Recognition and Enrichment of Archival Documents

D4.10
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	31.12.2016
Actual date of delivery	28.12.2016
Date of last update	15.12.2016
Deliverable number	D4.10
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, Günter Mühlberger
Author(s)	Louise Seaward
EC project officer	Martin Majek
Keywords	Crowdsourcing, Handwritten Text Recognition, Volunteering

Table of Contents

Table of Figures

Executive Summary

Transcribe Bentham is an award-winning crowdsourced transcription initiative which recruits members of the public to transcribe manuscripts written by the British philosopher Jeremy Bentham (1748-1832). This report explains the administrative workflow of the Transcribe Bentham Transcription Desk and the way in which volunteers interact with the platform. It provides details of lessons learned from the development of the TSX client, a prototype crowdsourcing platform which was constructed under the tranScriptorium project (http://transcriptorium.eu/). Finally, it indicates how READ will build on the TSX experiment to construct a new crowdsourcing interface where volunteers will be able to transcribe manuscripts with the assistance of HTR technology. This will be an open source platform which can be used and adapted by other institutions who wish to make their collections available for crowdsourcing.

1. Transcribe Bentham

1.1. Background

Transcribe Bentham was launched in 2010 and has become one of the longest running and most successful academic crowdsourcing initiatives. It is part of the Bentham Project at UCL, which is dedicated to editing and publishing the new authoritative edition of the *Collected Works* of Jeremy Bentham. Transcribe Bentham asks members of the public to access and transcribe images of Bentham's manuscripts at an online Transcription Desk (http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham). Transcripts produced by volunteer transcribers are used by researchers in the Bentham Project as part of their editorial work and are also uploaded to an open access digital repository. Transcribe Bentham helps to maximise the efficiency of the scholarly editing process by allowing researchers to concentrate on editing and research rather than the task of transcribing papers. It has also been hugely successful in preserving digital records of Bentham's manuscripts, spreading awareness of Bentham's philosophy and encouraging the public to take part in research.

1.2. User activity

The Transcribe Bentham Transcription Desk uses a Mediawiki framework. Users register for an account at the site and choose a manuscript page to transcribe. They can select material according to date, subject matter and the legibility of Bentham's handwriting. They transcribe their document as accurately as possible, applying TEI mark-up with a toolbar in order to tag features of the manuscript such as paragraphs, marginal notes and additions. Once the user is satisfied that they have completed a page, they send a message to notify the Transcribe Bentham team.

Since the platform was launched in 2010, we have had over 36,000 unique views of our website from 141 countries. There are a total of 522 registered users who have worked on at least one transcript. But Transcribe Bentham is primarily reliant upon the work of a small group of 28 'super-transcribers'. These active users have completed over 90% of the completed transcripts on the platform.

As a whole, our volunteer transcribers have worked on a total of 17,304 manuscripts (as of 15 December 2016). Over the past year, the volunteers have transcribed an average of 51 pages per week. This compares favourably with the work of a single researcher or student, who could probably transcribe around 50 to 60 transcripts in the same period. We have welcomed three new regular users to the platform over the past few months. Although this number may seem small, Transcribe Bentham has proved that crowdsourcing projects can flourish with a small but dedicated team of volunteers. The most productive new user is phil.fawcett who has transcribed a total of 228 pages (as of 30 November 2016). According to KPI-A3, the productivity levels of Transcribe Bentham will be used as a benchmark and compared with user activity on the new READ crowdsourcing interface.

The success of Transcribe Bentham is impressive but its dependence on a relatively small number of volunteers makes its future precarious. Our 'super-transcribers' do not all contribute at the same time and productivity would fall significantly if one or two of them stopped transcribing. Research into crowdsourcing projects has demonstrated that it is more difficult to recruit and retain volunteers to complete tasks with a high level of 'granularity' or difficulty. The user surveys we have conducted confirm this impression and suggest that the difficulty of two particular tasks constitute a barrier to participation in Transcribe Bentham. First, users must decipher Bentham's notoriously difficult handwriting. Second, users are required to encode their transcripts using TEI mark-up. The prospect of a new crowdsourcing platform, which integrates HTR technology directly into the transcription process promises to deal with these issues. A streamlined transcription interface would eliminate the need for TEI mark-up, while HTR technology could help volunteers to read Bentham's handwriting. We have a greater chance of attracting new volunteers if the process of transcription is made simpler in these ways.

Yet the necessity of applying layout analysis is something which needs to be considered, as we move Transcribe Bentham from Mediawiki into the Transkribus system. Working with Transkribus means that the images of the Bentham manuscripts must be segmented into baselines. The application of layout analysis is time-consuming and labour-intensive, as even automatically processed pages must be manually checked and corrected. What is more, the correction of erroneous baselines can currently only be performed within the Transkribus expert client. Although manual correction can usually be undertaken in just a few minutes, these steps make the administrative workflow more challenging.

1.3. Administrative workflow

Transcribe Bentham is an ongoing endeavour which requires continual administrative oversight. Each transcript submitted by the volunteers needs to be checked and corrected by Transcribe Bentham administrators. If a transcript is judged to be sufficiently accurate, it is 'locked' and then manually converted into a TEI-compliant XML file (a standard format for preserving digital files). If a transcript is judged to be incomplete, it is left open so that users can continue to work on it.

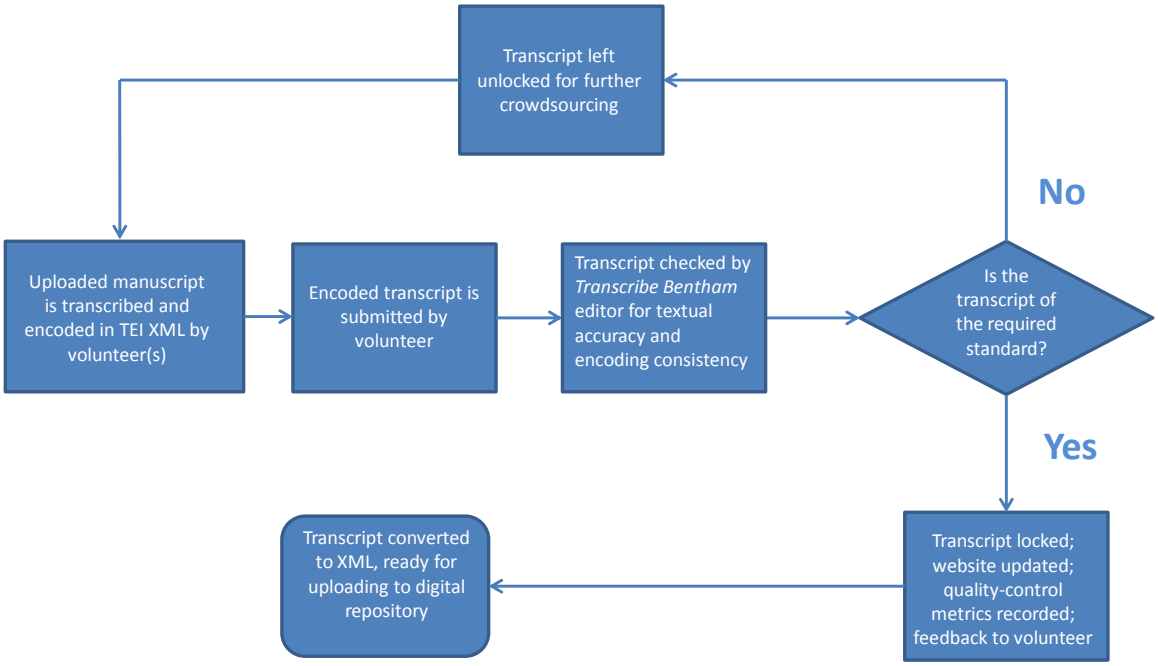

Figure 1 Transcribe Bentham quality control workflow

In addition to these quality control measures, it is necessary to provide users with feedback on their work and respond to any questions or comments they might have. Newly digitised material is also periodically uploaded to the Transcription Desk and 2,695 new images were made available in October 2016.

1.4. Promotion

We promote the project regularly, with a view to publicising our results and encouraging new volunteers to participate. We use the Transcribe Bentham blog, Twitter and Facebook accounts to share news of the transcribers' progress and discoveries. In November 2016 we printed a new leaflet to inform people about Transcribe Bentham's involvement with the READ project. We regularly give presentations about the project to academic audiences from the fields of scholarly editing, crowdsourcing and digital humanities. We also speak to public groups of students, retirees and history enthusiasts who might be interested in volunteering.

2. TSX

2.1. Background

Transcribe Bentham became involved with experiments with HTR technology under the tranScriptorium project. The idea was to establish a new crowdsourcing platform, where non-specialist users could transcribe documents with the assistance of HTR technology. It

was envisaged that this platform could be made freely available to other institutions who would like to get volunteers working on their collections.

UCL submitted 896 pages of transcripts in order to train the HTR engines to recognise the handwriting of the Bentham collection. UPVLC used this 'ground truth' training data to create two HTR models (based on Hidden Markov Models) which are capable of automatically transcribing manuscripts written by Bentham and his secretaries. The accuracy of the UPVLC model can be checked using measurements available in Transkribus: Word Error Rate (WER) and Character Error Rates (CER). The CER measurement is currently the most stable and this indicates that the second, more precise UPVLC model produces transcripts with a CER of around 18%.

ULCC then constructed a new version of the Transcription Desk called TSX (http://www.transcribe-bentham.da.ulcc.ac.uk/TSX/), which offered this HTR technology to users. TSX worked as a web-based overlay to Transkribus, meaning that the storage and management of the Bentham images was undertaken in the latter platform. 700 images from the Bentham collection were uploaded to Transkribus and segmented into text regions and baselines in preparation for the application of HTR technology. The HTR models created by UPVLC were used to produce word graphs for each of these images. These word graphs allowed users to request an automatically-generated transcript of a word, line or entire page of a Bentham manuscript. After a period of testing, TSX was opened to the public in February 2015.

Figure 2 TSX interface

2.2. User activity

Users register for an account at the TSX website and choose a manuscript page to transcribe. If the transcriber comes across a word or line that is difficult to read, they can ask TSX to suggest what the correct reading might be. TSX presents what it considers to be the best hypothesis, as well as a list of other word suggestions. It is also possible for users to

automatically generate the transcript of an entire page, which they would then check and correct.

Nearly 3,500 individuals visited the TSX website during 2015 but only 74 people actually registered for an account during this time. Technical issues with the platform and its lack of compatibility with the Mac operating system may have prevented more people from signing up. The accuracy of transcripts submitted on TSX was comparable to those completed on the Transcribe Bentham Transcription Desk. Most transcripts required between 3 and 6 alterations before they were accepted by a project administrator.

User behaviour indicated that volunteers did use HTR technology but that they preferred to ask TSX to suggest individual words, rather than entire lines or pages. Feedback from our 'super-transcribers' indicated that they were unlikely to engage much with HTR. These users are skilled transcribers who enjoy the intrinsic challenge of transcribing Bentham. The success of Transcribe Bentham has been built on the enormous efforts of our 'super-transcribers', so we need to ensure that they can continue to transcribe independently of HTR in any new version of the platform.

2.3. Administrative workflow

The quality control process for TSX worked similarly to that of the Transcription Desk. Submitted transcripts were checked and corrected in Transkribus. Unfortunately, there were no mechanisms to provide feedback to TSX users.

Quality control proved to be slightly more efficient in TSX. It took an average of 129 seconds (2 minutes and 9 seconds) to check a TSX transcript, compared to 141 seconds (2 minutes and 21 seconds) to check a Transcribe Bentham transcript. This was partly due to the nature of the images uploaded to TSX, where the writing on the page tended to be shorter in length. The segmentation of the images was another factor; verifying the accuracy of a transcript line-by-line is quicker than checking a free-text box. This increased efficiency is significant. Anyone who is thinking of launching a crowdsourcing endeavour needs to assess the resources needed for quality control.

3. READ Crowdsourcing Interface

3.1. Background

We originally planned to continue working with TSX during the course of the READ project. Staffing issues at ULCC prevented this and priority was instead given to the development of the web interface, where the READ crowdsourcing platform will eventually be hosted. The TSX site remains live but is no longer maintained by ULCC or UCL.

TSX showed that HTR technology could potentially make a useful contribution to crowdsourcing projects by helping users to read historical documents and increasing the efficiency of administrative oversight. But the TSX prototype had a number of issues which will need to be resolved in the new crowdsourcing platform being constructed by READ.

The issues with the TSX platform relate primarily to the three following areas:

- Usability. The interface must work smoothly with minimal bugs. Users should be able to move around the site and zoom in and out of images with ease.

- Integration of HTR technology. The presentation of the word suggestions from the HTR engine must be improved. Users should be offered short lists of plausible words when they use computer-assisted transcription. The accuracy of the UPVLC HTR models was also insufficient for integration into a crowdsourcing system. Tests carried out by UIBK indicate that it would be labour-intensive and unsatisfying for a human transcriber to check and correct text that has a CER of over 10%.

- TEI mark-up. It must be possible for users to transcribe documents without the additional complication of applying TEI mark-up. Users should instead be able to mark the features of the manuscript using a 'What You See Is What You Get (WYSIWYG)' toolbar. A version of this 'WYSIWYG' toolbar was under development in TSX but was not fully implemented.

READ will build on the lessons learned from the TSX prototype to create a new crowdsourcing platform which will be more appealing to volunteers. It will be made available via the forthcoming Transkribus web interface. Transcribe Bentham will be a test case for this interface but it will also be freely available to other collection holders.

1,471 images from different parts of the Bentham papers were segmented under tranScriptorium. This represents a significant collection of material with which to launch a new crowdsourcing platform. UCL is continuing to use Transkribus to segment additional material from the Bentham collection to ensure that volunteers have a wide selection of manuscripts to choose from.

3.2. HTR technology

The 'ground truth' training data which was processed by UPVLC during the tranScriptorium project has been reused by URO to create a third Bentham HTR model. This new model (based on Neural Networks) is more accurate according to WER and CER measurements available in Transkribus. The URO model can produce automatic transcripts of the Bentham papers with a CER of only 5-10%. User feedback from TSX revealed that the inaccuracy of words suggested by the HTR engine was a common concern. It therefore makes sense to work with this more accurate HTR model in the future incarnation of the crowdsourcing platform.

We also plan to strengthen the accuracy of the URO model still further by creating more 'ground truth' training data. Much of the tranScriptorium training material was comprised of papers written by Bentham's secretaries, which tend to be neat and legible. The HTR engine will need even more data before it is able to successfully recognise the most complicated manuscripts written in Bentham's own hand.

3.3. User activity

We envisage that this new crowdsourcing platform will work similarly to TSX but with some enhanced functionality. Like TSX, the new interface will be connected to Transkribus. Volunteers will sign up to work on a particular crowdsourced collection (e.g. Bentham). These users will be given the status of 'Volunteer' in Transkribus. 'Volunteers' are able to access 'crowd collections' in the Transkribus web interface but cannot view these documents in the Transkribus expert platform.

Building on some of the outcomes of the TSX experiment, NAF and ULCC are working on a prototype of a new transcription interface which will be integrated into the crowdsourcing platform. This interface is designed to be simpler and more user-friendly than TSX. It is partly modelled on the successful crowdsourcing initiative Shakespeare's World (https://www.shakespearesworld.org/#/) and so may be familiar to potential volunteers. Mostly notably, the image of the manuscript page will fill the entire screen. This will make it easier for users to read the handwriting and zoom in and out on difficult words. This interface is also designed to be flexible so users can move the different elements (image, transcription editor, toolbar) to a position of their liking.

The user chooses a document to transcribe and uses a 'WYSIWYG' toolbar to tag features of the manuscript (with the coding now hidden from view). Users can transcribe on their own or with support from the HTR engine. They can ask the computer to provide suggestions of words or lines, or to generate a complete page to check and correct. Once users are satisfied that they have finished a particular page, they will submit their transcript for approval. Users can view their progress and keep a record of all the documents they have worked on via their user page. If a user wants to continue working on a particular page in more than one session, they will be able to 'lock' their page for a limited amount of time (e.g. one week).

We are optimistic that the inclusion of HTR options will make transcription easier for potential volunteers. URO's HTR model should mean that accurate and useful words are presented to transcribers. Less-experienced transcribers might also find the prospect of checking and correcting an entire page an attractive option. The development of the e-learning app will complement the crowdsourcing platform. Cautious volunteers will be able to practice transcribing Bentham's handwriting before they begin participating in the crowdsourcing initiative.

We will need to be careful to support our existing transcribers. It will be helpful to them if a new platform mirrors the existing look and feel of our Transcription Desk. We also need to ensure that these users can continue to undertake the independent transcription that they currently enjoy doing.

3.4. Administrative workflow

The quality control workflow will be a little different to both Transcribe Bentham and TSX. The crowdsourcing interface will be connected to Transkribus but most of the administrative oversight will take place in the web interface. We will make it possible for Transkribus users to designate their documents as a 'crowd collection'. This allows them to open access to their documents, either to the general public or to a select group of volunteers. This means

that anyone who uploads documents to Transkribus is able to launch a crowdsourcing project should they wish to do so.

The 'Owner' of the documents in Transkribus will be able to conduct the administration of their crowdsourcing project through the new web interface. Project administrators use the web interface to interact with users, check the work of volunteers and export finished transcripts.

We envisage that the administrative workflow will be as follows:

1. 'Owner' uploads images of documents to Transkribus, with option to add extra document metadata in XML files.

2. 'Owner' pre-processes images using automatic layout analysis tools.

3. 'Owner' designates collection as a 'crowd collection'. Images are now accessible to 'Volunteers' in the crowdsourcing interface.

4. 'Owner' undertakes quality control process to proof-read transcripts submitted by 'Volunteers'.

5. If 'Owner' is satisfied with the quality of the transcript, they can set the status of the page to 'complete'. Pages with this status can no longer be edited by 'Volunteers'. If the submitted transcript still needs some work, it is left open for further crowdsourcing.

6. 'Owner' sends a feedback message to the 'Volunteer' to thank them for their submission and explain any changes that have been made to the transcript.

7. 'Owner' can export finished transcripts from Transkribus as TEI files, but also make them available via the REST service.

TSX showed that transcripts which were segmented into lines could be checked more quickly. The implementation of a 'WYSIWYG' toolbar for the transcribers could also help to improve the efficiency of the quality control process. In the current version of Transcribe Bentham, administrators spend most time correcting errors made in the TEI mark-up. When visible mark-up is eliminated from the platform, the time spent checking each page will be reduced and administrators can instead concentrate on ensuring the accuracy of the transcribed text. These questions of efficiency are crucial if we want to encourage other institutions to share their documents on the READ crowdsourcing platform.

The current state of the technology means that digital images must be segmented into text regions, lines and baselines in order for the HTR to work successfully. UCL used Transkribus to analyse the layout of 1,471 pages during the tranScriptorium project, using a mixture of automatic processing and manual correction. UIBK has already strengthened the automatic layout tool in Transkribus but the tool still struggles to process complex manuscript pages with complete accuracy. The process of layout analysis is time-consuming and represents a stumbling-block for other institutions who may be interested in using the READ

crowdsourcing platform to share their collections with the public. Nevertheless it is expected that during 2017 this process will be improved significantly so that the effort needed for correction will be reduced. We also envisage integrating segmentation into the workflow of volunteer contributors on the crowdsourcing platform. Volunteers will work with images which have been automatically segmented but will have the option of adding, deleting or moving the baselines provided by the Transkribus system. The process of drawing and moving lines might be made very straightforward, much as it is in the crowdsourcing platform Shakespeare's World.

4. Conclusion

Transcribe Bentham continues to thrive, engaging new and existing volunteers with historical documents and producing accurate transcripts which are useful for scholarly editing and research. The new interface developed by READ will demonstrate that Transkribus can produce the same benefits for public engagement and scholarship. The TSX prototype showed that users were interested in using HTR technology and that a HTR platform could make both transcription and quality control quicker and simpler. Our work with TSX has also suggested a number of improvements that would produce a new crowdsourcing platform which would be attractive to volunteers.

The READ crowdsourcing interface will be offering a 'WYSIWYG' toolbar, improved document and user management and more accurate HTR word suggestions. It is expected that this enhanced functionality will attract more people to the platform. Dedicated transcribers will still be able to transcribe independently, as they have done for years. Emphasis will also be placed on making sure that the READ crowdsourcing platform works for other institutions who might be interested in working with volunteers. The next phase of the project will involve developing and testing a prototype of this new platform.

5. References

[1] T. Causer, S. Arango, R. McNicholl, G. Mühlberger, P. Kahle and S. Colutto, 2015, 'tranScriptorium. D6.3.2: Evaluation of Public DIA, HTR & KWS Platforms', Public report, tranScriptorium: http://transcriptorium.eu/pdfs/deliverables/tranScriptorium-D6.3.2-31December2015.pdf

[2] 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, (in submission to *Digital Scholarship in the Humanities*)

[3] T. Causer and M. Terras, 2014, ''Many hand 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/

[4] 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

[5]S. Dunn and M. Hedges, 2012, *Connected Communities: Crowd-sourcing in the Humanities. A scoping study,* Report to the Arts and Humanities Research Council: http://www.ahrc.ac.uk/documents/project-reports-and-reviews/connected-communities/crowd-sourcing-in-the-humanities/

[6] R. Grayson, 2016, 'A Life in the Trenches? The Use of *Operation War Diary* and Crowdsourcing Methods to Provide an Understanding of the British Army's Day-to-Day Life on the Western Front', *British Journal for Military History,* 2, 2, pp. 160-185: http://bjmh.org.uk/index.php/bjmh/article/view/96/74

[7] K. Grint, 2013, 'tranScriptorium. D6.1: User needs', Public report, tranScriptorium: http://transcriptorium.eu/pdfs/deliverables/tranScriptorium-D6.1-30June2013.pdf

[8] J. Martin, S. Arango, R. Davis, G. Mühlberger, P. Kahle, S. Colutto, T. Causer and K. Grint, 2014, 'tranScriptorium. D6.2.2: Evaluation of Public DIA, HTR & KWS Platforms', Public report, tranScriptorium: http://transcriptorium.eu/pdfs/deliverables/tranScriptorium-D6.2.2-February2014.pdf