CROWD-SOURCING IN THE HUMANITIES: MAIN ISSUES AND QUESTIONS

Stuart Dunn and Mark Hedges
Department of Digital Humanities, Kings College London

Correspondence: stuart.dunn@kcl.ac.uk

www.humanitiescrowds.org

1. Introduction

1.1 Background

This report summarises a workshop held at King's College London on May 29th-30th, and seeks to frame the main questions and areas of interest being addressed by the AHRC-funded Crowd-Sourcing Scoping Study (www.humanitiescrowds.org). Crowd-sourcing projects recently have come to prominence in the physical sciences, with activities such as GalaxyZoo (http://www.galaxyzoo.org) achieving high profiles and impact both inside and outside the academy. Zooniverse (https://www.zooniverse.org), which facilitates the collection of tools, methods and applications from which other projects have emerged, has been successful in building a scientific community around the use of crowd-sourcing technology. It has achieved this by enabling the sharing of good practice, identifying common technological solutions, and providing a platform for advocacy and promotion. Against this background, a number of projects in the fields of history, literature, music, librarianship and archaeology have emerged in recent years which utilise crowd-sourcing methods of different kinds. As with Zooniverse, a current preoccupation in the digital humanities is to build communities around research questions, and to explicitly avoid the building of communities around technology. We suggest therefore that this frame the definition of crowd-sourcing for the humanities: that it should be understood as an activity which engages people without proactively recruiting them, or prescriptively organising or delimiting their contributions once these are under way.

A major challenge for the humanities as they engage with crowd-sourcing is to keep up with the sheer scale of digital information now available for humanistic research. The issue of the 'data deluge' in the humanities does not need rehearsing here[1], but it there is a general lack of infrastructure and resources to manage humanities data across institutions and communities. There is therefore a risk that individual researchers and research groups will be disempowered. Quite simply, such groups and individuals will find it increasingly difficult to form and control their own research agendas as more and more data is digitised and added to the available corpus. The success of Galaxy Zoo, from which Zooniverse emerged, is neatly summarised on the Zooniverse website:

[1] See, e.g. T. Blanke, M. Hedges and S. Dunn 2009: "Arts and humanities e-science—Current practices and future challenges", *Future Generation Computer Systems,* Vol 25, 4: 474-480.

> "Galaxy Zoo was important because not only was it incredibly popular, but it produced many unique scientific results, ranging from individual, serendipitous discoveries to those using classifications that depend on the input of everyone who's visited the site" (https://www.zooniverse.org/about).

Such clear-cut and definable success factors are not always in evidence in humanities crowd-sourcing projects, but it is likewise arguable that such success factors are not always in evidence in – or applicable to - the humanities themselves. The main challenge for humanities crowd-sourcing might therefore be characterised as being to replicate the scale and success of citizen science projects, whilst retaining control of the research agenda and questions, and doing this in the presence of the 'softer' and more incremental success factors by which the success of humanities enterprises are judged.

2.1 Projects represented and/or discussed

Transcribe Bentham (UCL) Transcribe Bentham is an award-winning participatory project based at University College London. Its aim is to engage the public in the online transcription of original and unstudied manuscript papers written by Jeremy Bentham (1748-1832), the great philosopher and reformer. We would like to encourage all those who have an interest in Bentham or those with an interest in history, politics, law, philosophy and economics, fields to which Bentham made significant contributions, to visit the site (from http://www.ucl.ac.uk/transcribe-bentham/about/).

SoundMap (British Library), sought to 'map the soundscape' of the UK, by allowing people to self-select ambient sounds to record and archive with the BL.

Year of Shakespeare (Shakespeare Institute / Shakespeare Birthplace Trust), which sought to collect online expert reviews of over 70 Shakespeare performances in the Globe Festival. This raises questions of who has the expertise/authority to provide such review, and how this should be expressed on an online platform. The general public were able to leave comments on the reviews.

Research on the use of Bodiam Castle (University of York), which used game-like visualisation to approach the question of Bodiam Castle's original function. Was it a private house or a military establishment?' The project allowed 'gamers' to make pathways through the building, and then aggregated and visualised those pathways in order to make inferences about which sort of function the building's structure lent itself to. There were no rewards whatsoever for participants.

Research on medieval graffiti in Durham Deanery (University of York) This project got 'causal collaborators' to trace the outlines of a mass of medieval graffiti discovered under the plaster in the deanery of Durham Cathedral. The main motivation for participants was the thrill of discovery.

HistoryPin (We Are What We Do) seeks to build a community by allowing the public to associate historic photographs with an online map, and enabling various kinds of comparison of the historic scenes depicted with those of the present day. This promotes contact and allows sharing experiences between the present and previous generations.

British Library Georeferencer's purpose was to "geo-enable" historic maps by asking participants to assign spatial coordinates to digitised images of maps. This task would have been too labour intensive for BL staff to undertake, so it was exposed to crowd-sourcing. Once digitised and

georeferenced, the maps can be viewed using online geographic technologies, and are geographically searchable due to the inclusion of latitude and longitude coordinates in the metadata. There was an element of instant gratification, in that users can see immediately the results of their work. The project had a 'citizens' forum' tab, which proved important for generating a sense of community among the participants. 725 maps were georefernced between 13 and 18 February 2012 by around 90 participants. Social media was considered to be key. The data quality was very good.

Old Weather (Met Office). Old Weather works with documentary data from the log books of historic ships, extracting weather/climate data in order to model the history of the climate in this period, which is extremely difficult to measure by other means. 300 years worth of log books at TNA are being processed in this way. The accuracy rate was 97%. A record is only frozen when two transcribers have come to an agreement on it.

What's the Score (University of Oxford) The principal aim of the project is to investigate a cost-effective approach to increasing access to music scores from the Bodleian's collections, to be achieved by a combination of rapid digitization and the creation of descriptive metadata through crowd-sourcing (from http://www.bodleian.ox.ac.uk/bodley/library/special/projects/whats-the-score). See the crowd-sourcing site at http://www.whats-the-score.org, and the delivery site http://scores.bodleian.ox.ac.uk.

2. Current state of the art in humanities crowd-sourcing

2.1 Key terms

Crowd-sourcing: The term was originally coined in 1996 in an article in *Wired* by Jeff Howe entitled *The Rise of Crowdsourcing* (http://www.wired.com/wired/archive/14.06/crowds.html). Most discussions of crowd-sourcing treat the term as being distinct from the theory of the 'Wisdom of Crowds' as advanced by James Surowiecki in 2004 (in *The Wisdom of Crowds: Why teh Many are Smarter than the Few*, Abacus 2004). This is because academic crowd-sourcing has, historically, consisted of activities conceived and directed for particular purposes, whereas wisdom of crowds, or 'collective intelligence', deals with the comparative facility of large groups of people to make decisions or address problems. For humanities crowd-sourcing, the situation is far more complex, since the information involved is typically fuzzy and/or extremely qualitative (although qualitative is a term from social science, and should always be used with extreme caution in humanities contexts), and it is more likely to speak directly to emotional and emotive responses in its recipients.

In order to be considered a humanities crowd-sourcing activity, the activity should have:

a) Some clearly defined humanities direction and/or research question. The question could be posed/designed by an academic team, or by an individual with particular knowledge and/or interests.

b) The potential for a group with unregulated membership to transform primary material

c) There needs to be some meaningful and replicable way of breaking the workflow down in to separate tasks. In general, this is relatively easy for certain kinds of project such as

d) It should be scalable, both to different volumes of data and different levels of participation.

A crowd-sourcing project t should have the capacity to allow large numbers of people to be involved, even if only a very small number of contributors end up being actively engaged (which is often the case). Indeed, most of the humanities crowd-sourcing projects represented at the meeting reported that a very small number of contributors generally do a very large percentage of the work. The point is that the body of contributors is self-organising and self-selecting, and there is not be a central(ised) recruitment process. As Trevor Owens has written:

> Most successful crowdsourcing projects are not about large anonymous masses of people. They are not about crowds. They are about inviting participation from interested and engaged members of the public. These projects can continue a long standing tradition of volunteerism and involvement of citizens in the creation and continued development of public goods (http://www.trevorowens.org/2012/05/the-crowd-and-the-library).

Participant: For the purposes of this report, people who contribute time, effort or input to a crowd-sourcing project are termed *participants*, rather than contributor. Some projects prefer the term 'volunteer'. It is important that participants are seen as collaborators and not as 'free labour'. However, as Mia Ridge has pointed out the humanities, and especially the fields of Cultural Heritage and museums, have long existing traditions of altruistic participation and volunteerism, meaning that crowd-sourcing projects can be seen as an extension of this (http://openobjects.blogspot.co.uk/search?updated-max=2012-07-13T20:21:00%2B01:00&max-results=5).

Contribution: A contribution is some particular intervention that a project participant has made. Depending on the content involved, a contribution can be very easy or very difficult to quantify/reify. A section of text which has been transcribed is a contribution, however the dispersed, and often anonymous, nature of crowd-sourcing means it can be difficult to apportion responsibility for outcomes, such as the successful georeferencing of an entire digital map. It should also be noted that the critical mass needed to meaningfully create something in a crowd-sourcing project is very difficult to identify in a generalised way. Also, 'contributions' can come in forms other than direct contribution to a project's workflow or content generation. For example they can also be made via contributions to a project discussion forum or blog.

Amateurism and **professionalism:** Both terms are extremely problematic in crowd-sourcing.

Crowd-sourcing is distinct from production of user-generated content (UGC) on platforms such as Google Earth, although such platforms can undoubtedly be used as components of crowd-sourcing projects. Equally, so-called transactional data, which is the harvesting and analysis of information about people's (usually online) activities, is not considered here to constitute crowd-sourcing. In

short, it is crowd-sourcing if some additional value is added to the whole dataset as a result of public participation.

It has also been stated that crowd-sourcing is distinct from 'crowd-funding', where large groups of people are invited to fund projects by individually contributing small sums of money, via sites such as kickstarter. However, it has become apparent that there are likely to be some crossovers where people who contribute financially to such projects are also offered the opportunity to get involved in some way. An intellectual contribution and a financial contribution are not mutually exclusive.

3. Who are 'the crowds'?

3.1 Homogeneity and heterogeneity

There is a perception that crowd-sourcing is a democratising phenomenon in academia, but in reality there are usually some barriers to participation, and certain skills that need to be picked up for an individual to participate. Participant groups in crowd-sourcing projects should therefore not be considered a 'core of experts with a penumbra of amateurs'. This, arguably, is a conception of 'the crowd' which has been inherited from the sciences, where there are different traditions of volunteerism and community involvement.

Generally, participants considered that it was very unusual for participants to switch between projects, or to work on more than one. This suggests one of two possibilities: a) humanities crowd-sourcing appeals to interest communities, rather than to people who contribute for the sake of contributing, or to feel that they are making a useful contribution to the common good; or b) a particular methodology of participation or interface appeals to an individual, and this draws them to the project. This is likely to appeal to individuals. It should be noted however that this does not always apply. The Trove project, which used crowd-sourcing to digitise an historic corpus of Australian newspapers found that standard motivational factors such as pleasure, short and long term goals, concentrating on outcomes, a sense of being trusted and respected, and the creation of challenges featured alongside subject-specific factors such as Australian and family history, or contributing to a 'worth cause'; however by the largest proportion of contributors were family historians, who wanted to investigate the material and help make it available to their peers.[2] At the other end of the scale completely is the project (http://www.digitalkoot.fi/en/splash) which relies on a 'gameified' interface to entertain participants while they contribute.

3.2 Changes/transformations in the crowd as a result of crowd-sourcing

Some projects noted that their participants had become far more IT literate as a result of their participation, and others reported that participants had built up significant areas of expertise. For example, some 'citizen scientists' in the Old Weather project are now expert in specialised aspects of naval history. This can contrast with the interests of the project team: in this case, for example, the team is interested in weather history, whereas others are interested in ships. This has two important implications: firstly, it must be recognised that humanities material is typically multi-dimensional and liable to provoke different kinds of interest in different people; secondly this confirms that point made above that the 'amateur/professional dichotomy typically in evidence in the sciences is far more permeable.

[2] R. Holley 2010: Crowdsourcing: how and why should libraries do it? *D-Lib Magazine* 16: 1-19.

4. Roles and levels of participation

4.1 Existing practice

Usually a crowd-sourcing project will be set for a particular task, and will have only one role for participants. This role will be mediated by some kind of interface, such as an editing window. However, the complexity of material in the humanities means it may be necessary to create multiple roles, which are likely to have differing levels of commitment in terms of time and effort. The Global Shakespeare project for example asks academic reviews to do the same task, but the public an be involved by leaving comments and responses to the reviews. This project is in its infancy however, and it was noted that the early participants are those who are contributing are generally academics who already blog.

Most projects noted that a small core of participants were most active, but it was noted that it is essential to engage a range of casual as well as 'uber' users to draw upon, so that the 'crowd' participating can self-organize.

The MarineLives project (http://www.marinelives.org), which is in its design stage (and is run by a private individual with no institutional affiliation), seeks to digitise, link and enhance the records of the High Court of the Admiralty, is experimenting with defining different roles, including project associate, project facilitator. Again, this has a parallel with roles of contributors to citizen journalism sites (as distinct from the websites of media organizations and newspapers), which adapts roles familiar from print media. One recent example of this is the Marlborough News online, which seeks to broaden the role of professional journalist by recruiting 'community stalwarts and would-be young journalists' to participate for a common good, rather than for a professional fee (see http://www.marlboroughnewsonline.co.uk/about-us).

5. What are the current objectives of humanities crowd-sourcing?

5.1 Open ended versus closed questions

Humanities crowd-sourcing projects can be divided in to three very broad categories: those which seek to answer humanistic research questions, those which seek to digitally process existing content (although this in itself is in itself a complex field with a problematic terminology – see http://englishplacenames.cerch.kcl.ac.uk/?p=65), and those which seek to create completely new content.

These goals are achieved using a variety of methods. Mia Ridge has compiled a typology of different types of crowd-sourcing objective for Cultural Heritage:

- Tagging
- Debunking
- Recording personal stories or histories
- Linking
- Stating preferences
- Categorising
- Creative responses

(http://openobjects.blogspot.co.uk/2012/06/frequently-asked-questions-about.html)

If one were to add 'transcribing', this would form a comprehensive set of labels, although many projects combine two or more of them. Old Weather, for example, relates directly to transcribing, but it is also concerned with scaling and linking data, describing (for example) the global contexts of heat waves in London, and giving rise to detailed historical analysis and narrative building. It was noted that some ships' logs had been previously transcribed, but this was done for a certain purpose, and thus did not include all the standardized information that the project is interested in.

Another objective is, therefore, to facilitate the linking of digital datasets and databases that people have already created. This lays stress on the importance of linking with existing groups, and could include the use of social media. A very small number of people are willing to go to the effort to share their data for the sake of sharing it, but putting it in context, cleaning it up and adding metadata might help to motivate people. This makes gaps in 'global' datasets more apparent, and this in turn makes them easier to address in a systematic way.

A key distinction in any project is whether the participant has any scope to influence the direction of the research, and the questions asked. In most transcription projects, for example, they will not be able to exert any such influence by participating as a transcriber only (even though this may in itself be a skilled task – for example, in the case of the Transcribe Bentham project, the handwriting is difficult to decipher). However, as noted above they might be able to contribute in this way by communicating with the team running the project, or by contributing to discussion fora. This however raises a separate set of project design issues, in that there needs to be a means of dealing with conflicts, dissent and controversy.

In a seminar given by Dr. Charlotte Tupman in the 2012 Digital Classicist seminar series, it was argued that an entirely separate objective of crowd-sourcing can be simply to involve the wider public in research, thereby increasing its profile and impact (see http://www.digitalclassicist.org/wip/wip2012-06ct.html).

5.2 Objectives for whom? How do projects achieve them?

If the multi-dimensionality of humanities research material is to be accessible to people with differing interests in it, then the interface must be well designed and presented. It is important to use standard, interoperable mark up formats, e.g. TEI and KML, depending on the nature of the resource to be crowd-sourced, and the kind of activity involved.

This basic need must be set in the context of the object-orientation which the post-Web 2.0 internet encourages. For example, the British Library's Wikipedia project is seeking to create articles dealing with all of the Library's most significant objects and collections. It is therefore all very object-oriented, and this reflects the structure of the collections. The question of how we object-orient data comes up again and again in the (digital) humanities, be it XML markup, textual chunks, CIDOC-CRM objects in museums, etc. One key objective of humanities crowd-sourcing, although expressed differently by different projects, is therefore to leverage this increasingly sophisticated objectification of humanities material in the digital world and promote public interaction with it.

5.3 Project design and identification of tasks

It is generally true that the 'deeper' the task, the fewer individuals will engage with it, since there will be higher cognitive barriers to participation. Where a task requires any combination of interpretation, *a priori* knowledge or significant commitment of time or effort, then it needs to be packaged by some intuitive means into micro and/or macro-tasks. This is analogous to the work done by the e-Science for the Study of Ancient documents project at Oxford, which sought to identify and map the activities of professional papyrologists as they deciphered manuscripts. In addition to standard software underpinning crowd-sourcing, such as Zooniverse, there needs to be a commonly understood set of standards to define usefully, but at the same time be sufficiently generic so as not to constrain, the cognitive tasks listed above. Such 'humanities crowd-sourcing primitives' will be critical to the realisation of future projects in the area.

5.4 Sustainability

A key question is what happens to crowd-sourced data after the project ends? This is an under-explored question, which raises technical, legal and ethical challenges. Much of this content is dynamic, so cannot be archived as static pages. Context is provided by the fact that many people proved suspicious of having their websites archived by the British Library's Internet Archive, which experienced a significant refusal rate among website owners it approached.

6. Incentive and reward in humanities crowd-sourcing

Motivation is a key question for all crowd-sourcing projects. Many participants are interested in the process, with the eventual outcome of the project being secondary. It was noted that volunteer motivations are very difficult to track, although there have been some quantitative and qualitative studies undertaken in the past (e.g. J. Raddick et al, 'Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers', *Astronomy Education Review* 9, 2010).

6.1 Competition

Competition is one possible motivation for people to participate in crowd-sourcing projects, although it is worth noting that very few participants studied in the qualitative research cited above admit to being motivated by competition with each other. For most projects it is possible to possible to track individual participants' contributions, and acquire statistics on who contributed the most etc. This can enable projects establish 'leader boards', indicating which participants have made the biggest contributions (in whatever terms the project is working with). In the case of the BL Georeferencer project for example, it displayed the handles of the users who processed the most maps. The 'winner' was invited to meet the BL's head of cartography, and it was considered that this kind of contact with such a prestigious institution was itself highly valued by the participant community. BL staff also felt that the project made the participants feel that they had a stake in the BL itself, and were part of the community it represents.

The nature of the material however means this can become complicated, for example where the kinds of content are not consistent. For example, in the BL Georeferencer project, some maps are more complex than others, so there was relatively little meaning in comparing the effort needed to georeference different ones. Another possibility raised was that different aspects of the same project could have different leader boards, thus reflecting a model of 'diffused competition'. As a solution however, the notion of encouraging competition should be qualified by the need not to

exclude potential participants who are not, by nature, competitive people, yet may have valuable knowledge or effort to bring.

Another qualification with using competition as a means of encouraging participation is the extent to which it encourages speed and volume at the expense of quality and care. As noted above, there is also an issue of how conflicts in participants' contributions are to be handled. This is likely to be especially so where creative/interpretive content is being gathered. This kind of research is also less likely to lend itself to the leader board type approach outlined above.

Where social media-like content is being created, it is likely that recent approaches to quantifying emotion, such as sentiment analysis, will have a role to play in gauging motivation and impetuses to contribute[3].

Another possible way of leveraging competition in crowd-sourcing projects is via gamification. Serious games are an emergent field in the digital humanities, and the kinds of competition that games engender can be a powerful means of encouraging people to complete tasks. The potential of this approach is particularly evident in the Bodiam Castle project, although this had a greater emphasis on visualisation than on competition.

6.2 Collaboration among participants

As noted, many crowd-sourcing projects have discussion fora attached to their sites, and this is often a valuable means for participants to communicate with one another. Some projects, such as Transcribe Bentham, actively encourage participants to report issues by these means.

6.3 Use of social media

It was noted that there is a large body of work on the deployment of social media in, e.g. political and advertising campaigns. If a humanities project has content which is capable of creating a 'spike' of interest, then this could be leveraged.

7. Other means of community engagement with humanities communities

7.1 Crowd-funding

Funding of projects is always an issue, and there are examples of humanities research areas, especially archaeology, which have turned to so-called 'crowd-funding', where members of the public are invited to support a project financially, usually in return for some kind of privileged access to the project or its data. While it has been successful in raising money in the short term (the Maeander project is a good example – see http://crmnews.org/2011/06/21/an-unusual-source-of-funding/), this approach clearly cannot be sustained satisfactorily. There are also implications for removing a project from the institutional framework of overheads, indemnity, ethical clearance etc; and whether the outcomes of a project funded in this way could be subject to academic peer-review.

[3] E.g. M. Thelwall, K. Buckley and G. Paltoglou 2011: Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology* 62.2: 406-418.

There are, examples of crowd-funding leading to excellent publicity: for example the New York Shakespeare Exchange, which needed $15000 funding to promote ‘Shakespearisation’ of New York.