Crowd-Sourcing Scoping Study

Engaging the Crowd with Humanities Research

Stuart Dunn and Mark Hedges

Project Administrator: Angela Waplington

Centre for e-Research, Department of Digital Humanities
King’s College London

www.humanitiescrowds.org

All URLs cited work as of 12th November 2012

A project of the AHRC Connected Communities Theme
Executive Summary

This project sought to establish a credible definition for, and the current state of the art of, crowd-sourcing in the humanities. The questions included what the humanities have learned from other research domains, where crowd-sourcing is being exploited, what the results are, why academics are motivated to undertake such activities, and why members of the public are willing to give up their time, effort and knowledge for free. We conducted a survey, supplemented by a set of follow-up interviews, of contributors’ motivations, which received 59 detailed responses with qualitative and quantitative information about why people contribute to humanities (see Appendix A). The project identified and reviewed 54 academic publications of direct relevance to the field, and a further 51 individual projects, activities and websites which document or present some application of humanities scholarship making use of crowd-sourcing (see Appendix B). Two workshops were held, one for academics making use of crowd-sourcing, and one for contributors to those projects.

Academics in the humanities undertake crowd-sourcing projects for a variety of reasons: to digitize content, to create or process content, to provide editorial or processing interventions, and so on. Judging the current value of crowd-sourcing in the humanities is therefore extremely difficult, even before issues of trust, reliability and academic rigour are accounted for. However, one common factor is that humanities crowd-sourcing succeeds where vibrant and interacting communities of contributors are created. Whilst the motivations of crowd-sourcing contributors are every bit as diverse as those of academics, passion for the subject (a characteristic shared with academics) is the dominant factor which draws them together into communities. These communities develop and perpetuate internal dynamics, self-correct, provide mutual support, and form their own relationships with the academic world. Despite the great diversity of humanities crowd-sourcing, it is possible to observe patterns in which such communities thrive: these patterns are dependent on the correct combinations of asset type (the content or data forming the subject of the activity), process type (what is done with that content) task type (how it is done), and the output type (the thing produced) desired. In this report, we propose a high-level typology which describes different instances of each of these, and identifies the combinations that are, on present evidence, most successful in achieving projects’ aims.
1. Introduction

1.1 Background

Crowd-sourcing, the process of leveraging public participation in or contributions to projects and activities, is relatively new to the academy, and even newer to the humanities. However, at a time when the web is simultaneously transforming the way in which people collaborate and communicate, and merging the spaces which the academic and non-academic communities inhabit, it has never been more important to consider the role which public communities - connected or otherwise - have come to play in academic humanities research. The purpose of this report is to present a review of literature of crowd-sourcing applications in the academic humanities domains, to assess its impact and development, to consider the motivations and aspects of community of those who choose to participate, and to present a typology which captures the different approaches which have emerged.

It should be emphasized that this report, the result of a nine-month study, does not claim to be comprehensive: there are bound to be important projects, publications, individuals and activities that we have missed. Just as inevitably, there is a strong UK and Anglophone focus on the activities studied. Non-inclusion in this document does not reflect any lack of importance or interest in an activity. We have come across many fascinating crowd-sourcing projects beyond Western Europe and the US, and hope that this report will act as a catalyst to developing a broader community of discussion and application beyond the projects and activities that we have become aware of while preparing it.

The study consisted of four components: a literature review of academic research in the humanities which has drawn on crowd-sourcing, as well as papers detailing research into crowd-sourcing itself as a method; two workshops held at King’s College London in May and October 2012 facilitating discussion between, respectively, academics in the humanities who use crowd-sourcing, and members of the public with records of contributing to such projects; a set of interviews with both academics and contributors, and an online survey of contributors exploring their backgrounds, histories, and motivations for participating. We also conducted an extensive web crawl, identifying projects using crowd-sourcing that may not yet have produced a tangible academic outcome; tools that facilitate crowd-sourcing, and relevant blogs and Twitter feeds. The projects and other activities identified from this are presented in Appendix B.

In order to explore the issues fully, we considered it necessary to conduct not just a review of research but also of research activity. The aim was to identify not just the kinds of humanities research that are engaging with wider communities, and how they are engaging, but also to look at how communities of contributors form (or are formed) and develop in the course of participation, and how notions of community matter both to them and to the academics running such projects.

1 In this report we follow the convention of hyphenating ‘crowd-sourcing’; other authors use ‘crowdsourcing’ or ‘crowd sourcing’. In quotations, we preserve the original form.
This is a complex and wide-ranging area, and it is necessary at the outset to define terms and the boundaries of the review. Public involvement in the humanities can take many forms – transcribing handwritten text into digital form; tagging photographs to facilitate discovery and preservation; entering structured or semi-structured data; commenting on content or participating in discussions, or recording one's own experiences and memories in the form of oral history – and the relationship between the public and the humanities is convoluted and poorly understood.

This diversity presents two immediate challenges for a review of crowd-sourcing as a research method. Firstly, in purely semantic terms, where should the boundaries of what is considered to be crowd-sourcing lie? And secondly, since humanities crowd-sourcing is in its very early stages, there is relatively little academic literature dealing with its application and outcomes to allow any firm judgements to be made about its potential to produce academically credible knowledge. Given this lack of evidence, we therefore do not seek to make value judgements on any individual cases, and we stress that equally this report does not seek to evangelize or promote crowd-sourcing as a method. It simply seeks to identify what, on present evidence, seems to work and what does not. Moreover, this underlines the need to examine other, less formal, sources of information, such as blogs and interviews, and emphasises that at this early stage, it is just as important to consider the academic validity of processes as well as outcomes.

2 Terminology and typologies

2.1 Analysis of prior research on terminology and typologies

The term crowd-sourcing is frequently used as a convenient label for a diverse range of activities. It was originally coined in 2006 in an article in Wired by Jeff Howe entitled The Rise of Crowdsourcing (Howe 2006). In this article, Howe draws a parallel between businesses farming out labour to cheaper markets in the developing world, and businesses utilising ‘the productive potential of millions of plugged-in enthusiasts’, with similar reduction in labour costs. In recent years, academics have come to use the power of the crowd to achieve research aims.

As a method of undertaking academic research, however, the term ‘crowd-sourcing’ is problematic. It is certainly less easy to define than the analogous term ‘citizen science’, which is commonly understood to refer to activities whereby members of the public undertake well-defined and (individually) small-scale tasks as part of much larger-scale scientific projects (Silvertown 2009), but which, in the past, has also been used to refer to more passive forms of participation such as making available unused CPU power of desktop machines for harvesting by research teams.

Furthermore, most discussions of crowd-sourcing treat it as being distinct from the concept of the ‘Wisdom of Crowds’ as originally advanced by James Surowiecki in 2004 in The Wisdom of Crowds.

---

2 The desk research for this review identified around sixty papers of potential relevance.
3 A good example is the Search for Extra-terrestrial Life project, see Anderson et. al. 2002; see also Anderson and Fedak 2006.
Crowds: Why the Many are Smarter than the Few (2004), which holds that large-scale collective decision-making can be superior to that of individuals, even experts, a thesis that lacks the elements of collaboration around activities conceived and directed for a common purpose that characterise crowd-sourcing as commonly understood. Although academic crowd-sourcing can be about decision making - and we make provision in our typology for such projects - the decisions involved are rarely as neatly packageable as those implied in the world of business, where the ‘good’ or ‘bad’ nature of a decision can be evaluated on the basis of profitability (Brabham 2008).

In their review of the field, Towards an integrated crowdsourcing definition, Estelle’s-Arola and González-Ladron-de-Guevara (2012) identify eight characteristics, distilled from 32 distinct definitions identified in the literature: the crowd; the task at hand; the recompense obtained; the crowdsourcer or initiator of the crowdsourcing activity; what is obtained by them following the crowdsourcing process; the type of process; the call to participate; and the medium. This extremely processual definition is comprehensive in identifying stages which map easily to business processes. For the humanities, the ‘type of process’ is both more significant and more problematic, given the great diversity of processes in the creation of humanities research material. A more task-oriented approach is that of Wiggins and Crowston (2011), who construct a typology for ‘citizen science’ activities. The use of the word ‘science’ (at least in the usual Anglophone sense) confines the activities reviewed (in terms of both the methods and the content) to a particular epistemic bracket which inevitably excludes some aspects of humanities research. Wiggins and Crowston identify five areas of application: Action, Conservation, Investigation, Virtual, and Education (2011). The factors that lead to an activity being assigned to a category are multivariate; and the categories’ identification was based on whether there is an occurrence in a category or not, rather than frequency of those occurrences. The coverage is therefore extremely broad. ‘Action’, for example, covers self-organising citizen groups that use web technologies to achieve a common purpose, often to do with campaigns on local issues.

One widely-quoted set of definitions for citizen science projects was presented by Bonney et al. in their report for Center for the Advancement of Informal Science Education (CAISE), Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education (Bonney et. al. 2009). This divided the field into three broad categories: Contributory projects, in which members of the public, via an open call, contribute along lines that are tightly defined and directed by scientists; Collaborative projects, which have a central design but to which members of the public contribute data, and may also help to refine project design, analyze data, or disseminate findings, and finally Co-created projects, which are designed by scientists and members of the public working together and for which at least some of the public participants are actively involved in most or all steps of the scientific process. This approach shares important characteristics of the ‘task type’ typology facet developed below, in that it is rooted in the complexity of the task being asked of the public, and the amount of effort, initiative and independent analysis required to make a contribution.

Certain subsets of the humanities disciplines have seen some efforts to develop their own typologies for crowd-sourcing, Most notable among these are the cultural heritage and
Galleries, Libraries, Archives and Museums (hereafter GLAM) sectors. This interest is hardly surprising since, unlike most humanities domains, these sectors are inherently public facing, and have long traditions of volunteerism and public engagement. Most museums, especially smaller ones, exist on competitive principles of attracting and engaging audiences to justify their funding; and ‘memory institutions’ such as national libraries and museums have formal duties to maintain access to their collections, both for scholars and the public. Against this background, approaches such as that of Tim Copeland have emphasised the importance of ‘constructivist’ approaches, where the public is encouraged to engage with the interpretation of collections, rather than ‘positivist’ approaches where they are passive recipients of knowledge organized by curators (Copeland 2004). Projects such as UCL's QRator have sought to achieve this by providing iPads as a channel for visitor feedback on the collections of the Petrie Museum (http://www.qrator.org/).

One typology for crowd-sourcing with a special focus on the GLAM sector has been suggested by Mia Ridge in a blog post (http://openobjects.blogspot.co.uk/2012/06/frequently-asked-questions-about.html). In this, Ridge proposes the categories Tagging, Debunking (i.e. correcting/reviewing content), Recording a personal story, Linking, Stating preferences, Categorizing, and Creative responses. Again, these categories imply a processual approach, and are, at least potentially, extensible across different types of online and physical-world content and collections. They are concerned with the type of task that the crowd is being requested to carry out.

An alternative typology for crowd-sourcing in the GLAM domain was developed by Oomen and Aroyo (2011). Their categories include Correction and Transcription, defined as inviting users to correct and/or transcribe outputs of digitisation processes (a category that Ridge’s Debunking’ partially, but not entirely, covers); Contextualisation, or adding contextual knowledge to objects, by constructing narratives or creating User Generated Content (UGC) with contextual data; Complementing Collections, which is the active pursuit of additional objects to be included in a collection; Classification, defined as the gathering of descriptive metadata related to objects in a collection (Ridge’s ‘tagging’ would be a sub-set of this); Co-curation, which is using inspiration/expertise of non-professional curators to create (Web) exhibits (somewhat analogous to Bonney et. al’s co-created projects category, but more task-oriented); and Crowdfunding, or the collective cooperation of people who pool their money and other resources together to support efforts initiated by others (see, e.g. BBC: Flag Fen hosts ‘crowdsourced’ Bronze Age Archaeology dig: http://www.bbc.co.uk/news/science-environment-19192220). Ridge (op. cit.) explicitly rejects crowdfunding as a component of crowd-sourcing.

2.2 Characteristics of crowd-sourcing identified by the review

In the workshop for academics using crowd-sourcing in their work, organised by this review in May 2012 (http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/09/workshop_report1.pdf), four factors were identified that characterise, very broadly speaking, crowd-sourcing for the humanities. These characteristics are not exhaustive, but they usefully highlight commonalities between some of the activities we have observed:
a) The existence of a 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. This seems to preclude some categories identified by some authors, such as the ‘Action’ category of Wiggins and Crowston (2011); and will preclude elements of others, such as the ‘Co-created’ projects of Bonney et. al. (2009). It is suggested here that this characteristic is especially significant, since the academic component of academic humanities crowd-sourcing implies some form of professional rigour. However, we do not assume that the source of that rigour must necessarily originate from a Higher Education Institution.

b) The potential for a group with open membership to transform or add value to primary material or the interpretation of this material. However, a distinction has been made elsewhere between ‘community sourcing’ and crowd-sourcing, with the latter typically dealing in open calls for participation (a key factor for Brabham 2008); and the former being more closed (see, e.g. http://millennialmarketing.com/2010/09/crowdsourcing-vs-community-sourcing).

c) There needs to be a definable task, or some meaningful and replicable way of breaking the workflow down into sets of definable tasks.

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

In the light of these characteristics, crowd-sourcing is considered to be distinct from the production of general user-generated content (UGC) on platforms such as Google Earth, as there is no clearly-defined direction or question, although such platforms could be components of crowd-sourcing projects if such a direction were present. Equally, the harvesting and analysis of so-called transactional data, that is information about people’s (usually online) activities, is not considered here to constitute crowd-sourcing, as whatever additional value is added to the data does not result from public participation.

It has also been suggested 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, and accordingly we have accommodated this aspect in our typology.

While (a) serves to distinguish crowd-sourcing from collaborative activities whose common purpose is primarily social or campaigning, it should be noted that these are not entirely disjoint categories. In particular, recent years have seen the emergence of several SMEs dedicated both to social goals and to the forms of principle behind humanities crowd-sourcing that we have uncovered in this review. An example of this is HistoryPin (http://www.historypin.com), a
website that allows people to georeference historical photographs on a modern map (see below).

3. Levels of participation

Rose Holley, an authority on the use of crowd-sourcing in mass digitization and archives, and formerly project manager of the TROVE project which used the crowd to correct and validate Optical Character Recognition (OCR)-derived digitisations of Australian newspapers (http://trove.nla.gov.au/newspaper), identifies a distinction between crowd-sourcing and ‘social engagement’. She states

Social engagement is about giving the public the ability to communicate with us and each other; to add value to existing library data by tagging, commenting, rating, reviewing, text correcting; and to create and upload content to add to our collections. This type of engagement is usually undertaken by individuals for themselves and their own purposes ... Crowdsourcing uses social engagement techniques to help a group of people achieve a shared, usually significant, and large goal by working collaboratively together as a group (Holley 2010).

She also notes that crowd-sourcing is likely to involve more effort, and by implication to require greater personal commitment to volunteer time for free, than social engagement - which, after all, is an extension of the kinds of online activities - Tweeting, providing content to Facebook etc - that millions do on a daily basis anyway. In a sense, this aligns crowd-sourcing with so-called ‘citizen science’, and implies a level of commitment and participation which goes beyond simple casual interest. Wiggins and Crowston (2011) develop this theme by highlighting a distinction between citizen science and community science, and stating as a key ingredient of the former that it is not self-organizing. As they state:

Citizen science does not represent peer production in the same sense as seen in prior work because the power structure of these projects is usually hierarchical. Furthermore, citizen science is not necessarily “open science,” a term that refers to open source-like practices in formal scientific research settings. Many citizen science projects share data, but may not make the full research process publicly viewable for comment and discussion (Wiggins and Crowston 2011)

A fundamental aspect of citizen science, therefore, is that the research goal is defined by a particular person or group (almost always as part of a professional academic undertaking), with participants recruited through an open call providing some significant effort towards achieving that goal or goals. The motivations that push people to contribute that effort is therefore critical, and this is a crucial distinction between ‘citizen’ projects in the sciences and the humanities. It is certainly true that the two different intellectual traditions embrace, and are embraced by, different kinds of non-academic community. This is particularly so in the domains of the
humanities related to the Cultural Heritage sectors where, as noted above, there are existing processes of interaction between the academy and the public. 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).

The term ‘participation’ itself has connotations of community and interaction. It tends to exclude notions of passivity and serendipity and, most importantly, implies a motivation that stems from interest in the subject and, by extension, community based discussion and exchange around that subject or issue.

A crowd-sourcing project should therefore 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 May 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.

4. Contributor motivations and engagement

4.1 Communities of crowd-sourcing

The foregoing discussion makes it clear that there can be no analysis of the role of crowd-sourcing in the humanities without detailed consideration of the motivations of those who participate in crowd-sourcing projects. This is intimately linked with notions of community, and the sense of community felt by participants. Much of the following section is derived from discussion of the workshop for crowd-sourcing contributors held at King’s College London on 18/10/2012. These contributors were a mixture of those identified by academic colleagues running crowd-sourcing projects (especially Old Weather and the British Library Georeferencer, to whom we record our thanks), and respondents from the online survey, who had ticked the box indicating they were happy to participate further in project activities.

4.2 Previous research

Most studies conclude that most crowd-sourcing contributors do not have a single motivation; the survey conducted for the current project indicated overwhelmingly (79%) that highly active contributors of the kind who responded have both personal and extrinsic motivations; that they do it both for themselves and for others. However in many cases it is possible to identify a single, dominant motivating factor, which is almost always concerned directly with the project or activity’s subject area. In an analysis of 207 forum posts and interview responses for example,
Galaxy Zoo found that the top motivation was an interest in astronomy (39%), a desire to contribute (13%) and a concern with the vastness of the universe (11%) (Raddik et. al. 2010). The first two of these align with our survey’s findings that motivations are both personal and extrinsic. This trend can be found reflected in far more niche areas. A similar study of volunteers on the Florida Fish and Wildlife Conservation Commission (FWC)’s Nesting Beach Survey project found that concern for sea turtle conservation was the overwhelming factor motivating volunteers (Bradford and Israel 2004). Moreover, studies of the motivations of the contributors to academic crowd-sourcing projects have focused on personal interest in the subject area concerned; and opportunities that projects provide to exercise that interest, and to engage with people who share it, without material benefit. Such interest is usually concerned with the outcome, but it can also be in the process, or some combination of both. For example, in her 2009 assessment of the motivation of volunteers to the TROVE project, Holley notes:

We noticed in our communication with text correctors that a large proportion was family history researchers. These people are highly motivated to learn new skills in order to get the information they need. They also have a sense of responsibility towards other genealogists to help not only themselves but other people where possible (Holley 2009).

In general therefore, it may be said that research into crowd-sourcing motivations suggest a clear primary, although certainly not exclusive focus on the subject or activity area, and that this focus can be altruistic, extrinsic or intrinsic. Our workshop of crowd-sourcing contributors also suggested that there is a distinction to be made between abstract interest in a subject area, such as mapping, and highly focused, or even obsessive, interest in a subset of that subject, e.g. maps of a particular period or area, often deriving from a personal or family connection.

4.3 Academic versus commercial crowd-sourcing

There is an obvious distinction to be made between motivations for crowd-sourcing and motivations driven by market economics, which suggest that people will only contribute effort or submit to regulation in return for some benefit, usually material. Academic participants in projects in universities have relatively clear motivations, including, but not limited to, the fact they are materially rewarded by salaries and grants, professional recognition in their field, career advancement, and publication. Most crowd-sourcing projects however do not reward their contributors in material or professional ways (at least not directly), and members of crowds who contribute to crowd-sourcing projects are not subject to discipline (in either sense of the term) or sanction in the way that members of conventionally configured research organizations are. That said, contributors may receive “social” rewards, for example through rankings, increased standing in the crowd-sourcing community, or (in the case of GalaxyZoo) being credited and named in publications. Similarly, contributors may be subjected to social sanctions, such as banning (e.g. removal of pages or blocking of accounts on Wikipedia), which can adversely affect their reputation and enjoyment, and may even in rare cases reflect on their professional standing. However, it is clear that the motivations of academic crowd-sourcing participants are more intrinsic to the activity.
This is a (further) important distinction from business models of crowd-sourcing, which offer either small-scale material recompense for input, or the prospect of larger rewards if a contributed design etc. is chosen for production. For example, in his review of business-oriented crowd-sourcing models, Brabham (2008) singles out the fact that participation in open source projects does not lead to material reward as evidence that OS does not provide a compelling model for crowd-sourcing as a business process. He states:

These questions cast some doubt on the open source model as a supreme model for product development. Crowdsourcing, however, overcomes these limitations in the open source model by providing a clear format for compensating contributors, a hybrid model that blends the transparent and democratizing elements of open source into a feasible model for doing profitable business, all facilitated through the web (Brabham 2008).

Again, though, we should not ignore indirect benefits. The knowledge of specific open source software products that one gains from contributing code to them on a volunteer basis may provide a significant advantage in the employment market. Moreover, some open source products have social structures built up around them, providing extrinsic, social motivations analogous to those noted above for crowd-sourcing projects, in addition to the intrinsic motivation of contributing to the software.

4.4 Gamification

Other approaches in the literature have emphasized the importance of tasks being enjoyable, and have focused on the development of games for crowd-sourcing of different kinds. Prestnopnik and Crowston (2011) discuss the role of games, and in particular possible approaches to creating an application for crowd-sourced natural history taxonomy classification using design science. They also note that ‘gamification’ has the potential to act as a disincentive to contributors who have expert knowledge or deep interest in the subject. The Bodiam Castle project provides a good use case of the power and the potential for use of games in the context of archaeological analysis of extant buildings, although this had a greater emphasis on visualisation than on competition (http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/04/Masinton.pdf).

However, gamification can also be a barrier for users who simply want to engage with the assets or processes in question; and furthermore it can trivialise the process of acquiring or processing data (see http://blog.tommorris.org/post/3216687621/im-not-an-experience-seeking-user-im-a for a combative assertion of this position). In their analysis of impact of the Cornell Ornithology Lab’s The Bird Network (TBN) project, where members of the public were given the opportunity to set up bird boxes near their homes on in their neighbourhoods and gather data about the use by birds which was then shared with the scientific team, Brossard et al (2005) note that participants’ interest in ornithology was likely to overshadow awareness of scientific
process; and was in turn likely to stymie efforts by the Lab to contribute to general scientific awareness and education in the US population (Tumbrull et. al. 1999).

4.5 Models of engagement

Even crowd-sourcing projects that are quite similar in terms of characteristics such as process, content or task type, can involve quite different models of collaboration and engagement with participants. Consider the following examples, each of which addresses the digitisation (via transcription) of handwritten, textual material that could not be processed by computer alone:

- **MarineLives (ML)** aims to create an academically respectable digital edition of 17th Century manuscripts originating in the High Court of Admiralty, London. Rather than recruiting volunteers widely via an open call, ML targets the subset of potential participants who are sufficiently dedicated to learn the required skills, commit their time, and persist when they encounter problems. Volunteers in ML make an explicit ‘deal’ with the project organisers, who require participants to commit three hours a week for fourteen weeks; in turn the project regards itself as responsible for investing in processes to sustain the contributors’ motivation and engagement. The project thus aims to recruit people who will be able to make a significant contribution from the start.

- **Old Weather (OW)**, on the other hand, takes a quite different approach to engagement, based on a ‘long tail’ type of model. It recruits many more volunteers, but a great proportion of the work is done by a small number of them, with a far greater number transcribing only a single page.

This difference in engagement models is to some extent driven by a difference in the nature of the work being carried out. In OW, the work can be split up into small components that can be carried out independently and do not individually require a lot of expertise. Each individual transcription is performed by 3 people and cross-checked for accuracy. This is not the case for ML, which aims to create an academically respectable digital edition. The British Library Georeferencer (BLG) is another example of a project taking the ‘long tail’ approach, although this has an emphasis on spatial metadata creation, which is a task requiring more specialist expertise.

Note however that in some projects it is possible for participants to move on to a different level, and to carry out more complex tasks; examples of this are the ships’ histories created in OW, or the biographies of notable botanists in Herbaria@Home.

4.6 Roles

It is commonly noted in crowd-sourcing projects that the roles played by contributors develop as their experience increases; however, the kind of model followed by a project influences the kinds of role that emerge. In OW, a “captain” can see every page that gets transcribed, which gives an overview of the resource. There is a strong element of competition among ‘super-contributors’ (and it should be noted that this particular motivation it is likely to be particularly strong among super-contributors, such as those present at the second workshop, and less so in the long tail component of the community). For example there is a single transcriber in OW who undertakes massive quantities of work, and participants are aware that if she ‘joins’ the ship
they are currently working on, their status is likely to be reduced by comparison. As this would mean that ‘their’ ship would be completed more quickly, it clearly indicates the level of personal competition among contributors.

ML has a different model in that it actively promotes collaboration, and provides a great deal of support via the team facilitators, who are identified individuals with clearly-defined roles. In OW there is a great deal of support via the forum (and the forum moderators are also volunteers, rather than project staff). It was noted that in OW there are tasks that require different skill sets, and conversely people who possess specific skills, and there is a degree of self-organisation whereby people match themselves to tasks. For example, some transcribers are particularly good at deciphering difficult handwriting.

In the BLG, the work was accomplished so swiftly, mainly by the super users, that a discernible collaboration model or infrastructure did not have time to emerge.

A sense of community is important. One participant in the second workshop considered joining OW but decided not to because she was concerned at being ‘just a number’ among the thousands of people involved; that the lack of a defined and developable role was an off-putting factor. This also contributed to a sense that her input would be more valuable elsewhere. It was noted that OW does not provide individual feedback from project staff, in contrast to Transcribe Bentham (TB) which does; and TB transcribers consider this to be important (this was also picked up in the interview with Contributor B, 10/10/2012).

Roles are also important in the development of specialised knowledge. OW in particular provides good examples of transcribers becoming expert in specialised areas of naval history. Indeed the development of roles is allied to both the development and the pre-existence of specialist knowledge and interest. This means it is necessary to nuance what is meant by subject area. For example, only a relatively small number of people are interested in maps of Leith in the eighteenth century and therefore in georeferencing only that type of map, but many more people are interested in maps and georeferencing generally. This means it is necessary to distinguish between interest in a question and interest in a subject area or particular kind of activity. At least two participants we spoke to compared crowd-sourcing work with crosswords as an activity that was pleasurable for its own sake. Contributor E, for example (interview 23/10/2012), discovered the Old Weather project via a piece on the radio. Whilst her initial interest was piqued by the climatology side of the project, her interests broadened as a result of her involvement, and she became interested also in synthesizing and editing the histories of particular ships. Some aspects of the material had personal geographic relevance to her, for example the WWI German Fleet being taken to the Firth of Forth after its capture. The editing part of the project gives a feeling of ‘having a whole ship to yourself’, and being able to ‘tell the story of a whole ship’.

A sense of support, either from a community or from the project is critical in recruiting and maintaining contributors. Two key questions therefore are how can sustainable support structures be set up for any given project, and how do communities develop around projects?
One participant in the second workshop observed the formation of ‘a transcription community, and a researcher community, with some overlap between the two’. In building infrastructure for crowd-sourcing projects, ‘the crowd’ is a difficult thing to predict, in terms of what kind of community will be formed, and what the group will achieve. It is hard to know how much support/input will materialise, making it difficult to formulate requirements. A good example is the BLG project, where 750 maps were transcribed in four days. This completely outstripped all infrastructural and support mechanisms the BL had put in place.

4.7 The role of 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 many projects it is possible to track individual participants’ contributions, and acquire statistics on who contributed the most etc (although not all; see below). Where this is the case, projects can establish ‘leader boards’, indicating which participants have made the biggest contributions (in whatever terms the project is working with). In the case of the BLG 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 this kind of contact with 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. However, in order for competition to be a significant factor in encouraging and sustaining crowd-sourcing, the tasks, and the result/outcome of accomplishing a particular task must be easily quantifiable and definable so that it can be compared to the outcomes of tasks completed by others.

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 BLG project, some maps are more complex than others, so the team felt there was relatively little meaning in comparing the effort needed to georeference different ones. Another general possibility is that different aspects of the same project could have different leader boards, thus reflecting a model of ‘diffused competition’ (see further discussion in the report of the May workshop, http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/09/workshop_report1.pdf). 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. One participant in the second workshop asked ‘Is there a prejudicial (or factual) connection between “free labour” and “low quality”? Even when data are vetted by software or by editors, can this idea still jeopardise the (actual or perceived) reliability of the outcome? 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 outputs are being generated. This kind of output is also less likely to lend itself to the leader board type approach outlined above.
In the second workshop, competition featured strongly as a motivating factor for participating; but this should be contextualised/qualified by the fact that all of those present were ‘super contributors’, who are likely to feel a sense of competition more keenly than those in the ‘long tail’ part of ‘the crowd’. Competition can be defined in various ways: in the quantitative sense, e.g. number of pages or other asset unit processed by individuals that can be displayed on a leader board. Others, however, are concerned with producing high-quality work in a more qualitative sense. ML for example has a volunteer (who works in a museum) whose main aim is to produce pages of excellent transcription with minimal help requests. This is not, however, incompatible with a sense of common purpose, e.g. in OW ‘you feel part of the ship’ you are working on.

4.8 Other motivations

It is possible for motivations to change over time, for example many OW volunteers are initially interested by the possibility to contribute to climate change research, but become interested in maritime history as they are exposed to the project’s content. Motivations can also change with the kind of task type. One participant remarked that ‘palaeography is only fun when you can’t do it’.

Different tasks also attract the motivation of simple curiosity, and the desire to locate new knowledge. One example given was ‘investigative’ crowd-sourcing, for example researching the corpus of MPs’ expenses published in the UK in 2009 (see http://www.niemanlab.org/2009/06/four-crowdsourcing-lessons-from-the-guardians-spectacular-expenses-scandal-experiment/)

Many participants in the second workshop felt that the crowd-sourcing activity they are engaged with is highly addictive. It was also considered important for recruitment that participants can do work on a project at any time (and put it down at any time); much like a computer game.

Personal factors are sometimes involved in the decision to engage in crowd-sourcing, for example redundancy frees up time to get involved and generates a need to keep the mind active, also displacement activity following bereavement has been cited. In this case crowd-sourcing formed part of a ‘restorative process’.

4.9 What initiates interest in a crowd-sourcing project?

Most super contributors at the second workshop discovered their projects serendipitously, for example via Twitter. The Zooniverse community has been very successful in generating interest across related projects, since it presents the projects in dashboard style and encourages cross-fertilization.

In many cases discussed, mass media exposure leads to a spike in uptake. For example in the Transcribe Bentham project a New York Times article about the project was published on 28th
December 2010⁴. This was, however, problematic. When the article appeared, many enquiries were made but the project team was not available due to the Christmas break. BBC Radio 4’s PM programme mentioned OW, which also led to a big spike.

4.10 Reward

Reward may be considered in terms of satisfying/fulfilling interest in the subject; or as a more tangible quid pro quo for being involved. This can take the form of status, some personal benefit such as training and experience (how can this be qualified and made provable?), or in instant/gamified gratification, or deferred gratification.

In many projects, the feedback loop, affirming that the contributions made were correct and valuable, is a very important component of the reward for engagement, and conversely lack of feedback can be very frustrating and discouraging.

Contributors mentioned a number of skills gained; these included general practical IT competencies, such as learning to edit wikis and to use Skype for distributed collaboration, as well as specialised skills such as XML encoding (e.g. through Transcribe Bentham). As noted above, many contributors gain domain knowledge; for example the material and opportunity to edit ships’ histories in OW for the Naval Histories website. One participant in the second workshop is now actively involved in editing a set of historical documents (ship’s history), which would not have happened without been involved in the OW project. This can contrast with the interests of the project team: in this case, for example, the team is interested in weather history, whereas some contributors are interested in ships (http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/04/Brohan.pdf). Other Old Weather participants highlighted multiple motivations, including the ability to learn about history and address the issue of global warming (Interview with Contributor A, 7/9/2012). Contributor F’s interest (interview, 29/10/2012) was aroused by a piece on the BBC news site. As with others, her initial interest centred on concern about climate change, but her motivations broadened as her familiarity with the material increased. She has ‘captained’ two ships, but noted that lack of time meant it was hard work staying captain, especially when other contributors have more free time. She also became interested in synthesizing and editing the histories of individual ships, and stressed her interest in exploring niches of history that had been hitherto unexplored. For her, interaction with other participants via the forum is extremely important, both for ‘exchanging chit chat’, and for discussing the practical and technical problems that the transcription process presents. She also noted that other Zooniverse projects had not piqued her interest in the same way; this is simply because she is more interested in the area of history.”

Participants can also pick up a basic grounding in research methods of collation, synthesis and analysis in the area of interest to them – contra observations made about some ‘citizen science’ projects, where a focus on the content often negates that on background method. See e.g. the Cornell Lab ornithology experiment (Trumbull et al 2009).

Where tasks develop in this way, people may gain the experience of working collaboratively, in contrast to situations where the tasks remain purely mechanical (e.g. anonymous marking up/transcribing etc. of records, as in FamilySearch).

5. Motivations of academics and other project organisers

At least part of the success of GalaxyZoo and other Zooniverse projects is that they catered to clear and present academic needs. In the case of Galaxy Zoo itself, the assets – photographs of galaxies from the Sloan Digital Sky Survey – were far too numerous to be examined individually by any research team, and the task – the classification of those galaxies’ morphology – was not one that could be performed by computer software, although for the most part could be carried out by a person without specialist expertise. In 2008, they reported that $4 \times 10^7$ individual classifications had been made by around $10^5$ participants (Lintott et. al. 2008). Quite simply, this is work that could not have been carried without large-scale public engagement and participation.

Since humanities crowd-sourcing is an emergent area, identifying the motivations of academics engaging with such work is not always so straightforward. In most cases, there is an academic research question or a need for a particular resource. For example, the Transcribe Bentham project was motivated by the fact that 40,000 folios of Bentham’s work was untranscribed, and that consequently (as the project states explicitly at [http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/04/Causer.pdf](http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/04/Causer.pdf))

> [o]ur understanding of Bentham and his thought – of importance to anyone studying the eighteenth or nineteenth centuries – is incomplete. In short, the Bentham Papers are a source of enormous historical and philosophical importance, yet much of the collection remains unknown, let alone adequately studied.

The British Library’s Georeferencer project is another example of this. The Library is able to make its map collections more searchable, and therefore more exploitable by using crowd-sourcing to georeference them. Again, this is not a task that a computer could do, and that a team tasked with metadata creation could only do over a very long period of time, and with prohibitive cost.

The OldWeather project provides an example where the motivations of the researchers involved are quite clear, namely to be able to use the information contained within the assets to explore historic weather patterns (although these motivations may not necessarily be shared by the participants):

> Climate researchers need the millions of historical weather records archived in Royal Navy ship's logbooks from decades and centuries ago, marine historians want the day-to-day records of people, ships and places in the same documents (see [http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/04/Brohan.pdf](http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/04/Brohan.pdf)).
Other researchers, particularly those in the GLAM sector, see crowd-sourcing as a means of filling gaps in cultural heritage institutions’ coverage, as noted by Terras (2010). This can be an effective way of obtaining information about assets (or obtaining the assets themselves) to which only certain members of the public have access, for example through personal or family connections. However, in order to be usable for academic purposes, the ‘circulation of knowledge’ (see Section 6.1) must be curated in some way, and this may involve expert input.

Finally, crowd-sourcing is sometimes seen as a means of community building, and thus increasing the wider impact of academic research activities. Moyle et al. 2012 note:

[The project is] helping to stimulate public engagement with scholarly archives and manuscript transcription – always a challenge, but perhaps carrying extra significance at a time when, in the UK at least, humanities research units and related services are under intensified pressure to quantify their social impact. It will also help to open up the thought of Jeremy Bentham to new audiences.

The motives for undertaking crowd-sourcing projects in the humanities are therefore less clear-cut that those for undertaking analogous ‘citizen science’ activities. In order to achieve a useful and/or usable academic outcome, it is necessary to correctly match the processes to be undertaken to the asset in question, to select the right kind of task by which those processes will be carried out, and to be clear about what the intended output is to be.

6. Outcomes of humanities crowd-sourcing

There was extensive discussion among participants in the second workshop concerning the sort of outcomes, whether anticipated or not, that arise from humanities crowd-sourcing projects. These outcomes fall broadly into two categories: creation of knowledge and creation of communities (and this is reflected in the typology).

6.1 Creation of knowledge

The following kinds of new knowledge, or new information that can lead to new knowledge, were identified in the discussion as potential outcomes of crowd-sourcing projects:

- Everyday ephemera that would not otherwise be accessible.
- Information that would normally only be accessible to local history groups.
- Specific kinds of information that is ‘locked away’ in projects’ assets. For example, the ships’ logs used in OW are a unique record of burials at sea, as such burials are not recorded anywhere else.
- Personal histories, for example from diaries. It is possible to source data about individuals and groups that are under-represented in official records and archives; such information can contribute a quite different angle on historical research.
- Personal family links to historical processes and events (e.g. ship histories, commonwealth war graves).
- Information relating to personal or family histories can be obtained indirectly as a result of a crowd-sourcing project rather than from the assets on which the project is working, for through connections that would not otherwise have been made. For example, one OW participant received a forum private message from someone whose grandfather died on the ship Odin, and who was able to relate some of the history around it.
- Identifying links between information or between physical objects.
- Sources such as working class diaries, and other social history resources. An example of research based on such diaries could be an exploration of changes in personal mobility changes brought about by the advent of the railway, a topic otherwise unexplored in railway history.
- Providing summaries of datasets to make them more discoverable (this may be regarded as a form of descriptive metadata).
- New syntheses of existing data.
- Knowledge of how to conduct collaborative research.
- Access to knowledge and skills that other individuals have but have not shared.
- Recording knowledge before it disappears, for example records of finds made with metal detectors (e.g. via the Portable Antiquities Scheme), or of damage to the Thames foreshore. There is a perceived value among participants in plugging holes in existing datasets.

In addition, participants can also learn practical skills; these can include ICT-related skills, but also the skills required for collaboration, as crowd-sourcing gives people the opportunity to work in large groups.

A key point raised is that crowd-sourcing can improve the circulation of knowledge. The internet, and especially social media, privileges a certain kind of knowledge circulation, which stems from very simply designed and widely understood tasks: applying hashtags, commenting, ‘liking’, conveying ephemeral information in text slang, and so on. These are not necessarily the best kind of channels for the discovery and development of humanities data. Crowd-sourcing can redress this imbalance through intelligent task design. The more clearly defined a task is, the more likely the crowd will be to participate. This needs to be accompanied by some direct value that will accrue, either to the individual or to the crowd.

It was also felt strongly that knowledge created through crowd-sourcing should be available on an open access basis. There were two aspects to this. Firstly, the ability to access the final results of a project to which the volunteer has freely contributed their efforts, and to show these results to friends or colleagues, can give a great level of satisfaction to the volunteer, and forms part of a project’s reward structure. Secondly, it was considered that the results of crowd-sourcing projects should be open for ethical reasons; if they are not open, this devalues the contributions of the volunteers.

6.2 Creation of communities
One characteristic of crowd-sourcing in the humanities is that it often leads to the creation of a new community around a humanities subject, question or topic, and in turn this community can carry out some interesting or valuable work that may go beyond what was envisaged by the project. This can be done in multiple ways, by selection, self-selection, tapping in to existing communities. For this to be possible, there needs to be wide, if not total, distribution of the project, usually via the internet. There needs to be a shared purpose in carrying out the activity both for the academic (or other organiser) and for the participants.

There also needs to be some element of peer review/quality control to reassure the community that their efforts create academically viable outputs; this may help to address the issue raised by Rose Holley (2010), who distinguishes between crowd-sourcing and social engagement. Quality control can be of collective outputs (e.g. quality assuring a mass produced dataset); or individual ones (e.g. keeping page images next to transcribed text to ensure correct transcription). In any case, a project should identify/specify what its quality requirements and processes are (and adhere to them), and also make the QA process evident. It probably is not necessary to aim for an “ideal” output, e.g. the comprehensive TEI markup of a text; rather crowd-sourcing can get a larger amount of material into the public domain, rather than a small amount of content in very great detail. For example, one of the shortcomings of Wikipedia is the lack of clearly-defined peer review roles, which results in a significant quantity of poorly-written and inaccurate material being published, as well as malicious or frivolous edits.

An interface enabling users to (for example) annotate/tag and suggest links without focus is not crowdsourcing; the focus on a shared task or purpose is critical. This relates to an observation that the more closely defined the task is, the more successful it will be. The more successful tasks are those where it is ‘easier not to be wrong’, and the outcome immediately verifiable or checkable. If the task is ambiguously or subjectively designed, then contributors worry about being wrong, or producing inaccurate information. This applies mainly to asset types which are empirical, e.g. dates, or weather observations: items that are either transcribed correctly or not. It does not apply to asset types which are more conjectural or creative in their nature. It depends on the purpose of the outcome type: In some cases the output is intended to digitise resources in order to answer a question (e.g weather readings to reconstruct historic weather); in others the focus can be on digitising material to make it accessible (e.g. an archive); or simply to record ephemera to make it available to posterity. In these cases, the questions are formed post hoc.

7. Typology

7.1 Introduction

This typology seeks both to bring together the research cited above in Section 2.1 and to reflect the experiences and processes uncovered elsewhere in this research review. It does not seek to provide an alternative set of categories or labels specifically for humanities crowd-sourcing; rather it recognises that there are a set of fluid and interchangeable categories within four key typological areas: asset type, process type, task type, and output type. Table A-1 fed in to its
development by demonstrating how participant-defined descriptions of crowd-sourcing activity can be mapped onto this structure.

It is the main conclusion of this research review that crowd-sourcing projects in the humanities – including the motivations of the participating communities and individuals – can best be understood by analysing them in terms of these four ‘primitive’ facets and of the relationships between them, and in particular by observing how the categories applicable in one facet are dependent on those in others.

Of course, not all projects will map straightforwardly onto single categories under of the four facets. HistoryPin (http://www.historypin.com), for example, is involved with georeferencing, images, metadata, impact, engagement and recording. While it operates outside the academic sector, it has developed strong links with the GLAM sector by providing a set of tools to allow embedding of HistoryPin content in cultural collections (see http://wearewhatwedo.org/press-releases/historypin-unleashes-new-tools). Such examples constantly challenge this typology, and provide the impetus that will guide its future evolution.

7.2 Processes

A process is a sequence of tasks (see Section 7.4), through which an output is produced by operating on an asset. It is conditioned by the kind of asset involved, and by the questions that are of interest to project stakeholders (both organisers and volunteers) and can be answered, or at least addressed, using information contained in the asset.

<table>
<thead>
<tr>
<th>PROCESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collaborative tagging</td>
</tr>
<tr>
<td>Linking</td>
</tr>
<tr>
<td>Correcting/modifying content</td>
</tr>
<tr>
<td>Transcribing(^5)</td>
</tr>
<tr>
<td>Recording and creating content</td>
</tr>
<tr>
<td>Commenting, critical responses and stating preferences</td>
</tr>
<tr>
<td>Categorising</td>
</tr>
<tr>
<td>Cataloguing</td>
</tr>
<tr>
<td>Contextualisation</td>
</tr>
<tr>
<td>Mapping</td>
</tr>
<tr>
<td>Georeferencing</td>
</tr>
<tr>
<td>Translating</td>
</tr>
</tbody>
</table>

Table 1: Process Types

**COLLABORATIVE TAGGING**

\(^5\) This category also includes marked-up transcriptions, e.g. as TEI XML.
Collaborative tagging may be thought of as crowd-sourcing the organisation of information assets by allowing users to attach tags to those assets. Tags can be based on existing controlled vocabularies, but are more usually derived from free text supplied by the users themselves. Such ‘folksonomies’ are distinguished from deliberately designed knowledge organisation systems by the fact that they are self-organising, evolving and growing as the crowd adds new terms. Research has also been carried out into extracting formal data structures from folksonomies (see Lin and Davies 2010).

Collaborative tagging can result in two concrete outcomes: it can make a corpus of information assets searchable using keywords applied by the user pool, and it can highlight assets that have particular significance, as evidenced by the number of repeat tags they are accorded by the pool. Previous research in this area has examined the patterns and information that can be extracted from folksonomies. Golder (2006) found that patterns generated by collaborative tagging are, on the whole, extremely stable, meaning that minority opinions can be preserved alongside more highly replicated, and therefore mainstream, concentrations of tags. Other research (Trant, Bearman and Chun 2007; Trant 2009) has shown that user-assigned tags in museums may be quite different from vocabulary terms assigned by curators, and that relating tags to controlled vocabularies can be very problematic, although it could be argued that this allows works to be addressed from a different perspective than that of the museum’s formal documentation. In any case, such approaches to knowledge organisation are likely to play a significant part in the organisation of humanities data in the future.

A good example of this is the BBC’s YourPaintings project, developed in collaboration with the Public Catalogue Foundation, which has amassed a collection of photographs of all paintings in public ownership in the UK. The public is invited to apply tags to these, which makes them searchable by keyword. See [http://www.bbc.co.uk/arts/yourpaintings/](http://www.bbc.co.uk/arts/yourpaintings/).

**Case study: the Prism Project**

[http://www.scholarslab.org/category/praxis-program/](http://www.scholarslab.org/category/praxis-program/)

One key assumption underlying the process of collaborative tagging is that the assets being tagged are themselves stable and clearly identifiable as distinct objects. This has been the case for most research focusing on this as a method, e.g. tagging pictures or museum objects. However, the rich TEL/XML-based markup used in many digital humanities projects has led to exploration of users defining semantically significant areas of text and then applying tags to them.

The Prism project has developed a tool which allows readers of texts to create collective interpretations of them by a combined process of tagging and highlighting. Users can highlight a section of text and then associate the highlighted section with a tag from a controlled vocabulary. This differs from conventional text markup, in which the text is marked up by a single editor in conformance to an XML schema; rather, multiple takes on the same passage can be created and overlaid. It also
differs from other text-focused crowd-sourcing projects in which texts are processed in a single way, for example transcribing handwriting to machine-readable assets. A colour-coding system is used, making it possible to visualise different contributors’ interpretations over time. Key to the design of the project is the relationship between instructor and students; the tagging, text selection and visualisation are an appropriate means of capturing debates in a classroom, and most of the discussions about the interpretations were undertaken offline.

A cross-section of different kinds of text was selected for the first release of the project, with the aim of representing different genres, including prose, poetry and archival material. The texts are:

*The Sneeches* by Dr. Seuss
*The Raven* by Edgar Allen Poe
*Notes on the State of Virginia* by Thomas Jefferson (excerpt)
*Portrait of the Artist as a Young Man* by James Joyce (excerpt).

The project raises interesting questions of professionalism versus amateurism. Whose interpretation is being presented? It could be a professional reader (e.g. a scholar) or an amateur enthusiast. The process of generating the interpretation is the same in each case, but there is no way of knowing if an interpreter is an emeritus professor or a member of the general public. The interpretations are also of sociological interest, to determine the reception of a text by its readership; there are therefore parallels with the Year of Shakespeare project [http://www.yearofshakespeare.com/](http://www.yearofshakespeare.com/), where reviews are commissioned from professional reviewers, and the public are able to comment on them.

Many people like to read texts, and this forms a core motivating factor for undertaking an activity such as this. There is also likely to be an intrinsic interest among groups of researchers in synthesising their interpretations of texts of common interest. User motivation is likely to increase when, in future releases, users can upload a text to Prism and pose their own questions, for example asking for a community’s interpretation of it. Giving people this agency to frame discussion and interpretation would be important an important motivating factor, and it reflects discussion in the May workshop on the role of crowd-sourcing contributors versus research teams in the design of both projects and research questions (see [http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/09/workshop_report1.pdf](http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/09/workshop_report1.pdf)).

The project team would like to be able to capture interpretations at a particular moment in time, e.g. by providing a function that allows a snapshot of the text, highlights and tags to be downloaded, navigated and visualised. They are also interested in extending the project to support collective interpretation of other media, for example segmenting video or annotating images.
A possible research question, which could lead to ‘conventional’ research outputs, would be to investigate whether there is a relationship between different kinds of texts, their structures, their stylistics/stylometrics, and people’s interpretations of different parts of the texts. Many ‘mainstream’ humanities crowd-sourcing activities, such as transcribing texts according to well-defined procedures, have identifiable completions, whereas interpretations can go on indefinitely. Furthermore, there is no right or wrong interpretation; the questions posed are therefore very open. The broader goal of Prism is not to see whether people interpreted a text correctly or not (in contrast, say, to whether they transcribed a text correctly or not), but rather to see how, in the aggregate, people read and give meaning to a text.

Source: Interview with Jeremy Boggs, 9/8/2012

LINKING

Linking can take the form of identifying and documenting links of a specified type between individual assets, or, far more commonly, of linking via semantic tags (where in this case the tags describe binary relationships); it is thus included in the typology as a subset of collaborative tagging.

CORRECTING/MODIFYING CONTENT

While digital content is increasingly ‘born digital’, projects for digitising existing analogue material abound. Many of the technologies for digitising research content on a large scale, such as Optical Character Recognition (OCR) and speech recognition, are generally error-prone, and factoring in quality control and error correction is essential for any such enterprise. Crowd-sourcing can be used for such error correction.

The TROVE project, which produced OCR-ed scans of newspapers from 1803 onwards held in the Australian National Archives, is an excellent example of this (Holley 2009; 2011). In this case, the volume of digitised material would simply have been too great for the library to undertake the corrections using its own staff, and if only page images were produced then there would have been little or no capability for searching the text, significantly reducing the benefits of having the material in digital form at all (Holley 2009).

There have also been attempts to support crowd-sourced correction and modification of content using automation. An example of this is COoperative eNgine for Correction of ExtRacted Text, or CONCERT (Karnin et. al. 2010), which tries to match tasks to contributors’ skills, and aims to implement robust quality assurance mechanisms.

TRANSCRIBING

Transcribing is closely linked to correction and modification, and is currently one of the most high-profile areas of humanities crowd-sourcing, as it addresses directly one of the most
fundamental problems with OCR: that handwriting, especially complex and/or difficult to read scripts, cannot be automatically rendered into machine-readable form using current technology. It can only be transcribed manually with the human eye and, in many cases, with human interpretation.

Two projects have contributed significantly to the prominence of transcription among crowdsourcing projects: Old Weather and Transcribe Bentham. The latter “invites the public to play a part in academic research and attempts to break down traditional barriers” (Causer et al. 2012; see also Moyle et al. 2012). The aim of Transcribe Bentham, funded by the AHRC during 2010/11, was to encourage volunteers to transcribe and engage with unpublished manuscripts by Jeremy Bentham (1748–1832), the philosopher and reformer, by rendering them into text marked up using TEI XML. These volunteer-transcribed manuscripts will contribute to the production of the Collected Works of Jeremy Bentham, and will be uploaded to UCL’s digital repository in order to make the collection searchable. However, as the project has stated, during this period the rate of volunteer transcription did not compare favourably with that of professional scholar-transcribers. 1009 manuscripts were transcribed in the six month testing period (8 Sept 2010 – 8 March 2011); in contrast to this, project staff estimated that the project’s research assistants could have transcribed 5000 manuscripts in a twelve month period (Causer et al. 2012). Several explanations are given for this: that Bentham’s handwriting can be extremely difficult to decipher, and that the material is extremely complex. It has also been noted that (as of May 2012) 304 (19%) registered users transcribed material, and around two-thirds of these have worked on only one manuscript (op. cit.). There is also an extremely high moderation overhead, with significant staff time needed to process the outputs, determine whether they are good enough to be locked for further editing and, crucially, to provide feedback to the contributors.

Since then, the volunteer transcription rate has improved. As of 2 November 2012, 4,612 manuscripts had been entirely or partially transcribed, of which 94% were of sufficient quality for editorial work and uploading to the digital repository. Between 8 September 2010 and 2 November 2012, an average of 41 manuscripts (c. 20,000 words) were worked on each week; however, between 28 January 2012 to 2 November 2012, the average was 51 manuscripts (c. 25,500 words) per week, which means that collectively the volunteers are now transcribing at a faster rate than a full-time researcher, so there is potential for avoiding significant costs in the future (see Causer and Wallace, 2012).

An interview with a contributor to Transcribe Bentham (interview with Contributor B 10/10/2012) stressed that getting feedback from the project on what was being done correctly and what needed improvement was essential to maintaining motivation and a sense that their contribution was being valued. The project also recognises this (op. cit.). Such a sense of

6 The project has received additional funding from the Andrew W. Mellon Foundation for the period 1 October 2012 to 30 September 2014, and it is hoped that improving the transcription interface (e.g. by developing a WYSIWYG interface) will increase the transcription rate and reduce staff time spent validating submissions (significant effort is spent correcting XML).
community is therefore critical in a project where the transcription involves the recognition of complex information.

*Old Weather* is also based on the need to digitise content that could not be digitised by computer programme, in this case the logs of ships of the British Royal Navy, which were transcribed in order to acquire and analyse the weather observations they contain. This information is of major scientific significance, as it allows researchers to reconstruct historical weather trends in a way, and to a level of detail, that could not be extrapolated from observations of the current physical environment. About 350,000 of pages of log-books held by The National Archives, covering the period 1914-23, have been photographed and are being transcribed by 12,000 volunteers, each page being transcribed three times independently to ensure accuracy (Brohan et. al. 2009). As with Transcribe Bentham, a sense of community is extremely important in this project, with the project’s discussion forum providing a critical channel of communication, both between participants and with project staff (*Interview with Contributor A, 7/9/2012*; see [http://forum.oldweather.org/](http://forum.oldweather.org)). Exchanges on technical aspects of the transcription process are very common on the forum, and roles are assigned to contributors based on their level of experience (i.e. the number of pages transcribed).

One question posed by a user with the role ‘Newbie’ is:

> I am currently working on the HMS Mantis which is an insect class river gunboat. I have been following the logs from summer of 1918 and I am now up to summer of 1919. Everything was going well until the ship went back to England for a month’s leave. Now that the logs have started back up they switched to what looks like blue ink. It is so faded that I can barely make out anything more than date and most of the weather entries. Most of the time I cannot make out the location (they are not using lat/lon location just listing a port or writing in to and from info) and I definitely cannot make out any events.

The first response, from a highly experienced ‘Hero Member’ is:

> Welcome to the forum.
> Sorry I haven’t got better news for you on your first visit here, but ... if you can grit your teeth for a couple of weeks of variable clarity, your logkeeper fills his pen, or gets a new bottle of ink, on 1st August.

> I managed to get all the weather readings, which is the important part, and your eyes are bound to be younger than mine. Some of our more savvy transcribers reckon that loading the page into Photoshop or a superior version of Word will allow you to adjust contrast/background and other stuff to improve legibility, but that is beyond my abilities.

> If you really are struggling, then abandon ship. There are others around to tax your ingenuity in different ways.
Remember, it's acceptable for transcribing to be a challenge but don't let it become a chore.

Good Luck, whatever you decide.


Such exchanges among community members are indicative of a high degree of collaborative, communal working in addressing problems that arise during the process.

The collaborative model needed for successful crowd-sourced transcription in the humanities varies according to the complexity of the material to be transcribed. Complex material, such as the two cases cited here, requires a high level of support, whether from a project team or from the participant’s peers in a forum; or more usually a combination of the two. Less complex material is likely to require less support, or at least less support in the form of contact with other users or project staff. For example, when transcribing the more structured data found in family records (e.g. http://www.familysearch.org), the information (text or integers) to be transcribed is presented to the user in small segments – names, dates, addresses etc. on birth certificates – and it requires different cognitive processes, which in turn are less dependent on the mode of community represented by feedback and engagement from ‘experts’.

RECORDING AND CREATING CONTENT

Typically processes in this category are concerned with recording ephemera and intangible culture, such as oral history or reminiscence. These frequently take the form of a GLAM or other cultural institution soliciting memories from the communities it serves, for example the Tenbury Wells Regal Cinema’s Memory Reel project (http://www.regaltenbury.org.uk/memory-reel/). Such processes can incorporate a form of editorial control or post hoc digital curation, and their outputs can be edited into more formal publications, or analysed/explored using methods such as sentiment analysis (see below).

This category is more likely to conform to the ‘social engagement’ model, in terms of Holley’s distinction of social engagement from crowd-sourcing (2010).

COMMENTING, CRITICAL RESPONSES AND STATING PREFERENCES

These processes are only likely to fall within the definitions of crowd-sourcing as set out in this review if there is some specific purpose for which people come together. One example of this is capturing audience response to the 2012 World Shakespeare Festival, which is the aim of the Shakespeare’s Global Communities project (www.yearofshakespeare.com). A key question of this project was ‘How do new social networking technologies reshape the ways in which diverse global communities connect with one another around a figure such as Shakespeare?’ (http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/09/workshop_report1.pdf). The question itself provides a focus for the activity, and although in and of itself it does not produce a verifiable
academic output, it provides a dataset for looking at research questions on the modern reception of Shakespeare. In some cases, appropriately presented blogging software provides a platform for focused scholarly interaction. For example, a review by Sonia Massia from King’s College London of King Lear on the Year of Shakespeare site attracted controversial responses, leading to an exchange about critical methods as well as content\(^7\). What differentiates such exchanges from amateur blogging is the scholarly focus and context provided by the project, and its proactive directing of the content creation. The project thus provides a tangible link between the crowd and the subject.

There is much potential in mining corpora of comments and critical responses, using techniques such as sentiment analysis (e.g. Chmiel et. al. 2007), which attaches quantitative weights to the positivity or negativity of units of text, such as tweets or commentaries, and allows collective analysis of these.

**CATEGORISING AND**

Categorising involves the placing of assets into predefined categories; it differs from collaborative tagging in that the latter is unconstrained.

**CATALOGUING**

Cataloguing – or, more expansively, the creation of structured, descriptive metadata (e.g. for cultural objects) – is a more open-ended process than categorising, but it is nevertheless constrained to follow accepted metadata standards and approaches. It frequently includes categorising as a sub-activity, e.g. according to LoC subject headings.

Cataloguing is a particularly time- and resource-consuming process for many GLAM institutions, and crowd-sourcing this activity has been explored as a cost-effective means of increasing access. For example the What’s the Score project at the Bodleian Library, whose principal aim 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 digitisation and the crowd-sourcing of descriptive metadata (see the crowd-sourcing site at [http://www.whats-the-score.org](http://www.whats-the-score.org), and the delivery site at [http://scores.bodleian.ox.ac.uk](http://scores.bodleian.ox.ac.uk)).

The cataloguing process type is linked to contextualisation, as ordering, arraying and describing assets will also make explicit some of their context.

**CONTEXTUALIZING**

Contextualization is typically a more broadly-conceived activity. It involves enriching an asset of a particular type by adding to it or associating with it other relevant information or content.

**MAPPING**

In the sense in which it is used here, *mapping* refers to the process of creating a spatial representation of some information asset(s). This could involve the creation of map data from scratch, as in the OpenStreetMap initiative ([http://www.openstreetmap.org/](http://www.openstreetmap.org/)), but it could also be applied to the visual or spatial mapping of concepts (as in a ‘mind map’). The precise sense will be highly dependent on the *asset type* to which mapping is being applied.

Note that *mapping* should not be confused with *georeferencing*.

**GEOREFERENCING**

Georeferencing refers to the process of establishing the location of un-referenced geographical information in terms of a modern real-world coordinate system (such as latitude and longitude). Georeferencing can be used to enrich significantly geographical datasets that do not include such information, but which could or should do, and there has been significant activity in this area in terms of crowd-sourcing and user engagement.

**Case study: the British Library Georeferencer (BLG) Project**

Georeferencing has been used in the UK in the BLG project (see also Section 3). BLG is based on the Georeferencer application, along with the Old Maps Online and TEMAP projects. The tool allows ‘free, crowd-sourced collaborative online georeferencing of map images from a number of libraries’, and has been applied to four other map collections in addition to those of the British Library itself (Fleet et. al. forthcoming). The purpose if its application by the British Library 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 themselves, 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.

For participants, there was an element of instant gratification as they could see the results of their work immediately. The project had a ‘citizens’ forum’ tab, which proved important for generating a sense of community among the participants. 725 maps were georeferenced between 13 and 18 February 2012 by around 90 participants, with very high data quality. Social media was considered to be a key factor in this astonishing success, as also was the institutional association with British Library, and the credibility that this gave to the processes and outcomes.

BLG therefore corresponds to the output type of *metadata* (see below) and, as no new map data was created, it is distinct from the process type of *mapping*. However, the researchers behind BLG have noted that
while the advantages that online georeferencing offers to cartographic collections are considerable, these very specialised benefits may be eclipsed by the broader ability to expose and share collections with the public in a new and much more engaging way than was before possible”. (Fleet et. al. forthcoming).

It is therefore evident that, as well as metadata, the more qualitative output types of engagement and knowledge/awareness also apply.

TRANSLATING

This process type covers the translation of content from one language to another. In many cases, a crowd-sourced translation will require a strongly collaborative element if it is to be successful, given the semantic interconnections and interdependencies between that can occur between different parts of a text. However, in cases where a large text can be broken up naturally into smaller pieces, a more independent mode of work may be possible. An example of this is the Suda On-Line project (http://www.stoa.org/sol/), which is (among other things) translating the entries in a 10th Century Byzantine lexicon/encyclopaedia. A more modern, although non-academic, example is provided by the phenomenon of ‘fansubbing’, where enthusiasts provide subtitles for television shows and other audiovisual material (Cintas 2006).

7.3 Asset type

The Asset in a crowd-sourcing project refers to the kind of content which is, in some way, transformed as a result of exposure to the crowd.

<table>
<thead>
<tr>
<th>ASSET</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geospatial</td>
</tr>
<tr>
<td>Text</td>
</tr>
<tr>
<td>Numerical or statistical information</td>
</tr>
<tr>
<td>Sound</td>
</tr>
<tr>
<td>Image</td>
</tr>
<tr>
<td>Video</td>
</tr>
<tr>
<td>Ephemera and intangible cultural heritage (e.g. oral history)</td>
</tr>
</tbody>
</table>

Table 2: Asset Types

GEOSPATIAL

A geospatial asset is a dataset or text that includes components that identify or refer to locations on the earth’s surface. A map is an obvious example, but the category also covers assets such as travelogues or gazetteers. The asset may refer to location absolutely (e.g. by using coordinates that refer to known locations on the earth’s surface), or relatively (e.g. by referring to an identifiable location or feature, such as ‘Birmingham, England’ or ‘the River Severn’), or
indeterminately, referring to spatial features or types of features in a way that is not transferrable directly to the earth’s surface (e.g. ambiguous place-names, or a reference to ‘rivers’).

There is an important distinction to be made in this context between geospatial assets created by expert organisations, such as the Ordnance Survey or the US Geological Survey, and those created by community-based initiatives. The former may have the authority of a governmental imprimatur, and the distinction of being officially endorsed for a variety of reasons (such as emergency response or guiding planning processes). However, the recent emergence of crowdsourced geospatial assets – a product of the recent global growth in the ownership of hand-held devices with the ability to record location using GPS (Goodchild 2007) – has led to the emergence of resources such as Open Street Map, which in turn has led to a discussion about the reliability of such resources. In general, it has been found that Open Street Map in particular is extremely reliable (Haklay and Weber 2008, Haklay 2010), but that the specifications for such resources must be carefully defined (Brando and Bucher 2010). The impact of Open Street Map on the cartographic community generally has been noted (Chilton 2009). The importance of mapping as a means of convening spatial significance means that this kind of asset is particularly open to different discourses, and possibly conflicting narratives. The digital realm, with the potential for many users to contribute many kinds of information, holds great potential for this area, but the issues raised are still emergent (see Fink 2011, Graham 2010).

**TEXT**

Our review suggests that history, especially social history, is currently by far the most common subject area for crowd-sourcing projects in the humanities, and consequently text is by far the most common asset type to be engaged with by humanities crowd-sourcing projects.

It should be noted that, as *asset type* describes the material that is transformed by crowd-sourcing, the *text* category refers to text that already exists, rather than the creation of new text, for example by recording personal experiences (this would fall under the asset type *ephemera and intangible cultural heritage*). This category of asset is likely to result in one or more of the output types *transcribed text, corrected text or enhanced text*.

**NUMERICAL OR STATISTICAL INFORMATION**

In humanities crowd-sourcing, the line between text and statistical/numerical information is not always clear, not only because source documents for statistical/numerical information frequently contain text as well, but also in terms of the organisation and operation of a project. Old Weather is an example of this. The logbooks contain both numerical information (e.g. weather observations) and free text, and the stated aim of the project was to extract the weather data for subsequent machine processing, yet much textual data has also been transcribed at the same time, and used for synthesis and interpretation of the histories of individual ships.

**SOUND**
The principal application to date of crowd-sourcing involving this asset type has been the gathering of audio recordings for various purposes. The British Library’s Sound Map, for example, has gathered examples of regional accents and dialects, wildlife sounds, environmental sounds, and other ‘soundscapes’ from around the UK (http://sounds.bl.uk/sound-maps/).

An emergent aspect of this area is the use of software tools to correct transcriptions of audio recordings. Automated transcription of recorded speech is currently highly error-prone, with error rates of 30% or more (Wald 2011). Application of crowd-sourcing techniques for processes of correcting/modifying content of recorded speech via transcription is likely to be very useful for a variety of humanities-oriented projects in the future.

**IMAGE**

Images are important asset type for humanities crowd-sourcing, and most current applications are concerned with the production of metadata for the purposes of enhancing search and discovery.

One key example is the Flickr Commons project, a collaboration between Flickr and the Library of Congress (Springer et. al. 2008) launched in 2008, which allows participating cultural heritage institutions to expose their images (if they have “no known copyright restrictions”) to the Web via Flickr, and users of Flickr then to tag them. This can improve access to and knowledge of these collections, and also allows the public to contribute aggregated knowledge. This can have unexpected extrinsic benefits; for example, in our interview with Contributor C, it emerged that he had tagged photos from the Netherlands National Library on Flickr Commons as part of his efforts to learn Dutch.

Another example is the YourPaintings project, developed by the Public Catalogue Foundation for the BBC (http://www.bbc.co.uk/arts/yourpaintings). This project has put online images of 200,000 oil paintings in public ownership in the UK, and exposed this super-collection to the public for them to tag, again to increase searchability and accessibility.

It should be noted that the ubiquity of the Web, and access to content creation and digitisation technologies, has led to the creation of non-professionally curated online archives, which are independent of conventional library and archive-derived curatorial narratives. These have a clear role to play in enriching and augmenting collections produced by ‘memory institutions’, and distributed via collaborative platforms such as Flickr Commons. As Melissa Terras has noted:

> Enthusiastic digitization by amateurs, a phenomenon previously ignored by information professionals, is providing a rich source of online cultural heritage content which often documents areas not covered via traditional institutions. Indeed, ephemera and popular culture materials are often better served by the pro-amateur community than memory institutions. The energy and zeal displayed by amateur digitizers is worthy of further consideration, as amateur collections often
complement existing collections, providing an alternative free discussion space for enthusiasts (Terras 2010).

VIDEO

Video as an asset type has to date demonstrated its potential for crowd-sourcing to a lesser degree than other kinds listed here.

However, YouTube has brought about a revolution in the communication of video media, and comments on YouTube videos provide an extensive and diverse source for the opinion and attitudes of (a subset of) the public. These comments have proved to be a rich seam for network analysis, where links between comments sharing the same characteristics can be mapped, and for sentiment analysis (Thelwall et al. 2012), where the positivity or negativity of comments’ sentiments can be quantitatively measured and analysed (op. cit.). It is also possible to discern the occurrence of discrete events depicted in the comment streams (Steiner et al. 2011).

In these examples the content created by the crowd provides raw material for the academic research, but its creation does not itself form part of an academic crowd-sourcing activity. However, it suggests that one avenue for video as an object of crowd-sourcing is afforded by the processes of (a) mapping and (b) commenting, critical responses and stating preferences, and that this will become particularly interesting when crowds interact around them.

Studies in this area have highlighted the positive connection between the attention the crowd gives to an individual user’s videos, and the productivity of that user, as measured by the numbers of videos they upload (Huberman et al. 2009). For such media to be of significant interest in humanities crowd-sourcing therefore, it is likely that they will have to be closely associated with human interaction, especially under the Commenting, critical responses and stating preferences process type.

EPHEMERA AND INTANGIBLE CULTURAL HERITAGE (E.G. ORAL HISTORY)

Ephemera and intangible cultural heritage form potentially the most productive category of asset for humanities crowd-sourcing, and yet also the most underdeveloped from a methodological point of view. Intangible cultural heritage covers any cultural manifestation that does not exist in tangible form; typically, crowd-sourcing is used to document such heritage through a set of processes and tasks, resulting in some form of tangible output. The importance of preserving intangible cultural heritage has been recognised by the UN (Kurin 2004), and the ways in which this can be documented and curated by distributed groups – indeed by ‘the crowd’ – is an important area for future research.

By ephemera, we understand cultural objects that are tangible, but are at risk of loss because of their transitory nature, for example home videos or personal photographs. There are a number of project addressing such assets, for example the Europeana 1914-1918 project.

8 This usage differs from the standard usage of the term by museums.
(http://www.europeana1914-1918.eu/en/contributor), which is collecting digitised personal artefacts relating to the First World War.

**Case study: the Scottish Words and Place-names (SWAP) Project**

The Scottish Words and Place-names project was a JISC-funded activity to gather words in Scots and, ultimately, to offer selected words for inclusion in the word collection of the Scottish Language Dictionaries resource. Scots is not a distinct language in the way that, for example, Gaelic is, but it has a community of people who believe strongly in its cultural and linguistic importance. The aim of the SWAP project was to see which Scots words are in current use, and where/how they are used.

The interface was relatively straightforward: a user could enter a word into a box on the website, and this was then harvested into a backend database. There was no moderation or checking done at this point, but words that the project felt were suitable were passed to the DSL (Dictionary of the Scots Language), where they were scrutinised by expert lexicographers.

Words were also gathered via Facebook and (to a lesser extent) Twitter. The Facebook page was an important venue for developing conversations around the forms and meanings of individual words. Contributions via Facebook were less formal and structured, taking the form of discussions with project staff in comment threads. This required more of an overhead in the time spent processing and extracting them. However, it generated valuable material, and allowed the project to target specific areas and questions that were of interest, such as local dialects and rhyming slang. For example:

“Thanks for all your place-name, alcohol and miscellaneous Scots info so far! This week we'd like to know about your local dialect rhyming slang. Some examples we already know about are: mammy mine - wine (Glasgow), Mick Jagger - lager (Gl), Lee Van (Cleef) - deif (Edinburgh), broon breid - deid (Ed & Gl), Mars bar - scar (Gl), Oscar (Slater) - later (Ed). Do you use these? Do you know any others? Please let us know!” (see http://www.facebook.com/scotswap/posts/230076623689593).

This generated responses such as:

“I remember Caroline Macafee's book on Glasgow: Language Varieties around the World collected some of these: askits (= shoulders) askit pooders, shouders); benny ( Lynch, cinch), etc. I never heard them used, though my godmother in Ayur [sic] still addresses me as 'china'.”

This highlights the importance of active engagement with (potentially small) interest communities for gathering contemporary linguistic content. In general, the project found that they received more useful information when they asked about specific words rather than framing open questions. For the same reason, acquiring information about specific
places was easier, as these are associated with a more concrete sense of belonging: ‘crowdsourcing informal place-names was relatively successful, perhaps because of a greater sense of involvement from participants, and social media does show good potential for further research of this sort’ (Hough et. al. 2011).

Information on Scottish place-names was also gathered by the project. Information about a Scots place-name element (e.g. 'liggat' meaning gate) would be provided from the Glossary of Scots Place-name Elements, with examples of place-names using this element, then other examples would be solicited from the public. These could then be added to the examples field of the Glossary (http://swap.nesc.gla.ac.uk/database).

The project was extremely successful in building up a community of followers and contributors among schoolchildren and teachers. This is largely because it was able to utilise and capitalise on the GLOW intranet network which connects Scottish schools. There was a slight potential ethical tension between the success of using Facebook to generate discussions, and the success of generating networks using GLOW: one is not allowed to have a Facebook account if under the age of 13. In June 2012 the project ran a competition for schoolchildren, with judges including Louise Welsh. This was again a highly successful way of engaging the community, and furthering the project’s aims of both capturing and encouraging the use of Scots.

For more information on the project, see the website at (http://swap.nesc.gla.ac.uk/), and especially the final JISC report (Hough et. al. 2011).

Source: interview with Ellen Bramwell and Jean Anderson, (14/09/2012)

7.4: Task type

A task is some particular activity that a project participant undertakes in order to create, process or modify an asset (usually a digital asset). Tasks can differ significantly as regards the extent to which they require initiative and/or independent analysis on the part of the participant, and the difficulty with which they can be quantified or documented. We have identified a number of task types with the aim of categorising this complexity.

These task types are identified in Table 3, and are ordered by the extent to which they require such independent analysis from the participant. Note that this refers only to what is required of the participant – a project requiring tasks that are in themselves mechanical will typically require a great deal of structure and planning on the part of the project team. Conversely, a task that is complex from the participant’s point of view may require less structure and input from the project team. Such a categorisation is not, of course, any reflection of the task’s value.
Superficially similar tasks can correspond to quite different task types, depending on the nature of the asset involved, and the requirements of the project; for example, the tasks involved in transcribing a section of text could be either *mechanical* or *editorial*.

<table>
<thead>
<tr>
<th>TASK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mechanical</td>
</tr>
<tr>
<td>Configurational</td>
</tr>
<tr>
<td>Editorial</td>
</tr>
<tr>
<td>Synthetic</td>
</tr>
<tr>
<td>Investigative</td>
</tr>
<tr>
<td>Creative</td>
</tr>
</tbody>
</table>

**Table 3: Task Types**

**MECHANICAL**

Processing discrete and granular units of information (typically quantitative/numerical/statistical information, or very short pieces of text) from one form into another, for example transcribing ages and dates of birth from birth certificates. Such tasks require little or no initiative. The FamilySearch site provides an example of this type of task; users simply transcribe numerical data from images of records into a web form (see [https://familysearch.org](https://familysearch.org)).

**CONFIGURATIONAL**

This category covers tasks that involve identifying structural patterns or ‘configurations’ in information, rather than processing individual pieces of information. Some such tasks will require a predisposition for working with quantitative data. The Bostonography project provides examples of such tasks (see [http://googlemapsmania.blogspot.co.uk/2012/07/crowdsourcing-neighborhood-boundaries.html?spref=tw](http://googlemapsmania.blogspot.co.uk/2012/07/crowdsourcing-neighborhood-boundaries.html?spref=tw)).

**EDITORIAL**

Modifying, or improving an existing asset or assets. This requires initiative, depending on the nature of the task and on the standards or procedures laid down by the project or activity. Feedback and interaction with peers is typically required, and also acts as a significant encouraging factor. Wikipedia and Wikisource are examples of editorial crowd-sourcing tasks.

**SYNTHETIC**

Synthesising information from different sources. This requires initiative and analytical ability, and is also likely to require interaction with peers. Carrying out tasks of this type may also lead to the development of expertise in some area of the humanities. Bringing together information on individual ships (e.g. information obtained in the OldWeather project) for the Naval History site ([http://www.naval-history.net](http://www.naval-history.net)) is an example of a synthetic task.
INVESTIGATIVE

Includes elements of the configuration and synthetic types, but more specifically involves seeking out information that is latent or implicit in corpora or datasets, but which may not be immediately apparent. The Guardian's 2009 experiment in crowd-sourcing documents relevant to the UK MPs’ expenses scandal provides examples of this, where members of the public identified documents in the corpus that were interesting enough to be investigated further (see http://www.niemanlab.org/2009/06/four-crowdsourcing-lessons-from-the-guardians-spectacular-expenses-scandal-experiment).

CREATIVE

Creation of new content, as for example in the Strandlines project (www.strandlines.net).

6.5: Output type

The output type refers to the thing an activity produces as the result of the application of a process, using tasks of a particular task type, to an asset. These outputs can be tangible and/or measurable in various ways, but we make allowance here also for intangible outcomes, such as awareness, knowledge etc. Our interviews and discussions with super-contributors indicated that these can be particularly important for the participants, although of course they do not lend themselves to quantification or measurement.

ORIGINAL TEXT

Text that is created as the result of a project or activity, and did not exist prior to the project or activity.

TRANSCRIBED TEXT

Text, almost always digital and machine-readable, that is created by processing other text that is not machine-readable. The transcribed text will have little or no semantic enhancement.

CORRECTED TEXT

Text that has been modified only to the extent of correcting errors in the input asset by manual intervention. Typically, this will involve a combination of mechanical and editorial task types, in a process of type correcting/modifying content.

ENHANCED TEXT

Text that has been semantically enhanced, for example by marking it up using TEI/XML.
**TRANSCRIBED MUSIC**

Music (in the form of musical notation, not audio) that has been transcribed into digital form and made retrievable from a digital system.

**METADATA**

Data about an asset, usually created to make it more discoverable, retrievable, curatable or usable, or generally to expose it via the Web more effectively. This category includes both descriptive metadata and preservation metadata, although most examples highlighted by our review are of the former.

**STRUCTURED DATA**

Data whose structure has been improved and/or made more explicit, for example by exposure to processes such as collaborative tagging or linking.

**KNOWLEDGE/AWARENESS**

Increased knowledge of a subject (including practical skills), or increased awareness of a project or topic, in the wider community.

**FUNDING**

Money that has been raised from open, distributed, voluntary sources. Frequently, such ‘crowd-funding’ also engenders knowledge/awareness, not only among the donors but also more generally through the ensuing publicity.

It has been shown that location is not always a significant factor in the ability of projects to attract money in this way (Agrawal et. al. 2011); however in the humanities, this phenomenon has been particularly associated with participation by funding in archaeological digs (see, e.g. [http://www.bbc.co.uk/news/science-environment-19192220](http://www.bbc.co.uk/news/science-environment-19192220)), and in such cases there is an obvious relationship between the participant and a particular location.

**SYNTHESIS**

A collection of pre-existing information that has been enriched and expanded by being combined, compared and developed through editorial or structural intervention.

**COMPOSITE DIGITAL COLLECTIONS WITH MULTIPLE MEANINGS**

The use of interoperable standards for markup, e.g. TEI (for text) and KML (for mapping), underpins many of the asset types identified here, and is essential for their subsequent searching and exploration. However, these standards are extrinsic to the crowd-sourcing
process, and have been separately developed for the purpose of organising and standardising data.

On the other hand, the crowd-sourced collection of ordered, but largely unstructured, information such as Wikipedia can also provide a means for navigating and describing content, in addition to the content itself. In Wikipedia’s case this is DBpedia (http://dbpedia.org/About), a linked data hub that identifies the *subjects* discussed in Wikipedia as unique entities, and allows these to be associated with external collections. Such complex information structures as the combined Wikipedia/DBpedia have their own value, and are thus defined here as a distinct asset type. In most cases, such outputs will include components of both *original/enhanced text* and *structured data*.

This type of output sits in the context of the collaborative creation of digital objects that the Web 2.0 internet encourages. For example, the British Museum’s Wikipedia project sought to assist the creation of articles dealing with all of the Library’s most significant objects and collections. The question of how objects should be described comes up again and again in the (digital) humanities, be it XML markup, textual chunks, museum schemata for objects such as the CIDOC CRM (http://www.cidoc-crm.org) etc. One key objective of humanities crowd-sourcing, although expressed differently by different projects, is therefore to leverage this increasingly sophisticated transformation of humanities content into digital objects and to promote public interaction with it.

**Case study: Wikisource**

Wikisource is a sister project of Wikipedia (http://wikisource.org/) based, like Wikipedia, on the MediaWiki platform. It aims to create a crowd-sourced, wiki-based library of digital texts. Scans of texts from Google Books and the Internet Archive are read by participants, and transcribed into machine-readable form, and edited by the community. There are no strictures on what text or texts members of the Wikisource community work on; thus the subject interest of individuals is easily catered for. There are relatively few ‘power users’ who do a great deal of work, and the Wikisource community in individual countries, such as Italy, is relatively small. A key source of traffic and interest is the fact that Wikisource resources are embedded in Wikipedia pages – e.g. the ‘rilettura del mese’ (rereading of the month) in the bottom left of the page at http://it.wikipedia.org/wiki/Pagina_principale – which allows the project to ‘exploit the power of the Wikipedia user base’.

A ranking is applied to each unedited page of each resource, with an orange square indicating that the page is formatted but not proofread, and a green one indicating that it is in a publishable condition, both corrected and validated. Wikisource provides a set of templates for structuring and ordering texts according to good editorial practice. It is noted however that it can be complex for users without programming knowledge or the ability to manipulate the code underlying the
templates. Although it has the same kind of cross-fertilising community functionality as the Zooniverse projects, it is not as easy to use as the Zooniverse interface.

Source: interview with Andrea Zanni, 22/10/2012

Such work is contextualised by the findings of Kittur et.al. (2007), who note that, while platforms such as Wikipedia generally start by being driven by large numbers of ‘super users’, both in terms of content created and of edits made, over time much of the content creation and modification is taken over by larger numbers of users with fewer time and effort commitments. This suggests that project-based funding models, which are inherently time limited, do not lend themselves to creating successful composite crowd-sourced archives of information.

8. Conclusion

The overall conclusion of this research review is that research involving humanities crowd-sourcing can be divided into the four facets of process, asset, task type and output type. Any robust set of replicable methodologies for creating or processing information by or for humanistic scholarship must be framed in these terms.

Depending on the project or activity, and what it aims to do, some categories, or indeed some facets, will have primacy. Outputs might be original knowledge, or they might be more ephemeral and difficult to identify: however, considering the processes of both knowledge and resource creation as comprising of these four facets gives a meaningful context to every piece of research, publication and activity we have uncovered in the course of this review. We hope the lessons and good practice we have identified here will, along with this typology, contribute to the development of new kinds of humanities crowd-sourcing in the future.

Most significantly, we have found that most humanities scholars who have used crowd-sourcing in its various forms now agree that it is not simply a form of cheap labour for the creation or digitization of content; indeed in a cost-benefit sense it does not always compare well with more conventional means of digitization and processing. In this sense, it has truly left its roots, as defined by Howe (2006) behind. The creativity, enthusiasm and alternative foci that communities outside that academy can bring to academic projects is a resource which is now ripe for tapping in to, and the examples shown in this report illustrate the rich variety of forms that tapping can take.

Acknowledgements

We are grateful to all those who have shared their knowledge and experience with us: Rebekkah Abraham, Anne Alexander, Donna Alexander, Jean Anderson, Keith Ball, Richard Blakemore, Jeremy Boggs, Ellen Bramwell, Phili Brohan, Geoff Browell, Ben Brumfield, Tim Causer, Sarah Cornell, David De Roure, Alison Dickens, Alexandra Eveleigh, Chris Fleet, Abigail Gilmore, Andrew Gray, Colin Greenstreet, Martin Holmes, Helen Julian, Kimberly Kowal, Sam Leon, Janet Lomas, Susan Major, Paola Marchionni, Anthony Masinton, Fabrizio Nevola,
References cited


*JISC* (online at [http://www.jisc.ac.uk/media/documents/programmes/digitisation/swapfinalreport.pdf](http://www.jisc.ac.uk/media/documents/programmes/digitisation/swapfinalreport.pdf)).


Oomen, J. and Aroyo, L. 2011: Crowdsourcing in the cultural heritage domain: opportunities and challenges. 

Prestopnik, N. R., & Crowston K. 2011: Gaming for (citizen) science: Exploring motivation and data quality in the context of crowdsourced science through the design and evaluation of a social-computational system. 


Surowiecki, J. 2004: The Wisdom of Crowds: Why the Many are Smarter than the Few, Abacus 2004


Trant, J. 2009: Tagging, Folksonomy, and Art Museums: Results of steve.museum’s research. (online at http://conference.archimuse.com/blog/jtrant/stevemuseum_research_report_available_tagging_f o)


Appendix A: Survey

The survey (approved by King’s College London Research Ethics Committee, project number REP-H/11/12-19; see Appendix A) was open to all, and was available on the project website at http://www.humanitiescrowds.org/survey from May 21st until October 30th 2012. Although the survey was publicised on various mailing lists in humanities communities, and was promoted to project participants by colleagues running humanities crowd-sourcing projects (especially those projects represented at the May workshop), it was completely self-selecting, and is therefore likely to have attracted only those with significant interests in crowd-sourcing, either ‘super contributors’ or people helping to run projects. It thus makes no claim to being statistically representative; it is purely a means of gathering further qualitative information about contributors’ backgrounds and motivations.

The survey was conducted online using SurveyMonkey, and included the following questions and possible responses. Where no response options are indicated, free text was allowed.

1. Are you male or female?
   Male, Female

2. Which category below includes your age?
   17 or younger, 18-25, 25-35, 35-45, 45-55, 55-60, 60+

3. Which of these best describes you?
   At school, Student, In work, Self-employed, Unemployed, Pensioner, Retired

4. Your nationality

5. Your current location (country, city)

6. Do you currently consider yourself to be an active contributor to a crowd-sourcing project or projects?

7. When did you last work on a crowd-sourcing project?
   In the last three days, In the last week, In the last month, In the last year

8. Which crowd-sourcing project(s) have you worked with?

9. How did you find out about the project(s) you have worked with?

10. What is (or was) the nature of your work with the project(s)?
11. Roughly how many hours a week do you (or did you) spend contributing?

Less than 1 hour, 1 to 4 hours, 5 to 10 hours, 10 to 20 hours, 15 to 20 hours, More than 20 hours

12. Typically how long does a single session last?

Less than 1 hour, 1 to 3 hours, More than 3 hours

13. Do you see your work with the project(s) as a solitary activity, or do you discuss it with others?

14. Do you use any social networks to discuss, disseminate or help your crowd-sourcing work?

15. Please describe what first interested you in working with crowd-sourcing projects.

16. What motivates you to do crowd-sourcing work?

17. Do you see your contribution as being for your own interest, or for the benefit of others?

18. What would increase your motivation to keep doing crowd-sourcing work, or encourage you to do more?

19. What would cause you to stop contributing?

Completion of project, Lack of time, Loss of interest, Other (please specify)

20. Have you ever created or edited a Wikipedia article?

59 responses were received. 58% of respondents were male, 42% female. Most were in the 35-45 age bracket (see Fig. A-2). A significant majority (59.6%) were in work, with students (22.8%) and retirees (15.8%) being the next largest categories (see Fig. A-3). Unsurprisingly, most of the respondents (25) were from the UK, followed by Italy (11), the USA (9), Ireland (2), and individual participants from Malta, Greece, Germany, Canada, Spain and Latvia. 80% were currently active as crowd-sourcing participants, with most having been active within the week before they completed the survey (see Fig. A-4).
Figure A-1: Age of respondents

Figure A-2: Employment status of respondents
Figure A-3: Time since last crowd-sourcing work

Figure A-4: Nature of crowd-sourcing work undertaken
Respondents were asked (Question 10) what kind of crowd-sourcing work they had undertaken, with options drawn from the early stages of our desk research. These categories were subsequently refined, but at that stage they were: transcribing, classifying, proofreading, tagging and commenting. These were incorporated into the typology described in Section 7 under the process type facet, with some modifications, e.g. proofreading was subsumed into Correcting/modifying content (see, e.g. Newby and Franks 2003). We also allowed participants to specify other kinds of activities, which identified some important lacunae in the typology, such as georeferencing and creating, which were subsequently incorporated. Table A-1 lists these additional responses, along with the process type to which they correspond in the typology. In some cases the fit is better than in others, and in all cases the process type needs to be qualified by the asset and task type involved (see Section 7).

<table>
<thead>
<tr>
<th>Response to Question 10</th>
<th>process type (from typology)</th>
</tr>
</thead>
<tbody>
<tr>
<td>data gathering and management</td>
<td>Linking, categorising and cataloguing</td>
</tr>
<tr>
<td>Research, online and library work.</td>
<td>Linking, categorising and cataloguing</td>
</tr>
<tr>
<td>OCR error-correction</td>
<td>Correcting/modifying content</td>
</tr>
<tr>
<td>aligning/reorienting images</td>
<td>Correcting/modifying content</td>
</tr>
<tr>
<td>Translating</td>
<td>Translating</td>
</tr>
<tr>
<td>Contributing Content (wikis)</td>
<td>Recording and creating content</td>
</tr>
<tr>
<td>Adding photos to iSpot. Writing for wikibooks.</td>
<td>Recording and creating content, linking</td>
</tr>
<tr>
<td>Editing, adding information</td>
<td>Correcting/modifying content</td>
</tr>
<tr>
<td>Technical work (Javascript etc.)</td>
<td>Recording and creating content</td>
</tr>
<tr>
<td>Sysop</td>
<td>Recording and creating content</td>
</tr>
<tr>
<td>Translating captions on flickr images</td>
<td>Translating</td>
</tr>
<tr>
<td>Creating</td>
<td>Recording and creating content</td>
</tr>
<tr>
<td>developing the infrastructure (technical, policies, etc.) of the whole project</td>
<td>Recording and creating content</td>
</tr>
<tr>
<td>Georeferencing - placing historic maps over current maps</td>
<td>Georeferencing</td>
</tr>
<tr>
<td>Georeferencing</td>
<td>Georeferencing</td>
</tr>
<tr>
<td>Georeferencing (BL) and OCR correction (Dickens)</td>
<td>Georeferencing</td>
</tr>
</tbody>
</table>

Table A-1: Nature of crowd-sourcing work undertaken

A majority of respondents (52.4%) stated that they spent less than an hour on an individual crowd-sourcing session, and 42.9% saying they spent between one and three hours. Only 4.8% of respondents who answered this question said they spent more than three hours, which possibly reflects the fact that most of the respondents were in work, and thus probably had less free time (many, but not all, of the super-contributors at the second workshop were not currently working for one reason or another), although it may also be the case that some tasks are not sufficiently absorbing to attract effort for very extended periods of time.

The question ‘Do you see your work with the project(s) as a solitary activity, or do you discuss it with others?’ was left open, but it is important as a measure of the sense of community within
crowd-sourcing. 17 responses can be classified as indicating that crowd-sourcing is a ‘solitary’ activity, whereas 23 felt that it was not. Many responses however raised more complex issues and could not be classified so simply. Some indicated that they discussed the project with others, for example ‘Solitary effort, but contributing to a group end. Discussed with others (family / friends) as it was interesting activity, but did not discuss with others in crowdsourcing project.’ Others, reflecting comments made in the second workshop, highlighted the importance of project discussion fora: ‘[I had discussions] only on discussion pages of projects if have a problem and need a second opinion or help.’ There was also reference to a sense of community arising indirectly, rather than through direct contact, for example through the visible achievement of common aims: ‘It was solitary but being able to view the stats. showing other participants progress enhanced the teamwork aspect and also added an element of competition’.

The responses to Question 15, which asked what first interested the respondent in contributing to crowd-sourcing, were analysed by classifying them under a number of headings, as shown in Table A-2. This analysis confirms that subject interest plays the greatest part in recruiting super-contributors, and that other motivations are largely altruistic in character.

<table>
<thead>
<tr>
<th>Heading</th>
<th>No. of Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject interest</td>
<td>24</td>
</tr>
<tr>
<td>Helping others learn</td>
<td>3</td>
</tr>
<tr>
<td>Want to contribute to science</td>
<td>2</td>
</tr>
<tr>
<td>Want to know how crowd-sourcing works</td>
<td>2</td>
</tr>
<tr>
<td>Wanting to be involved in voluntary work</td>
<td>1</td>
</tr>
<tr>
<td>Novelty</td>
<td>1</td>
</tr>
<tr>
<td>Unused CPU power might as well be used</td>
<td>1</td>
</tr>
</tbody>
</table>

Table A-2: reason for initial interest in crowd-sourcing

Question 16 asked respondents directly what motivates them to participate in crowd-sourcing work; the responses are listed in Table A-3 (note that spelling, etc. are as in the original responses). The responses indicate (a) that subject interest is paramount, but (b) that despite this, multiple motivations are usually in play, often involving a desire to contribute to a larger whole.

- Continuing to make a contribution to science
- The possibility to use the data we create
- It is useful and valuable
- I believe that knowledge is part of the common heritage of humanity.
- Feel it’s worthwhile and it’s easy to do

See also (Causer and Wallace, 2012).
Small individual effort has significant group benefits. I was also interested in the subject matter (Astronomy, antique maps).

<table>
<thead>
<tr>
<th>giving back to community</th>
</tr>
</thead>
<tbody>
<tr>
<td>Herbaria makes available information for anyone with a interest that would otherwise be to difficult to get at on the same scale - so makes the 1000's of specimens stored by our predecessors potentially more useful than at any time in the last 100+ years... Old weather should improve climate modelling and so refine our understanding of the world's greatest challenge: Climate Change. Both projects also provide fascinating insights help to mainly late 19th century naturalists lives and OW that of the RN in the early 20th Century.</td>
</tr>
</tbody>
</table>

The intrinsic value of the project

<table>
<thead>
<tr>
<th>Social contact/interest/altruism</th>
</tr>
</thead>
<tbody>
<tr>
<td>I like the idea of contributing to something of value to the world</td>
</tr>
<tr>
<td>A desire to be useful, I suppose, and sometimes a wish to see from farther in (if not precisely the inner circle!) how a project has been organized and to learn more about the content.</td>
</tr>
<tr>
<td>I always do ones that are part of subjects that I am interested in, so I suppose its a combination of entertainment and learning. I gain both from doing these tasks. I've not done a human comp thing that doesn't cover those; however I might in the future - with the caveat: I don't think I'd ever do something like Mechanical Turk for example, but I would get involved in something like the google hunt for the aeroplane.</td>
</tr>
</tbody>
</table>

Boredom while working public service point

| working on such projects is an opportunity to actively participate in digital culture in a meaningful way |
| Largely selfish enjoyment. I enjoy identifying other peoples photos and enjoy adding mine and having them classified. |
| Being part of wild interesting projects |
| My cultural interests. |
| Helping to build a better world |
| I don't know, I feel better when I do it |
| Sense of completion. "Someone is wrong on the Internet!" phenomenon. |
| Share my local and global interests |
| Passion |
| I'm motivated by the Wikimedia Foundation purpose: make every single human being able to share the sum of all knowledge. |

| help people |
| Reasons above. They are still the same, plus, in a way, I gained competences which gave me a job in the field and now I want to stay in contact with that world, also for professional reasons. |

| The opportunity to make available interesting resources |
| Interest in subject. Making material more accessible for others. |
| Appreciation for the democratic approach; the challenge (like a crossword puzzle), I have the time, and the effort does have some reasonably useful results. I find I can do it, unlike crosswords. |

| Providing resources for public viewing, and the possibility of research developing due to greater accessibility of materials |
### Table A-3: reason for initial interest in crowd-sourcing

<table>
<thead>
<tr>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>the interesting subject matter. making stuff better</td>
</tr>
<tr>
<td>Contribution to a worth-while project; progress up series of ranks</td>
</tr>
<tr>
<td>1. The subject matter 2. Being part of team 3. Believing the deliverable is useful &amp; will be used</td>
</tr>
<tr>
<td>Mostly the subject - the maps themselves, and the enjoyment of looking at what has changed between old and modern maps.</td>
</tr>
<tr>
<td>having recently retired, I have spare time to fill in, and I have found crowd-sourcing work that I have an interest in to be not only stimulating but very enjoyable to do</td>
</tr>
<tr>
<td>Interest in Dickens and in maps, keen on computing work</td>
</tr>
</tbody>
</table>

Question 17 asked whether participants see their contributions as being for their own interest or for the benefit of others, again a question that is concerned with the importance of a sense of community in humanities crowd-sourcing. The responses were free text, and of respondents who provided answers, only two said definitively that it was for their own benefit, whereas six said that it was for the benefit of others. Overwhelmingly (30 responses), the respondents stated that it was a combination of both, thus confirming that the motivations described in Table A-3 are, in fact, an alignment of the communal and the personal.
### Appendix B: Crowd-sourcing projects and activities

<table>
<thead>
<tr>
<th>Project Name</th>
<th>URL</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>AddressingHistory</td>
<td><a href="http://addressinghistory.edina.ac.uk/">http://addressinghistory.edina.ac.uk/</a></td>
<td>project</td>
</tr>
<tr>
<td>BBC YourPaintings</td>
<td><a href="http://www.bbc.co.uk/arts/yourpaintings/">http://www.bbc.co.uk/arts/yourpaintings/</a></td>
<td>project</td>
</tr>
<tr>
<td>Billion Graves</td>
<td><a href="http://billiongraves.com/">http://billiongraves.com/</a></td>
<td>project</td>
</tr>
<tr>
<td>British Library Georeferencer</td>
<td><a href="http://www.bl.uk/maps/">http://www.bl.uk/maps/</a></td>
<td>project</td>
</tr>
<tr>
<td>British Library Pin-a-tale</td>
<td><a href="http://www.bl.uk/pin-a-tale/pin-a-tale-map.aspx">http://www.bl.uk/pin-a-tale/pin-a-tale-map.aspx</a></td>
<td>project</td>
</tr>
<tr>
<td>Citizen Archivist Dashboard</td>
<td><a href="http://www.archives.gov/citizen-archivist/">http://www.archives.gov/citizen-archivist/</a></td>
<td>project</td>
</tr>
<tr>
<td>Crowdsourcing Australian Climate Change</td>
<td><a href="http://roseholley.blogspot.co.uk/2012/02/crowdsourcing-australian-climate-change.html">http://roseholley.blogspot.co.uk/2012/02/crowdsourcing-australian-climate-change.html</a></td>
<td>blog</td>
</tr>
<tr>
<td>Crowdsourcing Cultural Heritage: The Objectives Are Upside Down</td>
<td><a href="http://www.trevorowens.org/2012/03/crowdsourcing-cultural-heritage-the-objectives-are-upside-down/">http://www.trevorowens.org/2012/03/crowdsourcing-cultural-heritage-the-objectives-are-upside-down/</a></td>
<td>blog</td>
</tr>
<tr>
<td>Crowdsourcing Neighborhood Boundaries</td>
<td><a href="http://googlemapsmania.blogspot.co.uk/2012/07/crowdsourcing-neighborhood-boundsaries.html?spref-tw">http://googlemapsmania.blogspot.co.uk/2012/07/crowdsourcing-neighborhood-boundsaries.html?spref-tw</a></td>
<td>project</td>
</tr>
<tr>
<td>Crowdsourcing the AAM Annual Meeting</td>
<td><a href="http://futureofmuseums.blogspot.co.uk/2012/07/crowdsourcing-aam-annual-meeting.html">http://futureofmuseums.blogspot.co.uk/2012/07/crowdsourcing-aam-annual-meeting.html</a></td>
<td>blog</td>
</tr>
<tr>
<td>Crowdsourcing: Now With a Real Business Model!</td>
<td><a href="http://www.wired.com/business/2008/12/crowdsourcing-n/">http://www.wired.com/business/2008/12/crowdsourcing-n/</a></td>
<td>blog</td>
</tr>
<tr>
<td>DigitalKoot</td>
<td><a href="http://www.digitalkoot.fi/en/splash">http://www.digitalkoot.fi/en/splash</a></td>
<td>project</td>
</tr>
<tr>
<td>eBird</td>
<td><a href="http://ebird.org/content/ebird/">http://ebird.org/content/ebird/</a></td>
<td>project</td>
</tr>
<tr>
<td>Editathon, British Library</td>
<td><a href="http://uk.wikimedia.org/wiki/Editathon,British">http://uk.wikimedia.org/wiki/Editathon,British</a> Library</td>
<td>project</td>
</tr>
<tr>
<td>Editor’s Choice: Crowdsourcing and Cultural Heritage Round-up</td>
<td><a href="http://digitalhumanitiesnow.org/2012/06/editors-choice-crowdsourcing-and-cultural-heritage-round-up/">http://digitalhumanitiesnow.org/2012/06/editors-choice-crowdsourcing-and-cultural-heritage-round-up/</a></td>
<td>blog</td>
</tr>
<tr>
<td>Even Crowdsourcing Can Get Too Expensive</td>
<td><a href="http://scholarlykitchen.sspnet.org/2011/03/14/even-crowdsourcing-can-get-too-expensive/">http://scholarlykitchen.sspnet.org/2011/03/14/even-crowdsourcing-can-get-too-expensive/</a></td>
<td>blog</td>
</tr>
<tr>
<td>First World War Poetry Digital Archive</td>
<td><a href="http://www.oucs.ox.ac.uk/ww1lit/">http://www.oucs.ox.ac.uk/ww1lit/</a></td>
<td>project</td>
</tr>
<tr>
<td>Project Name</td>
<td>Website URL</td>
<td>Source Type</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Frequently Asked Questions about crowdsourcing in cultural heritage</td>
<td><a href="http://openobjects.blogspot.co.uk/2012/06/frequently-asked-questions-about.html">http://openobjects.blogspot.co.uk/2012/06/frequently-asked-questions-about.html</a></td>
<td>blog</td>
</tr>
<tr>
<td>Galaxy Zoo</td>
<td><a href="http://www.galaxyzoo.org/">http://www.galaxyzoo.org/</a></td>
<td>project</td>
</tr>
<tr>
<td>Gravestone Photographic Resource</td>
<td><a href="http://www.gravestonephotos.com/">http://www.gravestonephotos.com/</a></td>
<td>project</td>
</tr>
<tr>
<td>HistoryPin</td>
<td><a href="http://www.historypin.com/">http://www.historypin.com/</a></td>
<td>project</td>
</tr>
<tr>
<td>I'm not an experience-seeking user, I'm a meaning-seeking human person</td>
<td><a href="http://blog.tommorris.org/post/3216687621/im-not-an-experience-seeking-user-im-a">http://blog.tommorris.org/post/3216687621/im-not-an-experience-seeking-user-im-a</a></td>
<td>blog</td>
</tr>
<tr>
<td>iBayDigital</td>
<td><a href="http://www.ibaydigital.co.uk/btbond/">http://www.ibaydigital.co.uk/btbond/</a></td>
<td>project (concluded)</td>
</tr>
<tr>
<td>JISC and Crowdfunding</td>
<td><a href="http://www.jisc.ac.uk/blog/crowd/?utm_source=feedburner&amp;utm_medium=feed&amp;utm_campaign=Feed%3A+JISCBlog+%28JISC+Blog%29">http://www.jisc.ac.uk/blog/crowd/?utm_source=feedburner&amp;utm_medium=feed&amp;utm_campaign=Feed%3A+JISCBlog+%28JISC+Blog%29</a></td>
<td>blog</td>
</tr>
<tr>
<td>London riots app released to track down suspects</td>
<td><a href="http://www.london24.com/news/crime/london_riots_app_released_to_track_down_suspects_1_1421667">http://www.london24.com/news/crime/london_riots_app_released_to_track_down_suspects_1_1421667</a></td>
<td>news story</td>
</tr>
<tr>
<td>Meanification and crowdscaffolding: forget badges</td>
<td><a href="http://www.playthepast.org/?p=1027">http://www.playthepast.org/?p=1027</a></td>
<td>blog</td>
</tr>
<tr>
<td>Old Reel project</td>
<td><a href="http://www.regaltenbury.org.uk/memory-reel/">http://www.regaltenbury.org.uk/memory-reel/</a></td>
<td>project</td>
</tr>
<tr>
<td>Old Weather</td>
<td><a href="http://oldweather.org/">http://oldweather.org/</a></td>
<td>project</td>
</tr>
<tr>
<td>On Making, Use and Reuse in Digital Humanities</td>
<td><a href="http://melissaterras.blogspot.co.uk/2012/03/on-making-use-and-reuse-in-digital.html">http://melissaterras.blogspot.co.uk/2012/03/on-making-use-and-reuse-in-digital.html</a></td>
<td>blog</td>
</tr>
<tr>
<td>OurWikiBooks</td>
<td><a href="http://www.jisc.ac.uk/whatwedo/programmes/digitisation/communitycontent/ourwikibooks.aspx">http://www.jisc.ac.uk/whatwedo/programmes/digitisation/communitycontent/ourwikibooks.aspx</a></td>
<td>project</td>
</tr>
<tr>
<td>Portable Antiquities Scheme</td>
<td><a href="http://finds.org.uk/">http://finds.org.uk/</a></td>
<td>project</td>
</tr>
<tr>
<td>Prism project</td>
<td><a href="http://prism.scholarslab.org/">http://prism.scholarslab.org/</a></td>
<td>project</td>
</tr>
<tr>
<td>Project</td>
<td>URL</td>
<td>Type</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-------------------------------------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>Qurator</td>
<td><a href="http://www.qrator.org/">http://www.qrator.org/</a></td>
<td>project</td>
</tr>
<tr>
<td>Scripto</td>
<td><a href="http://scripto.org/">http://scripto.org/</a></td>
<td>tool</td>
</tr>
<tr>
<td>The Crowd and the Library</td>
<td><a href="http://www.trevorowens.org/2012/05/the-crowd-and-the-library/">http://www.trevorowens.org/2012/05/the-crowd-and-the-library/</a></td>
<td>blog</td>
</tr>
<tr>
<td>Transcribe Bentham</td>
<td><a href="http://www.ucl.ac.uk/transcribe-bentham/">http://www.ucl.ac.uk/transcribe-bentham/</a></td>
<td>project</td>
</tr>
<tr>
<td>Treading Water on Open Access</td>
<td><a href="http://www.dancohen.org/2012/09/25/treading-water-on-open-access/">http://www.dancohen.org/2012/09/25/treading-water-on-open-access/</a></td>
<td>blog</td>
</tr>
<tr>
<td>Ur Crowdsource</td>
<td><a href="http://urcrowdsource.org/omeka/">http://urcrowdsource.org/omeka/</a></td>
<td>project</td>
</tr>
<tr>
<td>Various resources</td>
<td><a href="http://www.delicious.com/stacking/crowdsourcing">http://www.delicious.com/stacking/crowdsourcing</a></td>
<td>links list</td>
</tr>
<tr>
<td>What’s the Score at the Bodleian</td>
<td><a href="http://www.whats-the-score.org/">http://www.whats-the-score.org/</a></td>
<td>project</td>
</tr>
<tr>
<td>Wikisource</td>
<td><a href="http://wikisource.org/wiki/Main_Page">http://wikisource.org/wiki/Main_Page</a></td>
<td>project</td>
</tr>
<tr>
<td>Year of Shakespeare</td>
<td><a href="http://www.yearofshakespeare.com/">http://www.yearofshakespeare.com/</a></td>
<td>project</td>
</tr>
<tr>
<td>Stranlines</td>
<td><a href="http://www.strandlines.net">http://www.strandlines.net</a></td>
<td>project</td>
</tr>
</tbody>
</table>