

Digitalkoot: Making Old Archives Accessible Using Crowdsourcing

Otto Chrons and Sami Sundell

Microtask

Bulevardi 1

00100, Helsinki, Finland

otto.chrons@microtask.com, sami.sundell@microtask.com

Abstract

In this paper, we present Digitalkoot, a system for fixing errors in the Optical Character Recognition (OCR) process of old texts through the use of human computation. By turning the work into simple games, we are able to attract a great number of volunteers to donate their time and cognitive capacity for the cause. Our analysis shows how untrained people can reach very high accuracy through the use of crowdsourcing. Furthermore we analyze the effect of social media and gender on participation levels and the amount of work accomplished.

Introduction

OCR problems in old archives

National libraries and other keepers of precious old archives have been busily converting material from paper and microfilm into digital domain. Newspapers, books, journals and even individual letters are finding themselves inside large scanners that try to reproduce the content to its finest detail. Reasons for these multi-million dollar projects are simple. Accessing fragile original documents is cumbersome and deteriorates the originals every time they are brought out from the storage rooms. Finding relevant information from these old archives is tedious and sometimes downright impossible. Having all that content in a digital format allows researchers and even laypeople to easily access all the information from the comfort of their desks.

After the scanning phase the archivists need to further process the image data into a structured format such as METS/ALTO (<http://www.loc.gov/ndnp/techspecs.html>) for it to be truly accessible and useful. The simplest approach is to perform full-page OCR, which gives searchability but provides little in the structure of the content. This is especially true with materials such as newspapers where each page contains multiple articles, advertisements and pictures. Software solutions are available for structuring the content with the help of human operators.

In this paper we focus on the problems of OCR with old newspaper material. Even though OCR algorithms can cope very well with modern typesetting, they struggle when presented with old material with strange typefaces, bad scan

quality or smudged originals. Our reference material was newspapers from the late 19th century, using a Fraktur typeface, in Finnish.

There have been other efforts in fixing bad OCR results before, most notably the reCAPTCHA project (von Ahn 2008), National Library of Australia's Trove project (<http://trove.nla.gov.au/newspaper>) and IMPACT EU (Balk 2009) in Europe. In all of these projects, computer algorithms have been augmented by the use of human computation to improve the results of the OCR process.

Distributed work

The traditional way of managing the library digitization projects has consisted of buying a lot of equipment and hiring personnel to run the program, or to outsource the whole process to a third party. Using these custom tools requires training and a skilled workforce. We show in this paper that some parts of that process can be distributed to a pool of unskilled volunteers with good results.

The concept of distributed work, in the context of very small *microtasks*, is relatively new. The rise of fast networks and cheap terminals created the opportunity to break down complex tasks and send them to a distributed workforce for processing. This was dubbed crowdsourcing by Jeff Howes (Howes 2006). Applications of crowdsourcing range from very creative design competitions to the mundane microtasking in reCAPTCHA.

Gamification

Since people don't like working on mundane activities for long but can waste hours in playing rather simple games, the idea of combining these two has been found to be an effective way to motivate people. Turning useful activities into games is called gamification and it has found its way into many uses such as education (Squire and Jenkins 2003), image tagging (von Ahn 2006) and many others (Ho, C-J. et al. 2009; Chandrasekar et al. 2009; Chang et al. 2010). Gamification works best when activities can be rewarded with scores, achievements or social benefits.

Digitalkoot concept

The concept came to be as we approached the National Library of Finland (NLF, <http://www.nationallibrary.fi>) with

an idea of improving their newspaper archives through the use of crowdsourcing. Since they are a government organization, they have very limited budget and it was obvious that participants could not be compensated monetarily. Finnish language has the word "talkoot", which is described in Wikipedia as "... a Finnish custom involving a group of people gathering to work together unpaid, for instance to build or repair something". What could be a better name for a concept, where volunteers gather together in a digital environment to repair newspaper archives, than Digitalkoot (<http://www.digitalkoot.fi>).

The digitization process used at the NLF has many steps, many of which would be suitable for crowdsourcing. We decided to start with the most easily accessible step, which is fixing OCR errors, as it could be done on material that had already gone through the whole digitization process. The NLF and many other libraries use an archive format known as METS/ALTO that has the benefit of storing not only the original images and OCR processed text but also the coordinates of each word and even a confidence value for every character and word. This allows the Digitalkoot system to crop individual words out of the newspaper page and send them as microtasks (figure 1) to volunteer workers. Since the NLF had already processed a few million newspaper pages into METS/ALTO format, we had no shortage of material.



Figure 1: A single microtask

The National Library of Australia (NLA) had launched a similar crowdsourcing effort in August 2008. Their approach is to show full articles to volunteers, who then fix them line by line. We decided to break the material into single words, losing all context and relevance to the original text. Main driver for this model was the desire to turn the whole thing into games.

Gaming the OCR fixing

Creating a good gamification concept requires balancing game play elements with task completion speed and accuracy. Game play is very important to keep people motivated in the game and to tap into a large population of potential gamers. The great challenge is to introduce some meaningful tasks into the game without breaking the game play mechanisms. One challenge that is especially applicable to crowdsourcing is the ability to give real-time feedback on player actions. In many distributed work schemes the task is verified by having many people do it (optimally at the same time) and then comparing the results. In reality this introduces a latency in the magnitude of several seconds or even minutes if the participation is low. Thus it puts pressure to design game concepts that can naturally adapt to varying feedback latencies.

To describe how we solved these issues, let's look at the games we built. Since we had very limited time to create the games we decided to go for simple and familiar concepts. The first game is intended for verifying OCR results and is called Mole Hunt (figure 2). The second game, Mole Bridge, is a bit more complex and involves typing in words that are shown to the player (figure 3).



Figure 2: Verifying OCR results in the Mole Hunt game

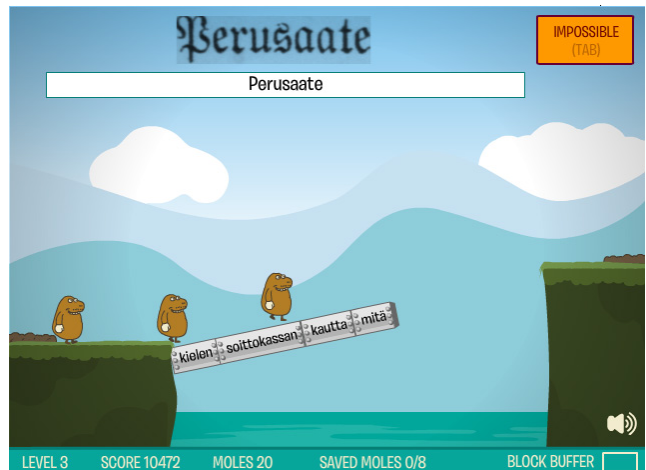


Figure 3: Performing human OCR in the Mole Bridge game

In Mole Hunt the player whacks moles (as in the Whack-a-mole game) by looking at the original word and OCR result of it, and comparing the two. The result of this task is simply a boolean value. Players are given instant feedback in the form of the mole disappearing from the scene, but evaluating the validity of the answer is postponed until the level is completed. We use a simple metaphor of "growing flowers" to indicate how well the player did in the game. For every correct task the flower blooms, for bad answers the mole eats the flower.

When playing Mole Bridge the goal of the player is to build a bridge to save moles from falling down. The bridge is composed of blocks and the player can create blocks by typing words shown to her. As with Mole Hunt there is an immediate feedback to typing words as blocks get appended to the bridge. Later, as the system is able to determine the correctness of the answer, wooden blocks transmute into steel (in case of correct answer) or explode (if wrong) taking a few neighboring blocks along. Once the bridge is complete and enough moles are saved, the level ends and the final score is calculated. In both games players are rewarded for correct answers and punished for incorrect ones.

In an ideal world there would be a lot of concurrent players, working on the same tasks with very low latencies, to figure out the results. In reality this is rarely the case, especially since every player advances at their own pace and thus they cannot be synchronized like in the ESP game. To alleviate this problem we introduce verification tasks to the task stream. Similar approach is used in reCAPTCHA, where one of the two words displayed is a verification word. These are special tasks that look just like the regular ones with the exception that the system already knows the correct answer. With the help of these pre-verified tasks the system can lower the latency of giving feedback and the game play is improved. The downside is that these tasks are not producing any new information. The rate at which verification tasks are fed to players varies according to the number of active players. This reduces the cost of lost work when there is a sufficient number of concurrent players.

There is, however, another use for the same verification tasks. When a new player enters the game, the system has no way of knowing the skill (or the benevolence!) of that player. If it's someone who just wants to fool around by answering every task with "asdf", it would be seriously detrimental to the system as a whole. As the system trusts the "wisdom of the crowds", it has to take into consideration also answers from these spammers, which results in additional task creation for other players to achieve good enough total confidence for the answer. To trap these problematic players the system begins the game by feeding the player only verification tasks, effectively keeping them in a sandbox. Once the player has proven to be doing proper work, the ratio of verification tasks is lowered in phases. The ratio never goes to zero so there is always a bit of monitoring going on in the system. The player is not informed about this mechanism in any way, so she has no way of knowing when to play "seriously" and when it would be ok to "cheat". In essence the system trusts no-one in the beginning, but players can earn the trust by playing properly.

To make the verification process fully automatic, the verification tasks themselves are created automatically. Every now and then a task is selected to become a candidate for a verification task. It is sent to several players (seven in our case) and only if all of them agree on the result, a new verification task is created. Note that it is entirely possible for all the seven people to make the same mistake and thus create an erroneous verification task. In the greater scheme of things this has very little impact on the performance or accuracy of the system.

The games utilize levels to break down the game play into manageable chunks. This creates a more diversified experience and gives players new challenges over time. Level design is a very delicate subject as the players have very different typing skills. Some level could be really easy for a good typist but would prove to be impossible for the elderly lady who just wants to pitch in for the sake of the National Library. In the two games we allow players to freely go back to previous levels or to retry the current level if they find it too difficult.

Results

Data was collected from the Digitalkoot system for 51 days from its launch on February 8th 2011 until March 31st. During this time, the site had 31,816 visitors and 4,768 people at least tried out one of the games. These users donated over 2,740 hours of effective game time and completed 2.5 million tasks. The typical Digitalkoot user spent 9.3 minutes on the games and completed 118 tasks.

User demographics

Because most of the users authenticated themselves using Facebook, we were able to gather some simple demographics of those users. Gender distribution of users is shown in Figure 4.

In the beginning, by far the largest complaint about the system was related to the mandatory Facebook login. This was remedied after the first two weeks by adding the possibility of e-mail login. In the end, however, only about one hundred users created an account using the e-mail option. Performance of these users was on the same level with Facebook users.

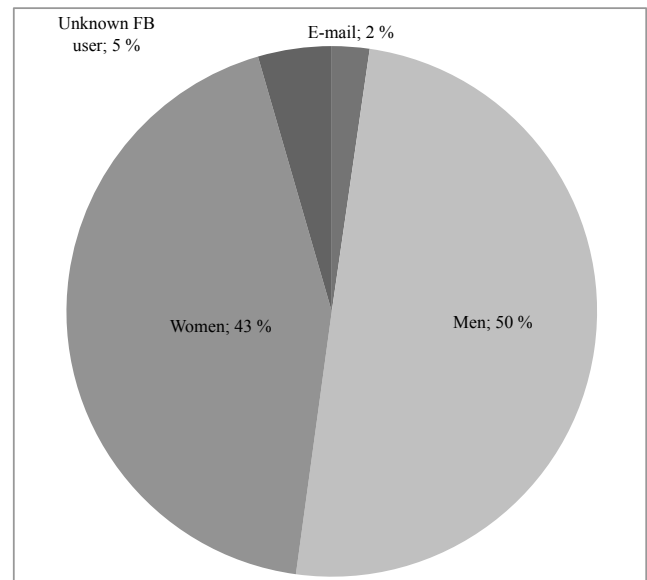


Figure 4: Half of the Digitalkoot users are male.

User contribution

Digitalkoot as a whole has been a huge contribution of time and work by thousands of individual users. In total, users have spent 2,740 hours playing the two games the site offers them. This has resulted in almost 2.5 million completed microtasks.

The amount of game play varied from trying a single game for a few seconds up to an individual spending more than one hundred hours trying to keep his name on the daily top lists. Median time spent on the games was 9 minutes 18 seconds. Even though men were more eager to join the cause, it was women who spent their time playing: the median time for female players was significantly higher than for the general population at 13 minutes 45 seconds (Figure 5). This is also reflected in the amount of tasks done. Median task count for female users was almost double that of men (Figure 6), and they did 54% of all tasks.

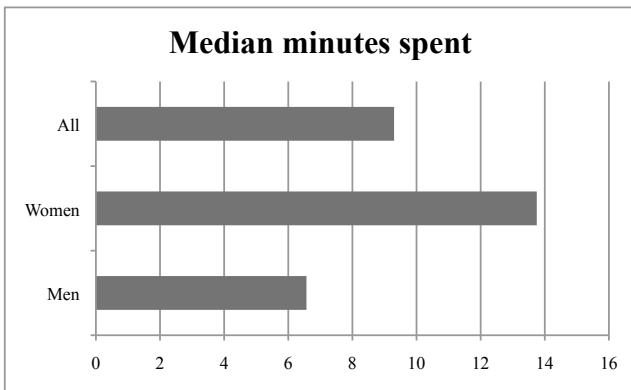


Figure 5: Female users tend to spend more time working in Digitalkoot.

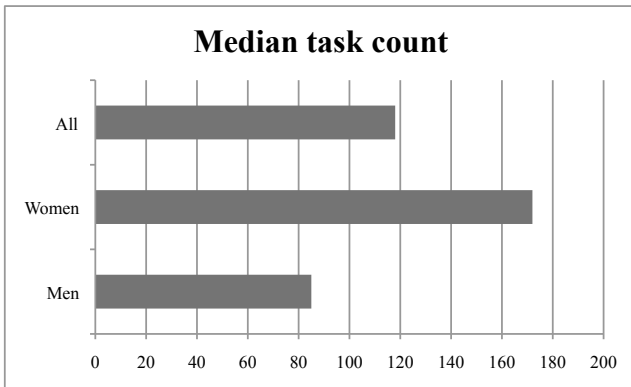


Figure 6: Median task counts reflect the amount of time spent playing games.

When it comes to actual accomplished work, the results represent an extreme case of the Pareto principle: the most active one percent of the users contributed almost one third of the total work (figure 7). Even if men contributed less

than women in general, this is where they got to shine: the hardest-working top 4 were all men, with the most active one completing almost 75,000 tasks in 101 hours of work.

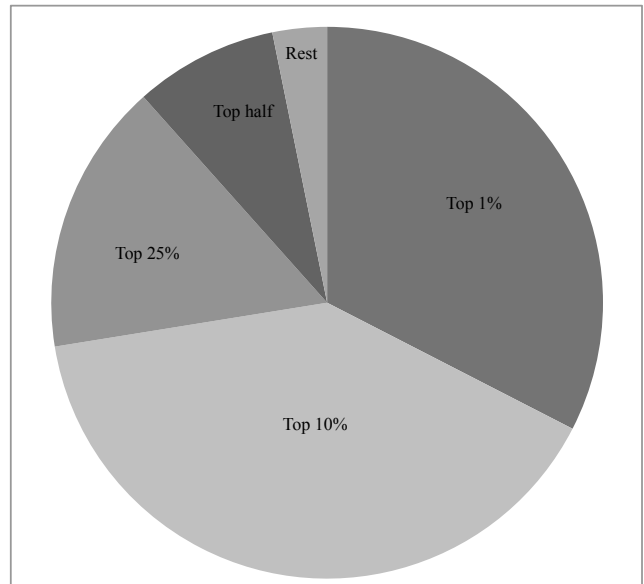


Figure 7: The hardest-working percent of the workforce did almost third of the work.

Accuracy

The source material used in Digitalkoot contained both the OCR version of the text as well as a confidence value for each word separately. This confidence value tells how certain the OCR system is about the word on a scale from 0.0 to 1.0. To estimate the accuracy of this confidence value we used a sample of about 25,000 words that were put into the system using high redundancy, so that multiple users would verify the words. When comparing Digitalkoot results for these 25,000 words and their OCR confidence, it appears that even when the OCR system is highly confident about the correctness (confidence higher than 0.8), up to 30% of the words ended up not matching the words written by human users.

The accuracy of completed work was gauged by randomly selecting two long articles from the processed newspapers and manually calculating the number of mistakes found first in the OCR and then in the article fixed by the Digitalkoot effort. The result was staggering: in a sample article of 1,467 words, Digitalkoot had produced only 14 mistakes. Another article of 516 words was even higher in quality, when only one single word in the whole article was considered wrong. By comparison, the OCR process had made a mistake in 228 words of the first article, and the second article had 118 mistakes after OCR. In other words, whereas OCR systems struggle to get pass 85% mark in accuracy, it seems possible to achieve well over 99% accuracy in digitizing words by simply playing games.

In a public system, there's always a possibility of foul

play. Digitalkoot system was built to use known verification tasks to both monitor the user performance and to keep up the game play. The effectiveness of this approach to keep off the malevolent users can be best described with anecdotal evidence: one particular user played for almost 1.5 hours and completed 5,692 tasks. However, the verification system detected that this particular user was sending mostly garbage – responses with only empty spaces, random characters, and wrong answers in Mole Hunt. As a result, only four out of all those tasks were considered for real. While this particular user was happily bounding his or her keyboard on verification tasks, the rest of the world was unaffected and continued to produce actual results by completing proper tasks.

In total, 10,324 game sessions was played in Digitalkoot, and 479 of those had more incorrect than correct answers. Superficial inspection suggests that malevolent users are even rarer than that: it seems that in many of these error-ridden cases the user just had systematic problems in recognizing Fraktur typeface.

Social aspects

Digitalkoot project was launched with Facebook integration, and that also shows in the user profile; 98% of users authenticated themselves using Facebook.

To gauge the social aspects of the effort – whether people are sharing their interest in Digitalkoot with their friends in Facebook – we measured the amount of friends who followed a new user into Digitalkoot. It should be noted that this certainly is not a clear indicator of social sharing. Heavy media coverage during the launch lured lots of people to joining independently of each other, so there may not always be any causal link between friends joining the effort.

Having said that, it's still interesting to see that during their first week of membership, 1,756 people had some friends joining in. This is more than a third of all Digitalkoot users logging in with Facebook. Of course this can also be turned around and note that two thirds of the user base do not show any indication of social interaction regarding Digitalkoot. For most users, the day of their registration was the most active when it comes to sharing information about the service to their friends, and only 341 people had friends joining after their first week in Digitalkoot.

Figure 8 shows that most of the people had only a few friends joining Digitalkoot, but the amount of social connections is still significant. Number of friends in Digitalkoot also reflects the size of person's social network: it's quite likely that the people with largest social network also benefit from the happy coincidence of their friends joining Digitalkoot of their own accord.

Media coverage and its effects

Because of its unique approach to improving newspaper archives, Digitalkoot garnered quite a lot of media publicity: up until the end of March, there had been more than 30 articles about Digitalkoot in various newspapers, magazines and online publications, as well as several national television appearances.

Every major appearance in the media also had its effect on the amount of new users, as can be seen from figure 9. The

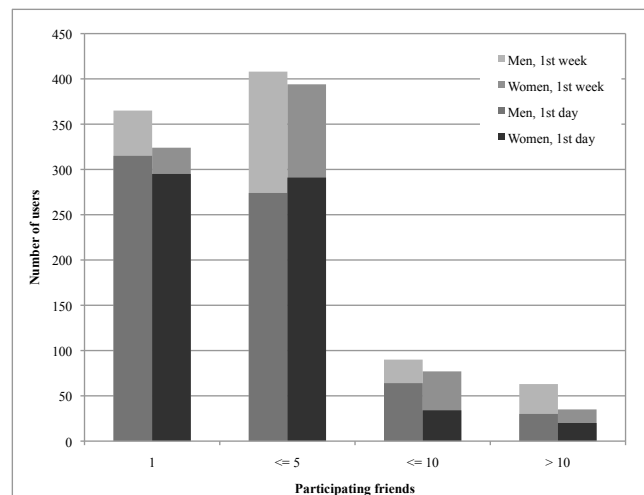


Figure 8: Third of the Facebook users had friends joining during their first week. Columns indicate users having a certain number of friends participating after the first day and the first week.

launch date obviously gathered the most prominent spike of new users, but there were several noticeable jumps in the user count after that. For example, on February 15th Digitalkoot was introduced in a national radio broadcast and appeared in several newspapers, which resulted in almost 200 new users. Around March 15th the effort was in local business newspapers and on 17th it was featured in a Wired.com column. On March 23rd Digitalkoot appeared in New York Times. All these articles raised public interest, which was mirrored in the site's user count.

Although the media coverage definitely helped in raising awareness of the effort and the games, there still needs to be a steady undercurrent of regular users to keep the effort going. During the inspection period, the amount of players stabilized into 300 individual users per week.

Conclusions and future work

When we started this project, we had no clue how it would succeed. Old newspaper archives aren't really the hottest topic in the modern social media circles and National Library sounds like a quiet room full of old, dusty books. Therefore it was a big surprise to see how well the Digitalkoot project was received by the media and especially by the general public. Obtaining almost 5,000 users who pitched over 2,740 hours of their free time in seven weeks is quite an accomplishment in a country with a population of 5.3 million. Technically the system worked almost flawlessly, requiring minimal maintenance, mainly adding new tasks to the pool from time to time.

Since the source material was especially difficult to read, even for humans, achieving over 99% accuracy was unexpected. It should be emphasized that the old Fraktur typeface is difficult to read to modern people and our workers were by no means professionals of historical texts. Using crowdsourcing may even have resulted in better quality than any

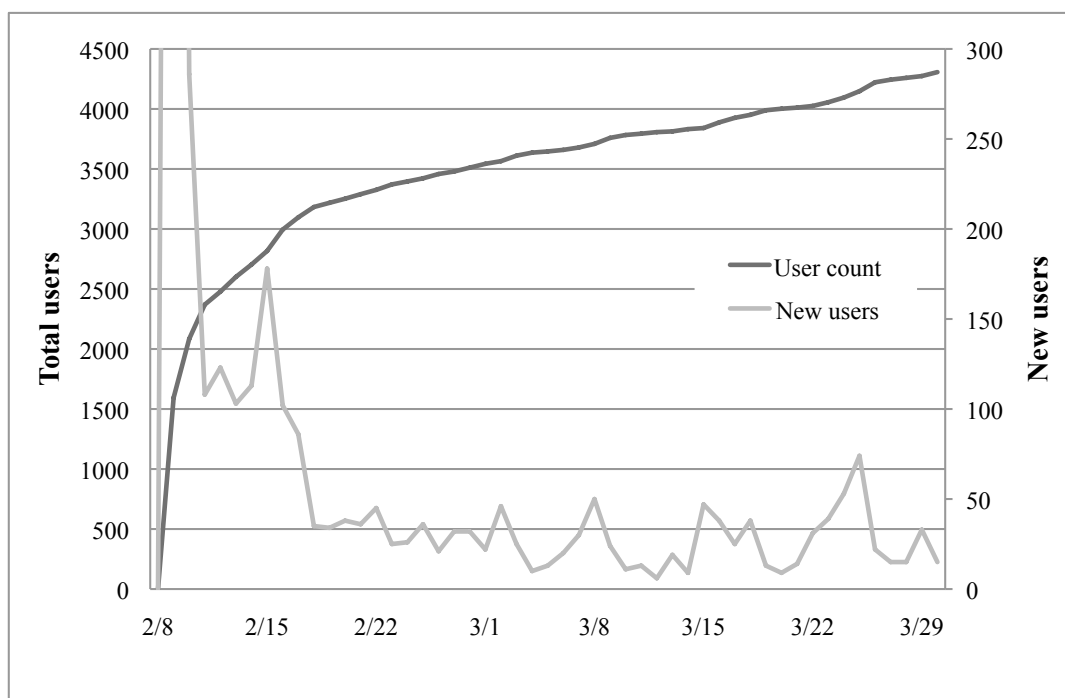


Figure 9: Digitalkoot user count over time.

single (professional) individual would have accomplished. At least the improvement over the original OCR was notable.

We had configured the system to produce a lot of redundancy and therefore the actual results, fixed words, were not as numerous as one would assume based on the number of tasks completed. With this information, however, we can now safely tune those configuration parameters to increase the yield, without sacrificing quality. An important nugget of data that we wished to obtain was to determine whether it makes sense to filter OCR results with another game before typing the words. The idea was to use the Mole Hunt game as a filter, to drop the words that OCR had identified correctly and leave only the incorrect words to the Mole Bridge game. As the source data was of such bad quality, it doesn't really make much sense to do this filtering, since so many words would go through anyway. Nevertheless, with better quality material, it would be an important optimization for the total performance of the system.

In addition to the basic games we had also designed social games on top of them. Unfortunately we didn't have the time to implement achievements, badges, competitions, top-lists and such, to increase the stickyness of the system. Adding these features would be an interesting experiment and we plan to do so in the future.

The games themselves are relatively simple and experience some difficulties due to their real-time nature. Other kinds of games with longer natural feedback times would suit better and would require less verification tasks to work properly.

Fixing OCR errors is just the first step in crowdsourcing

the archive digitization process. We have analyzed together with the NLF other aspects of that process and have identified several steps that could utilize the human computation power of a volunteer crowd. Now that the concept has been proven to work, it's time to focus on more complex and challenging things to help turn cultural treasures into a digital and easy-to-access format.

References

- Balk, H. 2009. Poor access to digitised historical texts: the solutions of the impact project. In *AND '09*, 1–1.
- Howes, J. 2006. The rise of crowdsourcing. *Wired* 14.
- Squire, K., and Jenkins, H. 2003. Harnessing the power of games in education. *InSight* 3.
- von Ahn, L. et al. 2008. recaptcha: Human-based character recognition via web security measures. *Science* 1465–1468.
- von Ahn, L. 2006. Games with a purpose. *IEEE Computer Magazine* 39(6):92–94.
- Ho, C-J. et al. 2009. KissKissBan: A Competitive Human Computation Game for Image Annotation. *Human Computation Workshop* 11–14.
- Chandrasekar R., Quirk C., Ma H., Gupta A. 2009. Page Hunt: Using Human Computation Games to Improve Web Search *Human Computation Workshop* 27–28.
- Chang T-H., Chan C-w., Hsu J. 2010. Human Computation Games for Commonsense Data Verification *AAAI Fall Symposium* 19–20.