Observatory of Public Sector Innovation

OECD BETTER POLICIES FOR BETTER LIVES

Horizon 2020 European Union Funding for Research & Innovation

Digitising Finnish history using crowdsourced volunteers

Received via Governance International

Published On: 18 November 2015

Organisation: Microtask

Country: Finland

Level of government: Central government

Sector: Recreation, culture and religion

Type: Digital, Methods, Public Service

Launched in: 2011

Overall development time: 1 year(s) 6 month(s)

Link to the innovation's website

Description

Once upon a time, gathering historical information meant visiting your local library, museum or university and painstakingly sifting through old records. Today, using the power of the internet, huge amounts of information is available and searchable at the click of a button. More and more data is being made available online by institutions every day, allowing unprecedented access to historical information.

But while it is simple enough for an institution to upload information that is already in digital format, converting hard copies of materials can take a long time, especially when it is old or in poor condition.

Today excellent software exists that allows computers to scan printed text and convert it into digital format. Unfortunately, when the source material is hard to read, this optical character recognition (OCR) software can make lots of mistakes. These mistakes make the material less useful and reliable when it is online, especially when people are using search engines to try to sift through it.

At the moment, such mistakes can only be corrected by people. The human brain is much better than machines at transcribing hard-to-read text. This makes it very expensive for institutions to accurately digitize large archives. If you have many millions of pages of material, hiring full time staff to correct all the errors could cost millions.

This Finnish solution, developed through a partnership between the National Library of Finland and crowdsourcing company Microtask, was to use online volunteers to fix the mistakes. The project is called Digitalkoot, and has so far encouraged more than 100,000 volunteers to donate over 400,000 minutes of time to helping correct OCR errors in the online version of its historical newspaper archive.

Why the innovation was developed

- The National Library of Finland is responsible for the collection, preservation and accessibility of Finland’s printed national heritage and the unique collections under its care. By digitizing this material and making it freely available and searchable online, the National Library aims to enhance the visibility, accessibility and usability of its unique collections. Once it is online it also means that the precious hard copies of the documents do not have to be used by people, which can damage them.
- At the moment the National Library has around 4 million digitized pages in its newspaper archive. This material was digitized using optical recognition software (OCR). The problem is that because much of the material is old and in hard to read type fonts, the OCR software has made lots of errors. Because only humans can correct these mistakes, this would usually be a hugely time consuming and expensive process.

Objectives

Develop staff capacity, Improve access, Improve efficiency, Increase citizen engagement

- The aim of Digitalkoot, this joint project between the National Library and Microtask, is to correct errors in its digital archives of old newspapers quickly and cheaply by using online volunteers to complete the tasks.
- The National Library also had the objective of encouraging Finnish people to get involved in the preservation of their culture.

Main beneficiaries

Academia, Civil Society, General population, Government staff, Students

- Citizens of Finland
- The National Library of Finland

Results

Effectiveness

- Digitalkoot was launched in February 2011. As of August 2012, 107,528 people have visited the Digitalkoot website; volunteers have contributed a total of 418,361 minutes of their time; and 7,634,174 microtasks have been completed by these volunteers.
- The accuracy of the tasks completed is estimated at over 99%! This figure 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, as 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 in 118 words of the second article. 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 getting people to play a computer game.
- Because of its unique approach to improving newspaper archives, Digitalkoot has attracted a lot of media publicity. Along with coverage in all the major Finnish newspapers, television and radio stations, it has also featured in major international publications including the New York Times and Wired.

Development

Design

The National Library had a huge problem on its hands: how to correct the mistakes optical character recognition (OCR) software had made digitizing millions of pages of historical archives. Instead of spending millions of Euros and many years trying to do this with library employees, the idea was to use a voluntary crowdsourced workforce. It did this by partnering with crowdsourcing company Microtask, and creating the Digitalkoot project.

Digitalkoot broke all the work up into small tasks and created an online platform where people could verify the words in a fun and engaging way from their own homes and workplaces. The project relied on the fact that people, especially Finns, would be interested in helping to preserve Finnish culture – as long as it was easy enough for them to do so.

Implementation

Tools used:

- A Microtask Platform identified and extracted the problematic words in old texts, and then presented them individually to volunteers online for checking. The platform would then automatically collect the answers, verify them, and insert them back into the digitized newspapers.
- But rather than asking the volunteers to spend time tediously reading through and typing out text that may not interest them, all they had to do was play social computer games. These games made helping correct the errors fun and competitive.
- In order to succeed in these games, players must accurately type in the difficult to read words cut from the original scanned version of the archive. To make sure that people were correctly entering the words, the same words were sent to different players simultaneously. This verified the results, ensuring a high level of accuracy.

Resources used:

- While the exact costs, including person hours contributed by the National Library of Finland and Microtask, are not available, it is clear that the project has been a huge success, both in terms of the cost and the time it has taken so far.
- The Library has 4 million digitized pages in its newspaper archives. Without the participation of the volunteers it would have taken 12 employees working full-time for 6 years to correct all the mistakes made by the OCR software.
- Not only would this have delayed the public’s access to the corrected archives, it would have been a very tedious and boring 6 years for those 12 people. Including employee benefits and all the associated costs, it would also have been prohibitively expensive!

Lessons Learned

Lessons Learned

- The challenges Microtask's engineers had to overcome were daunting. One problem, for example, was how to deal with malicious players who deliberately type in the wrong words (one tireless volunteer spent over an hour and a half doing just this).
- To identify such volunteers, the system begins the game feeding the player only "golden tasks", to which Microtask already knows the answer. Once the player demonstrates that they are actually trying to play properly, the ratio of these verification tasks gradually lowers. This process is completely invisible, so even if spammers understand the mechanism, they will not be able to cheat it.
- Other challenges included thorny mechanical issues like deciding the type of gameplay to implement (the first prototypes required input methods other than typing which proved to be very inefficient), issues around the number of parallel players required to crosscheck answers, and scalability of the system. There were also more mundane issues, for example some people were unwilling to use their Facebook account to sign in with the games (a very vocal minority requested login by email, so this was introduced a couple of weeks after the launch).

Other information

Microtask has identified a number of areas in which the next project can improve upon the current version of Digitalkoot. It wants to: take steps to minimize redundancy while checking the accuracy of tasks; be able to distinguish between words belonging to a title or to the text body; add soft keys for keyboards without the letters "å, ö, ä"; add new types of microtasks; and improve gameplay mechanics and reward mechanisms (i.e. figure out in which contexts the game interface undermines efficiency instead of increasing it).