Crowdsourcing: An integrated part of archival description

Anders Sode-Pedersen; National Archives; Copenhagen, Denmark

Abstract

This paper presents ideas of how to make crowdsourcing an integrated part of the workflow of an archival institution, as the founding principles of a Danish project in its making. Crowdsourcing is not about cheap labour and strict control of the crowd by the archival institution, but about entering in an open ended partnership with users of the archival records. Users of archival records are at least as competent as professional archivists in describing the content of archival records, and they should be allowed to do that freely in the Danish archival information system (Daisy). They should be allowed to choose the records they would like to describe on their own, and the users should be responsible for ensuring the quality of data themselves. The concrete project is about "indexing" and full transcription of digitized archival records, available on www.sa.dk/content/dk/ao-forside. The internet application for "indexing" the archival records, is about to be developed, and a Beta version is expected to be launched in the summer of 2014.

Crowdsourcing in general

Crowdsourcing is about obtaining needed services or content by soliciting contributions from a large group of people, for instance volunteers not working as professionals inside the archival sector. It can be, and normally is, an online community, but it doesn't have to involve IT at all.

Crowdsourcing has been going on for years, even before the actual term was coined, in Denmark for instance we have been doing it for about twenty years inside a project called "Kildeindtastningsprojektet". This project has involved many volunteers from all over the world, who have done transcriptions of mainly Danish public censuses offline, about 15.000.000 posts until now, and the transcriptions have later on been published on the internet in "Dansk Demografisk database" (Danish Demographic Database) www.ddd.dda.dk/.

But many online crowdsourcing projects have seen the day of light, most well known is of cause Wikipedia, and especially inside the cultural sector crowdsourcing have become popular all over the world, "Trove" in Australia (transcriptions of Australian newspapers) "VeleHanden" in the Netherlands (transcriptions of archival records from different local and regional archival institutions), "Transcribe Bentham" (transcriptions of the philosopher Jeremy Bentham's manuscripts) are just a few examples among many.

The reason for the large adaptation of crowdsourcing inside the cultural sector, are probably due to the fact that the cultural sector are on the one hand the keepers of vast amounts of analogue information (archival records, pictures et cetera), and on the other hand typically are poorly funded. Another important reason is probably that many institutions inside the cultural sector have among their users, people who are very enthusiastic users, who already spend a lot time studying the collections of the same institutions, not in the reading rooms, but on the internet if the information is made available online. This at least goes for archival institutions, with archival records of relevance for genealogists. The interest for genealogy is growing in these years, and many genealogists spend hours and hours everyday studying and transcribing or indexing archival records on their own computers.

Surveys of why people participate in crowdsourcing projects, at least in Denmark, shows that it is primarily elderly people, that they do it for altruistic purposes (they want to help others, by making the information available free of charge and as fast as possible). They don't want any payment for the job, but they want to get recognition for their work, and they expect the application to work smoothly, receive quick answers to questions they may have, and they expect errors corrected quickly whatever errors that might be. [1]

One of the most common reservations about crowdsourcing is about the quality of the work of the crowd. It is typically professional archivists who share this concern. Although errors are made in crowdsourcing projects dealing with transcribing or indexing archival records, I would say that these concerns are mostly irrelevant for several reasons.

First of all it rests on the assumption, that professional archivists are not prone to error, which obviously is not the case. Many genealogists are much more skilled in reading old gothic handwriting than young archivists in the 21th. Century, and their knowledge of the records in hand is often much larger than the knowledge of the archivists. In the "Old Weather" project, dealing with transcriptions of logs from British Royal Navy ships, the accuracy rate was around 97%. Errors will be made, whoever is responsible for the transcriptions, but not that many errors it seems. [2]

Secondly there are many ways to minimize errors. You can validate the entries in your database to a certain degree, and you can check the quality of the data after it has been entered. Ben Brumfeld, an American software developer and family historian has identified nine different methods of quality of data entry and review. He divides the methods in to two categories, single-track and multi-track methods. Single-track methods are mainly used for longer unstandardized format texts, where quality review is done by experts, and where errors are corrected in one version of the transcription. The multi-track methods are mostly used for structured data, and the records are transcribed in more than one version, and then later on the different versions of the data is compared to identify errors for correction. [3]

Thirdly maybe errors should not be that big issue at all. In the good old days where transcriptions was published in books, and access to the original records was difficult, accuracy was obviously of the upmost importance, because it was both difficult to correct errors or even identifying them. But that is not the case now. On the internet the transcription or the index information is made available alongside the original record, so every user of the transcription can check the accuracy on the spot for themselves, and if errors are being identified, readers can report them, and the errors can be corrected at low costs, because we don't have to republish a book, we only have to make the correction in our database. Quality control of data can be expensive and time consuming if you want to correct them all before making the information available, but errors will certainly in the course of time be identified of the users of the data. On the internet timing is often more important than quality, and with the limited funding of many archival institutions, you have to choose between making small amounts of information available slowly or larger amounts of information available quickly. As far as I can see the choice is easy, because errors will appear anyway, and errors will be corrected whatever you choose one way or another.

Crowdsourcing - The Danish project

The Danish crowdsourcing project deals with transcriptions of all digitized archival record made available on the internet by The Danish State Archives on "Arkivalieronline" (www.sa.dk/content/dk/ao-forside). Right now it amounts to about 20.000.000 pictures of archival records, and the number will grow to about 50.000.000 pictures of archival records in 2017. The archival records being among other things: Parish registries, public censuses, probation records, applications for honorary medals for Danish soldiers, all kinds of archival records from the former Danish Vest Indian Islands (The Virgin Islands) et cetera.

The records in hand are therefore very inhomogeneous in format and content, the number of records are vast and we must expect that many people with different interests and backgrounds will want to participate, some of them only for a short while and others for a longer period of time.

So the system we have to build must be: Flexible, generic, effective, cheap for The Danish State Archives to operate and take into account what we know about our users about usability, quick results (publication of data) and recognition of the efforts of the crowd. Not necessarily an easy task!

The way to do it, at least in Denmark, is to use our existing archival information system (Daisy) for the job. First of all how can we possible show the crowd more recognition, than by letting the crowd use the same system as archivists and government agencies are using to describe traditional and original archival records, you might say that we consider the crowd as citizen archivists. Secondly the transcriptions are after all a new "copy" of the original records or a digitized version of the original record. And thirdly with Daisy we already have most of the IT-infrastructure needed at our hands.

To understand the last point, some background information on Daisy is needed. Archival records in Daisy is described accordingly to The Danish Standard of Archival Description. This standard shows many similarities to The Australian Record Series System, but furthermore it demands a very structured and formatted way of describing the content of archival boxes. [4]

The content is not described in unstructured text, but in separate structured fields, using the original "identifiers" on the archival records themselves. All archival records have some original identifier: A name, a date, a file number et cetera. It may not be a totally unique identifier, but it separates one record from another record. If for instance we look at file with a file number: 1917/654, the file number consists of a year 1917 and a number 654. Which in Daisy means that you can describe the format of the file number by creating two separate fields "Year" and "File number" for describing all record files belonging to this record series. We do this already with the original records, and we have an internet application that government agencies use to describe the archival records in this way before the records are transferred to The Danish State Archives. And we have the software, that dynamically creates the interface for description of each and every record series.

When you describe for instance the file number of a file, you are already transcribing the file, or at least a small part of it. In principal or technically it is the same, it is only a question of adding more fields to make a full transcription possible. This is possible, because the records we want the crowd to transcribe within each record series have the same format, and many times across record series shares the same format. For instance all Danish parish registries have been formatted in the same way since 1812.

So what we have to do is to make an internet application that enhances functionality we already have, and combine it with our picture viewer showing the digitized version of the records we want the crowd to transcribe. And by doing it in this way, we can make large numbers of digitized versions of record series available for transcription in a well structured and cost efficiently way. Obviously there are more to it than this, but basically this should do it.

We plan to make all digitized versions of record series available, and letting each member of the crowd decide what records they want to describe. We do not want to present the crowd limitations, of which records they are allowed to transcribe, or how many pictures they have to transcribe. If a user only wants to do the job with one picture, it is acceptable for us, and if the picture is number 154 out of 435 pictures belonging to one series, that is okay too. One picture done is one less to go. So in this respect the system will be very open ended. We will probably try to organize small "project groups" dealing with certain archival records, so that whole record series can be completed, but it will be up to the members of the crowd themselves, if they want to participate in one of these groups.

As soon a picture is transcribed, the transcription will be available for everyone on the internet. Apart from some validation of data in certain fields at the moment of registration, data quality will be handled by a user driven single-track method. We will appoint certain members of the crowd to editors of the data. And the editors will have the responsibility of reviewing the data and correcting errors, but this will always happen after the data is made available to the public. By doing it in this way data reaches the public very quickly, and we believe that the appointment as editor will be considered as an appreciation and recognition by the people appointed.

It is not without costs to enter into crowdsourcing. Development of software, creation of the IT-infrastructure, digitization of archival records among other things are costly, but by using existing functionality and by accepting the crowd as a capable and competent partner, we believe an open ended model as the Danish model for crowdsourcing can and will be successful because it will attract many users, and because a lot of the responsibility for the quality of data and administration of the system will be handled by our new partner – the crowd itself.

References

[1] Charlotte S.H. Jensen, ER ARKIVER OG DERES BRUGERE DANMARKSMESTRE I CROWDSOURCING AF KULTURARV?, http://charlotteshj.wordpress.com/2011/07/15/er-arkiver-og-deres-brugere-danmarksmestre-i-crowdsourcing-af-kulturarv/

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

[3] Ben Brumfeld , Quality Control for Crowdsourced Transcription, http://manuscripttranscription.blogspot.dk/2012/03/quality-control-for-crowdsourced.html

[4] Anders Sode-Pedersen, Beskrivelsesmetoden, www.sa.dk/content/dk/forskning_og_udvikling/udviklingsprojekter/daisyprojektet/beskrivelsesmetoden

Author Biography

Anders Sode-Pedersen has a M.A. in History from The University of Copenhagen. He has been working in the The Danish State Archives since 1997. He has been involved in the development of The Danish Archival Information System (Daisy) and the Danish State Archives platform for making digitized archival records (Arkivalieronline), both as a project manager and as head of section for the it-development. He is currently employed as business architect in The Danish State Archives.