Recognition and Enrichment of Archival Documents

D4.11
Transcribe Bentham

Louise Seaward, UCL

Distribution: Public

http://read.transkribus.eu/

READ
H2020 Project 674943

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 674943

Project ref no.	H2020 674943
Project acronym	**READ**
Project full title	**Recognition and Enrichment of Archival Documents**
Instrument	H2020-EINFRA-2015-1
Thematic Priority	EINFRA-9-2015 - e-Infrastructures for virtual research environments (VRE)
Start date / duration	01 January 2016 / 42 Months

Distribution	Public
Contractual date of delivery	31.12.2017
Actual date of delivery	22.12.2017
Date of last update	22.12.2017
Deliverable number	D4.11
Deliverable title	Transcribe Bentham
Type	Demonstrator
Status & version	Public, Version 1
Contributing WP(s)	4
Responsible beneficiary	UCL
Other contributors	UIBK, ULCC, UPVLC, URO
Internal reviewers	Tobias Hodel
Author(s)	Louise Seaward
EC project officer	Martin Majek
Keywords	Crowdsourcing, Automated Text Recognition, Volunteering

Table of Contents

Executive Summary

Transcribe Bentham is a long-running crowdsourcing initiative based in the Bentham Project at UCL, which asks members of the public to transcribe papers written by the English philosopher Jeremy Bentham (1748-1832). This report summarises the latest progress in the initiative and plans to update the Transcribe Bentham website. It also explains the development of the Transkribus crowdsourcing interface, which will use Automated Text Recognition technology to make the task of transcription easier for volunteers.

1. Transcribe Bentham

1.1. Background

Transcribe Bentham was launched in 2010 and has become one of the most successful academic crowdsourcing initiatives. Transcribe Bentham asks members of the public to transcribe images of Bentham's manuscripts in text and TEI mark-up at an online Transcription Desk. Transcripts produced by volunteer transcribers are checked and used by researchers in the Bentham Project as part of their editorial work on Bentham's *Collected Works.* The transcripts are then made available in the open access digital repository of UCL's library. Transcribe Bentham helps to maximise the efficiency of the scholarly editing process, preserve records of Bentham's writings and spread awareness of his ideas amongst the general public. Through our talks, publications and blogs we have also become an example of best practice in the scholarly crowdsourcing community. For more details on Transcribe Bentham see *4.10 Transcribe Bentham*.

1.2. User activity

Since the platform was launched in 2010, we have had over 39,000 unique views of our website from 145 countries. Our volunteer transcribers have worked on a total of 19,249 manuscripts, at an average of 51 pages per week (as of November 2017). There are a total of 587 registered users who have worked on at least one transcript on the site. But in common with many crowdsourcing projects, Transcribe Bentham is largely reliant on the efforts of a relatively small group of volunteers. Our dedicated core of around 30 'super-transcribers' have completed over 90% of the finished transcripts on the platform.

1.3. User survey

The continued success of Transcribe Bentham is dependent upon these 'super-transcribers' who are willing to devote time and effort to the task of transcription. We are aware of the importance of being responsive to their needs and conduct regular online surveys of our most active users. Our most recent survey was conducted in August 2017 and was completed by 11 users. The survey was focused on the background, motivations and user experience of the volunteers.

The survey demonstrated that the super-transcribers tend to be older, well-educated people who have flexibility in their job or life (as a student, freelancer or retiree) that gives them the time to transcribe. The users stated that they liked the challenge of transcription, that they

have an interest in the subject matter of history, philosophy and law and perhaps most importantly, that they really enjoy the task of transcribing Bentham. They also suggested some improvements that could enhance their experience of using the site. These proposals comprised a mixture of technical requests like an improved display of untranscribed manuscript pages or the integration of computer-assisted transcription, as well as appeals for more communication and feedback from Transcribe Bentham team and other users. We presented the results of this survey in October 2017 at an academic conference on crowdsourcing at the University of Angers, Le Crowdsourcing: pour partager, enrichir et publier des sources patrimoniales. We are now making plans to use the feedback we received from users to rework elements of the Transcription Desk. We are focused on improving our communication with users as a first step as this is something that can be undertaken relatively simply at UCL. Technical improvements will take longer to realise because they require more resources and collaboration with other partners, both inside and outside UCL.

1.4. Bentham Hackathon

Proposals to improve Transcribe Bentham were also developed at the Bentham Hackathon, which took place at UCL on 20-22 October. This event was held in collaboration with IBM, which provided access to its technology and technical support for participants at the event. Around 40 attendees came together in six teams to consider how digital tools could facilitate research into Bentham's philosophy. The outputs included keyword searching of Bentham transcripts, a language model to predict Bentham's choice of words, a cleaner interface for navigating images and transcripts side-by-side and a sandbox area for the Transcription Desk where new users could practice transcribing manuscripts and receive immediate feedback on their efforts. These creations demonstrated how some of the desired innovations highlighted in the user survey could be put into effect on the Transcription Desk and the results of the event will help to inform our plans to enhance the usability of our platform.

1.5. Transcription Desk migration

An agreement has been reached to change the location where the Transcription Desk is hosted in 2018, from ULCC to UCL servers. This technical change will improve the stability and functionality of the Transcription Desk. It will make it easier for UCL to manage and upload new material to the site. It will also allow for an update of the platform's Mediawiki framework, with the additional possibility of some technical enhancements like stronger spam filters and automated transcription statistics.

1.6. Automated Text Recognition models

After various experiments in tranScriptorium and READ, we now have a powerful Automated Text Recognition model that is capable of processing papers from the Bentham collection. This model is based on Neural Network technology from URO and can produce transcripts with a Character Error Rate (CER) of between 5 and 10%. This model is publicly available to all Transkribus users under the title 'English Writing M1'. It has been applied to other eighteenth- and nineteenth-century documents written in English with some success, such as the papers of George III which are being transcribed by the Georgian Papers Programme.

This model was trained primarily on papers written by Bentham's secretaries and so it copes well with manuscripts that are in standard copperplate handwriting. We undertook further training this year in the hope of creating a model that will be capable of processing the most difficult papers in our collection, those written by Bentham himself in his old age. UCL prepared 197 pages of ground truth training data in Transkribus for this purpose. The resultant model was weaker than 'English Writing M1', with a CER of around 35%. We therefore plan to prepare and submit more pages of training data with a view to improving the accuracy of this model and making the automated transcription of Bentham's own handwriting possible.

We are also starting to create another model that would facilitate Bentham studies. This model will be based on the writings of Etienne Dumont, the Genevan liberal who edited Bentham's writings and helped to bring him to international attention in the early nineteenth century. 300 pages of Dumont's French writings are now being prepared for Text2Image matching where existing images and transcripts are automatically joined together as training data for Automated Text Recognition. Transcriptions from the resultant model should help us to analyse the way in which Dumont edited and presented Bentham's ideas. This experiment also represents a test case which will allow us to analyse the potential of Text2Image matching. If the technology works well on the Dumont papers, it could then be applied to the thousands of Bentham images and transcripts that we have collected in the course of the Transcribe Bentham initiative. This would enable us to generate a large amount of ground truth relatively easily and use this ground truth to improve the recognition of Bentham's own handwriting.

2. Using Transkribus in crowdsourcing

2.1. Transkribus crowdsourcing interface

The Web Interfaces Working Group is working on the technical implementation of a new crowdsourcing platform where volunteers will be able to transcribe manuscripts with the assistance of Automated Text Recognition technology.

This platform will be based on the tools for viewing and transcribing manuscripts that have been constructed for the Transkribus My Collections web interface. A prototype of this interface was demonstrated to users at the Transkribus User Conference in November 2017. We plan to ask a small group of conference attendees to test and give feedback on My Collections in January 2018.

My Collections allows users to view and work on all collections that they have access to in Transkribus. In the prototype version, users are able to view a digitised image of a segmented page and click on each line of the image to access a pop-up box where they can add, edit or view a transcription of each line. Future versions of My Collections will offer different viewing options to the users such as an image segmented into lines with a matching line-by-line transcription or an image presented with a block of text transcription on its left-hand side.

My Collections makes use of the improved line segmentation technology that has been developed in READ during 2017 (see D6.5 Basic Layout Analysis). This new technology is crucial for the crowdsourcing interface as it makes it more likely that institutions will be able to automatically and accurately segment their documents so that they are ready for line-by-line transcription by volunteers.

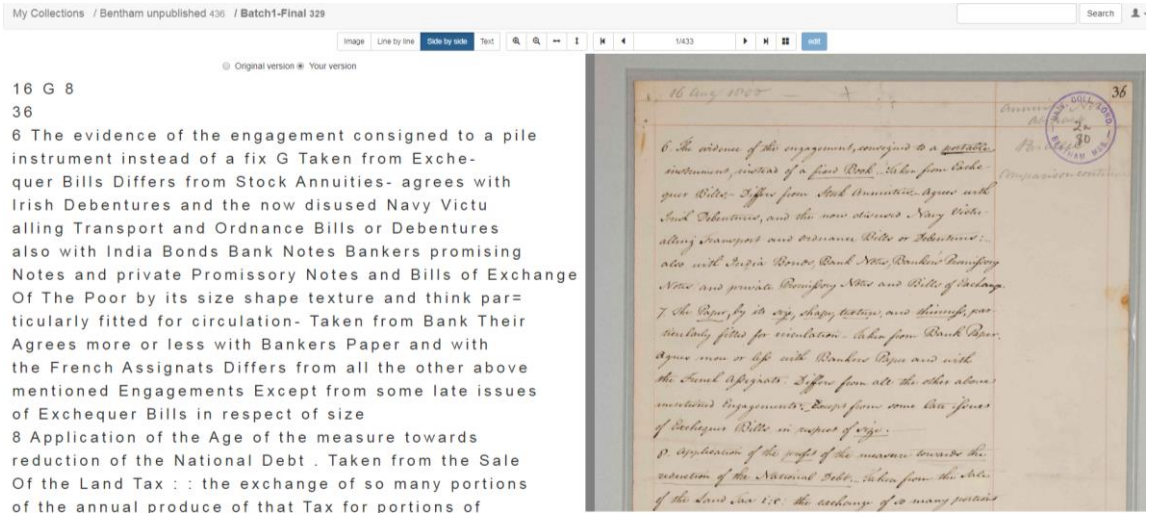

Figure 1 Document viewer in Transkribus My Collections

The Transkribus crowdsourcing platform will use this same interface for viewing and transcribing documents but will add enhanced functionalities for both volunteer users and the managers of crowdsourcing projects. UCL will test the platform using Bentham manuscripts and it will be freely available for anyone who would like to work with volunteers to transcribe a digital collection.

Volunteer users will benefit from the inclusion of Automated Text Recognition technology which will allow them to receive suggestions of unclear words when transcribing themselves or alternatively check and correct auto-generated transcripts. Feedback from the users of Transcribe Bentham and experiments with the TSX crowdsourcing platform under the tranScriptorium project indicate that these options would increase productivity and satisfaction amongst volunteers. Managers of crowdsourcing projects will also be able to get a clear overview of the status of the documents available for transcription and the activities of their volunteers.

2.2. Keyword Spotting

The recent integration of Keyword Spotting technology into Transkribus could also open up new opportunities for crowdsourcing projects like Transcribe Bentham. Keyword Spotting is a powerful searching tool where the technology analyses images of writing, rather than searching through transcriptions of these words generated either by humans or computers. This tool could therefore facilitate the searching of huge collections that have not yet been transcribed. This technology is particularly useful because it can work with Automated Text Recognition models that have higher error rates (of around 20-35% CER). This means that the new Bentham model, which is based on Bentham's most difficult handwriting, could be used for Keyword Spotting.

This technology will be useful for crowdsourcing projects in two ways. First, it could create a new task for volunteer users who would be asked to check the results generated by the Keyword Spotting engine and verify which words have been spotted correctly. This would be of interest to a broad range of users who would like to contribute to academic crowdsourcing but who do not necessarily have the time to transcribe difficult manuscripts page by page. An element of gamification could even be introduced, where a user tries to identify as many correct words as possible in a given time-frame. Second, Keyword Spotting could be used to pinpoint particular topics that might interest volunteer transcribers. We know that volunteers on Transcribe Bentham are motivated by their interest in the material they work on. It is likely that volunteer productivity would increase if we could guide users towards material that they might find intriguing.

We will begin experimenting with Keyword Spotting on a limited corpus of Bentham material in order to show its potential for organising and motivating volunteers from the crowd.

3. Conclusion

Transcribe Bentham continues to attract new and existing volunteers to the Transcription Desk. These volunteers contribute transcripts at an impressive rate, with high levels of accuracy. This crowdsourcing initiative makes a crucial contribution to scholarly editing, research and public engagement. In 2017, the user survey and Bentham Hackathon both proved useful in generating ideas that will help to update the Transcription Desk and improve the user experience. The back-end of the site will be updated during its forthcoming migration to UCL servers. Development work on the My Collections web interface is laying the groundwork for the production of a Transkribus crowdsourcing interface which will make transcription more convenient for volunteers. The development of Keyword Spotting technology should also open up the possibility of new workflows for such projects.

4. References

[1] T. Causer, K. Grint, A-M. Sichani and M. Terras, 2016, '"Making such bargain": Transcribe Bentham and the quality and cost-effectiveness of crowdsourced transcription', *Digital Scholarship in the Humanities*, forthcoming

[2] T. Causer and M. Terras, 2014, '"Many hands make light work. Many hands together make merry work": Transcribe Bentham and crowdsourcing manuscript collections', in *Crowdsourcing our Cultural Heritage,* ed. M. Ridge, Ashgate, pp. 57-88: http://discovery.ucl.ac.uk/1393567/

[3] T. Causer and V. Wallace, 2012, 'Building a volunteer community: results and findings from Transcribe Bentham', *Digital Humanities Quarterly,* 6, 2: http://www.digitalhumanities.org/dhq/vol/6/2/000125/000125.html