tranScriptorium

# D6.1:User Needs

Kris Grint, UCL

Distribution: Public

**tranScriptorium**
ICT Project 600707 Deliverable 6.1

June 30, 2013

Project funded by the European Community
under the Seventh Framework Programme for
Research and Technological Development

| | |
|---|---|
| Project ref no. | ICT-600707 |
| Project acronym | **tranScriptorium** |
| Project full title | **tranScriptorium** |
| Instrument | STREP |
| Thematic Priority | ICT-2011.8.2 ICT for access to cultural resources |
| Start date / duration | 01 January 2013 / 36 Months |

| | |
|---|---|
| Distribution | Public |
| Contractual date of delivery | June 30, 2013 |
| Actual date of delivery | June 30, 2013 |
| Date of last update | June 8, 2013 |
| Deliverable number | 6.1 |
| Deliverable title | User needs |
| Type | Report |
| Status & version | Final |
| Number of pages | 33 |
| Contributing WP(s) | 6 |
| WP / Task responsible | UCL |
| Other contributors | UIBK, ULCC |
| Internal reviewer | Joan Andreu Sánchez, Katrien Depuydt |
| Author(s) | Kris Grint, Guenter Muehlberger |
| EC project officer | Jose Maria del Aguila |
| Keywords | transcription, crowdsourcing, motivation, volunteer |

The partners in **tranScriptorium** are:

Universitat Politècnica de València - UPVLC (Spain)
University of Innsbruck - UIBK (Austria)
National Center for Scientific Research "Demokritos" - NCSR (Greece)
University College London - UCL (UK)
Institute for Dutch Lexicology - INL (Netherlands)
University London Computer Centre - ULCC (UK)

For copies of reports, updates on project activities and other **tranScriptorium** related information, contact:

The **tranScriptorium** Project Co-ordinator
Joan Andreu Sánchez,
Universitat Politècnica de València
Camí de Vera s/n. 46022 València, Spain
jandreu@dsic.upv.es
Phone (34) 96 387 7358 - (34) 699 348 523

Copies of reports and other material can also be accessed via the project's homepage: http://www.transcriptorium.eu/

## Executive Summary

The purpose of this report is to evaluate firstly the user needs of potential "crowdsourced volunteer contributors" (CVCs) who will participate in the Handwritten Text Recognition (HTR) technology manuscript transcription platform outlined in T6.2 of the *tranScriptorium* (*tS*) STREP proposal and secondly for archives/libraries who will make use of the technology in a "content provider platform" (CPP).

The evaluation of the first group of users is based on a survey of existing CVCs of a current manuscript transcription platform: the Transcription Desk (TD) of the *Transcribe Bentham* (*TB*) initiative developed by University College London (UCL) and the University of London Computer Centre (ULCC). A primary user need identified by this study is *motivation*, to which a large array of academic literature already exists in both a general and crowdsourcing-specific context. This report begins with a brief theoretical overview of this subject. It then aims to contribute to the debate by examining, primarily through the results of an online survey, how motivation and other needs of *TB*'s CVCs can be identified, managed and ultimately fulfilled.

The evaluation of the second group of users is based on some general considerations as well as on talks with representatives from archives and a survey carried out at the Archivschule Marburg – Hochschule für Archivwissenschaft.

## Contents

# PART 1 User needs: Crowdsourced Volunteer Contributors

## 1. Introduction

The existing collaborative project between University College London (UCL) and the University of London Computer Centre (ULCC), *Transcribe Bentham* (*TB*), facilitates the online transcription of the manuscripts of English philosopher and jurist Jeremy Bentham (1748–1832). UCL's library holds approximately 60,000 folios of these manuscripts, with the British Library holding a further 12,500. Prior to the launch of *TB* in 2010, an estimated 20,000 folios in these two collections had been manually transcribed by UCL's Bentham Project since its inception in 1958 (i.e. 20,000 folios transcribed in 52 years, or an average of approximately 385 folios transcribed per year). In the two-and-a-half years since *TB*'s launch, over 5,000 folios have been transcribed by "crowdsourced volunteer contributors" (CVCs) alone, significantly improving this rate of transcription. However, whilst *TB* currently boasts over 2,800 registered users, it is actually a very small fraction (16, 0.5%) of this user base who are responsible for the bulk (95%) of the crowdsourced transcription work carried out thus far. Indeed, *TB*'s three most prolific CVCs have submitted a combined total of over 3,500 transcripts—over 70% of the total output of the *TB* initiative since its launch.

Several previous studies of "massive virtual collaboration" (MVC) projects, "information pools" and online communities have shown that the phenomenon present in *TB* outlined above—a minority of individuals producing most of the desirable content—is not uncommon.[1] Wikipedia—perhaps the archetypical MVC project or information pool—is one pertinent example. By 2008 it had accrued 6 million registered users, but the median number of edits made by each user was less than 1, meaning most of its members stopped contributing on the day they joined.[2] This phenomenon has occasionally been referred to as "social loafing": whilst many people may consume a social good (for example, Wikipedia articles), few ever contribute to its maintenance (for example, through editing or correcting said articles). This description is arguably not applicable to academic projects such as *TB*, where an obvious distinction can be made between the small number of CVCs who produce output and the similarly-sized community of scholars interested in Jeremy Bentham who will make use of it.[3]

---

1 Kevin Crowston and Isabelle Fagnot, "The Motivational Arc of Massive Virtual Collaboration," in *Proceedings of the IFIP WG 9.5 Working Conference on Virtuality and Society: Massive Virtual Communities*, 2008, 2; Coye Cheshire and Judd Antin, "The Social Psychological Effects of Feedback on the Production of Internet Information Pools," *Journal of Computer-Mediated Communication* 13, no. 3 (2008): 722, doi:10.1111/j.1083-6101.2008.00416.x; Gerard Beenen et al., "Using Social Psychology to Motivate Contributions to Online Communities," in *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, 2004, 1.

2 Kevin Crowston and Isabelle Fagnot, "The Motivational Arc of Massive Virtual Collaboration," 7.

3 See for example the similar distinctions made about citizen science projects Oded Nov, Ofer Arazy, and David Anderson, "Technology-Mediated Citizen Science Participation: A Motivational Model," in *Proceedings of the AAAI International Conference on Weblogs and Social Media (ICWSM 2011)*, 2011, 250. *TB* makes available its crowdsourced transcriptions through an open access digital repository, meaning any

Nevertheless, there is something intrinsically interesting about people who donate significant amounts of time and effort to online knowledge production for little (or indeed no) discernible reward. In an attempt to formulate a "motivational arc" to analyse the behaviour of these types of participants in MVCs, one study has identified three different levels to which they can be ascribed. Non-participants who are attracted to the project, register, and became "initial contributors" comprise the first level; initial contributors who progress to become "sustained contributors" comprise the second; and sustained contributors who eventually become "meta-contributors" comprise the third.[4]Meta-contributors are distinguished by the fact that they are concerned with both the structure of the project in its entirety and the community surrounding the project, as opposed to their own contributions to it. The CVCs participating in *TB* can only be identified at the first or second level of this classification. The potential to create meta-contributors is, however, something that should be explored, especially given the envisaged changes to the workflow process that the introduction of *tranScriptorium*'s (*tS*) HTR technology will generate.

The sustained (second level) contributors of *TB* take on an additional label in this report: "super-transcribers". Since the super-transcribers carry out the bulk of the transcription work in *TB*, the survey instrument[5] developed for this study is geared towards understanding the demographics of this group, the motivations behind their participation, and their other needs. The primary aim of this report is to forecast how such motivations and other needs may adapt or change (or, indeed, remain consistent) following the introduction of *tS*'s HTR technology into the TD platform, the software which runs *TB*.[6] A subsidiary aim is to speculate on how more volunteers of a similar type to super-transcribers can be recruited, either from initial or first level contributors already within the project or externally. Increased membership at the sustained contributor level is likely to lead to increased output in the project's goal of increased manuscript transcription.

## 2. Literature review: Motivation

This study posits that motivation is a crucial user need of CVCs. Motivation has been described as one of the two pillars of online citizen science projects (the other being a technological pillar, or 'computer systems to manage large amounts of distributed resources'). It means 'attracting and retaining volunteers who contribute their skills, time, and effort to a scientific cause'.[7] Whilst the transcription of Bentham's manuscripts is not strictly citizen *science*, it is undoubtedly predicated on similar technological and motivational pillars (and one survey respondent—a retired microbiologist—even equated transcribing Bentham's hand to his experiences processing raw scientific data). Understanding the motivations of volunteers has been a topic of long-standing academic interest, and the recent popularity of crowdsourcing has meant that the motivations of CVCs in particular has also become a

interested person can make use of them. The major users of this repository, however, are members of UCL's Bentham Project engaged in research.

4 Kevin Crowston and Isabelle Fagnot, "The Motivational Arc of Massive Virtual Collaboration," 4.

5 See part 2.1 below.

6 The Transcription Desk software is a customised MediaWiki.

7 Oded Nov, Ofer Arazy, and David Anderson, "Technology-Mediated Citizen Science Participation: A Motivational Model," 249.

subject of scholarly inquiry.[8] It is especially relevant given that many research projects now themselves benefit from crowdsourced labour. Like all types of volunteer, CVCs by their very definition do not carry out work for monetary reward, and other reward motives also appear unimportant.[9] Knowing what kind of non-monetary incentives motivate them to devote their time and effort to projects such as *TB* can therefore enhance both the retention of current volunteers and the recruitment of future ones, both of which are means to *TB*'s overarching objective: increased transcription of Bentham's manuscripts.[10] According to Clary et al., motivations to volunteer can be both selfish or unselfish. They concern 'the agentic pursuit of ends and goals important to the individual, and … the precise ends and goals can and will vary with the specific activity.' Underlying the decision to volunteer is 'a process by which individuals come to see volunteerism in terms of their personal motivations.'[11] Such understanding of users' ends and goals—in other words their *needs*—is therefore crucial to the longevity of a MVC project, given that the retention rate of active volunteers in many crowdsourced projects is typically very low. If user needs are not fulfilled, motivation to participate in an MVC project is likely to quickly evaporate.[12]

To conclude our brief theoretical analysis of motivation as a principal user need, this study will discuss how Hackman and Oldham's (1980) model on work motivations, as depicted in figure 1 below, is applicable to both *TB*'s existing and *tS*'s envisioned crowdsourced transcription platforms. The discussion revolves around two of the three principles the model identifies as intrinsic to work: meaningfulness of work (attainable through task and skill granularity), and knowledge of outcomes of work (attainable through feedback).

[8] For the former see for example E Gil Clary et al., "Understanding and Assessing the Motivations of Volunteers: A Functional Approach," *Journal of Personality and Social Psychology* 74 (1998): 1516–1530; For the latter see for example M Jordan Raddick et al., "Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers," *Astronomy Education Review* 9, no. 1 (2010), doi:http://dx.doi.org/10.3847/AER2009036; Oded Nov, Ofer Arazy, and David Anderson, "Technology-Mediated Citizen Science Participation: A Motivational Model"; Kevin Crowston and Isabelle Fagnot, "The Motivational Arc of Massive Virtual Collaboration"; Tim Causer and Valerie Wallace, "Building a Volunteer Community: Results and Findings from Transcribe Bentham," *Digital Humanities Quarterly* 6 (2012), http://www.digitalhumanities.org/dhq/vol/6/2/000125/000125.html.

[9] Oded Nov, Ofer Arazy, and David Anderson, "Technology-Mediated Citizen Science Participation: A Motivational Model," 249.

[10] Samuel Shye, "The Motivation to Volunteer: A Systemic Quality of Life Theory," *Social Indicators Resear*ch 98, no. 2 (2010): 183.

[11] E Gil Clary et al., "Understanding and Assessing the Motivations of Volunteers: A Functional Approach," 1528.

[12] Oded Nov, Ofer Arazy, and David Anderson, "Technology-Mediated Citizen Science Participation: A Motivational Model," 249.

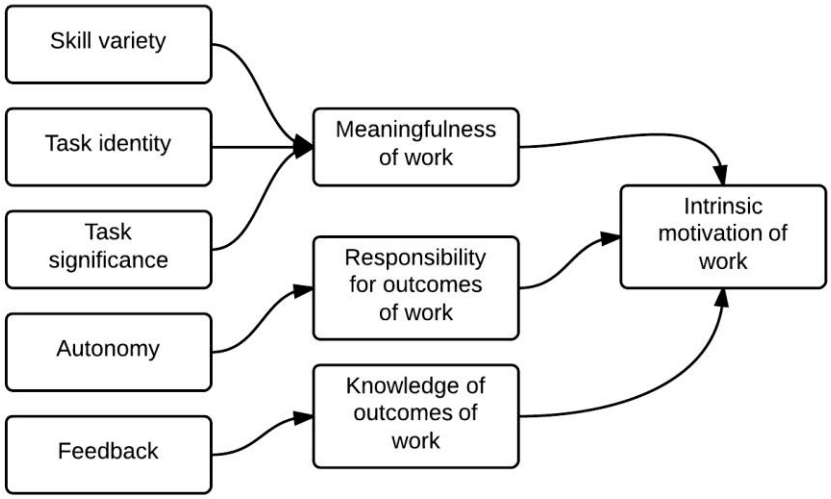


**Fig. 1 Work motivations, taken from Hackman and Oldham's (1980) model**

## 2.1 Meaningfulness

In their 2011 survey of MVCs, Nov, Arazy and Anderson utilised Benkler's 2006 definition of "task granularity" to distinguish between two different types of technology-mediated crowdsourced projects.[13] Task granularity, or 'the smallest individual investment necessary in order to make a contribution', was deemed very low in projects such as the volunteer computing initiative *SETI@home*, owing to the passive role the contributor played (users download and install a small program which then allocates their idle computer power to the analysis of radio signals, searching for signs of extra terrestrial intelligence). Conversely, high task granularity was identified in more active volunteer tasks such as the image analysis carried out by *Galaxy Zoo*, where users are tasked with the online classification of images of galaxies based on their shape. Nov et al. found that task granularity had a positive correlation with motivational levels: the greater the granularity of the task, the greater the level of motivation which was required in the participant to complete it. *TB* arguably possesses a very high—perhaps the highest—task granularity amongst MVCs, given the complicated nature of Bentham's handwriting and the need for its CVCs to not only transcribe this handwriting but encode it into XML. This complexity can be demonstrated by the time *TB*'s super-transcribers devote to working on individual transcripts. 80% of the survey sample said that on average they took at least 20 minutes to transcribe a single folio of Bentham's manuscript, and 40% of the sample said they took at least 30 minutes.[14] Motivation therefore appears to be a crucial factor to consider when evaluating user needs; if users are not sufficiently motivated, they will fail to perform the highly-granularity tasks required of them by projects such as *TB*.

[13] Ibid., 249–50.

[14] A folio is typically one single-sided page of manuscript, however some folios can comprise as many as four such pages.

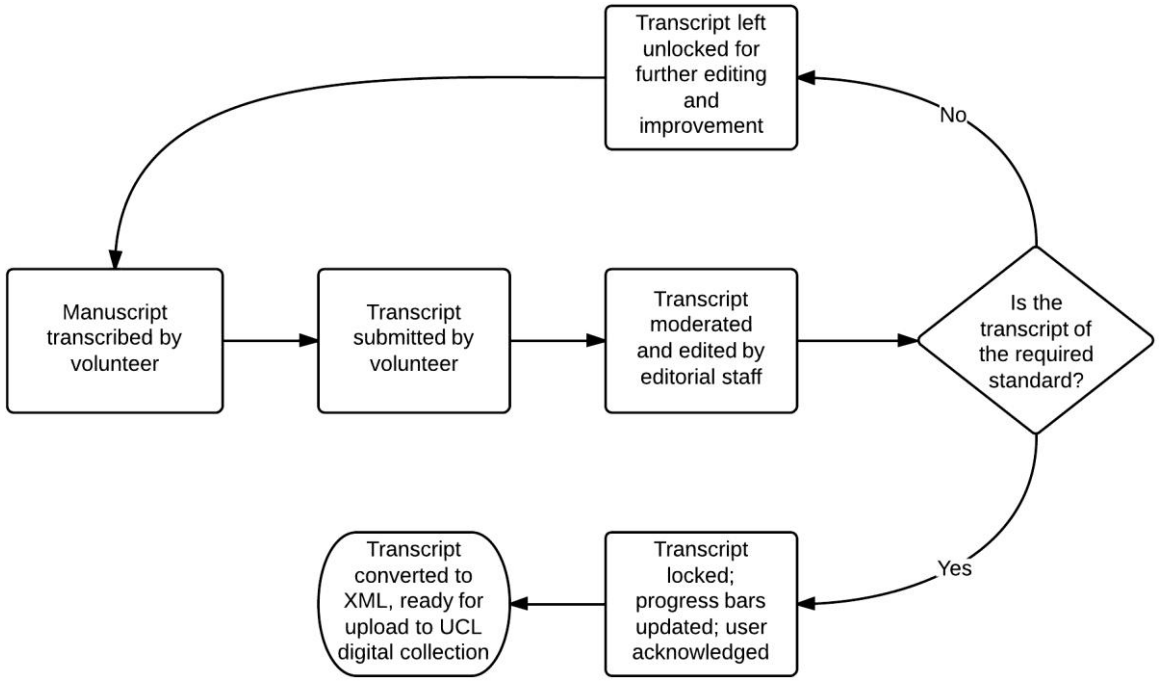


**Fig. 2 Current *TB* workflow**

It is envisaged that several additional process steps will be introduced in to the current *TB* workflow displayed in figure 2 above once the HTR technology being developed by *tS* is fully incorporated into the TD software. Potential new steps will allow for automated transcription of a selected manuscript to be carried out, for the user to check and edit the results of the automated transcription, and for corrections to be fed back into the HTR software for purposes of refinement. Critically, a CVC participating in the new, envisaged workflow may see their role redefined from one primarily of transcriber to one primarily of editor or checker (with only some attendant transcribing to correct errors or detect missing words). This means that there is potential for the granularity of the task to be lowered from the perspective of the CVC, if one accepts the argument that correcting an automatically transcribed manuscript is easier than manually transcribing it from scratch. Following Nov et al., this could correspond to a lower level of motivation required from the CVC to complete the task, as the transcribing process becomes "easier".

If it is likely that the introduction of HTR technology to the workflow will change the role of CVCs, it is possible that a future, potential *tS* user base may in fact resemble in composition that of other MVC projects such as the aforementioned *Galaxy Zoo*, whose task granularity is lower and, thus, requires less motivation in its volunteers to complete. This hypothesis is the source of much tension within the results of the super-transcriber survey. Several *TB* users have described manuscript transcription as akin to problem or puzzle-solving; this is not necessarily the same kind of person that would be attracted by a less-challenging crowdsourcing initiative that does not require the same level of intellectual effort to complete. To lower *TB*'s task granularity therefore runs the risk of alienating its current group of highly motivated sustained contributors, even if it may also open up participation to new and (in a strict sense) proportionally less-motivated CVCs.

A remedy to this risk may be found in Nov et al.'s argument that the relation between task granularity and motivations raises the need to create dynamic contribution environments that allow volunteers to start contributing at lower-level granularity tasks, and gradually progress to more demanding tasks and responsibilities. This is a very salient point regarding the potential of the HTR technology developed by *tS* to improve the user experience of *TB*. It makes possible the idea of two distinct levels of contributors—akin to Crowston and Fagnot's initial and sustained contributors—the former of which conduct relatively lower granularity editorial tasks such as the correction of the automated transcripts, the latter of which still carry out high-granularity transcription work, for example on complex documents that are unreadable by the HTR technology (such as Bentham's marginal summary sheets, or instances where the majority of Bentham's handwriting has been crossed out). There is also scope here to develop processes that devolve certain administrative responsibilities for the transcription platform to select volunteers—the role of the so-called meta-contributors mentioned in section 1 above.

A significant problem with the structure of this survey has therefore been identified: even if the user needs of current contributors to *TB* can be ascertained, it is likely that the introduction of HTR software into the workflow will redefine the user role, and in turn create a different demand on users: from one primarily of transcription to one primarily of editing or checking. How users would react to such a change was a key component of secondary, follow-up telephone interviews, outlined in section 3 and discussed in section 5 below.

### 2.2 Feedback

In *The Logic of Collective Action*, Mancur Olson argued that "selective incentives" could be used to encourage contribution to a public good and discourage the practice of so-called "free-riding" or social loafing, that is to say the consuming of a public good without contributing anything in return.[15] The obvious selective incentive for Olson was monetary benefit. In 2008, Cheshire and Antin built on Olson's hypothesis, postulating that social psychological processes, as opposed to monetary benefits, were also an effective selective incentive, and in turn defined one such social psychological process as the receiving of feedback.[16] Their study found that in cases where voluntary contributions to a task were solicited, the simple receiving of a response upon contribution was enough 'to prompt additional action by the original contributor', thus positively impacting upon the task's rate of completion.[17]

Feedback processes are central to the current workflow of *TB* and are similarly envisaged for the workflow of the platform incorporating the HTR technology of *tS*. It is central to the quality-control process that *TB*'s CVCs can access detailed feedback about their submitted transcript, and a summary of the changes made to a submission by the project's editorial staff is made available by the TD software. For example, if a user marks a word as unreadable in

[15] Mancur Olson, *The Logic of Collective Action: Public Goods and the Theory of Groups.* (Harvard University Press, 1965), 61–3.
[16] Coye Cheshire and Judd Antin, "The Social Psychological Effects of Feedback on the Production of Internet Information Pools," 709.
[17] Ibid., 711.

their submitted transcript but that word is subsequently identified by an editor in the checking stage, the user is able to view a side-by-side comparison of this and other changes within the context of their entire transcript once checking has been completed. The logic behind this type of feedback process is that CVCs will learn from their mistakes, and the process also shows the corrections made by the editors to the user's XML encoding for identical reasons. In Beenen et al.'s 2004 study, no similar feedback mechanism which delivered such tailored responses to individual contributions was found in their survey of online communities.[18] But whilst the results of the super-transcriber survey highlight the importance of this constructive feedback mechanism, for example in the refinement it gives to their transcription and encoding skills, Cheshire and Antin's study instead focuses on more basic kinds of feedback process. *TB* possess two such further instances in this regard, one individual and one collective. The first comprises the notifications CVCs receive once their submitted transcripts have been checked by editorial staff and subsequently marked as of a sufficient standard for academic use. This notification is typically a formulaic response left on a user's public profile page, noting the manuscripts they have transcribed and thanking them for contributing to the project. The second, which aligns rather closely with Hackman and Oldham's description of feedback as 'knowledge of the outcomes of work', is in the form of the weekly "Transcription Update" reports issued by *TB*'s editorial staff and published on the project's blog and social media accounts. Within this report, details are given about the collective progress of the project, such as number of transcripts completed and the current rate of transcription. Further community information about news, events and any significant discoveries found within newly-transcribed material are also included.

According to Cheshire and Antin, in situations where intrinsic motivations (such as altruism, fun, intellectual stimulation, a sense of obligation to contribute, etc.) are high, extrinsic motivations such as simple individual feedback to task completion can actually 'counteract the effects of the intrinsic motivation and lower overall contribution behaviour.'[19] Since super-transcribers of *TB* are considered highly intrinsically motivated given the high granularity of their task, the supposition to draw from this argument is that the simple feedback responses given by *TB* editors upon checking submitted transcripts either have zero impact or may actually impact negatively upon CVCs. Meanwhile, according to Nov et al., collective motives—such as identification with the project's goals—are only moderately related to intentions among "active-participation volunteers" such as *TB*'s CVCs.[20] Whilst that study found that the importance of the project's goals was a salient motivational factor for initial contributors, its effect dissipated amongst sustained contributors (such as super-transcribers), who were far more motivated by personal goals, or what Clary et al. has described as 'the agentic pursuit of ends and goals important to the individual'.[21] It is therefore theorized that simple feedback on collective progress in the form of the Transcription Update, much like simple feedback on individual progress in the form of

[18] Gerard Beenen et al., "Using Social Psychology to Motivate Contributions to Online Communities," 8.

[19] Coye Cheshire and Judd Antin, "The Social Psychological Effects of Feedback on the Production of Internet Information Pools," 721.

[20] Oded Nov, Ofer Arazy, and David Anderson, "Technology-Mediated Citizen Science Participation: A Motivational Model," 254.

[21] E Gil Clary et al., "Understanding and Assessing the Motivations of Volunteers: A Functional Approach," 1528.

notifications, may have a limited effect on motivating CVCs: they are not user needs. The feedback hypotheses of both Cheshire and Antin and Nov et al. represent sophisticated challenges to the classic argument by Bandura and Schunk that "consistent positive feedback should encourage high collective efficacy", as well as Crowston and Fagnot's related belief in a "virtuous cycle of contribution", whereby "as an individual contributes, they become more visible, which increases the likelihood of feedback and thus further contributions."[22]

# 3. Survey

## 3.1. Survey instrument

The online survey contained a total of 61 questions, however respondents only answered a maximum of 52 or 53 questions depending on their response to question 17, which determined whether they were an active transcriber (that is, whether they had contributed to *TB* in the past two months). The penultimate question of the survey asked whether participants would be willing to be contacted by telephone as a follow-up to their online responses.

The survey attempted to question the respondents on several different subjects: 1) their motivations for starting to participate in *TB*; 2a) their reasons for stopping their participation (if applicable); 2b) their current motivations for participating in *TB* (if applicable); 3) their experiences with using *TB* and how this experience could be improved; and 4) whether they engaged with *TB* through social media. After establishing basic demographical information such as age and nationality, the survey asked what drew the respondents to initially participate in *TB*. Seven of the questions posed in this part asked respondents to rate statements on a Likert scale of 1 to 7,[23] where 1 meant they felt the contents of the statement was not at all motivating to their initial participation in *TB*, and 7 meant the contents of the statement was very motivating. Examples of these statements were "I was interested in or wanted to find out more about Jeremy Bentham" and "I wanted to contribute to an academic research project". These questions were asked in a randomized order and were followed by a free-response question which asked the respondent if any other factors motivated their initial participation in *TB*. This was included to give respondents the opportunity to report motivations other than those listed on the survey instrument.

Depending on whether the respondent identified themselves as an active transcriber or not, they were directed to separate following sets of questions. Non-active transcribers were asked to rate how relevant statements were in explaining why they were no longer transcribing, again on a Likert scale of 1 to 7. Examples of these statements were "I am too busy with work, commitments or other projects" and "I did not feel adequately rewarded for my efforts". These questions were also followed by a free-response question where any other factors could be listed. Active transcribers meanwhile received the same 7 questions regarding initial motivations, transposed to the present—rather than past—tense (i.e. instead of "I was interested in history and/or philosophy" they received "I am interested in history

[22] Kevin Crowston and Isabelle Fagnot, "The Motivational Arc of Massive Virtual Collaboration," 11.

[23] Rensis Likert, "A Technique for the Measurement of Attitudes," *Archives of Psychology* 22, no. 140 (1932): 34.

and/or philosophy"), as well as a final free-response question. Again, the questions were presented in a randomized order.

Respondents were next asked about their experiences using *TB*, including how much time they spent transcribing a single folio, whether they had ever received formal training in palaeography or text encoding, how difficult they found these activities, and whether they found one more difficult than the other. They were then presented with a series of 8 statements about specific improvements that could be made to *TB*, and were asked to rate statements on a Likert scale of 1 to 7, where 1 meant they felt the contents of the statement was not at all relevant to improving their experience with *TB*, and 7 meant the contents of the statement was very relevant. Examples of these statements were "Assistance with reading Bentham's handwriting", "More feedback on my work from the editors of *Transcribe Bentham*" and "Better computer hardware from where I work (e.g. larger screen, faster PC, better internet connection)". The questions were presented in a randomized order and there was also a final free-response question.

The final set of questions asked respondents about whether they kept in contact with the editors of *TB*, and whether they followed the progress of the project on social media platforms such as Facebook. Respondents were also asked about their own sharing of *TB* news and updates within their social circles—for example whether they shared their progress with family members or friends.

### 3.2. Survey implementation

The online survey was carried out over a period of 4 weeks. The twenty top contributors to *TB* were contacted with a link to the online survey. This link contained a generated unique key to ensure only the invitee could respond. Reminders to complete the survey were sent out at the end of week 2.

By the end of the survey period, 10 responses, or 50% of the invited respondents, had been recorded. Of these ten, 8 initially expressed interest in participating in a follow-up telephone interview, however only five responded to further attempts to arrange potential times for this to take place. Two of this number subsequently could not be reached before this report's due date, but subsequent attempts will be made to contact them to augment the number of telephone interviews from 3 to 5.

The telephone interviews were unstructured, and respondents were asked to freely elaborate on any of their answers to the original online survey. One specific question that was asked, however, was how they felt about the introduction of an automated transcription tool (of the type envisioned by *tS*'s HTR technology) to the transcription workflow, and whether this would positively or negatively impact their opinion of their role as a super-transcriber (especially regarding the potential redefinition of role suggested in 2.1 above). The responses to the telephone interviews are discussed in section 5 below.

Whilst it is true that our sample size was small for the online survey and even smaller for the telephone interviews, this was unavoidable given the make-up of the *TB* user base and the prevalence of 'super-transcribers' within it who carry out the bulk of the transcription work.

## 4. Results

Of the survey respondents, 60% had contributed to *TB* in the form of some degree of transcription work during the past 2 months, the cut-off point after which respondents were considered "non-active" transcribers. Therefore, as outlined in section 3.1 above, 40% of respondents received questions 18–26 and 60% received questions 27–34.

A full outline of the survey results, including histograms of Likert scale responses, is available in the attendant "Comment report". In the sections below, results are elaborated for the questions pertaining to demographics, motivations, and other identifiable user needs.

### 4.1 Demographics

All super-transcribers are over 40 years old and 50% are retired. 70% of respondents were female and 30% were male. Half of respondents were from the UK and 40% were from the USA. Only one respondent resided elsewhere (in France). The frequency of responses is given in table 1 below.

| **Choices** | Absolute frequency | Relative frequency |
|---|---|---|
| **41 to 50 years of age** | 3 | 30% |
| **51 to 60 years of age** | 3 | 30% |
| **Over 60 years of age** | 4 | 40% |

**Table 1** Age range of respondents

More than half of respondents possessed a postgraduate qualification of some sort (MA, PhD, etc.), but 70% had never had any formal palaeography training and 60% were unaware of the principles of text-encoding prior to contributing to *TB*. This suggests that there are actually few skill pre-requisites to becoming a super-transcriber with the ability to decipher Bentham's hand and encode transcripts in XML mark-up. The results show that both these complex skills can be learnt relatively quickly through the instructive materials provided by the *TB* website or through intuitive use of the TD software. If correcting HTR results represents a complicated skill set that CVCs need to be trained in, it is unlikely that such skills could not be easily picked up by current *TB* super-transcribers. On a related note, respondents regarded text encoding as at worse an inconvenience, and nothing like as difficult as reading handwriting. This bodes well for the potential technical requirements of using HTR technology within a transcription workflow and also implies that assistance in reading Bentham's handwriting—an obvious goal of HTR—will be welcomed by CVCs.

Although 50% of respondents had some knowledge of other MVC projects such as *Galaxy Zoo*, only 1 respondent (i.e. 10% of the sample) had actually participated in another crowdsourcing project.

### 4.2 Motivations

#### *4.2.1 Initial motivations*

The highest rated motivations for participating in *TB* were "wanting to help" and "contributing to an academic project". 90% of respondents ranked "wanting to help" as a motivating factor at 6 or above on the Likert scale of 1–7. An identical number ranked

"contributing to an academic project" at 6 or above, with 50% of respondents giving this reason the highest possible ranking (7).

In taking a rating of 4 or above on a Likert scale of 1–7 as indicative of a motivating factor, three other motivations were ranked by at least 90% of the respondents: "wanting a challenge" (including 40% at 6 and 20% at 7 on the scale); "wanting to find out about Bentham" (100% were motivated by this factor, with 10% at 6 and 50% at 7); and "looking at material that few people had ever seen before" (30% at 6 and 10% at 7).

Responses to the free-response question stressed the respondents' desire to find something "constructive" or "mentally challenging" to do with their free time.

*4.2.2 Non-active transcribers*

The 40% of respondents who said they had not transcribed in the past 2 months were given a set of statements containing potential reasons for their lack of contribution, and asked to rate how relevant each statement was on a Likert scale of 1–7. The results here were unequivocal. 100% of respondents rated the statement "I am too busy with work, commitments or other projects" at 7 on the scale. 25% also said that being "no longer interested" in either Bentham or transcription was a factor (both rated at 4 on the scale). 50% said they found both transcription and encoding difficult or boring, and an identical figure found the transcription tools inadequate for the task required. 0% felt inadequately rewarded for their efforts—an expected although still welcome response given the nature of volunteer motivations outlined in part 2 above.

*4.2.3 Active transcribers*

The 60% of respondents who are currently still transcribing were given the same set of statements regarding their initial motivations but asked how they applied to their current motivations. The strongest motivating factors were wanting to help (82.5% at 6 or above on the scale), contributing to an academic project (83% at 6 or above), the challenge of transcribing (66% at 6 or above), or an interest in Bentham or history and/or philosophy (both at 66% at 6 or above).

By again taking a rating of 4 or above on a Likert scale of 1–7 as indicative of a motivating factor, comparisons can be drawn between initial and present motivations, as shown in figure 3 below. In this light, 100% of respondents said they were now motivated by the challenge (vs. 90% previously), by wanting to help (vs. 90%), by contributing to an academic project (vs. 90%) and by an interest in history and/or philosophy (vs. 100%). 82.5% of respondents said they were now motivated by an interest in Bentham, an increase of 2.5% on initial motivations, and 83% were now motivated by palaeography or transcription, an increase of 3%.

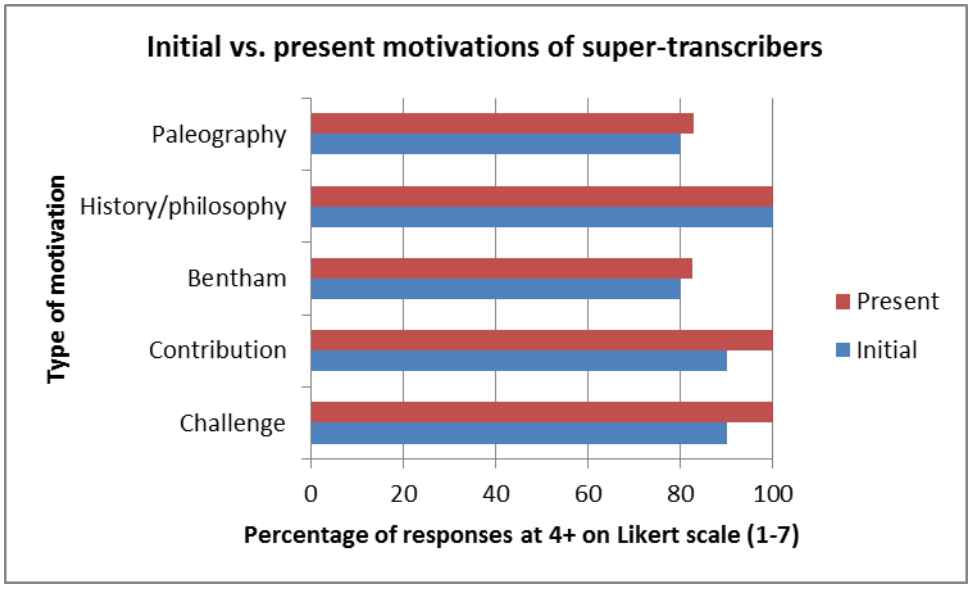


**Fig. 3** Initial vs. present motivations of super-transcribers

Responses to the free-response question mentioned the importance of personal contact and feedback with the editors of the project, a desire to continue learning, and the intrinsic reward of "deciphering a document".

### 4.3 Other Needs

All respondents, regardless of whether they were currently active in *TB* or not, were then asked to rate how relevant a series of statements were to their productivity (on a Likert scale of 1—7, with 1 being not at all relevant and 7 being very relevant), that is to say, what factors might allow them to transcribe more. This was a way of identifying specific technical user needs.

Unsurprisingly, 60% of respondents ranked having more time at the highest relevancy (7 on the scale). 60% of respondents said improved transcription tools such as better image zooming would also increase their productivity (6 or higher on the scale).

By taking a rating of 4 or above on a Likert scale of 1–7 as indicative that a statement was relevant, 80% of respondents said they would welcome assistance with reading Bentham's handwriting, 80% said they would welcome more feedback on their transcription efforts and 60% would welcome no longer having to encode the transcribed text.

Other needs identified by answers to the free-response question included "better documentation of XML's peculiarities", better feedback, and several technical feature requests such as the ability to rotate images and auto-saving of progress.

## 5. Discussion

This report commenced with a theoretical overview of what was considered a vital user need for CVCs in current transcription projects such as *TB* and envisioned transcription projects such as *tS*: motivation. Establishing what motivated *TB*'s CVCs was therefore the principal task of the super-transcriber survey. The survey results suggest that super-transcribers are highly motivated by altruistic reasons, such as the contribution they make to an academic

research project, as much as they are for personal reasons, such as their desire for something challenging to do, or to satiate their interest in history and/or philosophy. In this regard, super-transcribers conform to the profile of volunteer motivations advanced in recent academic studies such as Clary et al., i.e. they have both selfish and unselfish motives for participating. Indeed, the real hindrance to participation by super-transcribers in *TB* was not motivation (or a lack thereof), but time. A lack of time rather than motivation was cited by all non-active respondents as the primary reason for their current lack of contribution. Similarly, 70% of all respondents said they would contribute more to the project if they had more time.

Super-transcribers, however, already devote a lot of time to their endeavours. 80% of respondents say they spend on average more than 20 minutes transcribing a single folio—that is to say, completing a single transcription task. Of this figure, 20% said they spend over an hour, and 10% spend over two hours. More time is of course a difficult commodity to provide to CVCs, but this particular user need does chime with the envisaged benefit of the integration of *tS*'s HTR technology in to the TD software: the automated transcription of Bentham's manuscripts. This improvement is likely to eliminate the bulk of the transcription work for the user, meaning less time will need to be spent manually transcribing each individual manuscript. If super-transcribers were to continue to devote a similar amount of time to the future *tS* crowdsourced platform as they currently devote to *TB*, the overall rate of transcription is likely to increase provided that HTR makes the task more efficient.

One issue speculated on prior to the analysis of the survey's results was whether the envisioned simplification of the transcription task through the use of HTR technology, and the subsequent redefinition of the role of CVCs from transcribers to checkers or editors, could alienate existing *TB* users from the project, because their task would become less challenging. This argument follows the logic of part of Hackman and Oldham's hypothesis on intrinsic work motivation (cited in section 2 above) that they defined as work's meaningfulness, and this concern is borne out by the fact that the challenging nature of the transcription task was an important motivation CVCs gave for participating in *TB* (as a motivating factor it rose from 90% to 100% between initial motivations and current motivations of super-transcribers). There is also a slight paradox to resolve here between the argument advanced in Nov et al. stated in section 2.1—that the granularity of a task correlates with levels of motivation (in the sense that high motivation was required to complete high granularity tasks)—and the survey respondents' claims that it is the challenging nature of the task itself (shown, for example, in the average time they taken to complete a single transcript) that actually motivates them in the first place. Whilst the results of this survey certainly conform Nov et al.'s notion of a correlation between high task granularity and high user motivation, they also rearrange it. Rather than MVC projects with high task granularity seeking to motivate volunteers to complete tasks, it is actually the other way round: highly-motivated volunteers seek out high granularity tasks.

Despite the abovementioned concerns about the effects on super-transcribers of task granularity being lowered by the introduction of *tS*'s HTR technology, the results of the survey also highlight that 80% of respondents would welcome assistance with reading Bentham's handwriting, identifying this as a key factor that would enable them to contribute more to the project. Whilst such assistance might not have been envisaged as coming from

the introduction of *tS*'s HTR technology to the transcription workflow, it is doubtless that HTR represents the kind of support users would find valuable. In this regard, it is possible that the concerns about user alienation mentioned in the previous paragraph have been overstated. This sentiment was confirmed in the follow-up telephone interviews. When the possibility of HTR technology was described in detail to the interviewees, they were highly receptive to idea and excited about its potential to aid them in their transcription work. No respondent suggested they were perturbed by the redefinition of their role, from transcriber to editor, that such a development would engender. This was perhaps the most important conclusion to draw from the survey results.

If the new TD software means a change in the skillset required of the user, from one of primarily transcription to one more of checking or editing, an opportunity presents itself to recruit new volunteers who are not as intrinsically highly-motivated as current super-transcribers. In automating the transcription process, the envisaged transcription platform of *tS* may more closely resemble scientific crowdsourcing projects such as *Galaxy Zoo*, which comparatively-speaking have a lower task granularity than *TB* because they have a more mechanical or repetitive nature to their crowdsourced tasks. This again highlights the issue of time as a crucial user need. If the lowering of task granularity means task completion is quicker, users will receive the satisfaction of completion (and contribution) much more quickly than at present; the investment of time required by an initial contributor to contribute is reduced, and this may encourage them to become sustained contributors. One telephone interviewee provided particular insight on this point. Describing the initial attempts at transcription in *TB* as "very daunting", the respondent stated that they completed at least five transcripts 'before [they got] the feeling of satisfaction'. If the length of time it takes from contribution to satisfaction can be reduced, it is likely to have a positive impact on the contributing user base and a corresponding effect on the rate of transcription.

Alongside the intrinsic motivation for work provided by task and skill meaningfulness posited in Hackman and Oldham's model, represented in this study by analysis of *TB*'s current and *tS*'s potential task granularity, is the additional intrinsic motivation provided by knowledge of the outcomes of work, or in simpler terms: feedback. The consideration of feedback as a user need in the context of this survey of super-transcribers is interesting, given the dismissal of so-called "simple" feedback mechanisms found in parts of the current academic literature as outlined in section 2.2 above. 80% of respondents stated that they would like more feedback about their work, and this is taken to mean feedback in a complex sense, i.e. a detailed response to a user's submitted transcript, outlining their errors in transcription and text encoding. Such a feature is actually provided for in the TD software, but it became apparent that many respondents were unaware of its existence, and that it should be advertised better. There is thus potential to introduce an iteration of Crowston and Fagnot's "virtuous cycle of contribution" here, if super-transcribers use the detailed feedback made available by the TD software to refine their own transcription and text encoding skills.[24]

[24] Kevin Crowston and Isabelle Fagnot, "The Motivational Arc of Massive Virtual Collaboration," 11.

Respondents to the online survey offered no opinion on the simple personal feedback process of user notification on the acceptance of a submitted transcript, however the inference from those stating that they would like more feedback is that such a mechanism is insufficient as a source of motivation. One telephone respondent mentioned that they did look out for the user notification response, but only because they used it to update their overall progress in a notebook. Another respondent mentioned that one needed at least five such notifications before deriving any satisfaction from the transcription task. These results seems to support Cheshire and Antin's hypothesis that in situations where intrinsic motivation is high, extrinsic motivations such as simple individual feedback to task completion have no effect, however the survey did not test Cheshire and Antin's further supposition that this type of feedback actually had a negative effect on super-transcribers. The hypothesis of Nov et al., meanwhile, that collective feedback—such as information about the progress of the project's goals—is only moderately related to the motivations of CVCs is not supported by the results of the survey. The motivating factor of contributing to an academic project and wanting to help was rated highly in the survey's questions about both initial and present motivations. It was stated emphatically as the main motivating factor by one telephone respondent, and an interest in the project's wider or collective goals is also communicated by the fact that 80% of respondents said that they read the project's blog, which publishes the weekly Transcription Update.

## 6. Conclusion

This report draws conclusions on several issues. On the subject of motivation, the survey found that *TB*'s super-transcribers are highly-motivated for both personal and altruistic reasons, and that the combination of complexity of the transcription task and contribution to an academic research project is a powerful incentive to participate. The current high task granularity of *TB*, however, does represent something of a threat to its longevity. Since only highly-motivated CVCs are able to carry out the transcription tasks, maintaining a sufficient number of such super-transcribers is a significant challenge. The time it takes to typically complete a transcription task directly feeds into this issue, since it privileges those with a lot of time to dedicate to transcription—revealed in the demographic results of this study to be those over 40 and, significantly, those who are retired.

Whilst *tS*'s HTR technology may not be a panacea for all the concerns of maintaining adequately motivated CVCs for crowdsourced transcription projects such as *TB*, its theoretical application to the workflow of *TB* has shown that it comprehensively addresses a number of issues pertaining to the user needs outlined in this report. By reducing task granularity, *tS*'s HTR technology should lower the entry-point from which non-contributors can begin to contribute, thereby increasing the user base and improving the rate that these initial contributors become sustained contributors, or super-transcribers. Meanwhile, the simplification of the task (through automated transcription of Bentham's handwriting) will enable existing users to produce more transcriptions in the same amount of time they currently devote to the project, thereby positively impacting upon the overarching goal of increased transcription of Bentham's manuscripts. The threat of alienating existing *TB* super-transcribers through this redefinition of their role, moreover, is not supported by the telephone respondents who were asked this question directly.

As abovementioned, the survey reports that the demographic composition of *TB*'s super-transcribers is notable for its reliance on volunteers who are retired (50% of respondents). Efforts should be made to continue recruitment in this area, but efforts should also be made to diversify the user base. This is eminently possible with the envisaged lowering of task granularity through the introduction of *tS*'s HTR technology. However, even if task granularity is lowered, motivation will still need to be relatively high compared to most other MVC projects. Therefore, users who possess high intrinsic motivation still need to be recruited and maintained, and there is currently no straightforward solution to attracting such volunteers.

One key to keeping CVCs motivated is through the provision of detailed feedback, conforming to Hackman and Oldham's model that knowledge of outcomes of work is important to the intrinsic motivation to work. 80% of super-transcribers stated that they would like more feedback from editors regarding their performance, and this is certainly deliverable. High quality feedback has the potential to not only improve CVC retention, but also refine the skills of said CVCs.

Finally, there are three areas where the survey and this report could be notably improved. Firstly significant statistical analysis, absent in this report, should be conducted on the survey's results. Secondly, the study would clearly benefit from further telephone interviews which, owing to their less-structured nature, potentially offer more insight than their online equivalents. Thirdly, participants of other, similar MVC projects should also be surveyed with a similar instrument, to ascertain whether the user needs of *TB*'s super-transcribers are typical or not of other CVCs.

## 7. Bibliography

Gerard Beenen, Kimberly Ling, Xiaoqing Wang, Klarissa Chang, Dan Frankowski, Paul Resnick, and Robert E. Kraut. “Using Social Psychology to Motivate Contributions to Online Communities.” In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, 212–221, 2004.

Coye Cheshire, and Judd Antin. “The Social Psychological Effects of Feedback on the Production of Internet Information Pools.” *Journal of Computer-Mediated Communication* 13, no. 3 (2008): 705–727. doi:10.1111/j.1083-6101.2008.00416.x.

Kevin Crowston, and Isabelle Fagnot. “The Motivational Arc of Massive Virtual Collaboration.” In *Proceedings of the IFIP WG 9.5 Working Conference on Virtuality and Society: Massive Virtual Communities*, 2008.

E Gil Clary, Mark Snyder, Robert D. Ridge, John Copeland, Arthur A Stukas, Julie Haugen, and Peter Miene. “Understanding and Assessing the Motivations of Volunteers: A Functional Approach.” *Journal of Personality and Social Psychology* 74 (1998): 1516–1530.

M Jordan Raddick, Georgia Bracey, Pamela L Gay, Chris J. Lintott, Phil Murray, Kevin Schawinski, Alexander S. Szalay, and Jan Vandenberg. “Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers.” *Astronomy Education Review* 9, no. 1 (2010). doi:http://dx.doi.org/10.3847/AER2009036.

Mancur Olson. *The Logic of Collective Action: Public Goods and the Theory of Groups.* Harvard University Press, 1965.

Oded Nov, Ofer Arazy, and David Anderson. “Technology-Mediated Citizen Science Participation: A Motivational Model.” In *Proceedings of the AAAI International Conference on Weblogs and Social Media (ICWSM 2011)*, 2011.

Rensis Likert. “A Technique for the Measurement of Attitudes.” *Archives of Psychology* 22, no. 140 (1932).

Samuel Shye. “The Motivation to Volunteer: A Systemic Quality of Life Theory.” *Social Indicators Research* 98, no. 2 (2010): 183–200.

Tim Causer, and Valerie Wallace. “Building a Volunteer Community: Results and Findings from Transcribe Bentham.” *Digital Humanities Quarterly* 6 (2012). http://www.digitalhumanities.org/dhq/vol/6/2/000125/000125.html.

# PART 2: Archives as users of HTR (Content Provider Platform scenario)

## 1. Introduction

In addition to the "Crowdsourced Volunteer Contributors" two further user groups that are of interest for an HTR platform in a "Content Provider Platform" scenario are 1) researchers and 2) archives.

### 1.1 Researchers

The main characteristics of "researchers" with respect to their interest on handwritten texts are:

- That they come from different disciplines and that the transcription of handwritten text is a "daily" job in the course of carrying out historical investigations. For instance studies on the history of medicine, theology, biology, philosophy, etc. will always require the reading and transcription of historical documents.
- That in some cases the documents are so important that a scholarly edition will be published – as in the case with the Bentham papers at UCL – but that generally transcriptions will be done for the sake of taking a note from the content of a document, not for the purpose of actually transcribing the whole document.
- That researchers are often *the* experts for a document or a collection of documents. In contrast to CVCs, they have deep insight into the document, its importance in a larger context and they have a critical view on it (e.g. on the originality of the content or the history of the document itself).
- That researchers are "responsible" for their work and are therefore highly interested in quality. Again, in contrast to CVC users who enjoy a degree of anonymity, researchers must publish their results and stand with their full name behind their transcriptions. This is even more the case if the transcriptions are edited and published. Near perfect "quality" will therefore be one of their main requirements.

Nevertheless, it is rather obvious that the user needs of researchers, in terms of the interface for actually transcribing a manuscript, will in fact be very similar to those of CVCs. Effectiveness, transparency, performance and the chance to interact with others will play the same important roles. It is therefore not necessary to extend further on this issue, and instead assume that the user needs of researchers are in this respect the same as the needs of CVCs.

### 1.2 Archives

The needs analysis is completely different when considering archives as users of the HTR platform. Archives are considered as users for two primary reasons:

1. Archives are the "owners" of handwritten documents. Though major libraries will also contain some collections of handwritten material, e.g. codices from the middle ages or personal documents of famous persons, the overwhelming majority of handwritten documents reside in archives. And archives have their own tradition, history and identity, meaning the decision to digitise a document and to expose it via

an HTR platform for transcription will ultimately be taken by archives and not by libraries or researchers. Thus archives are not only the owners but also those who will deliver digital page images of documents to an HTR platform – and in order to convince them to do so their user needs must be taken into account as well.

2. Even comparatively poor Word Error Rates (WER) for a transcription tool still represent an enormous advantage over current archive systems and also support further digitisation activities in archives. Whilst the ultimate goal of a transcription project may be a complete and correct text, for mass-digitisation projects a comparatively poor text may also be an option. A WER of 30–50% may not be useful in speeding up the manual transcription process, since correcting a text is more work for an experienced user than writing it from the scratch, but it would allow researchers and other users to search in large handwritten collections (as is the case in printed text collections with poor OCR quality). Thus the way HTR technology can be integrated into the digital environment of archives also needs to take user needs into account.

In order to produce a detailed summary of the specific situation of archives with regard to digitisation in general and transcription specifically a survey instrument was produced.

## 2. Survey on digitisation and transcription

In 2013 the German Science Foundation started a three-year project called "Pilot project for the digitisation of archival sources". The project is led by the Archivschule Marburg – Archivwissenschaftliche Hochschule and comprises six well-known archives from across Germany. The Archivschule Marburg is a central institution with university status where all German archivists are educated. In November 2013 a two-day conference was held entitled "Digitisation in the archive", where the user needs survey was carried out. All 21 respondents to the survey were currently carrying out the university course as part of their regular professional training as archivists. Thus, whilst they may be regarded as "young archivists", they already have professional experience in the field.

The survey was structured into three main fields of interest:

a) What is the general attitude of archivists towards digitisation?
b) What do they think about an HTR platform?
c) What is their general attitude towards transcribing texts?

The survey presented a full sentence ("statement") with the option to fully agree, to slightly agree, to slightly disagree or to fully disagree. When evaluating the survey the counts for "fully agree" were weighted double compared to "slightly agree", in order to indicate the strength of emotion. Thus if two persons slightly disagreed with a statement, this was counted the same as if one person fully disagreed with it.

### 2.1 Selected Results

*Statement 1.1: Digitisation in archives is still at the very beginning.*

Agree: 20 Disagree: 4

*Statement 1.5: Digitisation is the most important task of archives in the future.*

Agree: 4 Disagree: 25

*Statement 1.14: In some years we will not talk about the sources which shall be digitised, but only about the sources which will not been digitised at that time.*

Agree: 7 Disagree: 20

In contextualizing these three questions it is rather obvious that a majority of the respondents perceive digitisation not as a "revolution" but still have the understanding that it relates just to some peripheral issues. A good example was mentioned by one of the respondents that digitisation will be similar to microfilming. Digitisation is just "another microfilm" – and microfilm campaigns were carried out for decades in archives for safeguarding their unique content. But microfilming remained just a very specific activity with nearly no impact on the archival work itself. It is important to understand that archivists do not seem to see the great potential impact of digitisation. The fact that it will change all processes in an archive, from indexing to searching, from communicating with a user to making material online available is not fully understood.

The second part comprised questions about an HTR platform. We described the platform according to our "scenario" of a CPP in the following way:

- An archive digitises 200,000 files from a law court from the years 1600 to 1900.
- The digital objects are made available freely to the public via a digital library.
- Simultaneously the archive delivers the digitised pages to a service platform. There the text of the page images is automatically recognised with Handwritten Text Recognition software. The word accuracy is between 50–70%.
- The archive encourages interested researchers, students and other persons to contribute in the correction of the (automated) transcription. Such interested parties register an account on the platform and – with the support of the software – are able to correct the transcription.
- The persons carrying out the corrections of transcriptions are allowed to use their text as well the text from others and to store it on their computers.
- The archive gets back the automated and the corrected transcriptions in a standard format. Based on this output users are able to search in the full-text of the archival material.

In the following we highlight some of the statements with regard to this scenario:

*Statement 2.1: This seems to be an interesting model and I would like to see its realisation.*

Agree: 14 Disagree: 8

*Statement 1.11: An archive should make available digitised material for free in the Internet only in exceptional cases.*

Agree: 11 Disagree: 19

*Statement 2.8: I am convinced, that on the long run an archive can benefit from voluntary contributions e.g. for transcription or for indexing.*

Agree: 21 Disagree: 4

*Statement 2.5: The "outsourcing of the transcription" may work for mass sources, but not for high-level sources from the early modern time.*

Agree: 23 Disagree: 5

In summary, it becomes clear that the respondents have difficulties finding a clear attitude towards the model. On the one hand, there is a majority that believes that an HTR platform that works independently from the archive is "an interesting model", but such a majority is weak. There is also the clear constraint that for "high level sources" this may not be feasible due to the lack of knowledge at the users. On the other hand, respondents believe that voluntary staff are able to make valuable contributions. Also their openness towards the making available of archival sources should be emphasized, although in this case the statement is not very strongly supported. In short: There are several prerequisites already in place which would foster the adaption of the HTR platform model, but a clear picture is missing.

This is summarized in another statement, where we asked implicitly for the tolerance on the performance of an HTR engine:

*Statement 1.10: If a software would be able to carry out the automatic recognition of handwritten text, than 50-70% of word accuracy would be sufficient.*

Agree: 5 Disagree: 31

It is very clear that the software is compared with a human being, and if a person is only able to read between 50–70% of the words of a text, it would not be regarded as sufficient. From an information retrieval point of view, however, the situation looks totally different. That there is now the chance to find 50–70% of the words of a text is a significant increase on what was possible before. Archivists, in a similar vein to librarians, do however have a different understanding of searching to that of a researcher: they wish to search in a complete set and the fear is "to miss anything". Both requirements cannot be fulfilled – the set will very likely never be complete in a large collection of digitised texts and therefore the fear to miss something is significant. Indeed, it is exactly this attitude which prevented many libraries to apply OCR technology to their digitised collections.

For the integration of HTR technology into archives this means that on the one hand unrealistic expectations will be raised on the accurateness of the process where probably even 90% word accuracy would be regarded as too erroneous for practical use and, on the other hand, the real advantages of (uncorrected) full text will not be appreciated in the right way – as being one important means to index a large collection in additional to the traditional finding aid of an archive.

Finally we asked also for the transcription process itself. Interestingly we got the strongest answers within this section. The statement with the highest emotion in the complete survey was the following:

*Statement 1.12: Many students who study history or similar studies have difficulties to read handwritten documents respectively to correctly transcribe them.*

Agree: 40 Disagree: 0

*Statement 3.9: I myself like to transcribe historical manuscripts, this is a challenging and interesting activity.*

Agree: 30 Disagree: 3

Though the first statement may reflect the general attitude (or sometimes even resentment) of archivists towards the study of humanities at universities, it has to be emphasized that in several personal talks with archivists and historians this observation can be taken as accurately describing the situation of "humanities students". More and more of them are very seldom confronted with handwritten texts and therefore the general capability of reading historical manuscripts decreases. As with many other capabilities it can be seen also in this case that – once someone has learned it – there is a sense of pride attached to the ability to read historical documents and to correctly transcribe them.

### 2.2 Conclusions

Some rather interesting consequences can be drawn from these results. The first is that people who have learned to read historical documents regard such a skill as something very positive. It is the chance to gain insight into very specific and unique documents and the ability to decipher a text is often a source of pleasure. The second is that there is a large number of university graduates who never acquired skills to transcribe handwritten documents and who are therefore unable to use an archival collection for their professional work. In other words, if (young) professionals could receive the chance to be trained for deciphering handwritten text, they would very likely take this up such an opportunity for reasons of both professional enrichment and personal pleasure. An HTR platform, with its standardized situation concerning the transcription of a text, should therefore take the need for e-learning facilities into account.

In drawing some consequences from the survey, the following user needs can be stated:

1. Digitisation of archival material is important, but the full impact is not seen from this surveyed user group.
2. Digitisation is seen as an "add on" but not as being the core activity of an archive.
3. The making available of digitised material is generally agreed on, but may be decided on a case-by-case basis.
4. Outsourced transcription platforms are seen as interesting model, but it is unclear if they will really be supported.
5. Contributions of volunteers are appreciated although not very much taken into account for the core work of archivists.

6. HTR technology is confronted with unrealistic expectations.
7. The advantages of raw, uncorrected text from HTR processes are not really acknowledged.
8. The capability of humanities students to read and transcribe handwritten texts is harshly criticised.
9. Transcription itself is regarded to be an challenging and interesting activity.

## 3. Pilot co-operation with Bozner Stadtarchiv

In spring 2013, whilst investigating the availability of interesting sources for the ground truth production phase of Work Package 2 (Data management), it was discovered that the Bozner Stadtarchiv in South Tyrol had already digitised 30,000 pages of its "Ratsprotokolle" – the minutes of the regular meetings of the town's council. Even more important was the fact that this digital collection was already in the process of being made available to the public via the Internet and this was being done under a Creative Commons license.

After several contacts with the head of the archive, Hannes Obermair, it was decided to officially cooperate with the archive as a partner in the pilot for our HTR platform for content providers. In this arrangement it is hoped that users who have already proved to be open towards modern technology will also be receptive to the development of the HTR platform.

The cooperation will investigate the following issues:

- Use of the digitised collection of the Bozner Ratsprotokolle as one of the test collections in the project.
- Involvement of researchers from the University Innsbruck as well as volunteers of the archive in the transcription platform.
- Regular feedback on the platform from experienced users/transcribers.
- Technology review and meetings with the technical provider of the archive in order to investigate the best methods for (technical) integration of the HTR platform into the technical environment of archives.
- Evaluation of the standard formats (METS, TEI, PDF, etc.) and their usefulness for digital archives.
- Discussion of a business model for sustaining an HTR platform.
- Collaborative organisation of a conference/workshop were the findings of the pilot project are presented to the archives community.

The cooperation was fixed in June 2013 and made public at the BOhisto website of the archive.

This cooperation will permit on the one hand the chance to test the HTR platform under "real world" requirements and to better understand user needs from the actual implementation of the software. On the other hand, the cooperation allows such testing to be done in a friendly and supportive environment. It must to be emphasized that the Stadtarchiv Bozen does not get any direct advantage from this cooperation nor will it be partner of the tranScriptorium project. Since it acts as a "volunteer", it is therefore independent from any project related constraints or limitations.

## 4. Appendix: Survey (in German)

# Hintergrund

- Im Rahmen des EU Forschungsprojekts transcriptorium sollen die Wünsche und Bedürfnisse von Archiven sowie Archivbenutzern im Zusammenhang mit der Transkription von Handschriften erhoben werden. Das Projekt wird von der Technischen Universität in Valencia koordiniert, das Institut für Germanistik der Universität Innsbruck ist als Forschungspartner beteiligt.

# Fragen an die TeilnehmerInnen

- Bitte beantworten Sie die Fragen spontan und ohne lange darüber nachzudenken!
- Uns ist klar, dass es sich hier um keine repräsentative oder standardisierte Umfrage handelt, sondern es geht uns um ein „Stimmungsbild“, das dazu beiträgt, dass wir die Anforderungen im Archivbereich für die Digitalisierung und Transkription von Archivmaterialien besser verstehen können.
- Die Daten dienen ausschließlich projektinternen Zwecken und werden nicht veröffentlicht.

# 1. Allgemeine Fragen zur Digitalisierung

| | Stimme völlig zu | Stimme eher zu | Stimme eher nicht zu | Stimme überhaupt nicht zu |
|---|---|---|---|---|
| Die Digitalisierung steht im Archivwesen noch ganz am Anfang. | | | | |
| Archive sollten die Voraussetzungen schaffen, damit ForscherInnen möglichst einfach Archivalien transkribieren können. | | | | |
| Die Massendigitalisierung von Archivgut schafft mehr Probleme als sie löst. | | | | |
| Die massenhafte Transkription von Archivmaterialien ist nicht die Aufgabe von Archiven. | | | | |
| Sollte ein Archiv jemals komplett digitalisiert sein, dann verliert es seine Benutzer vor Ort. | | | | |
| Die Digitalisierung ist die wichtigste Aufgabe von Archiven für die Zukunft. | | | | |
| Die Transkription von archivalischen Quellen sollte von ForscherInnen im Rahmen ihrer Arbeit erledigt werden. | | | | |
| Anders als bei Google Books wird niemals ein privater Anbieter Interesse an der umfassenden Digitalisierung von Archivbeständen besitzen. | | | | |
| Es gibt sicherlich viele HeimatforscherInnen/ChronistInnen die bereit wären, bei der Transkription von Quellen aus ihrem Landes- oder Ortsarchiv mitzuhelfen. | | | | |
| Wenn eine Software in der Lage wäre eine automatische Handschriftenerkennung durchzuführen, dann wären 50- 70% Wortgenauigkeit schon ausreichend. | | | | |
| Ein Archiv sollte digitalisierte | | | | |

| | | | | |
|---|---|---|---|---|
| Materialien nur ausnahmsweise für die Öffentlichkeit im Internet frei geben. | | | | |
| Viele StudentInnen, die Geschichte oder ähnliche Studien absolvieren, haben Schwierigkeiten, handschriftliche Akten zu lesen bzw. korrekt zu transkribieren. | | | | |

| | Stimme völlig zu | Stimme eher zu | Stimme eher nicht zu | Stimme überhaupt nicht zu |
|---|---|---|---|---|
| Ich kann mir gut vorstellen, dass ein Archiv digitalisierte Quellen des 18. oder 19. Jahrhunderts an eine Plattform übergibt, damit interessierte ForscherInnen und andere Personen online eine Transkription durchführen können. | | | | |
| In wenigen Jahren werden wir nicht mehr darüber reden, welche Quellen digitalisiert werden sollen, sondern nur noch darüber, welche Quellen NOCH NICHT digital verfügbar sind. | | | | |
| Die Beschlagwortung von Archivalien durch Benutzer, kann niemals der Arbeit eines Archivars gleichgestellt werden. | | | | |
| Die Langzeitarchivierung digitalisierter Archivmaterialien ist ein gelöstes Problem. | | | | |

Außerdem möchte ich zum Thema Digitalisierung noch festhalten, dass

(bitte ev. Rückseite benutzen)

# 1. Transkription von Archivmaterialien

Stellen Sie sich bitte folgendes Szenario vor.

- Ein Archiv digitalisiert 200.000 Seiten Gerichtsakten aus den Jahren 1600 bis 1900.
- Die Digitalisate werden im Rahmen einer vom Archiv betriebenen digitalen Bibliothek online frei zugänglich gemacht.
- Gleichzeitig liefert das Archiv die digitalisierten Seiten an eine Serviceplattform. Dort werden die Seiten automatisch mittels einer Handwritten Text Recognition Software texterkannt. Die Wortgenauigkeit beträgt rund 50-70%.
- Das Archiv fordert interessierte ForscherInnen, StudentInnen und andere Personen auf an der Korrektur der Transkription mitzuwirken. Das geschieht, indem sich diese Personen bei der Serviceplattform anmelden und dann – z.T. mit Unterstützung der Software – die Transkription korrigieren.
- Die Personen, die die Transkription erstellen, können ihre und die Arbeit aller anderen auf ihrem eigenen Computer speichern und für ihre eigenen Zwecke verwenden.
- Das Archiv erhält die automatisch erkannten und korrigierten Texte in einem Standardformat zurückübermittelt. Damit kann in der digitalen Bibliothek des Archivs auch im Volltext nach Archivmaterialien gesucht werden.

Wie beurteilen Sie dieses Szenario?

| | Stimme völlig zu | Stimme eher zu | Stimme eher nicht zu | Stimme überhaupt nicht zu |
|---|---|---|---|---|
| Das scheint ein interessantes Modell zu sein und ich würde eine Verwirklichung begrüßen. | | | | |
| Das kommt mir zu umständlich vor, besser wäre eine lokale Software, die jeder bei sich laufen lässt. | | | | |
| Ich kann mir nicht vorstellen, dass „archivfremde Leute“ einen nennenswerten Beitrag bei der Transkription von Handschriften leisten würden und damit ist dann der ganze Aufwand zu groß. | | | | |
| Ich fände es gut, wenn sich ein Archiv nicht mit einer komplizierten Software | | | | |

| | | | | |
|---|---|---|---|---|
| zur automatischen Erkennung herumschlagen müsste, sondern nur die Digitalisate liefert und die Transkriptionen in standardisierter Form zurückbekommt. | | | | |

| | Stimme völlig zu | Stimme eher zu | Stimme eher nicht zu | Stimme überhaupt nicht zu |
|---|---|---|---|---|
| Die „Auslagerung der Transkription“ mag für Massenquellen funktionieren, nicht aber für hochwertige Quellen aus der frühen Neuzeit. | | | | |
| Für das 20. Jahrhundert kann das nicht funktionieren, da das Persönlichkeitsrecht viel zu restriktiv ist und deshalb in den meisten Fällen die Digitalisate nicht öffentlich gemacht werden können. | | | | |
| Ein Dienstleister in Vietnam oder China könnte die Arbeit einfacher, rascher und billiger erledigen. | | | | |
| Ich bin überzeugt, dass auf lange Sicht ein Archiv von der ehrenamtlichen Mitarbeit in Form von Mitarbeit bei der Transkription und Vergabe von Metadaten profitieren könnte. | | | | |

# 2. Transkribieren

Zum Schluss noch einige Fragen zu Ihren persönlichen Erfahrungen beim Transkribieren handschriftlicher (historischer) Texte und Quellen.

| | Stimme völlig zu | Stimme eher zu | Stimme eher nicht zu | Stimme überhaupt nicht zu |
|---|---|---|---|---|
| Ich verwende MS Word bei der | | | | |

| | | | | |
|---|---|---|---|---|
| Transkription handschriftlicher Texte. | | | | |
| Ich habe bei der Transkription handschriftlicher Quellen große Schwierigkeiten beim Lesen. | | | | |
| Ich bin mit dem Umgang von TEI und XML bei der Transkription handschriftlicher Texte sehr gut vertraut. | | | | |
| Eine Buchedition einer handschriftlichen Quelle ist immer noch die beste Form der wissenschaftlichen Wissensvermittlung. | | | | |
| Ich bin mit dem Umgang von Oxygen bei der Transkription handschriftlicher Texte bestens vertraut. | | | | |
| | Stimme völlig zu | Stimme eher zu | Stimme eher nicht zu | Stimme überhaupt nicht zu |
| Die diplomatische Transkription mag für KorpuslinguistInnen interessant sein, für HistorikerInnen sind normalisierte Versionen besser zu benutzen. | | | | |
| Es gibt eindeutige und klar definierte Standards, die bei jeder Transkription handschriftlicher Texte beachtet werden müssen. | | | | |
| Die zügige Transkription handschriftlicher Quellen erfordert jahrelange Übung. | | | | |
| Ich transkribiere sehr gerne historische Handschriften, das ist eine anspruchsvolle und interessante Tätigkeit. | | | | |

# 3. Angaben zur Person

Alter: unter 20 / zwischen 20 und 30 / zwischen 30 und 50 / über 50

Geschlecht: w / m

Ausbildung:

Beruflicher Status: