Crowdsourcing & Digital Scholarly Editing

Joris van Zundert
Huygens Institute for the History of the Netherlands

It is disputable whether there is an accepted official or formal definition of crowdsourcing and what that definition should entail. Wikipedia attributes a definition to Daren C. Brabham stating that "Crowdsourcing is an online, distributed problem-solving and production model" (Brabham 2008). Coining the term crowdsourcing has been attributed to Jeff Howe (Howe, 2006).[1] The term was used to describe a trend resulting form the very poor results –due to cultural differences and language barriers– of outsourcing software and data production to cheap labour countries. Companies started to try to outsource repetitive and large volume tasks to volunteers or low payed contractors in a local context, thus increasing capacity and volume while actually reducing cost. The aspect of 'problem solving' is likely more related to the concept of 'wisdom of the crowd' than to crowdsourcing primarily. When utilizing wisdom of the crowd a specific problem is put to a general and large audience by crowd sourcing means. For instance: sort the countries of the European Union according to land mass. Given a large enough audience for voting and sorting this will yield the right answer. A key aspect here is the size of the audience. In wisdom of crowd games the active audience needs to be large enough to warrant a statistically viable median. This seems to differentiate 'wisdom of the crowds' considerably from 'crowdsourcing' as such. Where a production task is put to a indiscriminate audience, the size of the audience does not seem to be the most determining factor for a successful crowdsourcing project.

Indeed the considerable research that has been conducted over the last years on the phenomenon of crowdsourcing suggests that crowdsourcing projects share the characteristic of the actual labour or activity in the project being highly unevenly distributed. The curve describing activity per user in such cases is even exponential, and as a rule of thumb it can be taken that 10% of the labour force generates 90% of the output (Brumfield 2012a). The success of a crowdsourcing project is tied more to individuals maybe than to the crowd. As Rachel Stone puts it: "You need to find, among the vast number of vaguely interested, not very analytical people who look at web sites, the small number of tidy-minded obsessives who care deeply about [your data]" (Stone 2012). This seems true not only from the sheer perspective of labour resource, but also as to type of labour or role. A reflective study on semi-crowdsourcing with eLaborate[2] does also show the importance of an actively leading and guiding 'crowd master' (Beaulieu 2012).

[1] Google Books' NGram Viewer disagrees: it finds significant traces of the term at least as far back as 2003.

[2] A web based transcription environment developed and maintained by the Huygens Institute for the History of the Netherlands http://www.elaborate.huygens.knaw.nl

Another common aspect of the make up of 'crowds' is their relation to motivation and reward. Studies show that rewards or compensation are a non determining factor of crowdsourcing projects. In fact: simply paying for crowd sourced production seems to have adverse effects on quantity (Borst 2010, p.141) and quality (Organisciak 2010, Dobson 2012). The monetary aspect can actually profit the crowdsourcing party rather than the crowd: sometimes sources are put behind a pay wall and free access is allowed on the condition that a specific source is transcribed by the person requesting access (Ven 2010). But mostly payment compensation in any direction and magnitude defies the voluntary and playful nature of crowdsourcing (Organisciak 2010, pp.83–84). Rather contrary Borst suggests that the harder the challenge of the crowd sourced task, and the less the desire for compensation, the higher the output of a participant in a crowd sourced project. This ties in with findings that it are the intrinsic highly motivated people that drive production. Organiciak (2010) and Brumfield (2010) show that the most determining factors of a highly productive 'crowd sourcee' are the passion and interest in the subject and the importance felt about the subject or content.

As crowdsourcing production is driven by intrinsically motivated people it follows that one of the advantages of crowdsourcing may be found in its community forming power. Indeed experience at Huygens ING confirms this (Beaulieu 2012, Kets 2012). The eLaborate and Verwey-project cases of the Huygens ING show hybrid forms of applying crowd sourcing: the obvious benefits of a crowd (labour resource) is combined with a high care for quality. This results in community based outsourcing in which not anyone can access resources but only a selected community. This is motivated by the need to organize, control, correct and approve work by the 'crowd sourcees' to ensure scholarly quality. Beaulieu describes this as a positive development shying away from revolutionary tales and guiding new technologies towards a useful and critical application in humanities (Beaulieu 2012).

The argument of quality control is however often used to limit the opennes of crowd sourced projects. Scholars as supervisors of crowd sourced projects are caretakers of quality and thus are genuinely concerned that scholarly quality may falter under the effects of crowdsourcing (cf. Dunning 2011). Scholars seem to have little information to rely on when trying to establish the limits and practices of quality control in crowdsourcing environments: "despite the increased attention to crowdsourcing, so far little is known about how knowledge from the crowd is evaluated and the potential tension that issues of reliability or quality" (Prats Lopèz, 2013). Certainly however known and evaluated controls are available: from simple double keying to open ended community review (Brumfield, 2012b). That scholarly crowd sourced projects tend to opt for closed off communities notwithstanding may simply be motivated by the primary concern for quality. As Brokfeld (Brokfeld 2012, p.88) has shown and as common sense would confirm: crowd sourcing communities select the technology and project make up most suited to them. For scholarly projects this may justifiably mean fencing off the work force.

Public or community based, crowdsourcing does imply opening up at least partially the editorial process. Turning a previously often singular undertaking into a team or community based effort poses a number of questions to be considered on aspects of editorial organization and control. Patrick Sahle concludes that the role of the editor in crowd source projects indeed is changed from a sheer task of editing to providing structure, guidance, explicit quality controls, division of labour, the rational for practical editorial methodology, and even technical commodities as software and equipment (Sahle 2012, vol. 8, p235.). Another issue is that of ownership: if a crowd transcribes the sources of an edition, then who 'owns' that transcription? But even if ownership (e.g. in the case of public domain sources and works) is not a pivotal issue, crediting always is. Sufficient recognition of labour turns out to be essential in motivating a community (Borst 2010). The necessity to formally recognize and credit crowd sourcees' contributions is directly proportional to the complexity of the tasks outsourced.

Issues surrounding ownership of editions, the responsibility for the editorial work and methodology have direct implications for the openness of a (digital) edition. When Peter Robinson said "All readers may become editors too" he didn't refer to just a cheap labour force for transcribing sources, to be conveniently discarded at the moment a transcription phase is done (Robinson 2004). Instead, like Ray Siemens has proposed, he envisioned a 'social edition' that embodies the ideas of an open notebook science and renders all aspects of the editorial process –e.g. annotation, commenting, interpretation– open to public engagement (Siemens 2011). It is both challenging and exciting to think how far we can venture into opening up this process (and thus the digital edition) to informed communities and even public crowds. Scholarly editions provide much of the basic information and data on to which scholarly researchers of history and literature found their research. The analytic process involves in the case of literary analysis often painstakingly tracing names, annotating plot, clarifying meaning for instance. How much of this labour typically associated with high quality scholarly inference may actually be crowdsourced to the wisdom of the crowds? Current digital scholarly editions, seemingly inescapable and infinitely repeating bookish read only GUI metaphors, do not offer means to add information to the actual text by public users. As such they do more to hinder experiments with crowdsourcing than to support them. Harnessing the wisdom of crowds on our conceptions of narrative, of plot, of the development of characters at the very least will contrast and highlight our scholarly findings –more likely however they will reveal to us much more than we see now.

—JZ20130606

Bibliography

Beaulieu, A., Dalen-Oskam, van, K. & Zundert, van, J., 2012. Between Tradition and Web 2.0: eLaborate as a Social Experiment in Humanities Scholarship. In T. Takševa, ed. Social Software and the Evolution of

User Expertise: Future Trends in Knowledge Creation and Dissemination. IGI Global, pp. 112–129. doi: 10.4018/978-1-4666-2178-7.ch007.

Borst, I., 2010. Understanding Crowdsourcing: Effects of Motivation and Rewards on Participation and Performance in Voluntary Online Activities. PhD. Erasmus University Rotterdam. Available at: http://hdl.handle.net/1765/1 [Accessed June 4, 2013].

Brabham, Daren (2008), "Crowdsourcing as a Model for Problem Solving: An Introduction and Cases", Convergence: The International Journal of Research into New Media Technologies 14 (1): 75–90

Brokfeld, J., 2012. Die digitale Edition der „preußischen Zeitungsberichte“: Evaluation von Editionswerkzeugen zur nutzergenerierten Transkription handschriftlicher Quellen. Master. Potsdam: Fachhochschule Potsdam. Available at: http://opus4.kobv.de/opus4-fhpotsdam/files/331/masterarbeit_jbrokfeld.pdf [Accessed June 6, 2013].

Brumfield, B., 2010. NABPP Transcription User Survey Results. Collaborative Manuscript Transcription. Available at: http://manuscripttranscription.blogspot.nl/2010/12/nabpp-transcription-user-survey-results.html [Accessed June 6, 2013].

Brumfield, B., Klevan, D. & Vershbow, B., 2012a. Sharing Public History Work Using Crowdsourcing of both Data and Sources. Available at: http://www.imls.gov/about/webwise.aspx [Accessed June 6, 2013]. Also: Brumfield, B., 2012. Crowdsourcing at IMLS WebWise 2012. Collaborative Manuscript Transcription. Available at: http://manuscripttranscription.blogspot.nl/2012/03/crowdsourcing-at-imls-webwise-2012.html [Accessed June 6, 2013].

Brumfield, B., 2012b. Quality Control for Crowdsourced Transcription. Collaborative Manuscript Transcription. Available at: http://manuscripttranscription.blogspot.nl/2012/03/crowdsourcing-at-imls-webwise-2012.html [Accessed March 5, 2012].

Dobson, T. et al., 2012. Neither Bicycles Nor Sheep: Crowdsourcing Semantic Encoding for Elements of Plot. In Interedition: Scholarly Digital Editions, Tools and Infrastructure. The Hague: Huygens ING. Available at: http://www.interedition.eu/wp-content/bestanden/2012/03/7_3.pdf [Accessed June 6, 2013].

Dunn, S. & Hedges, M., 2012. Connected Communitie, Crowd-Sourcing in the Humanities: A scoping study, www.connectedcommunities.ac.uk. Available at: http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/12/Crowdsourcing-connected-communities.pdf.

Dunning, A., 2011. Crowdsourcing and Variant Digital Editions – some troubles ahead. JISC Digitisation and Content Programme. Available at: http://digitisation.jiscinvolve.org/wp/2011/07/18/crowdsourcing-and-variant-digital-editions-some-troubles-ahead/ [Accessed June 6, 2013].

Kets, A., 2013. Texts Worth Editing: Polyperspectival Corpora of Letters W. Mierlo, van, ed. Variants, 10, pp.93–103.

Organisciak, P., 2010. Why Bother? Examining the Motivations of Users in Large-Scale Crowd-Powered Online Initiatives. Master. Alberta: Univeristy of Alberta. Available at: https://era.library.ualberta.ca/public/datastream/get/uuid:af25a5dd-9da1-4536-9271-29946ab33b1c/DS1 [Accessed June 6, 2013].

Prats Lopèz, M. et al., 2013. Knowledge evaluation in organizations: A systematic review, VU Amsterdam. Available at: http://www.olkc2013.com/sites/www.olkc2013.com/files/downloads/140.pdf [Accessed June 6, 2012].

Robinson, P., 2004. Where We Are with Electronic Scholarly Editions, and Where We Want to Be. Available at: http://computerphilologie.uni-muenchen.de/jg03/robinson.html [Accessed June 6, 2013].

Sahle, P., 2013. Patrick Sahle: Digitale Editionsformen, Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels – Befunde, Theorie und Methodik, Norderstedt: Norderstedt: Books on Demand.

Siemens, Ray, Toward modeling the social edition: An approach to understanding the electronic scholarly edition in the context of new and emerging social media. Available at: http://llc.oxfordjournals.org/content/27/4/445.full [Accessed November 7, 2012].

Stone, R., 2012. What can the vulgus do? Crowd-sourcing for medievalists - Magistra et Mater. Magistra et Mater: Where history, religion and motherhood meet and have a long intellectual conversation. Available at: http://magistraetmater.blog.co.uk/2010/08/17/what-can-the-vulgus-do-crowd-sourcing-for-medievalists-9195007/ [Accessed June 6, 2013].

Ven, van der, C., 2010. De Digitale Archivaris: Crowdsourcen rond militieregisters. Available at: http://www.digitalearchivaris.nl/2010/07/crowdsourcen-rond-militieregisters.html#.UbBh4OubUUJ [Accessed June 6, 2013].