Manuscript transcription by crowdsourcing: Transcribe Bentham

Martin Moyle, Justin Tonra, Valerie Wallace
UCL (University College London)
m.moyle@ucl.ac.uk

Abstract

Transcribe Bentham is testing the feasibility of outsourcing the work of manuscript transcription to members of the public. UCL Library Services holds 60,000 folios of manuscripts of the philosopher and jurist Jeremy Bentham (1748-1832). Transcribe Bentham will digitise 12,500 Bentham folios, and, through a wiki-based interface, allow volunteer transcribers to take temporary ownership of manuscript images and to create TEI-encoded transcription text for final approval by UCL experts. Approved transcripts will be stored and preserved, with the manuscript images, in UCL's public Digital Collections repository.

The project makes innovative use of traditional Library material. It will stimulate public engagement with UCL's scholarly archive collections and the challenges of palaeography and manuscript transcription; it will raise the profile of the work and thought of Jeremy Bentham; and it will create new digital resources for future use by professional researchers. Towards the end of the project, the transcription tool will be made available to other projects and services.

This paper is based on a presentation given by the lead author at LIBER’s 39th Annual General Conference in Aarhus, Denmark, 2010.

Keywords

Crowdsourcing, TEI, digitisation, digital curation, digital humanities, palaeography, manuscript transcription.

Background

Transcribe Bentham is a participatory project based at UCL (University College London). Its aim is to engage the public in the online transcription of original and unstudied manuscript papers of the philosopher and jurist Jeremy Bentham (1748-1832). Transcribe Bentham is led by the UCL Bentham Project, the world centre for Bentham Studies. Founded in 1959, its main aim is the production of a complete scholarly edition of Bentham’s collected works. Bentham’s output over his long lifetime was prolific, and UCL Library Services Special Collections holds 60,000 Bentham folios. They are fully catalogued; around a third have previously been transcribed, albeit to various levels of quality and in different formats; but most are untranscribed and unstudied.

Transcribe Bentham aims to harness the power of crowdsourcing to complete the transcription of 12,500 of Jeremy Bentham’s manuscripts. A short definition of crowdsourcing from Wikipedia is:

> *Taking tasks traditionally performed by an employee or contractor, and outsourcing them to a group of people or community, through an "open call" to a large group of people (a crowd) asking for contributions.*

Wikipedia itself is, of course, heavily dependent on crowdsourced content. (Indeed, the idea of mass voluntary participation in scholarly endeavour dates at least as far back as the formative years of the Oxford English Dictionary in the latter half of the nineteenth century.) Rose Holley [Holley, 2010], in her article ‘Crowdsourcing: How and Why Should Libraries Do It?’, provides a valuable overview of the definition and purpose of crowdsourcing and its relevance to libraries, and examines several recent large-scale participatory projects to identify common characteristics for success. Holley's work helps to confirm that crowdsourcing is an increasingly popular way of gathering content. Nonetheless, Transcribe Bentham is the first effort to attempt to interest the international public in the hitherto somewhat recondite practice of manuscript transcription.

The project lasts for one year to May 2011. It is funded by the UK Arts and Humanities Research Council, led by the UCL Bentham Project, in collaboration with UCL Library Services and UCL Department of Information Studies. It is part of the work of the new UCL Centre for Digital Humanities. The technical development work is being carried out by the University of London Computing Centre.

Aims

Specifically, the aims of Transcribe Bentham are these:

- To digitise 12,500 previously unread Bentham manuscripts.
- To create a public transcription interface, with appropriate training tools, enabling crowdsourced TEI-encoded transcription.
- To promote the project to designated communities of volunteer transcribers.
- To convert any previously-created transcriptions to TEI format.
- To create a web-based 'Ideas Bank', based on the transcripts.
- To carry out a log analysis and a user study on how the public interacted with the project.
- To roll out a generic TEI transcription tool, for use by other transcription projects and services.
- The long-term curation of digitised Manuscripts and TEI-encoded transcripts in the UCL Library Services digital repository.

Project components

Figure 1 shows the interaction between the various project components.

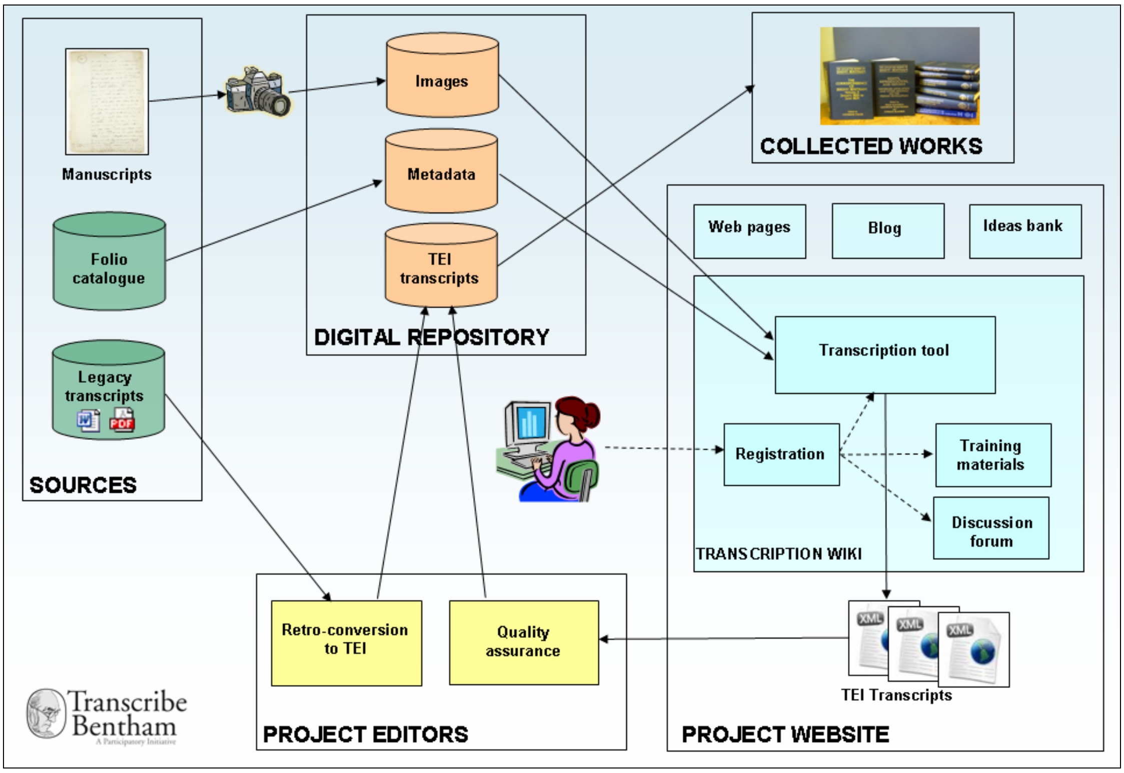


Fig.1 Project components - overview

Digitised manuscripts, records from the existing Bentham Manuscript catalogue and legacy transcripts are uploaded to the UCL Library Services digital repository. Derivative images and cut-down catalogue records are sourced from the repository into a wiki-based 'Transcription Desk', a customisation of open source Wikimedia software. Members of the public may register at the Transcription Desk, which enables them to access training tools, a variety of social networking activities, and to transcribe manuscripts. When a transcriber is happy with his or her work, the TEI-encoded transcript is passed to a project editor; when passed fit, it is uploaded to the digital repository for open access availability, alongside the relevant manuscript image, and long-term preservation. The project editors are also responsible for the TEI-conversion and upload of the legacy transcripts. Ultimately, the transcriptions will help the task of the UCL Bentham Project in producing future editions of Bentham's collected works.

Designing a transcription interface

Designing a public interface for manuscript transcription brings a number of initial challenges. In the first place, manuscript transcription can be difficult for the inexperienced; it is certainly so in the case of the Bentham corpus.

Figure 2 shows sample Bentham manuscript sheets:

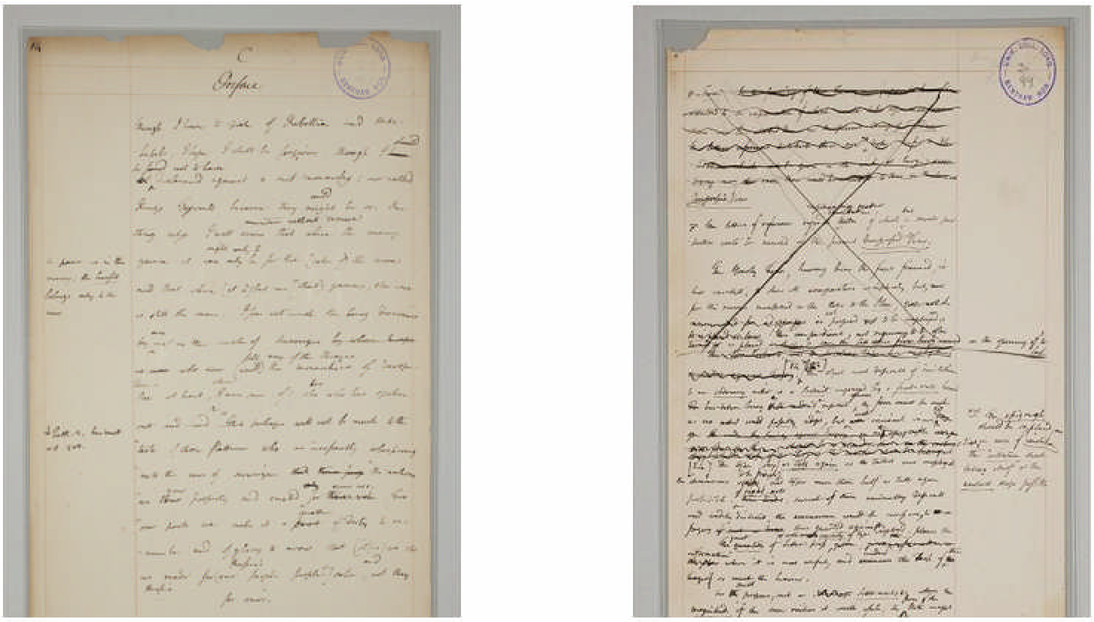
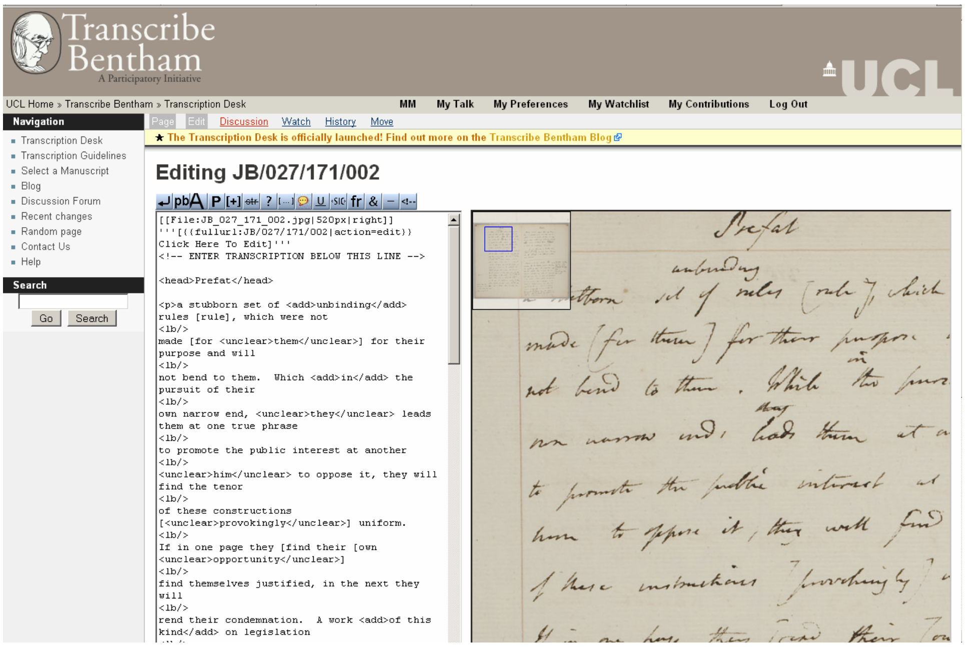


Fig.2. Excerpts from Bentham manuscripts

Legibility is often a challenge; moreover, Bentham’s hand deteriorated as he aged. (The page on the left is an 'early-career' manuscript; to the right is a later folio of "moderate" difficulty.) Additionally, in common with most manuscripts of the time there are deletions, additions, marginal notes, idioscyncratic and inconsistent spellings, and sprinklings of Latin, French and other languages, to name but a few of the obstacles confronting the novice transcriber. A second issue is that the requirement for detailed TEI markup, essential to the sustainability and added scholarly re-use value of contemporary transcription work, introduces yet more complexity for the beginner. Third, quality assurance here depends on expert human intervention, which is expensive, but to impose too high a quality threshold on untutored public volunteers would be quite unrealistic, and could deter participation. A project of this nature has to take care to find a practicable balance between the quantity and quality of the contributions and its running costs. Finally, while a contemporary, dynamic, wiki-based transcription environment is an essential platform for success, care must be taken not to let the technology lead the development: there are numerous potential demographic groups of participants for the project, and the site must strive to provide a welcoming, rewarding and addictive experience for volunteers of all ages and backgrounds.

The Transcription Desk

The Bentham Transcription Desk has been developed with these concerns in mind. A user, having registered, may select a manuscript for transcription. The available manuscripts are classified according to subject matter, date and difficulty - although most of the digitisation effort is deliberately focusing on the early-career, 'easy' folios. A user may look them over before finding one which they feel ready to edit. On selection, the manuscript image is passed to an editing interface (see Figure 3).



★ The Transcription Desk is officially launched! Find out more on the Transcribe Bentham Blog

Editing JB/027/171/002

```
[[File:JB_027_171_002.jpg|520px|right]]
'''[{{fullurl:JB/027/171/002|action=edit}}
Click Here To Edit]'''
<!-- ENTER TRANSCRIPTION BELOW THIS LINE -->

<head>Prefat</head>

<p>a stubborn set of <add>unbinding</add>
rules [rule], which were not
<lb/>
made [for <unclear>them</unclear>] for their
purpose and will
<lb/>
not bend to them.  Which <add>in</add> the
pursuit of their
<lb/>
own narrow end, <unclear>they</unclear> leads
them at one true phrase
<lb/>
to promote the public interest at another
<lb/>
<unclear>him</unclear> to oppose it, they will
find the tenor
<lb/>
of these constructions
[<unclear>provokingly</unclear>] uniform.
<lb/>
If in one page they [find their [own
<unclear>opportunity</unclear>]
<lb/>
find themselves justified, in the next they
will
<lb/>
rend their condemnation.  A work <add>of this
kind</add> on legislation
```

Fig.3. The editing interface

The editing interface presents the manuscript on the right of the page, and a text input box for transcript creation on the left. The manuscript viewer has a zoom facility, which gives the transcriber flexibility, whether to make a detailed inspection of a particular word or section, or to view the whole manuscript in its original layout. By default, the site is a true wiki, with open editing, but a transcription in progress may be locked down for a period by any user who is particularly keen to crack its code.

A toolbar has been developed to facilitate TEI encoding (Figure 4):

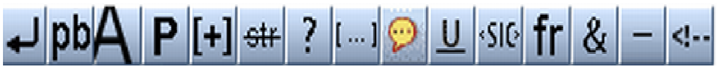


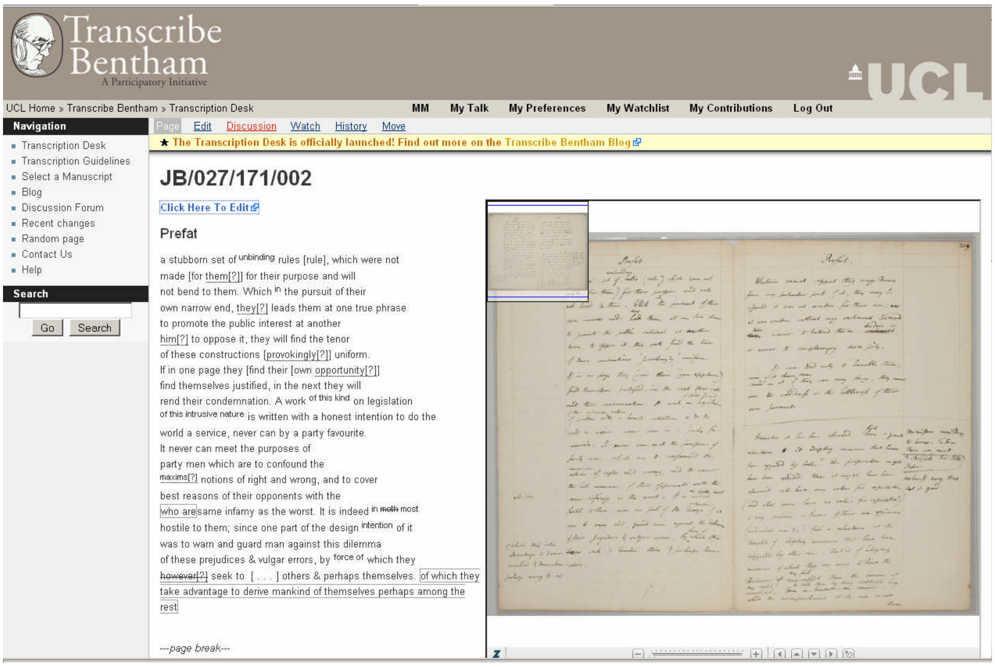
Fig.4. The TEI toolbar

From left to right, the buttons allow a user to annotate the text of a transcript with the following attributes:

Line Break – Page Break – Heading - Paragraph - Addition - Deletion - Unclear Reading - Illegible Text – Marginal Note - Underline - Unusual Spelling - Foreign Language - Ampersand – Long Dash - User Comment.

The TEI toolbar is an important innovation, which hides much of the complexity of TEI markup from the transcriber.

When a transcript is saved and viewed, the encoding is rendered as HTML to give a manuscript-like appearance to the finished work, as shown in Figure 5:



★ The Transcription Desk is officially launched! Find out more on the Transcribe Bentham Blog

JB/027/171/002

Prefat

a stubborn set of unbinding rules [rule], which were not
made [for them[?]] for their purpose and will
not bend to them. Which in the pursuit of their
own narrow end, they[?] leads them at one true phrase
to promote the public interest at another
him[?] to oppose it, they will find the tenor
of these constructions [provokingly[?]] uniform.
If in one page they [find their [own opportunity[?]]
find themselves justified, in the next they will
rend their condemnation. A work of this kind on legislation
of this intrusive nature is written with a honest intention to do the
world a service, never can by a party favourite.
It never can meet the purposes of
party men which are to confound the
maxims[?] notions of right and wrong, and to cover
best reasons of their opponents with the
who are same infamy as the worst. It is indeed in ~~meth~~ most
hostile to them; since one part of the design intention of it
was to warn and guard man against this dilemma
of these prejudices & vulgar errors, by force of which they
~~however[?]~~ seek to [. . . .] others & perhaps themselves. of which they
take advantage to derive mankind of themselves perhaps among the
rest

---page break---

Fig. 5. A completed transcript, with TEI encoding rendered as HTML

With diversity of audience in mind, the site has help and guidance in various formats, from static pages on the fundamentals of palaeography to a video tutorial. There is a discussion forum, including channels on Bentham, markup and transcription. Facebook-like transcriber ‘profiles’ are available to registered users. Reward is an important part of successful crowdsourcing: Transcribe Bentham has a points system, a topic of lively discussion on the forum, and a league table of transcribers is displayed on site. The site also displays the “Benthamometer”, a visual record of overall progress in the shared effort to read and encode Jeremy Bentham’s manuscript legacy.

Sourcing the crowds

The Transcription Desk is openly and freely available for use, and any prospective transcriber is welcome to register. However, in order to maximise interest, the project is actively being marketed to three broad priority groups of participant. One is the UK schools sector. The project is being promoted nationally to teachers, particularly those with responsibilities for 16-18 year-olds; additionally, workshops and classes are being arranged in the London schools with which UCL and the Bentham Project have existing outreach links. The project is also targeting the academic sector, both educators, in palaeography and humanities research methods, and researchers, in the digital humanities as well as in relevant traditional disciplines such as law, economics and history. The final target audience is the miscellaneous body of amateur historians, Bentham enthusiasts and interested members of the general public. Transcribe Bentham has communication plans for each group, with strategies ranging from workshops to press releases to paid-for advertisements in academic publications and magazines. Careful planning is essential: publication lead times and the academic cycle must be taken into account, and at 12 months, the project is relatively short. Alongside targeted marketing, the project also has strong network visibility, with Facebook and Twitter presence as well as the project's own website, blogs and feeds.

Outcomes

Transcribe Bentham promises a number of positive outcomes. It is helping to stimulate public engagement with scholarly archives and manuscript transcription - always a challenge, but perhaps carrying extra significance at a time when, in the UK at least, humanities research units and related services are under intensified pressure to quantify their social impact. It will also help to open up the thought of Jeremy Bentham to new audiences. Bentham was a widely-respected thinker, whose work challenged many institutions, practices and beliefs. His writings on social policy, law and economics were highly influential in the public administration reforms of the nineteenth century, and many are still relevant today. Transcribe Bentham will create an 'Ideas bank' to bring Bentham to new audiences, including policy makers, the media and the public. Another lasting output will be the creation of an open access corpus of digital manuscripts and transcripts, digitally preserved for the benefit of scholars for decaades to come. Transcribe Bentham will also make a tried and tested version of the Transcription

Desk tool available for re-use and adaptation by other projects and services. Finally, the project will collate quantitative and qualitative data on the public participation, enriching our understanding of how users interact with digital resources and helping to inform future digitisation programmes.

Conclusion

Transcribe Bentham is the first major crowdsourcing transcription project, from which we will learn much about the nature of community engagement with digitised resources and social media. The project meets most of the success criteria identified by Rose Holley [Holley 2010], but transcription is a relatively complex task, certainly more so than those required of the public in other major participatory projects to date, and notwithstanding the best efforts of the project team to test, help, guide and simplify. It remains to be seen how far the public's enthusiasm for the idea of historical manuscript transcription will be checked by difficulties in practice. Even failure to attract large numbers of high-quality completed transcripts would be a worthwhile research finding - the deep log analysis commissioned by the project will be revealing, and will help to show the way for future transcription undertakings - but the Transcribe Bentham team is confident, and progress so far has been encouraging, showing high levels of interest and engagement, lots of activity on the forum, and much transcription under way. Readers of LIBER Quarterly are warmly invited to visit Transcribe Bentham, join the crowd, and try online manuscript transcription for themselves!

References

Holley, R. (2010): Crowdsourcing: How and Why Should Libraries Do It? D-Lib Magazine, Volume 16, Number 3/4, 2010. [doi:10.1045/march2010-holley]

Web sites referred to in the text

Transcribe Bentham http://www.ucl.ac.uk/transcribe-bentham/
Text Encoding Iniative (TEI) http://www.tei-c.org/index.xml
UCL Bentham Project http://www.ucl.ac.uk/Bentham-Project/
UCL Library Services Digital Collections http://digital-collections.lib.ucl.ac.uk
UCL Special Collections http://www.ucl.ac.uk/Library/special-coll/index.shtml
Wikimedia http://www.wikimedia.org/