Is it a bird or is it a crow? The influence of presented tags on image tagging by non-expert users

Mieke H. R. Leyssen
Centrum Wiskunde en Informatica
P.O. Box 94079
1090 GB Amsterdam, The Netherlands
Mieke.Leyssen@cwi.nl

Myriam C. Traub
Centrum Wiskunde en Informatica
P.O. Box 94079
1090 GB Amsterdam, The Netherlands
Myriam.Traub@cwi.nl

Jacco van Ossenbruggen
Centrum Wiskunde en Informatica
P.O. Box 94079
1090 GB Amsterdam, The Netherlands
Jacco.van.Ossenbruggen@cwi.nl

Lynda Hardman
Centrum Wiskunde en Informatica
P.O. Box 94079
1090 GB Amsterdam, The Netherlands
Lynda.Hardman@cwi.nl

ABSTRACT

Cultural heritage institutes often make use of tags to facilitate searching their collections. While professionals associated with these institutes are able to add high quality descriptions to objects in the collections, both their time and their areas of expertise are limited. As a result, online tagging by non-professional users is more frequently becoming deployed to increase the number of tags. When these users are asked to tag objects in the collection, they can be confronted with tags submitted by other users. These tags may be of varying quality and present in differing numbers, both of which may influence users' tagging behavior. We report on a study on the impact of presenting different types of tags on the quality and quantity of tags added by users. We conclude that there is no difference in the quality and quantity of added tags in all experimental conditions, with the exception of the condition in which incorrect tags were presented. In this condition, the quality of the tags added by users decreased. We discuss the implications of these findings on the design of tagging interfaces.

Categories and Subject Descriptors

H.1.2 [**Information Interfaces and Presentation**]: User / Machine Systems; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces; M.0 [**Knowledge Management**]: Knowledge Acquisition; M.9 [**Knowledge Management**]: Knowledge Valuation

General Terms

Design, Experimentation, Human Factors.

Keywords

image tagging, cultural heritage, tagging support, user interface, user behavior, crowdsourcing

1. INTRODUCTION

Cultural heritage institutes such as museums or archives have always generated and curated metadata describing their assets to facilitate searching their collection. They employ multiple curators who describe different aspects of the artworks using predefined terms from specialist vocabularies. This metadata is generally considered to be high quality, precise and exhaustive with respect to art-historical aspects. However, these professionals cannot provide all specific knowledge to describe the variety of prints in detail since their time and areas of expertise are limited. Information that relates to the actual content of the depicted object (e.g. the species of a plant) can often not be provided by the curators. The knowledge required to add this information may not be available within the institute, but it can potentially be provided by external users from different fields of expertise.

To enable users from outside cultural heritage institutes to annotate artworks, we need a system that is easy to use, engaging and that does not require expert knowledge about different hierarchies and vocabularies. Such a system can be made available to selected domain experts to encourage annotations from specific fields, but also to the general public to attract larger numbers of contributions.

To gain an understanding of how to improve the design of tagging interfaces, we investigate the influence of presenting generic and specific tags on the quantity and quality of tags added by non-expert users. An example of a generic tag is "bird", and of a specific tag is "crow". Within the context

of a national project[1] in which cultural heritage institutes are partners[2], we conducted a user experiment using a selection of representative images that mimics the problem of a lack of domain-specific expertise. These were pre-tagged by ourselves to represent the varying quality of tags that one might find in an uncurated environment. We would like to encourage users to describe the object depicted in the image by adding specific tags and removing incorrect tags. We thus designed an experiment to investigate the effects of presenting correct generic tags, of presenting correct and incorrect specific tags, and of presenting no tags.

In the following section we discuss social tagging systems used by cultural heritage institutes and how they encourage users to provide high quality tags. In section 3, we describe the experimental design and we discuss the results of the experiment in section 4. We discuss the general findings and their implications in section 4. We describe the implementation of the system in section 6. Section 7 consists of a summary of our findings and future experiments.

2. RELATED WORK

Collaborative or social tagging is the principle of exploiting the knowledge of a large number of non-expert users to describe digital objects such as websites, images, music or videos [5]. In contrast to professional subject indexing as it is done in cultural heritage institutes, social tagging has normally no underlying thesaurus or vocabulary: users can choose their tags freely. As a consequence, the institutes need tools that enable them to steer the users to adding useful tags.

The photo sharing platform Flickr is a social tagging platform in which users can not only upload and tag their own photos, but also add tags and comments to photos of other users. Tags are added to provide additional information about the image and also to improve searching the collection of images [1]. In 2008, Flickr launched the project "The Commons"[3] which allows cultural heritage institutes to upload parts of their collections to the platform. Users can access these images and enrich the given information by adding tags, comments and even links to other data sources on the Web. Several national archives, libraries and museums are listed as participants, which shows that there is a need among these institutes to present their content online with the aim of harvesting knowledge from people outside the museum.

Another well-known social tagging platform is the ESP Game that was developed by Luis von Ahn [10]. In this game, two online gamers are randomly teamed up to find appropriate tags for images. The tags are later used to improve the image retrieval of the Google search engine. The gamers score if they independently add the same tags. For some images, the players are provided with a list of "taboo words". These are (not necessarily correct) words that the players are not allowed to use for tagging the image. With the ESP Game, Luis von Ahn has proved that it is possible to turn "tedious work into something people want to do" [10].

These examples show that well-designed tagging systems can lead to useful contributions from users. Cultural heritage institutes are starting to discover the potential of this, which is reflected in the growing number of crowd-sourced systems.

One of the first projects dealing with online tagging in cultural heritage is the Steve Tagger Project[4]. One of the challenges tackled in this project was to investigate the influence of the presentation of user-entered tags and museum metadata on user contributions [9]. One of their findings was that showing tags had a noticeable effect on whether newly added tags differed from those already assigned. Users tended not to duplicate tags shown with a work of art, and instead entered different tags, while they do duplicate information from the metadata provided by the museum when this is presented in the user interface. There seems to be no relationship between the usefulness of a tag and whether or not other tags were shown. In their research, they did not report whether or not the type of tags that were presented had an influence on user tagging. This partially motivated the design of the study we report here.

A more recent social tagging project within the cultural heritage area is the Your Paintings project[5]. This project aims at presenting "the entire national collection of oil paintings online for public enjoyment, learning, and research" [2]. One of the challenges of the project was the insufficient metadata that was available for building a search index for the envisioned 205,000 oil paintings. The basic metadata provided by the partner institutes had to be enriched by involving external users. Inspired by the Galaxy Zoo project[6], an elaborate online tagging platform was created. In their tagging platform, tags that users added were never presented to other taggers. This also leads us to question the usefulness of presenting tags to users.

Hildebrand et al. [4] carried out a user experiment in which museum professionals were asked to annotate museum objects. For this experiment, a tagging interface was developed that integrated internal and external thesauri with which the professionals were familiar. This tagging system was very suitable for the museum professionals, but not for people outside the museum since they do not have a good understanding of the structure of the thesauri.

Our goal is to investigate whether or not presenting different types of tags has an influence on the tagging behavior of lay users. The research carried out in the context of the Steve Tagger Project makes claims about the influence of presenting tags on user tagging. However, they do not report whether or not the type of tags that were presented had an influence.

To investigate which types of tags may influence user tagging behavior we first need to define useful tag types. In 1962 the art historian Erwin Panofsky published a model to describe renaissance art works in three levels: pre-iconic (generic), iconographic (specific) and iconological (abstract). This model was later proved to be applicable to any type of

[1] http://www.commit-nl.nl
[2] http://sealincmedia.wordpress.com
[3] http://www.flickr.com/commons
[4] http://tagger.steve.museum
[5] http://www.bbc.co.uk/arts/yourpaintings
[6] http://www.galaxyzoo.org

image and extended with further facets by Sara Shatford [8]. Schreiber et al [7] mention that "more specific" is also "at a lower level of the AAT hierarchy" which is often used by cultural heritage institutes. So by using the first two levels of the Panofsky/Shatford model, namely generic and specific, in our experiment, we can investigate the hierarchy without letting users make use of vocabularies and hierarchies that they are not familiar with. We choose not to include the third category of the Panofsky/Shatford model, namely abstract tags, to ensure a higher number of user agreement.

3. EXPERIMENT

The goal of this experiment is to investigate if and how different types of tags influence the quality and quantity of tags added by users.

The different types of presented tags accompanying images are generic and specific tags describing the depicted objects. We want to verify whether providing one type of tag encourages users to add the other type of tag.

We presented both correct and incorrect specific tags in the experiment to investigate whether users trust the presented tags or whether they trust their own knowledge and indicate this by replacing the incorrect tag. In general, there are no "incorrect" tags in social tagging, since everything that a user adds could be relevant and valuable. However, in this experiment we make use of images that depict an object for which there is a correct specific tag that is best suitable to describe it (e.g. "crow"). An incorrect specific tag in our study indicates a specific tag that is wrongly assigned to an object (e.g. "magpie" when a crow was depicted). Incorrect generic tags were not presented in the experiment to encourage the belief that the presented tags could have been entered by other users. However, it was not mentioned that the presented tags were tags that other users added.

All participants are asked to add tags to photographs and prints. Different types of images are used to ensure that the findings of the experiment are not restricted to one type of image and therefore it would be justified to generalize the results.

3.1 Participants

The data of 56 persons (26 male and 30 female, aged 19-61 years) are used for the analyses. In total there were more participants, but some of them did not complete the task and therefore their data is not used. All but two of the 56 participants speak Dutch as their native language and all participants have been living in a Dutch speaking country for the last 10 years (25 in Belgium and 31 in the Netherlands). Participants were recruited by social media and mailing lists.

3.2 Stimuli

We present 12 photographs and 12 prints to all participants. The photographs are collected from the Web and are all licensed under creative commons. The prints are provided by the Rijksmuseum Amsterdam and used with permission.

The images that were presented to the participants show everyday life objects[7]. Each selected image depicts one main object of which we expect the majority of the participants to know (or at least have heard of) the generic and specific tags describing the object.

3.3 Design

The four experimental conditions in which the images are presented to the user consisted of

- no tags (none),
- a correct generic tag (generic),
- a correct specific tag (correct specific) or
- a incorrect specific tag (incorrect specific).

In Figure 1, the image is presented in the generic condition, meaning that the generic tag describing the object is presented ("vogel" translates into "bird" in English). In the correct specific condition, the presented tag is replaced by the correct specific tag ("kraai" translates into "crow" in English) and in the incorrect specific condition, the presented tag is replaced by the incorrect specific tag ("ekster" translates into "magpie" in English). In the none condition, no tag is presented.

The conditions are randomly assigned to different images and every condition is presented six times to each participant. The experiment is balanced between participants, thus ensuring that every image is presented the same number of times in each condition.

3.4 Procedure

A website was created so that participants could carry out the experiment from their own computer. On the main page, participants are informed that they will see 24 images and that it is their task to describe what is depicted in the image. They are asked to be as specific as possible. Users were informed that by participating, they give their permission to use the resulting data for scientific purposes.

The introduction page offers instructions on how participants can add a term that describes the image. This is explained by presenting a screenshot of the tagging interface (see Figure 1) and an explanatory text. If they are uncertain about the correctness or appropriateness of the term they added, they can indicate this. The text states that they can provide as many tags as they like, but that they should not add phrases or sentences to describe the image.

Tags that are entered and submitted by users are immediately added to the tag list in the user interface. Users are given the option to add comments whenever they want to provide more information. Participants are informed that if they agree with the already presented tag, they do not need to enter it again. If, however, they do not agree with a term, they can remove it. When users delete a tag, a pop-up is presented in which they have the opportunity to add a comment.

[7]These images and their respective generic, correct specific and incorrect specific tags can be found in the online appendix: `http://sealincmedia.project.cwi.nl/papers/www2013`

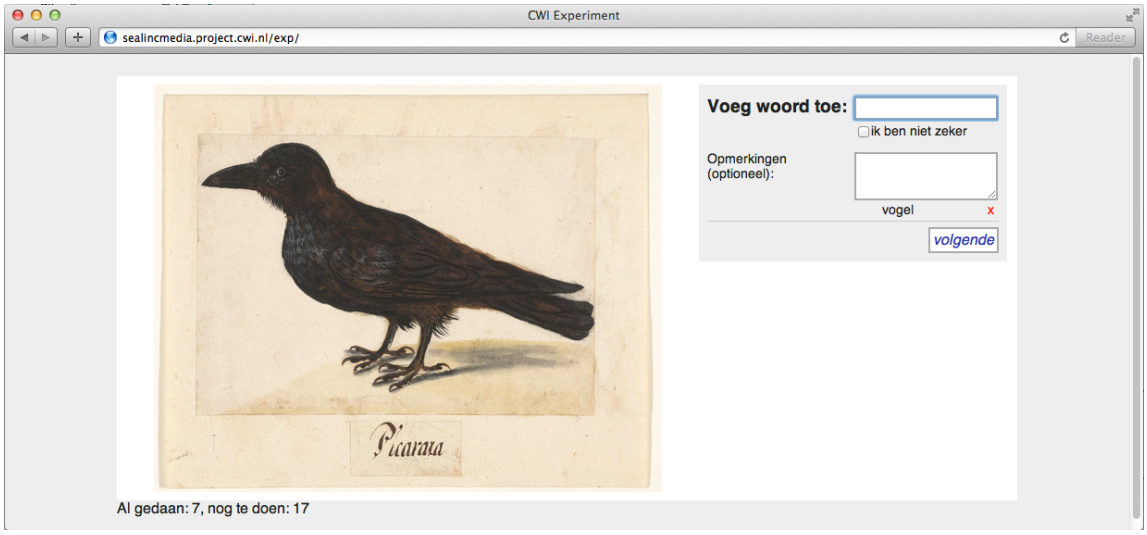

Figure 1: Screenshot of the online tagging interface used in the experiment. The Dutch "Voeg woord toe" translates into "add word" in English, "ik ben niet zeker" into "I am not sure" and "Opmerkingen (optioneel)" into "Comments (optional)". Image courtesy of Rijksmuseum Amsterdam, used with permission.

After having read the instructions, the participants can start tagging the first of 24 images. When finished, they are asked to fill in a demographic questionnaire.

4. RESULTS

In the first section, we focus on whether or not the participants added or deleted the generic, the correct specific tag or an incorrect specific tag to describe the object depicted in the image. After that we focus on the total number of tags that users added and deleted. Lastly, we discuss the quality of the tags that users added. In all these sections we will verify whether presenting different types of tags had an influence on the results.

Since there are no significant differences between the results for the photographs and the prints in all conditions in the three different analyses, we do not treat them separately in further analyses ($p > .001$ for the generic, correct specific and incorrect specific tags; $F < 1$ for the quantity of added tags; $p > .05$ for the quantity of deleted tags; and $p > .015$ for the quality of tags).

To analyze the results, a repeated-measures analysis of variance was used with the experimental conditions as within-images' factors and the proportion of generic, correct specific or incorrect specific tags as dependent variable. Which experimental conditions are taken into account and which type of tag, varies between analyses.

4.1 Generic, correct specific and incorrect specific tags

Here, we focus on whether or not the participants added the generic, the correct specific tag or an incorrect specific tag when describing the object depicted in the image. For each object, we only looked at users adding the one correct generic tag (the tag presented in the generic condition) and one correct specific tag (the tag presented in the correct specific condition). In contrast, we looked at all the incorrect specific tags that the user might have added. All other tags are not used in the analysis for this section.

For each participant and each image, we checked manually whether the generic tag, the correct specific tag or an incorrect specific tag was added or deleted[7].

To analyze the results, we calculated the proportion of generic, correct specific and incorrect specific tags for each of the 24 images.

A repeated-measures analysis of variance with three experimental conditions as within-images' factors (none, correct specific and incorrect specific) and the proportion of generic tags as dependent variable, clarifies that when no tag is presented to participants, the proportion of generic tags that they add is significantly higher than when a specific tag is presented ($F(2,46) = 7.189$, $p = .002$). This indicates that providing a specific tag has a negative effect on adding the generic tag compared to the condition in which no tag was presented. This may be because participants do not see the value of adding the generic tag to a presented specific tag. However, in 33.6% of all images, the generic tag was added when the specific tag was presented to the user, indicated that some users did find value in adding a more generic tag. An example of this is that when the tag "crow" was presented, several users still added the less informative tag "bird".

There is no significant difference between the proportion of added correct specific tags when no tag was presented and when the generic tag was presented ($F < 1$). This indicates that, unfortunately, providing a generic tag in our setting does *not* encourage people to provide the correct specific tag.

There is also no significant difference for the additions of incorrect tags whether or not the generic tag was presented

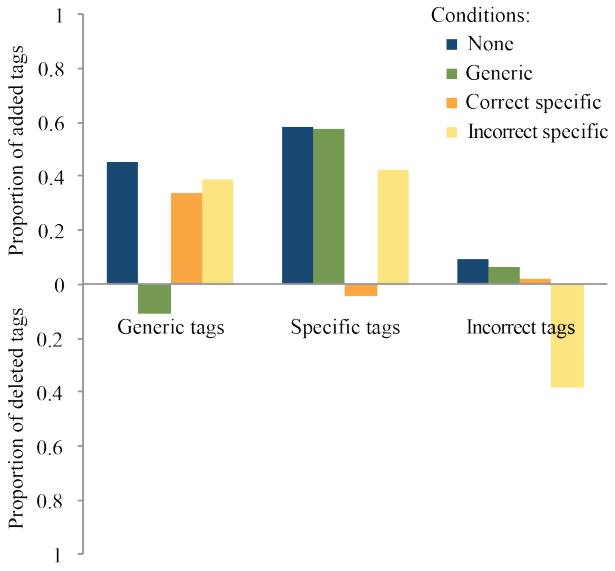

Figure 2: The proportion of generic, correct specific and incorrect specific tags that the participants added and deleted for all images in each of the four experimental conditions (none, generic, correct specific, incorrect specific).

($F(1,23) = 2.003$, $p = .170$).

All these findings (summarized in Figure 2), indicate that in this first analysis, we find no evidence that presenting existing tags to users is actually beneficial.

In the analysis above we only looked at tags explicitly added by the participants. It is also insightful to look at proportion of tag deletions (see Figure 2) and "submissions", which we define as the tags that remain when the user is finished tagging. We count a tag as submitted if it is added or given but not deleted. When the generic tag was given, the proportion of submitted generic tags is significantly higher ($F(3,69) = 112, 016$, $p < .001$) than the proportion of generic tags that were submitted when no tag, the correct specific tag or an incorrect specific tag was presented. It is also the case for the correct specific tag ($F(3,69) = 60.665$, $p < .001$) and the incorrect specific tag ($F(3,69) = 150.853$, $p < .001$). Overall, when a tag was given, the proportion of submitting that tag is higher in comparison to the proportion of adding that tag in the other conditions. This indicates that users were not inclined to delete tags that were presented to them. This is not a problem when the presented tag was the generic or the correct specific tag, but it is a problem when users refrain from deleting the incorrect specific tag. In several cases when the incorrect specific tag was presented, participants added the correct specific tag without removing the incorrect specific tag.

4.2 Quantity of tags

In the previous section we only made claims about the tags that were defined as generic, correct specific and incorrect specific tags describing the object depicted in the image. In this section we will take into account the total number of tags that the user added and deleted.

Table 1: Average number of all tags per image for each condition

	None	Generic	Correct specific	Incorrect specific
Presented tags	0.00	1.00	1.00	1.00
Added tags	2.37	1.81	1.72	1.99
Deleted tags	0.10	0.19	0.14	0.43
Total tags	2.27	2.62	2.58	2.56

There were strong differences between users. There is one user who added on average more than 10 tags per image and 40 users who added on average less than 2 tags per image. One user deleted 26 tags in total and another 0 (Average = 5.14, SD = 5.67). This indicates that the number of tags a user adds or deletes strongly depends on the individual.

The behavior of a user throughout the experiment was very consistent. A person who added a lot of tags for one image, also added a lot of tags for the other images. For this reason, and because each participant was presented the same number of images for each conditions, we are able to make general claims about the tagging behavior in the different conditions.

On average, participants added more tags when there was no tag presented to them then when there was a generic, correct specific or incorrect specific tag presented ($F(3,69) = 6.127$, $p = .001$). However, there is no significant difference between the four conditions for the total number of tags that a participant submitted ($F(3,69) = 2.671$, $p = .116$), which is calculated by subtracting the deleted tags from the sum of the given and added tags. An overview of these finding can be found in Table 1.

It is interesting to note that there are some deletions in the "none" conditions. This means that participants corrected tags that they entered themselves, probably because they made a spelling mistake. Deletions in the other conditions also include such corrections, and deletions of the given tag. The average number of tags that are deleted is higher when participants are presented with an incorrect tag ($F(3,69) = 39.464$, $p < .001$).

Only 2% of the tags that all users added were accompanied by a comment of the user. Analysis of the comments showed two kinds of comments: comments that were actually just other tags (e.g. "nature") and comments in which users explained why they choose to add the associated tag. When users elaborated on the reason for adding the tag, they sometimes added a link to the website as evidence to prove there tag was correct or mentioned the source that they consulted (e.g. "Used google images to check if I was right". For 26% of the tags that were deleted, a comment was added. The comment consisted out of tags that they thought were more suitable to describe the image (e.g. "This is not a magpie, but a crow") or to explain why they deleted that tag (e.g. "I made a spelling mistake"). Again, some participants provided a link to the website or mention the source they consulted, this time to prove or check they deleted tag was wrong. These findings indicate that in our setup, participants were more willing to comment on deletions than on

additions.

4.3 Quality of tags

In the previous section we looked at the total number of tags that users added. Of course, "more" does not necessarily mean "better" and therefore we will take into account the quality of tags that users added here. We will not discuss the quality of the generic, correct specific and incorrect specific tags since they have been discussed in section 4.1.

A list of all original tags that users added for the 24 images was reviewed by a native Dutch speaking person. She divided the tags into 6 different categories:

- Irrelevant: Nonsense tags that are irrelevant to such an extent that it is not even clear what the intention of the user was (e.g. "part of collection").
- Incorrect: Incorrect tags. Here the intention was clear, but most likely the user misjudged the object being depicted (e.g. "raven" when a crow was depicted).
- Subjective: Subjective tags that not everyone would agree on (e.g. "scary").
- Correct and possibly relevant: Correct tags that are not necessarily relevant for the image ("drawing").
- Correct and highly relevant: Correct tags that are highly relevant for the image (e.g. "beak").
- Spelling mistakes: Tags that are misspelled (e.g."b").

The tags of two randomly chosen photographs and two randomly chosen prints were reviewed by a second native Dutch speaking person. The overall agreement between the two reviewers according to Krippendorff's alpha is .536 on an nominal scale and .637 on an ordinal scale.

There is no significant difference of the number of tags between the experimental conditions (none, generic, correct specific, incorrect specific) for all the categories: irrelevant ($F < 1$), incorrect ($F(3,69) = 4.461$, $p = .006$), subjective ($F < 1$), correct and possibly relevant ($F < 1$), correct and highly relevant ($F < 1$).

From these findings we can conclude that the presence or absence of tags had no influence on the quality of the tags a user added (see Figure 3).

Users added much more correct tags than irrelevant, incorrect or subjective tags ($F(4,19) = 338.113$, $p < .001$)). There was no significant difference between the correct, possibly relevant tags and the correct, highly relevant tags ($F < 1$).

Several tags that users entered had spelling mistakes. There is no significant difference between the experimental conditions ($F(3,69) = 1.010$, $p = .394$). For some of these tags it was clear what the intention of the participant was, however, we did not correct these tags.

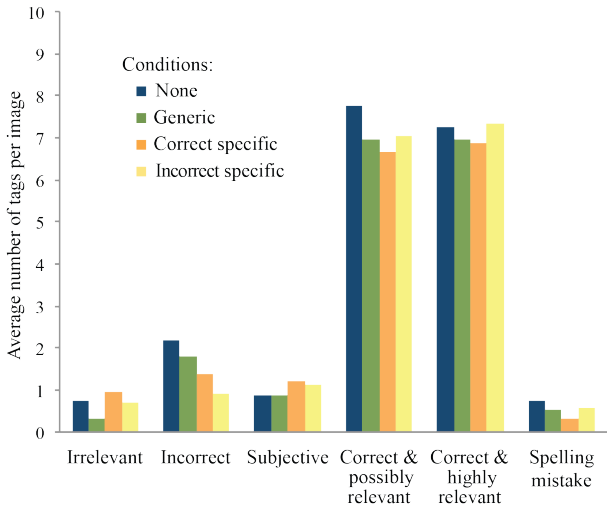

Figure 3: quality of all the tags that users added for an image in the four experimental conditions (none, generic, correct specific, incorrect specific).

5. DISCUSSION

In our experiment we did not find proof for the benefit of presenting tags on the quality and quantity of added tags.

Another finding is that users are very hesitant towards deleting tags. This is very problematic for the quality of the submitted tags when the tag that is presented to the user is an incorrect tag. Since the approach to ask users to delete tags they consider to be incorrect does not seem to work, we need to consider different approaches for quality assessment. One option would be that instead of letting users delete tags that they believe to be incorrect or inappropriate, we let them indicate whether or not they agree with tags that other users added.

The reason for our findings might be that the participants did not have any information about the provenance of the presented tags. For example, the interface might indicate that a tag originates from an authority in the field, from an employee of a cultural heritage institute, or that it has been generated by an automatic tool. It would be interesting to investigate whether users change their tagging behavior when such information is presented to them.

From the results it is clear that users comment more on deletions than on additions. This could be because the deletion of a tag activated a pop-up in which users could add a comment. These comments were very valuable since they often included the reasoning for deleting a tag with a link to the website they consulted.

When we compare the individual differences between the participants of our experiment in respect to the quantity of tags they entered, we find the same phenomenon as mentioned by [9]: We have very few "super-taggers" who enter an extraordinarily high number of tags for each image, but on the other hand, we have a large number of users who enter only very few tags.

The submission of tags that include spelling mistakes might partially be addressed by using autocompletion, allowing the user to select terms from a vocabulary. This requires selecting a vocabulary that suitable for the type of tagging required and matches the expertise of the user. Many vocabularies used in the cultural heritage domain require extensive knowledge of the complex structure of the vocabulary. Direct use of such vocabularies, as described in [4], may be effective in interfaces for professionals that are trained in using them, but less so for domain experts outside the cultural heritage institutes. The alternative, using common lexical dictionaries as in the Your Paintings project[5], has the disadvantage that it does not typically support named entities and domain-specific terms.

The images chosen for the experiment depict objects for which non-experts were expected to know the generic and specific tags that describe the depicted object and therefore we used non-expert users as participants. However, incorrect tags were added in each condition, indicating that adding the correct specific tag was more difficult than expected. This shows that including experts as users is desirable in a tagging system.

6. IMPLEMENTATION

The tagging interface is available as open source software and uses open Web standards where possible. The interface has been implemented on top of the ClioPatria Semantic Web application platform [11] as the `image_annotation` package[8]. It uses the Open Annotation format [6] to store all annotations in RDF, using the tag as a literal annotation body and the URI of the image as the annotation target.

The current implementation of the package has been updated with the insights obtained from this paper. It uses true tag deletion primarily to let users remove their own erroneous tags, while it enables taggers to rate tags of other users by agreeing or disagreeing with a tag. Users can also comment on all tags, or mark tags as questionable. Since all Open Annotations have a unique URI, such (dis)agreements, comments and questionable tags are simply implemented as annotations on annotations, using the comment or rating as the annotation body and the URI of the original annotation as the annotation target.

The Open Annotation format provides a commonly agreed upon format that allows us to store the URI representing the user doing the annotation and the time of annotation.

In addition to the simple single field interface used in this experiment, the package supports more complex tagging interfaces using multiple fields. Each tagging field is configurable as a free text field, or linked to a SKOS vocabulary from which the concepts can be used for autocompletion.

The interface is multilingual and is fully configurable in RDF using the patterns defined in [3].

7. CONCLUSION

The most important findings from the experiment are that presenting existing tags does not have a positive influence on user image tagging and that users are hesitant to delete existing incorrect tags, even when explicitly asked to do so.

The comments added by participants were useful since they sometimes included the reason why participants chose to add or delete a tag. Comments were more often added when users deleted a tag, than when users added a tag.

The overall aim of our research project is to design a tagging interface that is suitable for non-professionals, so non-experts in terms of the vocabularies and tagging systems, but experts in a particular domain. For that we will investigate the influences of further elements of the UI on the tagging behavior of users and the quality of the resulting annotations.

We will compare different methods of quality judgment by users (e.g. letting users indicate whether or not they agree with different tags instead of deleting these tags) and we will also investigate whether showing metadata of the presented tags has an influence on the tagging behavior of users. Furthermore, we will explore to what extent we can use vocabularies available on the Web (e.g. in SKOS) and in the cultural heritage institutes to provide suggestions in an autocompletion interface.

The results of the experiment described in this paper have already been taken into account for the prototype[9] that is being developed in the SEALINCMedia project. Thus far, the study has proved to be valuable for the project and the findings will be complimented by further studies on other UI elements.

Acknowledgments

This research is carried out in the context of the SEALINCMedia project[2] in the COMMIT research program[1]. We thank Archana Nottamkandath and Davide Ceolin for their constructive feedback on the qualitative analysis of the tags, and Jasper Oosterman and Chris Dijkshoorn for their contributions to the SEALINCMedia platform.

We would also like to thank all project members (especially Rijksmuseum Amsterdam[10]) for their contributions on the user interface and all participants in the experiment for their collaboration.

8. REFERENCES

[1] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 971–980, New York, NY, USA, 2007. ACM.

[2] A. Ellis, D. Gluckman, A. Cooper, and A. Greg. Your paintings: A nation's oil paintings go online, tagged by the public. In *Museums and the Web 2012*, 2012.

[3] M. Hildebrand and J. R. van Ossenbruggen. Configuring Semantic Web Interfaces By Data Mapping. In *Proceedings of the VISSW 2009*

[8] `http://cliopatria.swi-prolog.org/packs/image_annotation`

[9] `http://rma-accurator.appspot.com`

[10] `https://www.rijksmuseum.nl`

Workshop: Visual Interfaces to the Social and the Semantic Web, Proceedings of the VISSW Workshop: Visual Interfaces to the Social and the Semantic Web, February 2009.

[4] M. Hildebrand, J. R. van Ossenbruggen, L. Hardman, and G. Jacobs. Supporting subject matter annotation using heterogeneous thesauri, a user study in web data reuse. *International Journal of Human-Computer Studies*, 67(10):888 – 903, October 2009.

[5] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proceedings of the seventeenth conference on Hypertext and hypermedia*, HYPERTEXT '06, pages 31–40, New York, NY, USA, 2006. ACM.

[6] R. Sanderson, P. Ciccarese, H. V. de Sompel, T. Clark, T. Cole, J. Hunter, and N. Fraistat. Open annotation core data model. Technical report, W3C Community Draft, work in progress, May 9 2012.

[7] G. Schreiber, I. I. Blok, D. Carlier, W. P. C. v. Gent, J. Hokstam, and U. Roos. A mini-experiment in semantic annotation. In *Proceedings of the First International Semantic Web Conference on The Semantic Web*, ISWC '02, pages 404–408, London, UK, UK, 2002. Springer-Verlag.

[8] S. Shatford. Analyzing the subject of a picture: A theoretical approach. *Cataloging & Classification Quarterly*, 6(3):39–62, 1986.

[9] J. Trant. Tagging, folksonomy and art museums: Results of steve.museum's research. Technical report, Archives & Museum Informatics, January 2009.

[10] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 319–326, New York, NY, USA, 2004. ACM.

[11] J. Wielemaker, M. Hildebrand, J. R. van Ossenbruggen, and G. Schreiber. Thesaurus-Based Search In Large Heterogeneous Collections. In A. Sheth and et al, editors, *The Semantic Web - ISWC 2008*, volume 5318 of *Lecture Notes in Computer Science*, pages 483 – 498. Springer, October 2008.