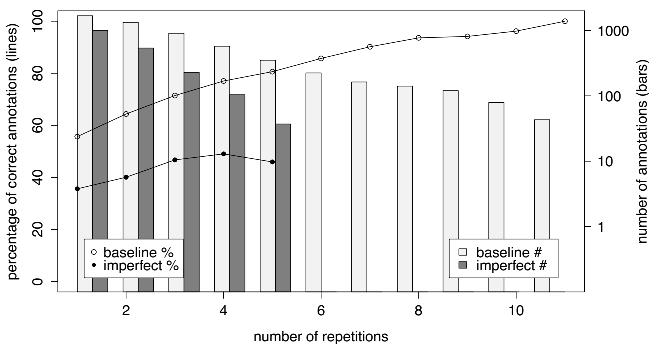


Fig. 7. Learning curves (lines) for the memorization effect of repeated images and numbers on annotations (bars) per repetition

success rate of 56% correct judgements. After seven repetitions, they judged 90% of the query images correctly. In the imperfect condition, the performance is consistently lower. The difference between the first appearance of an image (success rate of 36%) and the fifth appearance of an image (success rate of 46%) is lower than in the baseline experiment where we see an increase of 25 percent units. The lines in Fig. 7 were cut off after eleven repetitions for the baseline condition and five repetitions for the imperfect condition because the number of judgements dropped below 15. We further analyzed the results of a fixed homogeneous population of seven (baseline) and eight (imperfect) users. The outcomes were nearly identical for both conditions. These results show that users in the baseline condition improve on memorizing the correct subject type for a specific image. The differences between the two conditions indicate that users found it more difficult to learn the subject types in the imperfect condition.

Generalization - The judgement performance of users on the first appearances of images indicates whether they are able to generalize and apply the knowledge to unseen query images. If users learn to generalize, it is likely that they will improve over time at judging images that they have not seen before. Judgement precision increases throughout gameplay for both conditions (Fig. 8). While users in the baseline experiment started with a success rate of 44%, they reach 90% after about 250 images. Users in the imperfect condition started at a much lower rate of 33% and increase to 60%, after about 150 images. The declining number of images that are new to the user and the declining number of users that got so far in the game, lead to a drop in available judgements at later stages in the game. Therefore, we cut the graphs at sequence numbers 400 (baseline) and 160 (imperfect).

Our findings show that users can learn to accomplish the presented simplified expert task. This does not mean, however, that they would perform equally well if confronted with the "real" expert task. Users were given assistance by reducing the number of candidates from more than one hundred to six, they were provided

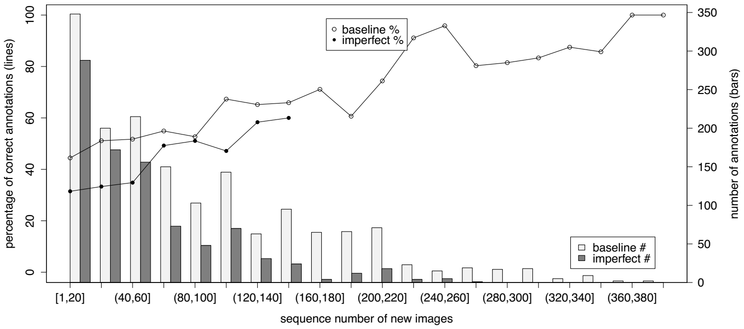


Fig. 8. Users' performance for first appearances of images that occur in different stages of the game (lines) and number of annotations

a visual key (example image) to aid memorization and a short description of the subject type. A way to increase the success rate in a realistic setting would be to train users on a "perfect" data set and after passing a predefined success threshold, introduce "imperfect" data into the game.

5 Conclusions

Our study investigates the use of crowdsourcing for a task that normally requires specific expert knowledge. Such a task could be relevant to facilitate search by improving metadata on non-textual data sets, but also in crowdsourcing relevance judgements for more complex data in a more classic IR setting.

Our main finding is that non-experts are able to learn to categorize paintings into subject types of the AAT thesaurus in our simplified set-up. We studied two conditions, one with the expert choice always present, and one in which the expert choice had been removed in 25% of the cases. Although the agreement between experts of the Rijksmuseum Amsterdam and non-experts for the first condition is higher, the agreement in the imperfect condition is still acceptably high. We found that the aggregation of votes leads to a noticeable "Wisdom of the Crowds" effect and increases the precision of the users' votes. While this removed many deviations of the users' judgements from the experts' judgements, on some images, the disagreement remained. We consulted an expert and identified two main reasons: Either the annotations by the experts were incomplete or incorrect or the correct classification required knowing context information of the paintings that was not given to the users.

The analysis of user performance over time showed that users learned to carry out the task with higher precision the longer they play. This holds for repeated images (memorization) as well as new images (generalization).

The next step is to balance the interdependencies of the three players: experts, automatic methods and gamers. We hope that reducing their weaknesses (scarce, requiring much training data, insufficient expertise) by directing the interplay of their strengths (ability to provide: high quality data, high quantity data, high quality when trained and assisted) can lead to a quickly growing collection of high quality annotations.

Acknowledgments. We would like to thank Sjaak Wouters for making his data collection available as a basis for our research, Daniel Horst for the re-evaluation of the expert judgements and all participants of the experiments for their time and effort, especially user 21 for playing 25 rounds. This publication is funded by the Dutch COMMIT/ program and EC FP7 grant 257024, the Fish4Knowledge project.

References

1. Carletti, L., Giannachi, G., McAuley, D.: Digital humanities and crowdsourcing: An exploration. In: MW 2013: Museums and the Web 2013 (2013)
2. Dijkshoorn, C., Leyssen, M.H.R., Nottamkandath, A., Oosterman, J., Traub, M.C., Aroyo, L., Bozzon, A., Fokkink, W., Houben, G.-J., Hovelmann, H., Jongma, L., van Ossenbruggen, J., Schreiber, G., Wielemaker, J.: Personalized nichesourcing: Acquisition of qualitative annotations from niche communities. In: 6th International Workshop on Personalized Access to Cultural Heritage (PATCH 2013), pp. 108–111 (2013)
3. Galton, F.: Vox populi. Nature 75(1949), 7 (1907)
4. Golbeck, J., Koepfler, J., Emmerling, B.: An experimental study of social tagging behavior and image content. Journal of the American Society for Information Science and Technology 62(9), 1750–1760 (2011)
5. He, J., van Ossenbruggen, J., de Vries, A.P.: Do you need experts in the crowd?: a case study in image annotation for marine biology. In: Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, OAIR 2013, Paris, France, pp. 57–60 (2013); Le Centre De Hautes Etudes Internationales D'Informatique Documentaire
6. Heimerl, K., Gawalt, B., Chen, K., Parikh, T., Hartmann, B.: Communitysourcing: engaging local crowds to perform expert work via physical kiosks. In: Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems, CHI 2012, pp. 1539–1548. ACM, New York (2012)
7. Hildebrand, M., van Ossenbruggen, J., Hardman, L., Jacobs, G.: Supporting subject matter annotation using heterogeneous thesauri: A user study in web data reuse. International Journal of Human-Computer Studies 67(10), 887–902 (2009)
8. Hosseini, M., Cox, I.J., Milić-Frayling, N., Kazai, G., Vinay, V.: On aggregating labels from multiple crowd workers to infer relevance of documents. In: Baeza-Yates, R., de Vries, A.P., Zaragoza, H., Cambazoglu, B.B., Murdock, V., Lempel, R., Silvestri, F. (eds.) ECIR 2012. LNCS, vol. 7224, pp. 182–194. Springer, Heidelberg (2012)
9. von Ahn, L., Dabbish, L.: ESP: Labeling images with a computer game. In: AAAI Spring Symposium: Knowledge Collection from Volunteer Contributors, pp. 91–98. AAAI (2005)
10. Wouters, S.: Semi-automatic annotation of artworks using crowdsourcing. Master's thesis, Vrije Universiteit Amsterdam, The Netherlands (2012)