Crowdsourcing Knowledge-Intensive Tasks
In Cultural Heritage

Jasper Oosterman∗, Archana Nottamkandath†, Chris Dijkshoorn†,
Alessandro Bozzon∗, Geert-Jan Houben∗, Lora Aroyo†

∗Delft University of Technology
Web Information Systems
P.O. Box 5, 2600 AA
Delft, The Netherlands
{j.e.g.oosterman,a.bozzon,
g.j.p.m.houben}@tudelft.nl
†VU University Amsterdam
De Boelelaan 1081a
1081 HV Amsterdam
Amsterdam, The Netherlands
{a.nottamkandath,c.r.dijkshoorn,
lora.aroyo}@vu.nl

ABSTRACT
Large datasets such as Cultural Heritage collections require detailed annotations when digitised and made available online. Annotating different aspects of such collections requires a variety of knowledge and expertise which is not always possessed by the collection curators. Artwork annotation is an example of a knowledge intensive image annotation task, i.e. a task that demands annotators to have domain-specific knowledge in order to be successfully completed. This paper describes the results of a study aimed at investigating the applicability of crowdsourcing techniques to knowledge intensive image annotation tasks. We observed a clear relationship between the annotation difficulty of an image, in terms of number of items to identify and annotate, and the performance of the recruited workers.

Categories and Subject Descriptors
H.4 [Information Systems Applications]: Miscellaneous

Keywords
Crowdsourcing; Cultural Heritage; Knowledge Intensive Tasks

1. INTRODUCTION
Some crowdsourcing tasks, referred to as knowledge intensive tasks[1], require specific knowledge to be successfully executed. An example is the annotation of flower images with the correct botanical names. Typically, cultural heritage institutions employ professionals, mostly art historians, who provide excellent annotations about the art-historical aspects of artworks, but lack domain expertise for other aspects such as the name of the depicted flowers. For these other aspects, people with the right domain expertise need to be found and engaged, a process called nichesourcing [2]. How to do nichesourcing effectively is an emerging research area.

Image annotation tasks require annotators to provide tags that describe the presence of one or more instance of a given generic entity (e.g. flowers, birds, etc.). This class of tasks is typically performed on photographic images, which carefully represent the real world; they are also quickly performed, as the knowledge and skills required for their execution are commonly available in the general population (e.g. the ability to recognise a flower). We claim the annotation of artworks to be a more involved process, especially when coupled with the need for accurate, possibly domain-specific labels. It is a good example of a knowledge intensive image annotation task, that features entity identification as an important source of complexity. The correct recognition of the targeted entities in an art image might be hindered by its artistic representation; for instance, colours or details might be missing; or the depicted content might be stylised or even completely abstract. This additional complexity increases the cognitive effort for an annotator by needing to 1) have the ability to “see” the raw content through the lens of a given style or technique; and 2) possess domain-specific knowledge.

In a previous work[1] we showed that niche sourcing can help with the creation of useful annotations in the domain of cultural heritage. In this paper, we study the performance of workers employed via crowdsourcing platforms for annotating a collection of prints depicting flowers from the Rijksmuseum Amsterdam1. We have created a testbed, similar to existing image annotation interfaces, where workers could tag prints with the specific names of flowers. The testbed measured the annotation behaviour of the workers and stored the annotations. In our study we try to answer the following questions: What is the relation between entity identification difficulty and crowd annotation behaviour?

2. EXPERIMENTAL SETUP
We manually selected 82 prints that were annotated with the Flowers taxonomy term, i.e. prints that were identified as containing at least one flower. A team of 3 trusted annotators inspected each print, labelling it as Prominent (P) (17 prints) or Non Prominent (NP) (65 prints), counting the number of contained flower instances, and counting the number of flower types. Despite the full focus and attention put in this annotation process, some prints proved challenging to label. To account for such difficulty, we classify prints

1http://rijksmuseum.nl
in three categories, according to the disagreement shown by the trusted annotators when counting the number of identified flower instances and flower types. Respectively: Easy: prints where all trusted annotators agreed on the same count; Average: prints where trusted annotators disagreed by 1 or 2 flowers/types; Hard: prints where the trusted annotators disagreed by more than 3 flowers/types. Table 1 reports the distribution of number of prints according to their combined identification difficulty.

The experiment was performed in February 2014, and it involved anonymous annotators drawn from the CrowdFlower² platform, which recruits workers worldwide. Of 732 workers that inspected the published task, 151 attempted execution (abandonment rate of 80%), and only 44 passed the qualification test and the designed quality checks. An extended description of the experimental setup and the resulting data is available online³.

Each print annotation required workers to indicate the number of flower instances, and the number of flower types they were able to identify. We compared the provided numbers with the corresponding assessment performed by our trusted annotators. With easy prints, we considered correct only values equal to the ground-truth ones. With average and hard prints, we considered correct values falling in the range of disagreement of the trusted annotators. Figure 1a reports the average precision of workers, aggregated according to the country of belonging (Western – Western Europe, plus USA and Canada – and non-western). Figure 1b and Figure 1c depicts the average error rate of workers according to the flower identification difficulty and to the flower prominence on the print. Flower type identification is generally performed more precisely than flower number identification, although hard and non-prominent prints generally lead to more errors. Flower type identification was also performed more precisely by Western workers, but errors affect every category of prints in a similarly distributed manner.

Acknowledgements. This publication was supported by the Dutch national program COMMIT.

4. REFERENCES


--

²http://crowdflower.com
³http://www.wis.ewi.tudelft.nl/WebScience2014