Manuscript Transcription: the Habits of Crowds

Title	Manuscript Transcription: the Habits of Crowds
Author(s)	Tonra, Justin
Publication Date	2013

Justin Tonra,
NUI Galway

Manuscript transcription: the habits of crowds.

1

I am here today to talk about my involvement in *Transcribe Bentham*, a project established to crowdsource transcriptions of the manuscripts of Jeremy Bentham, and to reflect more generally on the issues at stake in such an endeavour. I don't want to assume that you are all familiar with the project, so I will begin with a brief introduction. First of all, who was Jeremy Bentham? [SLIDE] He was a London-born philosopher and reformer, perhaps best known for devising the doctrine of Utilitarianism and conceiving the panopticon prison. Having donated his body to UCL, the university that was founded upon principles which he articulated, Bentham's clothed skeleton [SLIDE] (or auto-icon, as he described it) now sits on public display in the South Cloister. Less well-known are his writings on topics as diverse as contraception, convict transportation, and animal rights. His contributions to the disciplines of history, politics, law, philosophy, and economics amount to an important legacy for scholars who work in these fields.

The Bentham Project was established in 1959 to produce a new scholarly edition of the works and correspondence of Jeremy Bentham. It was conceived with the dual imperative that inspires many scholarly editions: to demonstrate the importance of Bentham's work and responsd to the inadequacy of existing texts that represented that work. Since 1968, twenty-seven volumes of the new *Collected Works* have been published by the project, but this editorial task is daunting. It involves exploring and navigating a very substantial body of manuscript material held at UCL (where there are 60,000 folios) and the British Library (which holds another 12,500 folios). Scholarly editing is a slow and scrupulous process, and beyond the identification of material for

editing in the Bentham archive, one of the more time-consuming tasks of the editorial process is transcribing manuscripts in Bentham's often erratic hand.

Enter *Transcribe Bentham*. The project emerged out of discussions between the Bentham Project and UCL Library, during which Martin Moyle (of the Library) proposed the establishment of a resource to facilitate crowdsourced transcription of the Bentham Papers. A consortium was formed with the University of London Computer Centre and UCL's Centre for Digital Humanities, and initial project funding was secured. Two objectives were central to the idea of soliciting members of the public to transcribe Bentham's papers:

- making Bentham's thought—a large proportion of which remains unstudied in unpublished manuscripts—more accessible to the world at large;
- providing transcribed texts to assist Bentham Project editors in the task of preparing further editions for the *Collected Works*.

2

The Transcription Desk, which forms the core of *Transcribe Bentham*, was designed and built between May and September 2010, and drew heavily on the different talents of the various member of the consortium (along with UCL's Creative Media Services, who set about the task of photographing Bentham's manuscripts).

[Brief demonstration of Transcription Desk, explaining design and implementation of Transcription Tool].

The general workflow of the project is illustrated by this image: [SLIDE]

1. Digitised manuscript images and records from the existing Bentham manuscript catalogue are uploaded to the UCL Library digital repository.
2. Images and catalogue records from the repository populate the Transcription Desk, where members of the public may viewing training material and transcribe manuscripts after registering.
3. When a transcriber is satisfied with his or her work, the TEI-encoded transcript is passed to a project editor for moderation. When passed fit, it is uploaded to the digital repository for open access availability, alongside the relevant manuscript image and metadata, and for long-term preservation.
4. In addition, the project staff were also responsible for converting to TEI (from MS Word) and uploading the Bentham Project's legacy transcripts.
5. Ultimately, the transcriptions aid the task of the UCL Bentham Project in producing future editions of Bentham's collected works.

3

If you build it, will they come? A crowdsourcing project is redundant without the involvement of a crowd, so how did *Transcribe Bentham* attract its community of volunteers? My colleague Valerie Wallace was responsible for a publicity campaign that targeted the general public, academic community, libraries and archives professionals, and schools (with a budget of £1,000). With the launch of the project, we sent a press-release to traditional media in the UK, and distributed leaflets at various conferences and institutions (leaving some in the care of Jeremy's auto-icon). We also promoted the project through social media, with Facebook and

Twitter accounts and regular blogging. Notifications were sent to a large number of academic and professional mailing lists, online forums, and the websites of academic societies. A portion of the website was tailored to explaining how *Transcribe Bentham* related to relevant A-Levels and Scottish Highers, including reading lists and direct links to groups of manuscripts of relevance to particular areas of study. Schools from East London and Chester visited the office and participated in user-testing before launch, and transcription thereafter. In terms of raising awareness of the initiative, the publicity campaign was a success. Both social and traditional media contributed to the development of the crowd. Social media accounted for much of the project's exposure, particularly in academic contexts, while a feature article in the *New York Times* online on 27 December 2010 (and in print the following day) resulted in a spike in registrations (more detailed analysis of usage trends can be found in the *LLC* and *DHQ* articles). Another significant factor in the success of the project, however, has been the serendipitous discovery of highly motivated and enthusiastic individual members of the crowd. If you take a look at the table of Top Contributors [HOMEPAGE], you will see three prodigiously productive transcribers without whom the practical outcomes of the project would be a good deal more modest. There is no accounting for discoveries like this. One might mount a publicity campaign costing millions of euro and fail to attract a volunteer of this kind. It seems to me to be purely a matter of chance.

4

From a practical perspective, it is worth noting some of the tasks that are involved in maintaining a project such as this. Because of the accuracy required of texts that form the raw materials of scholarly editions, checking volunteers' transcriptions forms a large part of the ongoing project.

To a greater or lesser degree, however, most crowdsourcing endeavours will need to dedicate some labour to monitoring submissions and quality control.

This [SLIDE] shows the time spent checking 639 transcripts, submitted by 25 individual volunteers, between 1 October 2012 and 22 February 2013. These transcripts total 212,521 words (average 332 words per transcript) not including mark-up, or 300,439 words (average 470) including mark-up. And this [SLIDE] details the level of correction that these submissions required. Project staff spent a total of 63 hours and 14 minutes (about eight days) worth of work checking these submissions over the course of (almost) five months. And this does not take into account time spent converting transcripts to well-formed XML files, updating the website, or writing the weekly progress reports for the project blog. Since 22 Feb, a new member of staff has begun checking transcripts—it may be interesting to compare the time spent checking by an experienced staff member and someone new to the process, but that is a topic for another discussion. The point of this information is simply to illustrate that in a crowdsourcing project of this kind, quality control is a continuous process, and one's investment therein must be proportionate to the desirability of accurate data.

5

Transcribe Bentham was initially funded for one year, and at the end of that period it was considered a qualified success (our articles in *LLC* and *DHQ* that are based on this initial period are quite reserved in judging the project's value for money). But as time passes and costs diminish to little more than maintaining a researcher's salary, the project's benefits will become

more apparent. This process has already begun, as I'll describe in a moment, but for now, what does the future hold for the project?

It has been funded for a further two years, from 1 October 2012, by the Andrew W. Mellon Foundation's 'Scholarly Communications Programme,' as part of a wider scheme entitled the 'Consolidated Bentham Papers Repository.' The British Library has also joined the project consortium. This grant will fund the digitisation of almost all of the remainder of the UCL Bentham Papers (the initial project funding did not cover the full expense of digitising all 60,000 folios in the UCL collection), and all of the British Library's Bentham manuscripts (around 12,500). Ultimately, these digital surrogates will all be stored in UCL's digital repository, thus reuniting the collection for the first time since Bentham's death.

Just as importantly, the Transcription Desk will be refined, modified, and relaunched during the next phase of the project. Taking into account feedback from volunteers and observations from staff, the most significant changes will be:

- to incorporate a more flexible image viewer, which will allow volunteers to rotate and resize the manuscript image to suit their requirements;
- general improvements to the appearance and functionality of the Transcription Desk, especially in making it more straightforward to select material for transcription. Volunteers have reported difficulties in distinguishing between untranscribed, partially-transcribed, and completed transcripts, and of not being able to sub-categorise (it is currently difficult to find, for example, a partially-transcribed panopticon manuscript of moderate difficulty. Faceted browsing would simplify this process a great deal);

- the introduction of an alternative WYSIWYG transcription interface, so that the TEI mark-up is under the hood, and volunteers can concentrate on transcription alone. Though the TEI toolbar has made the addition of TEI markup as straightforward as we initially thought possible, volunteers have reported that the TEI markup has either limited their participation, or dissuaded them from taking part at all. By introducing the WYSIWYG interface, we hope that user recruitment and retention can both be increased. It will also make the quality control process more efficient as volunteers using the WYSIWYG interface will be unable to amend the generated markup which previously appeared in the transcription box: about half the time spent checking a given transcript is generally expended upon checking the mark-up.
- The project will also experiment with recruiting 'volunteer moderators' to check the accuracy of transcripts. This will initially necessitate sampling their work to check that it meets the requirements of the project. This plan is potentially fraught with all kinds of issues, so it will need to be undertaken with a great deal of care and attention (it may even turn out to be counter-productive: only time and testing will tell).

6

One benefit of the project has been its provision of open-source code for other projects that wish to build similar crowdsourcing tools. It is currently being tested by a number of projects, including the Edvard Munch Archive in Oslo, and the George Boole Archive here in Cork. But what of its primary aims to increase access to Jeremy Bentham's thought and provide texts for his Collected Works?

- The UCL Bentham Papers consists of 60,000 folios, and the BL collection some 12,500. The Bentham Project has estimated that the combined collection will require about 100,000 transcripts before it is complete (some folios contain multiple pages).
- In the period 1959-2010, 20,000 folios (c.28K transcripts) were transcribed by Bentham Project staff, at average of 549 transcripts per year. At this rate, it would take another 131 years to transcribe them all.
- In the period from the site's launch on 8 September 2010 to 15 March 2013, volunteers completed 5,243 transcripts, at an average of 2,080 per year. At that rate, it would take 32 years to transcribe the remainder.
- If the proposed improvements to the Transcription Desk went live tomorrow, the project estimates that about 100 transcripts could be produced each week (c.5,200 per year). In this hypothetical scenario, the remainder of the collection could be transcribed in 12 years. So, taking into account the costs of launching the project, digitising the material, and staffing the website, if the remainder of the collection was transcribed by volunteers, there are still huge financial and labour costs to be avoided.

University College Cork
16 May 2013