Sourcing the cultural heritage crowd

Olaf Slijkhuis; Digitisation Consultant; Heiloo, the Netherlands

Abstract

The core business of an archive is to safeguard information for future generations but equally important to retrieve the information when needed. To retrieve information you have to know if you have the information and where it is stored in the archive. In other words you have to describe every object in the collection. Since archives grow in a tremendous rate it's inconceivable that archivists can describe everything in their possession. Especially with annual budget cuts hitting everybody in Western Europe volunteers are used more and more to assist in day to day operations and describing objects. But this was limited to the people that were able to come to the archive. There was no solid solution for people who were not able to come but who did have time and motivation to spend. Can this untapped power source be a potential problem solver?

How much info do you need?

Highly educated professionals are continuously cataloguing the items in their collection. Cataloguing is done with specially designed metadata sets like Dublin core, IPTC, Darwin core, etc. While this cataloguing is necessary to register what you have it doesn't provide a clear understanding of what it all means. You have to read it first to know if it's useful. To provide access to the content is the Holy Grail. Digitally born text and OCR on printed and scanned images can make text fully searchable but this only works for printed text and to be more specific printed text which is clearly readable. Everything else has to be transcribed or described. This is too large an amount of data for the professional to digest (depending on how long you can wait to have access to all the information). Analogous to a distributed computing network like for instance the SETI project whereby multiple computers from different locations form the processing power of a supercomputer you would like to have a network of human processing power equal to that of a superhuman to make sense of

Figure 1. *VeleHanden users on a global scale*

all the information. A network of seemingly random humans brought together for a single goal equals a crowd.

What is it about crowdsourcing?

Budget cuts in the cultural heritage field diminishes the use of in house resources to do all the work. There's simply too much to do with less and less people to do it. This is why museums and archives are more and more making use of volunteers. Voluntary in this case stands for 'Acting or done willingly and without constraint or expectation of reward' [1]. In June 2006 the term crowdsourcing came to life when Jeff Howe first coined the term in an article in Wired Magazine. He defines crowdsourcing [2] as:

"The act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call."

Although this isn't a new development in itself the concept has become more popular ever since. More and more companies, governmental bodies and institutions have looked and are looking at the possibilities of using the crowd for their benefit. With the rise of social network sites like MySpace, Facebook, Linkedin, Google+ reaching out to the crowd is easier than ever. Internet technology provides the spring board for crowdsourcing but still the question remains if this idea of using an anonymous crowd is the answer for cultural heritage institutions.

Is it as easy as it seems?

It all depends on how this raw power is channeled. One of the challenges is to make sure that this resource is not a temporary one. You have to motivate and entice the crowd to keep contributing. It would be very interesting if the crowd is able to work on different projects for different institutions without having to look somewhere else. It needs the effort and commitment of the institution to keep communicating and interact with the crowd. Cultural heritage institutions have to be actively involved in the project. Pay attention to the questions and problems of the volunteers and keep them updated about the progress and changes.

The other problem is the concern of information professionals about the reliability of the data that is produced through crowdsourcing initiatives. Is it not leading to more work because all the errors have to be corrected? How reliable the data entry of VeleHanden is, is convincingly demonstrated by Ellen Fleurbaay and Alexandra Eveleigh, in their paper "Crowdsourcing: Prone to Error?" [3].

There's also the fear that digitization will create an amount of data that is growing exponentially. Even with the help of the crowd it's going to take decades to get everything sorted. Maybe it's a Pandora's Box that has been opened and finding the right tool is the lid we need right now.

What should it look like?

Recognizing the value of crowdsourcing the Amsterdam City Archives in the Netherlands published a tender in late 2010 for the creation of a crowd sourcing solution. The tender explicitly called for a solution that would continue to be used after the first initial project had finished. It had to be a long lasting solution which would grow to fulfill other future crowd sourcing demands not only for the Amsterdam City Archives but for other projects in the country and in time projects from all around the world. With a public-private partnership between the archive and Picturae the 'VeleHanden' (translated as Many Hands) platform was born.

A long term and commercially viable solution which was user friendly, inviting to participate in and could serve different types of projects on one platform. It also had to be possible to customize the structure of each project in such a way that the goal was reached without changing the structure of platform. Participants have to have the opportunity to ask questions and discuss their results. A communication tool thus has to be an integral part of the crowdsourcing platform and the reaction time has to be short.

For some more challenging projects it is important to offer a training facility. The software can be configured to guide the volunteers through several stages of explanation depending on the complexity of the project. Data entry is explained online per field and more complex projects will start with a hands-on training to become an expert. This training can be in the form of a written manual or an online demonstration depicting all the necessary

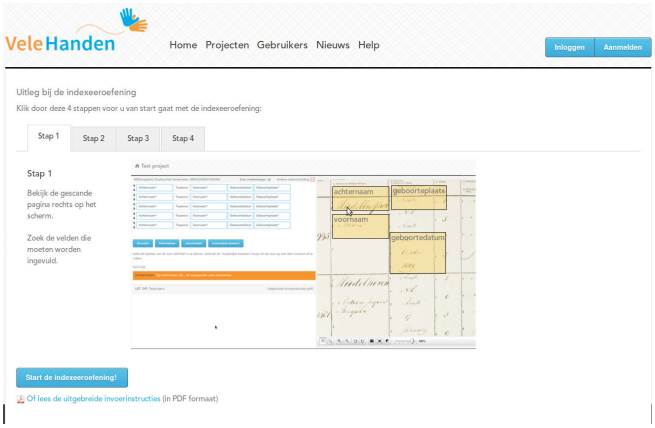

Figure 2. *Online training example*

steps. The communication tool is also part of the training facility, administrators are easily reachable for assistance Discussing problems with staff of the institution or other volunteers in a forum will lead to an increase in knowledge and thus contribute to the quality of the delivered data.

To motivate the participating volunteers, part of the tool is a reward structure whereby credits can be earned according to the amount of records that are contributed. These credits can be used for products or services which will tighten the bond with an institute (receive downloads, attend lectures, etc.). Using the crowd sourcing platform is in this way not only useful for acquiring data but also has the additional bonus for community building.

To insure the trustworthiness of the results several control mechanisms have to be in place. The structure of the platform provides the possibility to implement different types of control mechanisms depending on the type of project. It can use single-track or multi-track quality control methods. The basic control mechanism is the double entry system with a moderator as final approval but other more elaborate mechanisms can be applied if the need is expressed.

Picturae owns the platform and website (www.velehanden.nl) and charges a fee for the installation of new projects. By paying attention to the structure of the project, the way it is presented and the commitment of the institution to participate actively in the efforts of the crowd the results are more than encourage able sometimes even amazing. The first project which initiated the launch of the platform was the transcription of militia registers for the City Archives of Amsterdam, the Netherlands. What would have taken one person 16 years to complete was now finished in months with the help of over a thousand volunteers.

Practice makes perfect

Several projects have, since the start in 2010, been active on the 'VeleHanden' crowdsourcing platform. Transcribing registers for genealogy research, transcribing the information on

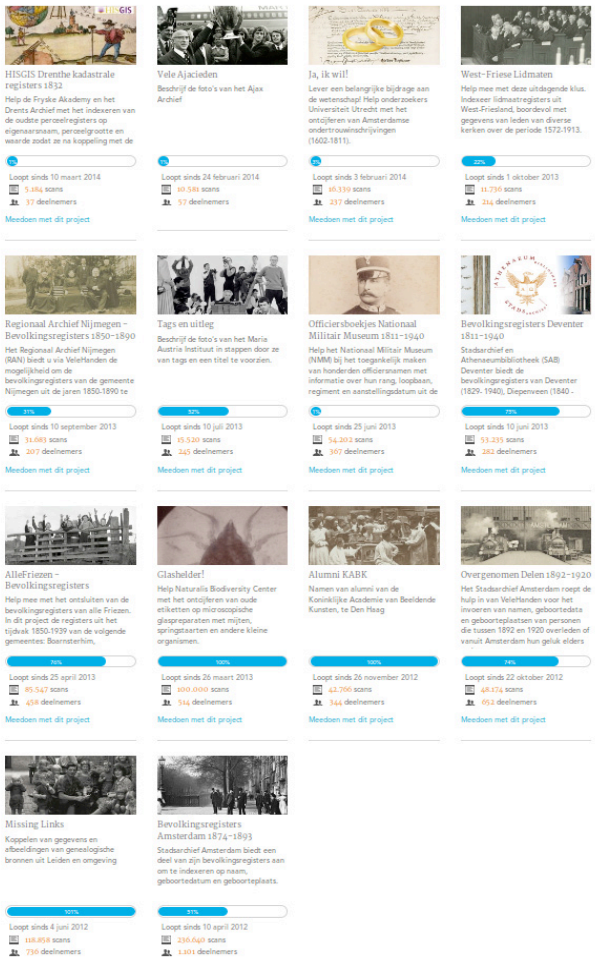

Figure 3. *Project overview on the VeleHanden platform*

microscopic plates for the Dutch biodiversity center Naturalis, tagging and describing a photographic collection for the Maria Austria Institute. More projects are added as we speak and new developments are introduced. The latest development is the possibility to select images from digitized contact sheets to facilitate a scanning on demand process. The crowd is choosing which images should be scanned in high resolution.

There are projects in which the data is of a more sensitive nature and is not suited for the crowd's eye. In these cases it is possible to assemble a private crowd with trusted experts.

The idea of a platform works to keep the volunteers participating to other projects. So far there are over three thousand people active on the platform and they are involved in more than ten projects. Quite a number of participants are contributing to more than one project. By presenting several projects in one environment the crowd can shop around and work on different subjects. On top of that it really creates a community spirit in which people will help each other in forums and have discussions with the staff of the institution. It turns out that the reward system of gaining credits is not an important motivating factor. People are not doing it for this kind of reward and they are often giving away their gained points to others on the platform.

Conclusion

In less than three years the 'VeleHanden' crowdsourcing platform has proven to be a valuable resource for cultural heritage institutions to provide online searchable content for parts of their collection that are otherwise only accessible to a few physical visitors. Now more people are able to use and find the hidden secrets of collections in the Netherlands and in the future from all over the world. The active involvement of contributing to the need of an institute creates an engagement which is more valuable than trying to reach the crowd using other social media sources.

References

[1] http://www.thefreedictionary.com/voluntary

[2] http://crowdsourcing.typepad.com/

[3] Ellen Fleurbaay and Alexandra Eveleigh, Crowdsourcing: Prone to error? http://ica2012.ica.org/files/pdf/Full%20papers%20upload/ica12Final00271.pdf

Author Biography

Olaf Slijkhuis's background is in Communication Science, Art History and Photography (practitioner). Since 2008 he works for Picturae (The Netherlands). First as data manager responsible for High-End scanning projects like Metamorfoze subsequently as production manager transparencies, as project manager for the 'Images for the Future', as a digitization consultant and as business developer for international projects.