Recognition and Enrichment of Archival Documents

D4.12
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	30.06.2019
Actual date of delivery	26.07.2019
Date of last update	26.07.2019
Deliverable number	D4.12
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	Günter Mühlberger
Author(s)	Louise Seaward
EC project officer	Christophe Doin
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 details the successful migration and update of the project website. It describes the latest achievements relating to the automated transcription and searching of Bentham's writings and gives an overview of work relating to the use of Transkribus for crowdsourcing.

1. Transcribe Bentham

1.1. Background

Transcribe Bentham is one of the most well-established scholarly crowdsourcing initiatives in the field. It asks members of the public to transcribe images of Bentham's manuscripts in text and TEI mark-up at an online Transcription Desk. Volunteer transcripts constitute a unique resource for researchers, help to maintain records of Bentham's writings and promote awareness of his philosophy. For more details on Transcribe Bentham see *4.10 Transcribe Bentham*.

1.2. User activity

Transcribe Bentham remains dependent on a group of 'super-transcribers' who have contributed around 95% of the transcribed material on the site. Since the beginning of the initiative in September 2010, volunteer transcribers have worked on 22,017 pages at an average of 49 pages per week (as of the end of June 2019). More than 7,000 pages have been transcribed over the course of the READ project. There are now 679 users who have transcribed something at least once. Transcribe Bentham remains dependent on a group of 'super transcribers' who have contributed around 95% of the transcribed material on the site. The number of active 'super-transcribers' has expanded slightly to comprise 38 users.

1.3. Digitisation

In May 2018, thanks to a collaboration with UCL Digital Media Services, UCL Library Special Collections and The British Library, the complete digitisation of the Bentham papers was finalised. The digital collection comprises over 95,000 images; around 80,000 from UCL and 15,000 from The British Library. Much of this material is now available online through the Transcription Desk and the digital repository of UCL Library, with additional material being added periodically. These digitised images are hugely significant in creating new opportunities for research and facilitating public engagement with Bentham's philosophy.

1.4. Improving user experience

In a 2017 survey we questioned the Transcribe Bentham 'super-transcribers' about their background, motivations and user experience. For a more detailed account of this survey see *D4.11 Transcribe Bentham*.

Based on this feedback, we implemented two new initiatives designed to increase engagement amongst the existing transcribers and encourage more like-minded people to become involved. In February 2018, we started a monthly Transcribe Bentham newsletter, which celebrates the work of individual volunteers and discusses project events, publications and research. The newsletter now reaches over 100 subscribers, including all of the most active *Transcribe Bentham* volunteers. In July 2018, we launched Transcribe Bentham's first ever 'transcription challenge' where volunteers were asked to prioritise the transcription of relatively complex material that was close to completion. Volunteers responded enthusiastically, transcribing the majority of the targeted material save for a few particularly difficult pages. In December 2018, we invited volunteers to begin a new challenge which was this time focused on the transcription of documents that are likely to be of use in editorial work on the next volumes of Bentham's *Collected Works.* We have received approving feedback about these changes from established volunteers and seen the participation of both new and existing users increase.

1.5. Game Jam

On 23-24 February 2019 we delivered another Hackathon event, this time in collaboration with The National Archives UK. The event was a 'Game Jam' where teams worked to design and create video games over a weekend. The key objective was the gamification of the task of transcription. The participants were challenged to use documents from The National Archives and the Bentham collection to create an imaginative game that made the task of transcribing historical documents fun and efficient. The gamification objective is particularly relevant for Transcribe Bentham because the work undertaken by our volunteers is difficult and time-consuming. If transcribing Bentham became more fun, it is likely that more people would take part. The ideas presented by the eight teams in the Game Jam included a mobile app where users gain access to Bentham's recipes by transcribing words and a website which rewarded transcribers with the chance to 'ride' a bike icon along the loops of an image of a handwritten word. Similar ideas could be integrated into the Transcribe Bentham workflow in future to make transcription more pleasurable.

1.6. Website migration

The results of the 2017 survey indicated that volunteers were frustrated with various technical aspects of the site, such as the quality of the image viewer and the difficulty of finding untranscribed pages to work on. In 2017, it was decided that we would migrate the Transcription Desk website from ULCC to UCL servers in order to improve its stability and functionality.

The website was migrated successfully to UCL in October 2018 by UCL's Research IT Services (RITS) team. Visitors to the old site were automatically redirected to the new version. UCL RITS updated the Mediawiki framework of the site, deleted thousands of spam users and fixed bugs associated with the migration. The team also indexed the site so that it became more visible in online search engines.

Post-migration work continues on outstanding snags relating to the display of images, TEI encoding and notification emails.

1.7. Handwritten Text Recognition models

Our attempts to train models to recognise Bentham's handwriting have continued and there are several success stories to report.
As reported in *4.11 Transcribe Bentham*, we already have a robust model that has been trained to recognise the easier pages from the Bentham collection (primarily those written by Bentham's secretaries). This model is based on Neural Network technology from URO and has a Character Error Rate (CER) of 3.66% on the test set. This model is publicly available in Transkribus, allowing users to experiment by recognising text or creating a new Handwritten Text Recognition (HTR) model that builds directly on the Bentham data.

Unfortunately the complexities of the Bentham papers mean that this model is unsuitable for transcribing the complete collection. The model was primarily trained on easier handwriting from Bentham's secretaries. It therefore struggles to recognise more difficult script: either from Bentham himself or his correspondents. We can usually expect a CER of 5-20% when applying the model to a random page from the collection.

In 2018 we focused on improving the recognition of Bentham's own handwriting, particularly the difficult style of his later years. The advances in Layout Analysis coming from URO were crucial here, as we were able to detect baselines in our images with much greater accuracy and therefore create ground truth more quickly.

We created an expanding set of ground truth manually, retyping existing transcripts in Transkribus. Our first milestone was a model which was trained on 57,000 transcribed words and which used 'English Writing M1' as a base model. This training resulted in a model with a CER of 26.53% on the test set.

Our latest model has been trained on 140,000 transcribed words, using HTR+ technology from URO. This model represents a dramatic improvement on our previous work, reaching a CER of just 9.53% on the test set. Although this model cannot yet produce transcripts of sufficient accuracy for the scholarly editing work of the Bentham Project, further improvements should be possible as the technology progresses and more pages of ground truth are created and used for training.

1.8. Keyword Spotting

In parallel to our experiments with URO HTR technology in Transkribus, we also worked with UPVLC to develop new HTR models for the Keyword Spotting of the Bentham papers.

We gave UPVLC access to all of the digitised images of the Bentham Collection (around 95,000 files), our metadata records and 1,200 pages of ground truth in Transkribus. The UPVLC team standardised the images, distinguishing between handwritten, printed and blank pages, and those written in different hands and languages. The images were uploaded to Transkribus, automatically segmented in batch mode and then a sample was checked for accuracy. UPVLC processed the ground truth pages with their own neural network HTR and probabilistic word indexing toolkit. The resulting models can produce transcripts of pages from the Bentham collection with a CER of 6.1% on easier material and 15.5% on more difficult material.

Based on these results, UPVLC have created a Keyword Spotting web interface where users can search 90,000 images from the Bentham collection (a figure reached following the removal of blank and printed pages). Tests show that the average search precision ranges from 79% to 94%. This website is a practical resource which will help researchers to find previously unknown references in pages that have not yet been transcribed.

This experiment has encouraged us to consider Keyword Spotting as a mechanism for volunteer engagement. If Keyword Spotting were to be integrated into the Transcription Desk, it could be used as a finding aid to help volunteers discover material that interests them. Volunteers could also be enlisted to verify which words have been spotted correctly by the machine.

We have been working with UPVLC to demonstrate a possible research use of the Keyword Spotting interface: the provenance of supposed neologisms invented by Bentham. A list of potential Bentham neologisms compiled by the Bentham Project was the starting point. We conducted Keyword Spotting for each word to discover the likely date of its first use by Bentham. We also searched for the words in the Google N-Gram viewer to determine if they had been used by other writers prior to Bentham. The results indicated that some neologisms might have appeared earlier than once thought, or were possibly coined by writers other than Bentham. UPVLC will report on this research at the ICDAR 2019 conference.

1.9. Text2Image matching

We have also been able to undertake large scale ground truth creation and training thanks to URO's Text2Image matching tool.

The Bentham Project has access to a significant amount of XML file transcripts that could theoretically be used for training HTR models. There are around 21,000 transcripts prepared by Transcribe Bentham volunteers and around 10,000 created by Bentham Project researchers.

Our first attempt to create ground truth by automatically matching a few hundred Bentham images and transcripts together did not create sufficiently accurate results.

After this initial failure, URO oversaw the Text2Img matching of a large set of thousands of transcripts and images. URO pre-processed the transcripts; converting the XML into raw text lines and removing TEI mark up. In order to facilitate the automated matching, lines containing unclear words tagged with 'unclear' or 'gap' tags were excluded from the training set. The resulting HTR model had a CER of 11.62% on easier images from the Bentham collection. This result is comparable with those generated by the 'English Writing M1' model and indicates that significant benefits can be gained by integrating large numbers of existing transcripts into HTR training. For more information on this Text2Image matching process see *D7.21 Model for Semi- and Unsupervised HTR training*.

1.10. Transkribus Learn

We have also taken advantage of the Transkribus Learn platform in an effort to make transcription easier for new volunteers. We uploaded two collections of Bentham material to the site – categorised as 'easier Bentham' (writing by Bentham and his secretaries) and 'difficult Bentham'. Both collections are an ideal training ground for new volunteers, offering the opportunity to practice transcribing different words in rapid succession. The Transcription Desk site now invites new volunteers to practice on Transkribus Learn before they attempt to transcribe an entire page of Bentham's writings.

1.11. Future plans

We will continue to work towards improvements in the recognition of Bentham's handwriting. As well as retraining our current model with additional pages of data, we want to create smaller models focused on the specific hands and languages in the Bentham collection.

We would also like to enhance the usability of the Keyword Spotting interface by connecting it to other digital resources such as the UCL library repository, the Bentham papers database and the Transcription Desk.

Our ultimate goal is still to integrate HTR technology directly into Transcribe Bentham, in order to make transcription easier for new and existing users. Volunteers will be able to check and correct automated transcripts or transcribe manually and receive computer-generated suggestions of words that are difficult to decipher.

2. Using Transkribus in crowdsourcing

2.1. Transkribus crowdsourcing interface

The Web UI working group have continued to develop the Transkribus Web interface; fixing bugs, streamlining its look and feel and ensuring transcription and annotation are more functional.

Collection owners have on request the option to use the Transkribus expert client to designate a collection open for crowdsourcing. The collection then becomes open to volunteer transcribers via Transkribus Web.

Rather than developing a Transkribus-based crowdsourcing platform, the team have concentrated on improving the performance of Transkribus Web tools for easy integration into external crowdsourcing projects. In March 2018 Amsterdam City Archives launched *'Crowd leert computer lezen'*, a new crowdsourcing project where volunteers are asked to correct automated transcripts of Dutch notarial records. The project is hosted on VeleHanden, a successful crowdsourcing platform created by the company Picturae. *Crowd leert computer lezen* is directly connected to the Transkribus web interface, meaning that any changes made by volunteers can be fed straight back into the system to improve the HTR**.**

For more on Transkribus Web and the crowdsourcing platform see *D5.7 Mobile Crowdsourcing Tools*.

3. Conclusion

In the final phase of the project, HTR+ has brought us much closer to the reliable recognition of Bentham's handwriting. The creation of a Keyword Spotting interface also opens up a host of new research possibilities for those interested in Bentham's philosophy. Our work in READ has laid the foundations for a future version of Transcribe Bentham, where the power of HTR will be combined with the expertise of volunteer transcribers.