Transcription maximized; expense minimized? Crowdsourcing and editing *The Collected Works of Jeremy Bentham**

Tim Causer
Bentham Project, University College London

Justin Tonra
Department of English, University of Virginia

Valerie Wallace
Bentham Project, University College London and Center for History and Economics, Harvard University

Abstract

This article discusses the crowdsourced manuscript transcription project *Transcribe Bentham*, and how it will impact upon long-established editorial practices at the Bentham Project, University College London, which is producing the new and authoritative edition of *The Collected Works of Jeremy Bentham*. We site *Transcribe Bentham* in the burgeoning field of scholarly crowdsourcing projects, and, by detailing our experiences of running and administering the project, attempt to assess the potential benefits of engaging the public in humanities research. The article examines the conceptualization and development of *Transcribe Bentham*, and how editorial practices at the Bentham Project may change as a result. We account for the design of the bespoke transcription tool which is at the project's heart, and which allows volunteers to transcribe the material and encode it in TEI-compliant XML. We attempt to answer five key questions: is crowdsourcing the transcription of complex manuscripts cost-effective? Is crowdsourcing exploitative? Are volunteer-produced transcripts of sufficient quality for editorial use and uploading to a digital repository, and what quality controls are required? Does crowdsourcing ensure sustainability and widen access to this priceless material? And finally, should the success of a project like *Transcribe Bentham* be measured solely according to cost-effectiveness or the volume of work produced, or do considerations of public engagement and access outweigh such concerns?

Correspondence:
Bentham Project, Faculty of Laws, University College London, Bentham House, Endsleigh Gardens, London, WC1H 0EG.
Email:
t.causer@ucl.ac.uk

**Transcribe Bentham* is a collaborative project involving several people. Professor Philip Schofield, Dr Tim Causer, Dr Justin Tonra, and Dr Valerie Wallace at the Bentham Project in the Faculty of Laws, UCL; Richard Davis at the University of London Computer Centre; Martin Moyle at UCL Library Services; Tony Slade at UCL Learning and Media Services; and Dr Melissa Terras at UCL's Department of Information Studies and Centre for Digital Humanities.

Literary and Linguistic Computing, Vol. 27, No. 2, 2012.

doi:10.1093/llc/fqs004 Advance Access published on 28 March 2012

1 Introduction

The philosopher and reformer Jeremy Bentham (1748–1832) is a figure of outstanding historical importance and contemporary significance in intellectual life. He was the founder of the modern doctrine of utilitarianism, a seminal figure in legal philosophy, a major theorist of representative democracy, and originator of contemporary notions of surveillance through his proposed panopticon prison. There is international interest in his writings, which are studied by philosophers, historians, political theorists, lawyers, and economists. The Bentham Project at University College London was established in 1958 with the aim of publishing the new authoritative scholarly edition of *The Collected Works of Jeremy Bentham* and thus far, has published twenty-nine of a projected seventy volumes.[1] This edition will replace the inadequate and incomplete eleven-volume edition published between 1838 and 1843 by Bentham's literary executor, John Bowring.[2] The Bowring edition contains a number of omissions and it is only during the last two years that the Bentham Project has begun to recover Bentham's thinking on religion and sexual morality, which will be published in a forthcoming *Collected Works* volume entitled *Not Paul, But Jesus*. Furthermore, the manuscripts relating to Bentham's panopticon prison scheme still require transcription, as do large amounts of important material on civil, penal, constitutional, and international law, on economics, and legal and political philosophy.

The volumes of the *Collected Works* are, to a large extent, based on edited transcripts of Bentham's unpublished manuscripts, 60,000 folios of which are housed in UCL Library's Special Collections, while a further 12,500 are held by the British Library. For over half a century, editorial staff at the Bentham Project have transcribed and edited Bentham's manuscripts according to a set of established editorial conventions, and transcribed an estimated 20,000 folios by hand, by typewriter, and by word processor. However, before the launch of *Transcribe Bentham* in September 2010, an estimated 40,000 manuscripts remained untranscribed and the completion of the *Collected Works*—the publication of a further forty-one volumes—seemed a long way off, and a great deal of material was unknown, let alone adequately studied.

Transcribe Bentham was established with a one-year grant from the Arts and Humanities Research Council, under their Digital Equipment and Database Enhancement for Impact (DEDEFI) scheme, with the intention of engaging the public with Bentham's thought and works, creating a searchable digital repository of the collection, and quickening the pace of transcription and publication by recruiting unpaid online volunteers to assist in transcribing the remaining manuscripts. Anyone, anywhere in the world with an internet connection can participate in *Transcribe Bentham*, and volunteers require no prior background knowledge or technical expertise. After registering a user account, participants transcribe Bentham's manuscripts into a text box and, using a customized toolbar (Fig. 2), encode the features of the manuscripts in Text Encoding Initiative (TEI)-compliant Extensible Mark-up Language (XML).[3] The transcripts produced by volunteers thus have a dual purpose: they will feed into the Bentham Project's editorial work and help form the basis of printed editions of the *Collected Works*, and are uploaded to UCL's digital Bentham repository where, owing to the TEI encoding, they render the collection fully accessible to all.

A number of prominent projects have put crowdsourcing to great effect to generate research and widen access to collections. Amongst others, the *Zooniverse* citizen science community—which includes, amongst others, the *Galaxy Zoo* and *Old Weather* projects—harnessed the public's enthusiasm to generate scientific research; the National Library of Australia successfully recruited amateur and family historians, and those with a general interest in Australian history, to correct Optical Character Recognition (OCR) text of digitized newspapers; and Ancestry.com's *World Archives Project* has utilized the enthusiasm and goodwill of genealogists to transcribe name indexes which will be of use to other researchers.

More recently, there have been explorations and attempts to crowdsource a more complex task: the transcription of manuscript collections, with the intention of engaging the public with, and adding value to, such archives. *Scripto*, developed by the

Center for History and New Media at George Mason University, and *T-Pen* from Saint Louis University, are open-source tools that aim to facilitate manuscript transcription. *Scripto* is currently being tested on the Papers of the American War Department, whereas *T-Pen* currently appears to be geared more towards transcribing medieval and renaissance manuscripts. The University of Iowa has crowdsourced the transcription of diaries written during the American Civil War, but without a specific transcription tool or any mark-up; instead the transcript is typed into a plain text box and emailed to the project administrators. Few projects, however, have tested the feasibility and practicalities of crowdsourcing manuscript transcription on a large scale within an academic context, or examined its impact on scholarly editing. Hitherto, academia has been a closed set and only scholars have played the starring roles. *Transcribe Bentham* invites the public to play a part in academic research and attempts to break down traditional barriers.

In order to assess the potential benefits of crowdsourcing, the *Transcribe Bentham* team set out to answer the following five key questions. First, would crowdsourcing the transcription of manuscripts be cost-effective? That is, would a sufficient number of transcripts be produced to justify the money and time spent on the development of a transcription tool, and the employment of two full-time Research Associates to coordinate the scheme? Second, is crowdsourcing exploitative, or do volunteers feel rewarded for their efforts? Third, would the crowdsourced transcripts be of a sufficient quality for editorial use at the Bentham Project, and for uploading to the digital repository? Or, put another way, what level of quality control is required? If no quality control is required, could crowdsourcing then render obsolete the skills of a professional scholar-transcriber? Fourth, would crowdsourcing ensure sustainability and widen access by creating a permanent collection of Bentham papers freely accessible to all? And finally, how is the success of a project like *Transcribe Bentham* to be measured? Do issues of sustainability and public engagement outweigh concerns about time and cost? Or is cost-effectiveness the main criterion against which the success of the project should be judged?

This article will attempt to answer these, and other questions by reviewing the experiences of the *Transcribe Bentham* team in conceptualizing and running the project. We shall outline how crowdsourcing worked—or did not—for us, and consider how viable it is for an academic editorial project. Section 2 will describe the development of our transcription tool; section 3 will assess the contributions made by our transcribers and outline the level of moderation undertaken by staff; and section 4 will judge whether the project can be considered a success, and assess the impact of *Transcribe Bentham* on the work of the Bentham Project. The article will conclude by reflecting on the potential implications of crowdsourcing for scholarly editing.

2 Project Development

2.1 Costs

The Bentham Project was awarded a grant of £262,673 from the Arts and Humanities Research Council, for a period of twelve months beginning in April 2010, and Fig. 1 illustrates the main points of expenditure. High-quality, manipulable images of Bentham manuscripts were unavailable prior to the start of the project, so a significant investment was made in photographing and digitizing the material, which was carried out by UCL Learning and Media Services. Another substantial portion of the grant was allocated to the University of London Computer Centre for assembling and hosting the Transcription Desk, and programming of the transcription tool. However, as can be seen, the two greatest expenses were substantial indirect costs associated with research grants, and academic salaries, primarily those of two full-time Research Associates, but also including a portion of the professorial salary of the project director, UCL Library Services' Digital Collections curator, and a consultant from UCL's Centre for Digital Humanities. Money was also assigned to UCL's Department of Information Studies to produce a server log analysis.[4]

2.2 Designing the tool

The interface that hosted the manuscript images and transcription tool—the 'Transcription Desk'—was a customized MediaWiki, programmed and put

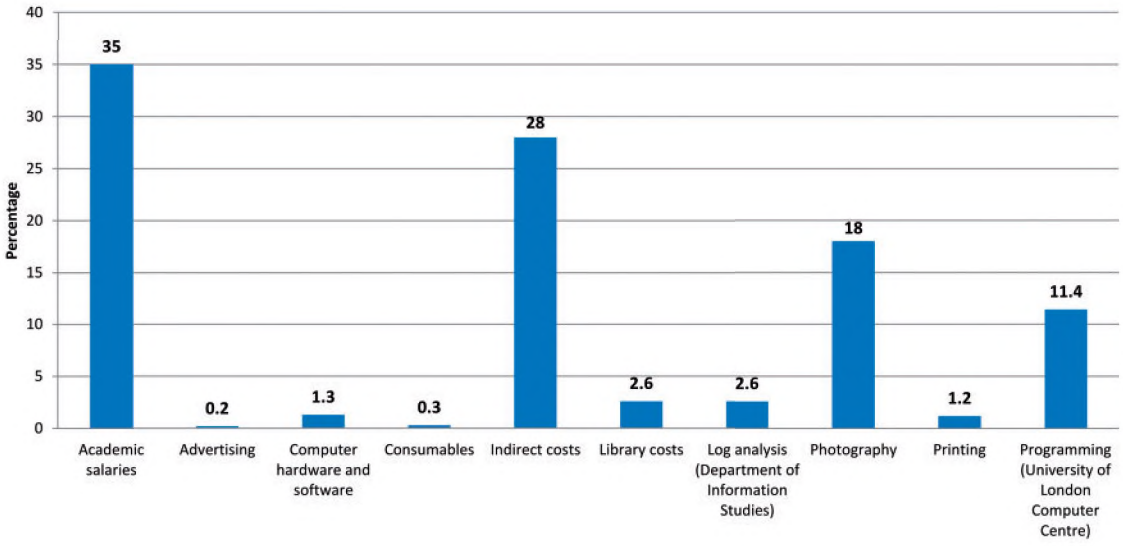

Fig. 1 Project grant allocation.

together by a team at the University of London Computer Centre (ULCC) led by Richard Davis. This not only provided the means of integrating these essential components, but also allowed the inclusion of pages on guidelines for users, project documentation, a discussion forum, and social media that enabled interaction and discussion between users and project staff. The MediaWiki tool captured the spirit of collaborative endeavour integral to the project. Since the basic unit for transcription is the manuscript page, it is not always possible for a user to complete a full transcription in one sitting. Thus, it was important to offer the option to save a transcript and return to it at a future time. It was also considered desirable to allow users to make corrections to one another's transcriptions, on the assumption that a second opinion might be valuable for more difficult interpretative decisions. MediaWiki's exhaustive revision history for any given page also meant that a record of all variant transcripts would be maintained, and it is an easy matter to revert to earlier versions in the case of spam, vandalism, or volunteers mistakenly overwriting the work of fellow transcribers.

In order to begin the process of crowdsourcing the transcription of Bentham's manuscripts, two components were vital: high-resolution digital images of the manuscripts, and a tool to allow users to transcribe the text. Images of the manuscripts were created by Tony Slade of UCL's Learning and Media Services photographic team at a resolution that would facilitate a high degree of magnification.[5] This factor was important because users are often presented with a large manuscript sheet to transcribe,[6] and magnification and navigation around the page were deemed necessary to enable a clear view of all of Bentham's handwriting.

The tool to enable users to transcribe manuscripts was developed with simplicity in mind. It was based around a plain-text box into which users would type their transcripts and either submit them for verification when complete, or save them for future completion. Acquiring the basic text of Bentham's writings was the primary goal of the crowdsourcing initiative, but during the development of the transcription tool, a further component was added, which allowed volunteers to add some basic formatting to their transcription. Thus, users had the opportunity to record both the linguistic and bibliographic codes of the manuscript they were transcribing.

The transcription toolbar (Fig. 2) was created for this purpose; while users are transcribing (or after they have completed the transcription), they can highlight a piece of text, or a position in the text, and click a button on the transcription toolbar in

Button	↵	pb	A	P	[+]	~~	?	[...]		U	x^2	sic	fr	&	—	<!--
Function	Line Break	Page Break	Heading	Paragraph	Addition	Deletion	Questionable Reading	Illegible Text	Marginal Note	Underline	Superscript	Unusual Spelling	Foreign Language	Ampersand	Long Dash	User Comment
Rendering	-	-	text	-	text	~~text~~	text[?]	[. . .]	text	text	te^{xt}	text	text	&	—	-

Fig. 2 Transcription toolbar.

order to identify a particular characteristic of the chosen portion. These include spatial and organizational features such as line breaks, page breaks, headings, and paragraphs; linguistic features like notes, unusual spellings, and foreign-language text; compositional features such as additions and deletions; and interpretive decisions about questionable readings and illegible text.

In order to determine which elements should be encoded, discussions took place with experienced *Collected Works* editors to identify which characteristics occurred most frequently in the manuscripts, and which would be most desirable to encode. A survey across the chronological span of the manuscripts which would be uploaded to the Transcription Desk confirmed the regular occurrence of these features, which were then assigned buttons on the transcription toolbar. If volunteers choose to include this type of information in their transcripts, they are adding a further layer of depth and complexity that assists in the service of making the transcripts searchable in a thorough and categorical fashion.

In practice, the transcription toolbar encodes the chosen text or position within the text by adding TEI XML tags to the existing transcript. This is an extensible, non-proprietary format that ensures the interoperability and non-obsolescence of the encoded Bentham transcripts once they are saved. If users notice that the word 'justice' has been deleted by Bentham in the manuscript they are transcribing, and they wish to record this fact, they highlight the word 'justice', click the 'Deletion' button on the toolbar, and a set of tags to identify the deletion surround the deleted word in the transcript, thus: ~~justice~~ (Fig. 3).

Because the transcription is entered into a plain-text box, it is not possible to distinguish the code from the transcript with coloured text or other formatting. To avoid obscuring users' transcriptions with code, minimal mark-up was employed, using only element names, and avoiding attributes and attribute values where possible. To advise users on identifying important features of the manuscript and the correct use of the transcription toolbar, the Transcription Desk carries a complete page of transcription guidelines, and a condensed version explaining the basics of encoding. The latter also included video tutorials to illustrate how the transcription tool should be used.

Once the user has saved or completed his or her transcription, a formatted version of the transcript is displayed opposite the manuscript image, with CSS[7] used to render the TEI XML-encoded text within the transcript.[8] Thus, for users who wish to read or search Bentham texts submitted by other transcribers, a code-free and simply rendered version is available.

2.3 Changes to editing practice

In the past, Bentham Project editors produced word-processed transcripts in a number of formats, including *Scientex*, *WordPerfect* and, most recently, in a bespoke, ten-year-old 'Bentham template', created in *Microsoft Word*. This template denotes some of the structural features of the manuscripts, including headings, subheadings, Bentham's footnotes, and marginal notes, thereby providing instructions to the typesetter of the *Collected Works* volumes. Editors have also developed specific conventions for transcribing other features of the manuscript, such as additions, deletions, and illegible text. For example, where Bentham provides an alternative reading above a word in the text, it is recorded between forward slashes: *in this way /manner/ it is possible.*

This amounts to a shorthand notation system, where forward slashes are used to identify alternative text. Indeed, in *Microsoft Word*, it would be possible to transcribe the appearance of the manuscript by formatting the text to make the word

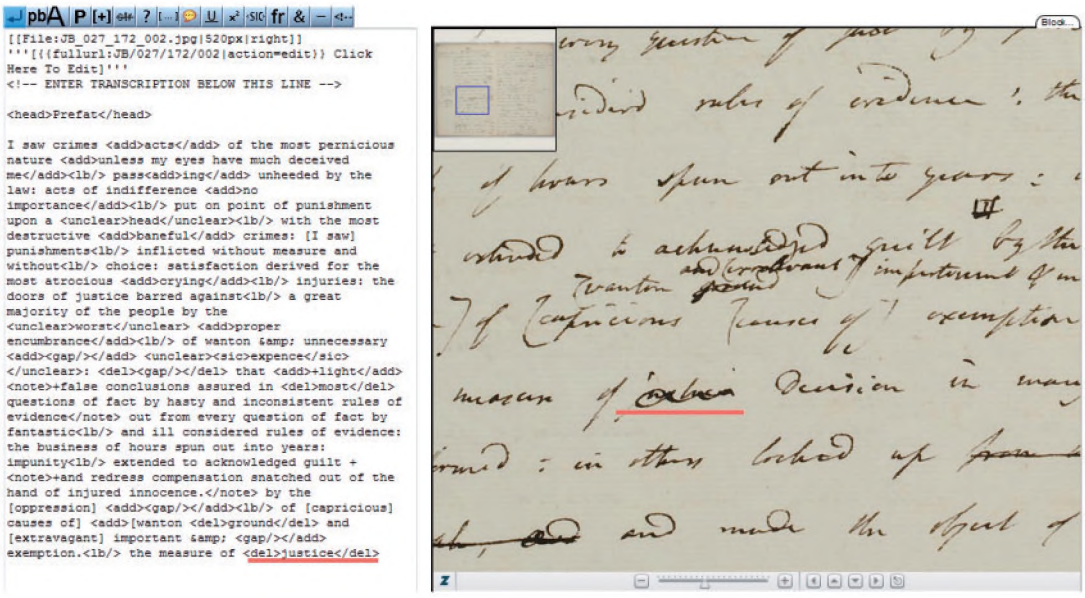

Fig. 3a The transcription interface showing manuscript JB/027/172/002, alongside the encoded transcript.

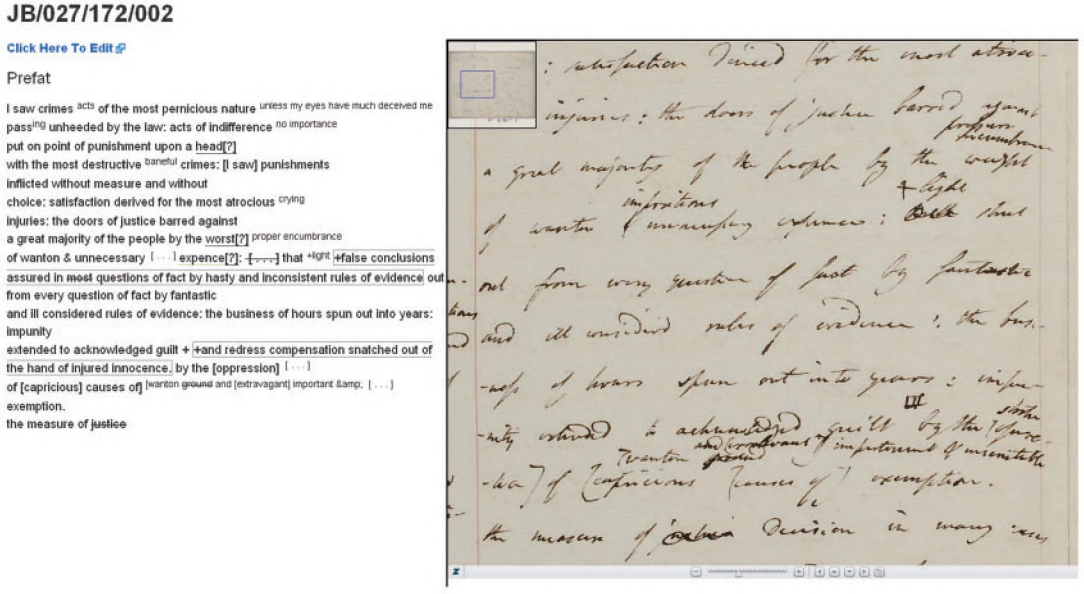

Fig. 3b Rendered version of the transcript of manuscript JB/027/172/002.

'manner' appear in superscript. However, neither of these methods identifies the meaning or function of the word 'manner'—instead, they provide a visual key for someone who is familiar with the transcription methods of the Bentham Project. By moving these transcriptions into TEI-compliant XML—a new method to the Bentham Project—the meaning of this authorial operation can be explicitly encoded into the text: *in this way* <add>*manner*</add> *it is possible.*[9]

The purpose of the Bentham transcripts that have already been completed (some 20,000 of about 60,000 folios) was to aid Bentham Project editors in the preparation of printed volumes for the *Collected Works*.[10] Up to 3,000 transcripts might be completed by editors in preparation for the production of one volume of the *Collected Works*. Most volumes are produced over a period of three years, sometimes more, and transcription occupies about one year of that time. When editing manuscripts in preparation for publication, editors have omitted some of the features of the text, including deleted text and additions. *Transcribe Bentham* thus intends to speed up production by crowdsourcing transcription, and will also retain the deleted features (which may be of interest to some scholars) in saved transcriptions, encoding them as such.

Transcripts submitted by volunteers, if considered to be of the requisite quality by project staff (see Section 3.2), are transferred from the Transcription Desk and easily converted into XML files using *oXygen* XML editor. They are subsequently uploaded to UCL's digital collection, which is a resource fully managed by UCL Library Services. Here, the full diplomatic transcriptions of Bentham manuscripts are made freely available to the public and to scholars anywhere in the world: these, along with the accompanying images, will be of interest to those keen on seeing the documents in their original form, before the intervention of the editor who creates the scholarly edition. In a dedicated 'Bentham Manuscripts' section of the repository, JPEG2000 manuscript images[11] are linked to corresponding TEI XML-encoded transcripts, and full metadata for the document and digital object. The TEI XML encoding will facilitate much more refined searching in the repository than a simple full-text search: for instance, a user may choose to see every instance in which Bentham deleted the word 'panopticon,' or search just within the text of the marginal notes.

The creation of this digital resource also guarantees the long-term curation and preservation of transcripts, as well as the manuscripts, beyond the end of the *Transcribe Bentham* initiative. Though the Bentham Project has assiduously backed-up and updated existing electronic transcripts to the most recent file formats on a shared network and offline, there is the risk that transcripts may be lost as formats become obsolete, or storage media degrades. Volunteers thus materially contribute to securing the collection's future and its accessibility.

Furthermore, if the Bentham Project should opt to publish the *Collected Works* digitally in the future, having transcripts readily available in a standardized and extensible format like TEI XML will preclude the necessity of large-scale document conversion. In the short term, the XML transcripts produced by *Transcribe Bentham* can be converted to *Word* by Bentham Project editors and saved to the Bentham template. Here, editors can amend the transcripts to produce texts in preparation for the typesetter. In future, editors may work solely with XML files and use an XSLT transformation to prepare from them a scholarly edition for print or digital publication, thus skipping steps currently required and rendering the editorial process more economical.[12] In the current climate, where the possibility of securing full funding for a further forty-one volumes seems increasingly remote, the ability to publish digitally becomes an important asset.[13]

3 Volunteer Contributions

3.1 Results

After a period of beta testing, the Transcription Desk went live to the public on 8 September 2010 with full staff support for six months until 8 March 2011.[14] During this six-month period, *Transcribe Bentham* attracted 1,207 registered users (excluding admin and project staff, and seven blocked spam accounts) who cumulatively transcribed 1,009 manuscripts, 569 (56%) of which were deemed to be complete and thus locked to prevent further editing. The Transcription Desk has been visited, according to Google Analytics reports, by people from ninety-one countries around the globe (Fig. 4); most visits to the Transcription Desk over the six months were from the USA, with Britain in second place.[15]

The publication of a feature article on *Transcribe Bentham* in the *New York Times*, online on 27 December and in print the following day (Cohen, 2010), had a significant impact upon *Transcribe*

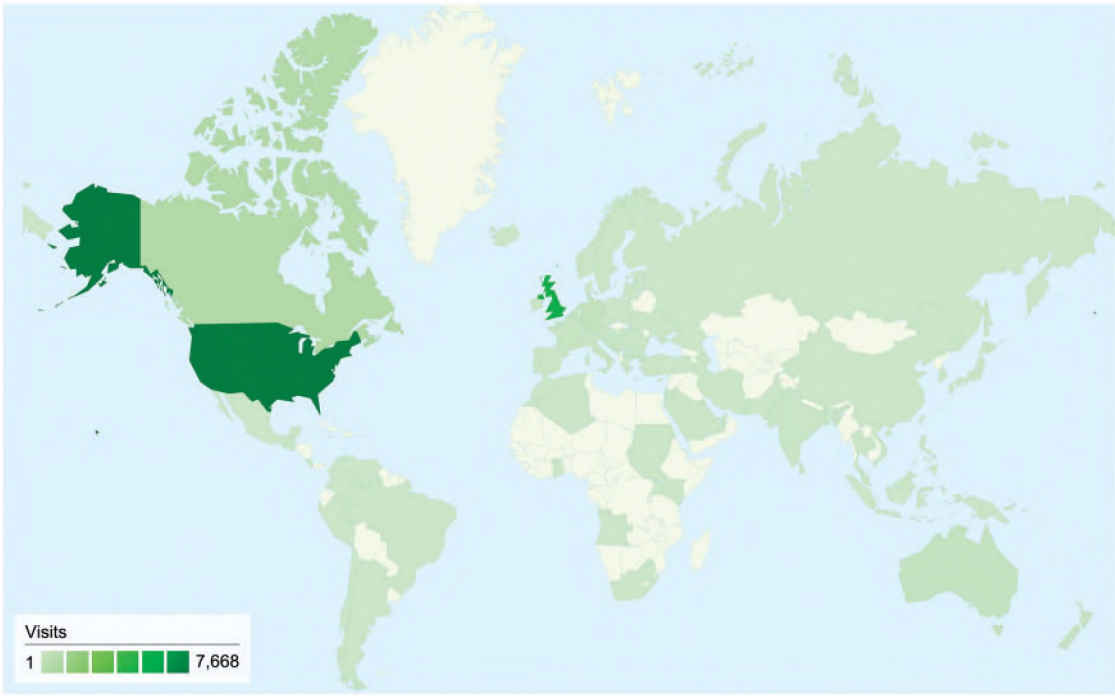

Fig. 4 Google Analytics map, showing global distribution of site visitors (8 Sep 2010 to 8 March 2010).

Bentham, and it is thus helpful to consider the testing period as being comprised of two parts. Period One, or the pre-*NYT* period, covers 8 September to 26 December 2010 (110 days); and Period Two, or the post-*NYT* period, covers 27 December 2010 to 8 March 2011 (72 days).[16] The Transcription Desk has remained available after 8 March, and will continue to be so for the foreseeable future, dependent on funding. As of 9 March 2012, there are 1,550 registered volunteers who have transcribed 2,975 manuscripts, of which 2,758 (92%) are complete. However, the findings below focus upon the six-month testing period.[17]

Of the 1,207 who registered an account between 8 September 2010 and 8 March 2011, only 259 (21%) did any transcription.[18] Almost two-thirds of those who transcribed, worked on a single manuscript, whereas over a quarter transcribed between two and five. Those who produced a significant number of transcripts were in a firm minority: fifteen volunteers edited between six and thirty manuscripts, another six regular users transcribed between sixty-three and eighty-two manuscripts each, and a further particularly dedicated volunteer transcribed 280, or 28% of all manuscripts then transcribed.[19] These seven most active volunteers—comprising a mere 0.6% of all registered users, or 2.7% of those who transcribed—produced between them a total of 709 transcripts by the end of the testing period, or 70% of all transcribed manuscripts.[20]

Period Two saw volunteer activity on the Transcription Desk increase dramatically, with 187 manuscripts transcribed between 27 December 2010 and 7 January 2011 alone, an increase of 43% on the end-of-Period One total. During Period Two, the seven most diligent volunteers transcribed large amounts of material on a regular basis, helping to produce an average of fifty-seven transcripts per week until 8 March.[21] Over the testing period as a whole, volunteers transcribed an average of thirty-five manuscripts each week;[22] if this rate were to be maintained, then 1,820 transcripts would be produced every twelve months. Taking Bentham's

Table 1 Number of manuscripts transcribed, 8 September 2010 to 8 March 2011

Month	Total manuscripts transcribed	Average no. of manuscripts transcribed per week
8 September to 24 September 2010	68	23
25 September to 29 October 2010	149	30
30 October to 26 November 2010	95	24
27 November to 23 December 2010	49	12
24 December 2010 to 28 January 2011	334	84
29 January to 25 February 2011	168	42
26 February to 8 March 2011	68	n/a

difficult handwriting, the complexity and length of the manuscripts, and the text-encoding into consideration, the volume of work carried out by Transcribe Bentham volunteers is quite remarkable (Table 1). Some manuscripts are only a few words long, whereas others are up to 2000 words in length, and we assume that the average manuscript is somewhere between 250 and 750 words long (plus mark-up). *Transcribe Bentham* volunteers thus transcribed somewhere between 250,000 and over 750,000 words (plus mark-up) during the testing period.[23]

To gain a greater insight into the volunteer base, we issued a user survey using *Opinio*, which was open to responses from 25 January to 24 March 2011. The survey consisted of a combination of multiple-choice questions and free-text boxes in which volunteers could record specific details and observations. We were keen to gauge the opinions of as wide a sample as possible, from regular transcribers to those who had looked at the site but not transcribed, and so the survey was open to all. It was advertised via the project blog, *Facebook* page and *Twitter* account, by a notice on the front page of the Transcription Desk, and by sending a message to each registered volunteer's user page. The survey received a total of 101 responses—or about 8% of those then registered with the project—of which 78 were fully completed. Although it is, therefore, not necessarily representative of the entire user base, the data collected was quite revealing.[24]

Transcribe Bentham volunteers were, according to our user survey, mostly motivated to take part by a specific interest in Bentham himself, general interest in philosophy and history, or an interest in crowdsourcing and the technology behind the project.[25] A significant number of survey respondents also reported that they were motivated by a sense of contributing to the greater good by contributing to the production of the *Collected Works* and making available Bentham's writings to others, whereas some even found transcribing fun. Competition and recognition were only secondary motivators. Significantly, no volunteer expressed any feeling of being exploited, and the regular feedback provided to volunteers on their work by project staff may have helped negate any feeling of exploitation.

That such a large proportion registered an account but did not transcribe suggests that certain factors dissuaded volunteers from taking part. The single most important factor in preventing or limiting participation was lack of time to dedicate to the initiative.[26] Others found themselves put off or intimidated by what seemed to many as extensive and complicated transcription instructions, the difficulty of reading Bentham's handwriting, and issues with the Transcription Desk such as being unable easily to identify untranscribed material. It seems that many found the task rather complicated. The majority of those who did transcribe had no prior experience of manuscript transcription, encoding, or reading archaic handwriting.[27] Feedback and a level of moderation were, therefore, important not only to maintain the pace and quality of transcription, but they were also a vital part of the general user experience, as we discovered when most regular transcribers ceased participating at the end of the fully-staffed testing period.[28]

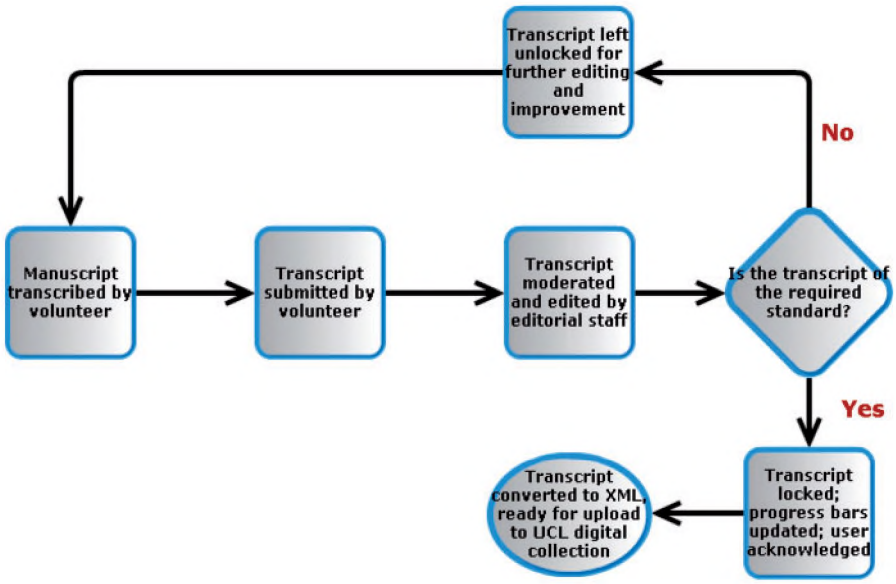

Fig. 5 Transcribe Bentham moderation workflow.

3.2 Quality control

Two Research Associates were employed full-time on *Transcribe Bentham* to coordinate the various aspects of the project, with one of their main priorities being the moderation of transcripts submitted by volunteers.

Figure 5 illustrates the transcription and moderation workflow, and the rigorous quality-checking process each transcript underwent. When a user completed a transcript and submitted it for moderation, it was checked for textual accuracy and encoding consistency using the anonymous 'TB_Editor' account. Changes would be made to the text and code, if necessary, and the moderator would then decide whether or not the transcript was completed to a satisfactory degree, as the transcripts needed to present as full a version of the text as possible for uploading to UCL's digital repository. If it was decided that no appreciable improvements could be made to the transcript through further crowdsourcing, then the transcript would be locked to prevent further editing, though the formatted and rendered transcript would remain present for viewing and searching. For example, a transcript would be deemed complete if there were few or no gaps, or if there were few or no unclear words in the transcribed text. Other crowdsourced transcription projects have demanding quality controls; as Romeo and Blaser (2011) note, transcripts submitted to the *Old Weather* project are checked three times before being accepted as finished. *Transcribe Bentham* also has rigorous quality control mechanisms. Transcripts uploaded to the digital repository will have been inspected by at least two people: one or more volunteer transcribers, and one member of staff. Those transcripts that are used for a *Collected Works* volume will be ultimately worked on by at least four people: one or more transcribers, the *Transcribe Bentham* moderator, a *Collected Works* editor, and the edition's General Editor. In the latter instance, this will be a bare minimum, as Bentham Project editors frequently consult each other regarding complex manuscripts.

Locked transcripts were then converted into TEI-compliant XML, and stored on a shared drive ready for uploading to UCL's digital repository. We were aware that locking transcripts creates some tension vis-à-vis the open-access and collaborative principles of *Transcribe Bentham* and MediaWiki,

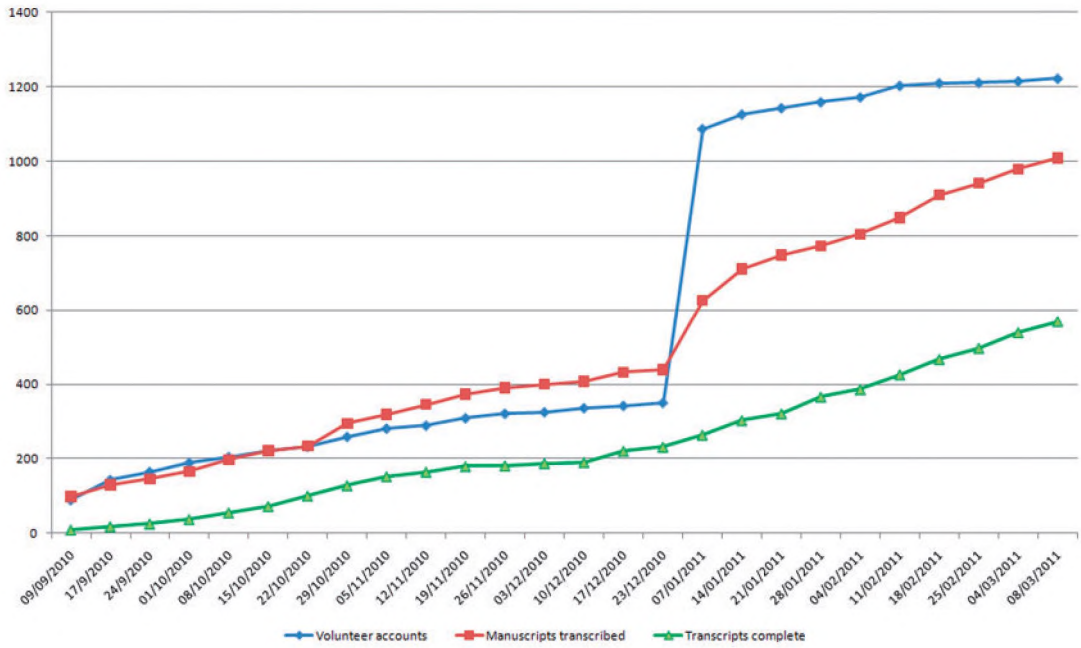

Fig. 6 User accounts registered, manuscripts transcribed, and manuscripts locked (8 September 2010 to 8 March 2011).

but some cut-off point is required to signal the decision that a transcript is as good as it is possibly going to be, and to provide the volunteers with a sense of achievement. If the moderator decided that a submitted transcript was incomplete and could be improved by further editing from volunteers, then it would remain available for editing on the Transcription Desk. A link would be added to an 'Incomplete Folios' page, which displayed a list of manuscripts upon which transcription had begun, but which required further work.

A transcript was generally left unlocked if, after moderation, it contained several gaps and/or unclear words, or if the manuscript had only been partially transcribed. Whether a submission was locked or not, volunteers were informed of the outcome by a standard TB_Editor message left on their user pages,[29] which also acted as an acknowledgement of their work. Further feedback might be provided, either via user pages or email, and was often requested by volunteers seeking further assistance or clarification of edits made to their transcripts. Each submission also necessitated the updating of the progress bars, and the manuscript identifier had to be added to the incomplete or complete transcripts pages. These tasks have yet to be automated and are extremely laborious, particularly when there have been a high number of submissions.

The moderation workload and level of Transcription Desk maintenance varied, as can be seen in Fig. 6. From 8 September 2010 to 26 December 2010 (Period One), an average of twenty-three manuscripts was transcribed each week. In Period Two, this rose to a weekly average of fifty-seven manuscripts, though this is slightly distorted by the extremely high level of activity in the three weeks immediately after the *NYT* article was published on 27 and 28 December. As a result, over the December break, almost 200 manuscripts were transcribed; it took around 10 days to catch up fully, and in the meantime, more were submitted.

Over the six-month period, 56% of submitted transcripts were deemed 'lockable' and ready for uploading to the repository. Despite users being on the whole inexperienced encoders and transcribers, the standard of submissions was, in general,

surprisingly high, though every single submission did require some level of moderation.[30] We estimate that during the testing period, the Research Associates each spent the equivalent of a month's full-time labour on moderating submissions and the associated upkeep of the Transcription Desk. If they had spent this time transcribing manuscripts, the Associates could have produced around 400 transcripts between them.[31]

Though moderation is evidently a time-consuming process, it was—and remains—an indispensable part of *Transcribe Bentham*. Many volunteers found the transcription process complicated, and as users sought to improve the quality of their work and their skills, feedback about submissions, transcribing, and encoding practices, was greatly appreciated. Likewise, the messages left on volunteers' user pages acted as acknowledgement that their work was valued. Indeed, it is clear that volunteers' enthusiasm can and will fall away when feedback and acknowledgement are not given, or not given in a timely fashion.[32] A great number of volunteers signed up with the project and transcribed material over the December break following publication of the *NYT* article, when staff were away from the office and many submissions were not checked and acknowledged for a week or two; we fear we may have missed out on recruiting several regular volunteers by being unable to issue prompt feedback. This suggests that project staff must devote some time to moderation and quality control; otherwise, users may lose interest, and feel undervalued, or exploited.

However, with an improved transcription tool—the introduction, for example, of a What-You-See-Is-What-You-Get (WYSIWYG) interface—and accessible online training in deciphering handwriting, quality control may become less of a priority. *Transcribe Bentham*'s testing period was too short for users to gain enough experience and feel confident enough to form a self-moderating community.[33] If *Transcribe Bentham* had been able to continue with full support for a longer period, eventually staff time could have been utilized in other ways. Indeed, transcribing might one day be omitted completely from the editing process, and the role of professional transcriber rendered obsolete, if volunteers are able in future to transcribe accurately with minimal moderation. Rather than transcribing, or moderating transcription, Bentham Project editors will be able to devote their time to other tasks and the editorial process will become more cost-effective, though providing staff support for *Transcribe Bentham*, or crowdsourced manuscript transcription more generally, can never be fully dispensed with.

4 Success?

Transcribe Bentham produced 1,009 transcripts during its six-month testing period, of which 56% were considered of high enough quality by moderators to have been locked, converted to XML files, and stored for uploading to UCL Library's digital Bentham repository. It is estimated that between 15,000 and 20,000 manuscript images will be photographed by the time the money from the AHRC grant assigned to this purpose is exhausted.[34] In the 2011 Prix Ars Electronica, *Transcribe Bentham* was honoured with an Award of Distinction, which carries a prize of €5,000 which will be spent on photography and, therefore, an additional 2,000 or more images will be produced.[35]As of 9 March 2012, 6,009 images have been uploaded to the Transcription Desk, and 2,102 of the 20,000 or so already existing transcripts produced by Bentham Project staff during the last 25 years have been encoded in TEI-compliant XML, ready for linking to manuscript images in the digital repository.

But has the project been a success? In terms of quickening the pace of transcription, perhaps not, or at least not thus far. The number of transcripts produced seems, on the surface, to compare unfavourably with the hundred million galaxy classifications produced by citizen scientists taking part in the *Galaxy Zoo* project, or the 36.5 million lines of OCR text corrected by volunteers involved in the National Library of Australia's historic newspaper digitization programme (Raddick *et al.*, 2010). Furthermore, we estimate that had the Research Associates been employed purely to transcribe manuscripts on a full-time basis, they could have transcribed about 5,000 manuscripts between them

over twelve months,[36] or two-and-a-half times as many as the volunteers would have produced had they continued transcribing at the same rate. Without having to invest in digitization or programming, the AHRC grant could have employed both Research Associates for three years, allowing almost half of the remaining 40,000 UCL Bentham Papers to be transcribed. Instead, they spent the equivalent of a month's full-time labour moderating 1,009 submissions, with the rest of their time spent in development and testing of the interface, volunteer recruitment, publicity, maintenance, the conversion of legacy transcripts, and other editorial tasks. On the face of it, this makes the products of *Transcribe Bentham* seem rather insignificant, and the project not particularly cost-effective, at least in the short-term. However, this point is somewhat moot for one main reason: no funding body would ever provide a grant for mere transcription alone.

Moreover, each manuscript which *Transcribe Bentham* volunteers are asked to transcribe can be up to 2,000 words long, and even short ones may be complicated to a great extent by deletions, marginalia, interlineal additions, and other features. Bentham's handwriting and style can often seem impenetrable, and text encoding is daunting to many beginners. In this sense, and given that the task is much more onerous and demanding than in most other crowdsourcing projects, the amount of work completed by *Transcribe Bentham* volunteers is quite remarkable.

By making a few improvements to the transcription tool—such as introducing a WYSIWYG transcription interface as an alternative, so that volunteers see their transcript rendered as they type with the encoding remaining 'under the hood'—we are confident that volunteer recruitment and retention will be increased and transcripts will be produced at a much quicker rate, thereby increasing the ongoing cost-effectiveness of the project. In total, 55% of respondents to the volunteer survey agreed that a WYSIWYG interface would be beneficial to transcribing, particularly for new participants. Although a handful of respondents expressed satisfaction with the TEI toolbar, one volunteer—perhaps speaking for many others who abandoned the project without transcribing—noted that he had 'given up because of the encoding'. Looking at *Transcribe Bentham* in the long-term, the initial cost of establishing the project could be offset by future savings. For instance, were transcripts of all relevant manuscripts available for a particular *Collected Works* volume, we estimate that *Transcribe Bentham* could save between six months and one year of an editor's time.

Crowdsourced manuscript transcription might be made more cost-effective by doing away with editorial moderation altogether, thereby negating the need to pay academics to run the site. However, in *Transcribe Bentham*, which deals with a complex set of manuscripts as part of a wider academic editorial project, there is a clear need for some editorial expertise in order to moderate submissions and to advise volunteers on their work; this was particularly the case during the project's infancy. Indeed, quality control is an integral part of *Transcribe Bentham*, fulfilling the volunteer requirement for feedback, encouragement, and feeling of connection and contribution to the wider activities of the Bentham Project. If moderation was removed entirely, viable transcripts might still be produced, but their standard might be poorer and the rate of production far slower, and the process would be explicitly exploitative.

Indeed, in order to avoid a sense of exploitation, a crowdsourcing venture like *Transcribe Bentham* must be based on mutual respect and trust: volunteers are provided with material, instructions, and a transcription tool, and trusted accurately to transcribe; in return, project staff must treat volunteers and their work with respect, respond to enquiries promptly, keep them informed of project news and progress, and not waste their time. It is thus impossible to imagine *Transcribe Bentham* ever doing away with the moderation process, but with a larger and more cohesive user community, it may be possible to encourage certain experienced volunteers to take on limited editorial duties and community supervisory roles, such as dealing with queries that new volunteers might have about the transcription process. In this scenario, expenditure on moderating the Transcription Desk might be scaled back—to a point—with volunteer–moderators correcting submissions; these submissions

would then only require a brief checking over by editorial staff before being locked, thereby saving an enormous amount of time. With moderation being partly carried out by volunteers in this way, transcription might in future be a very minor part of the Bentham Project's editorial practice, allowing editors to focus on other aspects of the publication process.

As others have noted, crowdsourcing, particularly crowdsourcing manuscript transcription, is plainly not necessarily a cheap, quick, or easy solution (Anderson, 2011; Parry, 2011; Zou, 2011). Editors of the *Papers of Abraham Lincoln* found that, when experimenting with using non-academic transcribers, they spent more time correcting errors than they would have had they carried out the transcription themselves. Staff at the University of Iowa's *Civil War Diaries Transcription Project* had a similar experience (Cohen, 2010; Zou, 2011). As *Transcribe Bentham* necessitated building the tool from scratch, significant investment was required in programming, infrastructure, and digitization; the first half of the twelve-month grant was spent developing, testing, and publicizing the transcription tool, and the project in general. We were very fortunate that the metadata for UCL's Bentham collection compiled for the *Bentham Papers Database Catalogue* already existed; the collation of metadata would have required significant further investment.[37] The recent proliferation of open-source transcription tools, will at least allow new projects to reduce the time and money spent on development.

Within a short space of time, the *Transcribe Bentham* project team worked closely and successfully to produce the Transcription Desk. However, it was available to the public with full support for just a six-month period, and only began to produce its most impressive results during the final three months after the publication of the *New York Times* article. We had gone some way to recruiting a core group of volunteers, seven regulars who produced more than 70% of all transcripts and were transcribing at an impressive rate by the time the testing period came to an end. Out of respect for those who had so generously given their time and effort to the project, we felt duty-bound to inform volunteers that we would have to end full-time staffing of the Transcription Desk after 8 March 2011. Unsurprisingly, most regulars ceased transcribing.[38]

There is a clear danger here. If academic crowdsourcing projects are funded only in the short-term or have no post-grant sustainability plan, thereby precluding the establishment and recruitment of a wide enough pool of volunteers to ensure the project's self-sufficiency, or the refinement of their websites and software in response to user feedback, potential volunteers will become increasingly suspicious. What motivation would anyone have for taking part in a project with a long-term goal, if they suspect the project might cease six months down the line due to a grant ending? This would amount to a betrayal of trust, and could kill a project attempting to establish itself.

In an ideal world, large-scale manuscript crowdsourcing projects such as *Transcribe Bentham* would have sufficient funding for at least twelve months—and ideally fifteen to eighteen months with full staff support. A period of six months is simply not enough—unless the project crowdsources material that is of more general interest than the Bentham Papers, or receives prominent, international media coverage from the beginning[39]—fully to recruit and establish a wide group of regular volunteers. Had we had twelve months in which to staff the project full-time, we could have added to the regulars—assuming the existing ones remained—and identified those who could act as volunteer moderators. We have been fortunate that UCL has recognized the importance of *Transcribe Bentham*, and provided some funding for limited staffing of the Transcription Desk for a further twelve months while we explore further opportunities.

Transcribe Bentham's success, however, should not be measured solely according to the number of transcripts it has produced. In terms of sustainability and public engagement, *Transcribe Bentham* has made a significant impact, of which we are very proud and which is of vital importance for bodies funding academic projects. The project has resulted in significant publicity for Bentham studies, history, and philosophy more generally, and for crowdsourcing. In his round-up of crowdsourced transcription during 2010, Ben Brumfield (2011) noted that

Transcribe Bentham has done 'more than any other transcription tool to publicize the field'. *Transcribe Bentham* has, to date, engaged people from ninety-one countries round the globe, has been mentioned in dozens of blogs, two radio broadcasts,[40] and eleven press articles, and has been utilized in teaching and learning at Queen's University Belfast, Bloomsburg University, the University of Virginia, and King's College London.[41] The code for the transcription tool is freely available for download[42] and *Transcribe Bentham*, and our outreach activities, are being used as a model for other similar initiatives in Europe and North America.[43] The project was shortlisted for the 2011 Digital Heritage Award, and received an Award of Distinction in the 2011 Prix Ars Electronica;[44] this is further testament to *Transcribe Bentham*'s international impact. The jury commended the project for its 'potential to become a standard tool for scholarly crowdsourcing projects', and its 'potential to create the legacy of participatory education and the preservation of heritage or an endangered culture' (Achaleke *et al.*, 2011, p. 206).

Most importantly perhaps, the creation of a freely available digital collection of the Bentham Papers to complement the printed texts, will widen access to, and encourage user participation with, Bentham's writings, while ensuring the long-term preservation of the manuscripts. User-generated transcripts—currently being produced at an average rate of thirty-five per week—will be added progressively to the digital repository; the collection is thus being further enhanced as time goes on. This is something of which Bentham would have approved. Indeed, Bentham requested that his manuscripts be publicly displayed after his death alongside his Auto-Icon: 'conceive', he said, 'the old philosopher preserved in some safe repository [...] accompaniments of it, his unedited and unfinished manuscripts, lodged in an appropriate case of shelves [...] In this far-famed receptacle, there would be no want of matter of wonder and admiration' (Bentham, 1842, p. 15). Thanks to the efforts of volunteer transcribers, *Transcribe Bentham* is helping to honour Bentham's last wish.[45]

With adequate funding to support development costs, and enough time to mature, a crowdsourced manuscript transcription initiative like *Transcribe Bentham* could be enormously beneficial to a scholarly editorial project like the Bentham Project. An improved transcription tool would relieve volunteers from being overly-concerned with encoding and allow them to concentrate upon deciphering Bentham's manuscripts, and result in the submission of a greater number of transcripts at a faster pace. The speedy production of high-quality transcripts would then quicken the pace of editing and publishing the anticipated forty-one printed volumes of Bentham's works yet to appear. Our results suggest that a longer time-scale would allow the community of transcribers to develop and become more self-sufficient, requiring less feedback and quality control from staff, thereby rendering the project more cost-effective. Nevertheless, despite the difficulties involved in transcribing Bentham's manuscripts, and despite the short time-frame in which the tool was developed, publicized, and made available to the public, *Transcribe Bentham* has engaged a wide range of people and produced a significant number of usable transcripts (and continues to do so). This underlines the great potential of crowdsourcing manuscript transcription: if untrained volunteers are able to transcribe the papers of Jeremy Bentham, some of which border on the illegible, they can transcribe almost anything.

Funding

This work was supported by the Arts and Humanities Research Council's Digital Equipment and Database Enhancement for Impact (DEDEFI) scheme, under which *Transcribe Bentham* was established.

Acknowledgements

The authors would like to thank our *Transcribe Bentham* colleagues: Philip Schofield, Richard Davis and his team, Martin Moyle, Tony Slade, and Melissa Terras; and all of the volunteer transcribers who have given their time and effort to the project, and who continue to contribute to *Transcribe Bentham's* ongoing success. We are also

very grateful to Oliver Harris, Michael Quinn, Philip Schofield, Melissa Terras, and Catherine Pease-Watkin for reading earlier drafts of this article, and offering many very helpful comments. Thanks also to our two referees, whose comments and suggestions have also been extremely helpful in finalizing the article.

References

Achaleke, B., Harwood, G., Koblin, A., Yan, L., and Peixoto, T. (2011). Guinea pigs and apples: statement of the Digital Communities Jury. In Leopoldseder, H., Schöpf, C., and Stocker, G. (eds), *Prix Ars Electronica International Compendium: CyberArts 2011*. Ostfildern: Hatje Cantz, pp. 202–07 and 210–11.

Ancestry.com's World Archives Project, http://community.ancestry.com/wap/download.aspx (accessed 30 March 2011).

Anderson, K. (2011). Even crowdsourcing can get too expensive. *Scholarly Kitchen* http://scholarlykitchen.sspnet.org/2011/03/14/even-crowdsourcing-can-get-too-expensive/ (accessed 16 March 2011).

Ars Electronica: Winners of the 2011 Prix Ars Electronica, http://new.aec.at/prix/en/gewinner/2011/ (accessed 26 May 2011).

Bentham, J. (1842). *Auto-Icon; or, Farther Uses of the Dead to the Living*, unpublished. Reprinted in Crimmins, J. E. (2002), *Jeremy Bentham's Auto-Icon and Related Writings*. Bristol: Thoemmes Continuum.

Bentham Papers Database Catalogue, http://www.benthampapers.ucl.ac.uk/ (accessed 15 November 2010).

Bentham Project: Jeremy Bentham's Auto-Icon, http://www.ucl.ac.uk/Bentham-Project/who/autoicon (accessed 15 September 2011).

Bloomsburg University Digital Humanities 2010 Honours Class Blog, http://stephanieschlitz.com/dh/2010/ (accessed 20 March 2011).

Brumfield, B. (2011). The year of crowdsourcing transcription. *Collaborative Manuscript Transcription* http://manuscripttranscription.blogspot.com/2011/02/2010-year-of-crowdsourcing.html (accessed 4 February 2011).

Causer, T. and Wallace, V. (forthcoming, 2012). Building a Volunteer Community: Results and Findings from *Transcribe Bentham*. *Digital Humanities Quarterly*.

Civil War Diaries Transcription Project, http://digital.lib.uiowa.edu/cwd/transcripts.html (accessed 20 April 2011).

Cohen, P. (2010). Scholars recruit public for project. *New York Times* published 27 December 2010: http://www.nytimes.com/2010/12/28/books/28transcribe.html?pagewanted=all (accessed 28 December 2011).

Dickens Journals Online, http://www.djo.org.uk/ (accessed 1 August 2011).

Digital Strategies for Heritage Conference 2011: List of Nominees for 2011 Digital Heritage Award, http://www.dish2011.nl/news/digital-heritage-award-2011-nominees (accessed 12 January 2012).

National Library of Australia: Behind the Scenes Blog, http://blogs.nla.gov.au/behind-the-scenes/ (accessed 15 May 2011).

National Library of Australia, Historic Newspaper Digitisation Project, http://trove.nla.gov.au/newspaper (accessed 30 March 2011).

Old Weather, http://www.oldweather.org/ (accessed 30 March 2011).

Papers of the War Department, 1784 to 1800, http://wardepartmentpapers.org/index.php (accessed 17 March 2011).

Parry, M. (2011). Facing budget woes, prominent crowdsourcing project will scale back. *The Chronicle of Higher Education* published 11 March 2011: http://chronicle.com/blogs/wiredcampus/facing-budget-woes-prominent-crowdsourcing-project-will-scale-back/30322 (accessed 11 March 2011).

Praxis Programme, Scholars' Lab at the University of Virginia Library, http://praxis.scholarslab.org/topics/evaluating-digital-work/ (accessed 15 September 2011).

Raddick, M. J., Szalay, A. S., Vandenberg, J. *et al.* (2010). Galaxy Zoo: exploring the motivations of citizen science volunteers. *Astronomy Education Review*, **9**, http://dx.doi.org/10.3847/AER2009036 (accessed 20 April 2011).

Romeo, F. and Blaser, L. (2011). Bringing citizen scientists and historians together. In Trant, J. and Bearman, D. (eds), *Museums and the Web 2011: Proceedings*. Toronto: Archives and Museum Informatics.

Schofield, P. (2009). *Bentham: A Guide for the Perplexed*. London: Continuum Books.

Scripto Transcription Tool, http://scripto.org/ (accessed 17 March 2011).

The Frederick Law Olmsted Papers Project, http://www.olmsted.org/flo (accessed 20 August 2011).

Galaxy Zoo, http://www.galaxyzoo.org/ (accessed 30 March 2011).

T-Pen Transcription Tool Blog, http://digital-editor.blogspot.com/ (accessed 2 November 2011).

Transcribe Bentham: Project Blog, http://www.ucl.ac.uk/transcribe-bentham (accessed 30 March 2011).

Transcribe Bentham: Transcription Desk, http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham (accessed 30 March 2011).

Transcribe Bentham: Basic Transcription Guidelines, http://www.transcribe-bentham.da.ulcc.ac.uk/td/Getting_Started (accessed 30 March 2011).

Transcribe Bentham: Benthamometer, http://www.transcribe-bentham.da.ulcc.ac.uk/td/Benthamometer (accessed 30 March 2011).

Transcribe Bentham: Full Transcription Guidelines, http://www.transcribe-bentham.da.ulcc.ac.uk/td/Help:Transcription_Input_Form (accessed 30 March 2011).

Transcribe Bentham: Incomplete Folios Category, http://www.transcribe-bentham.da.ulcc.ac.uk/td/Category:Incomplete_Folios (accessed 30 March 2011).

Transcribe Bentham: MediaWiki Plugins, Code and Instructions, http://code.google.com/p/tb-transcription-desk/ (accessed 9 July 2011).

UCL Library Services: Bentham Digital Collection, http://www.ucl.ac.uk/library/bentham (accessed 15 September 2011).

University of Virginia etext collection, http://search.lib.virginia.edu/catalog?f[source_facet][]=Digital+Library (accessed 14 December 2011).

Zooniverse Community, http://www.zooniverse.org/ (accessed 30 March 2011).

Zou, J. J. (2011). Civil war project shows pros and cons of crowdsourcing. *The Chronicle of Higher Education*, published 14 June 2011: http://chronicle.com/blogs/wiredcampus/civil-war-project-shows-pros-and-cons-of-crowdsourcing/31749 (accessed 14 June 2011).

Notes

1 One volume is currently in press, with another dozen at advanced stages of preparation.

2 The Bowring edition contains most of the works that Bentham published during his lifetime (except those on religion); texts edited by Bentham's 'disciples' during his lifetime, including translations of recensions of Bentham's work in French by Étienne Dumont; and a number of unpublished works from the manuscripts. For more information, see Schofield (2009, p. 14–15, 20–22).

3 Extensible Mark-up Language (XML) is a set of rules for encoding machine-readable texts. The Text Encoding Initiative (TEI) is a consortium that publishes a widely used set of guidelines for encoding machine-readable texts in the Humanities: it provides schemas for encoding Humanities texts within the structure and confines of XML.

4 Owing to staffing problems in the Department of Information Studies, the log analysis was not completed.

5 Images were captured as 16.7 m pixel Canon CR2 files, and processed as 50mb High Resolution TIFF files. These were converted to 4096px JPEGs for uploading to the Transcription Desk, and JPEG2000 format for archiving in UCL Library Service's digital Bentham repository.

6 On other occasions, the image photographed is a larger sheet of paper that has been folded by Bentham to create multiple 'pages' (see, for example, http://www.transcribe-bentham.da.ulcc.ac.uk/td/JB/096/116/001), or a single sheet divided into columns and containing Bentham's marginal summaries.

7 Cascading Style Sheets: a style-sheet language used to describe the appearance and formatting of a document written in a mark-up language such as XML or HTML.

8 For instance, text marked as deleted was rendered with a strikethrough, questionable readings were followed by a question mark in parentheses, and notes were surrounded by a box.

9 While the TEI does provide the means for distinguishing between 'added' and 'alternative' text, (e.g. *<add type="alternative">manner</add>*), in order to keep the transcription text-box uncluttered, we refrained from adding attributes and attribute values, where possible.

10 For more information on editing practices at the Bentham Project, see Schofield (2009, ch. 2).

11 See endnote 5.

12 Fully completed and checked transcripts are made available through the UCL digital repository. These will then be used by editors to produce the print volumes, though no decision has been made as to whether the *Collected Works* volumes will be made available via the web. Unlike other projects, such as the University of Virginia's e-text collection (1992–2007), we will not automatically create e-books directly from transcripts, as these will always require collating, editing, and working into a coherent scholarly text.

13 The National Historical Publications and Records Commission (NHPRC) in the USA has recently

declared its priority of supporting editorial projects that have plans to publish their results as a free online resource, and the support of the NHPRC is crucial for the feasibility of multi-volume documentary editions. For example, the twelve-volume *Papers of Frederick Law Olmsted* editorial project has recently changed its 2011 proposal to the NHPRC in order to meet the new guidelines. The project has until now been a print-only edition, but has made plans to publish past volumes free and open to the public via its new website: http://www.olmsted.org/flo.

14 The call for the AHRC's DEDEFI scheme was issued in September 2009, and the results were announced in February 2012. This was a 'one-off' call, with a maximum duration of one year, and specifically intended to build on projects that had been initially funded by the AHRC's Resource Enhancement Scheme (which itself is now closed). Interviews for the Research Associate positions were held in March 2010. The grant expired at the end of April 2011, and the last month was allocated to compiling and writing up the project's findings.

15 Visits from *Transcribe Bentham* staff offices were discounted when analysing site statistics.

16 Whereas the *NYT* article had a significant impact upon *Transcribe Bentham*, it has not skewed our long-term projections for the rate of transcription that are based upon an assessment of volunteer activity during the eleven months since the article's publication.

17 These results are compiled from a Google Analytics account, analysis of the Transcription Desk statistics, and qualitative information from a user survey. More detailed findings about *Transcribe Bentham* volunteers and their motivations can be found in Causer and Wallace (2012).

18 Though it is possible to view manuscripts and search transcripts without a volunteer account, it is necessary to register in order to see and access the transcription interface. The remaining 79% who registered, but did no transcription, were at least curious enough about the project to sign up.

19 As of 9 March 2012, the most prolific volunteer has transcribed 976 manuscripts (32% of the current total), and the next most regular has produced 469 transcripts (15%), despite having only begun participating in mid-September 2011. Another volunteer has transcribed 454 manuscripts (15%).

20 For detailed information on the backgrounds of *Transcribe Bentham* volunteers, see Causer and Wallace (2012).

21 This mean average for Period Two is distorted somewhat by the high rate of transcription between 27 December 2010 and 7 January 2011. The median average for Period Two is 43.

22 The median average for the testing period was 27, and the modal average was 32.

23 As of 9 March 2012, on the same calculation, they have transcribed somewhere between 750,000 and 2,230,000 words, plus mark-up. It is thus estimated that the 40,000 folios of material which were untranscribed at the start of the project contain between 10 million and 30 million words.

24 Further research may involve interviewing the most prolific volunteers. A full list of the questions asked in the survey is available upon request from the authors.

25 There were forty-three responses to a question asking 'If you consider yourself to be a regular user of Transcribe Bentham, what motivates you to keep returning to transcribe material?'

26 There were thirty-four responses to a question asking 'If you do not consider yourself a regular user of Transcribe Bentham or you have signed up but not transcribed, could you let us know what has put you off or dissuaded you from transcribing?'
We also asked 'How long, on average, do you think you spend transcribing a single folio?' This question is rather impressionistic, though of forty-three responses, less than 40% said 'Up to 1 hour', and just under a quarter said 'More than 2 hours'.

27 There were eighty-five respondents to a question asking 'Have you had any palaeography training, or have you transcribed manuscripts before taking part in Transcribe Bentham?'. Twenty-eight said 'yes', fifty-seven said 'no'. There were eighty-four respondents to a question asking 'Were you familiar with the principles of text encoding prior to taking part in Transcribe Bentham'. Thirty-five said 'yes', forty-nine said 'no'.

28 Volunteer feedback is covered in more detail in Causer and Wallace (2012).

29 For example, see http://www.transcribe-bentham.da.ulcc.ac.uk/td/User_profile:TimCauser.

30 As of 9 March 2012, in the majority of instances, submitted transcripts are of such quality that they require little intervention from a moderator before being locked. No transcript has ever been left unlocked purely because of errors in text encoding.

31 On the assumption of producing an average of ten transcripts per day.

32 After 8 March 2011, the end of the testing period and cessation of full-time staffing, we expected the number

of submissions to decrease and some volunteers to cease participating. While the latter certainly did occur, the former did not happen to the extent we feared. From 9 March 2011 to 9 March 2012 overall, an average of thirty-eight manuscripts were transcribed each week, compared with an average of thirty-six per week during the testing period. The rate of transcription has increased noticeably from the period ending 16 September to 9 March 2012, since when an average of forty-six manuscripts have been transcribed each week. This increase is especially pronounced during the period 3 February 2012 to 9 March 2012, when the weekly rate of manuscripts transcribed was 63.

33 Bentham Project staff are currently exploring further avenues of funding, in order to complete the digitization of the collection and continue to crowdsource the material.

34 As photography is completed in-house, the costs were reduced from £3 per image to £2. Thus, more images will be produced for the same sum.

35 The Prix Ars Electronica is the world's premier digital arts competition. Transcribe Bentham was entered into the 'Digital Communities' category, which received 407 entries, and the Award of Distinction constitutes joint second place. For more details see http://new.aec.at/prix/en/gewinner/2011/, and http://new.aec.at/press/files/2011/05/Prix-Ars-Electronica-2011_EN.pdf. Photographs of the event are available (http://www.flickr.com/photos/arselectronica), as well as a video recording of the Digital Communities winners' forum (http://www.youtube.com/watch?v=vvD3ipgZCTQ).

36 At an average rate of ten transcripts per day.

37 The Catalogue was compiled by Dr Deborah Colville (née McVea) between 2003 and 2006. See http://www.benthampapers.ucl.ac.uk/. Without this database, *Transcribe Bentham* would not have been possible, or would have required further extensive investment in time and money to establish.

38 Despite losing all but three of the regular transcribers after the end of the testing period, as of 9 March 2012, *Transcribe Bentham* has ten volunteers producing transcripts on a regular basis, and at a rate of almost double that at the end of the testing period.

39 For example, the *Dickens Journals Online* project received national and international media attention, and crowdsources OCR-generated text of Dickens's journals *Household Words* and *All the Year Round*, material which is of far wider popular appeal than Bentham's manuscripts.

40 Austrian ORF radio's 'Matrix' programme, broadcast on 29 January 2012 (http://oe1.orf.at/programm/294290), investigated the state of the digital humanities, and featured an interview with *Transcribe Bentham* staff.

41 The Honours Class in Digital Humanities at Bloomsburg University concentrated on crowdsourcing and *Transcribe Bentham* during one week of the course. The class blog features responses from the students; see, for example, http://stephanieschlitz.com/dh/2010/11/07/whatam-i-learning-in-dh/ and http://stephanieschlitz.com/dh/2010/10/29/thebeauty-of-transcribing/. Searching the site for 'Bentham' will bring up more responses. See also the 2011–2012 Praxis Programme in the Scholars' Lab at the University of Virginia Library, in which various digital humanities resources are assessed: http://praxis.scholarslab.org/topics/evaluating-digital-work/. Students do not appear to be required to transcribe material in these examples, but critically assess the methods the project team have used. For more information on the use of *Transcribe Bentham* in teaching, see Causer and Wallace (2012).

42 For the code for the MediaWiki plugins, please see: http://code.google.com/p/tb-transcription-desk/.

43 *Transcribe Bentham* staff have been contacted for consultation by, amongst others, the University of Southampton, the University of Exeter, the London School of Economics, the Natural History Museum, the National Library of the Netherlands, Library and Archives Canada, the Royal Library in Denmark, San Diego Natural History Museum, the University of Texas, the J. Paul Getty Research Institution, and the Church of Jesus Christ of Latter-Day Saints.

44 See http://www.dish2011.nl/news/digital-heritage-award-2011-nominees.

45 The Auto-Icon display in UCL's South Cloisters was recently updated to include a new interpretation panel and a touch-screen computer, through which visitors can explore aspects of Bentham's life and thought, and view *Transcribe Bentham* and the digital Bentham Papers collection. A 'Virtual Auto-Icon' is now available at: http://www.ucl.ac.uk/Bentham-Project/who/autoicon/Virtual_Auto_Icon.