Connected Communities

Crowd-Sourcing in the Humanities

A scoping study

Stuart Dunn, Mark Hedges

CONNECTED COMMUNITIES

Background

Executive Summary

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 the *Crowd-Sourcing Scoping Study* was to review crowd-sourcing practices in the academic humanities, to assess their impact and development, to consider the motivations and aspects of community among those who participate, and to present a typology that captures the various approaches that have emerged.

In this Discussion Paper, we focus on describing the typology developed by the study, with a view to stimulating discussion its effectiveness as a conceptual framework for describing, analysing and planning crowd-sourcing activities in the humanities, and as a focus for making proposals for future research suggested by the study as a whole. The other topics addressed by the study may be found in the full report, which may be found here.

Researchers and Project Partners

Mark Hedges, Stuart Dunn
Centre for e-Research, Department of Digital Humanities, King's College London

Project website
http://humanitiescrowds.org

The study would not have been possible without the participation of all those, whether academics or contributors to crowd-sourcing projects, who shared their knowledge and experience with us, and in particular those who agreed to be interviewed, or participated in the workshops, or provided feedback on the project report.

Particular thanks are due to the following:

Rebekkah Abraham
Anne Alexander
Donna Alexander
Jean Anderson
Keith Ball
Richard Blakemore
Jeremy Boggs
Ellen Bramwell
Phil 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
Maurice Nicholson
Bethany Nowviskie
Nicola Osborne
Daniel Pett
David Price
Mia Ridge
Anna-Maria Sichani
Nick Stanhope
Su Startin
Lea Stern
David Stuart
Erin Sullivan
Mike Thelwall
David Tomkins
Charlotte Tupman
Mark Van Harmelen
Valeria Vitale
Claire Warwick
Jill Wilcox
Stella Wisdom
Chris Woodings
Andrea Zanni

Key words

Crowd-sourcing
Citizen science
Typology
Motivation
Incentives

Methodology

The study's methodology had four main components: a literature review covering academic humanities research that has involved crowd-sourcing, as well as research into crowd-sourcing itself as a method, and less formal outputs such as blogs and project websites; two workshops held at King's College London in May and October 2012, facilitating discussion between, respectively, humanities academics who have used crowd-sourcing, and contributors to crowd-sourcing projects; a set of interviews with both academics and contributors; and an online survey of contributors exploring their backgrounds, histories, and motivations.

The study does not claim to be comprehensive: there are inevitably important projects, publications, individuals and activities that have been omitted, and there is a strong Anglophone focus on the activities studied. The projects investigated by the study, and details of the survey, may be found in the full report.

Typology of crowd-sourcing

Our literature review identified a number of proposed categorisations of various aspects of crowd-sourcing activities, as well as related concepts such as 'citizen science'. One of the major outcomes of the study was a typology for crowd-sourcing in the humanities, which brought together this earlier work with the experiences and processes uncovered during the study. This typology does not aim to provide an alternative set of categories specifically for the humanities, in competition with these other typologies; rather, the aim was to propose a model for describing and understanding crowd-sourcing projects in the humanities by analysing them in terms of four key 'primitive' facets – **asset type**, **process type**, **task type**, and **output type** – and of the relationships between them, and in particular by observing how the applicable categories in one facet are dependent on those in other facets.

Figure 1: Typology framework

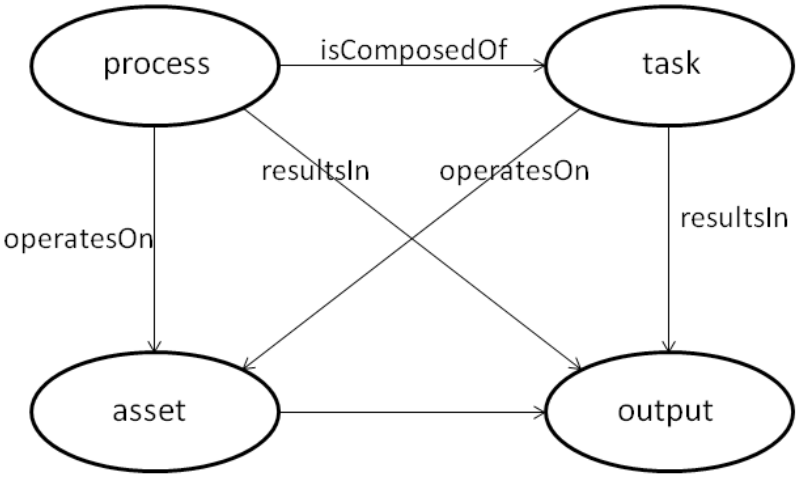

Figure 1 illustrates the four facets and their interactions.

- A *process* is composed of *tasks* through which an *output* is produced by operating on an *asset*. The most effective choice of process 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.
- An *Asset* refers to the content that is, in some way, transformed as a result of processing by a crowd-sourcing activity. It can be either tangible or intangible.
- A *task* is an 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. The *task types were* identified with the aim of categorising this complexity, and are listed below in approximately increasing order.
- The *output* is what is produced as the result of applying a process to an asset. Outputs can be tangible and/or measurable, but we make allowance also for intangible outcomes, such as awareness or knowledge etc.

Process Type
Collaborative tagging
Linking
Correcting/modifying content
Transcribing
Recording and creating content
Commenting, critical responses and stating preferences
Categorising
Cataloguing
Contextualisation
Mapping
Georeferencing
Translating

Task Type
Mechanical
Configurational
Editorial
Synthetic
Investigative
Creative

Asset Type
Geospatial
Text
Numerical or statistical information
Sound
Image
Video
Ephemera and intangible cultural heritage

Output Type
Original text
Transcribed text
Corrected text
Enhanced text
Transcribed music
Metadata
Structured data
Knowledge/awareness
Funding
Synthesis
Composite digital collections with multiple meanings

The tables above list the categories identified under each facet, based largely on an examination of existing crowd-sourcing practice, so it is to be expected that the lists will be extended and/or challenged by future work. Detailed descriptions may be found in (Dunn and Hedges 2012); for brevity, here we provide an overview focused around the process facet.

At the core of the typology is the *Process* facet, as processes, effectively, correspond to humanities research methods. In order to facilitate participation from distributed networks of people, these have to be articulated clearly, and made explicit, just as the methodologies are in any research project.

Collaborative tagging

Collaborative tagging may be regarded as crowd-sourcing the organisation of information. Tags can be based on controlled vocabularies, but are more usually derived from free text supplied by the users. Such 'folksonomies' are distinguished from deliberately designed knowledge organisation systems by the fact that they are self-organising, evolving and growing as contributors add new terms (Lin and Davies 2010).

Collaborative tagging can have 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. Golder (2006) found that patterns generated by collaborative tagging are, on the whole, extremely stable, meaning that minority opinions can be preserved alongside more mainstream ones. Other research (Trant 2009) has shown that user-assigned tags in museums may be quite different from 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.

An example is the BBC's YourPaintings project (*www.bbc.co.uk/arts/yourpaintings/*), 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 both improves discovery and enables the creation of an aggregation of specialised knowledge.

The Prism project (*www.scholarslab.org/category/praxis-program/*) provides a more complex example. Collaborative tagging typically assumes that the assets being tagged are themselves stable and clearly identifiable. Prism, however, allowed readers to select and tag sections of text at will, and thus build up a collective interpretation of the text.

Transcribing

Transcribing is currently one of the most prominent areas of humanities crowd-sourcing, as it can be used to address a fundamental problem with digitisation, namely the difficulty of rendering handwriting into machine-readable form using current technology. Typically, such transcription requires the human eye and, in many cases, human interpretation.

Two projects have contributed significantly to this prominence: *Old Weather* (Brohan et. al. 2009) and *Transcribe Bentham* (Causer *et al.* 2012; Causer and Wallace, 2012). OW involved the transcription of ships' log-books held by The National Archives, in order to obtain access to the weather observations they contain, information that is of major significance for climate research. TB encouraged volunteers to transcribe and engage with unpublished manuscripts by the philosopher and reformer Jeremy Bentham, by rendering them into text marked up using TEI XML.

The collaborative model needed for successful crowd-sourced transcription depends on the complexity of the source material. Complex material such as these requires a high level of support, whether from peers or the project team. Simpler material is likely to require less support; for example, when transcribing the more structured data found in family records (e.g. *www.familysearch.org*), the information can be presented to the user in small segments – e.g. names, dates, addresses – whose transcription requires different cognitive processes that are less dependent on interaction with peers and experts.

Note that this category includes marked-up transcriptions. There will be a point however at which the addition of mark-up will go beyond mere transcription, and will count as a form of *collaborative tagging* or *linking*.

Correcting/modifying content

Mass-digitisation technologies, such as Optical Character Recognition (OCR), can be error-prone, and any such enterprise needs to factor in quality control and error correction, which can potentially make use of crowd-sourcing.

The TROVE project, which produced OCR-ed scans of newspapers from the Australian National Archives, is an excellent example of this (Holley 2009; 2010). The volume of digitised material precluded the corrections being undertaken by the Archive's its own staff, and using uncorrected text would have significantly reduced the benefits of digitisation, as search capability would have been very restricted.

Another potential application is correcting automated transcriptions of recorded speech,

which is currently highly error-prone, with error rates of 30% or more (Wald 2011).

Linking

Linking covers the identification and documentation of relationships (usually typed) between individual assets. Most commonly, this takes the form of linking via semantic tags, where the tags describe binary relationships, in which case it is analogous to collaborative tagging. In principle, this could also include the identification of n-ary relationships.

Recording and creating content

These processes frequently deal with intangible cultural heritage, which covers any cultural manifestation that does not exist in tangible form. The importance of such heritage has been recognised by the UN (Kurin 2004), and typically, crowd-sourcing is used to document and preserve it in a tangible output.

This frequently takes the form of a cultural institution soliciting memories from the communities it serves, for example the Tenbury Wells Regal Cinema's Memory Reel project (*www.regaltenbury.org.uk/memory-reel/*). Another example is the Scottish Words and Place-names project (*http://swap.nesc.gla.ac.uk/*; Hough et. al. 2011), which gathered words in Scots, determining which words were in current use and where/how they were used. Such processes can incorporate a form of editorial control or post hoc curation, and their outputs can be edited into more formal publications.

Similar processes can address ephemera, here understood as cultural objects that are tangible, but are at risk of loss because of their transitory nature, for example personal artefacts such as photographs. An example is the Europeana 1914-1918 project (*www.europeana1914-1918.eu/en/contributor*).

The ubiquity of the Web, and access to content creation and digitisation technologies, has led to the creation of non-professionally curated online archives. These have a clear role to play in enriching, augmenting and complementing collections in memory institutions, and in developing curatorial narratives independent of those of professionals (Terras 2010).

Commenting, critical responses and stating preferences

These processes are likely to count as crowd-sourcing only if there is some specific purpose around which people come together. For example, the *Shakespeare's Global Communities* project (*www.yearofshakespeare.com*) captured responses to the 2012 World Shakespeare Festival, to investigate how social networking is reshaping the ways in which diverse global audiences for such a figure connect with one another. The question provides a focus for the activity, which results in a dataset for addressing questions on the modern reception of Shakespeare.

Appropriately managed blogs can provide a platform for focused scholarly interactions of this type. For example, a review of *King Lear* on the Year of Shakespeare site led to an exchange about critical methods as well as content (*http://bloggingshakespeare.com/year-of-shakespeare-king-lear-at-the-almeida*). What differentiates this from blogging in general is the scholarly context provided by the project, and its proactive directing of content creation, which provides a link between the crowd and the subject.

Categorising

Categorising involves assigning assets to predefined categories; it differs from collaborative tagging in that the latter is unconstrained.

Cataloguing

Cataloguing – or the creation of structured, descriptive metadata – is a more open-ended process than categorising, but is nevertheless constrained to following accepted metadata standards and approaches. It frequently includes categorising as a sub-activity, e.g. by LoC subject headings.

Cataloguing is a resource-intensive process for many GLAM institutions, and crowd-sourcing has been explored as a means of addressing this. For example, the What's the Score project at the Bodleian investigated a cost-effective approach to increasing access to their music scores through a combination of rapid digitisation and crowd-sourcing descriptive metadata (*www.whats-the-score.org*, *http://scores.bodleian.ox.ac.uk*).

Contextualising

Contextualising is typically a more broadly-conceived activity than the related process types of cataloguing or linking, and it involves enriching an asset by adding to it or associating with it other relevant information or content.

Georeferencing

Georeferencing is the process of establishing the location of un-referenced geographical information in terms of a modern coordinate system such as latitude and longitude, thus enriching assets that contain such information, e.g. maps, gazetteers or travelogues.

An example is the British Library Georeferencer project, which aimed to 'geo-enable' historical maps in its collections by asking participants to assign spatial coordinates to digitised map images. Once georeferenced, the digitised maps are searchable geographically (Fleet et. al. forthcoming).

Mapping

Mapping (in the sense of this typology) refers to the process of creating a spatial representation of some information asset(s). This could involve the creation of map data from scratch, but could also be applied to the spatial mapping of concepts, as in a 'mind map'. The precise sense will depend on the asset type to which mapping is being applied.

The global growth in ownership of hand-held devices with GPS capabilities (Goodchild 2007) has led to the emergence of community-based mapping resources such as Open Street Map (*www.openstreetmap.org/*). Although there has been much discussion about the reliability of such resources, in contrast to those created by expert organisations, it has been found that Open Street Map in particular is extremely reliable (Haklay and Weber 2008, Haklay 2010).

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 its potential for accommodating multiple, diverse, contributions and interpretations, holds great potential for such material (see Fink 2011, Graham 2010).

Translating

Typically, a crowd-sourced translation from one language to another will require a strongly collaborative element if it is to be successful, given the semantic interdependencies that can occur between different parts of a text. However, in cases where the text can be broken up naturally into smaller pieces, a more independent mode of work may be possible. For example, Suda On-Line (*www.stoa.org/sol/*) is translating the entries in a 10th Century Byzantine lexicon/encyclopaedia. A more modern, although non-academic, example is the phenomenon of 'fansubbing', where enthusiasts provide subtitles for television shows and films (Cintas 2006).

Recommendations for Future Research

This typology, especially the typology of processes, will develop and evolve as the field of humanities crowd-sourcing itself evolves. However, even at this early stage, we suggest that it is stable enough for funders such as the AHRC to adopt it as a framework of advice for projects involving crowd-sourcing.

More research is needed in the field of crowd-sourcing itself. While a reactive research review such as this can highlight significant questions and issues, there is a need for proactive, systematic experiments to test fully the potential for crowd-sourcing in the humanities, and to explore how the processes identified here can be developed.

Specific research questions include:

- What kinds of communities form in the course of humanities crowd-sourcing projects, and how do these communities interact and change over time?
- How is engagement with humanities crowd-sourcing projects motivated and rewarded, both for participants and academics?
- How does crowd-sourcing stimulate broader circulation of information and knowledge, and how can it make the exchange of information and knowledge more democratic?
- What issues of trust and provenance are raised by the use of crowd-sourced information? How do these affect traditional models of peer review or research assessment?
- How can evolving humanities publication models, involving digital outputs beyond the journal article/monograph model, address publication of crowd-sourced outputs? What issues of citation, accreditation and IPR are raised, and how should they be dealt with?

References and external links

A full list of all the publications and projects referenced during the study may be found in (Dunn and Hedges, 2012).

Brohan, P., Allan, R., Freeman, J. E., Waple, A. M., Wheeler, D. Wilkinson, C. Woodruff, S. 2009: Marine observations of old weather. *Bulletin of the American Meteorological Society*, Vol 90, Issue 2, 219-230

Causer, T., Tonra, J. and Wallace, V. 2012: Transcription maximized; expense minimized? crowdsourcing and editing The Collected Works of Jeremy Bentham. *Literary and Linguistic Computing* Vol 27 Issue 2, 1-19

Causer, T. and Wallace, V. 2012: Building A Volunteer Community: Results and Findings from Transcribe Bentham. *Digital Humanities Quarterly*, Vol.6 Number 2. www.digitalhumanities.org/dhq/vol/6/2/000125/000125.html

Cintas, J.D. and Sanchez, P.M. 2006: Fansubs: Audiovisual Translation in an amateur Environment. *Journal of Specialised Translation*, 6, pp.37-52

Dunn S. and Hedges M, 2012: Crowd-Sourcing Scoping Study: Engaging the Crowd with Humanities Research, AHRC Report. *http://humanitiescrowds.org/wp-uploads/2012/12/Crowdsourcing-connected-communities.pdf*

Fink, C. 2011: Mapping Together : On collaborative implicit cartographies, their discourses and space construction. *Journal for Theoretical Cartography*, Vol 4. 1-14

Fleet, C., Kowal, K. C. and Pridal, P. forthcoming: Georeferencer – crowdsourced georeferencing for map library collections. Forthcoming in *D-Lib Magazine*

Golder, S. 2006: Usage patterns of collaborative tagging systems. *Journal of Information Science* Vol 32 Issue 2, 198-208

Goodchild, M. 2007: Editorial : Citizens as Voluntary Sensors : Spatial Data Infrastructure in the World of Web 2.0. *International Journal of Spatial Data Infrastructures Research*, Vol. 2. 24-32

Graham, M. 2010: Neogeography and the Palimpsests of Place: Web 2.0 and the Construction of a Virtual Earth. *Tijdschrift voor Economische en Sociale Geografie*, Vol 101, Issue 4. 422-436

Haklay, M. 2010: How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, Vol. 37, Issue 4. 682-703.

Haklay, M. and Weber, P. 2008: Openstreetmap: User-generated street maps. *Pervasive Computing, IEEE*, Viol 7, Issue 7. 12-18.

Holley, R. 2010: Crowdsourcing: How and Why Should Libraries Do It? *D-Lib Magazine* March/April 2010, Vol 16, Number 3/4 (online at *www.dlib.org/dlib/march10/holley/03holley.html*)

Holley, R. 2009: Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers. National Library of Australia 2009 (online at *www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf*)

Hough, C., Bramwell, E. and Grieve, D. 2011: Scots Words and Place-Names Final Report. *JISC* (online at *www.jisc.ac.uk/media/documents/programmes/digitisation/swapfinalreport.pdf*)

Kurin, R. 2004: Safeguarding Intangible Cultural Heritage in the 2003 UNESCO Convention: a critical appraisal. *Museum International*, Vol 56, Issue 1-2, 66–77

Lin, H. and Davis, J. 2010: Computational and Crowdsourcing Methods for Extracting Ontological Structure from Folksonomy. *The Semantic Web: Research and Applications, Lecture Notes in Computer Science*, 2010, Vol. 6089/2010, 472-477, DOI: 10.1007/978-3-642-13489-0_46.

Terras, M. 2010: Digital Curiosities: Resource Creation Via Amateur Digitisation. *Literary and Linguist Computing* 25 (4): 425-438. doi: 10.1093/llc/fqq019.

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_fo*)

Wald, M. 2011: Crowdsourcing correction of speech recognition captioning errors. Proceedings of the *International Cross-Disciplinary Conference on Web Accessibility – W4A '11* (online at *http://eprints.soton.ac.uk/272430/1/crowdsourcecaptioningw4allCRv2.pdf*).

The Connected Communities

Connected Communities is a cross-Council Programme being led by the AHRC in partnership with the EPSRC, ESRC, MRC and NERC and a range of external partners. The current vision for the Programme is:

"to mobilise the potential for increasingly inter-connected, culturally diverse, communities to enhance participation, prosperity, sustainability, health & well-being by better connecting research, stakeholders and communities."

Further details about the Programme can be found on the AHRC's Connected Communities web pages at:

www.ahrc.ac.uk/FundingOpportunities/Pages/connectedcommunities.aspx

www.connectedcommunities.ac.uk