International Council on Archives Conference 2012
Brisbane, Australia

Crowdsourcing: Prone to Error?

Ellen Fleurbaay, Amsterdam City Archives
Alexandra Eveleigh, University College London

Abstract

This paper considers the issue of quality control in ‘crowdsourcing’ or online user participation projects in archives. Drawing on the practical experience of the Dutch *Many Hands* (*VeleHanden*) crowdsourcing service collaborative, we begin by discussing the numerous options for error prevention, and then for error correction, for data submitted online by volunteer transcribers or indexers. This leads us to debate what actually counts as an error, and on issues surrounding the accuracy and reliability of the underlying historical record. We conclude with some reflections about how crowdsourcing might be leading to a re-focusing of some of the functions and roles of the professional archivist.

Introduction - the *VeleHanden* Project[1]

VeleHanden was established last year for ‘crowdsourcing’ information from archive documents. The way the site works is as follows: any archive service in The Netherlands is able to make scanned documents available on *VeleHanden* and ask for volunteers (the ‘crowd’) to help with indexing these documents, or transcribing them, or tagging photographs, or matching up data to scans - or indeed any kind of task that the archive service thinks people might be interested in doing online.

The website was an initiative of Amsterdam City Archives, and was designed as an innovative public-private partnership between the archives and Pictura, a commercial digitisation company in The Netherlands. Pictura brings to the project expertise in software development and the hosting of large-scale image banks, as well as services for mass digitisation. Pictura actually owns the *VeleHanden* website, and when an archive service wants to use *VeleHanden,* it has to pay a service fee that is related to the size, complexity and duration of the envisaged project. If new functionality is required for a particular project, it is up to Pictura to develop this functionality in dialogue with the archival institution. However, the archive service retains control over both the digital images and any metadata created by the volunteers during the project. We believe that the partnership therefore combines a commercial imperative for Pictura to support,

[1] http://velehanden.nl/

develop and sustain *VeleHanden* with the archival institutions' mission to promote online access and public engagement with archives.

Right from the start, *VeleHanden* proved extremely popular. Currently there are 3 projects running and another 2 will follow before the site celebrates its first birthday this coming November. The 'crowd' have racked up some pretty impressive achievements in this short time. There are 1389 members and jointly, over the last 9 months, they have contributed roughly 16 person-years to the project[2].

***VeleHanden* Production Statistics**

Project	**Extent**	**Indexing Activity**	**Duration**
Militia Registers	283,732 scans	1,532,643 records	253 days
Population Registers	236,640 scans	367,443 records	91 days
Missing Links	84.455 scans	22.485 records	36 days

The personal profiles which participants can set up on the *VeleHanden* site give us some idea of the people who take part in *VeleHanden.* Of course, not everyone wishes to give out their personal details. But we can say that demographically this 'crowd' looks a lot like the visitors to the reading room at the City Archives: most of them are men and they are of a 'riper' age. It is quite interesting to see that quite a few of the participants, about 4%, live outside the Netherlands.

***VeleHanden* Participants**

	Men	Women	No sex	
<20	2	2	0	
20-50	143	141	18	
50-70	355	166	67	
>70	78	17	17	
No age	113	114	156	
Total	691	440	258	**1389**

[2] Counting 40 working hours per week, at a rate of 1 minute for each record transcribed or checked.

***VeleHanden* Geographical Reach**

Netherlands	322
Rest of Europe	22
America	13
Australia	9
Africa & Asia	2

Although the project appeared an instant hit with the volunteers, it didn't succeed in attracting as many archivist colleagues as had been hoped. Only 15 archival institutions joined Amsterdam City Archives in the first pilot project, the project for indexing the Miltia Registers - even though for this pilot they didn't have to pay the fee for using *VeleHanden*, just the costs of digitising their own militia records. It was difficult to find a satisfactory explanation for this, until, when a presentation about *VeleHanden* was being given, somebody tweeted: 'Prone to error?' Maybe this question reveals the kind of doubts that Dutch archivists have about *VeleHanden*, and indeed, a lot of other people have similar concerns on crowdsourcing in general. That is what gave the inspiration for the title of this paper.

There are four main sections to this examination of reliability and control issues in archival crowdsourcing:

1. Starting with the principle that prevention is better than cure, we take a look at how *VeleHanden* and similar volunteer participation initiatives try to prevent the mistakes creeping in in the first place through careful site design and support facilities for participants.
2. Secondly, acknowledging that some human error is inevitable in any project, we consider various mechanisms for correcting and quality-assuring participants' contributions, which then leads us to the question, what counts as a mistake anyway?
3. So is there - thirdly - a sense in which archivists perhaps need simply to learn to live with error, and to consider instead how to go about highlighting errors (both historical and contemporary) and make appropriate judgements based on knowledge of these errors? And to help our users to do likewise.
4. Finally, we conclude with some reflections on what impact this all might have upon the role and identity of the professional archivist.

Error Prevention

So firstly, let us look at what precautions were taken with *VeleHanden* to prevent as many errors as possible.

Obviously participants need to be guided into inputting accurately from the start. Instructions are provided in several different places and formats. For instance, when a participant clicks on 'invoeren' (meaning:

‘enter data’) and the first scan appears on the screen, underneath the data entry box a single-word instruction appears explaining what information from the document scan should be typed into that box, with a corresponding one line instruction at the bottom of the data entry form detailing the format that should be applied - for instance, dates are entered in figures as dd-mm-yyyy. When moving the cursor to the next data field, another instruction appears, and if the participant tries to move to the next entry without formatting the input in the prescribed fashion, a pop-up warning appears at the top of the screen.

When the site was first set up, the system prevented the submission of incomplete or incorrectly formatted dates. At the time, this seemed a useful intervention, but it turned out to be a nuisance and the testers asked for it to be removed. There appeared to be many more technically ‘incorrect’ dates in the source material than had been expected, and this proved to be a problem with the automated checking procedure.

There is also a pdf manual available for printing, and a help section on the site. And for people who are a bit nervous, there is a sample on the homepage of the site that provides an opportunity to try out the interface, even before registering as a member.

However, questions still came up, regardless of how much instruction was given, and people wanted to discuss what they had found. So a ‘remarkable-button’ (= opmerkelijk) was added at the bottom of the data entry form. This enables members to send an email to the project coordinators, and is primarily meant to enable participants to draw attention to unusual things encountered in the documents. But project staff do also offer individual assistance via this route when necessary.

And there is a forum facility to which members can post questions. The more experienced members have proved to be an enormous help on the forum. They show endless patience in answering the questions of newcomers. They even set up a help list of urls on where to find useful information, for instance about the names of hamlets or townships that have disappeared from use because they have been incorporated into larger cities.

Supporting volunteers in these kinds of ways is important, not only because of the significant benefits in terms of quality of data submitted, but also in keeping people motivated to participate. Interviews with project participants repeatedly evidence the dedication and determination to ‘get it right’ of the most committed volunteers. Motivated volunteers pick up on anomalies, decipher abbreviations, check external sources for corroboration and to add additional context, and draw attention to those interesting bits and bobs found with archival documents that a commercial indexer would probably just ignore because they don’t fit a standardised format for data entry. Volunteer participants are generally interested in the subject and context of a particular project, unlike, for instance, a commercial transcription service whose focus is solely on the documentary text and a formal level of accuracy or margin of error.

Error Correction

But of course, no matter how much help is organised, mistakes DO occur. No human work can ever be faultless, and anyway, not everyone wants to get involved in the forum discussion on how to deal with every last spelling variant or has the time to keep checking back for the latest advice and guidance. That is why most crowdsourcing sites have built in facilities for also checking the data after it has been entered.

VeleHanden uses a double-entry system - i.e. two different people independently index the same scan. And then on top of that, a third person checks the data that the first two have entered. This third volunteer can see both the scanned document and the two sets of data that have already been entered. The website highlights where the entered data are not the same. The third person's job is to check and decide on the right choice. Or, in an exceptional case when the checker is not sure either, there is another special 'problem' button which sends an email to the project leader.

Ben Brumfield, a software developer and family historian from Texas who writes a blog dedicated to collaborative manuscript description, has identified no fewer than 9 methods of quality control over data entry and review[3]. Brumfield's 9 methods are divided in 2 categories: single-track and multi-track methods. The multi-track methods are mostly used for easily atomised and structured data, as we have seen on *VeleHanden,* whereas with single-track methods, all the corrections are made to a single transcription of the data. Single-track methods are mainly used for longer, unstandardized format texts, which call for a quality review process similar to traditional manuscript publication or expert peer review mechanisms.

UCL's *Transcribe Bentham* project[4] operates exactly such an expert review process designed to result in a single, authoritative transcription of each manuscript page written by the philosopher Jeremy Bentham. Volunteers can work collaboratively on the initial draft - although most frequently a single participant takes responsibility for a whole page - and when the volunteer deems the transcription to be 'finished', it can be submitted to a staff editorial panel for review. If the expert staff reviewer considers the volunteer's work meets the required transcription standards, the document is locked and cannot be edited further; if not, the transcript remains unlocked so that it can be revised further. *Wikisource[5],* Wikipedia's sister project for historical documents, similarly enables pages to be 'protected' from further editing, although there are also

[3] http://manuscripttranscription.blogspot.co.uk/2012/03/quality-control-for-crowdsourced.html

[4] http://www.ucl.ac.uk/transcribe-bentham/

[5] http://wikisource.org/wiki/Main_Page

archival examples of a more open-ended review process, for example with sources transcribed on The (U.K.) National Archives *Your Archives*[6] wiki.

The *VeleHanden* quality control process (double data entry followed by community review) is by contrast a multi-track method, as the source material used meets the criteria for easily isolated and structured data. Again, these multi-track methods can also be taken to further extremes than have been implemented to date on *VeleHanden.*

The *Old Weather* project[7], for example, originally required five independent transcriptions of each page of weather data derived from British Royal Navy ships' logs, but quickly reduced this to three independent transcribers upon realising that individual transcriptions had an accuracy rate of 97%, and that the remaining disagreements between transcribers were only occurring because of mistakes or illegibility in the original documents[8]. Another U.K. based project, the *Crew Lists Indexing Project* (or *CLIP*)[9]*,* has devised an even more elaborate combination of automated and manual checks and volunteer monitor balances to meet stringent data quality targets of a 2% error rate in the record of each individual seafarer recorded by the project, and a mere 0.13% error rate for each of the data fields which make up this record. This is an entirely volunteer-led initiative which accomplishes disproportionately large-scale indexing with fewer than 40 participants currently on the books. The CLIP quality control system not only helps to maintain the quality of the CLIP indexing data, but also functions as a progression hierarchy for keeping the volunteers motivated and honing their transcription skills.

Accuracy and Reliability

As these examples demonstrate, issues over quality control are a recurring challenge for every crowdsourcing site. Now archivists of course - and perhaps particularly archivists working for government institutions - attach great importance to reliability, and being sensitive to its value, our instinct with crowdsourcing is generally to tighten the controls, to demonstrate that we have made every effort to try to avoid and correct errors in the data. Reliability here then is also closely linked to issues of institutional and archival reputation and authority, whereby the archivist acts as a trusted intermediary between the archives and its users.

But on the other hand, we cannot deny that mistakes are simply a fact of life. No matter how much we try to prevent them of try to correct them, they do occur now and then, and they will in the future as they did

[6] http://yourarchives.nationalarchives.gov.uk/
[7] http://www.oldweather.org/
[8] http://blog.oldweather.org/2011/03/31/better-than-the-defence/
[9] http://www.crewlist.org.uk/

in the past. This fact brings up new questions: just how bad are these mistakes anyway? How do we know a mistake to be a mistake? What IS a mistake? And how can we learn to live with these mistakes?

Errors: Contemporary or Historical?

These might seem simple questions, but that is not so. On the *VeleHanden* forum, project staff often have to dampen the enthusiasm of our volunteers to fill in omissions or correct mistakes that we think should not be classified as mistakes. Here is an example from the Militia Registers Project of a mistake and an omission in the source material where, with the help of additional knowledge from other sources, the gap could be filled in:

Abraham Elias Visser, born in 1809 or in 1811, is one of many boys who are registered in the Militia Registers at the beginning of the 19th century with only the *year* in which they were born. Abraham is registered twice in different years, and in both registrations no month or day of birth are mentioned. He was born before Napoleon introduced an official birth registration. It's also possible of course that he simply didn't know his birthday! Some research brought to light that he was a Jewish boy, known in to his family as: *Abraham Elias Fieshl.* His name (this was common practice in those days) was just 'translated' for the Dutch officials. Abraham had a miserable life. He became an orphan when he was only 9 years old and he turned into a petty thief. He was sent to prison 4 times before he was 18… Perhaps he hoped to find a job and a home in the Militia and invented a year of birth that suited this ambition?

Abraham's interesting and colourful story (incidentally, he was not allowed to serve in the army as he proved to be too small, which means he was under 1.55 metres in height - and he again went in jail for theft) leaves us with a problem: should we 'correct' the mistaken name? Should we correct Visser into Fieshl so that other researchers can find him more easily? And should we fill in the omission of the day and month in which he was born? We don't think so. Instead, we have started to look for ways in which additional information like the story of Abraham can be published alongside or linked to the Militia Registers index.

There are many more examples of 'mistakes-that-are-no-mistakes', some can easily be traced, others not. Those with dates can be traced rather easily. Here is another example from the Militia Registers.

Hendrikus Ebeling cannot have been born on June 12th in the year 1. As he is registered in 1861, and all boys had to sign on at the age of 19, it is obvious that Hendrikus must have been born in 1842, as were all the other boys registered on the same page. But the strict instruction for the volunteers is to copy exactly what they read in the scan. So in the index you can find him born in the year 1 and this is not a mistake made during indexing.

According to the Militia Registers index, there were 3 boys born in the year 1, and although these entries are without any doubt mistakes, in all 3 cases it was not the crowd that made the mistake, it was the 19th century official.

The search system can help detect mistakes like these in dates. In the 1,162,135 records in the Militia-index there are 316 of these 'impossible' years of birth, but this is only 0.027% of the index, a reassuring percentage.

Learning to Live with Errors

The question remains though: what to DO with these mistakes?

The *VeleHanden* staff feel that to correct any of these 'mistakes' or 'semi-mistakes' would be to jeopardise the authenticity of the indexed data, and so, knowing that errors and slip-ups have been made in the past and will be made in the future – in other words realising that mistakes are just a fact of life - it is wise not only to think of ways of *avoiding* or *correcting* mistakes, but also we have to think of ways of how to *deal* with mistakes. In short, we have to learn to live with them.

For the Militia Registers this is not so difficult if we keep in mind the reason we wanted to create an index in the first place. The main purpose of this index is to help family historians find a document in which they can find additional information about a person, already more or less known to them by name, date and place of birth. These researchers are not generally interested in the index as a whole. What they are looking for is a fast, user friendly and ingenious search system that will help them find an individual mentioned in the Militia Registers, at the same time overcoming any mistakes or limitations in the source data or in the index itself. And that is exactly what the Militia Registers search system does.

Let us use Ellen's family name as an example: Fleurbaay. In the Netherlands, 'ij' and 'y' are quite commonly used interchangeably, so the official version of the name is actually Fleurbaaij. When the name is typed into the Militia Registers search system, it retrieves 28 boys named Fleurbaaij and 1 with the name Fleurbuuy. This 'smells' like a mistake. Just comparing the given names, one might expect this last boy to be a Fleurbaaij as well. We might have anticipated the ij / y –variants, but the Fleurbuuy is unusual. Even comparing with the original document, it is not clear whether the relevant letters are 'uu' or 'aa' - but thanks to the autocomplete function in the search system, this entry is findable anyway.

Archivists therefore have to balance an instinct for descriptive consistency with a mission to provide user access. Then perhaps the future of our collections knowledge systems lies not so much in control - in suppressing anomalies in the data - but in communication. In exposing the diversity and differences found in archives, and applying our professional expertise to helping the user navigate and filter an abundance of online historical sources.

What is important here is that *archivists* as well as participants reflect and learn from crowdsourcing projects, and that we work together to identify areas where current professional processes and services might be improved to meet user needs. In this guise, archival authority relies not so much on control, as on a balance of interaction, and of reciprocal trust and understanding between archivist and user.

Conclusion: The Archivist's Role in Crowdsourcing

So to conclude, on to the last question we would like to raise for discussion with you, the one that we suspect is the real (if perhaps unconscious) reason behind professional anxieties about the accuracy and reliability of crowsourcing sites like *VeleHanden*. What implications might the crowdsourcing of indexes and other finding aids to archival collections have for the functions and role of the professional archivist? Could there be a certain fear that the great workforce of the 'crowd' is going to make our work less interesting or even superfluous? We don't think so at all, but we do think it will bring about some changes of focus in our work. Let us have a closer look at three suggested aspects of this change:

The first and perhaps most profound change concerns our relationship with the users of archive collections. The archivist will no longer be the sole expert helping the individual researcher who visits the archive in person to find documents relevant for his research. The focus of our services will shift from the single individual to the community of users. We will be facilitating their research online just as we have to organise and facilitate their participation on the crowdsourcing website. We have to think of devices to motivate people to work for the community, and consider what benefits can be offered to participants to stimulate this collaborative behaviour. On *VeleHanden.nl* for instance, participants can earn points from their indexing work which can be exchanged for free downloads of digitised archive documents. The 'payment' amounts to no more than €0.50 an hour, which as a salary nobody could accept, but it plays to the private interests of the volunteers, many of whom are genealogists, and seems to be appreciated.

This first change of focus includes a change in the 'tone' of the relationship between the archivist and the user. It is no longer the user thanking the archivist for helping him; archivists have to learn to say thanks to the users for helping them open up the archives for the benefit of us all. In short, the archivist's role changes from gatekeeper to facilitator; to enabling a web of contrasting (or even competing) interests, opinions and contexts to flourish around the archival record.

The second change of focus then concerns the classical core business of our profession: archival description, a function central to the identity of the professional archivist. In the past the focus of the archivist was primarily on the control of masses of documents, on creating a singular overview of an archive by uncovering its structure, and describing the elements that together form this structure as simply and efficiently as possible. In Amsterdam, for most of the collections, we have to be satisfied if we can accomplish this at all, as we always have more archives to deal with, than archivists to do the description. There is seldom time to create indexes. But with the help of the 'crowd' it may yet be possible to index or even to transcribe individual documents, and to enrich basic descriptions with new contextual knowledge. Archival description therefore becomes more than merely a means-to-an-end, a process that leads to the production of reading room finding aids, and more an end in itself, a way of making sense of archives[10] in the complex, hyperlinked environment of the Internet. Once again, the role of the archivist here changes - from having a singular responsibility for description towards being one actor within a network of different contributors and perspectives.

All this additional metadata will make the content searchable, and that will be a great advantage in itself. However, when transcribing 40 kilometres of documents in this way, the advantage may diminish again. Merely creating a huge quantity of metadata is not enough: archivists need to work together with information technologists to consider ways to link all of this information together and to devise new research tools to support the creation of new knowledge out of this abundance. To give a simple example, when you want to look up the history of the house you live in, you expect to be able to find it at a present-day address. But in the past this address may have been completely different. It requires structure and knowledge embedded in a search system to link a house in Amsterdam that is now situated in the *Vrijheidslaan* (Avenue of Freedom), to the same house on the *Stalinlaan* between 1945 (the end of World War II) and 1956 (the Hungarian Revolution) and on the *Amstellaan* (Avenue of the river Amstel) before 1945. The third change therefore shifts the archivist's focus away from data input (creating catalogues and authoring definitive guides to records) towards the creation of more flexible and open-ended tools for structuring, filtering and understanding the archival world in all its richness.

[10] See further J.J.Bunn (2011) *Multiple Narratives, Multiple Views: Observing Archival Description* PhD Thesis, University College London, http://discovery.ucl.ac.uk/1322455/1/1322455.pdf